

Rough Table of Contents for notes that accompany the lectures

Topic	Pages
Simple linear regression (SLR) - heights example	1-9
Simple linear regression	10
Correlation	12-20
Probability Model for SLR	21
Multiple linear regression	33
Matrix formulation for MLR	36-40
Variable selection in MLR	41-52
Partial and seq. sums of squares	47-50
ANOVA as regression	69
General linear model	72
ANCOVA	82-88
Lack-of-fit	89-92
One-way ANOVA, contrasts	93-102
Multiple comparisons	103-109
Expected mean squares, part I	110-111
Sample size computation	112-116
Orthogonal polynomial contrasts	117-121
Multi-factorial expts	122
2×2 expts	122-133
More than 2 levels	134
More than 2 factors	141
Unbalanced example	149
Blocking	156-165
Latin squares	166-175
One-way random effects	176-189
Mixed models	190-235
Split-plot expts	236-255

ST512, Fall, 2009 - Syllabus

Instructor: Jason A. Osborne (osborne@stat.ncsu.edu)
SAS Hall, Room 5238, Phone number is 515-1922

Office Hours: To be determined

Course Website: www4.stat.ncsu.edu/~osborne/st512r/

Lecture: MWF: 9:10 a.m. - 10:00 a.m., 1216 SAS Hall

Labs : 1107 SAS Hall (SICL lab), Th 3:00-4:15pm

TAs: To be determined

Text: *Statistical Research Methods in the Life Sciences* by P.V. Rao, 1998, 2007 Brooks/Cole.

Lecture notes: Online or at SirSpeedy (834-8128)

Computing: The statistical software package SAS will be used extensively. The labs are intended in part to facilitate the learning of this package. (SAS is available free of charge for NCSU students, see course website for info.)

Course Description: ST512 is an applied course that introduces statistical methods based on linear models commonly used in designed experiments with continuous response variables. Examples include multiple linear regression, factorial designs, and split-plot experiments. It is a prerequisite for most advanced courses in statistics.

Topic	Chapter
Simple and multiple linear regression	10,11
General linear models, ANCOVA	12
Factorial experiments	13
Random and mixed effects models	14
Blocking: the RCBD and latin square	15
Repeated measures and split-plots	16

Policies of the instructor

Attendance: Regular attendance and oral participation are strongly encouraged. They will not be considered in the grading process.

Graded Coursework:

- Homework: 5 assignments, the avg. of which counts 25% of total grade
- Quiz 1, (Sept 21), 25% of total grade
- Quiz 2, (Oct 30), 25% of total grade
- Final exam, (Dec 11, 8:00 a.m. - 11:00 a.m) 20% of total grade
- Students achieving

≥ 90% of the total points will receive at least an A-

≥ 80% of the total points will receive at least a B-

≥ 70% of the total points will receive at least a C- .

Homework: Working together on homework is acceptable, but assignments must be submitted individually. If SAS printouts are included in work, output and page counts must be kept to a minimum, and pertinent elements of the output must be explained clearly.

Quizzes: Some quizzes may be in-class, some may be take-home, depending upon the discretion of the instructor. Notes may be written on a standard size sheet of paper, front and back, and used for in-class quizzes, a single sheet for each quiz. Aside from this allowance, in-class quizzes are closed book. No make-up quizzes will be allowed.

Academic Integrity: Academic misconduct, such as cheating on exams, will not be tolerated. Please see the NCSU policy at the link below.

[http://www.ncsu.edu/policies/student_services/
\(continued ...\)/student_discipline/POL11.35.1.php](http://www.ncsu.edu/policies/student_services/(continued...)/student_discipline/POL11.35.1.php)

Calendar for ST512, Spring, 2009 - Osborne

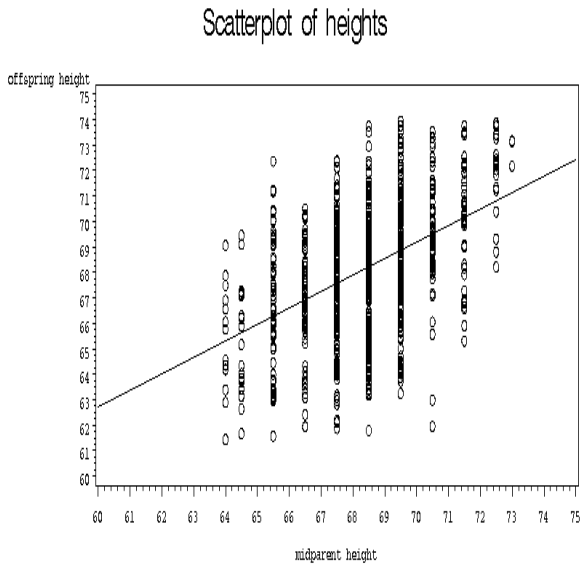
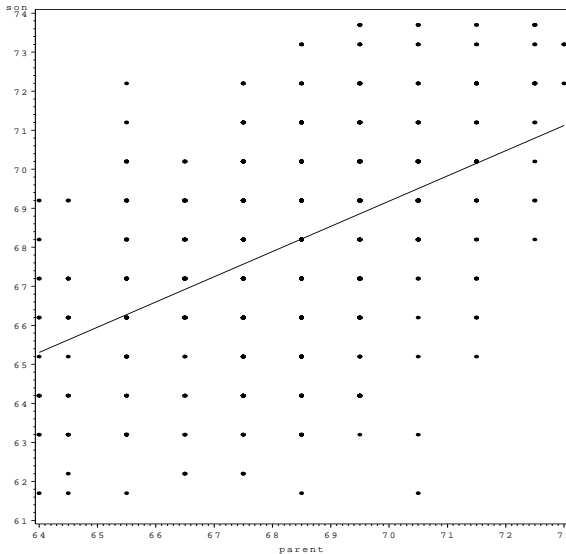
Monday	Tuesday	Wednesday	Thursday	Friday
		Aug 19 First Lecture	Aug 20 Lab 1	Aug 21
Aug 24	Aug 25	Aug 26	Aug 27 Lab 2	Aug 28
Aug 31	Sept 1	Sept 2	Sept 3 Lab 3	Sept 4
Sept 7 Labor Day	Sept 8	Sept 9	Sept 10 Lab 4	Sept 11
Sept 14 HW1 due	Sept 15	Sept 16	Sept 17 Lab 5	Sept 18
Sept 21 Quiz 1	Sept 22	Sept 23	Sept 24 Lab 6	Sept 25
Sept 28	Sept 29	Sept 30	Oct 1 Lab 7	Oct 2
Oct 5	Oct 6	Oct 7 HW2 due	Oct 8 Fall Break	Oct 9 Fall Break
Oct 12	Oct 13	Oct 14	Oct 15 Lab 8	Oct 16
Oct 19	Oct 20	Oct 21	Oct 22 Lab 9	Oct 23
Oct 28 HW3 due	Oct 27	Oct 28	Oct 29 Lab 10	Oct 30 Quiz 2
Nov 2	Nov 3	Nov 4	Nov 5 Lab 11	Nov 6
Nov 9	Nov 10	Nov 11	Nov 12 Lab 12	Nov 13
Nov 16	Nov 17	Nov 18	Nov 19 Lab 13	Nov 20
Nov 23 HW4 due	Nov 24 Thanksgiving	Nov 25 Thanksgiving	Nov 26 Thanksgiving	Nov 27 Thanksgiving
Nov 30	Dec 1	Dec 2	Dec 3 Lab 14	Dec 4 Classes end HW5 due
Dec 7	Dec 8	Dec 9	Dec 10	Dec 11 Final Exam

ST 512: Exptl Stats for Biol. Sciences II - Osborne

Week 1: Simple linear regression (MLR) - height example

Reading Chapter 10

The association between height of adults and their parents



```

/* -----
| Stigler , History of Statistics pg. 285 gives Galton's famous data |
| on heights of sons (columns,Y) and average parents' height (rows,X) |
| scaled to represent a male height (essentially sons' heights versus |
| fathers' heights). Taken from Dickey's website. |
\ -----*/

```

	61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	73.7
73.0	0	0	0	0	0	0	0	0	0	0	0	1	3	0
72.5	0	0	0	0	0	0	0	1	2	1	2	7	2	4
71.5	0	0	0	0	1	3	4	3	5	10	4	9	2	2
70.5	1	0	1	0	1	1	3	12	18	14	7	4	3	3
69.5	0	0	1	16	4	17	27	20	33	25	20	11	4	5
68.5	1	0	7	11	16	25	31	34	48	21	18	4	3	0
67.5	0	3	5	14	15	36	38	28	38	19	11	4	0	0
66.5	0	3	3	5	2	17	17	14	13	4	0	0	0	0
65.5	1	0	9	5	7	11	11	7	7	5	2	1	0	0
64.5	1	1	4	4	1	5	5	0	2	0	0	0	0	0
64.0	1	0	2	4	1	2	2	1	1	0	0	0	0	0

(The points in the scatterplot on the right have been *jittered* to convey the frequencies of heights in the dataset.)

Consider a *statistical model* for these data, randomly sampled from some population of interest. In particular, choose a model which accounts for the apparent linear dependence of the mean height of sons on midparent height X . Let Y_1, \dots, Y_n denote the sons' heights. Given $X = x_i$,

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad \text{for } i = 1, \dots, n (n = 928).$$

where E_1, \dots, E_n are

- independent,
- identically and
- normally distributed random variables with mean 0 and error variance σ^2 .

(Write $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$.)

This implies

1. $\mu(x) = E(Y|X = x) = \beta_0 + \beta_1 x$
2. $\text{Var}(Y|X = x) = \sigma^2$ (Three unknown parameters $\beta_0, \beta_1, \sigma^2$ quantify the whole population of interest.)

Question: Suppose we ignore midparent height x . Consider estimating the mean $E(Y)$. Propose a model. Propose a method for obtaining a confidence interval for the mean height of the sons in the population from which these data were randomly sampled. Use summary statistics on page 6 to complete this naive analysis.

Many questions to answer using regression analysis:

1. What is the meaning, in words, of β_1 ?
2. True/false: (a) β_1 is a statistic (b) β_1 is a parameter (c) β_1 is unknown.
3. What is the observed value of $\hat{\beta}_1$?
4. True/false: (a) $\hat{\beta}_1$ is a statistic (b) $\hat{\beta}_1$ is a parameter (c) $\hat{\beta}_1$ is unknown.
5. Is $\hat{\beta}_1 = \beta_1$?
6. How much does $\hat{\beta}_1$ vary about β_1 from sample to sample? (Provide an estimate of the standard error, as well as an expression indicating how it was computed.)
7. What is a region of plausible values for β_1 suggested by the data?
8. What is the line that best fits these data, using the criterion that smallest sum of squared residuals is “best?”
9. How much of the observed variation in the heights of sons (the y -axis) is explained by this “best” line?
10. What is the estimated average height of sons whose midparent height is $x = 68$?
11. Is this the true average height in the whole population of sons whose midparent height is $x = 68$?
12. Under the model, what is the true average height of sons with midparent height $x = 68$?

13. What is the estimated standard deviation among the population of sons whose parents have midparent height $x = 68$? Would you call this standard deviation a “standard error?”
14. What is the estimated standard deviation among the population of sons whose parents have midparent height $x = 72$? Bigger, smaller, or the same as that for $x = 68$? Is your answer obviously supported or refuted by inspection of the scatterplot?
15. What is the estimated standard error of the estimated average for sons with midparent height $x = 68$, $\hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1$? Provide an expression for this standard error.
16. Is the estimated standard error of $\hat{\mu}(72)$ bigger, smaller, or the same as that for $\hat{\mu}(68)$?
17. Is the observed linear association between son’s height and midparent height strong? Report a test statistic.
18. What quantity can you use to describe or characterize the linear association between height and midparent height in the whole population? Is this a parameter or a statistic?
19. Let Y denote the height of a male randomly sampled from this population and X his midparent height. Is it true that the population correlation coefficient ρ satisfies

$$\rho = E\left[\left(\frac{Y - \mu_Y}{\sigma_Y}\right) \times \left(\frac{X - \mu_X}{\sigma_X}\right)\right]?$$
20. Define $\mu_Y, \sigma_Y, \mu_X, \sigma_X, \rho$. Parameters or statistics?
21. What are plausible values for ρ suggested by the data?
22. Is $\boxed{E_1, \dots, E_{928} \stackrel{iid}{\sim} N(0, \sigma^2)}$ a reasonable assumption?

```

options ls=75 nodate;

data Galton;
array cdata(14);
if _n_ = 1 then input cdata1-cdata14 @ ;
retain cdata1-cdata14; drop cdata1-cdata14 i;
input parent @;
do i = 1 to 14; input count @ ; son=cdata(i);
output; end;
cards;
  61.7 62.2 63.2 64.2 65.2 66.2 67.2 68.2 69.2 70.2 71.2 72.2 73.2 73.7
73.0 0 0 0 0 0 0 0 0 0 0 1 3 0
72.5 0 0 0 0 0 0 1 2 1 2 7 2 4
71.5 0 0 0 0 1 3 4 3 5 10 4 9 2 2
70.5 1 0 1 0 1 1 3 12 18 14 7 4 3 3
69.5 0 0 1 16 4 17 27 20 33 25 20 11 4 5
68.5 1 0 7 11 16 25 31 34 48 21 18 4 3 0
67.5 0 3 5 14 15 36 38 28 38 19 11 4 0 0
66.5 0 3 3 5 2 17 17 14 13 4 0 0 0 0
65.5 1 0 9 5 7 11 11 7 7 5 2 1 0 0
64.5 1 1 4 4 1 5 5 0 2 0 0 0 0 0
64.0 1 0 2 4 1 2 2 1 1 0 0 0 0 0
;
proc print data=Galton(obs=100);
run;

data big; set galton; drop j count;
do j=1 to count;output; end;
proc print data=big(obs=20);

proc means; var son parent;

data questions;
/* these values used for prediction or estimation at x=68,x=72 */
input parent son;
cards;
68 .
72 .
;
run;

data big;
set big questions;
run;
proc reg;
model son=parent/clb;
output out=out1 residual=r p=yhat ucl=pihigh lcl=pilow uclm=cihigh lclm=cilow
stdp=stdmean;

data questions; set out1;
if son=.;

proc print;
title "questions regarding prediction, estimation when x=68, x=72";
run;

```

```

data fisherz;
  n=928;
  r=sqrt(0.2105);
  rratio=(1+r)/(1-r);
  z=probit(0.975);
  expon=probit(0.975)/sqrt(n-3);
  rlow=(rratio*exp(-2*expon)-1)/(rratio*exp(-2*expon)+1);
  rhigh=(rratio*exp(2*expon)-1)/(rratio*exp(2*expon)+1);
run;

proc print;run;

*goptions dev=pslepsf colors=(black);

symbol1 i=r1
  value=dot;

proc gplot;
  plot son*parent;
run;
quit;

/*
proc univariate data=out1 normal plot;
  title "residual analysis";
  var r;
run;
*/

```

The SAS System

1

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
son	928	68.0884698	2.5179414	61.7000000	73.7000000
parent	928	68.3081897	1.7873334	64.0000000	73.0000000

The REG Procedure
 Dependent Variable: son

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1236.93401	1236.93401	246.84	<.0001
Error	926	4640.27261	5.01109		
Corrected Total	927	5877.20663			

Root MSE	2.23855	R-Square	0.2105
Dependent Mean	68.08847	Adj R-Sq	0.2096
Coeff Var	3.28770		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	23.94153	2.81088	8.52	<.0001
parent	1	0.64629	0.04114	15.71	<.0001

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	18.42510	29.45796
parent	1	0.56556	0.72702

questions regarding prediction, estimation when x=68, x=72 3

Obs	parent	son	yhat	stdmean	cilow	cihigh	piLOW	pihigh	r
1	68	.	67.8893	0.07457	67.7429	68.0356	63.4936	72.2849	.
2	72	.	70.4745	0.16871	70.1434	70.8056	66.0688	74.8801	.

Obs	n	r	r ratio	z	expon	rLOW	rhigh
1	928	0.45880	2.69551	1.95996	0.064443	0.40645	0.50815

Answers to questions from simple linear regression
analysis of Galton's height data

1. Change in average son's height (inches) per one inch increase in midparent height (in the whole population.)
2. β_1 is an unknown parameter.
3. $\hat{\beta}_1 = 0.65$ son inches/midparent inch (from output.)
4. $\hat{\beta}_1 = 0.65$ is an observed value of a statistic.
5. β_1 is the slope of the population mean, $\hat{\beta}_1$ is the slope from the SLR of the observed data. $\hat{\beta}_1 = \beta_1$ is unlikely.
6. $\widehat{SE}(\hat{\beta}_1) = \sqrt{MS[E]/S_{xx}} = 0.04$ (from output.)
7. Add and subtract about 2 SE to get (0.57, 0.73)
8. $y = 23.9 + 0.65x$
9. $r^2 = 21\%$
10. $\hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1 = 67.9$ (from output also.)
11. Not sure, as $\mu(68) = \beta_0 + 68\beta_1$ is unknown.
12. $\mu(68) = \beta_0 + 68\beta_1$.

13. $\sqrt{MS[E]} = 2.24$. Not a SE.
 14. $\sqrt{MS[E]} = 2.24$. (Assume homoscedasticity.)
 15. $SE(\hat{\beta}_0 + 68\hat{\beta}_1) = 0.07$. Expressions given by

$$\begin{aligned}\widehat{SE}(\hat{\mu}(68)) &= \sqrt{MS[E] \left(\frac{1}{n} + \frac{(68 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)} \\ &= \sqrt{(1, 68)' MS[E] (X'X)^{-1} (1, 68)}\end{aligned}$$

X a (928×2) design matrix.

16. $\widehat{SE}(\hat{\mu}(72)) > \widehat{SE}(\hat{\mu}(68))$
 17. $r = \sqrt{0.21} = 0.46$, moderate, positive.
 18. ρ is a population correlation coefficient.
 19. True.
 20. These parameters describe the bivariate population of son and midparent heights.
 21. Using the complicated expression in Rao and in notes, the confidence interval is
- $$\left(\frac{\frac{1+r}{1-r} e^{-2z/\sqrt{n-3}} - 1}{\frac{1+r}{1-r} e^{-2z/\sqrt{n-3}} + 1}, \frac{\frac{1+r}{1-r} e^{2z/\sqrt{n-3}} - 1}{\frac{1+r}{1-r} e^{2z/\sqrt{n-3}} + 1} \right)$$
- or $0.41 < \rho < 0.51$.
 22. Residuals reasonably symmetric, no heavy tails.

ST 512: Exptl Stats for Biol. Sciences II**Week 2** Simple linear regression**Reading:** Ch. 10

An example: The association between corn yield and rainfall

Yields y (in bushels/acre) on corn raised in six midwestern states from 1890 to 1927 recorded with rainfall x (inches/yr).

y_1, \dots, y_{38} and x_1, \dots, x_{38} .

Year	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899
Yield	24.5	33.7	27.9	27.5	21.7	31.9	36.8	29.9	30.2	32
Rainfall	9.6	12.9	9.9	8.7	6.8	12.5	13	10.1	10.1	10.1
Year	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909
Yield	34	19.4	36	30.2	32.4	36.4	36.9	31.5	30.5	32.3
Rainfall	10.8	7.8	16.2	14.1	10.6	10	11.5	13.6	12.1	12
Year	1910	1911	1912	1913	1914	1915	1916	1917	1918	1919
Yield	34.9	30.1	36.9	26.8	30.5	33.3	29.7	35	29.9	35.2
Rainfall	9.3	7.7	11	6.9	9.5	16.5	9.3	9.4	8.7	9.5
Year	1920	1921	1922	1923	1924	1925	1926	1927		
Yield	38.3	35.2	35.5	36.7	26.8	38	31.7	32.6		
Rainfall	11.6	12.1	8	10.7	13.9	11.3	11.6	10.4		

A *scatterplot* provides some indication of an association between x and y . In particular, yields increase with rainfall.



Some questions:

- How can we describe the association between yield and rainfall? Does it appear *linear*?
- How can we measure the strength of the linear association?
- To what degree is the variability in yield described or explained by its association with rainfall?
- How can we use this association to estimate average yield, given a certain level of rainfall?
- How can we use this association to predict future yield, if we have an idea about what the rainfall will be?

Correlation

Definition: The sample correlation coefficient r_{xy} of the paired data

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

is defined by

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})/(n - 1)}{\sqrt{\sum(x_i - \bar{x})^2/(n - 1) * \sum(y_i - \bar{y})^2/(n - 1)}} = \frac{s_{xy}}{s_x s_y}$$

s_{xy} is called the sample covariance of x and y :

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}.$$

Some properties of r_{xy}

- r_{xy} is a measure of the linear assn. between x and y in a dataset.
- correlation coefficients always between -1 and 1:

$$-1 \leq r_{xy} \leq 1$$

- The closer r_{xy} is to 1, the stonger the positive linear association between x and y
- The closer r_{xy} is to -1, the stonger the negative linear association (y tends to be smaller than avg. when x bigger than avg.)
- The bigger $|r_{xy}|$, the stronger the linear association
- If $|r_{xy}| = 1$, then x and y are said to be perfectly correlated.

Summary statistics for corn yields data:

$$\bar{x} = 10.8, \quad s_x^2 = 5.13 \quad s_x = 2.27$$

$$\bar{y} = 31.9, \quad s_y^2 = 19.0 \quad s_y = 4.44$$

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{147.3}{38 - 1} = 3.98$$

Applying the formula for r_{xy} , we get

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{3.98}{\sqrt{5.13 \times 19.0}} = \frac{3.98}{9.87} = 0.40$$

The population correlation coefficient ρ

Just as \bar{x} can be used to say things about a population mean μ_x , r_{xy} can be used for inference about the *population correlation coefficient* ρ_{xy} . This parameter refers to the correlation among x and y in the population from which the sample was drawn:

$$\rho_{XY} = E \left[\frac{(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y} \right] = \rho.$$

A test statistic useful for inference about ρ is

$$Z(\rho) = \left(\frac{1}{2}\sqrt{n-3}\right) \left(\log \frac{1+R}{1-R} - \log \frac{1+\rho}{1-\rho}\right).$$

Asymptotically, Z has the standard normal ($N(0, 1)$) distribution so that it can be used to derive methods of inference for ρ (testing and confidence intervals).

Under $H_0 : \rho = 0$,

$$\left(\frac{1}{2}\sqrt{n-3}\right) \log \frac{1+R}{1-R} \sim N(0, 1).$$

A large-sample test of H_0 with level α then rejects H_0 whenever

$$\frac{1}{2}\sqrt{n-3} \log \left(\frac{1+R}{1-R}\right) > z_{\alpha/2}$$

or

$$\frac{1}{2}\sqrt{n-3} \log \left(\frac{1+R}{1-R}\right) < z_{1-\alpha/2}$$

where z_α satisfies $\alpha = \Pr(Z > z_\alpha)$ with $Z \sim N(0, 1)$.

An approximate $100(1-\alpha)\%$ confidence interval for ρ can be obtained by inverting the *Fisher transformation*:

$$\psi = \frac{1}{2} \log \left(\frac{1+\rho}{1-\rho}\right).$$

The probability statement

$$1-\alpha = \Pr\left(z_{1-\alpha/2} < \sqrt{n-3} \frac{1}{2} \left(\log \left(\frac{1+R}{1-R}\right) - \log \left(\frac{1+\rho}{1-\rho}\right) \right) < z_{\alpha/2}\right)$$

can be rearranged to yield an approximate $100(1-\alpha)\%$ confidence interval for ψ :

$$\frac{1}{2} \log \left(\frac{1+R}{1-R}\right) \pm \frac{z_{\alpha/2}}{\sqrt{n-3}}.$$

Note that ρ and ψ are related by

$$\rho = \frac{e^{2\psi} - 1}{e^{2\psi} + 1}$$

Evaluating ρ at the limits for ψ leads to the interval

$$\left(\frac{\frac{1+r}{1-r} e^{-2z/\sqrt{n-3}} - 1}{\frac{1+r}{1-r} e^{-2z/\sqrt{n-3}} + 1}, \frac{\frac{1+r}{1-r} e^{2z/\sqrt{n-3}} - 1}{\frac{1+r}{1-r} e^{2z/\sqrt{n-3}} + 1} \right).$$

For the corn yields data, $r = 0.4$ and $n = 38$, and a 95% interval is given by

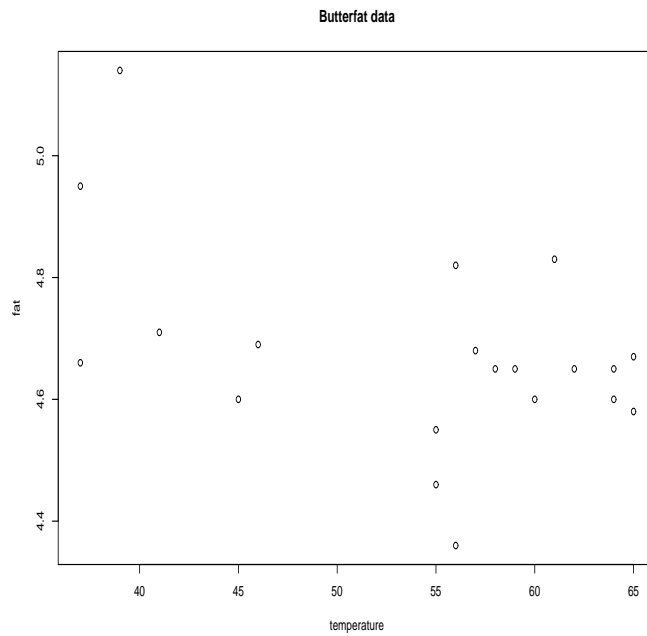
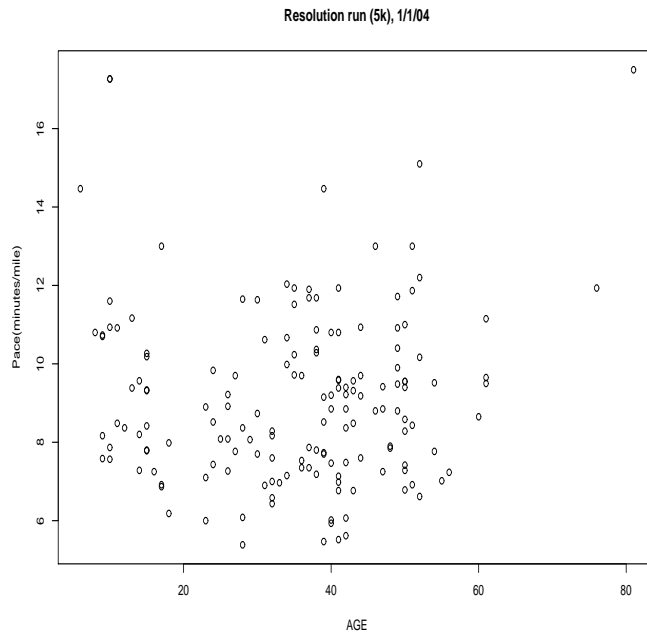
$$(0.09 < \rho < 0.64).$$

(There is a one-to-one correspondence between testing and interval estimation here, so that $H_0 : \rho = 0$ would be rejected at $\alpha = 0.05$.)

Exercise:

1. Examine the butterfat and temperature data plotted on the next page. Is there evidence of linear association? The sample correlation coefficient is $r = -0.45$ based on randomly sampled days. Carry out an appropriate test. Obtain a 95% confidence interval for the population correlation coefficient describing the linear association between % butterfat and temperature.
2. Suppose that two variables X and Y have correlation $\rho = 0.6$. What is the probability that a random sample of $n = 30$ observations from this bivariate population will yield a sample correlation coefficient of 0.7 or higher?

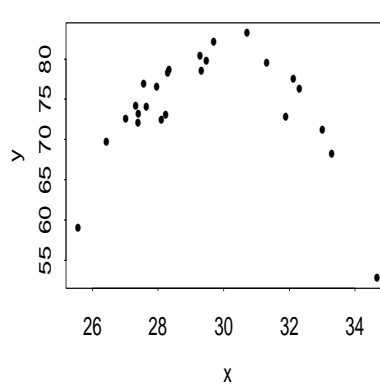
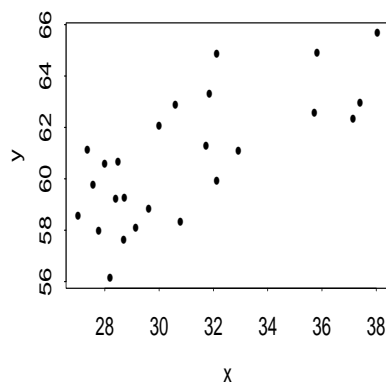
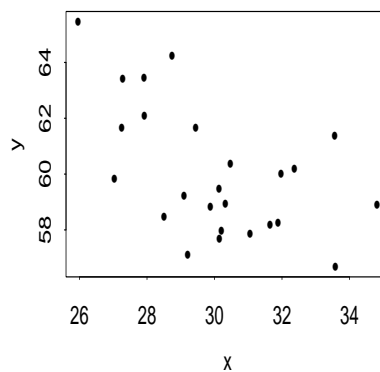
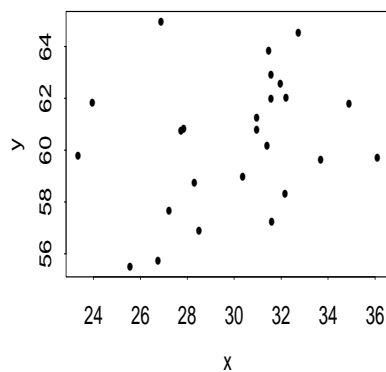
Some example scatterplots ($r_1 = 0.04$ and $r_2 = -0.45$)



An exercise/activity:

Label the four plots below with the four sample correlation coefficients:

1. $r = 0.3$
2. $r = 0.7$
3. $r = 0.1$
4. $r = -0.6$



Correlation does not imply causation

Famous examples of *spurious correlations*:

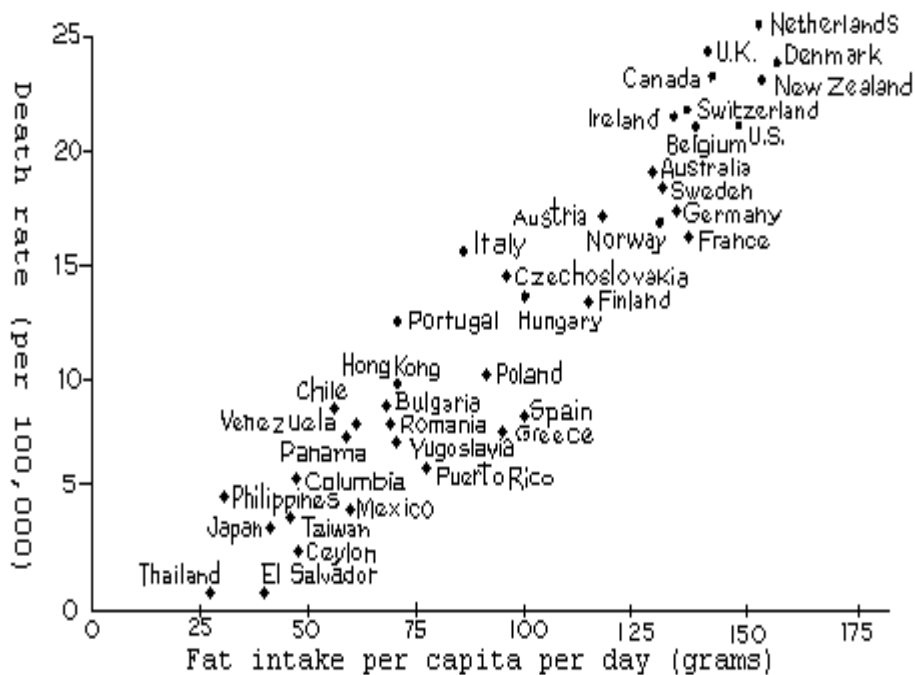
- A study finds a high positive correlation between coffee drinking and coronary heart disease. Newspaper reports say the fragrant essence of the roasted beans of *Coffea arabica* are a menace to public health.
- In a city, if you were to observe the amount of damage and the number of fire engines for enough recent fires, you would likely see a positive and significant correlation among these variables. Obviously, it would be erroneous to conclude that fire engines cause damage.
- *Lurking variable* - a third variable that is responsible for a correlation between two others. (A.k.a. confounding factor.) An example would be to assess the association between say the reading skills of children and other measurements taken on them, such as shoesize. There may be a statistically significant association between shoe size and reading skills, but that doesn't imply that one causes the other. Rather, both are positively associated with a third variable, *age*.
- Among 50 countries examined in a dietary study, high positive correlation among fat intake and cancer (see figure, next page). This example is taken from from *Statistics* by Freedman, Pisani and Purves.

In countries where people eat lots of fat like the United States rates of breast cancer and colon cancer are high. This correlation is often used to argue that fat in the diet causes cancer. How good is the evidence?

Discussion. If fat in the diet causes cancer, then the points in the diagram should slope up, other things being equal. So the diagram is some evidence for the theory. But the evidence is quite weak, because other things aren't equal. For example, the countries with lots of fat in the diet also have lots of sugar. A plot of colon cancer rates against sugar consumption would look just like figure 8, and nobody thinks that sugar causes colon cancer. As it turns out, fat and sugar are relatively expensive. In rich countries, people can afford to eat fat and sugar rather than starchier grain products. Some aspects of the diet in these countries, or other factors in the life-style, probably do cause certain kinds of cancer and protect against other kinds. So far, epidemiologists can identify only a few of these factors with any real confidence. Fat is not among them.

(p. 152, *Statistics* by Friedman, Pisani, Purves and Adhikari)

Figure 8. Cancer rates plotted against fat in the diet for a sample of countries



Source: K. Carroll. "Experimentalevidence of dietary factors and hormone-dependent cancers
Cancer Research vol. 35 (1975) p.3379. Copyright by Cancer Research. Reproduced by
permission

A linear model for regression

Observe n independent pairs $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$

A probabilistic model for Y conditional on $X = x$:

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\substack{\text{deterministic} \\ \text{component}}} + \underbrace{E_i}_{\substack{\text{random error} \\ \text{component}}}$$

where E_1, \dots, E_n are independent and identically distributed normal random variables with mean 0 and variance σ^2 .

(Write $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$.)

Note that this implies

1. $\mu(x) = E(Y|X = x) = \beta_0 + \beta_1 x$
2. $\text{Var}(Y|X = x) = \sigma^2$

Definitions:

- response or dependent variable Y (left side of regression equation)
- independent variable or predictor variable X (right side)
- intercept term $\beta_0 = E(Y|X = 0)$ (where $\mu(x)$ crosses y -axis.)
- slope term β_1 , average change in $E(Y)$ per unit increase in x
- error variance σ^2

β_0, β_1 and σ^2 are modelled as fixed, unknown parameters which can be estimated from the data using simple linear regression.

Nonlinear regression: other models for $E(Y|X = x)$ such as

$$\mu(x) = \beta_0 x^{\beta_1}.$$

Fitting a linear model

- Choose “best” values for β_0, β_1 .

Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that

$$SS[E] = \sum_1^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \sum_1^n (y_i - \hat{y}_i)^2 = \sum_1^n e_i^2$$

is minimized. These are “least squares” (LS) estimates.

Definitions:

- Predicted value of response Y_i given $X = x_i$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

- residual for the i^{th} observation:

$$e_i = y_i - \hat{y}_i$$

Elementary calculus can show that $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimize the sum of squared residuals $SS[E]$ are given by

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum (x_j - \bar{x})(y_j - \bar{y})}{\sum (x_j - \bar{x})^2} \\ &= \frac{S_{xy}}{S_{xx}} \quad (\text{Rao notation, p. 396}) \\ &= \frac{s_{xy}}{s_x^2} \quad (\text{sample covariance } \div \text{ sample variance}) \\ &= r_{xy} \frac{s_y}{s_x} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

An *unbiased* estimate of σ^2 is given by

$$\hat{\sigma}^2 = MS[E] = \frac{SS[E]}{n - 2}.$$

Definition: The line satisfying the equation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

is called the linear regression of y on x .

(It is also called the least squares line.)

For the corn yield data, recall that

$$\bar{x} = 10.8 \text{ inches} \quad \bar{y} = 31.9 \text{ bushels per acre}$$

$$s_x^2 = 5.1, s_y^2 = 19.0, s_{xy} = 3.98, r_{xy} = 0.40$$

so that

$$\begin{aligned} \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\ &= \frac{3.98}{5.13} \\ &= r_{xy} \frac{s_y}{s_x} \\ &= (0.40) \sqrt{\frac{19.0}{5.1}} \\ &= 0.776(\text{ bushels per acre } \div \text{ inches per year}) \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= 31.9 - 0.776(10.8) \\ &= 23.5 \text{ bushels per acre} \end{aligned}$$

yielding the least squares line of

$$\hat{y} = 23.5 + (0.776)x.$$

Note that

1. $\sum_1^n (y_i - \hat{y}_i) = 0$
2. $\sum_1^n (y_i - \hat{y}_i)^2$ is minimized

The ANOVA table from simple linear regression

Observed variability in the response Y is measured by the total sum of squares $SS[TOT]$ and can be partitioned into independent components: the sum of squares due to regression, $SS[R]$ and the sum of squares due to error, $SS[E]$.

Source	Sum of squares	df	Mean Square	F-Ratio
Regression	$SS[R]$	1	$MS[R]$	$MS[R]/MS[E]$
Error	$SS[E]$	$n - 2$	$MS[E]$	
Total	$SS[TOT]$	$n - 1$		

The sums of squares are defined by

$$\begin{aligned}
 SS[TOT] &= \sum_{i=1}^n (y_i - \bar{y})^2 \\
 SS[R] &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
 &= \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}))^2 \\
 &= \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\
 &= \hat{\beta}_1^2 S_{xx}
 \end{aligned}$$

$$\begin{aligned}
 SS[E] &= \sum e_i^2 \\
 &= \sum (y_i - \hat{y}_i)^2 \\
 &= SS[TOT] - SS[R]
 \end{aligned}$$

- The F ratio can be used to test for “significance of regression” or to test the null hypothesis that the slope parameter, β_1 is zero:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

at level α . The critical value for F is the upper α percentile from the F distribution with 1 numerator and $n - 2$ denominator degrees of freedom. This F test is equivalent to a T test based on the statistic we’re about to discuss:

$$T = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}. \quad T^2 = F = \frac{MS[R]}{MS[E]}$$

- The mean square for error, $MS[E]$, is an unbiased estimator for σ^2 , the common variance of the response variable Y *conditional* on an observed independent variable X . ($E(MS[E]) = \sigma^2 = \text{Var}(Y|X)$). (Here, *conditional* means for those elements in the population with independent variable $X = x$.) As such, it can be used to construct confidence intervals for β_0 and β_1 . It is based on $n - 2$ degrees of freedom.
- The ratio of $SS[R]$ to $SS[TOT]$ is called the *coefficient of determination*, or sometimes, simply “r-square”. It represents the proportion of variation observed in the response variable y which can be “explained” by its linear association with x . In simple linear regression, “r-square” is in fact equal to r_{xy}^2 . (But this isn’t the case in multiple regression.) It is also equal to the squared correlation between y_i and \hat{y}_i . (This is the case in multiple regression.)

Confidence intervals for β_0, β_1

Important results for sampling distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ (given x_1, \dots, x_n)

- unbiasedness:

$$E(\hat{\beta}_1|x_1, \dots, x_n) = \beta_1 \quad \text{and} \quad E(\hat{\beta}_0|x_1, \dots, x_n) = \beta_0$$

- for normal data, $\bar{Y} \perp \hat{\beta}_1$ which leads to

$$\begin{aligned} \text{Var}(\hat{\beta}_1|x_1, \dots, x_n) &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \\ \text{Var}(\hat{\beta}_0|x_1, \dots, x_n) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right) \end{aligned}$$

Take $\sqrt{\quad}$ and substitute $MS[E]$ for σ^2 to get estimated standard errors:

$$\begin{aligned} \widehat{SE}(\hat{\beta}_1) &= \sqrt{\frac{MS[E]}{S_{xx}}} \\ \widehat{SE}(\hat{\beta}_0) &= \sqrt{MS[E] \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right)} \end{aligned}$$

100(1 - α)% confidence intervals for β_1 and β_0 are given by

$$\begin{aligned} \hat{\beta}_1 \pm t(n-2, \alpha/2) \sqrt{\frac{MS[E]}{S_{xx}}} \\ \hat{\beta}_0 \pm t(n-2, \alpha/2) \sqrt{MS[E] \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}. \end{aligned}$$

Any hypothetical slope, like $H_0 : \beta_1 = \text{slope}_0$ may be tested using the T -statistic below with $df = n - 2$:

$$T = \frac{\hat{\beta}_1 - \text{slope}_0}{\widehat{SE}(\hat{\beta}_1)}$$

Confidence interval for $E(Y|X = x_0)$

The conditional mean $E(Y|X = x_0)$ can be estimated by evaluating the regression function $\mu(x_0)$ at the estimates $\hat{\beta}_0, \hat{\beta}_1$. The conditional variance of the expression isn't too difficult:

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 | X = x_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

This yields a confidence interval of the form

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t(n - 2, \alpha/2) \sqrt{MS[E] \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

Exercise: derive these variances.

The yield on corn by rainfall example

Source	Sum of squares	df	Mean Square	F-Ratio
Regression	114	1	114	6.95
Error	591	36	16.4	
Total	705	37		

The $\alpha = 0.05$ critical value is $F(1, 36, 0.05) = 4.11$. Therefore, there is a significant, positive linear association between yield and rainfall.

For 95% CIs, use $t(26, 0.025) = 2.028$. For β_1 , note that

$S_{xx} = (n - 1)s_x^2 = 5.13(38 - 1) = 189.8$ so that a c.i. is given by

$$0.776 \pm (2.028)\sqrt{\frac{16.4}{189.8}}$$

or

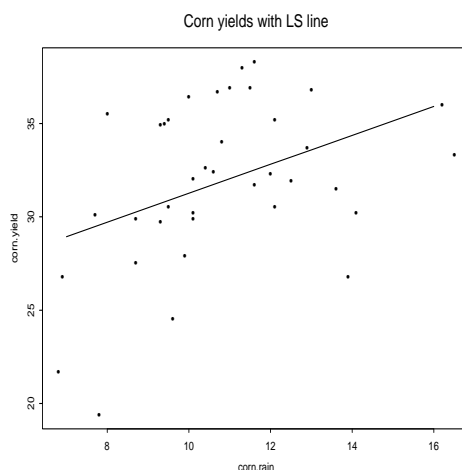
$$0.776 \pm (2.028)(0.294) \quad \text{or} \quad 0.776 \pm 0.596$$

For $\hat{\beta}_0$,

$$23.5 \pm (2.028)\sqrt{16.4 \left(\frac{1}{38} + \frac{(10.8)^2}{189.8} \right)}$$

or

$$23.5 \pm (2.028)(3.24) \quad \text{or} \quad 23.5 \pm 6.57$$



Prediction

Often, prediction of the response variable Y for a given value, say x_0 , of the independent variable is of interest. In order to make statements about future values of Y , we need to take into account

- the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$
- the randomness of a future value Y .

We've seen that the predicted value of Y based on the linear regression is given by $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

In order to form a 95% prediction interval take

$$\hat{Y}_0 \pm t(n-2, \alpha/2) \sqrt{MS[E] \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}.$$

Example: Suppose that one year rainfall is $x_0 = 14$ inches, but that yield Y_0 from the six states hasn't been measured. Obtain a 95% prediction interval for Y_0 , using the model.

$$23.5 + 0.776(14) \pm 2.028 \sqrt{16.4 \left(1 + \frac{1}{38} + \frac{(14 - 10.8)^2}{189.8} \right)}$$

or

$$34.4 \pm 8.5 \quad \text{or} \quad (25.9, 42.9)$$

The 95% prediction interval is (25.9, 42.9).

A 95% confidence interval for $E(Y|X = 14)$ is given by

$$23.5 + 0.776(14) \pm 2.028 \sqrt{16.4 \left(\frac{1}{38} + \frac{(14 - 10.8)^2}{189.8} \right)}$$

or

$$34.4 \pm 2.32 \quad \text{or} \quad (32.1, 36.7)$$

What is the difference?

Exercise taken from Dickey's notes:

An industrial quality control expert takes 200 hourly measurements on an industrial furnace which is under control and finds that a 95% confidence interval for the mean temperature is (500.35, 531.36). As a result he tells management that the process should be declared out of control whenever hourly measurements fall outside this interval and, of course, is later fired for incompetence. (Why and what should he have done?)

A note of caution:

Mark Twain, in *Life on the Mississippi*:

In the space of 176 years the Lower Mississippi has shortened itself 252 miles. That is an average of a trifle more than one mile and a third per year. Therefore any calm person, who is not blind or idiotic, can see that in 742 years from now, the lower Mississippi will be one mile and three quarters long, and Cairo, Ill., and New Orleans will have joined their streets together.

It is not safe to extrapolate the results of a linear regression for the purposes of prediction beyond the range of observed independent variables.

1. Butterfat by temperature in cows:

- y_j : average “percent butterfat” for 10 cows on date j
- x_j : temperature on date j
- $n = 20$ successive days

Date	1	2	3	4	5	6	7	8	9	10
x_j	64	65	65	64	61	55	39	41	46	59
y_j	4.65	4.58	4.67	4.60	4.83	4.55	5.14	4.71	4.69	4.65
Date	11	12	13	14	15	16	17	18	19	20
x_j	56	56	62	37	37	45	57	58	60	55
y_j	4.36	4.82	4.65	4.66	4.95	4.60	4.68	4.65	4.6	4.46

2. Hybrid duck data:

- Mallard and Pintail ducks were crossed yielding $n = 11$ second generation males with attributes as given in the table
- y_j : Behavioral index
- x_j : Plumage index
- A 0 corresponds to a purely mallard phenotype and a 15 corresponds to a purely pintail phenotype.
- The same scoring is used to quantify duck behavioral traits.

x_j	7	13	14	6	14	15	4	8	7	9	14
y_j	3	10	11	5	15	15	7	10	4	9	11

3. Cricket Data

- y_j Chirps per second
- x_j Temperature ($^{\circ}F$)
- Striped ground cricket.

y_j	20	16	19.8	18.4	17.1	15.5	14.7	17.1	15.4	16.2
x_j	88.6	71.6	93.3	84.3	80.6	75.2	69.7	82	69.4	83.3
y_j	15	17.2	16	17	14.4					
x_j	79.6	82.6	80.6	83.5	76.3					

ST 512

Week 2-3 Multiple linear regression (MLR): Intro., Model Selection

Reading Ch. 11

Multiple linear regression - an example

A random sample of students taking the same exam:

IQ	Study TIME	GRADE
105	10	75
110	12	79
120	6	68
116	13	85
122	16	91
130	8	79
114	20	98
102	15	76

Consider a regression model for the GRADE of subject i , Y_i , in which the mean of Y_i is a linear function of two independent variables $X_{i1} = \text{IQ}$ and $X_{i2} = \text{Study TIME}$ for subjects $i = 1, \dots, 8$:

$$Y = \beta_0 + \beta_1 \text{IQ} + \beta_2 \text{TIME} + \text{error}$$

or

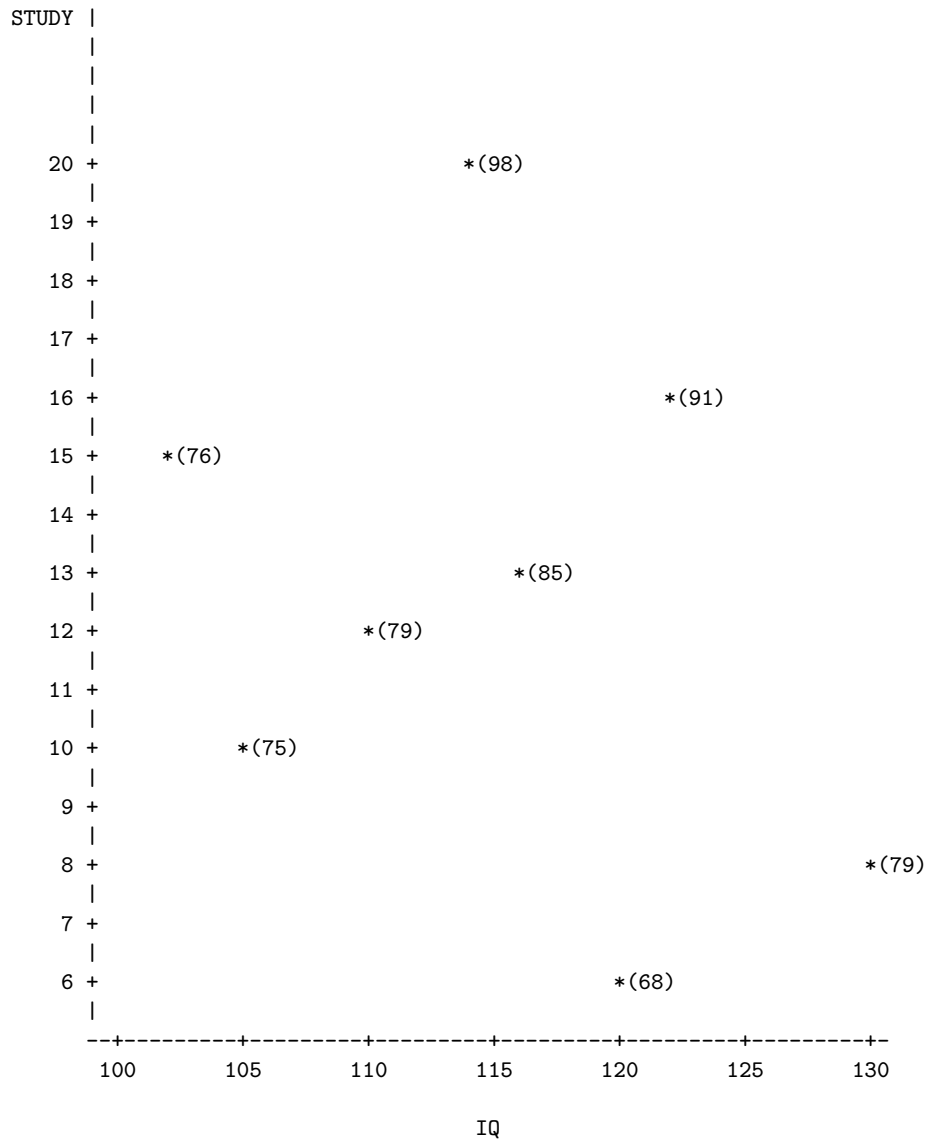
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + E_i$$

or

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + E_1 \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + E_2 \\ &\vdots = \vdots \\ Y_8 &= \beta_0 + \beta_1 X_{81} + \beta_2 X_{82} + E_8 \end{aligned}$$

GRADE AND STUDY TIME EXAMPLE FROM ST 512 NOTES

Plot of STUDY*IQ. Symbol used is '*'.



A multiple linear regression (MLR) model w/ p independent variables

Let p independent variables be denoted by x_1, \dots, x_p .

- Observed values of p independent variables for i^{th} subject from sample denoted by $x_{i1}, x_{i2}, \dots, x_{ip}$
- response variable for i^{th} subject denoted by Y_i
- For $i = 1, \dots, n$, MLR model for Y_i :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + E_i.$$

- As in SLR, $E_1, \dots, E_n \stackrel{iid}{\sim} N(0, \sigma^2)$,

Least squares estimates of regression parameters minimize $SS[E]$:

$$SS[E] = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2$$

$$\hat{\sigma}^2 = \frac{SS[E]}{n-p-1}$$

Interpretations of regression parameters:

- σ^2 is unknown error variance parameter.
- $\beta_0, \beta_1, \dots, \beta_p$ are $p + 1$ unknown regression parameters:
 - β_0 : average response when $x_1 = x_2 = \dots = x_p = 0$
 - β_i is called a partial slope for x_i . Represents mean change in y per unit increase in x_i *with all other independent variables held fixed*.

For this example, with $p = 2$ and $n = 8$,

$$\hat{\beta}_0 = 0.74, \hat{\beta}_1 = 0.47, \hat{\beta}_2 = 2.1$$

What is the uncertainty associated with these parameter estimates?

Matrix formulation of MLR

Let a $(1 \times (p + 1))$ vector for p observed independent variables for individual i be defined by

$$x_{i\cdot} = (1, x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip}).$$

The MLR model for Y_1, \dots, Y_n is given by

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1p} + E_1 \\ Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_p x_{2p} + E_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{np} + E_n \end{aligned}$$

This system of n equations can be expressed using matrices:

$$Y = X\beta + E$$

where

- Y denotes a response vector ($n \times 1$)
- X denotes a design matrix ($n \times (p + 1)$)
- β denotes a vector of regression parameters ($(p + 1) \times 1$)
- E denotes an error vector ($n \times 1$)

Here, the error vector E is assumed to follow a multivariate normal distribution with variance-covariance matrix $\sigma^2 I_n$

For individual i ,

$$Y_i = x_{i\cdot}\beta + E_i$$

Some simplified expressions: (a is a known $p \times 1$ vector)

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ \text{Var}(\hat{\beta}) &= \sigma^2(X'X)^{-1} \\ &= \Sigma \\ \widehat{\text{Var}}(\hat{\beta}) &= MS[E](X'X)^{-1} \\ &= \hat{\Sigma} \\ \widehat{\text{Var}}(a'\hat{\beta}) &= a'\hat{\Sigma}a\end{aligned}$$

(What are the dimensions of each of these quantities?)

- Rao calls $(X'X)^{-1}$ the S -matrix.
- $\hat{\Sigma}$ is the estimated variance-covariance matrix for the estimate of the regression parameter vector $\hat{\beta}$
- X is assumed to be of full *rank*

Some more simplified expressions:

$$\begin{aligned}\hat{Y} &= X\hat{\beta} \\ &= X(X'X)^{-1}X'y \\ &= Hy \\ e &= Y - \hat{Y} \\ &= Y - X\hat{\beta} \\ &= (I - H)Y\end{aligned}$$

- \hat{Y} is called the vector of fitted or predicted values
- $H = X(X'X)^{-1}X$ is called the hat matrix
- e is the vector of residuals

For the IQ, Study TIME example, with $p = 2$ independent variables and $n = 8$ observations, consider

$X, Y, (X'X)^{-1}, (X'X)^{-1}X; Y, X(X'X)^{-1}X'Y:$

$$X = \begin{pmatrix} 1 & 105 & 10 \\ 1 & 110 & 12 \\ 1 & 120 & 6 \\ 1 & 116 & 13 \\ 1 & 122 & 16 \\ 1 & 130 & 8 \\ 1 & 114 & 20 \\ 1 & 102 & 15 \end{pmatrix}$$

and

$$X'X = \begin{pmatrix} 8 & 919 & 100 \\ 919 & 106165 & 11400 \\ 100 & 11400 & 1394 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 28.90 & -0.23 & -0.22 \\ -0.23 & 0.0018 & 0.0011 \\ -0.22 & 0.0011 & 0.0076 \end{pmatrix}$$

$$(X'X)^{-1}X'Y = \begin{pmatrix} 0.74 \\ 0.47 \\ 2.10 \end{pmatrix} = ?$$

$$SS[E] = e'e = (Y - \hat{Y})'(Y - \hat{Y}) = 45.8, \quad e'e/df = 9.15 = ?$$

$$\hat{\Sigma} = MS[E](X'X)^{-1} = \begin{pmatrix} 264.45 & -2.07 & -2.05 \\ -2.07 & 0.017 & 0.010 \\ -2.05 & 0.010 & 0.070 \end{pmatrix}$$

Some questions - use preceding page

1. What is the estimate for β_1 ? Interpretation?
2. What is the standard error of $\hat{\beta}_1$?
3. Is $\beta_1 = 0$ plausible, while controlling for possible linear associations between Test Score and Study time? ($t(0.025, 5) = 2.57$)
4. Estimate the mean grade among the population of ALL students with $IQ = 113$ who study $TIME = 14$ hours.
5. Report a standard error
6. Report a 95% confidence interval

Some answers

1. $\hat{\beta}_1 = 0.47$ (second element of $(X'X)^{-1}X'Y$, exam points per IQ point for students studying the same amount)
2. $\sqrt{0.017} = 0.13$ (square root of middle element of $\hat{\Sigma}$)
3. $H_0 : \beta_1 = 0$, T-statistic: $t = (\hat{\beta}_1 - 0)/SE(\hat{\beta}_1)$
Observed value is $t = .47/\sqrt{.017} = .47/.13 = 3.6 > 2.57$,
 (“ $\hat{\beta}_1$ differs significantly from 0.”)
4. Unknown population mean: $\theta = \beta_0 + \beta_1(113) + \beta_1(14)$
Estimate : $\hat{\theta} = (1, 113, 14) * \hat{\beta} = 83.6$
5. $\text{Var}((1, 113, 14) * \hat{\beta}) = (1, 113, 14)\widehat{\text{Var}}(\hat{\beta})(1, 113, 14)'$
or $(1, 113, 14)\hat{\Sigma}(1, 113, 14)' = 1.3$ or $SE(\hat{\theta}) = \sqrt{1.3} = 1.14$
6. $\hat{\theta} \pm t(0.025, 5)SE(\hat{\theta})$ or $83.6 \pm 2.57(1.14)$ or $(80.7, 86.6)$

```
DATA GRADES; INPUT IQ STUDY GRADE @@; CARDS;
105 10 75 110 12 79 120 6 68 116 13 85 122 16 91 130 8 79 114 20 98 102 15 76
DATA EXTRA; INPUT IQ STUDY GRADE; CARDS;
113 14 .
DATA BOTH; SET GRADES EXTRA;
PROC REG; MODEL GRADE = IQ STUDY/P CLM XPX INV COVB;
```

The SAS System
The REG Procedure

Model Crossproducts X'X X'Y Y'Y

Variable	Intercept	IQ	STUDY	GRADE
Intercept	8	919	100	651
IQ	919	106165	11400	74881
STUDY	100	11400	1394	8399
GRADE	651	74881	8399	53617

X'X Inverse, Parameter Estimates, and SSE

Variable	Intercept	IQ	STUDY	GRADE
Intercept	28.898526711	-0.226082693	-0.224182192	0.7365546771
IQ	-0.226082693	0.0018460178	0.0011217122	0.473083715
STUDY	-0.224182192	0.0011217122	0.0076260404	2.1034362851
GRADE	0.7365546771	0.473083715	2.1034362851	45.759884688

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	596.11512	298.05756	32.57	0.0014
Error	5	45.75988	9.15198		
Corrected Total	7	641.87500			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.73655	16.26280	0.05	0.9656
IQ	1	0.47308	0.12998	3.64	0.0149
STUDY	1	2.10344	0.26418	7.96	0.0005

Covariance of Estimates

Variable	Intercept	IQ	STUDY
Intercept	264.47864999	-2.069103589	-2.051710248
IQ	-2.069103589	0.016894712	0.010265884
STUDY	-2.051710248	0.010265884	0.0697933458

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean	Residual
1	75.0000	71.4447	1.9325	66.4770 76.4124	3.5553
(abbreviated)					
8	76.0000	80.5426	1.9287	75.5847 85.5005	-4.5426
9	.	83.6431	1.1414	80.7092 86.5771	.

Model Selection

x_1, x_2, x_3 denote p independent variables. Consider several models:

1. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1x_1$
2. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2x_2$
3. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_3x_3$
4. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$
5. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1x_1 + \beta_3x_3$
6. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1x_1 + \beta_2x_2$
7. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2x_2 + \beta_3x_3$

A is nested in B means model A can be obtained by restricting (e.g. setting to 0) parameter values in model B .

True or false:

- Model 1 nested in Model 4
- Model 1 nested in Model 5
- Model 2 nested in Model 4
- Model 4 nested in Model 1
- Model 3 nested in Model 4
- Model 5 nested in Model 4

A nested in $B \longrightarrow A$ called *reduced*, B called *full*.

p - number of regression parameters in full model

q - number of regression parameters in reduced model

$p - q$ - number of regression parameters being tested.

Recall:

$$\begin{aligned}
 SS[R] &= \sum (\hat{Y}_i - \bar{Y})^2 \\
 SS[E] &= \sum (\hat{Y}_i - Y_i)^2 \\
 SS[Total] &= \sum (Y_i - \bar{Y})^2
 \end{aligned}$$

Model Selection - concepts

In comparing two models, suppose

β_1, \dots, β_q in reduced model (A)

$\beta_1, \dots, \beta_q, \beta_{q+1}, \dots, \beta_p$ in full model (B).

Comparison of models A and B amounts to testing

$H_0 : \beta_{q+1} = \beta_{q+2} = \dots = \beta_p = 0$ (model A ok)

$H_1 : \beta_{q+1}, \beta_{q+2}, \dots, \beta_p$ not all 0 (need model B)

$$\text{Let } F = \frac{(SS[E]_r - SS[E]_f)/(p - q)}{MS[E]_f} = \frac{MS[H_0]}{MS[E]}$$

(r and f abbreviate *reduced* and *full*, respectively.)

Difference in the numerator called an *extra regression sum of squares*:

$$R(\beta_{q+1}, \beta_{q+2}, \dots, \beta_p | \beta_0, \beta_1, \beta_2, \dots, \beta_q) = SS[R]_f - SS[R]_r.$$

(ok to suppress β_0 in these extra SS terms.)

Theory gives that if H_0 holds (model A is appropriate) F behaves according to the F distribution with $p - q$ numerator and $n - p - 1$ denominator degrees of freedom.

Extra SS terms for comparing some of the nested models on preceding page:

- Model 1 in model 4: $R(\beta_2, \beta_3 | \beta_1)$
- Model 2 in model 4 ?
- Model 3 in model 4 ?
- Model 1 in model 5: $R(\beta_3 | \beta_1)$
- Model 5 in model 4: ?

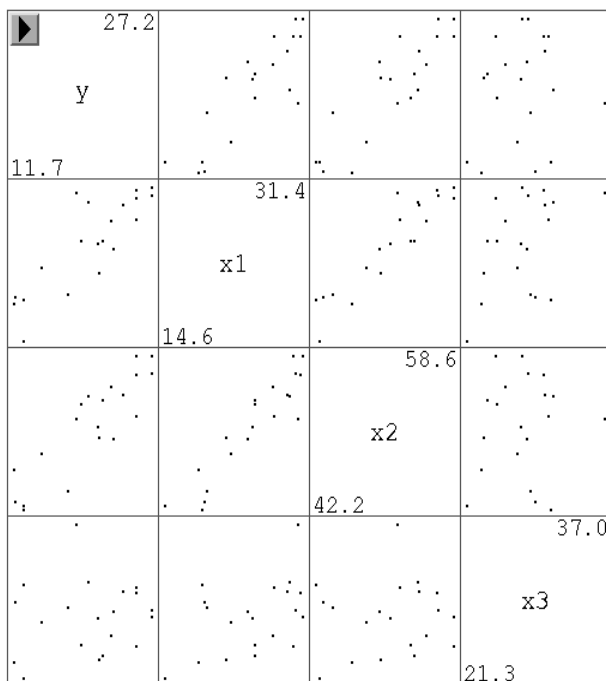
An example: How to measure body fat?

For each of $n = 20$ healthy individuals, the following measurements were made: bodyfat percentage y_i , triceps skinfold thickness, x_1 , thigh circumference x_2 , midarm circumference x_3

x1	x2	x3	y
19.5	43.1	29.1	11.9
24.7	49.8	28.2	22.8
30.7	51.9	37.0	18.7
29.8	54.3	31.1	20.1
19.1	42.2	30.9	12.9
25.6	53.9	23.7	21.7
31.4	58.5	27.6	27.1
27.9	52.1	30.6	25.4
22.1	49.9	23.2	21.3
25.5	53.5	24.8	19.3
31.1	56.6	30.0	25.4
30.4	56.7	28.3	27.2
18.7	46.5	23.0	11.7
19.7	44.2	28.6	17.8
14.6	42.7	21.3	12.8
29.5	54.4	30.1	23.9
27.7	55.3	25.7	22.6
30.2	58.6	24.6	25.4
22.7	48.2	27.1	14.8
25.2	51.0	27.5	21.1

Summary statistics:

Symbol	Variable	mean	st. dev.
y	Body fat	20.2	5.1
x1	Triceps	25.3	5.0
x2	Thigh Circ.	51.2	5.2
x3	Midarm Circ.	27.6	3.6



Pearson Correlation Coefficients, N = 20
 Prob > |r| under H0: Rho=0

	y	x1	x2	x3
y	1.00000	0.84327 <.0001	0.87809 <.0001	0.14244 0.5491
x1	0.84327 <.0001	1.00000	0.92384 <.0001	0.45778 0.0424
x2	0.87809 <.0001	0.92384 <.0001	1.00000	0.08467 0.7227
x3	0.14244 0.5491	0.45778 0.0424	0.08467 0.7227	1.00000

Marginal associations between y and x_1 and between y and x_2 are highly significant, providing evidence of a strong $r \approx 0.85$ linear association between average bodyfat and triceps skinfold and between average bodyfat and thigh circumference.

Multicollinearity: linear associations among the independent variables; causes problems such as inflated sampling variances for $\hat{\beta}$.

```

data bodyfat;
  input x1 x2 x3 y;
  cards;
19.5 43.1 29.1 11.9
24.7 49.8 28.2 22.8
  (data abbreviated)
22.7 48.2 27.1 14.8
25.2 51.0 27.5 21.1
;

proc reg data=bodyfat;
  model y=x1 x2 x3;
  model y=x1;
  model y=x2;
  model y=x3;
  *model y=x1 x2 x3/xpx i covb corrb;

```

Yields the following (abbreviated) output

The SAS System
The REG Procedure

1

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	117.08469	99.78240	1.17	0.2578
x1	1	4.33409	3.01551	1.44	0.1699
x2	1	-2.85685	2.58202	-1.11	0.2849
x3	1	-2.18606	1.59550	-1.37	0.1896

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-1.49610	3.31923	-0.45	0.6576
x1	1	0.85719	0.12878	6.66	<.0001

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-23.63449	5.65741	-4.18	0.0006
x2	1	0.85655	0.11002	7.79	<.0001

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	14.68678	9.09593	1.61	0.1238
x3	1	0.19943	0.32663	0.61	0.5491

Model Selection - examples

In the bodyfat data, consider comparing the simple model that Y depends only on x_1 (triceps) and not on x_2 (thigh) or x_3 (midarm) after accounting for x_1 versus the full model that it depends on all three.

$$\text{Model } A : \mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1$$

$$\text{Model } B : \mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

or the null hypothesis

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_1 : \beta_2, \beta_3 \text{ not both } 0$$

after accounting for x_1 .

$$F = \frac{(143.1 - 98.4)/2}{6.15} = \frac{22.4}{6.2} = 3.64$$

How many df associated with this F ratio? (Recall $n = 20$.) The 95th percentile is $F(0.05, \quad, \quad) = 3.63$.

Q: Conclusion from this comparison of nested models?

After accounting for variation in bodyfat explained by triceps, there is still some association between mean bodyfat and at least one of x_2, x_3 (thigh, midarm).

To get this F -ratio in SAS, try

```
proc reg data=bodyfat;
  model y=x1 x2 x3;
  test x2=0,x3=0;
run;
```

Adding x_2, x_3 to the model leads to overfitting (a model too complex for its own good)

PROC GLM can replace PROC REG to get the SUMS OF SQUARES for use in model selection as in the following output:

The GLM Procedure					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.9846118	132.3282039	21.52	<.0001
Error	16	98.4048882	6.1503055		
Corrected Total	19	495.3895000			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x1	1	352.2697968	352.2697968	57.28	<.0001
x2	1	33.1689128	33.1689128	5.39	0.0337
x3	1	11.5459022	11.5459022	1.88	0.1896

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x1	1	12.70489278	12.70489278	2.07	0.1699
x2	1	7.52927788	7.52927788	1.22	0.2849
x3	1	11.54590217	11.54590217	1.88	0.1896

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	117.0846948	99.78240295	1.17	0.2578
x1	4.3340920	3.01551136	1.44	0.1699
x2	-2.8568479	2.58201527	-1.11	0.2849
x3	-2.1860603	1.59549900	-1.37	0.1896

Note agreement between p -values from Type III F tests and p -values from t tests from parameter estimates from MLR.

Type I sums of squares - sequential (order of selection matters)

Type III sums of squares - partial

Type II sums of squares - partial (change in SSE due to adding term A to model with all other terms not 'containing' A)

In the output on the preceding page,

$$R(\beta_1|\beta_0) = 352.3$$

$$R(\beta_2|\beta_0, \beta_1) = 33.2$$

$$R(\beta_3|\beta_0, \beta_1, \beta_2) = 11.5$$

$$R(\beta_1|\beta_0, \beta_2, \beta_3) = 12.7$$

$$R(\beta_2|\beta_0, \beta_1, \beta_3) = 7.5$$

Type III test for β_j - test of partial association between y and x_j
after accounting for all other x_i

Type III F -ratios from bodyfat data for x_1, x_2, x_3 , respectively:

$$F = \frac{12.7/1}{6.15} = 2.07, \quad F = \frac{7.5/1}{6.15} = 1.22, \quad F = \frac{11.5/1}{6.15} = 1.88.$$

(Partial) effects significant ? (Use $F(0.95, 1, 16) = 4.49$.)

Exercise: specify the comparison of nested models that corresponds to each of these F -ratios.

In GLM output, which models are the type I tests comparing?

1. Type I SS for x_1 from PROC GLM appropriate for SLR of y on x_1 .
2. Type I SS for x_2 from PROC GLM appropriate for test of association between y and x_2 after accounting for x_1 .
3. Type I test for x_3 from PROC GLM same as type III test for x_3 .

In all three of these tests, $MS[E]$ computed from full model (#4).

Some model comparison examples

1. Compare models 1 and 6
2. Compare models 2 and 6

For 1. use $R(\beta_2|\beta_0, \beta_1)$ in the F ratio:

$$\begin{aligned}
 F &= \frac{R(\beta_2|\beta_0, \beta_1)}{MS[E]_6} \\
 &= \frac{33.2}{(SS[Total] - R(\beta_1, \beta_2|\beta_0))/(20 - 2 - 1)} \\
 &= \frac{33.2}{(495.4 - 352.3 - 33.2)/(20 - 2 - 1)} \\
 &= \frac{33.2}{109.9/17} \\
 &= 5.1
 \end{aligned}$$

Note that

$$SS[E]_f = (SS[Total] - SS[R]_f) \text{ and } SS[R]_f = SS[R]_r + R(\beta_2|\beta_0, \beta_1)$$

$F(0.05, 1, 17) = 4.45$: nested model 1 is rejected in favor of model 6: there is evidence ($p = 0.037$) of association between y and x_2 after accounting for dependence on x_1 .

To compare models 2 and 6, we need $SS[R]_r = R(\beta_2|\beta_0) = 382.0$ which cannot be gleaned from preceding output. You could also get it from $r_{yx_2}^2 \times SS[Total]$ or from running something like

```
proc reg;
  model y=x1 x2/ss1 ss2;
run;
```

The REG Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	385.43871	192.71935	29.80	<.0001
Error	17	109.95079	6.46769		
Corrected Total	19	495.38950			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS	Type II SS
Intercept	1	-19.17425	8.36064	-2.29	0.0348	8156.76050	34.01785
x1	1	0.22235	0.30344	0.73	0.4737	352.26980	3.47289
x2	1	0.65942	0.29119	2.26	0.0369	33.16891	33.16891

$$\begin{aligned}
 F &= \frac{R(\beta_1|\beta_0, \beta_2)/(\Delta df)}{MS[E]_f} \\
 &= \frac{(SS[R]_f - SS[R]_r)/1}{6.5} \\
 &= \frac{352.3 + 33.2 - 382.0}{6.5} \\
 &= \frac{3.4}{6.5} \approx 0.5
 \end{aligned}$$

Conclusions?

- x_2 gives you a little when you add it to model with x_1
- x_1 gives you nothing when you add it to model with x_2
- Take model with x_2 . (Has higher r^2 too.)

Note that all of these comparisons of nested models are easy to carry out using the **TEST** statement in **PROC REG**.

Another example, revisiting test scores and study times
 Consider this sequence of analyses:

1. Regress GRADE on IQ.
2. Regress GRADE on IQ and TIME.
3. Regress GRADE on TIME IQ TI where $TI = \text{TIME} \cdot \text{IQ}$.

ANOVA (Grade on IQ)

SOURCE	DF	SS	MS	F	<i>p</i> -value
IQ	1	15.9393	15.9393	0.153	0.71
Error	6	625.935	104.32		

It appears that IQ has nothing to do with grade, but we did not look at study time. Looking at the *multiple* regression we get

The REG Procedure

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	596.11512	298.05756	32.57	0.0014
Error	5	45.75988	9.15198		
Corrected Total	7	641.87500			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.73655	16.26280	0.05	0.9656
IQ	1	0.47308	0.12998	3.64	0.0149
study	1	2.10344	0.26418	7.96	0.0005

Now the test for dependence on IQ is significant $p = 0.0149$. Why?

The interaction model

		The SAS System				1	
		The REG Procedure					
Analysis of Variance							
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F		
Model	3	610.81033	203.60344	26.22	0.0043		
Error	4	31.06467	7.76617				
Corrected Total	7	641.87500					

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS	Type II SS
Intercept	1	72.20608	54.07278	1.34	0.2527	52975	13.84832
IQ	1	-0.13117	0.45530	-0.29	0.7876	15.93930	0.64459
study	1	-4.11107	4.52430	-0.91	0.4149	580.17582	6.41230
IQ_study	1	0.05307	0.03858	1.38	0.2410	14.69521	14.69521

Discussion of the interaction model. We call the product $I*S = IQ*STUDY$ an "interaction" term. Our model is

$$\hat{G} = 72.21 - 0.13 * I - 4.11 * S + 0.0531(I * S)$$

Now if $IQ = 100$ we get

$$\hat{G} = (72.21 - 13.1) + (-4.11 + 5.31)S$$

and if $IQ = 120$ we get

$$\hat{G} = (72.21 - 15.7) + (-4.11 + 6.37)S.$$

Thus we expect an extra hour of study to increase the grade by 1.20 points for someone with $IQ = 100$ and by 2.26 points for someone with $IQ = 120$ if we use this interaction model. Since the interaction is not significant, we may want to go back to the simpler "main effects" model. (*This example taken from Dickey's ST512 notes.*)

Some questions about design matrices

Recall three models under consideration for the bodyfat data

$$M_1 : \mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1$$

$$M_2 : \mu(x_1, x_2, x_3) = \beta_0 + \beta_2 x_2$$

$$M_6 : \mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Q: $MS[E]_{M_6} < MS[E]_{M_1}$ and $MS[E]_{M_6} < MS[E]_{M_2}$ but the partial slopes have larger standard errors in M_6 . Why?

Design matrices

$$X_{M_6} = \begin{pmatrix} 1 & 19.5 & 43.1 \\ 1 & 24.7 & 49.8 \\ \vdots & \vdots & \vdots \\ 1 & 22.7 & 48.2 \\ 1 & 25.2 & 51.0 \end{pmatrix} \quad X_{M_1} = \begin{pmatrix} 1 & 19.5 \\ 1 & 24.7 \\ \vdots & \vdots \\ 1 & 22.7 \\ 1 & 25.2 \end{pmatrix}$$

(similarly for X_{M_2}).

$$(X'X)_{M_6} = \begin{pmatrix} ? & 506.1 & 1023.4 \\ & 13386.3 & 26358.7 \\ & & 52888.0 \end{pmatrix}$$

$$(X'X)_{M_1} = \begin{pmatrix} ? & ? \\ & ? \end{pmatrix} \quad (X'X)_{M_1}^{-1} = \begin{pmatrix} 1.39 & -0.053 \\ & 0.002 \end{pmatrix}$$

$$(X'X)_{M_2} = \begin{pmatrix} ? & ? \\ & ? \end{pmatrix} \quad (X'X)_{M_2}^{-1} = \begin{pmatrix} 5.08 & -0.098 \\ & 0.0019 \end{pmatrix}$$

$$(X'X)_{M_6}^{-1} = \begin{pmatrix} 10.8 & 0.29 & -0.35 \\ & 0.014 & -0.012 \\ & & 0.013 \end{pmatrix}$$

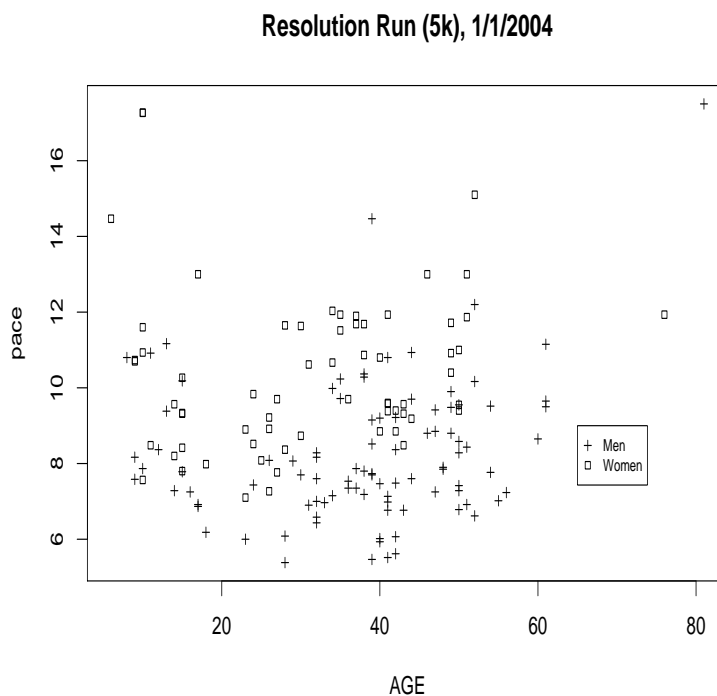
Q: Why is $\text{Var}(\hat{\beta}_0)$ bigger in M_2 than in M_1 ?

Recall the Resolution Run 5k race data

Obs	age	sex	race	pace
1	28	M	16.6833	5.38333
2	39	M	16.9500	5.46667
3	41	M	17.1333	5.51667
4	42	M	17.4000	5.61667
(abbreviated)				
157	52	F	46.8833	15.1000
158	10	F	53.6000	17.2667
159	10	F	53.6167	17.2667
160	81	M	54.3167	17.5000

Summary statistics ($n = 160$):

Symbol	Variable	mean	st. dev.	variance
y	Pace	9.1	2.2	5.0
x	Age	35.1	14.7	216.5



Quadratic model for pace (Y) as a function of age (x):

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i \quad \text{for } i = 1, \dots, 160$$

where $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$.

- $\beta = (\beta_0, \beta_1, \beta_2)'$ is a vector of unknown regression parameters
- σ^2 is the unknown error variance of paces given age x .

Compare this model with the (previously discarded) SLR model

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad \text{for } i = 1, \dots, 160$$

Q1: Does β_1 have the same interpretation in both models?

Q2: How can we compare the two models?

A2: Using F -ratios to compare nested models (see output next page).

$$\begin{aligned} F &= \frac{R(\beta_2 | \beta_0, \beta_1)}{MS[E]_{full}} \\ &= \frac{(SS[R]_{full} - SS[R]_{red})/1}{MS[E]_{full}} \\ &= \frac{(113.6 - 1.1)/1}{4.3} \\ &= \frac{(SS[E]_{red} - SS[E]_{full})/1}{MS[E]_{full}} \\ &= \frac{(787.0 - 674.4)/1}{4.3} \\ &= 26.2 \\ &= \left(\frac{\hat{\beta}_1}{SE} \right)^2 \end{aligned}$$

with $F(0.05, 1, 157) = 3.90$. Since $26.2 \gg 3.9$, the linear model is implausible when compared to the quadratic model.

```

/* age2 defined in data step as age*age */
PROC REG;
  MODEL pace=age;
  MODEL pace=age age2/ss1; /* ss1 generates sequential sums of squares */
/* only the 2nd model statement really necessary */
RUN;

```

1

The SAS System
The REG Procedure
Model: MODEL1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.09650	1.09650	0.22	0.6396
Error	158	786.99821	4.98100		
Corrected Total	159	788.09472			

Root MSE	2.23182	R-Square	0.0014
Dependent Mean	9.12063	Adj R-Sq	-0.0049

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	8.92271	0.45724	19.51	<.0001
age	1	0.00564	0.01203	0.47	0.6396

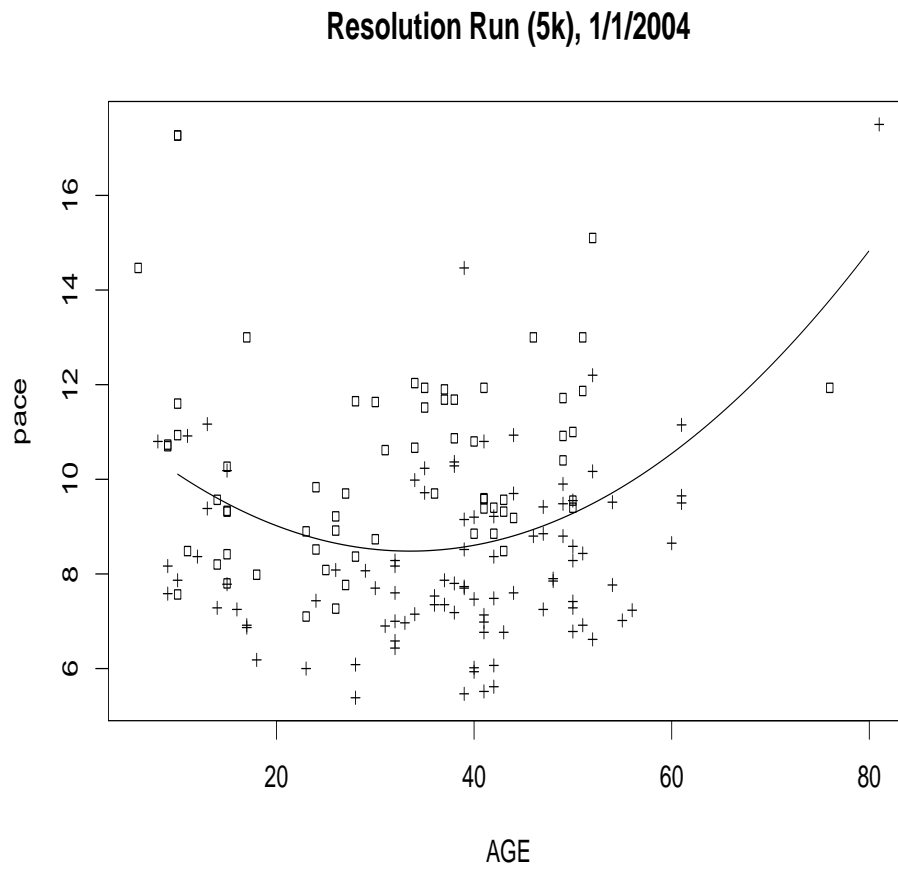
Model: MODEL2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	113.64500	56.82250	13.23	<.0001
Error	157	674.44972	4.29586		
Corrected Total	159	788.09472			

Root MSE	2.07265	R-Square	0.1442
Dependent Mean	9.12063	Adj R-Sq	0.1333

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	11.78503	0.70216	16.78	<.0001	13310
age	1	-0.19699	0.04113	-4.79	<.0001	1.09650
age2	1	0.00294	0.00057380	5.12	<.0001	112.54850



Fitted model is

$$\hat{\mu}(x) = 11.785 - 0.197x + 0.00294x^2$$

or

$$\hat{\mu}(\text{age}) = 11.785 - 0.197\text{age} + 0.00294\text{age}^2.$$

Inference for response Y given predictor x_i .

A photography outfit specializing in portraits of children operates studios in $n = 21$ cities. They're considering expansion into other cities. For city i , define

- x_{i1} : thousands of people aged ≤ 16
- x_{i2} : community disposable income per capita
- Y_i : company sales.

Given x_1 and x_2 , a MLR model for these data is given by

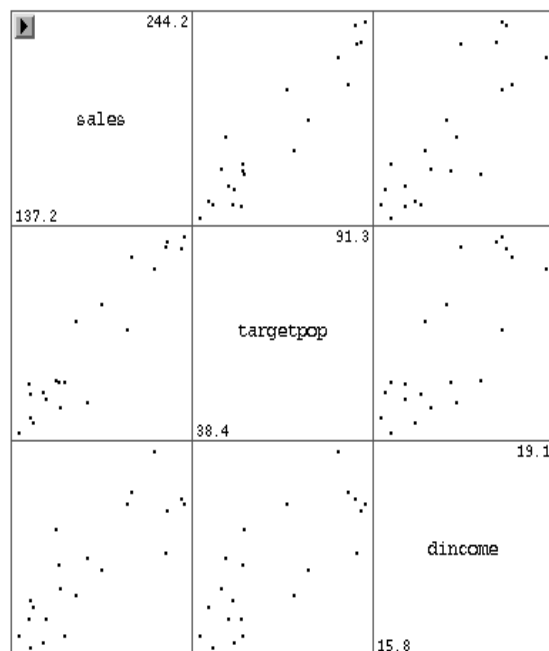
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + E_i \text{ for } i = 1, \dots, n$$

where errors are assumed iid normal w/ constant variance σ^2 .

For a city with x_1, x_2 the model for mean sales is

$$\mu(x_1, x_2) = E(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

A scatterplot matrix



Summary statistics (using SIMPLE option in PROC REG statement)

The REG Procedure					
Descriptive Statistics					
Uncorrected					
Variable	Sum	Mean	SS	Variance	Standard Deviation
Intercept	21.00000	1.00000	21.00000	0	0
x1	1302.40000	62.01905	87708	346.71662	18.62033
x2	360.00000	17.14286	6190.26000	0.94157	0.97035
y	3820.00000	181.90476	721072	1309.81048	36.19130

Some questions

Consider cities with $x_{01} = 90k$ kids and $x_{02} = 18k$ \$ per capita.

- Estimate the mean of the sales function among such cities, along with a standard error and 95% confidence interval.
- Obtain a 95% prediction interval of y_0 , the sales which would be observed for such an (individual) city.

SAS generates $\hat{\beta}$ and $\widehat{Var}(\hat{\beta}) = MS[E] * (X'X)^{-1}$

The SAS System

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-68.85707	60.01695	-1.15	0.2663
x1	1	1.45456	0.21178	6.87	<.0001
x2	1	9.36550	4.06396	2.30	0.0333

Covariance of Estimates

Variable	Intercept	x1	x2
Intercept	3602.0346743	8.7459395806	-241.4229923
x1	8.7459395806	0.0448515096	-0.672442604
x2	-241.4229923	-0.672442604	16.515755794

Moments of linear combinations of random vectors (Appendix B)

Let W denote a $p \times 1$ random vector with mean μ_W and covariance matrix Σ_W . Suppose a is a $p \times 1$ (fixed) vector of coefficients. Then

$$\begin{aligned} E(a'W) &= a'\mu_W \\ \text{Var}(a'W) &= a'\Sigma_W a. \end{aligned}$$

(See http://www.stat.ncsu.edu/people/dickey/st512/crsnotes/notes_1.htm for review of matrices and random vectors.)

Inference for the mean response in MLR

Consider all cities with 90000 youngsters and average disposable income of $18k$. To estimate mean sales and report a standard error, take $x'_0 = (1, 90, 18)$ and consider $\hat{\mu}(x_0) = x'_0 \hat{\beta}$.

$$\begin{aligned} E(x'_0 \hat{\beta}) &= x'_0 \beta \\ \text{Var}(x'_0 \hat{\beta}) &= x'_0 \hat{\Sigma} x_0 \end{aligned}$$

Substitution of $\hat{\beta}$ and $\hat{\Sigma} = MS[E](X'X)^{-1}$ gives the estimates:

$$\begin{aligned} \hat{\mu}(x_0) &= (1, 90, 18) \begin{pmatrix} -68.857 \\ 1.455 \\ 9.366 \end{pmatrix} \\ &= 230.7 \\ \widehat{\text{Var}}(\hat{\mu}(x_0)) &= (1, 90, 18) \begin{pmatrix} 3602.035 & 8.746 & -241.423 \\ 8.746 & 0.04485 & -0.67244 \\ -241.423 & -0.67244 & 16.516 \end{pmatrix} \begin{pmatrix} 1 \\ 90 \\ 18 \end{pmatrix} \\ &= 20.9 \\ \widehat{SE}(\hat{\mu}(x_0)) &= \sqrt{20.9} = 4.6 \end{aligned}$$

which can be obtained using PROC REG and the missing y trick:

```
Obs  targetpop  dincome    y    x0  x1  x2    p    semean    r
    1     90.0     18.0    .    1  90.0  18.0  230.632  4.55677  .
```

Partial correlations

The partial correlation coefficient for x_1 in the MLR

$$E(Y|x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

is defined as the correlation coefficient between the residuals computed from the two regressions below:

$$\begin{aligned} Y &= \beta_0 + \beta_2 x_2 + \dots + \beta_p x_p + E \\ X_1 &= \beta_0 + \beta_2 x_2 + \dots + \beta_p x_p + E \end{aligned}$$

Call these sets of residuals $e_{y \cdot 2,3,\dots,p}$ and $e_{1 \cdot 2,3,\dots,p}$ respectively. The *partial correlation* between y and x_1 after accounting for the linear association between y and x_2, x_3, \dots, x_p is defined as

$$r_{y1 \cdot 2,3,\dots,p} = \text{correlation between } e_{y \cdot 2,3,\dots,p} \text{ and } e_{1 \cdot 2,3,\dots,p}.$$

The *partial coeff. of determination* is $r_{y1 \cdot 2,3,\dots,p}^2$.

Note also (see Figure 11.7) that

$$r_{y1 \cdot 2,\dots,p}^2 = \frac{R(\beta_1 | \beta_0, \beta_2, \dots, \beta_p)}{SS[Total] - R(\beta_2, \beta_3, \dots, \beta_p | \beta_0)}.$$

Bodyfat data, compare models 1,2 and 6 (ignore x_3 .)

bodyfat data									
Obs	x1	x2	y	py_1	ey_1	e2_1	py_2	ey_2	e1_2
1	19.5	43.1	11.9	15.2190	-3.31903	-2.48145	13.2827	-1.38267	1.34939
2	24.7	49.8	22.8	19.6764	3.12360	-0.78756	19.0215	3.77847	0.60956
3	30.7	51.9	18.7	24.8195	-6.11952	-4.46384	20.8203	-2.12028	4.74782
4	29.8	54.3	20.1	24.0481	-3.94805	-1.19739	22.8760	-2.77599	1.72013
5	19.1	42.2	12.9	14.8762	-1.97616	-2.99637	12.5118	0.38822	1.74728
(abbreviated)									
20	25.2	51.0	21.1	20.1050	0.99500	-0.06892	20.0494	1.05061	0.04571

The partial correlation coefficient between y and x_1 after accounting for x_2 is $r_{y1.2} = 0.17$ and the partial for x_2 after accounting for x_1 is $r_{y2.1} = 0.48$. The partial coefficients of determination are

$$r_{y1.2}^2 = 0.03062 \text{ and } r_{y2.1}^2 = 0.23176.$$

Q: If you had to choose one variable or the other from x_1 and x_2 , which would it be?

Q: Anything wrong with throwing both x_1 and x_2 in the final model?

Q: Write the coefficients of determination in terms of extra sums of squares, using $R(\cdot|\cdot)$ notation.

Note: partial correlations obtained in SAS using **PCORR2** option:

Variable	DF	Corr	Squared Partial Type II
Intercept	1		.
x1	1	0.17	0.03062
x2	1	0.48	0.23176

Partial regression plots

A plot of the residuals from the regression

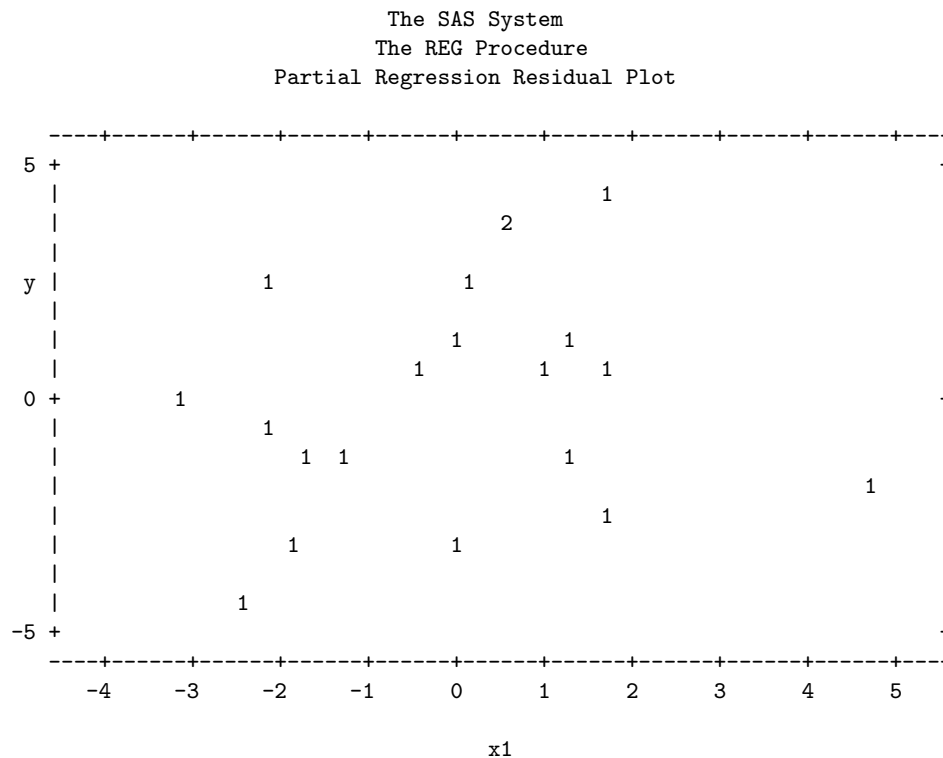
$$Y = \beta_0 + \beta_2 x_2 + \cdots + \beta_p x_p + E$$

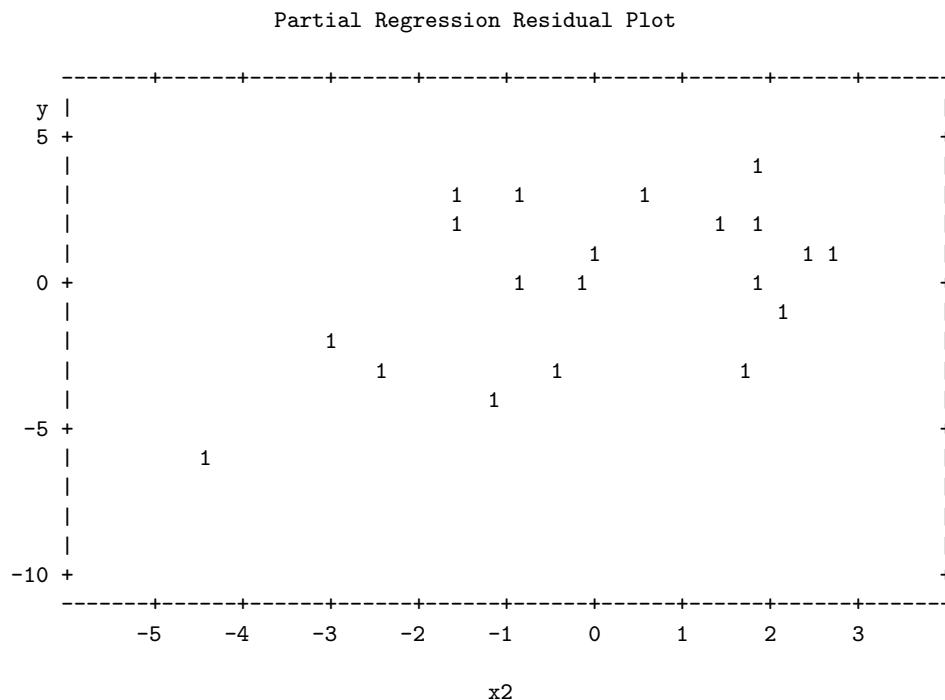
versus the residuals from the regression

$$X_1 = \beta_0 + \beta_2 x_2 + \cdots + \beta_p x_p + E$$

is called a partial regression plot for x_1 or a partial leverage plot of x_1 in the MLR. They can be generated

- in SAS/INSIGHT by clicking
 - Analyze • Fit XY • Output tab • Plots: (Partial Leverage).
- using the **PARTIAL** command in the MODEL statement of PROC REG





Q: What can these plots tell us?

A1: They convey info. about linear associations between y and a candidate independent variable x_i after accounting for linear associations between y and other independent variables $x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_p$

A2: They can convey info about nonlinear associations between y and x_i after accounting for linear associations with other variables.

A3: They can illuminate possible outliers.

Some exercises (hint: use matrix algebra or SAS).

1. Suppose you are a local 32 yr-old male runner. Regarding these data as randomly sampled from the population of all local runners, fit a quadratic regression function and use it to obtain an estimate of the mean 5k pace in your cohort of all 32 yr-old male runners. Report a standard error and 95% confidence interval.
2. Obtain a 95% prediction interval for your time if you are about to run the race.
3. Explain the difference between the two intervals in questions 1 and 2.
4. At what rate is $\mu(x)$ changing with age? Estimate the appropriate function.
5. Estimate θ , the peak age to run a 5k in the fastest time. Is θ a linear function of regression parameters? Can you obtain an unbiased estimate of the standard error of θ ?

Residual diagnostics

- Residuals can be plotted against independent variables to check for model inadequacy.
- Residuals can be plotted against predicted values to look for inhomogeneity of variance (heteroscedasticity).
- The sorted residuals can be plotted against the normal inverse of the empirical CDF of the residuals in a normal plot to assess the normal distributional assumption. A nonlinear association in such a q-q plot indicates nonnormality.

1. Obtain the observed quantiles by ordering the residuals:

$$e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(n)}.$$

2. For each $i = 1, \dots, n$ compute the expected quantile from

$$q_{(i)} = z\left(1 - \frac{i}{n+1}\right).$$

3. Plot the (ordered) residuals on the vertical axis versus the (ordered) theoretical quantiles on the horizontal axis.

As an illustration, we'll obtain the e_{ij} and the $q_{(i)}$ for the data in a table. To do this, we'll need the *ranks* of the residuals

$$\text{Rank of } e_{(i)} = i.$$

The *empirical cumulative probability* associated with $e_{(i)}$ is

$$p_i = \frac{\text{Rank of } e_{(i)}}{N+1}.$$

These can be used to obtain the corresponding theoretical quantiles via

$$q_{(i)} = z(1 - p_{(i)}).$$

```

proc reg simple outest=three;
  model pace=male age age2;
  output out=two p=yhat r=resid;
run;

proc rank out=two;
  ranks rankresid;
  var resid;
run;

data two;
  set two;
  ecdf=rankresid/130;
  q=probit(ecdf);
run;

proc print;
  var age pace yhat resid rankresid q;
run;

proc gplot;
  plot resid*q;
run;

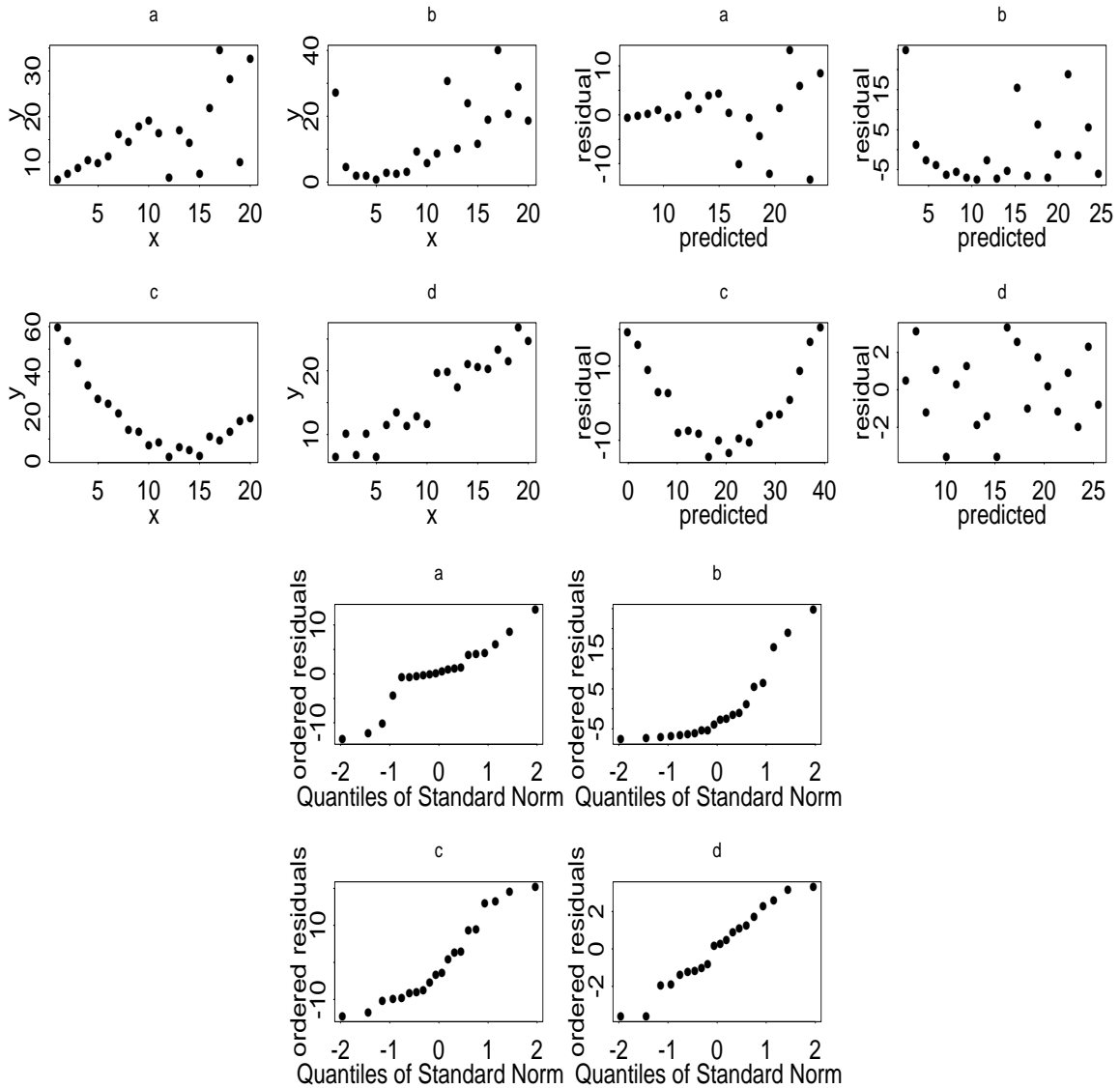
```

The SAS System

1

Obs	age	pace	yhat	resid	rankresid	q
1	21	5.26667	8.34197	-3.07531	2	-2.16004
2	33	5.28333	7.88456	-2.60123	8	-1.54199
3	19	5.40000	8.48871	-3.08871	1	-2.42320
4	20	5.41667	8.41283	-2.99616	3	-1.99398
5	20	5.48333	8.41283	-2.92949	5	-1.76883
			...			
38	46	8.10000	8.20740	-0.10740	66	0.01928
52	57	9.01667	9.14535	-0.12869	63	-0.03857
55	32	9.41667	7.89498	1.52169	108	0.95721
74	46	11.7500	8.2074	3.54260	128	2.16004
76	12	13.1667	9.1610	4.00572	129	2.42320
129	71	14.1333	12.4141	1.71922	112	1.08726

An exercise: Match up letters a,b,c,d with the model violation



1. Heteroscedasticity
2. Nonlinearity
3. Nonnormality
4. Model fits

ST 512

Weeks 4-5 The general linear model

Reading Ch. 8, 9, 12, 13

ANOVA revisited:

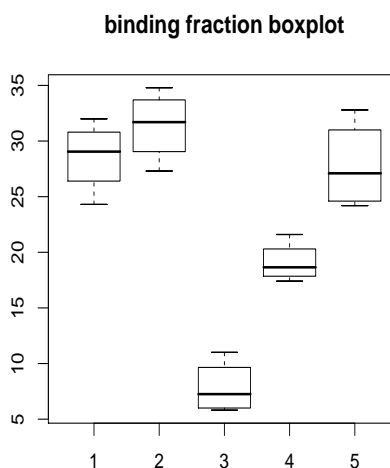
Following data come from study investigating binding fraction for several antibiotics using $n = 20$ bovine serum samples:

Antibiotic	Binding Percentage				Sample mean
Penicillin G	29.6	24.3	28.5	32	28.6
Tetracyclin	27.3	32.6	30.8	34.8	31.4
Streptomycin	5.8	6.2	11	8.3	7.8
Erythromycin	21.6	17.4	18.3	19	19.1
Chloramphenicol	29.2	32.8	25	24.2	27.8

A *completely randomized design* (CRD) was used.

Q: Are the population means for these 5 treatments plausibly equal?

Q: Do these (sample) treatment means differ significantly?



Modelling the binding fraction expt

One model parameterizes antibiotic effects as differences from mean:

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

for $i = 1, \dots, 5$ and $j = 1, \dots, 4$, where E_{ij} are i.i.d. $N(0, \sigma^2)$ errors.

Unknown parameters

- μ - overall population mean (avg of 5 treatment population means)
- τ_i - difference between (population) mean for treatment i and μ
- σ^2 - (population) variance of bf for a given antibiotic

To test $H_0 : \tau_1 = \tau_2 = \dots = \tau_5 = 0$, we just carry out one-way ANOVA:

Source	d.f.	Sum of squares	Mean Square	F
Treatments	4	1481	370	41
Error	15	136	9	
Total	19	1617		

Conclusion? (Use $F(0.05, 4, 15) = 3.06$)

Parameter estimates $\hat{\mu}, \hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3, \hat{\tau}_4, \hat{\tau}_5$? ($\bar{y}_{++} = 22.935$)?

Standard errors of parameter estimates?

Table for balanced one-way ANOVA

Y_{ij} denotes j^{th} observation receiving level i of treatment factor with t levels, for a total of N observations.

Source	d.f.	Sum of squares	Mean Square	F
Treatments	$t - 1$	$SS[T]$	$MS[T] = \frac{SS[T]}{(t-1)}$	$F = \frac{MS[T]}{MS[E]}$
Error	$N - t$	$SS[E]$	$MS[E] = \frac{SS[E]}{(N-t)}$	
Total	$N - 1$	$SS[TOT]$		

where

$$\begin{aligned}
 SS[T] &= \sum \sum (\bar{y}_{i+} - \bar{y}_{++})^2 \\
 SS[E] &= \sum \sum (y_{ij} - \bar{y}_{i+})^2 \\
 SS[TOT] &= \sum \sum (y_{ij} - \bar{y}_{++})^2
 \end{aligned}$$

The linear model $\mu_{ij} = E(Y_{ij}) = \mu + \tau_i$ could be fit using MLR with **5 indicator variables** x_1, \dots, x_5 for the 5 antibiotics. Let

$$x_{ij} = \begin{cases} 1 & \text{if treatment } j \\ 0 & \text{else} \end{cases}$$

The MLR model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + E_i \quad i = 1, \dots, 20$$

where a design matrix X of 1's and 0's of dimension (20×6) could be specified. Note that $\beta_0 = \mu$ and $\beta_j = \tau_j$.

- problem: $(X'X)^{-1}$ does not exist
- standard errors for parameter estimates ($\hat{\beta}$) can't be obtained
- model is *overparameterized* (6 parameters, 5 means)

A general linear model

Models which parameterize the effects of classification factors this way are general linear models. One-way ANOVA and linear regression models are general linear models. The linearity pertains to the parameters, not the explanatory variables.

Here, reparameterizing using $5 - 1$ indicator variables leads to a general linear model. Define x_1, x_2, x_3, x_4 as before. Then the MLR model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + E_i \quad i = 1, \dots, 20$$

where $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$. The X matrix looks like

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Remarks:

- $(X'X)^{-1}$ exists
- continuous covariates (as opposed to indicators) can be added and it is still a general linear model

For the one-way ANOVA,

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} 27.8 \\ 0.8 \\ 3.6 \\ -20.0 \\ -8.7 \end{pmatrix}$$

Estimates for the five treatment means obtained by substitution of $\hat{\beta}$ into $\mu(x_1, x_2, x_3, x_4) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4$

$$\begin{aligned} \hat{\mu}(1, 0, 0, 0) &= \hat{\beta}_0 + \hat{\beta}_1 = 28.6 \\ \hat{\mu}(0, 1, 0, 0) &= \hat{\beta}_0 + \hat{\beta}_2 = 31.4 \\ \hat{\mu}(0, 0, 1, 0) &= \hat{\beta}_0 + \hat{\beta}_3 = 7.8 \\ \hat{\mu}(0, 0, 0, 1) &= \hat{\beta}_0 + \hat{\beta}_4 = 19.1 \\ \hat{\mu}(0, 0, 0, 0) &= \hat{\beta}_0 = 27.8 \end{aligned}$$

(Compare with page 4 69.)

For standard errors, use $\hat{\Sigma}$:

$$\hat{\Sigma} = MS[E](X'X)^{-1} = \begin{pmatrix} 2.3 & -2.3 & -2.3 & -2.3 & -2.3 \\ & 4.5 & 2.3 & 2.3 & 2.3 \\ & & 4.5 & 2.3 & 2.3 \\ & & & 4.5 & 2.3 \\ & & & & 4.5 \end{pmatrix}$$

Let a, b, c, d be defined by

$$a' = (1, 1, 0, 0, 0), b' = (1, 0, 1, 0, 0), c' = (1, 0, 0, 1, 0), d' = (1, 0, 0, 0, 1).$$

Then

$$\begin{aligned} \hat{\mu}(1, 0, 0, 0) &= \hat{\beta}_0 + \hat{\beta}_1 = a'\hat{\beta} \\ \hat{\mu}(0, 1, 0, 0) &= \hat{\beta}_0 + \hat{\beta}_2 = b'\hat{\beta} \\ \hat{\mu}(0, 0, 1, 0) &= \hat{\beta}_0 + \hat{\beta}_3 = c'\hat{\beta} \\ \hat{\mu}(0, 0, 0, 1) &= \hat{\beta}_0 + \hat{\beta}_4 = d'\hat{\beta} \\ \hat{\mu}(0, 0, 0, 0) &= \hat{\beta}_0 = \hat{\beta}_0 \end{aligned}$$

and

$$a'\hat{\Sigma}a = b'\hat{\Sigma}b = c'\hat{\Sigma}c = d'\hat{\Sigma}d = \hat{\Sigma}_{11} = 2.3 = \widehat{\text{Var}}(\hat{\beta}_0) = \widehat{\text{Var}}(\hat{\beta}_0 + \hat{\beta}_j)$$

so the estimated SE for any sample treatment mean is $\sqrt{2.3} = 1.5$.

Recall from one-way ANOVA that

$$\widehat{SE}(\bar{y}_{i+}) = \sqrt{\frac{MS[E]}{n_i}} = \sqrt{\frac{9.1}{4}} = \sqrt{2.3} = 1.5$$

A general linear model for 5k times of men AND women
 (Resolution Run, Jan. 1, 2004, Centennial Campus)

Quadratic model $\mu(x) = \beta_0 + \beta_1x + \beta_1x^2$ was used for the association between mean pace for male runners and age x . Consider modelling female race times as well. How could the model be extended to incorporate sex differences? Let $x_2 = x^2$ and let an indicator variable x_3 be defined by

$$x_3 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

Some candidate models:

$$\mu(x_1, x_2, x_3) = \beta_0$$

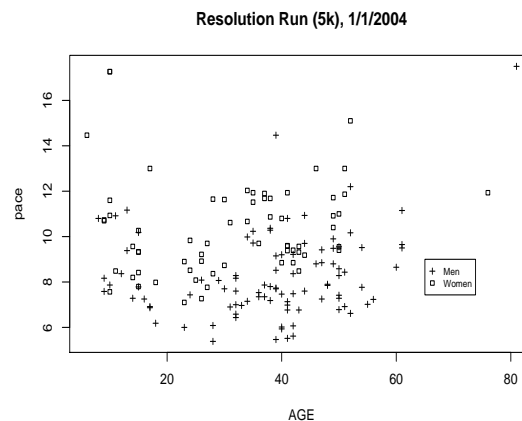
$$\mu(x_1, x_2, x_3) = \beta_0 + \beta_3x_3$$

$$\mu(x_1, x_2, x_3) = \beta_0 + \beta_1x_1$$

$$\mu(x_1, x_2, x_3) = \beta_0 + \beta_1x_1 + \beta_2x_2$$

$$\mu(x_1, x_2, x_3) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$$

$$\mu(x_1, x_2, x_3) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4 \underbrace{x_1x_3}_{x_4} + \beta_5 \underbrace{x_2x_3}_{x_5}$$



```

data race5k;
  set race5k;
  sexf=(sex="F");
  age2=age*age; agef=age*sexf; age2f=age2*sexf;
run;
proc reg data=one ;
  model pace=;
  model pace=sexf; /* equivalent to two-sample t-test */
  model pace=age age2;
  model pace=sexf age age2;
  model pace=sexf age age2 agef age2f;
  test agef=0, age2f=0;
run;

```

The REG Procedure
Model: MODEL1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	159	788.09472	4.95657		
Corrected Total	159	788.09472			

Root MSE 2.22634 R-Square 0.0000

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	9.12063	0.17601	51.82	<.0001

Model: MODEL2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	170.74137	170.74137	43.70	<.0001
Error	158	617.35335	3.90730		
Corrected Total	159	788.09472			

Root MSE 1.97669 R-Square 0.2167

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	8.26614	0.20280	40.76	<.0001
sexf	1	2.10335	0.31819	6.61	<.0001

(For MODEL3 output, see linear and quadratic fits from multiple regression notes)

Model: MODEL4

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	113.64500	56.82250	13.23	<.0001
Error	157	674.44972	4.29586		
Corrected Total	159	788.09472			

Root MSE 2.07265 R-Square 0.1442

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	11.78503	0.70216	16.78	<.0001
age	1	-0.19699	0.04113	-4.79	<.0001
age2	1	0.00294	0.00057380	5.12	<.0001

Model: MODEL5

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	290.34851	96.78284	30.33	<.0001
Error	156	497.74621	3.19068		
Corrected Total	159	788.09472			

Root MSE 1.78625 R-Square 0.3684

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	10.18317	0.64228	15.85	<.0001
sexf	1	2.19792	0.29535	7.44	<.0001
age	1	-0.17146	0.03562	-4.81	<.0001
age2	1	0.00281	0.00049481	5.67	<.0001

Fitted model is

$$\mu(x) = \begin{cases} \beta_0 + \beta_1 x + \beta_2 x^2 & \text{for men} \\ \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 & \text{for women} \end{cases}$$

$$= \begin{cases} \boxed{10.18 - 0.17x + 0.0028x^2} & \text{for men} \\ \boxed{10.18 + 2.20 - 0.17x + 0.0028x^2} & \text{for women} \end{cases}$$

Model: MODEL6

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	293.52828	58.70566	18.28	<.0001
Error	154	494.56644	3.21147		
Corrected Total	159	788.09472			

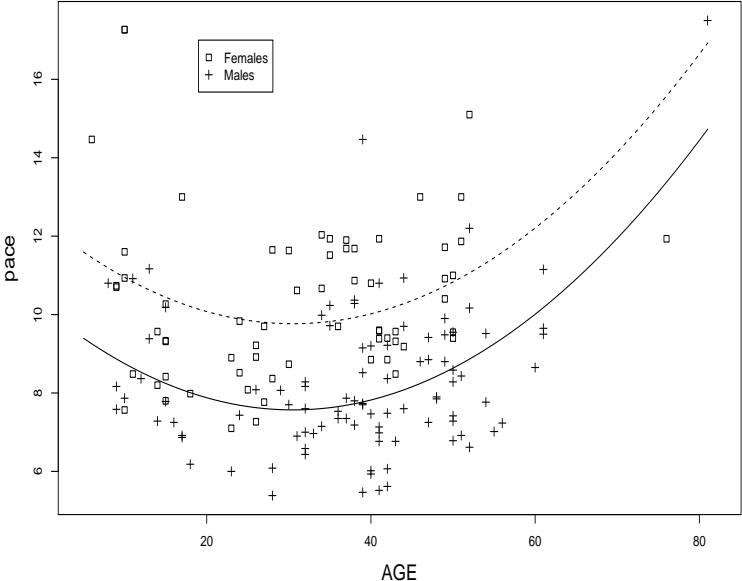
Root MSE 1.79206 R-Square 0.3725

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	10.60848	0.88641	11.97	<.0001
sexf	1	1.25728	1.23237	1.02	0.3092
age	1	-0.19986	0.04842	-4.13	<.0001
age2	1	0.00321	0.00064628	4.96	<.0001
agef	1	0.06882	0.07298	0.94	0.3471
age2f	1	-0.00103	0.00103	-0.99	0.3217

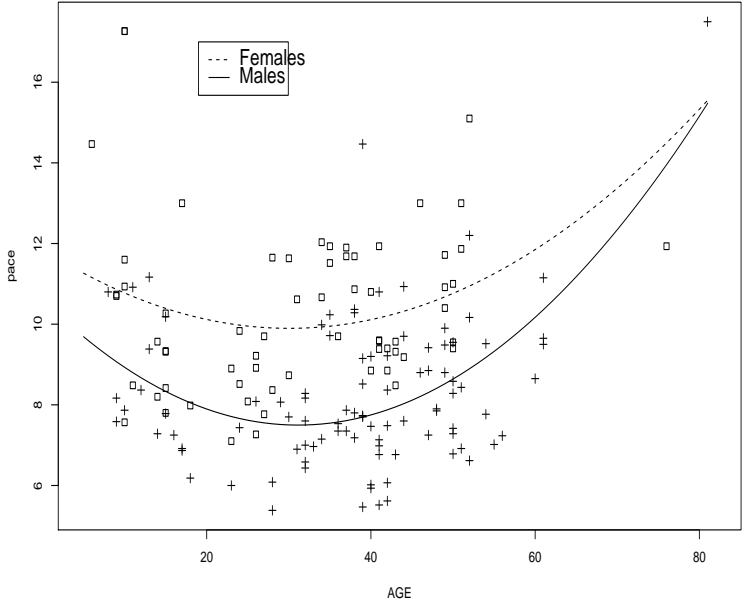
$$\mu(x) = \begin{cases} \beta_0 + \beta_1x + \beta_2x^2 + \beta_3(0) + \beta_4(0) + \beta_5(0) & \text{men} \\ \beta_0 + \beta_1x + \beta_2x^2 + \beta_3(1) + \beta_4(x) + \beta_5(x^2) & \text{women} \end{cases}$$

$$\begin{cases} 10.61 - 0.20x + 0.0032x^2 \\ 10.61 + 1.25 + (-0.20 + 0.07)x + (0.0032 - 0.0010)x^2 \\ 11.86 - 0.13x + 0.0022x^2 \end{cases}$$

Model 5



Model 6



Which model is “better”? What do we mean by “better?” Is there a test we can use to compare these models?

Comparison of models 5 and 6

reduced: $\mu(x_1, x_2, x_3) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$

full: $\mu(x_1, x_2, x_3) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_3 + \beta_5x_2x_3$

Extra sum of squares:

$$R(\beta_4, \beta_5 | \beta_0, \beta_1, \beta_2) = SS[R]_f - SS[R]_r = 293.5 - 290.3 = 3.0$$

The F -ratio

$$F = \frac{R(\beta_4, \beta_5 | \beta_0, \beta_1, \beta_2, \beta_3) / (5 - 3)}{MS[E]_f} = \frac{3.2/2}{3.21} = \frac{1.6}{3.21} = 0.5$$

The observed F -ratio is not significant on $df = 2, 154$.

In SAS, you could use

```
proc reg;
  model pace=age age2 sexf agef age2f;
  test agef=0, age2f=0;
run;
```

to get the following model selection F -ratio in the output:

The REG Procedure				
Model: MODEL6				
Test 1 Results for Dependent Variable pace				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	1.58988	0.50	0.6105
Denominator	154	3.21147		

Which model do we choose at this point?

More stuff

Estimate the “peak” running age for men and for women. Is it different? θ_M and θ_W denote peak running ages for men and women respectively. Using calculus on the model 6 regression,

$$\theta_M = \frac{-\beta_1}{2\beta_2} \quad \theta_W = \frac{-(\beta_1 + \beta_4)}{2(\beta_2 + \beta_5)}$$

These are nonlinear functions of regression parameters. Note that acceptance of any model but 6 implies equality of these peak ages.

$$\hat{\theta}_W = \begin{cases} 30.5 & \text{different intercepts model (5)} \\ 30.1 & \text{full model (6)} \end{cases}$$

$$\hat{\theta}_M = \begin{cases} 30.5 & \text{different intercepts model (5)} \\ 31.1 & \text{full model (6)} \end{cases}$$

Things to ponder:

Q: Which of these estimates is “better”?

Q: Which of these estimates are closest to the true peak(s)?

Q: What criterion can we use to assess the estimation?

Analysis of covariance, ANCOVA

Covariates are predictive responses. Associations between covariates z and the main response variable of interest y can be used to reduce unexplained variation σ^2 .

An nutrition example

A nutrition scientist conducted an experiment to evaluate the effects of four vitamin supplements on the weight gain of laboratory animals. The experiment was conducted in a completely randomized design with $N = 20$ animals randomized to $a = 4$ supplement groups, each with sample size $n \equiv 5$. The response variable of interest is weight gain, but calorie intake z was measured concomitantly.

Diet	$y(g)$	Diet	y	Diet	y	Diet	y
1	48	2	65	3	79	4	59
1	67	2	49	3	52	4	50
1	78	2	37	3	63	4	59
1	69	2	75	3	65	4	42
1	53	2	63	3	67	4	34
1	$\bar{y}_{1+} = 63$	2	$\bar{y}_{2+} = 57.8$	3	$\bar{y}_{3+} = 65.2$	4	$\bar{y}_{4+} = 48.8$
1	$s_1 = 12.3$	2	$s_2 = 14.9$	3	$s_3 = 9.7$	4	$s_4 = 10.9$

Q: Is there evidence of a vitamin supplement effect?

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	797.800000	265.933333	1.82	0.1836
Error	16	2334.400000	145.900000		
Corrected Total	19	3132.200000			

But calorie intake z was measured concomitantly:

Diet	y	z	Diet	y	z	Diet	y	z	Diet	y	z
1	48	350	2	65	400	3	79	510	4	59	530
1	67	440	2	49	450	3	52	410	4	50	520
1	78	440	2	37	370	3	63	470	4	59	520
1	69	510	2	73	530	3	65	470	4	42	510
1	53	470	2	63	420	3	67	480	4	34	430

Q: How and why could these new data be incorporated into analysis?

A: ANCOVA can be used to reduce unexplained variation.

Model, given z_i ,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_z z_i + E_i \quad \text{for } i = 1, \dots, 20$$

where x_{ij} is an indicator variable for subject i receiving vitamin supplement j :

$$x_{ij} = \begin{cases} 1 & \text{subject } i \text{ receives supplement } j \\ 0 & \text{else} \end{cases}$$

and errors $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$.

Exercise: specify the parametric mean weight gain for the first subject in each treatment group, conditional on their caloric intakes.

Proceeding with MLR analysis of this general linear model:

```

The GLM Procedure
Class Level Information
Class          Levels  Values
diet           4      1 2 3 4

```

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1951.680373	487.920093	6.20	0.0038
Error	15	1180.519627	78.701308		
Corrected Total	19	3132.200000			

R-Square	Coeff Var	Root MSE	y Mean
0.623102	15.11308	8.871376	58.70000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
diet	3	797.800000	265.933333	3.38	0.0463
z	1	1153.880373	1153.880373	14.66	0.0016

Source	DF	Type III SS	Mean Square	F Value	Pr > F
diet	3	1537.071659	512.357220	6.51	0.0049
z	1	1153.880373	1153.880373	14.66	0.0016

To test for a diet effect: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, use the type III F -ratio, on 3 and 15 numerator and denominator degrees of freedom. (Note that this is a comparison of nested models.)

Q: Conclusion?

FYI: this model was fit with the following code:

```

proc glm;
  class diet;
  model y=diet z;
  means diet;
  lsmeans diet/stderr;
run;

```

NOTE: the drop in \sqrt{MSE} (was $\hat{\sigma} \approx 12g$ is $\hat{\sigma} \approx 9g$)

Adjusted and unadjusted means

Recall the sample mean weight gains for the four diets (generated by the `means diet;` statement in `proc glm`):

The GLM Procedure

Level of diet	N	-----y-----		-----z-----	
		Mean	Std Dev	Mean	Std Dev
1	5	63.000000	12.2678441	442.000000	58.9067059
2	5	57.800000	14.8727940	434.000000	61.0737259
3	5	65.200000	9.6540147	468.000000	36.3318042
4	5	48.800000	10.8949530	502.000000	40.8656335

These means y are computed without taking z into account, so they are called *unadjusted* means.

Unadjusted means do not make any adjustment for the facts that

1. caloric intake may vary by diet (presumably by chance, not because of diet)
2. weight gain depends on caloric intake

Adjusted means are estimated mean weight gains at a common reference value (sample mean, \bar{z}) of the covariate, z .

Here, $\bar{z} = (442 + 434 + 468 + 502)/4 = 461.5$. The adjusted means are then just

$$\begin{aligned}\bar{y}_{1,a} &= \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_z(461.5) \\ \bar{y}_{2,a} &= \hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_z(461.5) \\ \bar{y}_{3,a} &= \hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_z(461.5) \\ \bar{y}_{4,a} &= \hat{\beta}_0 + \hat{\beta}_z(461.5)\end{aligned}$$

To get SAS to report the estimated regression parameter vector $\hat{\beta}$, use the **solution** option in the model statement. The default parameterization is the one we've adopted here where β_0 is the mean of the last level of the classification treatment factor:

Parameter		Estimate	Standard Error	t Value	Pr > t
Intercept		-35.66310108 B	22.41252629	-1.59	0.1324
diet	1	24.29519136 B	6.19932022	3.92	0.0014
diet	2	20.44121688 B	6.35678835	3.22	0.0058
diet	3	22.12060844 B	5.80625371	3.81	0.0017
diet	4	0.00000000 B	.	.	.
z		0.16825319	0.04394140	3.83	0.0016

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Substitution of $\hat{\beta}$ into the expressions for adjusted means yields

$$\bar{y}_{1,a} = -35.7 + 24.3 + 0.17(461.5) = 66.3$$

$$\bar{y}_{2,a} = -35.7 + 20.4 + 0.17(461.5) = 62.4$$

$$\bar{y}_{3,a} = -35.7 + 22.1 + 0.17(461.5) = 64.1$$

$$\bar{y}_{4,a} = -35.7 + 0.0 + 0.17(461.5) = 42.0$$

Standard errors of $\bar{y}_{j,a}$

Consider $\bar{y}_{2,a}$. What vector c is needed so that $c'\hat{\beta} = \bar{y}_{2,a}$?

What is the standard error of $c'\hat{\beta}$?

To get SAS to produce the adjusted means and estimated standard errors, use an `lsmeans` statement for the factor `diet`

```

The GLM Procedure
Least Squares Means

```

diet	y LSMEAN	Standard Error	Pr > t
1	66.2809372	4.0588750	<.0001
2	62.4269627	4.1473443	<.0001
3	64.1063543	3.9776677	<.0001
4	41.9857458	4.3482563	<.0001

Concerns:

Aside from the usual residual-based checks for model adequacy, does treatment affect the covariate? To check this, one could carry out a one-way ANOVA treating z as a response variable and check for a diet effect on the mean of z :

```

The GLM Procedure
Dependent Variable: z

```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model (diet)	3	14095.00000	4698.33333	1.84	0.1798
Error	16	40760.00000	2547.50000		
Corrected Total	19	54855.00000			

A: No evidence that treatment affects covariate.

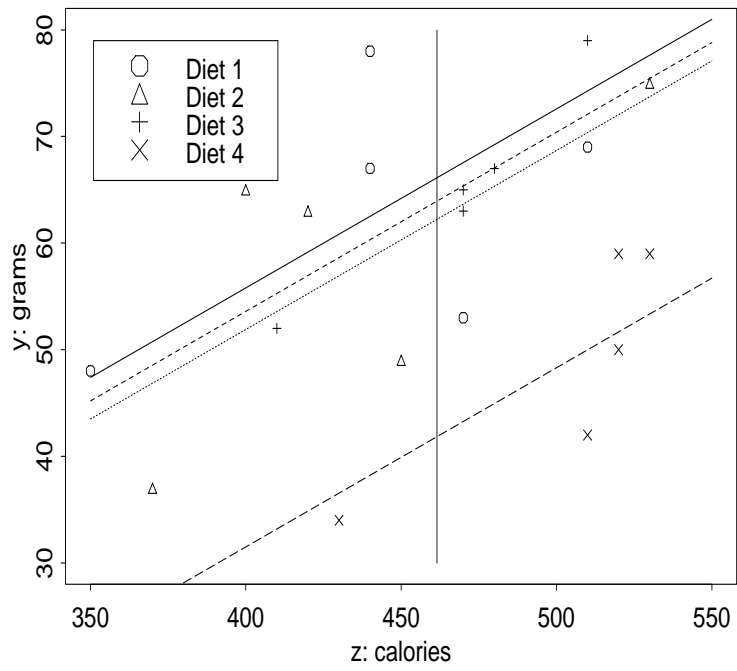
Q: Among the diets, which we've concluded are different, what are the differences? (Look at the means, have a guess.)

Q: If you are a lab animal and you want to gain weight, which diet(s) would you choose?

Q: Why are the standard errors for the adjusted means different?

Q: Which adjusted means require the most adjustment?

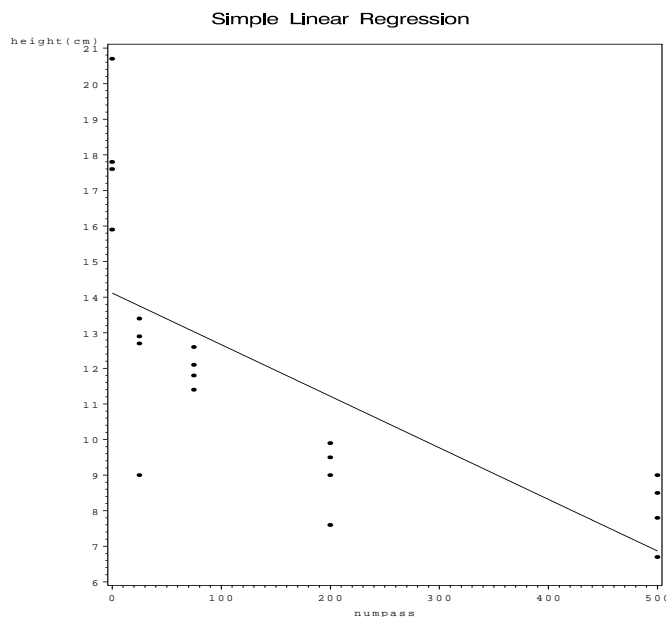
Vitamin supplement ANCOVA



Lack-of-fit of a SLR model (supplementary to textbook)

Hiking example: completely randomized experiment involving alpine meadows in the White Mountains of New Hampshire. $N = 20$ lanes of dimension $0.5m \times 1.5m$ randomized to 5 trampling treatments:

i : trt group	x : Number of passes	y_{ij} : Height(cm)
1	0	20.7 15.9 17.8 17.6
2	25	12.9 13.4 12.7 9.0
3	75	11.8 12.6 11.4 12.1
4	200	7.6 9.5 9.9 9.0
5	500	7.8 9.0 8.5 6.7



Two models for mean plant height:

$$\text{SLR model : } \mu(x) = \beta_0 + \beta_1 x$$

$$\text{one-factor ANOVA model : } \mu_{ij} = \mu + \tau_i$$

```
proc reg data=one;
  model y=numpass;
run;

proc glm data=one;
  class cnumpass;
  model y=numpass cnumpass;
run;
```

The SAS System 1
 The REG Procedure
 Dependent Variable: y height(cm)
 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	141.29532	141.29532	19.15	0.0004
Error	18	132.79418	7.37745		
Corrected Total	19	274.08950			

Root MSE	2.71615	R-Square	0.5155
Dependent Mean	11.79500	Adj R-Sq	0.4886

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	14.11334	0.80592	17.51	<.0001
numpass		1	-0.01449	0.00331	-4.38	0.0004

The GLM Procedure
 Class Level Information

Class	Levels	Values
cnumpass	5	0 25 75 200 500

Dependent Variable: y height(cm)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	243.1620000	60.7905000	29.48	<.0001
Error	15	30.9275000	2.0618333		
Corrected Total	19	274.0895000			

R-Square	0.887163	Coeff Var	12.17387	Root MSE	1.435909	y Mean	11.79500
----------	----------	-----------	----------	----------	----------	--------	----------

Source	DF	Type I SS	Mean Square	F Value	Pr > F
numpass	1	141.2953228	141.2953228	68.53	<.0001
cnumpass	3	101.8666772	33.9555591	16.47	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
numpass	0	0.0000000	.	.	.
cnumpass	3	101.8666772	33.9555591	16.47	<.0001

When the t treatments have an interval scale, the SLR model, and all polynomials of degree $p \leq t - 2$, are nested in one-factor ANOVA model with t treatment means.

F-ratio for lack-of-fit

To test for lack-of-fit of a polynomial (*reduced*) model of degree p , use extra sum-of-squares F-ratio on $t - 1 - p$ and $N - t$ *df*:

$$F = \frac{SS[\text{lack of fit}]/(t - 1 - p)}{MS[\text{pure error}]}$$

where

$$MS[\text{pure error}] = MS[E]_{full}$$

and

$$\begin{aligned} SS[\text{lack-of-fit}] &= SS[Trt] - SS[R]_{poly} \\ &= SS[E]_{poly} - SS[E]_{full} \\ &= SS[E]_{poly} - SS[\text{pure error}] \end{aligned}$$

In a simple linear ($p = 1$) model for the meadows data,

$$SS[\text{lack of fit}] = 243.163 - 141.295 = 101.867 \text{ on } t - 1 - p = 3df$$

and the sum of squares for pure error is $SS[E]_{full} = 30.93$ yielding

$$F = \frac{101.867/3}{30.93/15} \approx \frac{34}{2.1} = 16.5.$$

(highly significant since $F(0.01, 3, 15) = 5.42$.)

\implies model misspecified: SLR model suffers from lack of fit.

Next step: either go with the one-factor ANOVA model or specify some other model, such as quadratic.

ST 512

Weeks 7-8 Completely randomized factorial designs

Reading Ch. 9,13

- This packet
 - Introduction, notation, jargon:
Terms: factors, levels, treatments, treatment combinations, main effects, interaction effects, crossed factors, nested factors, contrasts, orthogonal contrasts, expected mean squares, multiplicity of comparisons, familywise or experimentwise error rates, power.
 - Specific topics:
 - * multiple comparisons,
 - * expected mean squares
 - * power computations
- Next packet
 - 2×2 experiments
 - $a \times b$ experiments
 - three-factor ANOVA
 - nested vs. crossed designs

Comparisons (contrasts) among means

Definition: In the one-way ANOVA layout:

$$Y_{ij} = \mu_i + E_{ij}, i = 1, 2, \dots, t, \text{ and } j = 1, 2, \dots, n_i$$

with $E_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$,

a linear function of the group means of the form

$$\theta = c_1\mu_1 + c_2\mu_2 + \dots + c_t\mu_t$$

is called a linear combination of the treatment means.

Definition: The c_j s are the coefficients of the linear combination.

If

$$c_1 + c_2 + \dots + c_t = \sum_1^t c_j = 0,$$

the linear combo is called a contrast.

Definition: Contrasts in which only two of the coefficients are nonzero are called simple or pairwise contrasts.

Definition: Contrasts in with more than two nonzero coefficients are called complex contrasts.

Result: The *best* estimator for a contrast of interest can be obtained by substituting treatment group sample means \bar{y}_{i+} for treatment population means μ_i in the contrast θ :

$$\hat{\theta} = c_1\bar{Y}_{1+} + c_2\bar{Y}_{2+} + \dots + c_t\bar{Y}_{t+}.$$

Example For the binding fraction data, consider the pairwise contrast comparing penicillin (population) mean to Tetracyclin mean:

$$\theta = \mu_1 - \mu_2 = (1)\mu_1 + (-1)\mu_2 + (0)\mu_3 + (0)\mu_4 + (0)\mu_5$$

Using the result, point estimator of θ is

$$\hat{\theta} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{Y}_{1+} - \bar{Y}_{2+}$$

Recall the binding fraction data and ANOVA table:

Antibiotic	Binding Percentage				Sample mean	Sample variance
Penicillin G	29.6	24.3	28.5	32	28.6	10.4
Tetracyclin	27.3	32.6	30.8	34.8	31.4	10.1
Streptomycin	5.8	6.2	11	8.3	7.8	5.7
Erythromycin	21.6	17.4	18.3	19	19.1	3.3
Chloramphenicol	29.2	32.8	25	24.2	27.8	15.9

Source	d.f.	Sum of squares	Mean Square	F
Treatments	4	1481	370	41
Error	15	136	9.05	
Total	19	1617		

Substitution of \bar{y}_{1+} and \bar{y}_{2+} yields $\hat{\theta} = 28.6 - 31.4 = -2.8$.

Q: How *good* is this estimate?

Sampling distribution of $\hat{\theta}$

Q: What is the sampling distribution of $\hat{\theta}$?

Q': That is, what are $E(\hat{\theta})$, $SE(\hat{\theta})$ and shape of distribution of $\hat{\theta}$?

$$\hat{\theta} \sim N(\theta, \text{Var}(\hat{\theta}))$$

Normality follows because $\hat{\theta}$ is a linear function of normal data Y_{ij} .

Standard error:

$$SE(\hat{\theta}) = \sqrt{\frac{c_1^2}{n_1}\sigma^2 + \frac{c_2^2}{n_2}\sigma^2 + \dots + \frac{c_t^2}{n_t}\sigma^2} = \sqrt{\sigma^2 \sum_{i=1}^{i=t} \frac{c_i^2}{n_i}}$$

estimated by

$$\hat{SE}(\hat{\theta}) = \sqrt{MS[E] \sum_{i=1}^{i=t} \frac{c_i^2}{n_i}}$$

To test $H_0 : \theta = \theta_0$ (often 0) versus $H_1 : \theta \neq \theta_0$ a t use t -test:

$$t = \frac{\text{est} - \text{null}}{\hat{SE}} = \frac{\hat{\theta} - \theta_0}{\hat{SE}(\hat{\theta})} \stackrel{H_0}{\sim} t_{N-t}.$$

At level α , the critical value for this test is $t(N - t, \alpha/2)$.

100(1 - α)% confidence interval for a contrast $\theta = \sum c_i \mu_i$ given by

$$\boxed{\sum c_i \bar{y}_{i+} \pm t(N - t, \alpha/2) \sqrt{MS[E] \sum \frac{c_i^2}{n_i}}}$$

Here,

$$\widehat{SE}(\hat{\theta}) = \sqrt{\left(\frac{1^2}{n_1} + \frac{(-1)^2}{n_2}\right) (9.05)} = \sqrt{\frac{9.05}{2}} = 2.127$$

So that the t statistic becomes

$$\frac{-2.8}{2.127} = -1.32$$

which is not in the critical region, so that the sample mean binding fractions for Penicillin G and Tetracyclin do not differ significantly.

A 95% confidence interval is given by

$$-2.8 \pm 2.13(2.127) \text{ or } (-7.3, 1.7)$$

Code (next page) estimates all pairwise contrasts involving Pen. G:

- $\theta_1 = a'\mu = (1, -1, 0, 0, 0)\mu$
- $\theta_2 = b'\mu = ?$
- $\theta_3 = c'\mu = ?$
- $\theta_4 = d'\mu = ?$

along with the complex contrast comparing Pen G. with mean of other four antibiotics:

$$\theta_5 = (\quad , \quad , \quad , \quad , \quad)\mu$$

Here $\mu = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)'$.

```

proc glm data=one;
  class drug;
  model y=drug/clparm;
  estimate "theta1" drug 1 -1;
  estimate "theta2" drug 1 0 -1;
  estimate "theta3" drug 1 0 0 -1;
  estimate "theta4" drug 1 0 0 0 -1;
  estimate "theta5" drug 4 -1 -1 -1 -1/divisor=4;
run;

```

The GLM Procedure
Class Level Information

Class	Levels	Values				
drug	5	1	2	3	4	5

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1480.823000	370.205750	40.88	<.0001
Error	15	135.822500	9.054833		
Corrected Total	19	1616.645500			

R-Square	Coeff Var	Root MSE	y Mean
0.915985	13.12023	3.009125	22.93500

Source	DF	Type III SS	Mean Square	F Value	Pr > F
drug	4	1480.823000	370.205750	40.88	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > t
theta1	-2.7750000	2.12777270	-1.30	0.2118
theta2	20.7750000	2.12777270	9.76	<.0001
theta3	9.5250000	2.12777270	4.48	0.0004
theta4	0.8000000	2.12777270	0.38	0.7122
theta5	7.0812500	1.68215202	4.21	0.0008

Parameter	95% Confidence Limits	
theta1	-7.3102402	1.7602402
theta2	16.2397598	25.3102402
theta3	4.9897598	14.0602402
theta4	-3.7352402	5.3352402
theta5	3.4958278	10.6666722

Orthogonal contrasts

In the same way the $SS[TOT]$ can be partitioned into independent components $SS[Trt]$ and $SS[E]$, the sum of squares for treatments, $SS[Trt]$ can be partitioned into $t - 1$ independent components.

Let two contrasts θ_1 and θ_2 be given by

$$\theta_1 = c_1\mu_1 + \cdots + c_t\mu_t \quad \text{and} \quad \theta_2 = d_1\mu_1 + \cdots + d_t\mu_t$$

or

$$\theta_1 = \sum_{i=1}^t c_i\mu_i \quad \text{and} \quad \theta_2 = \sum_{i=1}^t d_i\mu_i$$

Definition: The two contrasts θ_1 and θ_2 are mutually orthogonal if the products of their coefficients sum to zero:

$$c_1d_1 + \cdots + c_t d_t = \sum_{i=1}^t c_i d_i = 0$$

Consider several contrasts, say k of them: $\theta_1, \dots, \theta_k$. The *set* is mutually orthogonal if all pairs are mutually orthogonal.

Examples:

$$(-1, 1, 0, 0, 0) \text{ and } (0, 0, -1, 1, 0) \text{ orthogonal?}$$

$$(1, -1/2, -1/2, 0, 0) \text{ and } (0, 0, 0, -1, 1) \text{ orthogonal?}$$

$$(-1, 1, 0, 0, 0) \text{ and } (0, -1, 1, 0, 0) \text{ orthogonal?}$$

θ_i and θ_j orthogonal $\implies \hat{\theta}_i$ and $\hat{\theta}_j$ are statistically independent.

Sums of squares for contrasts

In the same way $SS[Trt]$ was obtained for a treatment effect, a sum of squares term can be obtained for a contrast:

$$SS[\hat{\theta}_1] = \frac{\hat{\theta}_1^2}{\left(\frac{c_1^2}{n_1} + \cdots + \frac{c_t^2}{n_t}\right)}$$

If $\theta_1, \dots, \theta_{t-1}$ are $t - 1$ mutually orthogonal contrasts, then

$$SS[Trt] = SS(\hat{\theta}_1) + SS(\hat{\theta}_2) + \cdots + SS(\hat{\theta}_{t-1})$$

There is one *df* associated with a sum of squares for an individual contrast $\hat{\theta}_j$ and if $\theta_j = 0$, then it can be shown that

$$E(SS[\hat{\theta}_j]) = \sigma^2.$$

To test $H_0 : \theta_j = 0$ versus $H_1 : \theta_j \neq 0$ use

$$F = \frac{SS[\hat{\theta}_j]}{MS[E]}$$

on 1 numerator degree of freedom and $N - t$ denominator degrees of freedom. For $\theta_1 = \mu_1 - \mu_2$ in the binding fractions,

$$F = \frac{(-2.8)^2}{MS[E] \left(\frac{1}{4} + \frac{(-1)^2}{4} + 0 + 0 + 0\right)} = 1.73.$$

(Using $F(0.05, 1, 15) = 4.54$, is the value $\theta_1 = 0$ plausible?)

A new dataset:

Number of contaminants in IV fluids made by $t = 3$ pharmaceutical companies

	Cutter	Abbott	McGaw
	255	105	577
	264	288	515
	342	98	214
	331	275	413
	234	221	401
	217	240	260
\bar{y}_{i+}	273.8	204.5	396.7

Source	d.f.	Sum of squares	Mean Square	F
Treatments (or pharmacies)	2	113646	56823	5.81
Error	15	146753	9784	
Total	17	260400		

Consider the following 2 contrasts:

$$\theta_1 = \mu_M - \mu_A \quad \text{and} \quad \theta_2 = \mu_C - \frac{\mu_M + \mu_A}{2}$$

Q: Are these contrasts orthogonal?

Q: Are the estimated contrasts $\hat{\theta}_1$ and $\hat{\theta}_2$ independent?

Exercise: Compute $SS[\hat{\theta}_1]$ and $SS[\hat{\theta}_2]$. Add em up.

```

proc format;
  value firmfmt 1="Cutter" 2="Abbott" 3="McGaw";
run;

data one;
  infile "pharmfirm.dat";
  input firm con;
  format firm firmfmt.;
run;

proc glm order=formatted;
  title "contaminant particles in IV fluids";
  class firm;
  model con=firm;
  contrast 'C - avg of M and A' firm -0.5 1 -0.5;
  contrast 'McGaw - Abbott' firm -1 0 1;
  estimate 'C - avg of M and A' firm -0.5 1 -0.5;
  estimate 'McGaw - Abbott' firm -1 0 1;
run;

```

contaminant particles in IV fluids
The GLM Procedure

1

Class	Levels	Values				
firm	3	Abbott Cutter McGaw				
Sum of						
Source	DF	Squares	Mean Square	F Value	Pr > F	
Model	2	113646.3333	56823.1667	5.81	0.0136	
Error	15	146753.6667	9783.5778			
Corrected Total	17	260400.0000				
R-Square Coeff Var Root MSE con Mean						
	0.436430	33.91268	98.91197	291.6667		
Contrast						
	DF	Contrast SS	Mean Square	F Value	Pr > F	
C - avg of M and A	1	2862.2500	2862.2500	0.29	0.5965	
McGaw - Abbott	1	110784.0833	110784.0833	11.32	0.0043	
Standard						
Parameter	Estimate	Error	t Value	Pr > t		
C - avg of M and A	-26.750000	49.4559849	-0.54	0.5965		
McGaw - Abbott	192.166667	57.1068524	3.37	0.0043		

Multiple Comparisons

- Can't go carrying out many many tests of significance willy-nilly
- e.g. consider the case with $t = 5$ (antibiotic treatments): all simple (pairwise) contrasts of the form $\theta = \mu_i - \mu_j$
- $\binom{5}{2} = 10$ tests of significance each at level $\alpha = 0.05$
- probability of committing at least one type I error ?

Definition: When testing k contrasts, the experimentwise error rate (or familywise) is

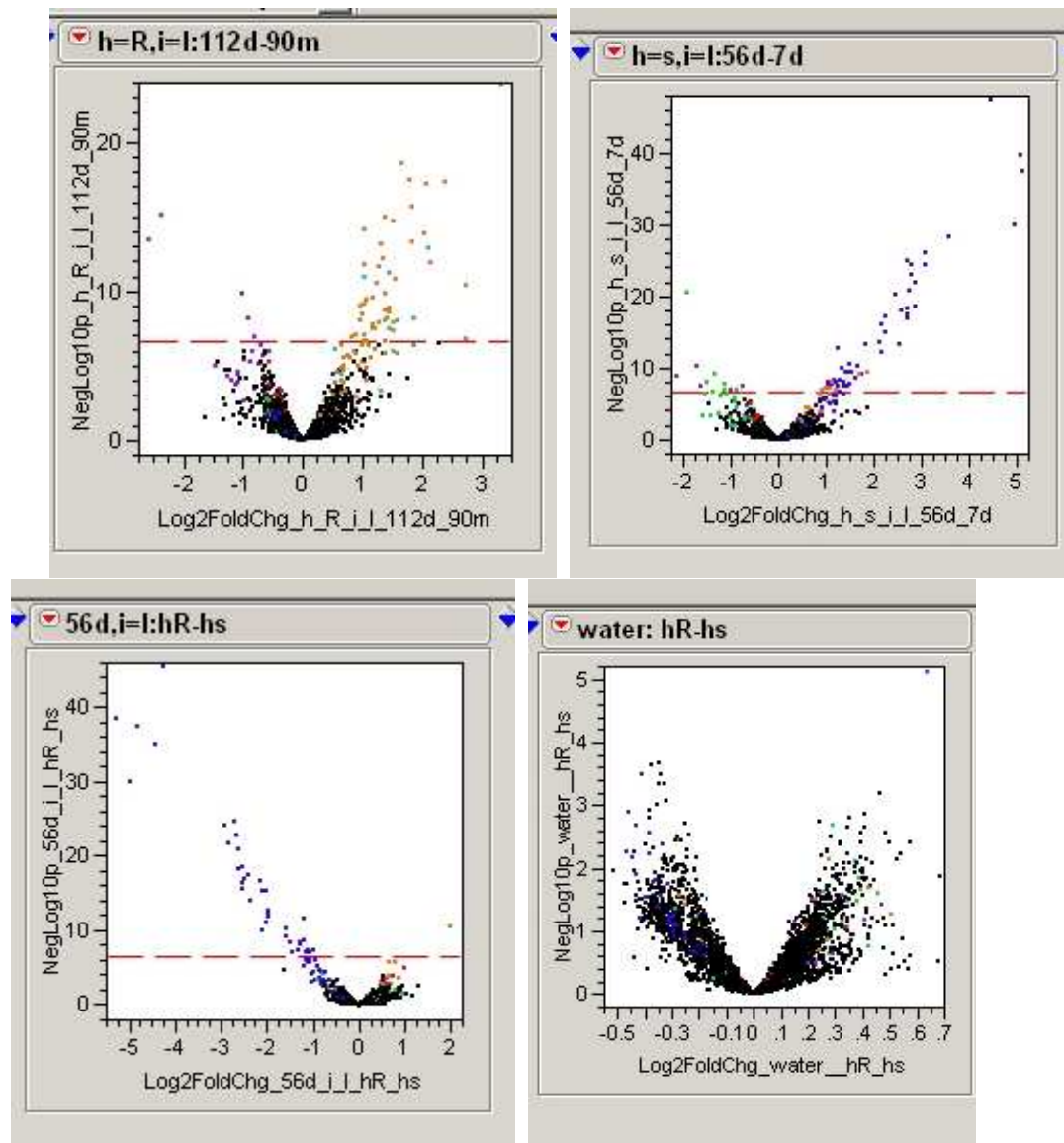
$$fwe = \Pr(\text{at least one type I error})$$

Methods for simultaneous inference for multiple contrasts include

- Scheffé
- Bonferroni
- Tukey

A context in which multiplicity is a big issue:

Microarray experiments, which may involve thousands of genes and tests



(Data courtesy of Cassi Myburg)

Bonferroni

Suppose interest lies in exactly k contrasts. The Bonferroni adjustment to α which controls *fw* is

$$\alpha' = \frac{\alpha}{k}$$

Simultaneous 95% confidence intervals for the k contrasts given by

$$a_1\bar{Y}_{1+} + a_2\bar{Y}_{2+} + \cdots + a_t\bar{Y}_{t+} \pm t\left(\frac{\alpha'}{2}, \nu\right) \sqrt{MS[E] \sum \frac{a_j^2}{n_j}}$$

and

$$b_1\bar{Y}_{1+} + b_2\bar{Y}_{2+} + \cdots + b_t\bar{Y}_{t+} \pm t\left(\frac{\alpha'}{2}, \nu\right) \sqrt{MS[E] \sum \frac{b_j^2}{n_j}}$$

⋮

$$k_1\bar{Y}_{1+} + k_2\bar{Y}_{2+} + \cdots + k_t\bar{Y}_{t+} \pm t\left(\frac{\alpha'}{2}, \nu\right) \sqrt{MS[E] \sum \frac{k_j^2}{n_j}}$$

where ν denotes *df* for error. $t(\frac{\alpha'}{2}, \nu)$ might have to be obtained using software.

For the binding fraction example, consider only pairwise comparisons with Penicillin:

$$\theta_1 = \mu_1 - \mu_2, \theta_2 = \mu_1 - \mu_3, \theta_3 = \mu_1 - \mu_4, \theta_4 = \mu_1 - \mu_5$$

We have $k = 4$, $\alpha' = 0.05/k = 0.0125$, and $t(\frac{\alpha'}{2}, 15) = 2.84$.

Substitution leads to

$$t(\alpha', 15) \sqrt{MS[E] \left(\frac{(-1)^2}{4} + \frac{(-1)^2}{4} + \frac{0^2}{4} + \cdots + \frac{0^2}{4} \right)} = 2.84 \sqrt{(9.05) \frac{2}{4}} = 6.0$$

so that simultaneous 95% confidence intervals for $\theta_1, \theta_2, \theta_3, \theta_4$ take the form

$$\bar{y}_1 - \bar{y}_i \pm 6.0$$

In SAS, an adjustment for $k = 4$ can be achieved with care:

```
proc glm data=one;
  title "Bonferroni correction for 4 contrasts";
  class drug;
  model y=drug/clparm alpha=.0125;
  estimate "theta1" drug -1 1;
  estimate "theta2" drug -1 0 1;
  estimate "theta3" drug -1 0 0 1;
  estimate "theta4" drug -1 0 0 0 1;
run;
```

```

          Bonferroni correction for 4 contrasts
                The GLM Procedure
          Class Level Information
```

Parameter	Estimate	Standard					t Value	Pr > t
		Error						
theta1	2.7750000	2.12777270					1.30	0.2118
theta2	-20.7750000	2.12777270					-9.76	<.0001
theta3	-9.5250000	2.12777270					-4.48	0.0004
theta4	-0.8000000	2.12777270					-0.38	0.7122

Parameter	98.75% Confidence Limits	
theta1	-3.2606985	8.8106985
theta2	-26.8106985	-14.7393015
theta3	-15.5606985	-3.4893015
theta4	-6.8356985	5.2356985

(actually simultaneous **95%** confidence intervals)

Another method: Scheffé

For **simultaneous** 95% confidence intervals for **ALL** contrasts, use

$$\boxed{\sum_1^t c_i \bar{y}_{i+} \pm \sqrt{(t-1)(F^*)MS[E] \sum_1^t \frac{c_i^2}{n_i}}}$$

where, $F^* = F(\alpha, t-1, N-t)$. For a pairwise comparisons of means, μ_j and μ_k , this yields

$$\bar{y}_{j+} - \bar{y}_{k+} \pm \sqrt{(t-1)(F^*)MS[E](1/n_j + 1/n_k)}$$

Using $\alpha = 0.05$, need to specify

- t (from the design)
- F^* (same critical value as for $H_0 : \tau_i \equiv 0$).
- $MS[E]$ (from the data)
- $\bar{y}_{i+}, \bar{y}_{j+}$
- n_i, n_j (from the data)

For binding fraction data,

$$\sqrt{(t-1)(F^*)MS[E] \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} = \sqrt{(5-1)(3.06)9.05 \left(\frac{1}{4} + \frac{1}{4} \right)} = 7.44$$

If any two sample means differ by more than 7.44, they differ significantly.

For IV fluids,

$$\bar{y}_{.A} = 204.5, \quad \bar{y}_{.M} = 396.67 \quad \bar{y}_{.C} = 273.83$$

and

$$\sqrt{(t-1)(F^*)MS[E] \left(\frac{1}{n_j} + \frac{1}{n_k} \right)} = \sqrt{(3-1)(3.68)9784 \left(\frac{1}{6} + \frac{1}{6} \right)} = 154.9$$

conclusion about pairwise contrasts? (compare w/ Bonferroni)

Tukey

Tukey's method is better than Scheffé's method when making all pairwise comparisons in balanced designs ($n = n_1 = n_2 = \dots = n_t$). It is conservative, controlling the experimentwise error rate, and has a lower type II error rate in these cases than Scheffé. (It is more powerful.)

For simple contrasts of the form

$$\theta = \mu_j - \mu_k$$

to test

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta \neq 0$$

reject H_0 at level α if

$$|\hat{\theta}| > q(t, N - t, \alpha) \sqrt{\frac{MS[E]}{n}}$$

where $q(t, N - t, \alpha)$ denotes α level *studentized range* for t means and $N - t$ degrees of freedom. These studentized ranges can be found in Table C.11 of Rao.

For the IV data, $q(3, 15, 0.05) = 3.67$. Tukey's 95% honestly significant difference (HSD) for pairwise comparisons of treatment means in this balanced design are

$$3.67 \sqrt{\frac{9784}{6}} = 148.3$$

```
proc glm;
  class firm;
  model con=firm;
  means firm/scheffe tukey;
run;
```

Tukey's Studentized Range (HSD) Test for con

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	15
Error Mean Square	9783.578
Critical Value of Studentized Range	3.67338
Minimum Significant Difference	148.33

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	firm
A	396.67	6	McGaw
B A	273.83	6	Cutter
B	204.50	6	Abbott

The GLM Procedure

Scheffe's Test for con

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	15
Error Mean Square	9783.578
Critical Value of F	3.68232
Minimum Significant Difference	154.98

Means with the same letter are not significantly different.

Scheffe Grouping	Mean	N	firm
A	396.67	6	McGaw
B A	273.83	6	Cutter
B	204.50	6	Abbott

Expected mean squares

Definition: The treatment mean square is given by

$$MS[Trt] = \frac{SS[Trt]}{t-1} = \frac{1}{t-1} \sum_i \sum_j (\bar{y}_{i+} - \bar{y}_{++})^2$$

$$\left(\bar{y}_{++} = \bar{y}_{++} \quad \bar{y}_{i+} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \right)$$

Q: Why are there $t-1$ degrees of freedom associated with $MS[Trt]$?

A: Note that the terms in the $\Sigma\Sigma$ above do not depend on j , so it is a sum of squares from t independent sample treatment means, leaving $t-1$ *df* for assessing variability.

It can be shown that

$$\begin{aligned} E[MS[Trt]; H_1] &= E[SS[Trt]/(t-1); H_1] \\ &= \sigma^2 + \frac{1}{t-1} \sum n_i (\mu_i - \mu)^2 \\ &= \sigma^2 + n \frac{1}{t-1} \sum (\mu_i - \mu)^2 \text{ (balanced case)} \\ &= \sigma^2 + n\psi_T^2 \end{aligned}$$

where

$$\psi_T^2 = \frac{1}{t-1} \sum (\mu_i - \mu)^2.$$

Note that under $H_0 : \mu_i \equiv \mu$ and $\psi_T^2 = 0$ so that

$$\begin{aligned} E[MS[Trt]; H_0] &= E[SS[Trt]/(t-1); H_0] \\ &= \sigma^2 \end{aligned}$$

Definition: The error mean square is given by

$$MS[E] = \frac{SS[E]}{N - t}$$

This is just a generalization of the pooled variance S_p^2 to the case of more than $t = 2$ groups:

$$\begin{aligned} MS[E] &= \frac{SS[E]}{N - t} \\ &= \frac{1}{N - t} \sum_{i=1}^t \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i+})^2 \\ &= \frac{1}{N - t} \sum_{i=1}^t (n_i - 1) s_i^2 \\ &= \left(\frac{n_1 - 1}{N - t} \right) s_1^2 + \left(\frac{n_2 - 1}{N - t} \right) s_2^2 + \cdots + \left(\frac{n_t - 1}{N - t} \right) s_t^2 \\ &= \text{“}S_p^2\text{”} \end{aligned}$$

Since $E(S_i^2) = \sigma^2$, $MS[E]$ is unbiased for σ^2 regardless of H_0 or H_1 :

$$\begin{aligned} E(S_i^2) &= \sigma^2 \\ \implies \\ E[MS[E]] &= \left(\frac{n_1 - 1}{N - t} \right) \sigma^2 + \left(\frac{n_2 - 1}{N - t} \right) \sigma^2 + \cdots + \left(\frac{n_t - 1}{N - t} \right) \sigma^2 \\ &= \sigma^2 \end{aligned}$$

$$\boxed{E[MS[E]] = \sigma^2}$$

Sample size computations for one-way ANOVA

Now consider the null hypothesis in a balanced experiment using one-way ANOVA to compare t treatment means and $\alpha = 0.05$:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t = \mu$$

versus the alternative

$$H_1 : \mu_i \neq \mu_j \text{ for some } i \neq j$$

Q: Suppose that we intend to use a balanced design. How big does our sample size $n_1 = n_2 = \dots = n_t = n$ need to be?

Of course, the answer depends on lots of things, namely, σ^2 and how many treatment groups t we have and how much of a difference among the means we hope to be able to detect, and with how big a probability.

Given α , μ_1, \dots, μ_t , and σ^2 , we can choose n to ensure a power of at least β using the *noncentral F distribution*.

Recall that the critical region for the statistic $F = MS[T]/MS[E]$ is everything bigger than $F(\alpha, t - 1, N - t) = F^*$.

The power of the F -test conducted using $\alpha = 0.05$ to reject H_0 under this alternative is given by

$$1 - \beta = \Pr(MS[T]/MS[E] > F^*; H_1 \text{ is true}). \quad (1)$$

Let $\tau_i = \mu_i - \mu$ for each treatment i so that

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$$

When some H_1 is true and the sample size n is used in each group, it can be shown that the F ratio has the noncentral F distribution with noncentrality parameter

$$\gamma = \sum_{j=1}^t n_j \left(\frac{\tau_j}{\sigma} \right)^2 = n \sum_{j=1}^t \left(\frac{\tau_j}{\sigma} \right)^2$$

This is the parameterization for the F distribution used in both SAS and S+.

One way to obtain an adequate sample size is trial and error. Software packages can be used to get probabilities of the form (1) for various values of n . Russ Lenth's website is also terrific and helpful:

<http://www.stat.uiowa.edu/~rlenth/Power/>

We'll write SAS code to do some computations but one could also use the procedure `PROC POWER` or other software such as R.

An example: suppose that a balanced completely randomized design (CRD) is to be used to test for a difference in the number of contaminant particles in IV fluid for three pharmaceutical companies. It is believed that the standard deviation on a given observation is about 100 particles for each company. In order to test $H_0 : \mu_1 = \mu_2 = \mu_3$ at level $\alpha = 0.05$, how large does the common sample size, n , need to be?

Q: “What alternative to H_0 would be meaningful? What is σ ? ”

A: The alternative $H_1 : \mu_1 = \mu_2 = (\mu - 30) = 230, \mu_3 = \mu + 60 = 320$ would be meaningful. Assume $\sigma \approx 100$.

Q: “What is an acceptable type II error rate, or what kind of power are we looking for?”

A: Suppose that $1 - \beta = 0.8$ should be good enough.

To obtain probabilities of the form (1) we need the noncentrality parameter γ :

$$\gamma = n\left[\left(\frac{\tau_1}{\sigma}\right)^2 + \left(\frac{\tau_2}{\sigma}\right)^2 + \left(\frac{\tau_3}{\sigma}\right)^2\right] = n[-30^2 + -30^2 + (60)^2]/100^2 = 0.54n$$

The $\alpha = 0.05$ critical value for H_0 is given by

$$F^* = F(3 - 1, 3(n - 1), 0.05).$$

We need the area to the right of F^* for the noncentral F distribution with degrees of freedom 2 and $3(n - 1)$ and noncentrality parameter $\gamma = 0.54n$. The following printout suggests the sufficiency of $n = 19$ for power of $1 - \beta = 0.8$

```

data one;
  do n=3 to 25; output; end;
run;
data one;
  set one;
  t=3;
  nu1=t-1;
  nu2=t*(n-1);
  sumtau2=(-30)**2 + (-30)**2 + 60**2;
  sigma2=10000;
  *sigma2u=(2/3)*var(100,100,190);
  *ncp=t*sigma2u/(2*sigma2);
  ncp=n*sumtau2/sigma2;
  qf=finv(0.95,nu1,nu2);
  pf=probf(qf,nu1,nu2,ncp);
  power=1-pf;
run;

proc print;run;

```

OBS	N	T	NU1	NU2	SUMTAU2	SIGMA2	SIGMA2U	NCP	QF	PF	POWER
1	3	3	2	6	5400	10000	1800	1.62	5.14325	0.86663	0.13337
2	4	3	2	9	5400	10000	1800	2.16	4.25649	0.81604	0.18396
3	5	3	2	12	5400	10000	1800	2.70	3.88529	0.76402	0.23598
4	6	3	2	15	5400	10000	1800	3.24	3.68232	0.71152	0.28848
5	7	3	2	18	5400	10000	1800	3.78	3.55456	0.65935	0.34065
6	8	3	2	21	5400	10000	1800	4.32	3.46680	0.60817	0.39183
7	9	3	2	24	5400	10000	1800	4.86	3.40283	0.55853	0.44147
8	10	3	2	27	5400	10000	1800	5.40	3.35413	0.51085	0.48915
9	11	3	2	30	5400	10000	1800	5.94	3.31583	0.46546	0.53454
10	12	3	2	33	5400	10000	1800	6.48	3.28492	0.42257	0.57743
11	13	3	2	36	5400	10000	1800	7.02	3.25945	0.38233	0.61767
12	14	3	2	39	5400	10000	1800	7.56	3.23810	0.34481	0.65519
13	15	3	2	42	5400	10000	1800	8.10	3.21994	0.31002	0.68998
14	16	3	2	45	5400	10000	1800	8.64	3.20432	0.27795	0.72205
15	17	3	2	48	5400	10000	1800	9.18	3.19073	0.24850	0.75150
16	18	3	2	51	5400	10000	1800	9.72	3.17880	0.22160	0.77840
17	19	3	2	54	5400	10000	1800	10.26	3.16825	0.19712	0.80288
18	20	3	2	57	5400	10000	1800	10.80	3.15884	0.17493	0.82507
19	21	3	2	60	5400	10000	1800	11.34	3.15041	0.15489	0.84511
20	22	3	2	63	5400	10000	1800	11.88	3.14281	0.13684	0.86316
21	23	3	2	66	5400	10000	1800	12.42	3.13592	0.12065	0.87935
22	24	3	2	69	5400	10000	1800	12.96	3.12964	0.10616	0.89384
23	25	3	2	72	5400	10000	1800	13.50	3.12391	0.09324	0.90676

Another example: suppose we want to test equal mean binding fractions among antibiotics against the alternative

$$H_1 : \mu_P = \mu + 3, \mu_T = \mu + 3, \mu_S = \mu - 6, \mu_E = \mu, \mu_C = \mu$$

so that

$$\tau_1 = 3, \tau_2 = 3, \tau_3 = -6, \tau_4 = \tau_5 = 0.$$

Assume $\sigma = 3$ and we need to use $\alpha = \beta = 0.05$.

The noncentrality parameter is given by

$$\gamma = n\left[\left(\frac{3}{3}\right)^2 + \left(\frac{3}{3}\right)^2 + \left(\frac{-6}{3}\right)^2\right].$$

The following code should do the trick

```
data one;
  do n=2 to 10; output; end;
run;
data one;
  set one;
  t=5; nu1=t-1; nu2=t*(n-1);
  sumtau2=3**2+3**2+(-6)**2;
  sigma2=9;
  ncp=n*sumtau2/sigma2;
  qf=finv(0.95,nu1,nu2);
  pf=probf(qf,nu1,nu2,ncp);
  power=1-pf;
run;
proc print;run;
```

OBS	N	T	NU1	NU2	SUMTAU2	SIGMA2	NCP	QF	PF	POWER
1	2	5	4	5	54	9	12	5.19217	0.59246	0.40754
2	3	5	4	10	54	9	18	3.47805	0.22465	0.77535
3	4	5	4	15	54	9	24	3.05557	0.06437	0.93563
4	5	5	4	20	54	9	30	2.86608	0.01533	0.98467
5	6	5	4	25	54	9	36	2.75871	0.00319	0.99681
6	7	5	4	30	54	9	42	2.68963	0.00060	0.99940
7	8	5	4	35	54	9	48	2.64147	0.00010	0.99990
8	9	5	4	40	54	9	54	2.60597	0.00002	0.99998
9	10	5	4	45	54	9	60	2.57874	0.00000	1.00000

Orthogonal polynomial contrasts

Example: poultry science experiment measures bodyweights of chickens from $a = 4$ diet groups, characterized by protein concentration in diet.

- Y : 21-day bodyweights of chickens
- completely randomized design with one factor, protein in diet, with four *equally spaced* levels.
- thanks to P. Plumstead for data.
- $n = 18$ pens, $N = 72$.

diet group	x : level of protein	diet mean \bar{y}_{i+}	diet std. dev. s_i	Tukey grouping
1	21.8	993	38	
2	23.5	1003	28	
3	25.2	1022	39	
4	26.9	1050	32	

One-way ANOVA table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	34311.7666	11437.2555	9.57	<.0001
Error	68	81279.4678	1195.2863		
Corrected Total	71	115591.2344			
R-Square	Coeff Var	Root MSE	AMBW21D Mean		
0.296837	3.399254	34.57291	1017.073		

Some omitted exam questions:

1. Sketch a plot of mean bodyweight at 21 days against protein content.
2. Consider the following three sample contrasts

$$\hat{\theta}_1 = -3\bar{y}_{1+} - \bar{y}_{2+} + \bar{y}_{3+} + 3\bar{y}_{4+}$$

$$\hat{\theta}_2 = \bar{y}_{1+} - \bar{y}_{2+} - \bar{y}_{3+} + \bar{y}_{4+}$$

$$\hat{\theta}_3 = -\bar{y}_{1+} + 3\bar{y}_{2+} - 3\bar{y}_{3+} + \bar{y}_{4+}$$

- (a) True/false: These estimated contrasts are orthogonal.
- (b) True/false: If contrast sums of squares are obtained, then

$$SS[\hat{\theta}_1] + SS[\hat{\theta}_2] + SS[\hat{\theta}_3] = SS[Trt]?$$

- (c) Report $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$.
 - (d) Provide an expression for the standard error of $\hat{\theta}_2$.
 - (e) Estimate the standard error of $\hat{\theta}_2$.
 - (f) Report $SS(\hat{\theta}_1)$.
3. Fitting the SLR model leads to $SS[Reg] = 32742$ and $SS[Tot] = 115591$.
 - (a) Report the F -ratio for testing for a lack-of-fit of the linear model.
 - (b) The appropriate critical value for a test with level $\alpha = 0.05$ is $F^* = 3.13$. Draw a conclusion about the adequacy of the linear model using $\alpha = 0.05$: is there evidence that the linear model is inadequate?

The contrasts in problem 2 are called *orthogonal polynomial contrasts*. The table below gives coefficients for orthogonal polynomial contrasts for balanced single-factor experiments with 3, 4, or 5 equally spaced levels.

Factor levels	Poly. Degree	contrast	Coefficients for					$SS(\hat{\theta}_i)$
			\bar{y}_{1+}	\bar{y}_{2+}	\bar{y}_{3+}	\bar{y}_{4+}	\bar{y}_{5+}	
3	1	$\hat{\theta}_1$	-1	0	1			$R(\beta_1 \beta_0)$
	2	$\hat{\theta}_2$	1	-2	1			$R(\beta_2 \beta_0, \beta_1)$
4	1	$\hat{\theta}_1$	-3	-1	1	3		$R(\beta_1 \beta_0)$
	2	$\hat{\theta}_2$	1	-1	-1	1		$R(\beta_2 \beta_0, \beta_1)$
	3	$\hat{\theta}_3$	-1	3	-3	1		$R(\beta_3 \beta_0, \beta_1, \beta_2)$
5	1	$\hat{\theta}_1$	-2	-1	0	1	2	?
	2	$\hat{\theta}_2$	2	-1	-2	-1	2	?
	3	$\hat{\theta}_3$	-1	2	0	-2	1	?
	4	$\hat{\theta}_4$	1	-4	6	-4	1	?

Rightmost column indicates extra SS in MLR of the form

$$\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots .$$

The contrast corresponding to a polynomial of degree p can be used to test for a p^{th} degree association:

- large $|\hat{\theta}_1|$ indicates linear association between y and x .
- large $|\hat{\theta}_2|$ indicates quadratic association between y and x .
- large $|\hat{\theta}_3|$ indicates cubic association between y and x .

This is computationally (and otherwise) easier than fitting polynomial regressions of various degrees.

```

Proc glm;
  title "protein concentration and chicken weights";
  class cp;
  MODEL AMBW21D=cp;
  contrast 'cp linear' cp -3 -1 1 3;
  contrast 'CP quadratic' CP 1 -1 -1 1;
  contrast 'CP cubic' CP -1 3 -3 1;
  contrast 'all three' CP -3 -1 1 3,
                    cp 1 -1 -1 1,
                    cp -1 3 -3 1;
  /* 'all three' tests that the 3-vector of contrasts is (0,0,0)' */
  estimate 'cp linear' cp -3 -1 1 3;
  estimate 'CP quadratic' CP 1 -1 -1 1;
  estimate 'CP cubic' CP -1 3 -3 1;
RUN;
proc glm; /* no class statement will fit regression model*/
  model ambw21d=cp cp*cp cp*cp*cp;
run;
proc glm;
  model ambw21d=cp;
run;

```

protein concentration and chicken weights 1
 The GLM Procedure

Class	Levels	Values
CP	4	21.8 23.5 25.2 26.9

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	34311.7666	11437.2555	9.57	<.0001
Error	68	81279.4678	1195.2863		
Corrected Total	71	115591.2344			

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
cp linear	1	32741.55648	32741.55648	27.39	<.0001
CP quadratic	1	1568.66674	1568.66674	1.31	0.2560
CP cubic	1	1.54337	1.54337	0.00	0.9714
all three	3	34311.76658	11437.25553	9.57	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > t
cp linear	190.734127	36.4430498	5.23	<.0001
CP quadratic	18.670635	16.2978273	1.15	0.2560
CP cubic	1.309524	36.4430498	0.04	0.9714

protein concentration and chicken weights

3

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	34311.7666	11437.2555	9.57	<.0001
Error	68	81279.4678	1195.2863		
Corrected Total	71	115591.2344			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CP	1	32741.55648	32741.55648	27.39	<.0001
CP*CP	1	1568.66674	1568.66674	1.31	0.2560
CP*CP*CP	1	1.54337	1.54337	0.00	0.9714

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	1060.706771	17690.23708	0.06	0.9524
CP	11.320308	2192.80559	0.01	0.9959
CP*CP	-1.630049	90.32122	-0.02	0.9857
CP*CP*CP	0.044424	1.23628	0.04	0.9714

protein concentration and chicken weights

6

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	32741.5565	32741.5565	27.66	<.0001
Error	70	82849.6779	1183.5668		
Corrected Total	71	115591.2344			

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	743.8748249	52.10077470	14.28	<.0001
CP	11.2196545	2.13317368	5.26	<.0001

Note MSE. Linear regression on x preferred to one-factor model for Plumstead's data. Multiple comparisons among treatment means might be unnecessary.

ST 512: Exptl Stats for Biol. Sciences II

Weeks 9-10 Multi-factor ANOVA

Problems Ch. 13

- 2×2 experiments
- $a \times b$ experiments
- three-factor ANOVA
- nested vs. crossed designs (not described in packet)

An example of a 2×2 study

Cholesterol measurements for random samples of $n_j \equiv 7$ people from four populations are given in the table below. The groups (cohorts) are defined as follows:

- I The population of women younger than 50
- II The population of men younger than 50
- III The population of women 50 years or older
- IV The population of men 50 years or older

Group	Cholesterol level							avg	std. dev.
I	221	213	202	183	185	197	162	$\bar{y}_I = 194.7$	$s = 20$
II	271	192	189	209	227	236	142	$\bar{y}_{II} = 209.4$	$s = 41$
III	262	193	224	201	161	178	265	$\bar{y}_{III} = 212.0$	$s = 40$
IV	192	253	248	278	232	267	289	$\bar{y}_{IV} = 251.3$	$s = 32$

One-way ANOVA Model:

$$\begin{aligned} Y_{ij} &= \mu_i + E_{ij} \\ &= \mu + \tau_i + E_{ij} \end{aligned}$$

$i = 1, 2, 3, 4$ $j = 1, 2, \dots, 7$ and E_{ij} i.i.d. $N(0, \sigma^2)$

Parameters: $\mu, \tau_1, \tau_2, \tau_3, \tau_4, \sigma^2$, with $\sum_1^4 \tau_i = 0$ constraint.

One-way ANOVA table:

The GLM Procedure

Class	Levels	Values			
cohort	4	I	II	III	IV
Number of observations		28			
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12280.85714	4093.61905	3.46	0.0323
Error	24	28434.57143	1184.77381		
Corrected Total	27	40715.42857			
	R-Square	Coeff Var	Root MSE	y Mean	
	0.301627	15.87245	34.42054	216.8571	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
cohort	3	12280.85714	4093.61905	3.46	0.0323

Conclusion so far is the cohort means, μ_i or $\mu + \tau_i$, are not plausibly equal (using $\alpha = 0.05$).

Some terminology

Definition: A factor in an experiment or study is a variable whose effect on the response is of primary interest. The values that a factor takes in the experiment are called factor levels or treatments.

Definition: In completely randomized designs experimental units are randomly assigned to factor levels, or treatment groups.

Note: The cholesterol study is NOT a completely randomized design, as randomization of subjects to different levels of AGE and GENDER isn't possible.

Definition: When the same number of units are used for each treatment, the design is balanced.

In one-way analysis of cholesterol data, COHORT is the only factor. This factor can be broken down into two factors in a two-way analysis: AGE (factor A) and GENDER (factor B).

Definition: If there are observations at all combinations of all factors, the design is complete, otherwise it is incomplete.

Exercise

1. Estimate the mean difference in cholesterol between young men and young women.
2. Estimate the mean difference between old men and old women.
3. Estimate the mean difference between men and women.
4. Estimate the mean difference between older and younger folks.
5. Estimate the mean difference between the differences estimated in 1. and 2.
6. Provide standard errors for all of these estimated contrasts
7. Specify the vectors defining these contrasts. For example, the first contrast of cohort means can be written

$$\theta_1 = (-1, 1, 0, 0)' \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} = \mu_2 - \mu_1$$

Consider the following contrasts of the cohort cholesterol means in the population:

$$\theta_3 = (-1, 1, -1, 1)' \mu$$

$$\theta_4 = (-1, -1, 1, 1)' \mu$$

$$\theta_5 = (-1, 1, 1, -1)' \mu$$

Q: Are these contrasts orthogonal?

Q: True/False: $SS(\hat{\theta}_3) + SS(\hat{\theta}_4) + SS(\hat{\theta}_5) = SS[Trt]$

Another exercise:

1. Compute the sums of squares for the estimated contrasts in 3., 4. and 5. using the exercise just completed and the fact that if $\hat{\theta} = \sum c_i \bar{y}_{i+}$ then

$$SS[\hat{\theta}] = \frac{\hat{\theta}^2}{\sum \frac{c_i^2}{n_i}}$$

2. Formulate a test of $H_0 : \theta_i = 0$ for each of these three contrasts. Obtain the F -ratio for each of these tests.
3. Obtain the $\alpha = 0.05$ critical region for each test. Compare the observed F -ratios to critical value and draw conclusions about
 - (a) an age effect
 - (b) a gender effect
 - (c) an age \times gender interaction

Types of effects

Two-way ANOVA model for the cholesterol measurements:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$$

$$i = 1, 2 = a \text{ and } j = 1, 2 = b \text{ and } k = 1, 2, \dots, 7 = n.$$

$E_{ijk} \stackrel{iid}{\sim} N(0, \sigma^2)$. Parameter constraints: $\sum_i \alpha_i = \sum_j \beta_j = 0$ and $\sum_i (\alpha\beta)_{ij} \equiv 0$ for each j and $\sum_j (\alpha\beta)_{ij} \equiv 0$ for each i .

Factor A: AGE has $a = 2$ levels - A_1 : younger and A_2 : older

Factor B: GENDER has $b = 2$ levels - B_1 : female and B_2 : male

Three kinds of effects in 2×2 designs:

1. Simple effects are simple contrasts.

- $\mu(A_1B) = \mu_{II} - \mu_I$ - simple effect of gender for young folks.
- $\mu(AB_1) = \mu_{III} - \mu_I$ - simple effect of age for women

2. Interaction effects are differences of simple effects:

$$\mu(AB) = \mu(AB_2) - \mu(AB_1) = (\mu_{IV} - \mu_{II} - (\mu_{III} - \mu_I))$$

- difference between simple age effects for men and women
- difference between simple gender effects for old and young folks
- interaction effect of AGE and GENDER.

3. Main effects are averages or sums of simple effects

$$\mu(A) = \frac{1}{2} (\mu(AB_1) + \mu(AB_2))$$

$$\mu(B) = \frac{1}{2} (\mu(A_1B) + \mu(A_2B))$$

Exercise: Classify the contrasts in the last exercise as simple, interaction or main effects.

Partitioning the treatment SS into $t - 1$ orthogonal components

$$12281 = SS[Trt] = SS[\hat{\theta}_3] + SS[\hat{\theta}_4] + SS[\hat{\theta}_5] = 5103 + 6121 + 1056$$

- $(a - 1)(b - 1)$ df for AB interaction
- $(a - 1)$ df for main effect of A
- $(b - 1)$ df for main effect of B

F test for interaction effect

To test for interaction,

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{21} = (\alpha\beta)_{22} = 0$$

vs.

$$H_1 : (\alpha\beta)_{ij} \neq 0 \text{ for some } i, j$$

use $\theta_5 = \mu(AB)$ and

$$F = \frac{SS(\hat{\theta}) / ((a - 1)(b - 1))}{MS[E]}$$

on 1 and $28 - 4 = 24$ numerator, denominator df . For cholesterol data the estimated interaction effect is

$$\hat{\theta}_5 = \hat{\mu}(AB) = (251.3 - 209.4) - (212 - 194.7) = 41.9 - 17.3 = 24.6$$

the associated sum of squares is

$$SS(\hat{\theta}_5) = \frac{(24.6)^2}{\frac{1}{7} + \frac{(-1)^2}{7} + \frac{(-1)^2}{7} + \frac{1}{7}} = \frac{(24.6)^2}{\frac{4}{7}} = 1056$$

and

$$F = 1056/1185 = 0.9$$

which isn't significant at $\alpha = 0.05$ on 1,24 df .

F test for main effects

To test for main effect of A: AGE

$$H_0 : \alpha_1 = \alpha_2 = 0 \text{ vs. } H_1 : \alpha_1 \neq 0 \text{ or } \alpha_2 \neq 0$$

use $\theta_4 = \mu(A)$ and

$$F = \frac{SS(\hat{\theta}_4)}{MS[E]}$$

on 1, 24 *df*. The estimated main effect of AGE is

$$\hat{\mu}(A) = \frac{(251.3 - 209.4)}{2} + \frac{(212 - 194.7)}{2} = \frac{59.2}{2} = 29.6$$

the associated sum of squares is

$$SS(\hat{\theta}_4) = \frac{(29.6)^2}{\frac{(\frac{1}{2})^2}{7} + \frac{(\frac{1}{2})^2}{7} + \frac{(\frac{1}{2})^2}{7} + \frac{(\frac{1}{2})^2}{7}} = \frac{(29.6)^2}{\frac{1}{7}} = 6121$$

and

$$F = 6121/1185 = 5.2$$

since $F(0.05, 1, 24) = 4.26$ AGE effect significant at $\alpha = 0.05$.

Similarly for the main effect of B: gender

$$H_0 : \beta_1 = \beta_2 = 0 \text{ vs. } H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

use $\theta_3 = \mu(B)$ on 1 and 24 *df*

$$\hat{\mu}(B) = \frac{209.4 - 194.7}{2} + \frac{251.3 - 212}{2} = 27$$

$$SS(\hat{\theta}_3) = \frac{(27)^2}{\frac{(\frac{1}{2})^2}{7} + \frac{(\frac{1}{2})^2}{7} + \frac{(\frac{1}{2})^2}{7} + \frac{(\frac{1}{2})^2}{7}} = \frac{(27)^2}{\frac{1}{7}} = 5103$$

and

$$F = 5103/1185 = 4.3$$

since $F(0.05, 1, 24) = 4.26$ GENDER effect significant at $\alpha = 0.05$.

Confidence intervals for effects

If $\theta = c'\mu$, $100(1 - \alpha)\%$ confidence interval given by

$$\hat{\theta} \pm t(\alpha/2, N - t) \sqrt{MS[E] \sum \frac{c_i^2}{n_i}}$$

For the cholesterol data, with $t(0.025, 24) = 2.06$ we have a 95% confidence interval for the AGE \times GENDER interaction effect:

$$24.6 \pm 2.06 \sqrt{\frac{4}{7} 1185} \quad \text{or} \quad 24.6 \pm 2.06(26.0) \quad \text{or} \quad (-29, 78)$$

a 95% confidence interval for the AGE effect:

$$29.6 \pm 2.06 \sqrt{\frac{1}{7} 1185} \quad \text{or} \quad 29.6 \pm 2.06(13.0) \quad \text{or} \quad (2.7, 56.4)$$

and a 95% confidence interval for the GENDER effect:

$$27.0 \pm 2.06 \sqrt{\frac{1}{7} 1185} \quad \text{or} \quad 27.0 \pm 2.06(13.0) \quad \text{or} \quad (0.15, 53.9).$$

The term under the $\sqrt{\quad}$ is the estimated standard error of the estimated contrast:

$$\widehat{SE}(\sum c_i \bar{y}_{i+}) = \sqrt{MS[E] \sum \frac{c_i^2}{n_i}}$$

SAS code for cholesterol problem

```

data one;
  input cohort $ @;
  do subj=1 to 7;
    input y @;
    if cohort="I" then do; gender="W"; age="y"; end;
    else if cohort="II" then do; gender="M"; age="y";end;
    else if cohort="III" then do; gender="W" ; age="o";end;
    else if cohort="IV" then do; gender="M" ; age="o";end;
    output;
  end;
cards;
I  221  213  202  183  185  197  162
II 271  192  189  209  227  236  142
III 262  193  224  201  161  178  265
IV  192  253  248  278  232  267  289
;
run;
proc glm;
  class cohort;
  model y=cohort/clparm;
  contrast "main effect of age  " cohort -1 -1 1 1;
  contrast "main effect of gender" cohort -1 1 -1 1;
  contrast "interaction effect  " cohort -1 1 1 -1;
  estimate "main effect of age  " cohort -1 -1 1 1/divisor=2;
  estimate "main effect of gender" cohort -1 1 -1 1/divisor=2;
  estimate "interaction effect  " cohort -1 1 1 -1;
run;

proc glm;
  class gender age;
  model y=age|gender;
run;

```

(SAS will overlook misspelling of **contrast**.)

SAS output (abbreviated) for cholesterol problem

The SAS System
The GLM Procedure

1

Class Level Information

Class	Levels	Values			
cohort	4	I	II	III	IV

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12280.85714	4093.61905	3.46	0.0323
Error	24	28434.57143	1184.77381		
Corrected Total	27	40715.42857			

R-Square	Coeff Var	Root MSE	y Mean
0.301627	15.87245	34.42054	216.8571

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
main effect of age	1	6121.285714	6121.285714	5.17	0.0323
main effect of gender	1	5103.000000	5103.000000	4.31	0.0488
interaction effect	1	1056.571429	1056.571429	0.89	0.3544

Parameter	Estimate	Standard Error	t Value	Pr > t
main effect of age	29.5714286	13.0097426	2.27	0.0323
main effect of gender	27.0000000	13.0097426	2.08	0.0488
interaction effect	-24.5714286	26.0194851	-0.94	0.3544

Parameter	95% Confidence Limits	
main effect of age	2.7206396	56.4222175
main effect of gender	0.1492111	53.8507889
interaction effect	-78.2730065	29.1301493

$a \times b$ designs

An example: Entomologist records energy expended (y) by $N = 27$ honeybees at $a = 3$ temperature (A) levels (20, 30, 40°C) consuming liquids with $b = 3$ levels of sucrose concentration (B) (20%, 40%, 60%) in a balanced, completely randomized crossed 3×3 design.

Temp	Suc	Sample		
20	20	3.1	3.7	4.7
20	40	5.5	6.7	7.3
20	60	7.9	9.2	9.3
30	20	6	6.9	7.5
30	40	11.5	12.9	13.4
30	60	17.5	15.8	14.7
40	20	7.7	8.3	9.5
40	40	15.7	14.3	15.9
40	60	19.1	18.0	19.9

The SAS System
The GLM Procedure

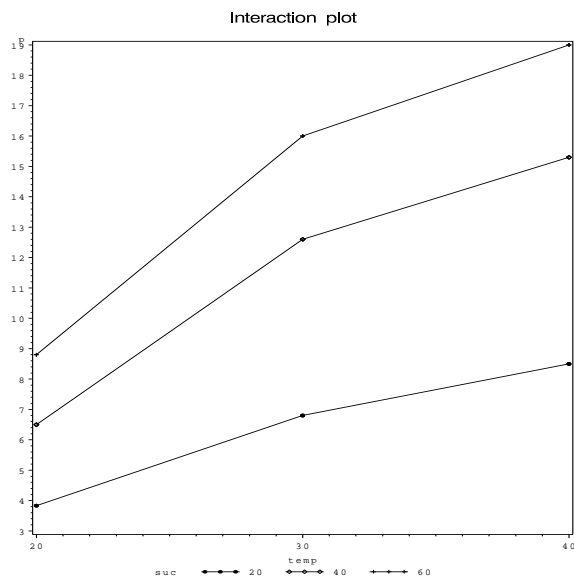
Class	Levels	Values
TEMP	3	20 30 40
SUC	3	20 40 60

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	630.2474074	78.7809259	87.07	<.0001
Error	18	16.2866667	0.9048148		
Corrected Total	26	646.5340741			

R-Square	Coeff Var	Root MSE	y Mean
0.974809	8.795505	0.951218	10.81481

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TEMP	2	293.1585185	146.5792593	162.00	<.0001
SUC	2	309.9585185	154.9792593	171.28	<.0001
TEMP*SUC	4	27.1303704	6.7825926	7.50	0.0010

3 × 3 honeybee example continued



Unlike 2×2 study, not possible to express interaction between factors A: TEMP and B: SUCROSE using a single number (w/ 1 *df*).

Level of TEMP	Level of SUC	n	Mean	SD
20	20	3	3.8333333	0.80829038
20	40	3	6.5000000	0.91651514
20	60	3	8.8000000	0.78102497
30	20	3	6.8000000	0.75498344
30	40	3	12.6000000	0.98488578
30	60	3	16.0000000	1.41067360
40	20	3	8.5000000	0.91651514
40	40	3	15.3000000	0.87177979
40	60	3	19.0000000	0.95393920

The plot above is called an interaction plot. In 2×2 designs, Rao distinguishes between qualitative and quantitative interactions, depending on whether or not the sign of the two simple effects is the same or different at the two levels of the other factor.

Exercise: Obtain an interaction plot for the cholesterol data. Characterize the observed interaction as qualitative or quantitative.

Partitioning $SS[Total]$ in $a \times b$ design

Two-way ANOVA Model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$$

$$(i = 1, 2, \dots, a \text{ and } j = 1, 2, \dots, b \text{ and } k = 1, 2, \dots, n)$$

Deviations:

total : $y_{ijk} - \bar{y}_{+++}$

due to level i of factor A: $\bar{y}_{i++} - \bar{y}_{+++}$

due to level j of factor B: $\bar{y}_{+j+} - \bar{y}_{+++}$

due to levels i of factor A and j of factor B after subtracting main effects:

$$\bar{y}_{ij+} - \bar{y}_{+++} - (\bar{y}_{i++} - \bar{y}_{+++}) - (\bar{y}_{+j+} - \bar{y}_{+++}) = \bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++}$$

$$SS[Total] = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{+++})^2$$

$$SS[A] = \sum_i \sum_j \sum_k (\bar{y}_{i++} - \bar{y}_{+++})^2$$

$$SS[B] = \sum_i \sum_j \sum_k (\bar{y}_{+j+} - \bar{y}_{+++})^2$$

$$SS[AB] = \sum_i \sum_j \sum_k (\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2$$

$$SS[E] = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij+})^2$$

$$y_{ijk} = \bar{y}_{+++} + (\bar{y}_{i++} - \bar{y}_{+++}) + (\bar{y}_{+j+} - \bar{y}_{+++}) + (\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++}) + y_{ijk} - \bar{y}_{ij+}$$

$$y_{ijk} - \bar{y}_{+++} = (\bar{y}_{i++} - \bar{y}_{+++}) + (\bar{y}_{+j+} - \bar{y}_{+++}) + (\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++}) + y_{ijk} - \bar{y}_{ij+}$$

Square both sides, \times -products vanish. (See output, p. 133, note that SS terms add up to model SS. Also, Type I and III SS equal.)

$a \times b$ example continued

Test for interaction effect generalizes from p.127:

$$H_0 : (\alpha\beta)_{ij} \equiv 0 \text{ vs. } H_1 : (\alpha\beta)_{ij} \neq 0 \text{ for some } i, j$$

$$F = \frac{MS[AB]}{MS[E]}$$

on $(a - 1)(b - 1)$ and $N - ab$ numerator, denominator df .

For honeybee data,

$$SS[AB] = n \sum_{i=1}^3 \sum_{j=1}^3 (\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2 = 27.1$$

$$F = \frac{27.1/4}{0.9} = 7.5$$

which is highly significant ($p = 0.001$) on 4,18 df .

We could proceed to test for main effects, but we won't.

Q: Why not?

A: Because effect of one factor depends on the level of the other factor, it doesn't make sense to talk about main effects.

If one insists on main effects, the appropriate F -ratios are

$$F_A = \frac{SS[A]/(a - 1)}{MS[E]} \text{ on } a - 1, N - ab \text{ } df$$

$$F_B = \frac{SS[B]/(b - 1)}{MS[E]} \text{ on } b - 1, N - ab \text{ } df$$

Another $a \times b$ design - no interaction

Yields on 36 tomato crops from balanced, complete, crossed design with $a = 3$ varieties (A) at $b = 4$ planting densities (B) :

Variety	Density $k/hectare$	Sample		
1	10	7.9	9.2	10.5
2	10	8.1	8.6	10.1
3	10	15.3	16.1	17.5
1	20	11.2	12.8	13.3
2	20	11.5	12.7	13.7
3	20	16.6	18.5	19.2
1	30	12.1	12.6	14.0
2	30	13.7	14.4	15.4
3	30	18.0	20.8	21.0
1	40	9.1	10.8	12.5
2	40	11.3	12.5	14.5
3	40	17.2	18.4	18.9

ANOVA table

The SAS System
The GLM Procedure

1

Class	Levels	Values
a	3	1 2 3
b	4	10 20 30 40

Number of observations 36

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	422.3155556	38.3923232	24.22	<.0001
Error	24	38.0400000	1.5850000		
Corrected Total	35	460.3555556			

R-Square	Coeff Var	Root MSE	y Mean
0.917368	9.064568	1.258968	13.88889

Source	DF	Type I SS	Mean Square	F Value	Pr > F
a	2	327.5972222	163.7986111	103.34	<.0001
b	3	86.6866667	28.8955556	18.23	<.0001
a*b	6	8.0316667	1.3386111	0.84	0.5484

Analysis of replicated two (or more) factor designs often proceed according to the following directions:

1. Check for interaction
2. If no interaction, analyze main effects
3. If interaction, analyze simple effects

Since there is no evidence of interaction, we proceed to analyze main effects. The F -ratios for factors A and B are each highly significant ($p < 0.0001$).

Level of		-----y-----	
a	N	Mean	Std Dev
1	12	11.3333333	1.88309867
2	12	12.2083333	2.34887142
3	12	18.1250000	1.73369023

Level of		-----y-----	
b	N	Mean	Std Dev
10	9	11.4777778	3.75458978
20	9	14.3888889	2.96835158
30	9	15.7777778	3.36480972
40	9	13.9111111	3.53250777

A conventional look at main effects is just to make pairwise comparisons among marginal means, after averaging over other factors. Pairwise comparisons of density means using Tukey's procedure with $\alpha = 0.05$ are given below.

(Use `means b/tukey;` to obtain the output.)

The GLM Procedure

Tukey's Studentized Range (HSD) Test for y

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	24
Error Mean Square	1.585
Critical Value of Studentized Range	3.90126
Minimum Significant Difference	1.6372

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	b
A	15.7778	9	30
A			
B A	14.3889	9	20
B			
B	13.9111	9	40
C	11.4778	9	10

A three-factor example

In a balanced, complete, crossed design, $N = 36$ shrimp were randomized to $abc = 12$ treatment combinations from the factors below:

A1: Temperature at 25° C

A2: Temperature at 35° C

B1: Density of shrimp population at 80 shrimp/40l

B2: Density of shrimp population at 160 shrimp/40l

C1: Salinity at 10 units

C2: Salinity at 25 units

C3: Salinity at 40 units

The response variable of interest is weight gain Y_{ijkl} after four weeks.

Three-way ANOVA Model:

$$\begin{aligned} Y_{ijkl} = & \mu + \alpha_i + \beta_j + \gamma_k \\ & + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \\ & + (\alpha\beta\gamma)_{ijk} + E_{ijkl} \end{aligned}$$

$$i = 1, 2$$

$$j = 1, 2$$

$$k = 1, 2, 3$$

$$l = 1, 2, 3$$

$$E_{ijkl} \stackrel{iid}{\sim} N(0, \sigma^2)$$

Many constraints such as:

$$\sum_i n_{i++} \alpha_i = \sum_j n_{+j+} \beta_j = \sum_k n_{++k} \gamma_k = 0$$

where n_{i++} denotes the number of observations at the i^{th} level of factor A .

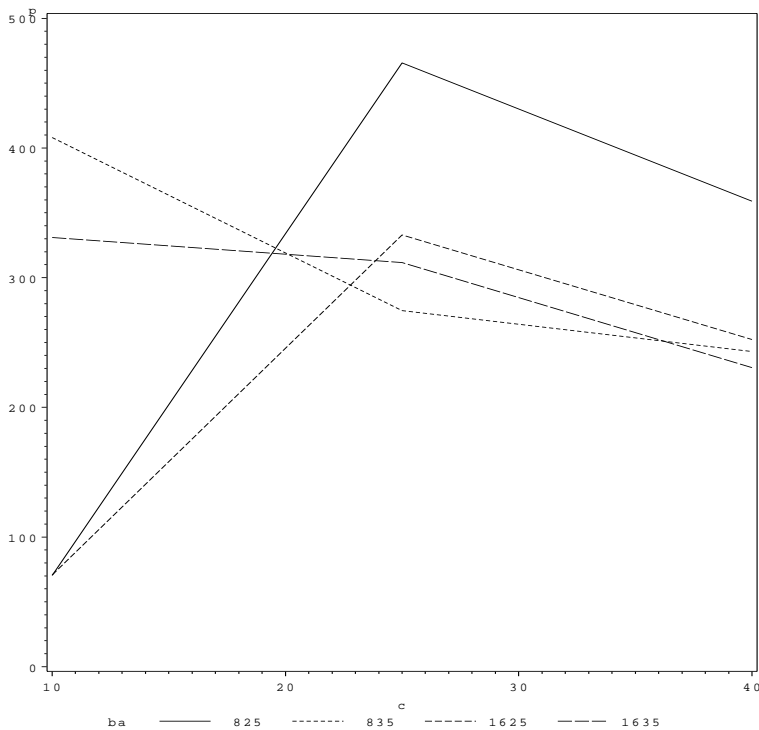
The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	467636.3333	42512.3939	14.64	<.0001
Error	24	69690.6667	2903.7778		
Corrected Total	35	537327.0000			

R-Square Coeff Var Root MSE y Mean
 0.870301 19.30270 53.88671 279.1667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
a	1	15376.0000	15376.0000	5.30	0.0304
b	1	21218.7778	21218.7778	7.31	0.0124
a*b	1	8711.1111	8711.1111	3.00	0.0961
c	2	96762.5000	48381.2500	16.66	<.0001
a*c	2	300855.1667	150427.5833	51.80	<.0001
b*c	2	674.3889	337.1944	0.12	0.8909
a*b*c	2	24038.3889	12019.1944	4.14	0.0285

Interaction plot for shrimp wt gains



Level of a	Level of b	N	Mean	Std Dev
25	80	9	298.333333	185.106051
25	160	9	218.666667	128.739077
35	80	9	308.555556	85.475305
35	160	9	291.111111	57.953525

Level of a	Level of c	N	Mean	Std Dev
25	10	6	70.500000	15.109600
25	25	6	399.333333	114.206246
25	40	6	305.666667	69.987618
35	10	6	369.500000	56.450864
35	25	6	293.166667	45.375838
35	40	6	236.833333	38.096807

Level of b	Level of c	N	Mean	Std Dev
80	10	6	239.166667	188.065326
80	25	6	370.166667	122.218520
80	40	6	301.000000	77.415761
160	10	6	200.833333	144.240655
160	25	6	322.333333	74.529636
160	40	6	241.500000	32.788718

Level of a	Level of b	Level of c	N	Mean	Std Dev
25	80	10	3	70.333333	17.156146
25	80	25	3	465.666667	87.648921
25	80	40	3	359.000000	59.858166
25	160	10	3	70.666667	16.623277
25	160	25	3	333.000000	108.282039
25	160	40	3	252.333333	11.372481
35	80	10	3	408.000000	51.117512
35	80	25	3	274.666667	47.961790
35	80	40	3	243.000000	36.166283
35	160	10	3	331.000000	30.116441
35	160	25	3	311.666667	42.665365
35	160	40	3	230.666667	46.971623

Interpretation of second order interaction

1st order interaction is between two factors

2nd order interaction is between three factors

Consider the AB interaction at each of three levels, C_1, C_2, C_3 .

To do this, look at three 2×2 tables as follows:

		B	
$(C = 1)$		B1	B2
A	A1	70	71
A2		408	331
		B	
$(C = 2)$		B1	B2
A	A1	466	333
A2		275	312
		B	
$(C = 3)$		B1	B2
A	A1	359.0	252
A2		243	231

Q: How is the ABC interaction manifested here?

A: We could compute $\hat{\mu}(ABC_1)$, $\hat{\mu}(ABC_2)$, $\hat{\mu}(ABC_3)$ and see if these first order interactions, with C fixed, are the same. (We know they are not by the F_{ABC} ratio and p -value.)

$$\hat{\mu}(ABC_1) = 408 - 70 - (331 - 71) \approx 77$$

Exercise: Obtain $\hat{\mu}(ABC_2)$ and $\hat{\mu}(ABC_3)$ as well as AB interaction plots for $C = 1$, $C = 2$ and $C = 3$. Interpret the plots.

Getting interaction contrasts using the ESTIMATE statement in GLM

To get SAS to estimate the interaction like $\mu(ABC_1)$, the AB interaction at $C = 1$, you must specify the parameters involved.

We saw on the last page that

$$\hat{\mu}(ABC_1) = \underbrace{\bar{y}_{211+} - \bar{y}_{111+}}_{\text{effect of } A \text{ at } B_1C_1} - \underbrace{\bar{y}_{221+} - \bar{y}_{121+}}_{\text{effect of } A \text{ at } B_2C_1}$$

Using the model to specify parameters, we can write

$$\begin{aligned} E(\bar{y}_{211+}) &= \mu + \alpha_2 + \beta_1 + \gamma_1 + (\alpha\beta)_{21} + (\alpha\gamma)_{21} + (\beta\gamma)_{11} + (\alpha\beta\gamma)_{211} \\ E(-\bar{y}_{111+}) &= -\mu - \alpha_1 - \beta_1 - \gamma_1 - (\alpha\beta)_{11} - (\alpha\gamma)_{11} - (\beta\gamma)_{11} - (\alpha\beta\gamma)_{111} \\ E(-\bar{y}_{221+}) &= -\mu - \alpha_2 - \beta_2 - \gamma_1 - (\alpha\beta)_{22} - (\alpha\gamma)_{21} - (\beta\gamma)_{21} - (\alpha\beta\gamma)_{221} \\ E(\bar{y}_{121+}) &= \mu + \alpha_1 + \beta_2 + \gamma_1 + (\alpha\beta)_{12} + (\alpha\gamma)_{11} + (\beta\gamma)_{21} + (\alpha\beta\gamma)_{121} \end{aligned}$$

Add these all up to get the contrast we're interested in, $\mu(ABC_1)$. Note that all terms vanish except the [second](#) order parameters and first order $(\alpha\beta)_{ij}$ parameters:

$$\begin{aligned} \mu(ABC_1) &= (\alpha\beta)_{21} - (\alpha\beta)_{11} - (\alpha\beta)_{22} + (\alpha\beta)_{12} \\ &\quad + (\alpha\beta\gamma)_{211} - (\alpha\beta\gamma)_{111} - (\alpha\beta\gamma)_{221} + (\alpha\beta\gamma)_{121} \end{aligned}$$

These can be rearranged so that they agree with the ordering of the treatment combinations employed by the ESTIMATE statement:

$$\begin{aligned} \mu(ABC_1) &= -(\alpha\beta)_{11} + (\alpha\beta)_{12} + (\alpha\beta)_{21} - (\alpha\beta)_{22} \\ &\quad - (\alpha\beta\gamma)_{111} + (\alpha\beta\gamma)_{121} + (\alpha\beta\gamma)_{211} - (\alpha\beta\gamma)_{221} \end{aligned}$$

ESTIMATE statement with two two-level factors

If A appears before B in the CLASS statement, SAS uses the following ordering for $(\alpha\beta)$ terms when specifying contrasts in ESTIMATE (or CONTRAST) statements:

$$A1B1, A1B2, A2B1, A2B2$$

ESTIMATE statement with two two-level and a three level factor

Similarly, with a **CLASS a b c;** statement, the following order is used for second-order interaction parameters

$$A1B1C1, A1B1C2, A1B1C3, A1B2C1, A1B2C2, A1B2C3, \dots \\ \dots, A2B1C1, A2B1C2, A2B1C3, A2B2C1, A2B2C2, A2B2C3$$

If $(\alpha\beta)$ is a vector of AB interaction effects with the default ordering:

$$(\alpha\beta) = \begin{pmatrix} (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{22} \end{pmatrix}$$

and likewise for $(\alpha\beta\gamma)$, then the contrast $\mu[ABC_1]$ on p. 25 can be written

$$\mu[ABC_1] = (-1, 1, 1, -1)(\alpha\beta) + (-1, 0, 0, 1, 0, 0, 1, 0, 0, -1, 0, 0)(\alpha\beta\gamma).$$

Similarly for $\mu[ABC_2]$ and $\mu[ABC_3]$:

$$\mu[ABC_2] = (-1, 1, 1, -1)(\alpha\beta) + (0, -1, 0, 0, 1, 0, 0, 1, 0, 0, -1, 0)(\alpha\beta\gamma).$$

$$\mu[ABC_3] = (-1, 1, 1, -1)(\alpha\beta) + (0, 0, -1, 0, 0, 1, 0, 0, 1, 0, 0, -1)(\alpha\beta\gamma).$$

```

proc glm;
  class a b c;
  model y=a|b|c;
  estimate "theta1: ABC1" a*b -1 1 1 -1
      a*b*c -1 0 0 1 0 0 1 0 0 -1 0 0;
  estimate "theta2: ABC2" a*b -1 1 1 -1
      a*b*c 0 -1 0 0 1 0 0 1 0 0 -1 0;
  estimate "theta3: ABC3" a*b -1 1 1 -1
      a*b*c 0 0 -1 0 0 1 0 0 1 0 0 -1;
  estimate "t1-av(t2+t3)" a*b*c -2 1 1 2 -1 -1 2 -1 -1 -2 1 1/divisor=2;
  means a|b|c;
run;

```

Parameter	Estimate	Standard Error	t Value	Pr > t
theta1: ABC1	77.333333	62.2230159	1.24	0.2259
theta2: ABC2	-169.666667	62.2230159	-2.73	0.0118
theta3: ABC3	-94.333333	62.2230159	-1.52	0.1426
t1-av(t2+t3)	209.333333	76.2073196	2.75	0.0112
t2-av(t1+t3)	-161.166667	76.2073196	-2.11	0.0450

The F_{ABC} ratio indicates the three contrasts estimated above are not plausibly ($p = 0.05$) equal. ($\hat{\mu}(ABC_1)$, $\hat{\mu}(ABC_2)$ and $\hat{\mu}(ABC_3)$ differ significantly). Interaction plots from p.24 suggest the comparison

$$\theta = \mu(ABC_1) - \frac{1}{2}(\mu(ABC_2) + \mu(ABC_3))$$

Adding up all the coefficients in this combination yields the contrast below to use with an ESTIMATE statement:

$$\left(-1, \frac{1}{2}, \frac{1}{2}, 1, -\frac{1}{2}, -\frac{1}{2}, 1, -\frac{1}{2}, -\frac{1}{2}, -1, \frac{1}{2}, \frac{1}{2}\right)$$

PS: Three-factor interactions are not easily interpreted. Effects can sometimes be made additive through a transformation of the response.

Activity w/ ESTIMATE statement

A linear function $\theta(\beta)$ of parameters is *estimable* if and only if there is a linear combination of Y whose expected value is θ .

Exercise: identify the estimable contrasts in each of the ESTIMATE statements in the correspondence below, which pertains to a 3×2 study with factors and levels

Factor	Levels
A : additive	acetic, nothing, sorbate
B : uv	0,1.

To: osborne@stat.ncsu.edu

Subject: non estimatable estimate statements

I am still having trouble with the estimate statements, the only ones that work for the additive*uv interaction are where we contrast the same additive over the uv, can anything be done about this??

```
proc glm;
  class additive uv;
  model ycount=additive uv uv*additive;
  estimate 'acetic uv=0 vs acetic uv=1' uv 1 -1 uv*additive 1 -1 0 0 0 0;
  estimate 'acetic uv=0 vs nothing uv=0' uv 1 -1 uv*additive 1 0 -1 0 0 0;
  estimate 'acetic uv=0 vs nothing uv=1' uv 1 -1 uv*additive 1 0 0 -1 0 0;
  estimate 'acetic uv=0 vs sorbate uv=0' uv 1 -1 uv*additive 1 0 0 0 -1 0;
  estimate 'acetic uv=0 vs sorbate uv=1' uv 1 -1 uv*additive 1 0 0 0 0 -1;
  estimate 'acetic uv=1 vs nothing uv=0' uv 1 -1 uv*additive 0 1 -1 0 0 0;
  estimate 'acetic uv=1 vs nothing uv=1' uv 1 -1 uv*additive 0 1 0 -1 0 0;
  estimate 'acetic uv=1 vs sorbate uv=0' uv 1 -1 uv*additive 0 1 0 0 -1 0;
  estimate 'acetic uv=1 vs sorbate uv=1' uv 1 -1 uv*additive 0 1 0 0 0 -1;
  estimate 'nothing uv=0 vs nothing uv=1' uv 1 -1 uv*additive 0 0 1 -1 0 0;
  estimate 'nothing uv=0 vs sorbate uv=0' uv 1 -1 uv*additive 0 0 1 0 -1 0;
  estimate 'nothing uv=0 vs sorbate uv=1' uv 1 -1 uv*additive 0 0 1 0 0 -1;
  estimate 'nothing uv=1 vs sorbate uv=0' uv 1 -1 uv*additive 0 0 0 1 -1 0;
  estimate 'nothing uv=1 vs sorbate uv=1' uv 1 -1 uv*additive 0 0 0 1 0 -1;
  estimate 'sorbate uv=0 vs sorbate uv=1' uv 1 -1 uv*additive 0 0 0 0 1 -1;
  estimate 'uv=0 vs uv=1' uv 1 -1;
  estimate 'acetic vs nothing' additive 1 -1;
  estimate 'acetic vs sorbate' additive 1 0 -1;
  estimate 'nothing vs sorbate' additive 0 1 -1;
```

ST 512: Exptl Stats for Biol. Sciences II - Fall, 2003

Dr. Jason A. Osborne

Supplement on design imbalance

Recall the 2×2 cholesterol study. Suppose the study is unbalanced and the data are given by

Age	Gender		Marginal mean
	Male	Female	
young	271,192,189,209, 227,236	162	$\bar{y}_{1++} = 212.3$
old	289	262,193,224,201 161,178,265	$\bar{y}_{2++} = 221.6$
	$\bar{y}_{+1+} = 230.4$	$\bar{y}_{+2+} = 205.8$	

$$\bar{y}_{11} = 220.7, \bar{y}_{12} = 162, \bar{y}_{21} = 289, \bar{y}_{22} = 212.$$

Consider an additive two-factor ANOVA model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$$

Exercise: finish parametric expressions for expected values below:

$$E(\bar{Y}_{1++}) = \mu + \alpha_1 + \frac{1}{7}(6\beta_1 + \beta_2)$$

$$E(\bar{Y}_{2++}) = \mu + \alpha_2 + \frac{1}{8}(\beta_1 + 7\beta_2)$$

$$E(\bar{Y}_{+1+}) =$$

$$E(\bar{Y}_{+2+}) =$$

Marginal sample means are not real useful in this unbalanced study.

Q: How are group population means estimated then?

A: *Least squares means* (what would be estimated by marginal means if design *were* balanced).

Parametric expressions for all the population means of interest are given below for the additive model:

Population group	effect of interest	estimate
Young folks	$\mu + \alpha_1 + \frac{1}{2}(\beta_1 + \beta_2)$	188.03
Older folks	$\mu + \alpha_2 + \frac{1}{2}(\beta_1 + \beta_2)$	
Men	$\mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \beta_1$	
Women	$\mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \beta_2$	
Young men	$\mu + \alpha_1 + \beta_1$	
Older men	$\mu + \alpha_2 + \beta_1$	
Young women	$\mu + \alpha_1 + \beta_2$	
Older women	$\mu + \alpha_2 + \beta_2$	

Invoking the command `lsmeans age gender;` will report least squares estimates for the first four means above.

The GLM Procedure
Least Squares Means

gender	y LSMEAN	Standard Error	Pr > t
m	251.525773	16.233482	<.0001
w	183.597938	15.842256	<.0001

age	y LSMEAN	Standard Error	Pr > t
jr	188.025773	16.233482	<.0001
sr	247.097938	15.842256	<.0001

All of these quantities are estimated using linear combinations of the treatment means of the form:

$$\hat{\theta} = c_{11}\bar{y}_{11+} + c_{12}\bar{y}_{12+} + c_{21}\bar{y}_{21+} + c_{22}\bar{y}_{22+}.$$

The coefficients are chosen so that $E(\hat{\theta}) = \theta$ and $\sum \frac{c_{ij}^2}{n_{ij}}$ is minimized.

Example: What are the coefficients for the contrast which estimates the population mean for young folks, $\mu + \alpha_1 + \frac{1}{2}(\beta_1 + \beta_2)$ with minimum variance?

$$\begin{aligned}c_{11} + c_{12} &= 1(\text{coeff for } \alpha_1) \\c_{21} + c_{22} &= 0(\text{coeff for } \alpha_2) \\c_{11} + c_{21} &= \frac{1}{2}(\text{coeff for } \beta_1) \\c_{12} + c_{22} &= \frac{1}{2}(\text{coeff for } \beta_2)\end{aligned}$$

Variance is then proportional to

$$\frac{c_{11}^2}{n_{11}} + \frac{c_{12}^2}{n_{12}} + \frac{c_{21}^2}{n_{21}} + \frac{c_{22}^2}{n_{22}} = \frac{c_{11}^2}{n_{11}} + \frac{(1 - c_{11})^2}{n_{12}} + \frac{(\frac{1}{2} - c_{11})^2}{n_{21}} + \frac{(c_{11} - \frac{1}{2})^2}{n_{22}}$$

which is minimized at $c_{11} = \frac{66}{97}$ by setting the derivative to zero and solving. The least squares mean is then

$$\hat{\theta} = \frac{66}{97}\bar{y}_{11+} + (1 - \frac{66}{97})\bar{y}_{12+} + (\frac{1}{2} - \frac{66}{97})\bar{y}_{21+} + (\frac{66}{97} - \frac{1}{2})\bar{y}_{22+} = 188.03.$$

Similarly for old folks, the contrast with minimum variance has

$$c_{11} = 18/97 = -c_{12}$$

and

$$c_{21} = \frac{1}{2} - 18/97, c_{22} = \frac{1}{2} + 18/97$$

so that the estimate for the old folks mean is

$$\frac{18}{97}\bar{y}_{11+} - \frac{18}{97}\bar{y}_{12+} + (\frac{1}{2} - \frac{18}{97})\bar{y}_{21+} + (\frac{1}{2} + \frac{18}{97})\bar{y}_{22+} = 247.1.$$

Exercise: obtain least squares estimators and estimates of marginal means for men and women as well as for each age \times gender combination.

Q: Is there an age effect? Should we base our conclusion on

$$\sum_i \sum_j \sum_k (\bar{y}_{i++} - \bar{y}_{+++})^2 = 325.6?$$

A: Might not be a good idea if factor B has an effect.

(This is the type I SS for Age if Age is the first factor entered into the model, or the so-called unadjusted sum of squares for age.)

Alternatively, consider the contrast

$$\theta_{age} = \alpha_1 - \alpha_2.$$

We can obtain the coefficients of the LS estimate of this contrast and then use them to get $SS[\hat{\theta}_{age}]$, which is the sum of squares for the age effect adjusted for gender, or type II sum of squares for age:

$$\hat{\theta}_{age} = \hat{\alpha}_1 - \hat{\alpha}_2 = c_{11}\bar{y}_{11+} + c_{12}\bar{y}_{12+} + c_{21}\bar{y}_{21+} + c_{22}\bar{y}_{22+}$$

where

$$\begin{aligned} c_{11} + c_{12} &= 1(\text{coeff for } \alpha_1) \\ c_{21} + c_{22} &= -1(\text{coeff for } \alpha_2) \\ c_{11} + c_{21} &= 0(\text{coeff for } \beta_1) \\ c_{12} + c_{22} &= 0(\text{coeff for } \beta_2) \end{aligned}$$

$\text{Var}(\hat{\theta}_{age})$ is minimized when $c_{11} = \frac{48}{97}$ which leads to

$$\hat{\theta}_{age} = -59.07$$

with

$$SS[\hat{\theta}_{age}] = \frac{(-59.07)^2}{\frac{(\frac{48}{97})^2}{6} + \frac{(-1 - \frac{48}{97})^2}{1} + \frac{(\frac{-48}{97})^2}{1} + \frac{(-1 + \frac{48}{97})^2}{7}} = 6044.$$

```

/*
I   221  213  202  183  185  197  162
II  271  192  189  209  227  236  142
III 262  193  224  201  161  178  265
IV  192  253  248  278  232  267  289
*/
options ls=75;
data one;
  input gender $ age $2. @;
  do i=1 to 7;
    input y @@;
    output;
  end;
cards;
w jr . . . . . 162
m jr 271 192 189 209 227 236 .
w sr 262 193 224 201 161 178 265
m sr . . . . . 289
;
run;

proc glm;
  class age gender;
  model y=age gender/solution;
  lsmeans gender age/stderr;
  estimate "lsmean for young folks" intercept 2 age 2 0 gender 1 1/divisor=2;
  estimate "lsmean for older folks" intercept 2 age 0 2 gender 1 1/divisor=2;
  estimate "lsmean for men" intercept 2 age 1 1 gender 2 0/divisor=2;
  estimate "lsmean for women" intercept 2 age 1 1 gender 0 2/divisor=2;
  estimate "lsmean for young men" intercept 1 age 1 0 gender 1 0;
  estimate "lsmean for young women" intercept 1 age 1 0 gender 0 1;
  estimate "lsmean for old men" intercept 1 age 0 1 gender 1 0;
  estimate "lsmean for old women" intercept 1 age 0 1 gender 0 1;
  contrast "age effect" age 1 -1;
  estimate "age effect" age 1 -1;
  contrast "gender effect" gender 1 -1;
  means gender age;
run;

```

The SAS System
The GLM Procedure
Class Level Information

1

Class	Levels	Values
age	2	jr sr
gender	2	m w

Number of observations 28

NOTE: Due to missing values, only 15 observations can be used in this analysis.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8318.06735	4159.03368	3.42	0.0669
Error	12	14606.86598	1217.23883		
Corrected Total	14	22924.93333			

R-Square	Coeff Var	Root MSE	y Mean
0.362839	16.05812	34.88895	217.2667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	325.629762	325.629762	0.27	0.6144
gender	1	7992.437592	7992.437592	6.57	0.0249

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	6044.348306	6044.348306	4.97	0.0457
gender	1	7992.437592	7992.437592	6.57	0.0249

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	213.1340206 B	12.77243521	16.69	<.0001
age jr	-59.0721649 B	26.50916484	-2.23	0.0457
age sr	0.0000000 B	.	.	.
gender m	67.9278351 B	26.50916484	2.56	0.0249
gender w	0.0000000 B	.	.	.

Least Squares Means

gender	y LSMEAN	Standard Error	Pr > t
m	251.525773	16.233482	<.0001
w	183.597938	15.842256	<.0001

age	y LSMEAN	Standard Error	Pr > t
jr	188.025773	16.233482	<.0001
sr	247.097938	15.842256	<.0001

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
age effect	1	6044.348306	6044.348306	4.97	0.0457
gender effect	1	7992.437592	7992.437592	6.57	0.0249

Parameter	Estimate	Standard Error	t Value	Pr > t
lsmean for young folks	188.025773	16.2334818	11.58	<.0001
lsmean for older folks	247.097938	15.8422561	15.60	<.0001
lsmean for men	251.525773	16.2334818	15.49	<.0001
lsmean for women	183.597938	15.8422561	11.59	<.0001
lsmean for young men	221.989691	13.7197962	16.18	<.0001
lsmean for young women	154.061856	26.2714097	5.86	<.0001
lsmean for old men	281.061856	26.2714097	10.70	<.0001
lsmean for old women	213.134021	12.7724352	16.69	<.0001
age effect	-59.072165	26.5091648	-2.23	0.0457

(These notes were adapted from one of Dr. Dickey's lectures:)

http://www.stat.ncsu.edu/~st512_info/dickey/crsnotes/rnotes_unbal.htm

Block Designs

Reading Ch. 15.3,15.4,15.6

Motivation - sometimes the variability of responses among experimental units is large, making detection of differences among treatment means $\mu_1, \mu_2, \dots, \mu_t$ difficult

In a randomized block design (RBD),

1. matched sets of experimental units are formed, each consisting of t units. Goal is reduced variability of response within a block. That is, the units within a block are homogeneous. Variance between blocks is ok.
2. Blocks are randomly assigned to each of the t treatments. (*Restricted randomization*, as opposed to a *completely randomized* design).

RBD - first example

Acrophobia can be treated in several ways.

- “Contact desensitization” - activity/task demonstrated then walked through while a therapist is in constant contact with the subject.
- “Demonstration participation” - therapist talks subject through task without any contact.
- “Live Modeling” - subject simply watches completion of task

Severity of acrophobia measured by HAT (Height Avoidance Test) scores. The point of the study is to investigate the effectiveness of the three therapies. The study will measure HAT scores before and after therapy. There is considerable heterogeneity in the degree to which acrophobia afflicts subjects. So $N = 15$ subjects will be put into **blocks** according to their original HAT score, then one from each block will be randomly assigned to a therapy: Let Y_{ij} denote the change in HAT score for subject in block j assigned to treatment i . (Bigger score means bigger reduction in fear.)

Block j	Therapy			\bar{y}_{+j}
	Contact Desensitization	Demonstration Participation	Live Modeling	
1	8	2	-2	2.67
2	$y_{12} = 11$	1	0	4
3	9	12	6	9
4	16	11	2	9.67
5	24	19	11	18
Avg \bar{y}_{i+}	13.6	9	3.4	

RBD example

Source	Sum of Squares	d.f.	Mean Square	F
A: Therapies	260.9	2	130.5	15.3
B: Blocks	438	4	109.5	12.8
Error	68.4	8	8.6	
Total	767.3	14		

(Data taken from Larsen and Marx, 1986)

$$SS[Total] = SS[A] + SS[B] + SS[E]$$

$$SS[Total] = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{++})^2$$

$$SS[A] = \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{i+} - \bar{y}_{++})^2 = b \sum_{i=1}^a (\bar{y}_{i+} - \bar{y}_{++})^2$$

$$SS[B] = \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{+j} - \bar{y}_{++})^2 = a \sum_{j=1}^b (\bar{y}_{+j} - \bar{y}_{++})^2$$

$$SS[E] = \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{++})^2$$

Note that

$$y_{ij} - \bar{y}_{++} = \underbrace{(\bar{y}_{i+} - \bar{y}_{++})}_{\text{therapy effect}} + \underbrace{(\bar{y}_{+j} - \bar{y}_{++})}_{\text{block effect}} + \underbrace{(y_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{++})}_{\text{error}}$$

F-tests in the RBD

A model for RBD with fixed treatment (therapy) effects is

$$Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}$$

where $i = 1, \dots, a$ $j = 1, \dots, b$ and $E_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$

Mean squares obtained by dividing SS by df :

$$\begin{aligned} MS[A] &= \frac{SS[A]}{a - 1} \\ MS[B] &= \frac{SS[B]}{b - 1} \\ MS[E] &= \frac{SS[E]}{N - a - b + 1} \end{aligned}$$

The primary hypothesis of interest is for a therapy effect:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad \text{vs} \quad H_1 : \text{not all equal.}$$

Using level α , reject H_0 if

$$F = \frac{MS[A]}{MS[E]} > F(\alpha, a - 1, N - a - b + 1)$$

The EMS for error is $E(MS[E]) = \sigma^2$, but only under the *additivity* assumption that there is no block-trt interaction. This assumption is required for inference about treatment effects in the absence of replication, common to block designs.

For the HAT scores, $F_A = MS[A]/MS[E] = 130.5/8.6 = 15.3$ which has $p < 0.01$ on 2, 8 df, providing strong evidence of a therapy effect. Inference, including MCPs, for CONTRASTS involving fixed effects is the same in the complete RBD as it is for other factorial experiments with fixed effects. (E.g. $\widehat{SE}(\bar{Y}_{i+}) = \sqrt{MS[E]/b}$)

Multiple comparisons among means in the RBD

Scheffè simultaneous 95% confidence intervals for contrasts like

$$c_1\mu_1 + c_2\mu_2 + \cdots + c_a\mu_a$$

look like

$$c_1\bar{y}_{1+} + c_2\bar{y}_{2+} + \cdots + c_a\bar{y}_{a+} \pm \sqrt{(a-1)(F^*)MS[E] \sum \frac{c_i^2}{b}}$$

where $F^* = F(0.05, a-1, N-a-b+1)$. For simultaneous pairwise differences, these look like

$$\bar{y}_{i+} - \bar{y}_{j+} \pm \underbrace{\sqrt{(a-1)(F^*)MS[E] \frac{2}{b}}}_{\text{“minimum significant difference”}}$$

For the HAT scores,

$$\bar{y}_{1+} = 13.6, \quad \bar{y}_{2+} = 9, \quad \bar{y}_{3+} = 3.4$$

and

$$\sqrt{(a-1)(F^*)(MS[E])(1/5 + 1/5)} = \sqrt{(3-1)(4.46)(8.6)(2/5)} = 5.5$$

with \bar{y}_{LM+} significantly different from the other two. (LM brings about significantly less improvement than the other two therapies.)

The Tukey minimum significant difference term in the RBD is

$$q(a, n-a-b+1, \alpha) \sqrt{\frac{MS[E]}{b}}$$

For the acrophobia RBD, the term is $4.04 * \sqrt{\frac{8.6}{5}} = 5.3$

means therapy/scheffe tukey; will get the job done in SAS.

Tukey's Studentized Range (HSD) Test for variable: DIFF

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 8 MSE= 8.55
 Critical Value of Studentized Range= 4.041
 Minimum Significant Difference= 5.2843

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	TREAT
A	13.600	5	Contact Desensit
A			
A	9.000	5	Demonstration Pa
B	3.400	5	Live Modelling

Scheffe's test for variable: DIFF

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than REGWF for all pairwise comparisons

Alpha= 0.05 df= 8 MSE= 8.55
 Critical Value of F= 4.45897
 Minimum Significant Difference= 5.5226

Scheffe Grouping	Mean	N	TREAT
A	13.600	5	Contact Desensit
A			
A	9.000	5	Demonstration Pa
B	3.400	5	Live Modelling

Another example, blocks are random

(this material to be covered after random effects have been introduced)

A study investigates the efficiency of four different unit-dose injection systems. For each system, an individual subject (pharmacist or nurse) measures the average time it takes to remove a unit of each system from its outer package, assemble it, and simulate an injection. (Data from Larsen and Marx, 1986.)

Average times (seconds) for implementing systems

Subject	Standard	Vari-Ject	Unimatic	Tubex	\bar{y}_{+j}
1	35.6	17.3	24.4	25.0	25.6
2	31.3	16.4	22.4	26.0	24.0
3	36.2	18.1	22.8	25.3	25.6
4	31.1	17.8	21	24	23.5
5	39.4	18.8	23.3	24.2	26.4
6	34.7	17	21.8	26.2	24.9
7	34.1	14.5	23	24	23.9
8	36.5	17.9	24.1	20.9	24.9
9	40.7	16.4	31.3	36.9	31.3
\bar{y}_{i+}	35.5	17.1	23.8	25.8	25.6

Model

$$Y_{ij} = \mu + \alpha_i + B_j + E_{ij}$$

- $i = 1, \dots, 4 = a$ and $j = 1, \dots, 9 = b$
- α_i denote fixed system effects
- $B_j \stackrel{iid}{\sim} N(0, \sigma_B^2)$ and $E_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$ denote random subject (block) and error effects ($B \perp E$).

```

data one;
  input subject system time;
  cards;
1 1 35.6
2 1 31.3
  ...
8 4 20.9
9 4 36.9
;
run;

proc mixed method=type3;
  class system subject;
  model time=system/ddfm=satterth;
  random subject;
  lsmeans system/adj=tukey cl pdiff;
run;

```

The SAS System
The Mixed Procedure
Model Information

1

Data Set	WORK.ONE
Dependent Variable	time
Covariance Structure	Variance Components
Estimation Method	Type 3
Residual Variance Method	Factor
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Satterthwaite

Class	Levels	Values
system	4	1 2 3 4
subject	9	1 2 3 4 5 6 7 8 9

Total Observations 36

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square
system	3	1559.20222	519.734074	Var(Residual) + Q(system)
subject	8	177.405000	22.175625	Var(Residual) + 4 Var(subject)
Residual	24	148.472778	6.186366	Var(Residual)

Source	Error Term	Error DF	F Value	Pr > F
system	MS(Residual)	24	84.01	<.0001
subject	MS(Residual)	24	3.58	0.0072
Residual

Covariance Parameter Estimates

Cov Parm Estimate

subject	3.9973
Residual	6.1864

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
system	3	24	84.01	<.0001

Least Squares Means

Effect	system	Estimate	Standard Error	DF	t Value	Pr > t	Alpha
system	1	35.5111	1.0637	21.9	33.38	<.0001	0.05
system	2	17.1333	1.0637	21.9	16.11	<.0001	0.05
system	3	23.7889	1.0637	21.9	22.36	<.0001	0.05
system	4	25.8333	1.0637	21.9	24.29	<.0001	0.05

Least Squares Means

Effect	system	Lower	Upper
system	1	33.3044	37.7178
system	2	14.9266	19.3400
system	3	21.5822	25.9956
system	4	23.6266	28.0400

Clearly, the injection system effects are highly significant, as is the random block (or subject) effect, which has an estimated variance component of

$$\hat{\sigma}_B^2 = \frac{1}{a}(MS[B] - MS[E]) = \frac{1}{4}(22.2 - 6.2) = 4 \text{ (squared seconds)}$$

Differences of Least Squares Means

Effect	system	_system	Estimate	Standard		DF	t Value	Pr > t	Adjustment
				Error					
system	1	2	18.3778	1.1725		24	15.67	<.0001	Tukey-Kramer
system	1	3	11.7222	1.1725		24	10.00	<.0001	Tukey-Kramer
system	1	4	9.6778	1.1725		24	8.25	<.0001	Tukey-Kramer
system	2	3	-6.6556	1.1725		24	-5.68	<.0001	Tukey-Kramer
system	2	4	-8.7000	1.1725		24	-7.42	<.0001	Tukey-Kramer
system	3	4	-2.0444	1.1725		24	-1.74	0.0940	Tukey-Kramer

Differences of Least Squares Means

Effect	system	_system	Adj P	Alpha	Lower	Upper	Adj	
							Lower	Upper
system	1	2	<.0001	0.05	15.9579	20.7977	15.1433	21.6122
system	1	3	<.0001	0.05	9.3023	14.1421	8.4878	14.9567
system	1	4	<.0001	0.05	7.2579	12.0977	6.4433	12.9122
system	2	3	<.0001	0.05	-9.0755	-4.2356	-9.8900	-3.4211
system	2	4	<.0001	0.05	-11.1199	-6.2801	-11.9345	-5.4655
system	3	4	0.3242	0.05	-4.4644	0.3755	-5.2789	1.1900

Note the *df* columns:

- For difference of means, pesky mean random effects wash out
- For means, pesky mean random effects don't wash out, necessitating a Satterthwaite *df* approximation

$$\begin{aligned}\bar{Y}_{i_1+} &= \mu + \alpha_{i_1} + \bar{B} + \bar{E}_{i_1+} \\ \bar{Y}_{i_2+} &= \mu + \alpha_{i_2} + \bar{B} + \bar{E}_{i_2+} \\ \bar{Y}_{i_1+} - \bar{Y}_{i_2+} &= \alpha_{i_1} - \alpha_{i_2} + \bar{E}_{i_1+} - \bar{E}_{i_2+} \\ SE(\bar{Y}_{i_1+}) &= \sqrt{\frac{1}{b}(\sigma_B^2 + \sigma^2)} \\ SE(\bar{Y}_{i_1+} - \bar{Y}_{i_2+}) &= \sqrt{\frac{2}{b}\sigma^2}\end{aligned}$$

Latin squares for experiments with two blocking factors

- Experiment with ~ 30 plants, 3 fertilizer trts.
- blocks to control for variability in exptl. units
- location on bench a second blocking factor

Non-randomized design (number indicates trt):

1	2	3
3	1	2
2	3	1

Fertilizer trts randomized to row \times column (or *sunlight* \times *iheight*) combinations by randomly permuting the

1. columns
2. rows

Eg, random number generator permutes (1, 2, 3) to get (2, 3, 1). Placing the columns in this order leads to

2	3	1
1	2	3
3	1	2

Another random permutation of $(1, 2, 3)$ is $(3, 2, 1)$. Placing the rows in this order leads to an unreplicated $3 \times 3 \times 3$ design:

2	3	1	→	3	1	2
1	2	3		1	2	3
3	1	2		2	3	1

Suppose nine rows are available. Three latin squares may be generated, one below the other. The one on top is closest to sunlight, the one on bottom furthest. Columns correspond to initial height blocks.

3	1	2
1	2	3
2	3	1

2	3	1
3	1	2
1	2	3

2	1	3
3	2	1
1	3	2

SAS code to illustrate how such an experiment might go follows. To consider an unreplicated $3 \times 3 \times 3$ Latin square, ignore squares 2 and 3. The fixed effects model generated by the code is:

$$Y_{ijk} = \mu + \tau_k + \rho_i + \kappa_j + E_{ijk}$$

where

- $\mu = 13, (\tau_1, \tau_2, \tau_3) = (-1, 0, 1)$ are trt effects
- $(\rho_1, \rho_2, \rho_3) = (1, 0, -1)$ are sunlight effects.
- $(\kappa_1, \kappa_2, \kappa_3) = (-2, 0, 2)$ are initial ht. effects.
- $E_{ijk} \stackrel{iid}{\sim} N(0, \sigma^2)$ with $\sigma^2 = 1$.

Exercises:

1. specify the theoretical mean in each of the nine cells of the unreplicated design in square 1.
2. specify the marginal means for trt, row and column

Fixed effect model replicated design has extra term:

$$Y_{ijkl} = \mu + \tau_l + \rho_{i(k)} + \kappa_j + \beta_k + E_{ijkl}$$

where

- $(\beta_1, \beta_2, \beta_3) = (3, 0, -3)$ are square effects
- For each k , $(\rho_{1(k)}, \rho_{2(k)}, \rho_{3(k)}) = (3, 0, -3)$ are nested row effects

```

data latinsq;
  array sqware{3} (3,0,-3);
  array slight1{3} (1,0,-1); /* in square 1 */
  array slight2{3} (1,0,-1); /* in square 2 */
  array slight3{3} (1,0,-1); /* in square 3 */
  array iheight{3} (-2,0,2); /* initial height effects */
  array treatment{3} (-1,0,1); /* fertilizer effects */
  input square row col trt;
  growth=round(growth,0.1);
  sigma=1; /* try various values of sigma to generate the data */
  if square=1 then do;
    growth = 10 + sqware{square} + slight1{row} + iheight{col}
      + treatment{trt} + sigma*rannor(1234);
  end;
  else if square=2 then do;
    growth = 10 + sqware{square} + slight2{row} + iheight{col}
      + treatment{trt} + sigma*rannor(1234);
  end;
  else do;
    growth = 10 + sqware{square} + slight3{row} + iheight{col}
      + treatment{trt} + sigma*rannor(1234);
  end;
  height=col;
  cards;
  1 1 1 3
  1 1 2 1
  1 1 3 2
  1 2 1 1
  1 2 2 2
  1 2 3 3
  1 3 1 2
  1 3 2 3
  1 3 3 1
  2 1 1 2
  2 1 2 3
  2 1 3 1
  2 2 1 1
  2 2 2 2
  2 2 3 3
  2 3 1 3
  2 3 2 1
  2 3 3 2
  3 1 1 2
  3 1 2 1
  3 1 3 3
  3 2 1 3
  3 2 2 2
  3 2 3 1
  3 3 1 1
  3 3 2 3
  3 3 3 2 ;

```

```
data one; set latinsq; if square=1; run;
proc glm data=one;
  class square row col trt;
  model growth = row col trt;
  lsmeans row col trt;
run;
```

The SAS System 1
The GLM Procedure

Class	Levels	Values
square	1	1
row	3	1 2 3
col	3	1 2 3
trt	3	1 2 3

Number of observations 9

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	31.55333333	5.25888889	10.45	0.0899
Error	2	1.00666667	0.50333333		
Corrected Total	8	32.56000000			

R-Square Coeff Var Root MSE growth Mean
0.969083 5.415724 0.709460 13.10000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
row	2	9.94666667	4.97333333	9.88	0.0919
col	2	20.16000000	10.08000000	20.03	0.0476
trt	2	1.44666667	0.72333333	1.44	0.4103

growth

row	LSMEAN
1	14.5000000
2	12.8333333
3	11.9666667

growth

col	LSMEAN
1	14.7000000
2	13.5000000
3	11.1000000

growth

trt	LSMEAN
1	12.5333333
2	13.3666667
3	13.4000000

```
proc glm data=latinsq;
  class square row col trt;
  *model growth = row*square col trt;
  model growth = square row(square) col trt;
  lsmeans square row(square) col trt;
run;
```

The SAS System
The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	301.0822222	25.0901852	27.49	<.0001
Error	14	12.7777778	0.9126984		
Corrected Total	26	313.8600000			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
square	2	155.3355556	77.6677778	85.10	<.0001
row(square)	6	33.5777778	5.5962963	6.13	0.0025
col	2	96.6200000	48.3100000	52.93	<.0001
trt	2	15.5488889	7.7744444	8.52	0.0038

square	growth LSMEAN
1	13.1000000
2	9.7555556
3	7.2444444

row	square	growth LSMEAN
1	1	14.5000000
2	1	12.8333333
3	1	11.9666667
1	2	11.3333333
2	2	9.5000000
3	2	8.4333333
1	3	8.6333333
2	3	7.1333333
3	3	5.9666667

col	growth LSMEAN
1	12.3666667
2	10.0000000
3	7.7333333

trt	growth LSMEAN
1	9.0444444
2	10.1666667
3	10.8888889

Exercise

Let Y_{ij} denote the observation in row i column j . Let \bar{Y}_k denote the treatment mean for level k of the treatment factor. For an unrepliated latin square, identify these sums of squares:

$$\sum_{i=1}^a \sum_{j=1}^a (\bar{y}_{i+} - \bar{y}_{++})^2 = SS[\quad]$$

$$\sum_{i=1}^a \sum_{j=1}^a (y_{ij} - \bar{y}_{++})^2 = SS[\quad]$$

$$a \sum_{j=1}^a (\bar{y}_{+j} - \bar{y}_{++})^2 = SS[\quad]$$

$$\sum_{i=1}^3 \sum_{j=1}^3 (y_{ij} - \bar{y}_{i+} - \bar{y}_{+j} - \bar{y}_k + 2\bar{y}_{++})^2 = SS[\quad]$$

$$a \sum_{k=1}^a (\bar{y}_k - \bar{y}_{++})^2 = SS[\quad]$$

Note that \bar{y}_k is determined by the i, j combination. In the $3 \times 3 \times 3$ scheme used for our plant heights, fertilizer $k = 3$ was assigned to the first row and column so that in the last sum of squares above, for $i = 1, j = 1$, \bar{y}_k is the third fertilizer treatment mean, $\bar{y}_3 = 13.40$.

A $4 \times 4 \times 4$ example taken from Ott and Longnecker

- Blocking factors: plot rows, plot columns
- Treatment factor: Fertilizer (4 levels, 2 factors)
(1=broad A, 2=broad B, 3=band A, 4=band B)

1	3	4	2
2	1	3	4
4	2	1	3
3	4	2	1

```

data watermelons;
  input row col trt yield; cards;
  1 1 1 1.75
  1 2 3 1.43
  1 3 4 1.28
  1 4 2 1.66
  2 1 2 1.7
  2 2 1 1.78
  2 3 3 1.40
  2 4 4 1.31
  3 1 4 1.35
  3 2 2 1.73
  3 3 1 1.69
  3 4 3 1.41
  4 1 3 1.45
  4 2 4 1.36
  4 3 2 1.65
  4 4 1 1.73
;
proc glm;
  class row col trt;
  model yield = row col trt;
  estimate "micronutrient effect A-B" trt 1 -1 1 -1/divisor=2;
  contrast "micronutrient effect A-B" trt 1 -1 1 -1;
  estimate "placement effect" trt 1 1 -1 -1/divisor=2;
  contrast "placement effect" trt 1 1 -1 -1;
  estimate "placement-x-nutrient interaction " trt 1 -1 -1 1;
  contrast "placement-x-nutrient interaction " trt 1 -1 -1 1;
  lsmeans row col trt; run;

```

The SAS System
The GLM Procedure

1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	0.49335000	0.05481667	438.53	<.0001
Error	6	0.00075000	0.00012500		
Corrected Total	15	0.49410000			

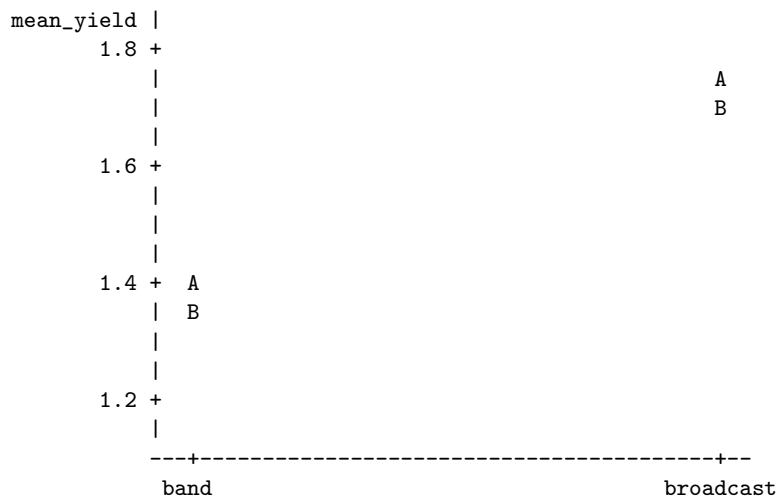
Source	DF	Type I SS	Mean Square	F Value	Pr > F
row	3	0.00085000	0.00028333	2.27	0.1810
col	3	0.01235000	0.00411667	32.93	0.0004
trt	3	0.48015000	0.16005000	1280.40	<.0001

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
micro.effect A-B	1	0.02250000	0.02250000	180.00	<.0001
place.effect	1	0.45562500	0.45562500	3645.00	<.0001
interaction	1	0.00202500	0.00202500	16.20	0.0069

Parameter	Estimate	Standard Error	t Value	Pr > t
micro.effect A-B	0.07500000	0.00559017	13.42	<.0001
place.effect	0.33750000	0.00559017	60.37	<.0001
interaction	-0.04500000	0.01118034	-4.02	0.0069

trt	yield LSMEAN
1	1.73750000
2	1.68500000
3	1.42250000
4	1.32500000

Plot of mean_yield*placement. Symbol is value of micronutrient.



```

proc glm data=watermelons;
  class row col micronutrient placement;
  model yield = row col micronutrient|placement;
  lsmeans micronutrient|placement;
run;

```

The SAS System
The GLM Procedure

Class	Levels	Values
row	4	1 2 3 4
col	4	1 2 3 4
micronutrient	2	A B
placement	2	band broadcast

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	0.49335000	0.05481667	438.53	<.0001
Error	6	0.00075000	0.00012500		
Corrected Total	15	0.49410000			

R-Square	Coeff Var	Root MSE	yield Mean
0.998482	0.724819	0.011180	1.542500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
row	3	0.00085000	0.00028333	2.27	0.1810
col	3	0.01235000	0.00411667	32.93	0.0004
micronutrient	1	0.02250000	0.02250000	180.00	<.0001
placement	1	0.45562500	0.45562500	3645.00	<.0001
micronutri*placement	1	0.00202500	0.00202500	16.20	0.0069

micronutrient	yield LSMEAN
A	1.58000000
B	1.50500000

placement	yield LSMEAN
band	1.37375000
broadcast	1.71125000

micronutrient	placement	yield LSMEAN
A	band	1.42250000
A	broadcast	1.73750000
B	band	1.32500000
B	broadcast	1.68500000

ST 512: Exptl Stats for Biol. Sciences II

Weeks 11-12 Mixed Models for factorial designs

Reading Ch. 14.1,14.2,14.3

- One-way random effects model to study *variances*
- Mixed effects models
- Subsampling
- Expected mean squares for mixed models

An example using one-way random effects model

- Genetics study w/ beef animals. Measure birthweight Y (*lbs*).
- $t = 5$ sires, each mated to a separate group of $n = 8$ dams.
- $N = 40$, completely randomized.

Sire #	Level	Birthweights								\bar{y}_{i+}	s_i
		Sample									
177	1	61	100	56	113	99	103	75	62	83.6	22.6
200	2	75	102	95	103	98	115	98	94	97.5	11.2
201	3	58	60	60	57	57	59	54	100	63.1	15.0
202	4	57	56	67	59	58	121	101	101	77.5	25.9
203	5	59	46	120	115	115	93	105	75	91.0	28.0

Q: Statistical model for these data?

A: One-way fixed effects model?

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

where τ_i denotes the difference between the mean birthweight of population of offspring from sire i and μ , mean of whole population.

The one-way random effects model

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{T_i}_{\text{random}} + \underbrace{E_{ij}}_{\text{random}} \quad \text{for } i = 1, 2, \dots, t \text{ and } j = 1, \dots, n$$

with

- $T_1, T_2, \dots, T_t \stackrel{iid}{\sim} N(0, \sigma_T^2)$
- $E_{11}, \dots, E_{tn} \stackrel{iid}{\sim} N(0, \sigma^2)$
- T_1, T_2, \dots, T_t independent of E_{11}, \dots, E_{tn}

Features

- T_1, T_2, \dots denote *random* effects, drawn from some population of interest. That is, T_1, T_2, \dots is a random sample!
- σ_T^2 and σ^2 are called variance components
- conceptually different from one-way fixed effects model

For beef animal genetic study, with $t = 5$ and $n = 8$, the random effects T_1, T_2, \dots, T_5 reflect sire-to-sire variability.

No particular interest in $\tau_1, \tau_2, \dots, \tau_5$ from the fixed effects model:

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{\tau_i}_{\text{fixed}} + \underbrace{E_{ij}}_{\text{random}} \quad \text{for } i = 1, 2, \dots, t \text{ and } j = 1, \dots, n$$

with

- $\tau_1, \tau_2, \dots, \tau_t$ unknown model parameters
- $E_{11}, \dots, E_{tn} \stackrel{iid}{\sim} N(0, \sigma^2)$

One-way random effects model continued

Exercise: Using the random effects model, specify

$$E(Y_{ij}) \text{ and } \text{Var}(Y_{ij})$$

- Two *components* to variability in data: σ^2, σ_T^2
- T_1, T_2, T_3, T_4, T_5 a random sample of sire effects
- Sire effects is a population in its own right.

Contrast this situation with the binding fractions. Why not model antibiotic effects random? Why fixed? (See Ch. 17 for more discussion.)

Model parameters: $\sigma^2, \sigma_T^2, \mu$

Sums of squares and mean squares - same as in one-way fixed effects ANOVA:

$$\begin{aligned} SS[T] &= \sum \sum (\bar{y}_{i+} - \bar{y}_{++})^2 \\ SS[E] &= \sum \sum (y_{ij} - \bar{y}_{i+})^2 \\ SS[Total] &= \sum \sum (y_{ij} - \bar{y}_{++})^2 \end{aligned}$$

The ANOVA table is almost the same, it just has a different expected mean squares column:

Source	SS	df	MS	Expected MS
Treatment	$SS[T]$	$t - 1$	$MS[T]$	$\sigma^2 + n\sigma_T^2$
Error	$SS[E]$	$N - t$	$MS[E]$	σ^2
Total	$SS[Total]$	$N - 1$		

Estimating parameters of one-way random effects model

$$\begin{aligned}\hat{\mu} &= \bar{y}_{++} \\ \hat{\sigma}^2 &= MS[E] \\ \hat{\sigma}_T^2 &= \frac{MS[T] - MS[E]}{n}\end{aligned}$$

For sires data, $\bar{y}_{++} = 82.6$ and

Source	SS	df	MS	Expected MS
Sire	5591	4	1398	$\sigma^2 + 8\sigma_T^2$
Error	16233	35	464	σ^2
Total	21824	39		

$$\begin{aligned}\hat{\mu} &= 82.6 \\ \hat{\sigma}^2 &= 464 \text{ (lbs}^2\text{)} \\ \hat{\sigma}_T^2 &= \frac{1398 - 464}{8} \\ &= 117 \text{ (lbs}^2\text{)}\end{aligned}$$

Specific questions pertaining to this study:

Consider the birthweight of a randomly sampled calf.

1. What is the estimated variance of such a calf?
2. Estimate how much of this variation is due to the sire effect.
3. Estimate how much of this variation is not due to the sire effect.

General questions:

1. Is it possible for an estimated variance component to be negative?
2. How?
3. What do you do in that case?

Other parameters of interest in random effects models

Coefficient of variation (CV):

$$CV(Y_{ij}) = \frac{\sqrt{\text{Var}(Y_{ij})}}{|E(Y_{ij})|} = \frac{\sqrt{\sigma_T^2 + \sigma^2}}{|\mu|}$$

Note: this is *not* estimated by **Coef** **Var** in PROC GLM output.

Intraclass correlation coefficient

$$\rho_I = \frac{\text{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\text{Var}(Y_{ij})\text{Var}(Y_{ik})}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2}$$

- Interpretation: the correlation between two responses receiving the same level of the random factor.
- Bigger values of ρ_I correspond to (bigger/smaller?) random treatment effects.

For sires,

$$\begin{aligned} \widehat{CV} &= \frac{\sqrt{117+464}}{82.6} = 0.29 \\ \hat{\rho}_I &= \frac{117}{117+464} = 0.20 \end{aligned}$$

Interpretations:

- The estimated standard deviation of a birthweight, 24.1 is 29% of the estimated mean birthweight, 82.6.
- The estimated correlation between any two calves with the same sire for a male parent, or the estimated *intrasire* correlation coefficient, is 0.20

Using PROC GLM for random effects models

```

data one;
  input sire @;
  do i=1 to 8;
    input bw @; output;
  end;
  cards;
177 61 100 56 113 99 103 75 62
200 75 102 95 103 98 115 98 94
201 58 60 60 57 57 59 54 100
202 57 56 67 59 58 121 101 101
203 59 46 120 115 115 93 105 75
;
run;

proc glm;
  class sire;
  model bw=sire;
  random sire;
run;

```

The GLM Procedure

	Class	Levels	Values			
	sire	5	177 200 201 202 203			
Source		DF	Sum of Squares	Mean Square	F Value	Pr > F
Model		4	5591.15000	1397.78750	3.01	0.0309
Error		35	16232.75000	463.79286		
Corrected Total		39	21823.90000			
	R-Square	Coeff Var	Root MSE	bw Mean		
	0.256194	26.08825	21.53585	82.55000		
Source	Type III Expected Mean Square					
sire	Var(Error) + 8 Var(sire)					

($\sigma^2 = \text{Var}(\text{Error})$ and $\sigma_T^2 = \text{Var}(\text{sire})$.)

Testing a variance component - $H_0 : \sigma_T^2 = 0$

Recall that $\sigma_T^2 = \text{Var}(T_i)$, the variance among the population of treatment effects.

$$F = \frac{MS[T]}{MS[E]}$$

reject H_0 at level α if $F > F(\alpha, t - 1, N - t)$

For the sires,

$$F = \frac{1398}{464} = 3.01 > 2.64 = F(0.05, 4, 35)$$

so H_0 is rejected at $\alpha = 0.05$. (The p -value is 0.0309)

Q: “Isn’t this just just like the F -test for one-way ANOVA with *fixed* effects?”

A: “Yes.”

Interval Estimation of some model parameters

A 95% confidence interval for μ derived by consideration of $SE(\bar{Y}_{++})$:

$$\begin{aligned}\bar{Y}_{++} &= \frac{1}{N} \sum_{i=1}^t \sum_{j=1}^n Y_{ij} \\ &= \frac{1}{N} \sum_{i=1}^t \sum_{j=1}^n (\mu + T_i + E_{ij}) \\ &= \mu + \bar{T}_+ + \bar{E}_{++}\end{aligned}$$

where $\bar{T}_+ = (T_1 + \cdots + T_t)/t$ and $\bar{E}_{++} = (\sum \sum E_{ij})/N$, so that

$$\begin{aligned}\text{Var}(\bar{Y}_{++}) &= \text{Var}(\bar{T}_+ + \bar{E}_{++}) \\ &= \frac{\sigma_T^2}{t} + \frac{\sigma_2}{nt} \\ &= \frac{1}{nt} (n\sigma_T^2 + \sigma^2) \\ &= \frac{1}{nt} E(MS[T]).\end{aligned}$$

If the data are normally distributed, then

$$\frac{\bar{Y}_{++} - \mu}{\sqrt{\frac{MS[T]}{nt}}} \sim t_{t-1}$$

and a 95% confidence interval for μ given by

$$\boxed{\bar{Y}_{++} \pm t(0.025, t-1) \sqrt{\frac{MS[T]}{nt}}}$$

Sires data: $\bar{y}_{++} = 82.6$, $MS[T] = 1398$, $nt = 40$. Critical value $t(0.025, 4) = 2.78$ yields the interval

$$82.6 \pm 2.78(5.91) \quad \text{or} \quad (66.1, 99.0).$$

Confidence interval for ρ_I

A 95% confidence interval for ρ_I can be obtained from the expression

$$\frac{F_{obs} - F_{\alpha/2}}{F_{obs} + (n-1)F_{\alpha/2}} < \rho_I < \frac{F_{obs} - F_{1-\alpha/2}}{F_{obs} + (n-1)F_{1-\alpha/2}}$$

where $F_{\alpha/2} = F(\frac{\alpha}{2}, t-1, N-t)$ and F_{obs} is the observed F -ratio for treatment effect from the ANOVA table.

For the sires, $F_{obs} = 3.01$ and $F_{0.025} = 3.179$, $F_{0.975} = 0.119$. The formula gives $(-0.01, 0.75)$.

Note the asymmetry and disagreement with test of $H_0 : \sigma_T^2 = 0$

These formulas arrived at via some distributional results:

- $(t-1) \frac{MS[T]}{\sigma^2 + n\sigma_T^2} \sim \chi_{t-1}^2$
- $(N-t) \frac{MS[E]}{\sigma^2} \sim \chi_{N-t}^2$
- $MS[T]$ and $MS[E]$ are independent
- Ratio of independent χ^2 RVs divided by df has an F distribution
-

$$\left(\frac{MS[T]}{\sigma^2 + n\sigma_T^2} \right) / \left(\frac{MS[E]}{\sigma^2} \right) \sim F_{t-1, N-t}$$

(which explains the F test for $H_0 : \sigma_t^2 = 0$)

- Rearranging the probability statement below

$$1-\alpha = \Pr \left(F\left(1 - \frac{\alpha}{2}, t-1, N-t\right) < \frac{\frac{MS[T]}{\sigma^2 + n\sigma_T^2}}{\frac{MS[E]}{\sigma^2}} < F\left(\frac{\alpha}{2}, t-1, N-t\right) \right)$$

so that ρ_I gets left in the middle yields the confidence interval yields the c.i. at the top o' the page.

Using PROC MIXED for random effects models

```
proc mixed cl;
  class sire;
  model bw=;
  random sire;
  estimate "mean" intercept 1/cl;
run;
```

The SAS System
The Mixed Procedure
Model Information

1

Dependent Variable	bw
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class	Levels	Values
sire	5	177 200 201 202 203

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
sire	116.75	0.05	29.9707	7051.37
Residual	463.79	0.05	305.11	789.17

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > t	Alpha
mean	82.5500	5.9114	4	13.96	0.0002	0.05

Estimates

Label	Lower	Upper
mean	66.1373	98.9627

More interval estimation for variance components

The estimated residual variance component for the sire data was $\hat{\sigma}^2 = MS[E] = 464 \text{ lbs}^2$.

A 95% confidence interval for this variance component is given by

$$\left(\frac{(40 - 5)464}{53.2} < \sigma^2 < \frac{(40 - 5)464}{20.6} \right)$$

or

$$\left(\frac{35}{53.2}464 < \sigma^2 < \frac{35}{20.6}464 \right)$$

or $(305.2, 789.5)\text{lbs}^2$

This can be derived using the distributional result

$$(N - t) \frac{MS[E]}{\sigma^2} \sim \chi_{N-t}^2$$

setting up the probability statement

$$1 - \alpha = \Pr \left(\chi^2 \left(1 - \frac{\alpha}{2}, N - t \right) < (N - t) \frac{MS[E]}{\sigma^2} < \chi^2 \left(\frac{\alpha}{2}, N - t \right) \right)$$

Rearranging to get σ^2 in the middle yields the $100(1 - \alpha)\%$ confidence interval for σ^2 :

$$\left(\frac{(N - t)MS[E]}{\chi_{\alpha/2}^2}, \frac{(N - t)MS[E]}{\chi_{1-\alpha/2}^2} \right).$$

Q: What are the mean and variance of the χ_{35}^2 distribution?

Interval estimation for σ_T^2

The estimated variance component for the random sire effect was $\hat{\sigma}_T^2 = 117$.

Q: How can we get a 95% confidence interval for σ_T^2 ?

A: In a similar fashion, but the confidence level based on Satterthwaite's approximation to the degrees of freedom of the linear combination of *MS* terms:

$$\left(\frac{\hat{df} \hat{\sigma}_T^2}{\chi_{\alpha/2, \hat{df}}^2}, \frac{\hat{df} \hat{\sigma}_T^2}{\chi_{1-\alpha/2, \hat{df}}^2} \right)$$

where

$$\hat{df} = \frac{(n\hat{\sigma}_T^2)^2}{\frac{MS[T]^2}{t-1} + \frac{MS[E]^2}{N-t}}$$

For the sire data,

$$\hat{df} = \frac{(8 \times 117)^2}{\frac{1398^2}{4} + \frac{464^2}{35}} = 1.76$$

Using the **CL** option in the **MIXED** statement will request this confidence interval and will use this approximation to *df* and will not round to the nearest integer *df*:

$$\chi_{0.975, 1.76}^2 = 0.029, \quad \chi_{0.025, 1.76}^2 = 6.87$$

yielding the 95% confidence interval

$$\left(\frac{1.76(117)}{6.87}, \frac{1.76(117)}{0.29} \right)$$

or

$$(30, 7051)$$

Review of one-way random effects ANOVA

The model

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{T_i}_{\text{random}} + \underbrace{E_{ij}}_{\text{random}} \quad \text{for } i = 1, 2, \dots, t \text{ and } j = 1, \dots, n$$

with

$$T_1, T_2, \dots, T_t \stackrel{iid}{\sim} N(0, \sigma_T^2) \quad \text{independent of} \quad E_{11}, \dots, E_{tn} \stackrel{iid}{\sim} N(0, \sigma^2)$$

Remarks:

- (T_1, T_2, \dots) randomly drawn from pop'n of treatment effects.)
- Only three parameters: μ, σ, σ_T^2
- Several functions of these parameters of interest

$$- CV(Y) = \frac{\sqrt{\sigma^2 + \sigma_T^2}}{\mu}$$

$$- \rho_I = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_T^2}{\sigma^2 + \sigma_T^2}$$

- Two observations from same treatment group not independent

Exercise: match up the formulas for confidence intervals below with their targets, $\rho_I, \sigma^2, \sigma_T^2, \mu$:

$$\begin{aligned} & \bar{Y}_{++} \pm t(0.025, t-1) \sqrt{\frac{MS[T]}{nt}} \\ & \left(\frac{F_{obs} - F_{1-\alpha/2}}{F_{obs} + (n-1)F_{1-\alpha/2}}, \frac{F_{obs} - F_{\alpha/2}}{F_{obs} + (n-1)F_{\alpha/2}} \right) \\ & \left(\frac{(N-t)MS[E]}{\chi_{\alpha/2}^2}, \frac{(N-t)MS[E]}{\chi_{1-\alpha/2}^2} \right) \\ & \left(\frac{\hat{df} \hat{\sigma}_T^2}{\chi_{\alpha/2, \hat{df}}^2}, \frac{\hat{df} \hat{\sigma}_T^2}{\chi_{1-\alpha/2, \hat{df}}^2} \right) \end{aligned}$$

Modelling factorial effects: fixed, or random?
A guide

	Random	Fixed
Levels		
- selected from conceptually ∞ popn of collection of levels	X	
- finite number of possible levels		X
Another expt		
- would use same levels		X
- would involve new levels sampled from same popn	X	
Goal		
- estimate varcomps	X	
- estimate longrun means		X
Inference		
- for these levels used in this expt		X
- for the popn of levels	X	

ST 512: Exptl Stats for Biol. Sciences II

Weeks 11-12 Mixed Models for factorial designs (read Ch. 17)

Two-factor designs

with factors that are fixed/random and nested/crossed

1. Entomologist records energy expended (y) by $N = 27$ honeybees
 - at three TEMPERATURES (20, 30, 40°C)
 - consuming three levels of SUCROSE (20%, 40%, 60%)

Temp	Suc	Sample		
20	20	3.1	3.7	4.7
20	40	5.5	6.7	7.3
20	60	7.9	9.2	9.3
30	20	6	6.9	7.5
30	40	11.5	12.9	13.4
30	60	17.5	15.8	14.7
40	20	7.7	8.3	9.5
40	40	15.7	14.3	15.9
40	60	19.1	18.0	19.9

2. Experiment to study effect of drug and method of administration on fasting blood sugar in a random sample of $N = 18$ diabetic patients. (see Rao exercise 13.35, dataset=`blsugar.dat`)
 - First factor is drug: brand I tablet, brand II tablet, insulin injection
 - Second factor is type of administration (see table)

Drug (i)	Type of Administration (j)	Mean $\bar{y}_{j(i)}$	Variance $s_{j(i)}^2$
$(i = 1)$ Brand I tablet	$(j = 1)$ 30mg \times 1	15.7	6.3
	$(j = 2)$ 15mg \times 2	19.7	9.3
$(i = 2)$ Brand II tablet	$(j = 1)$ 20mg \times 1	20	1
	$(j = 2)$ 10mg \times 2	17.3	6.3
$(i = 3)$ Insulin injection	$(j = 1)$ before breakfast	28	4
	$(j = 2)$ before supper	33	9

3. An experiment is conducted to determine variability among laboratories (interlaboratory differences) in their assessment of bacterial concentration in milk after pasteurization. Milk w/ various degrees of contamination was tested by randomly drawing four samples of milk from a collection of cartons at various stages of spoilage. Y is colony-forming units/ μl . Labs think they're receiving 8 independent samples

Lab	Sample			
	1	2	3	4
1	2200	3000	210	270
	2200	2900	200	260
2	2600	3600	290	360
	2500	3500	240	380
3	1900	2500	160	230
	2100	2200	200	230
4	2600	2800	330	350
	4300	1800	340	290
5	4000	4800	370	500
	3900	4800	340	480

(Data from Oehlert, 2000)

4. An expt measures *Campylobacter* counts in $N = 120$ chickens in a processing plant, at four locations, over three days. Means (std) for $n = 10$ chickens sampled at each location tabulated below:

Day	Location			
	Before Washer	After Washer	After mic. rinse	After chill tank
1	70070.00	48310.00	12020.00	11790.00
	(79034.49)	(34166.80)	(3807.24)	(7832.05)
2	75890.00	52020.00	8090.00	8690.00
	(74551.32)	(17686.27)	(4848.01)	(5526.19)
3	95260.00	33170.00	6200.00	8370.00
	(03176.00)	(22259.08)	(5028.81)	(5720.15)

Data courtesy of Michael Bashor, General Mills

Transformation?

5. An experiment to assess the variability of a particular acid among plants and among leaves of plants:

Plant i	1			2			3			4		
Leaf j	1	2	3	1	2	3	1	2	3	1	2	3
$k = 1$	11.2	16.5	18.3	14.1	19.0	11.9	15.3	19.5	16.5	7.3	8.9	11.3
$k = 2$	11.6	16.8	18.7	13.8	18.5	12.4	15.9	20.1	17.2	7.8	9.4	10.9
$k = 3$	12.0	16.1	19.0	14.2	18.2	12.0	16.0	19.3	16.9	7.0	9.3	10.5

Data from Neter, et al (1996)

6. Study of effect of salinity on barley growth in a controlled medium.

Salinity	Container	Weights (g)		
c	1	11.29	11.08	11.1
c	2	7.37	6.55	8.5
6b	1	5.64	5.98	5.69
6b	2	4.2	3.34	4.21
12b	1	4.83	4.77	5.66
12b	2	3.28	2.61	2.69

- Two containers for each level of salinity treatment factor (for a total of 6 containers).

Six types of two-factor models

Fixed and/or random effects that are either crossed or nested

1.	$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + E_{ijk}$	crossed/random
2.	$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + E_{ijk}$	nested/fixed
3.	$Y_{ijk} = \mu + A_i + B_{j(i)} + E_{ijk}$	nested/random
4.	$Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}$	crossed/mixed
5.	$Y_{ijk} = \mu + \alpha_i + B_{j(i)} + E_{ijk}$	nested/mixed
6.	$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$	crossed/fixed

In the models above, (which are not ordered according to the six prior datasets)

- GREEK symbols parameterize FIXED, unknown treatment means
- CAPITAL letters represent RANDOM effects
- for Model 1, $\sum \alpha_i = \sum \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} \equiv 0$
- for Model 2, $\sum \alpha_i = \sum_j \beta_{j(i)} \equiv 0$
- for Model 3, $A_i, B_i, (AB)_{ij}$ are all independent
- for Model 4, $\sum \alpha_i = 0$ and $B_j, (\alpha B)_{ij}$ are all independent
- for Model 5, $A_i, B_{j(i)}$ are all independent
- for Model 6, $\sum \alpha_i = 0$

Recall

- RANDOM effects are used when it makes sense to think of LEVELS of factor as random sample from a population.

Identifying the appropriate model for our 6 examples:

1. Energy expended by honeybees.

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model:

$$Y_{ijk} = \mu + \quad + E_{ijk}$$

2. Change in fasting blood sugar for diabetics

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model:

$$Y_{ijk} = \mu + \quad + E_{ijk}$$

3. Measuring bacterial concentration in milk

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model:

$$Y_{ijk} = \mu + \quad + E_{ijk}$$

4. Measuring bacteria counts in chickens at processing plant

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model:

$$Y_{ijk} = \mu + \quad + E_{ijk}$$

5. Acids in leaves of plants

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model:

$$Y_{ijk} = \mu + \quad + E_{ijk}$$

6. Barley growth

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model:

$$Y_{ijk} = \mu + \quad + E_{ijk}$$

Tables of expected means squares (EMS): (see Rao table 14.1)

When factors A and B are CROSSED, and no sum-to-zero assumptions are made on random effects, expected means associated with sums of squares are given in the table below:

Source	df	A, B fixed	A, B random	A fixed B random
A	$a - 1$	$\sigma^2 + nb\psi_A^2$	$\sigma^2 + nb\sigma_A^2 + n\sigma_{AB}^2$	$\sigma^2 + nb\psi_A^2 + n\sigma_{\alpha B}^2$
B	$b - 1$	$\sigma^2 + na\psi_B^2$	$\sigma^2 + na\sigma_B^2 + n\sigma_{AB}^2$	$\sigma^2 + na\sigma_B^2 + n\sigma_{\alpha B}^2$
AB	$(a - 1)$ $\times (b - 1)$	$\sigma^2 + n\psi_{AB}^2$	$\sigma^2 + n\sigma_{AB}^2$	$\sigma^2 + n\sigma_{\alpha B}^2$
Error	$ab(n - 1)$	σ^2	σ^2	σ^2

When factor B is NESTED in factor A , expected means associated with sums of squares are given in the table below:

Source	df	A, B fixed	A, B random	A fixed B random
A	$a - 1$	$\sigma^2 + nb\psi_A^2$	$\sigma^2 + nb\sigma_A^2 + n\sigma_{B(A)}^2$	$\sigma^2 + nb\psi_A^2 + n\sigma_{B(A)}^2$
$B(A)$	$a(b - 1)$	$\sigma^2 + n\psi_{B(A)}^2$	$\sigma^2 + n\sigma_{B(A)}^2$	$\sigma^2 + n\sigma_{B(A)}^2$
Error	$ab(n - 1)$	σ^2	σ^2	σ^2

where ψ^2 and σ^2 values are defined on the next page.

Help with computing expected mean squares
(without sum-to-zero assumptions on random effects)

1. If a factor X with index i is random then $EMS(X)$ is a linear combo of σ^2 and varcomps for all random effects containing index i . Coefficients for varcomps are limits of indexes NOT listed (summed over) in random effects.
2. If a factor X is fixed. Treat it like it is random and then just replace the varcomp for X with the effect size, ψ_X^2 .

$$\psi_A^2 = \frac{1}{a-1} \sum_1^a \alpha_i^2 \quad \text{effect size of factor } A$$

$$\psi_B^2 = \frac{1}{b-1} \sum_1^b \beta_i^2 \quad \text{effect size of factor } B$$

$$\psi_{AB}^2 = \frac{1}{(a-1)(b-1)} \sum_{i=1}^a \sum_{j=1}^b (\alpha\beta)_{ij}^2 \quad \text{effect size of interaction}$$

$$\psi_{B(A)}^2 = \frac{1}{a(b-1)} \sum_{i=1}^a \sum_{j=1}^b \beta_{j(i)}^2 \quad \text{effect size of factor } B$$

$$\sigma_A^2 = \text{Var}(A_i) \quad \text{variance component for factor } A$$

$$\sigma_B^2 = \text{Var}(B_i) \quad \text{variance component for factor } B$$

$$\sigma_{AB}^2 = \text{Var}((AB)_{ij}) \quad \text{variance component for interaction}$$

$$\sigma_{B(A)}^2 = \text{Var}(B_{j(i)}) \quad \text{variance component for factor } B$$

$$\sigma^2 = \text{Var}(E_{ijk}) \quad \text{error variance}$$

The term *effect size* is often used in power considerations and sometimes involves division by σ^2 .

Using expected mean squares to analyze data in mixed models

- *EMS* tables dictate which F -ratios test which effects
- *EMS* tables yield estimating equations for variance components

Milk example (p.2):

 F -tests and estimating variance components.

1. To test for interaction effect, use $F_{AB} = \frac{MS[AB]}{MS[E]}$
2. To test for main effect of A, use $F_A = \frac{MS[A]}{MS[AB]}$
3. To test for main effect of B, use $F_B = \frac{MS[B]}{MS[AB]}$

Note the departure from fixed effects analysis, where $MS[E]$ is always used in the denominator.

The SAS System					
The GLM Procedure					
Dependent Variable: ly = log(y)					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	56.03510844	2.94921623	191.44	<.0001
sample	3	53.18978788	17.72992929	1150.89	<.0001
lab	4	2.30248803	0.57562201	37.37	<.0001
sample*lab	12	0.54283253	0.04523604	2.94	0.0161
Error	20	0.30810726	0.01540536		
Corrected Total	39	56.34321569			

The wrong F -ratio and p -value for testing for random LAB (A) effect:

$$F = \frac{MS[A]}{MS[E]} = \frac{0.5756}{0.0154} = 37.37(p < 0.0001)$$

The correct F -ratio and p -value for testing for random LAB (A) effect:

$$F = \frac{MS[A]}{MS[AB]} = \frac{0.5756}{0.0452} = 12.72(p = 0.0003)$$

Estimating variance components

The estimated variance components satisfy the following system of equations:

$$\begin{aligned}
 MS[E] &= \hat{\sigma}^2 \\
 MS[AB] &= \hat{\sigma}^2 + n\hat{\sigma}_{AB}^2 \\
 &= \hat{\sigma}^2 + 2\hat{\sigma}_{AB}^2 \\
 MS[A] &= \hat{\sigma}^2 + nb\hat{\sigma}_A^2 + n\hat{\sigma}_{AB}^2 \\
 &= \hat{\sigma}^2 + 8\hat{\sigma}_A^2 + 2\hat{\sigma}_{AB}^2 \\
 MS[B] &= \hat{\sigma}^2 + na\hat{\sigma}_B^2 + n\hat{\sigma}_{AB}^2 \\
 &= \hat{\sigma}^2 + 10\hat{\sigma}_B^2 + 2\hat{\sigma}_{AB}^2
 \end{aligned}$$

Substitution of

$$\begin{aligned}
 MS[E] &= 0.0154 \\
 MS[AB] &= 0.0452 \\
 MS[A] &= 0.5756 \\
 MS[B] &= 17.7299
 \end{aligned}$$

into the system of equations yields estimated variance components:

$$\begin{aligned}
 \hat{\sigma}^2 &= MS[E] = 0.0154 \\
 \hat{\sigma}_{AB}^2 &= \frac{MS[AB] - MS[E]}{2} = \frac{0.0452 - 0.0154}{2} = 0.01492 \\
 \hat{\sigma}_A^2 &= \frac{MS[A] - MS[AB]}{nb} = \frac{0.5756 - 0.0452}{8} = 0.0663 \\
 \hat{\sigma}_B^2 &= \frac{MS[B] - MS[AB]}{na} = \frac{17.7299 - 0.0452}{10} = 1.768
 \end{aligned}$$

```

data one;
  infile "milk.dat" firstobs=4;
  input sample lab y;
  ly=log(y);
run;

proc glm;
  class lab sample;
  model ly=sample|lab;
  random sample lab sample*lab;
  test h=lab sample e=sample*lab;
  lsmeans sample*lab;
run;

```

The GLM Procedure

Dependent Variable: ly

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	56.03510844	2.94921623	191.44	<.0001
Error	20	0.30810726	0.01540536		
Corrected Total	39	56.34321569			

R-Square	0.994532	Coeff Var	1.821098	Root MSE	0.124118	ly Mean	6.815577
----------	----------	-----------	----------	----------	----------	---------	----------

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sample	3	53.18978788	17.72992929	1150.89	<.0001
lab	4	2.30248803	0.57562201	37.37	<.0001
lab*sample	12	0.54283253	0.04523604	2.94	0.0161

Source	Type III Expected Mean Square
sample	Var(Error) + 2 Var(lab*sample) + 10 Var(sample)
lab	Var(Error) + 2 Var(lab*sample) + 8 Var(lab)
lab*sample	Var(Error) + 2 Var(lab*sample)

Tests of Hypotheses Using the Type III MS for lab*sample as an Error Term

Source	DF	Type III SS	Mean Square	F Value	Pr > F
lab	4	2.30248803	0.57562201	12.72	0.0003
sample	3	53.18978788	17.72992929	391.94	<.0001

```
proc varcomp;
  class sample lab;
  model y=sample|lab;
run;
```

Variance Components Estimation Procedure

Variance Component	ly
Var(sample)	1.76847
Var(lab)	0.06630
Var(sample*lab)	0.01492
Var(Error)	0.01541

Q: At the end of the day, what is the conclusion from the analysis of this crossed, random effects experiment?

- There is evidence of variability due to laboratory \times sample interaction; interlaboratory effects vary by sample.
- The estimated parameters (μ + variance components) of the model

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + E_{ijk}$$

are

$$\begin{aligned}\hat{\sigma}^2 &= 0.0154 \\ \hat{\sigma}_{AB}^2 &= 0.0149 \\ \hat{\sigma}_A^2 &= 0.0663 \\ \hat{\sigma}_B^2 &= 1.7685 \\ \hat{\mu} &= 6.82(\text{log scale})\end{aligned}$$

- The standard error of \bar{Y}_{+++} can be derived by

$$\begin{aligned}\bar{Y}_{+++} &= \mu + \bar{A}_+ + \bar{B}_+ + \overline{(AB)}_{++} + \bar{E}_{+++} \\ \text{Var}(\bar{Y}_{+++}) &= \text{Var}(\bar{A}_+) + \text{Var}(\bar{B}_+) + \text{Var}(\overline{(AB)}_{++}) + \text{Var}(\bar{E}_{+++}) \\ &= \frac{\sigma_A^2}{a} + \frac{\sigma_B^2}{b} + \frac{\sigma_{AB}^2}{ab} + \frac{\sigma^2}{abn}\end{aligned}$$

Estimation of standard error and approximation of df

The standard error

$$SE(\bar{Y}_{+++}) = \sqrt{\frac{\sigma_A^2}{a} + \frac{\sigma_B^2}{b} + \frac{\sigma_{AB}^2}{ab} + \frac{\sigma^2}{abn}}$$

can be estimated by substitution of estimated variance components ($\hat{\sigma}^2$), which leads to

$$\begin{aligned} \widehat{SE}(\bar{Y}_{+++}) &= \sqrt{\frac{\hat{\sigma}_A^2}{a} + \frac{\hat{\sigma}_B^2}{b} + \frac{\hat{\sigma}_{AB}^2}{ab} + \frac{\hat{\sigma}^2}{abn}} \\ &= \text{lots of algebra and cancellations} \\ &= \sqrt{\frac{1}{nab} (MS[A] + MS[B] - MS[AB])} \end{aligned}$$

For the milk data, we have

$$\widehat{SE}(\bar{Y}_{+++}) = \sqrt{\frac{1}{40} (0.58 + 17.73 - 0.05)} = 0.6757$$

For a 95% confidence interval, we have a problem: we don't know how many df are associated with a t statistic based on this estimated SE .

ST511 Flashback:

Unequal variances independent samples t -test:

Example: Suspended particulate matter Y (in micrograms per cubic meter) in homes with smokers (Y_1) and without smokers (Y_2):

smokers	133	128	136	135	131	131	130	131	131	132	147
no smokers	106	85	84	95	104	79	72	115	95		

Summary statistics:

$$\bar{y}_1 = 133.2, s_1^2 = 26.0$$

$$\bar{y}_2 = 92.8, s_2^2 = 195.4$$

$$n_1 = 11, n_2 = 9$$

Assumptions:

- Y_{11}, \dots, Y_{1n_1} and Y_{21}, \dots, Y_{2n_2} are independent random samples from normal distributions with unknown $\mu_1, \mu_2, \sigma_1, \sigma_2$ and $\sigma_1^2 \neq \sigma_2^2$. Note the large difference in the sample variances.

$$H_0 : \mu_1 - \mu_2 = 0 \text{ v. } H_1 : \mu_1 - \mu_2 \neq 0$$

Consider the test statistic

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

For small n_1, n_2 , this quantity does not have the standard normal distribution. nor does the version where S_p^2 is used in the denominator. An approximate solution is to use the student t distribution with df approximated by the Satterthwaite approximation:

$$\hat{df} = \frac{(c_1 MS_1 + c_2 MS_2)^2}{(c_1 MS_1)^2/df_1 + (c_2 MS_2)^2/df_2}$$

where $MS_i = S_i^2$ and $c_i = 1/n_i$.

ST511 Flashback continued:

For the air pollution in homes with a smoking occupant data, $c_1MS_1 = 26/11 = 2.36$, $c_2MS_2 = 195.4/9 = 21.71$ and

$$\widehat{df} = \frac{(2.36 + 21.71)^2}{\frac{2.36^2}{10} + \frac{21.71^2}{8}} = 9.74$$

The 97.5th percentile of the t distribution with $df = 9.74$ is $t(0.025, 9.74) = 2.236$

A 95% confidence interval for the mean difference between homes with and without a smoking occupant, $\mu_1 - \mu_2$ is given by

$$133.2 - 92.8 \pm 2.236\sqrt{26/11 + 195.4/9}$$

or

$$40.4 \pm 2.236(4.91)$$

or

$$40.4 \pm 10.97$$

or

$$(29.4, 51.4)$$

These data would lead to the rejection of $H_0 : \mu_1 = \mu_2 = 0$ versus the two-tailed alternative. The observed test statistic is given by

$$t_{obs} = \frac{133.2 - 92.8}{\sqrt{26/11 + 195.4/9}} = \frac{40.4}{4.91} = 8.2 \quad (p < 0.0001)$$

This problem aka the Behrens-Fisher problem.

```

data one;
  infile "smokers.dat" firstobs=2;
  input y smoke;
  label y="suspended particulate matter";
run;

```

```

proc ttest;
  class smoke;
  var y;
run;

```

The SAS System 1
The TTEST Procedure
Statistics

Variable	smoke	N	Lower CL Mean	Mean	Upper CL Mean
y	0	9	82.032	92.778	103.52
y	1	11	129.76	133.18	136.6
y	Diff (1-2)		-49.91	-40.4	-30.9

Variable	smoke	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
y	0	9.443	13.98	26.783	4.66
y	1	3.5603	5.0955	8.9422	1.5363
y	Diff (1-2)	7.6046	10.064	14.883	4.5235

T-Tests

Variable	Method	Variances	DF	t Value	Pr > t
y	Pooled	Equal	18	-8.93	<.0001
y	Satterthwaite	Unequal	9.74*	-8.23*	<.0001

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
y	Folded F	8	10	7.53	0.0045

The two-way random effects model for milk data
Satterthwaite's approximation (cont'd)

To approximate the df associated with a t statistic based on a standard error of the form

$$\sqrt{c_1MS_1 + c_2MS_2 + \cdots + c_kMS_k}$$

(a linear combination of mean square terms), use the

Satterthwaite approximation:

$$\widehat{df} = \frac{(c_1MS_1 + c_2MS_2 + \cdots + c_kMS_k)^2}{(c_1MS_1)^2/df_1 + (c_2MS_2)^2/df_2 + \cdots + (c_kMS_k)^2/df_k}$$

Recall that for the milk data, we have

$$\begin{aligned} \widehat{SE}(\bar{Y}_{+++}) &= \sqrt{\frac{1}{40}(MS[A] + MS[B] - MS[AB])} \\ &= \sqrt{\frac{1}{40}(0.58 + 17.73 - 0.05)} \\ &= 0.6757 \end{aligned}$$

The degrees of freedom associated with this linear combination is approximated by

$$\widehat{df} = \frac{(0.6757)^4}{(\frac{1}{40}17.73)^2/3 + (\frac{1}{40}0.58)^2/4 + (\frac{1}{40}0.045)^2/12} = 3.18$$

Using $t(0.025, 3.18) = 3.08$, a 95% confidence interval for the mean μ among the population of all labs and samples is given by

$$6.82 \pm 3.08(0.6757)$$

or

$$6.82 \pm 2.08$$

(log scale)

```
data one;
  infile "milk.dat" firstobs=4;
  input sample lab y;
  ly=log(y);
run;
```

```
proc mixed cl;
  class sample lab;
  model ly=/s ddfm=satterth cl;
  random sample lab sample*lab;
run;
```

The SAS System
The Mixed Procedure
Model Information

1

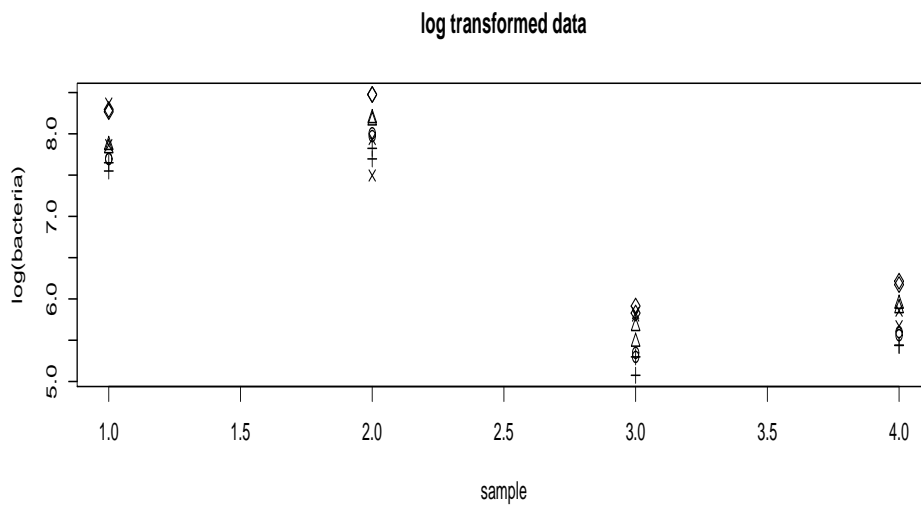
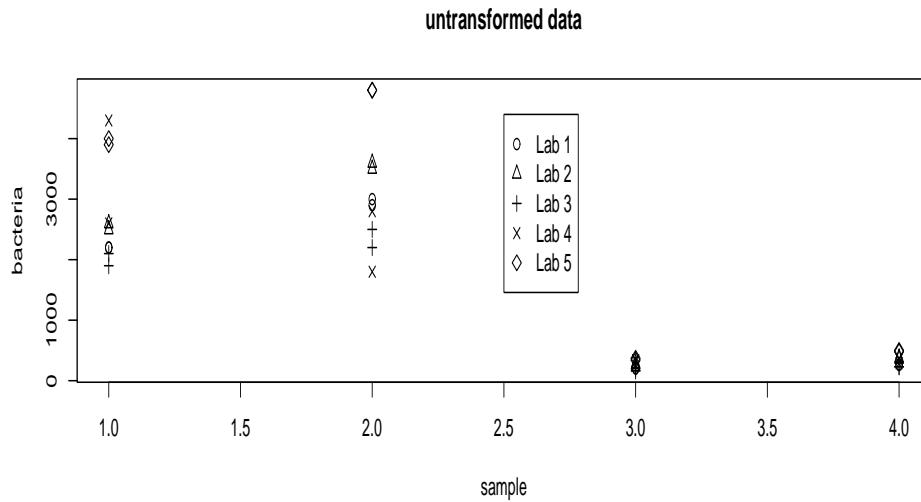
Dependent Variable	ly
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Satterthwaite

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
sample	1.7685	0.05	0.5664	24.8486
lab	0.06630	0.05	0.02233	0.7260
sample*lab	0.01492	0.05	0.005761	0.09261
Residual	0.01541	0.05	0.009017	0.03213

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t	Alpha
Intercept	6.8156	0.6757	3.18	10.09	0.0016	0.05
Effect						
Intercept			4.7325		8.8987	



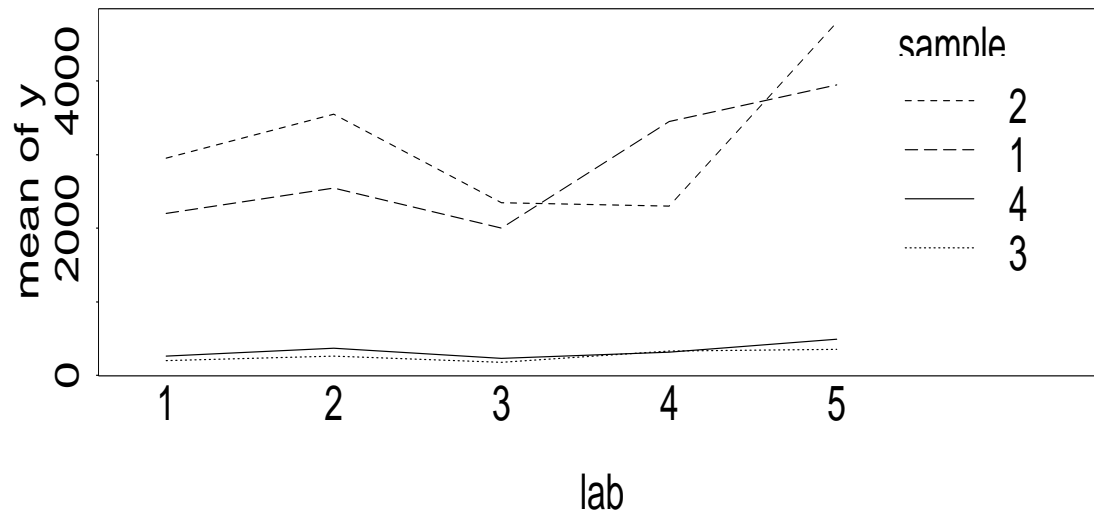
```

milk.data <- read.table("milk.dat",skip=3,col.names=c("sample","lab","bacteria"))
attach(milk.data)
postscript(file="milkplot1.ps")
par(mfrow=c(2,1)) # A 2x1 template (for two plots in a single column)
plot(x=sample,y=bacteria,pch=lab)
title("untransformed data")
legend(2.5,4400,pch=1:5,legend=c("Lab 1","Lab 2","Lab 3","Lab 4","Lab 5"))

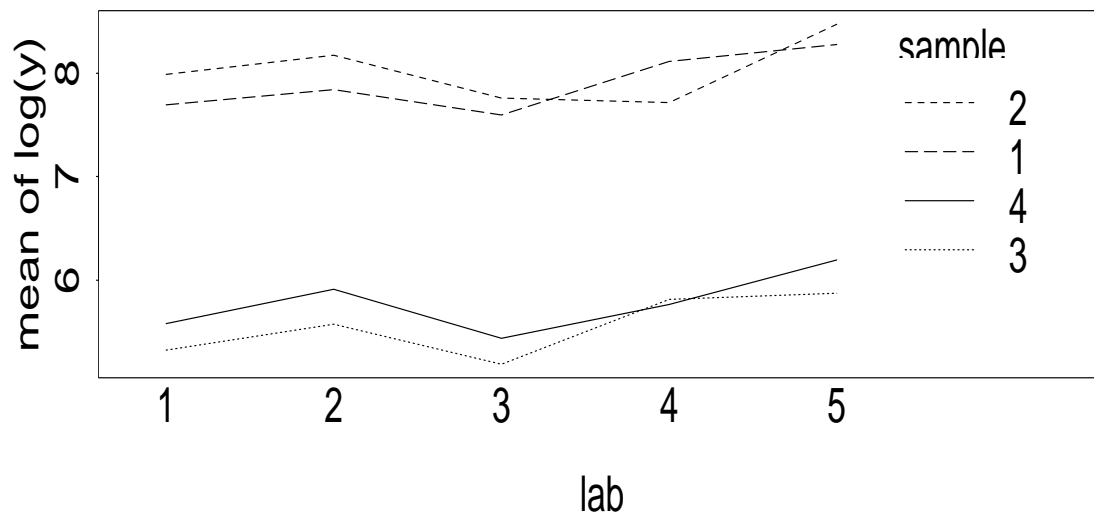
plot(x=sample,y=log(bacteria),pch=lab)
title("log transformed data")
postscript()

```

Interaction plot for milk - raw counts



Interaction plot for milk - log(counts)



A nested design

Experiment to study effect of drug and method of administration on fasting blood sugar in diabetic patients

- First factor is drug: brand I tablet, brand II tablet, insulin injection
- Second factor is type of administration (see table)

Drug (i)	Type of Administration (j)	Mean $\bar{y}_{j(i)}$	Variance $s_{j(i)}^2$	Mean $\bar{y}_{+(i)}$
Brand I tablet	$30mg \times 1$	15.7	6.3	17.7
	$15mg \times 2$	19.7	9.3	
Brand II tablet	$20mg \times 1$	20	1	18.7
	$10mg \times 2$	17.3	6.3	
Insulin injection	before breakfast	28	4	30.5
	before supper	33	9	

(This is exercise 13.35. Grand mean is $\bar{y}_{+++} = 22.3$.)

Definition: Factor B is nested in factor A if there is a new set of levels of factor B for every different level of factor A .

Analysis of variance in nested designs

Consider a two-factor design in which factor B is nested in factor A . Let Y_{ijk} denote the k^{th} response at level j of factor B within level i of factor A . A model:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + E_{ijk}$$

for $i = 1, 2, \dots, a$, $j = 1, 2, \dots, b_i$, $k = 1, 2, \dots, n$

$SS[Total]$ can be broken down into components reflecting variability due to A , $B(A)$ and variability not due to either factor ($SS[E]$):

$$SS[Total] = SS[A] + SS[B(A)] + SS[E]$$

$$SS[Total] = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{+++})^2$$

$$SS[A] = \sum_i \sum_j \sum_k (\bar{y}_{i++} - \bar{y}_{+++})^2$$

$$SS[B(A)] = \sum_i \sum_j \sum_k (\bar{y}_{ij+} - \bar{y}_{i++})^2$$

$$SS[E] = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij+})^2$$

The ANOVA table looks like

Source	d.f.	Sum of squares	Mean Square	F
A	$a - 1$	$SS[A]$	$MS[A] = \frac{SS[T]}{(a-1)}$	$F_A = \frac{MS[A]}{MS[E]}$
$B(A)$	$\sum_i (b_i - 1)$	$SS[B(A)]$	$MS[B(A)] = \frac{SS[B(A)]}{\sum (b_i - 1)}$	$F_{B(A)} = \frac{MS[B(A)]}{MS[E]}$
Error	$N - \sum b_i$	$SS[E]$	$MS[E] = \frac{SS[E]}{(N-t)}$	
Total	$N - 1$	$SS[TOT]$		

If $b_1 = b_2 = \dots = b_a = b$ then $\sum (b_i - 1) = a(b - 1)$ and $df_E = ab(n - 1)$.

Inference from nested designs

To test $H_0 : \alpha_i \equiv 0$, use F_A on $a - 1$ and df_E degrees of freedom.

To test $H_0 : \beta_{j(i)} \equiv 0$, for all i, j , use $F_{B(A)}$ on $\sum(b_i - 1)$ and df_E degrees of freedom.

For the diabetics blood sugar data, with $\bar{y}_{+++} = 22.3$ and means

Drug (i)	Type of Administration (j)	Mean $\bar{y}_{j(i)}$	Variance $s_{j(i)}^2$	Mean $\bar{y}_{+(i)}$
Brand I tablet	$30mg \times 1$	15.7	6.3	17.7
	$15mg \times 2$	19.7	9.3	
Brand II tablet	$20mg \times 1$	20	1	18.7
	$10mg \times 2$	17.3	6.3	
Insulin injection	before breakfast	28	4	30.5
	before supper	33	9	

$$\begin{aligned} SS[A] &= 2(3)[(17.7 - 22.3)^2 + (18.7 - 22.3)^2 + (30.5 - 22.3)^2] \\ &= 611.4 \end{aligned}$$

$$\begin{aligned} SS[B(A)] &= 3[(15.7 - 17.7)^2 + (19.7 - 17.7)^2 + (20.0 - 18.7)^2 \\ &\quad + (17.3 - 18.7)^2 + (28 - 30.5)^2 + (33 - 30.5)^2] \\ &= 72.2 \end{aligned}$$

$$SS[E] = 72$$

Q1: How many df associated with $SS[A]$?

Q2: How many df associated with $SS[B(A)]$?

Q3: How many df associated with $SS[E]$?

```

data one;
  infile "blsugar.dat" firstobs=2 dlm='09'x;
  input a b rep y;
  drug=a;  admin=b;
run;

proc glm;
  class a b;
  model y=a b(a);
  output out=two p=p r=r;
  means a b(a)/lsd;
  estimate "effect of B within A=1" b(a) -1 1;
  estimate "effect of B within A=2" b(a) 0 0 -1 1;
  estimate "effect of B within A=3" b(a) 0 0 0 0 -1 1;
  estimate "A=1 mean - A=2 mean" a 1 -1;
  estimate "A=1 mean - A=3 mean" a 1 0 -1;
  estimate "A=2 mean - A=3 mean" a 0 1 -1;
run;

```

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	683.6111111	136.7222222	22.79	<.0001
Error	12	72.0000000	6.0000000		
Corrected Total	17	755.6111111			

R-Square	Coeff Var	Root MSE	y Mean
0.904713	10.99522	2.449490	22.27778

Source	DF	Type I SS	Mean Square	F Value	Pr > F
a	2	611.4444444	305.7222222	50.95	<.0001
b(a)	3	72.1666667	24.0555556	4.01	0.0344

Parameter	Estimate	Standard Error	t Value	Pr > t
effect of B within A=1	4.0000000	2.0000000	2.00	0.0687
effect of B within A=2	-2.6666667	2.0000000	-1.33	0.2072
effect of B within A=3	5.0000000	2.0000000	2.50	0.0279
A=1 mean - A=2 mean	-1.0000000	1.41421356	-0.71	0.4930
A=1 mean - A=3 mean	-12.8333333	1.41421356	-9.07	<.0001
A=2 mean - A=3 mean	-11.8333333	1.41421356	-8.37	<.0001

Conclusions?

- The administration effect B (nested in the type of drug effect A) is statistically significant ($p = 0.0344$). This is due mostly to the before breakfast/supper difference, which is estimated to be

$$\bar{y}_{32+} - \bar{y}_{31+} = 5mg/dl$$

with an (estimated) standard error of $SE = 2 = ?$.

- The effect of type of drug (factor A) is highly significant ($p < 0.0001$). Unadjusted pairwise comparisons indicate that the insulin injections yield greater changes, on average, in blood sugar than either pill and the mean changes brought by the pills don't differ significantly.
- The following contrasts may be of interest:

$$\theta_1 = \mu_{1(3)} - \frac{1}{4}(\mu_{1(1)} + \mu_{2(1)} + \mu_{1(2)} + \mu_{2(2)})$$

$$\theta_2 = \mu_{2(3)} - \frac{1}{4}(\mu_{1(1)} + \mu_{2(1)} + \mu_{1(2)} + \mu_{2(2)})$$

Exercise: Estimate them and test their significance ($H_0 : \theta_i = 0$).

More Two-factor mixed models (Ch. 14)

- Expt measures *campylobacter* counts in $N = 120$ chickens in a processing plant
 - Crossed design with two factors
 - * Location (4 levels)
 - * Day (3 levels)
 - 4×3 layout, $n = 10$ chickens per combo

Day	Location			
	Before Washer	After Washer	After mic. rinse	After chill tank
1	70070.00 (79034.49)	48310.00 (34166.80)	12020.00 (3807.24)	11790.00 (7832.05)
2	75890.00 (74551.32)	52020.00 (17686.27)	8090.00 (4848.01)	8690.00 (5526.19)
3	95260.00 (03176.00)	33170.00 (22259.08)	6200.00 (5028.81)	8370.00 (5720.15)

Data courtesy of Michael Bashor, General Mills

- An experiment to assess the variability of a particular acid among plants and among leaves of plants:

Plant i	1			2			3			4			
	Leaf j	1	2	3	1	2	3	1	2	3	1	2	3
$k = 1$		11.2	16.5	18.3	14.1	19.0	11.9	15.3	19.5	16.5	7.3	8.9	11.3
$k = 2$		11.6	16.8	18.7	13.8	18.5	12.4	15.9	20.1	17.2	7.8	9.4	10.9
$k = 3$		12.0	16.1	19.0	14.2	18.2	12.0	16.0	19.3	16.9	7.0	9.3	10.5

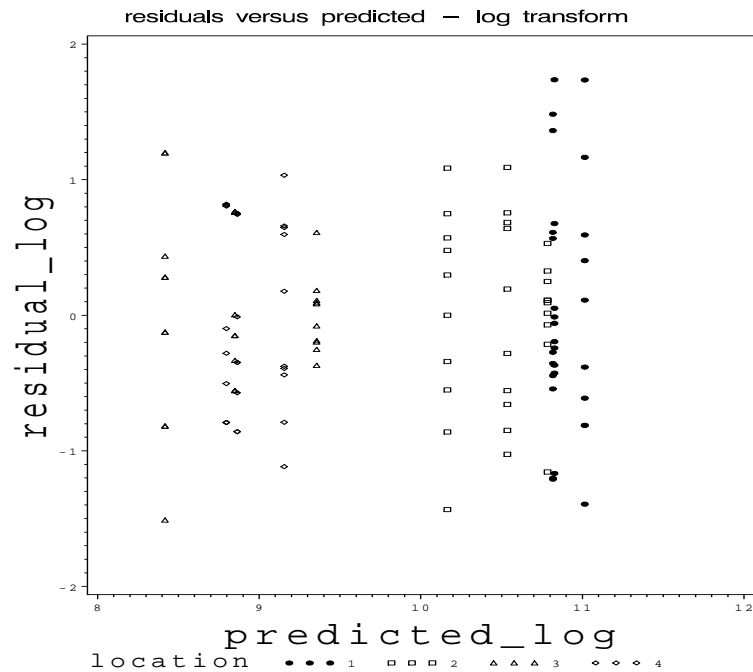
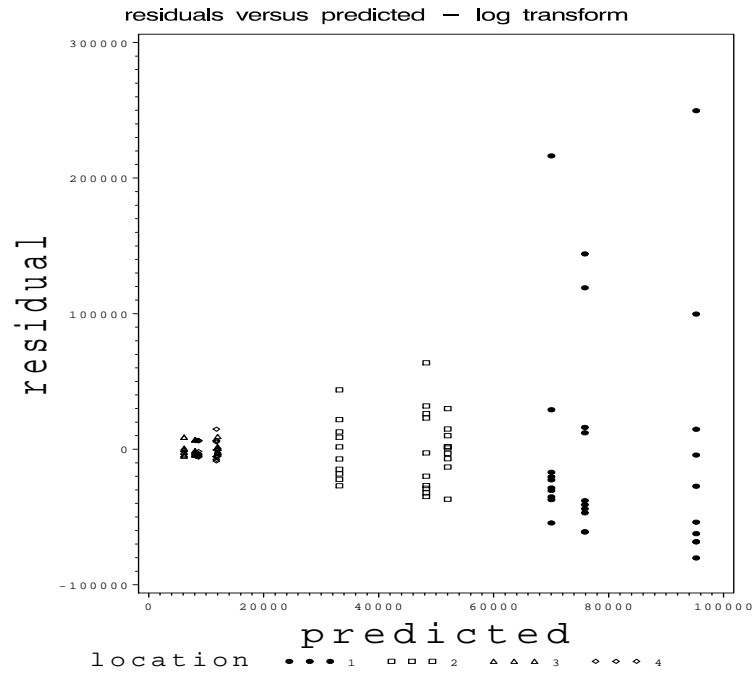
- Study of effect of salinity on barley growth in a controlled medium.

Salinity	Container	Weights (g)		
c	1	11.29	11.08	11.1
c	2	7.37	6.55	8.5
6b	1	5.64	5.98	5.69
6b	2	4.2	3.34	4.21
12b	1	4.83	4.77	5.66
12b	2	3.28	2.61	2.69

- Total of 6 containers

Analysis of *Campylobacter* counts on chickens data

Residual plots (resid .vs \hat{y}) for bacteria counts, after fitting two factor fixed effects models (similar plots for mixed models):



```

data one;  /* Bashor data */
  infile "bashor.dat" firstobs=3;
  input day location y;          ly=log(y);
run;

proc glm;
  class day location;
  model y ly=location|day;
  output out=two r=residual residual_log p=predicted predicted_log;
run;
/*
symbol1 value=dot color=black;      symbol2 value=square color=black;
symbol3 value=triangle color=black; symbol4 value=diamond color=black;

axis1 offset=(1,1) label=(height=3);
axis2 offset=(1,1) label=(height=3 angle=90);
legend1 label=(height=2);

proc gplot data=two;
  title "residuals versus predicted";
  plot residual_log*predicted_log=location/haxis=axis1 vaxis=axis2 legend=legend1;
  plot residual_log*predicted_log=location/haxis=axis1 vaxis=axis2 legend=legend1;
run; */
proc mixed method=type3 cl;
  class day location;
  model ly=location/ddfm=satterth outp=predz;
  random day day*location;
  lsmeans location/adj=tukey;
run;
/*proc glm;  * the old way of doing things, before PROC MIXED ;
  class day location;
  model ly=day|location;
  random day day*location;
  test h=location e=day*method;
  lsmeans location/pdiff;  *wrong;
run;*/

proc mixed method=type3;  * to get ANOVA table with EMS terms;
*proc mixed cl;  * to get asymmetric confidence intervals ;
  class day location;
  model ly=location/ddfm=satterth;
  random day day*location;
  lsmeans location/adj=tukey;
run;

```

The SAS System
The Mixed Procedure
Model Information

1

Data Set	WORK.ONE
Dependent Variable	ly
Covariance Structure	Variance Components
Estimation Method	Type 3
Residual Variance Method	Factor
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Satterthwaite

Class Level Information

Class	Levels	Values
day	3	1 2 3
location	4	1 2 3 4

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square
location	3	97.865388	32.621796	Var(Residual) + 10 Var(day*location) + Q(location)
day	2	2.787355	1.393677	Var(Residual) + 10 Var(day*location) + 40 Var(day)
day*location	6	4.533565	0.755594	Var(Residual) + 10 Var(day*location)
Residual	108	59.254946	0.548657	Var(Residual)

Type 3 Analysis of Variance

Source	Error Term	Error DF	F Value	Pr > F
location	MS(day*location)	6	43.17	0.0002
day	MS(day*location)	6	1.84	0.2375
day*location	MS(Residual)	108	1.38	0.2303

(generated by 2nd run of PROC MIXED)

*	Cov Parm	Estimate	Alpha	Lower	Upper
*	day	0.01595	0.05	0.002071	1156981
*	day*location	0.02069	0.05	0.002844	145734
*	Residual	0.5487	0.05	0.4274	0.7303

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
location	3	6	43.17	0.0002

Least Squares Means

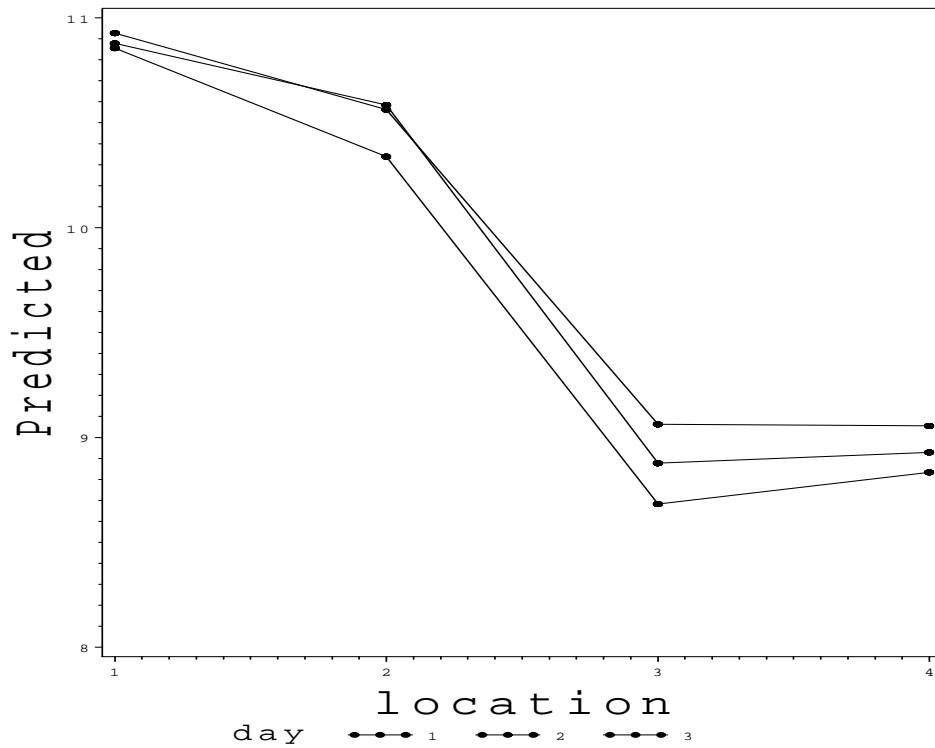
Effect	location	Estimate	Standard Error	DF	t Value	Pr > t
location	1	10.8870	0.1747	7.33	62.33	<.0001
location	2	10.4953	0.1747	7.33	60.09	<.0001
location	3	8.8745	0.1747	7.33	50.81	<.0001
location	4	8.9394	0.1747	7.33	51.18	<.0001

Differences of Least Squares Means

Effect	location	_location	Estimate	Standard Error	DF	t Value	Pr > t
location	1	2	0.3917	0.2244	6	1.75	0.1316
location	1	3	2.0125	0.2244	6	8.97	0.0001
location	1	4	1.9476	0.2244	6	8.68	0.0001
location	2	3	1.6208	0.2244	6	7.22	0.0004
location	2	4	1.5559	0.2244	6	6.93	0.0004
location	3	4	-0.06488	0.2244	6	-0.29	0.7823

Differences of Least Squares Means

Effect	location	_location	Adjustment	Adj P
location	1	2	Tukey-Kramer	0.3801
location	1	3	Tukey-Kramer	0.0004
location	1	4	Tukey-Kramer	0.0005
location	2	3	Tukey-Kramer	0.0015
location	2	4	Tukey-Kramer	0.0018
location	3	4	Tukey-Kramer	0.9907

Theory for mixed/crossed model used to analyze *Campylobacter* dataDiscussion of MIXED output

Model

$$Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}$$

w/ variance components $\sigma_B^2, \sigma_{\alpha B}^2, \sigma^2$.

Campylobacter analysis, continued

Fixed Factor A: location

Random Factor B: day

To test $H_0 : \sigma_{\alpha B}^2 = 0$, use

$$F_{AB} = \frac{MS[AB]}{MS[E]} = \frac{0.76}{0.55} = 1.38$$

on $(a - 1)(b - 1) = 6$ and $ab(n - 1) = 108$ *df*. The *p*-value is 0.2303. providing no evidence of a random day \times location interaction effect. The variance component for this random effect is estimated by

$$\hat{\sigma}_{\alpha B}^2 = \frac{MS[AB] - MS[E]}{n} = \frac{0.76 - 0.55}{10} = 0.021$$

Intepretation: there is no evidence that day-to-day variability varies by location. The estimated variance component is itself very small.

$$\begin{aligned} \hat{\sigma}^2 &= MS[E] = \boxed{0.55} \\ \hat{\sigma}_{\alpha B}^2 &= \frac{MS[AB] - MS[E]}{n} \\ &= \frac{0.76 - 0.55}{10} = \boxed{0.021} \\ \hat{\sigma}_B^2 &= \frac{MS[B] - MS[AB]}{na} \\ &= \frac{1.39 - 0.76}{40} = \boxed{0.016} \end{aligned}$$

Implied correlation structure

What is the correlation of two observations taken on the same day

- at the same location?
- at different locations?

Recall that $Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}$.

$$\begin{aligned}
 \text{Corr}(Y_{ijk_1}, Y_{ijk_2}) &= \frac{\text{Cov}(Y_{ijk_1}, Y_{ijk_2})}{\sigma^2 + \sigma_B^2 + \sigma_{\alpha B}^2} \\
 &= \frac{\text{Cov}(B_i, B_i) + \text{Cov}((\alpha B)_{ij}, (\alpha B)_{ij})}{\sigma^2 + \sigma_B^2 + \sigma_{\alpha B}^2} \\
 &= \frac{\sigma_B^2 + \sigma_{\alpha B}^2}{\sigma^2 + \sigma_B^2 + \sigma_{\alpha B}^2} \\
 \text{Corr}(Y_{1jk_1}, Y_{2jk_2}) &= \frac{\text{Cov}(Y_{1jk_1}, Y_{2jk_2})}{\sigma^2 + \sigma_B^2 + \sigma_{\alpha B}^2} \\
 &= \frac{\text{Cov}(B_i, B_i)}{\sigma^2 + \sigma_B^2 + \sigma_{\alpha B}^2} \\
 &= \frac{\sigma_B^2}{\sigma^2 + \sigma_B^2 + \sigma_{\alpha B}^2}
 \end{aligned}$$

Estimates of these correlations are

- $\frac{0.016+0.021}{0.016+0.021+0.55} = \frac{0.037}{.587} = 0.063$
- $\frac{0.016}{0.016+0.021+0.55} = \frac{0.016}{.587} = 0.027$

Which is which?

What about the correlation of two observations on different days?

Some analysis of fixed effects

Consider testing for a fixed effect of location. That is, test the hypothesis that average bacteria counts are constant across the locations,

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

$$F_A = \frac{MS[A]}{MS[AB]} = \frac{32.6}{0.76} = 43.2$$

on $a - 1 = 3$ and $(a - 1)(b - 1) = 6$ *df*, which is significant ($p = 0.0002$).

Campylobacter analysis, continued

To estimate the a pairwise comparison among location means, such as, $\alpha_4 - \alpha_3$, consider

$$\hat{\theta} = \bar{y}_{4++} - \bar{y}_{3++} = 8.940 - 8.875 = -0.065$$

Note that

$$\text{Var}(\bar{Y}_{4++} - \bar{Y}_{3++}) \neq \sigma^2\left(\frac{1}{nb} + \frac{1}{nb}\right)$$

(Why not?)

What is $SE(\hat{\theta})$ and how can it be estimated?

$$\begin{aligned} \hat{\theta} &= \bar{Y}_{2++} - \bar{Y}_{1++} \\ &= \alpha_2 + \bar{B} + \overline{\alpha B}_{2+} + \bar{E}_{2++} \\ &\quad - (\alpha_1 + \bar{B} + \overline{\alpha B}_{1+} + \bar{E}_{1++}) \\ &= \alpha_2 - \alpha_1 + \overline{\alpha B}_{2+} - \overline{\alpha B}_{1+} + \bar{E}_{2++} - \bar{E}_{1++} \end{aligned}$$

which has variance

$$\begin{aligned} \text{Var}(\hat{\theta}) &= \text{Var}(\overline{\alpha B}_{2+}) + \text{Var}(\overline{\alpha B}_{1+}) + \text{Var}(\bar{E}_{2++}) + \text{Var}(\bar{E}_{1++}) \\ &= 2\frac{\sigma_{\alpha B}^2}{b} + 2\frac{\sigma^2}{nb} \\ &= \frac{2}{nb}(\sigma^2 + n\sigma_{\alpha B}^2) \end{aligned}$$

which can be estimated nicely on $(a-1)(b-1) = 6df$ by

$$\hat{\text{Var}}(\hat{\theta}) = \frac{2}{nb}MS[AB]$$

for the chickens, where $\bar{y}_{4++} - \bar{y}_{3++} = -0.06$ the SE is

$$\sqrt{\hat{\text{Var}}(\hat{\theta})} = \sqrt{\frac{2}{3 * 10}0.76} = 0.22$$

Since $t(0.025, 6) = 2.45$, a 95% c.i. for θ given by $-0.06 \pm 2.45(0.22)$.

Campylobacter analysis, continued

Reporting standard errors for sample means of levels of fixed factor, like LOCATION means, is a little messier:

$$\begin{aligned} \bar{Y}_{i++} &= \mu + \alpha_i + \bar{B} + \bar{\alpha B}_{i+} + \bar{E}_{i++} \\ \text{Var}(\bar{Y}_{i++}) &= \text{Var}(\bar{B}) + \text{Var}(\bar{\alpha B}_{i+}) + \text{Var}(\bar{E}_{i++}) \\ &= \frac{\sigma_B^2}{b} + \frac{\sigma_{\alpha B}^2}{b} + \frac{\sigma^2}{nb} \\ &= \frac{1}{nb}(n\sigma_B^2 + n\sigma_{\alpha B}^2 + \sigma^2) \\ &\quad \text{estimated by} \\ \widehat{\text{Var}}(\bar{Y}_{i++}) &= \frac{1}{nb}(n\hat{\sigma}_B^2 + n\hat{\sigma}_{\alpha B}^2 + \hat{\sigma}^2) \\ &= \text{algebra yields a linear combo of multiple EMS terms} \\ &= \frac{1}{nab}\{(a-1)EMS[AB] + EMS[B]\} \end{aligned}$$

The standard error is estimated easily enough:

$$\begin{aligned} \widehat{SE}(\bar{Y}_{i++}) &= \sqrt{\frac{1}{nab}\{(a-1)MS[AB] + MS[B]\}} \\ &= \sqrt{\frac{1}{120}\{(4-1)0.76 + 1.39\}} \\ &= \sqrt{0.03} = 0.175 \end{aligned}$$

but the df must be approximated using the Satterthwaite approach

$$\hat{df} = \frac{0.175^4}{\frac{1}{120^2} \left(\frac{((4-1)0.76)^2}{6} + \frac{1.39^2}{2} \right)} = 7.33$$

with $df_{AB} = 6$, $df_B = 2$. Since $t(0.025, 7.33) = 2.34$, a 95% c.i. for the population mean of location 1, for example, is $\boxed{10.9 \pm 2.34(0.175)}$.

SAS code to fit two-factor random effects model for plant acid data

Nested or crossed?

```
options ls=75 nodate;

data one;
  infile "plantacid.dat";
  input y plant leaf rep;
run;

proc mixed cl method=type3;
*proc mixed cl;
  class plant leaf;
  model y=/s cl;
  random plant leaf(plant);
run;

goptions colors=(black) dev=pslepsz;
*goptions colors=(black);

axis1 value=(h=2) offset=(10);

symbol1 value=dot h=1.5;
symbol2 value=diamond h=1.5;
symbol3 value=plus h=1.5;

proc gplot;
  title "plant acids";
  plot y*plant=leaf/haxis=axis1;
run;
```

The Mixed Procedure
Class Level Information

Class	Levels	Values
plant	4	1 2 3 4
leaf	3	1 2 3

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square
plant	3	343.178889	114.392963
leaf(plant)	8	187.453333	23.431667
Residual	24	3.033333	0.126389

Source	Expected Mean Square	Error Term	Error DF
plant	Var(Residual) + 3 Var(leaf(plant)) + 9 Var(plant)	MS(leaf(plant))	8
leaf(plant)	Var(Residual) + 3 Var(leaf(plant))	MS(Residual)	24
Residual	Var(Residual)	.	.

Source	F Value	Pr > F
plant	4.88	0.0324
leaf(plant)	185.39	<.0001

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
plant	10.1068	0.05	-10.3930	30.6066
leaf(plant)	7.7684	0.05	0.1142	15.4227
Residual	0.1264	0.05	0.07706	0.2446

/*Covariance Parameter Estimates*/

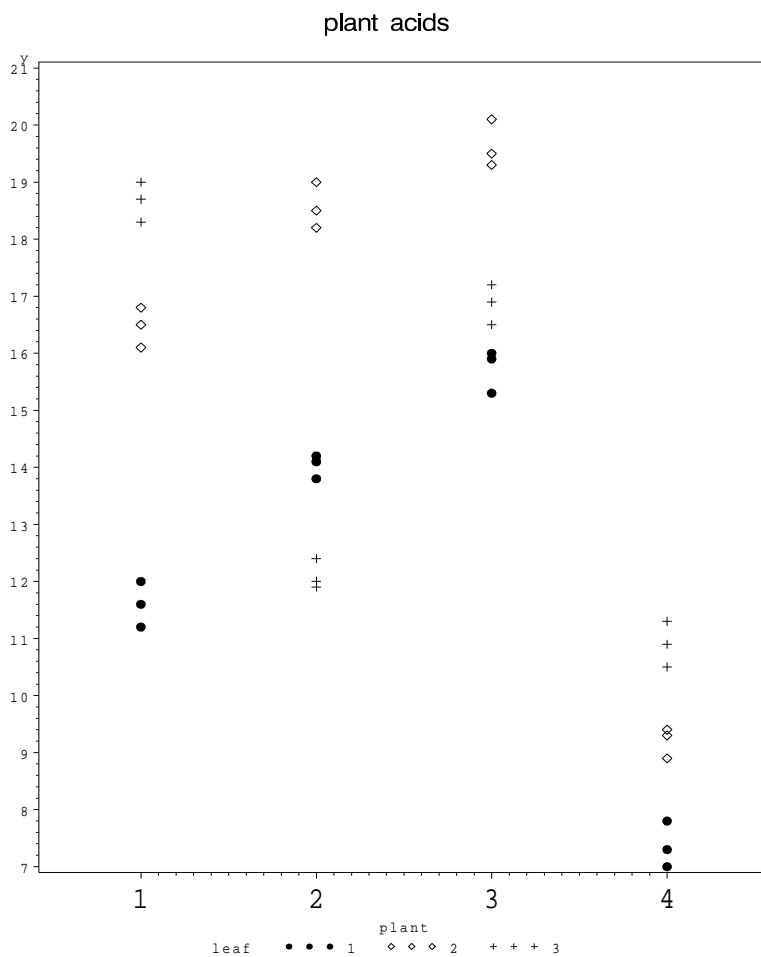
Cov Parm	Estimate	Alpha	Lower	Upper
plant	10.1068	0.05	2.6599	499.70
leaf(plant)	7.7684	0.05	3.5322	28.7787
Residual	0.1264	0.05	0.07706	0.2446

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t	Alpha
Intercept	14.2611	1.7826	3	8.00	0.0041	0.05

Solution for Fixed Effects

Effect	Lower	Upper
Intercept	8.5882	19.9341



Discussion of MIXED output and analysis of plant acid data

Random, nested model

$$Y_{ijk} = \mu + A_i + B_{j(i)} + E_{ijk}$$

w/ variance components $\sigma^2, \sigma_A^2, \sigma_{B(A)}^2$.

To test for random effect of nested factor B (leaf), $H_0 : \sigma_{B(A)}^2 = 0$,

$$F = \frac{MS[B(A)]}{MS[E]} = \frac{23.4}{0.13} = 185.4$$

on $(b - 1)a = 8$ and $(n - 1)ab = 24$ df (p -value < 0.0001).

To test for random effect of factor A (plant), $H_0 : \sigma_A^2 = 0$,

$$F = \frac{MS[A]}{MS[B(A)]} = \frac{114.4}{23.4} = 4.88$$

on $a - 1 = 3$ and $(b - 1)a = 8$ df with $p = 0.0324$.

Reminder: Watch that denominator MS !

$$\begin{aligned} \hat{\sigma}^2 &= MS[E] = \boxed{0.13} \\ \hat{\sigma}_{B(A)}^2 &= \frac{MS[B(A)] - MS[E]}{n} \\ &= \frac{23.4 - 0.13}{3} = \boxed{7.8} \\ \hat{\sigma}_A^2 &= \frac{MS[A] - MS[B(A)]}{nb} \\ &= \frac{114.4 - 23.4}{9} = \boxed{10.1} \end{aligned}$$

So there is some evidence of both a random plant effect and a random leaf effect, nested in plant. The magnitudes of these effects are quantified by the estimated variance components. The statistical significance addressed by the p -values.

Implied correlation structure for plant acids

What is the correlation of two observations taken from the same plant

- and the same leaf?
- and different leaves?

Recall that $Y_{ijk} = \mu + A_i + B_{j(i)} + E_{ijk}$.

$$\begin{aligned}
 \text{Corr}(Y_{ijk_1}, Y_{ijk_2}) &= \frac{\text{Cov}(Y_{ijk_1}, Y_{ijk_2})}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2} \\
 &= \frac{\text{Cov}(A_i, A_i) + \text{Cov}(B_{j(i)}, B_{j(i)})}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2} \\
 &= \frac{\sigma_A^2 + \sigma_{B(A)}^2}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2} \\
 \text{Corr}(Y_{ij_1k_1}, Y_{ij_2k_2}) &= \frac{\text{Cov}(Y_{ij_1k_1}, Y_{ij_2k_2})}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2} \\
 &= \frac{\text{Cov}(A_i, A_i)}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2} \\
 &= \frac{\sigma_A^2}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2}
 \end{aligned}$$

Estimates of these correlations are

- $\frac{10.1+7.8}{10.1+7.8+0.13} = \frac{17.9}{18.0} = 0.99$
- $\frac{10.1}{10.1+7.8+0.13} = \frac{10.1}{18.0} = 0.56$

This means that two measurements taken on the same leaf are almost perfectly correlated. Almost all the variation in any measurement can be explained by the leaf and plant effects.

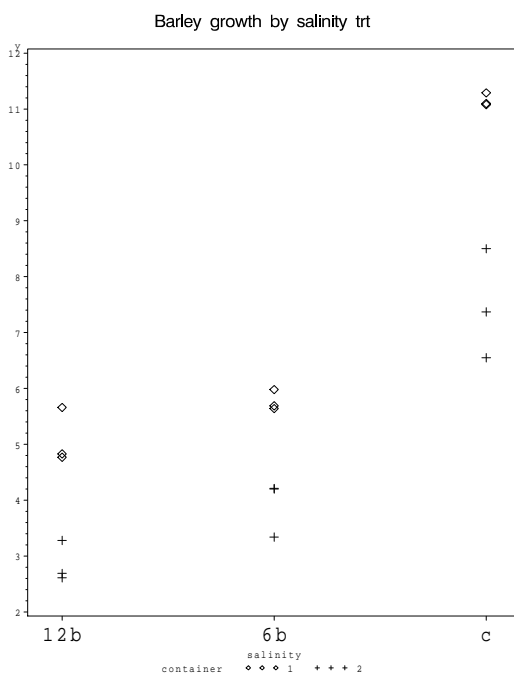
SAS code to analyze data from two-factor Barley growth experiment

Nested or crossed?

```

data one;    /* Barley growth and salinity */
  input salinity $3. container @;
  do rep=1 to 3; input y @;  output; end;
  cards;
c   1  11.29 11.08 11.1
c   2   7.37  6.55  8.5
6b  1   5.64  5.98  5.69
6b  2   4.2  3.34  4.21
12b 1   4.83  4.77  5.66
12b 2   3.28  2.61  2.69
;
run;
proc mixed cl method=type3; /* proc mixed cl; */
  class salinity container;
  model y=salinity/s cl ddfm=satterth;
  random container(salinity);
  lsmeans salinity/tdiff pdiff;
run;
proc gplot;
  plot y*salinity=container;
run;

```



The SAS System
The Mixed Procedure

1

Class	Levels	Values
salinity	3	12b 6b c
container	2	1 2

Total Observations 18

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square
salinity	2	98.572211	49.286106
container(salinity)	3	32.939750	10.979917
Residual	12	3.273067	0.272756

Source	Expected Mean Square
salinity	Var(Residual) + 3 Var(container(salinity)) + Q(salinity)
container(salinity)	Var(Residual) + 3 Var(container(salinity))
Residual	Var(Residual)

Source	Error Term	Error		
		DF	F Value	Pr > F
salinity	MS(container(salinity))	3	4.49	0.1254
container(salinity)	MS(Residual)	12	40.26	<.0001
Residual

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
/*container(salinity)	3.5691	0.05	1.1223	55.2528*/
container(salinity)	3.5691	0.05	-2.2885	9.4266
Residual	0.2728	0.05	0.1403	0.7432

Solution for Fixed Effects

Effect	salinity	Estimate	Standard Error	DF	t Value	Pr > t	Alpha
Intercept		9.3150	1.3528	3	6.89	0.0063	0.05
salinity	12b	-5.3417	1.9131	3	-2.79	0.0683	0.05
salinity	6b	-4.4717	1.9131	3	-2.34	0.1015	0.05
salinity	c	0

Solution for Fixed Effects

Effect	salinity	Lower	Upper
Intercept		5.0099	13.6201
salinity	12b	-11.4300	0.7467
salinity	6b	-10.5600	1.6167
salinity	c	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
salinity	2	3	4.49	0.1254

Least Squares Means

Effect	salinity	Estimate	Standard Error	DF	t Value	Pr > t
salinity	12b	3.9733	1.3528	3	2.94	0.0606
salinity	6b	4.8433	1.3528	3	3.58	0.0373
salinity	c	9.3150	1.3528	3	6.89	0.0063

Differences of Least Squares Means

Effect	salinity	_salinity	Estimate	Standard Error	DF	t Value	Pr > t
salinity	12b	6b	-0.8700	1.9131	3	-0.45	0.6802
salinity	12b	c	-5.3417	1.9131	3	-2.79	0.0683
salinity	6b	c	-4.4717	1.9131	3	-2.34	0.1015

Test for container effect:

$$F = \frac{MS[B(A)]}{MS[E]} = \frac{10.98}{0.27} = 40.3$$

on $a(b - 1) = 3$ and $ab(n - 1) = 12$ *df*. (Here, $p < 0.0001$) Test for salinity treatment effect:

$$F = \frac{MS[A]}{MS[B(A)]} = \frac{49.3}{10.98} = 4.49$$

on $a - 1 = 2$ and $a(b - 1) = 3$ *df*. (Here, $p = 0.1254$, which is not significant.)

Note that using the wrong error term leads to a different conclusion:

$$F_{wrong} = \frac{MS[A]}{MS[E]} = \frac{49.3}{0.27} = 180.7$$

with $p < 0.0001$ on 2 and 12 *df*.

Is there a treatment effect going on? Why or why not? Which F ratio is appropriate? Is it a matter of modelling preference?

Estimated variance components:

$$\hat{\sigma}^2 = MS[E] = 0.27$$

$$\hat{\sigma}_{B(A)}^2 = (MS[B(A)] - MS[E])/n = (10.98 - 0.27)/3 = 3.57$$

Implied correlation structure: for two different ($k_1 \neq k_2$) observations from the same container,

$$\text{corr}(Y_{ijk_1}, Y_{ijk_2}) = \frac{\sigma_{B(A)}^2}{\sigma^2 + \sigma_{B(A)}^2}$$

which is estimated by

$$\hat{\rho} = \frac{3.57}{0.27 + 3.57} = 0.93$$

Inference for fixed effects is clean, without need for Satterthwaite:

$$\bar{Y}_{i++} = \mu + \alpha_i + \bar{B}_{+(i)} + \bar{E}_{i++}$$

which has

$$SE(\bar{Y}_{i++}) = \sqrt{\frac{\sigma_B^2}{b} + \frac{\sigma^2}{nb}}$$

which can be estimated cleanly on $(b-1)a = 3$ *df* by

$$\widehat{SE}(\bar{Y}_{i++}) = \sqrt{\frac{1}{nb}MS[B(A)]} = \sqrt{\frac{1}{6}11.0} = \boxed{1.35}$$

and differences are just as easy:

$$\bar{Y}_{i_1++} - \bar{Y}_{i_2++} = \alpha_{i_1} - \alpha_{i_2} + \bar{B}_{+(i_1)} - \bar{B}_{+(i_2)} + \bar{E}_{i_1++} - \bar{E}_{i_2++}$$

which has

$$SE(\bar{Y}_{i_1++} - \bar{Y}_{i_2++}) = \sqrt{2\frac{\sigma_B^2}{b} + 2\frac{\sigma^2}{nb}} = \sqrt{\frac{2}{nb}(\sigma^2 + n\sigma_B^2)}$$

which can be estimated cleanly on $(b-1)a = 3$ *df* by

$$\widehat{SE}(\bar{Y}_{i_1++} - \bar{Y}_{i_2++}) = \sqrt{\frac{2}{nb}MS[B(A)]} = \sqrt{\frac{2}{6}11.0} = \boxed{1.91}$$

(This is a good place to go back and look at RBD expt w/ random block effects. The experiment measured assembly times for iv injection systems.)

ST 512: Exptl Stats for Biol. Sciences II

Weeks 13-14 Split-plots: a repeated measures design

Reading Ch. 16.

Repeated measures models

Consider an experiment to study effects of irrigation and aerially sprayed pesticide on yields of different varieties of corn.

Factors:

- A : Pesticide treatment, $a = 3$ levels
- B : Irrigation treatment, $b = 4$ levels (called ‘treatment, trt’)
- Plots, $n = 2$ per level of A , total of $na = 6$ plots

For the moment, ignore B , or fix the level of B .

plot	pest	y
1	1	53.4
2	1	46.5
1	2	54.3
2	2	57.2
1	3	55.9
2	3	57.4

One-way ANOVA for A effect and plot effects:

Source	df
A : Pesticide	$a - 1 = 2$
Error or “plots”	$(n - 1)a = (2 - 1)3 = 3$
Total	$an - 1 = 5$

Split-plot design

Levels of factor B are randomly assigned to $b = 4$ subplots within each of the $na = 6$ plots in a *split-plot* design:

pest	plot	B1	B2	B3	B4
1	1	53.4	53.8	58.2	59.5
1	2	46.5	51.1	49.2	51.3
2	1	54.3	56.3	60.4	64.5
2	2	57.2	56.9	61.6	66.8
3	1	55.9	58.6	62.4	64.5
3	2	57.4	60.2	57.2	62.7

Each row corresponds to one of $na = 6$ plots. Each plot is divided into $b = 4$ subplots and levels of factor B are assigned to these at random.

The ANOVA table on the preceding page is at the whole plot level. Sources of variation for the split-plot level:

Source	df
B: treatments	$b - 1 = 3$
$A \times B$	$(a - 1)(b - 1) = 6$
$B \times \text{plot}(A)$ aka Subplot error	$(b - 1)(n - 1)a = 9$

- A - a *between plots* or *between subjects* factor
- B - a *within plots* or *within subjects* factor
- Plots are ‘subjects’ in repeated measures terminology, where time is often the within subjects factor.

Suggestion: draw a picture of the layout.

Source	df	EMS
A : Pesticide	$a - 1 = 2$	$\sigma^2 + b\sigma_s^2 + bn\psi_A^2$
Plot(A)	$(n - 1)a = (2 - 1)3 = 3$	$\sigma^2 + b\sigma_s^2$
B : treatments	$b - 1 = 3$	$\sigma^2 + na\psi_B^2$
$A \times B$	$(a - 1)(b - 1) = 6$	$\sigma^2 + n\psi_{AB}^2$
$B \times \text{plot}(A)$	$(b - 1)(n - 1)a = 9$	σ^2
Subplot error		
Total	$abn - 1 = 23$	

Where variance components and size effects pertain to the model for a *completely randomized split-plot design*:

$$Y_{ijk} = \underbrace{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}}_{\mu_{ij}:(\text{fixed component})} + \overbrace{S_{k(i)} + E_{ijk}}^{\text{random error component}} .$$

Here, $i = 1, \dots, a$ and $j = 1, \dots, b$ and $k = 1, \dots, n_i$ where n_i denotes the number of plots treated with level i of factor a . If n_i is constant, call it n .

Random effects and variance components:

$$S_{k(i)} \stackrel{iid}{\sim} N(0, \sigma_s^2)$$

$$E_{ijk} \stackrel{iid}{\sim} N(0, \sigma^2)$$

Size effects for fixed factors same as in prior 2-factor models. For our example, α_i denote pesticide effects, β_j denote irrigation effects, $(\alpha\beta)_{ij}$ are interactions. F -tests for fixed effects guided by EMS column above

For the corn yields data on p. 2,

Source	MS	df	EMS	F	<i>p</i> -value
<i>A</i> : Pesticide	128.1	2	$\sigma^2 + b\sigma_s^2 + bn\psi_A^2$	3.9	0.1452
Whole plot error	$MS[S(A)] = 32.6$	3	$\sigma^2 + b\sigma_s^2$	10.1	0.0031
<i>B</i> : treatments	60.2	3	$\sigma^2 + na\psi_B^2$	18.7	0.0003
<i>A</i> × <i>B</i>	4.1	6	$\sigma^2 + n\psi_{AB}^2$	1.3	0.3607
<i>B</i> × plot(<i>A</i>) (Subplot error)	$MS[E] = 3.2$	9	σ^2		
Total		23			

- $MS[S(A)]$ denotes mean square for WHOLE plots (nested in *A*)
- $MS[E]$ denotes error or subplot mean square

For pesticide effect, on 2, 3 df:

$$F = MS[A]/MS[S(A)] = 128.1/32.6$$

For irrigation effect, on 3, 9 df:

$$F = MS[B]/MS[E] = 60.2/3.2$$

For pesticide by irrigation interaction, on 6, 9 df:

$$F = MS[AB]/MS[E] = 4.1/3.2$$

For random effect of whole plots, on 3, 9 df:

$$F = MS[S(A)]/MS[E] = 32.6/3.2$$

Estimated varcomps:

$$\hat{\sigma}^2 = MS[E] = 3.2 \quad \text{and} \quad \hat{\sigma}_s^2 = (MS[S(A)] - MS[E])/4 = 7.3$$

Pairwise comparisons

Several kinds of pairwise comparisons of treatment means:

1. Main effects of A : $\bar{y}_{i_1++} - \bar{y}_{i_2++}$
2. Main effects of B : $\bar{y}_{+j_1+} - \bar{y}_{+j_2+}$
3. Simple effects of A : $\bar{y}_{i_1j+} - \bar{y}_{i_2j+}$
4. Simple effects of B : $\bar{y}_{ij_1+} - \bar{y}_{ij_2+}$
5. Interaction effects: $\bar{y}_{i_1j_1+} - \bar{y}_{i_2j_2+}$

Skipping the algebra, the standard errors for all of these comparisons, save # 3 and #5, can be estimated ‘cleanly.’ That is, with single MS terms and integer df . (See table 16.6, careful of errata)

Comparison	Variance	Estimate	df
$\bar{Y}_{i_1++} - \bar{Y}_{i_2++}$	$\frac{2}{nb}(\sigma^2 + b\sigma_s^2)$	$\frac{2}{nb}MS[S(A)]$	$(n-1)a$
$\bar{Y}_{+j_1+} - \bar{Y}_{+j_2+}$	$\frac{2}{na}\sigma^2$	$\frac{2}{na}MS[E]$	$(n-1)(b-1)a$
$\bar{Y}_{i_1j+} - \bar{Y}_{i_2j+}$	$\frac{2}{n}(\sigma^2 + \sigma_s^2)$	$\frac{2}{n}(\hat{\sigma}^2 + \hat{\sigma}_s^2)$	messy
$\bar{Y}_{ij_1+} - \bar{Y}_{ij_2+}$	$\frac{2}{n}\sigma^2$	$\frac{2}{n}MS[E]$	$(n-1)(b-1)a$
$\bar{Y}_{i_1j_1+} - \bar{Y}_{i_2j_2+}$	$\frac{2}{n}(\sigma^2 + \sigma_s^2)$	$\frac{2}{n}(\hat{\sigma}^2 + \hat{\sigma}_s^2)$	messy

To analyze data from a CRSPD in SAS, consider using PROC MIXED instead of PROC GLM:

```
proc mixed method=type3;
  class field pest trt irr cv;
  model y=trt|pest/ddfm=satterth;
  random field(pest);
  *parms /nobound;
  lsmeans trt pest/pdiff; /* can use adj=bon; to adjust for multiplicity */
  *lsmeans trt|pest/pdiff; /* if there were interaction */
run;
```

```
/* parms statement can be used to keep SAS from dropping random
effects w/ negative estimated varcomps */
```

The SAS System
The Mixed Procedure
Model Information

1

Data Set	WORK.ONE
Dependent Variable	y
Covariance Structure	Variance Components
Estimation Method	Type 3
Residual Variance Method	Factor
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Satterthwaite

Class Level Information

Class	Levels	Values
field	2	1 2
pest	3	1 2 3
irr	2	1 2
cv	2	1 2
trt	4	1 2 3 4

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square
trt	3	180.697917	60.232639	Var(Residual) + Q(trt,pest*trt)
pest	2	256.275833	128.137917	Var(Residual) + 4 Var(field(pest)) + Q(pest,pest*trt)
pest*trt	6	24.490833	4.081806	Var(Residual) + Q(pest*trt)
field(pest)	3	97.806250	32.602083	Var(Residual) + 4 Var(field(pest))
Residual	9	29.058750	3.228750	Var(Residual)

Type 3 Analysis of Variance

Source	Error Term	Error DF	F Value	Pr > F
trt	MS(Residual)	9	18.66	0.0003
pest	MS(field(pest))	3	3.93	0.1452
pest*trt	MS(Residual)	9	1.26	0.3607
field(pest)	MS(Residual)	9	10.10	0.0031

Covariance Parameter Estimates

Cov Parm	Estimate
field(pest)	7.3433
Residual	3.2287

Type 3 Tests of Fixed Effects

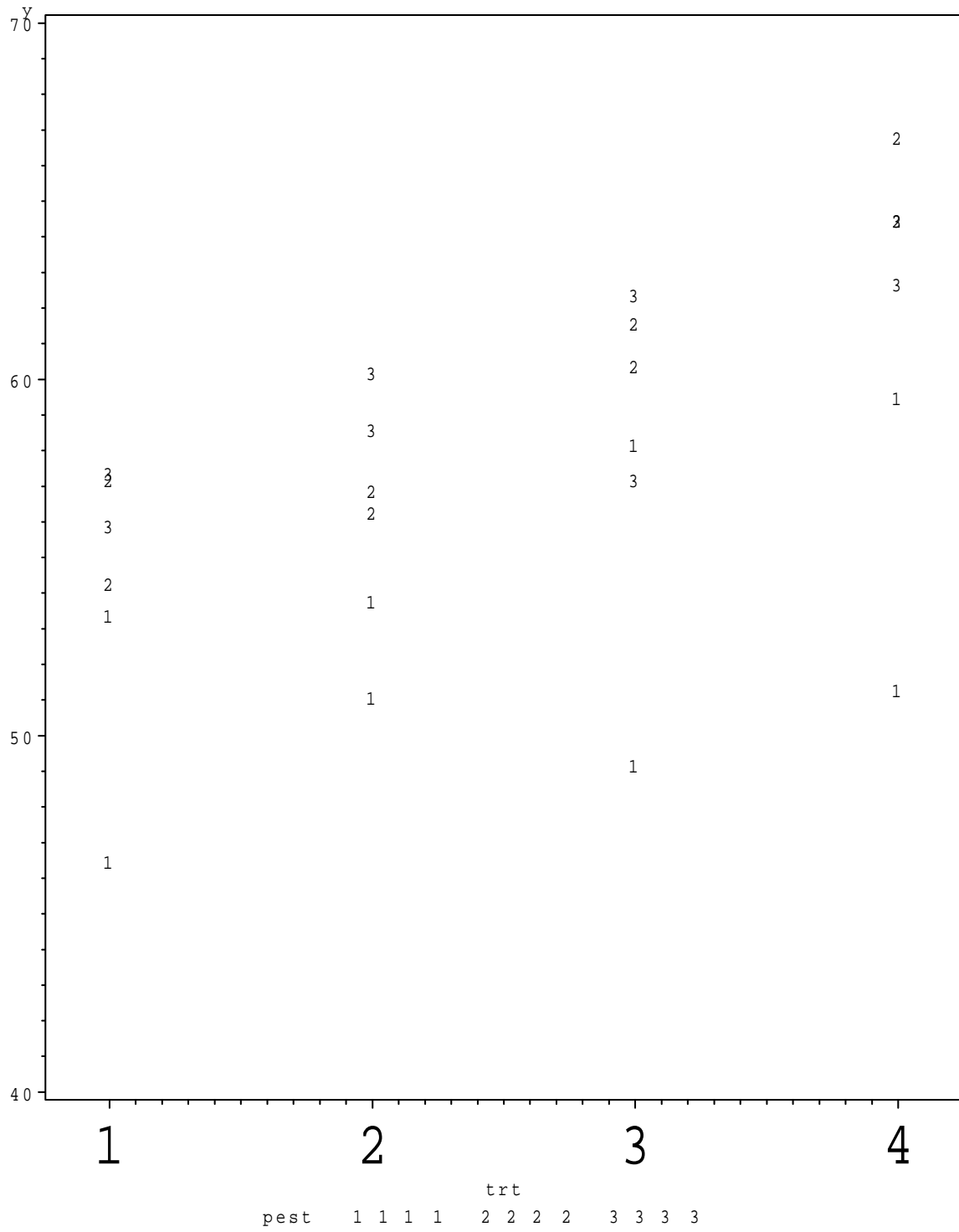
Effect	Num DF	Den DF	F Value	Pr > F
trt	3	9	18.66	0.0003
pest	2	3	3.93	0.1452
pest*trt	6	9	1.26	0.3607

Least Squares Means

Effect	pest	trt	Estimate	Standard Error	DF	t Value	Pr > t
trt		1	54.1167	1.3274	4.9	40.77	<.0001
trt		2	56.1500	1.3274	4.9	42.30	<.0001
trt		3	58.1667	1.3274	4.9	43.82	<.0001
trt		4	61.5500	1.3274	4.9	46.37	<.0001
pest	1		52.8750	2.0187	3	26.19	0.0001
pest	2		59.7500	2.0187	3	29.60	<.0001
pest	3		59.8625	2.0187	3	29.65	<.0001

Differences of Least Squares Means

Effect	pest	trt	_pest	_trt	Estimate	Standard Error	DF	t Value	Pr > t
trt		1		2	-2.0333	1.0374	9	-1.96	0.0816
trt		1		3	-4.0500	1.0374	9	-3.90	0.0036
trt		1		4	-7.4333	1.0374	9	-7.17	<.0001
trt		2		3	-2.0167	1.0374	9	-1.94	0.0838
trt		2		4	-5.4000	1.0374	9	-5.21	0.0006
trt		3		4	-3.3833	1.0374	9	-3.26	0.0098
pest	1			2	-6.8750	2.8549	3	-2.41	0.0952
pest	1			3	-6.9875	2.8549	3	-2.45	0.0919
pest	2			3	-0.1125	2.8549	3	-0.04	0.9710

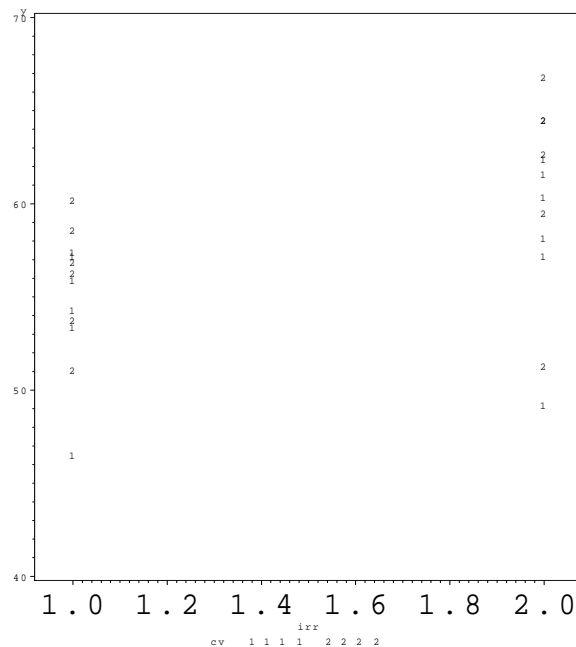


Corn yield, irrigation, pesticide and cultivars continued

So the treatment effect B is highly significant ($p = 0.0003$). Are there particular comparisons among the three treatments that are of interest? There are because real experiment is actually slightly more complicated than previously described. The factor B is really a 2×2 combination of irrigation and cultivar:

B	Irr	CV
1	no	1
2	no	2
3	yes	1
4	yes	2

The 3 df for the within plot factor B can be broken up into three 1 df components due to main effect of irr, main effect of CV and interaction. Same with the AB interaction. The plot below averages over pesticide and field:



```

proc mixed method=type3;
  class field pest irr cv trt;
  *model y=trt|pest/ddfm=satterth;
  model y=irr|cv|pest/ddfm=satterth;
  random field(pest);
  *parms /nobound;
  *lsmeans trt pest/pdiff adj=tukey;
  lsmeans irr cv /pdiff;
  lsmeans irr*cv;
run;

```

The SAS System
The Mixed Procedure
Class Level Information

1

Class	Levels	Values
field	2	1 2
pest	3	1 2 3
irr	2	1 2
cv	2	1 2
trt	4	1 2 3 4

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square
irr	1	133.953750	133.953750	Var(Residual) + Q(irr, irr*cv, pest*irr, pest*irr*cv)
cv	1	44.010417	44.010417	Var(Residual) + Q(cv, irr*cv, pest*cv, pest*irr*cv)
irr*cv	1	2.733750	2.733750	Var(Residual) + Q(irr*cv, pest*irr*cv)
pest	2	256.275833	128.137917	Var(Residual) + 4 Var(field(pest)) + Q(pest, pest*irr, pest*cv, pest*irr*cv)
pest*irr	2	17.747500	8.873750	Var(Residual) + Q(pest*irr, pest*irr*cv)
pest*cv	2	1.385833	0.692917	Var(Residual) + Q(pest*cv, pest*irr*cv)
pest*irr*cv	2	5.357500	2.678750	Var(Residual) + Q(pest*irr*cv)
field(pest)	3	97.806250	32.602083	Var(Residual) + 4 Var(field(pest))
Residual	9	29.058750	3.228750	Var(Residual)

Type 3 Analysis of Variance

Source	Error Term	Error DF	F Value	Pr > F
irr	MS(Residual)	9	41.49	0.0001
cv	MS(Residual)	9	13.63	0.0050
irr*cv	MS(Residual)	9	0.85	0.3815
pest	MS(field(pest))	3	3.93	0.1452
pest*irr	MS(Residual)	9	2.75	0.1171
pest*cv	MS(Residual)	9	0.21	0.8109
pest*irr*cv	MS(Residual)	9	0.83	0.4670
field(pest)	MS(Residual)	9	10.10	0.0031
Residual

Covariance Parameter
Estimates

Cov Parm	Estimate
field(pest)	7.3433
Residual	3.2287

Least Squares Means

Effect	irr	cv	Estimate	Standard Error	DF	t Value	Pr > t
irr	1		55.1333	1.2219	3.61	45.12	<.0001
irr	2		59.8583	1.2219	3.61	48.99	<.0001
cv		1	56.1417	1.2219	3.61	45.95	<.0001
cv		2	58.8500	1.2219	3.61	48.16	<.0001
irr*cv	1	1	54.1167	1.3274	4.9	40.77	<.0001
irr*cv	1	2	56.1500	1.3274	4.9	42.30	<.0001
irr*cv	2	1	58.1667	1.3274	4.9	43.82	<.0001
irr*cv	2	2	61.5500	1.3274	4.9	46.37	<.0001

Differences of Least Squares Means

Effect	irr	cv	_irr	_cv	Estimate	Standard Error	DF	t Value	Pr > t
irr	1		2		-4.7250	0.7336	9	-6.44	0.0001
cv		1		2	-2.7083	0.7336	9	-3.69	0.0050

Split-plot in blocks (RCBSPD)

In the randomized block split-plot design, sets of homogeneous plots are formed and levels of the whole plot factor are assigned to the plots within these sets in a restricted randomization. Assignment of levels of the split-plot factor are as in the CRSPD.

In the split-plot experiment with pesticide as the whole plot factor and irrigation \times CV as the split-plot factor, suppose the six plots come from two farms, with three plots in each farm. Suppose that the three pesticide treatments are randomized to plots within farms. Renumbering plots (1,2,1,2,1,2) as (1,2,3,4,5,6) and supposing plots (2,3,6) come from farm 1 and plots (1,4,5) from farm 2, the data are given as

The SAS System								1
Obs	farm	pest	plot	B1	B2	B3	B4	
1	2	1	1	53.4	53.8	58.2	59.5	
2	1	1	2	46.5	51.1	49.2	51.3	
3	1	2	3	54.3	56.3	60.4	64.5	
4	2	2	4	57.2	56.9	61.6	66.8	
5	2	3	5	55.9	58.6	62.4	64.5	
6	1	3	6	57.4	60.2	57.2	62.7	

At the whole plot level (ignoring the split-plot factor), the df in an ANOVA for pesticide effects are given by

Source	df
A : Pesticide	2
Farms	
Error	
Total	5

so that an F -ratio for the pesticide effect is based on $df = 2, 2$

In general, for a RCBSPD with a levels of a whole-plot level (A) randomized to r blocks (for a total of ra plots) and b levels of a split-plot factor (B) within each plot, the model and ANOVA table are given by

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + R_k + \beta_j + (\alpha\beta)_{ij} + (SR)_{ik} + E_{ijk} \\ &= \mu_{ij} + R_k + (SR)_{ik} + E_{ijk} \end{aligned}$$

where

- i denotes level of A ,
- j denotes level of B ,
- k denotes block.

$R_k \stackrel{iid}{\sim} N(0, \sigma_r^2)$ and $SR_{ik} \stackrel{iid}{\sim} N(0, \sigma_{sr}^2)$. All random errors are mutually independent.

Source	df	EMS
A	$a - 1$	$\sigma^2 + b\sigma_{sr}^2 + br\psi_A^2$
Blocks	$r - 1$	$\sigma^2 + b\sigma_{sr}^2 + ab\sigma_r^2$
Whole plot error (Block \times A)	$(r - 1)(a - 1)$	$\sigma^2 + b\sigma_{sr}^2$
B	$b - 1$	$\sigma^2 + ar\psi_B^2$
AB	$(a - 1)(b - 1)$	$\sigma^2 + r\psi_{AB}^2$
Error ($B \times$ Blocks(A))	$a(b - 1)(r - 1)$	σ^2
Total	$abr - 1$	

```

data one; /* 'one' is original dataset */
  set one;
  if plot in (2,3,6) then farm=1;
  else farm=2;
run;
proc mixed method=type3;
  class farm plot pest irr cv trt;
  model y=pest|trt;
  random farm farm*pest;
run;

```

The Mixed Procedure

Class	Levels	Values
farm	2	1 2
plot	6	1 2 3 4 5 6
pest	3	1 2 3
trt	4	1 2 3 4

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square
pest	2	256.275833	128.137917	Var(Residual) + 4 Var(farm*pest) + Q(pest,pest*trt)
trt	3	180.697917	60.232639	Var(Residual) + Q(trt,pest*trt)
pest*trt	6	24.490833	4.081806	Var(Residual) + Q(pest*trt)
farm	1	59.220417	59.220417	Var(Residual) + 4 Var(farm*pest) + 12 Var(farm)
farm*pest	2	38.585833	19.292917	Var(Residual) + 4 Var(farm*pest)
Residual	9	29.058750	3.228750	Var(Residual)

Cov Parm	Estimate
farm	3.3273
farm*pest	4.0160
Residual	3.2287

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
pest	2	2	6.64	0.1309
trt	3	9	18.66	0.0003
pest*trt	6	9	1.26	0.3607

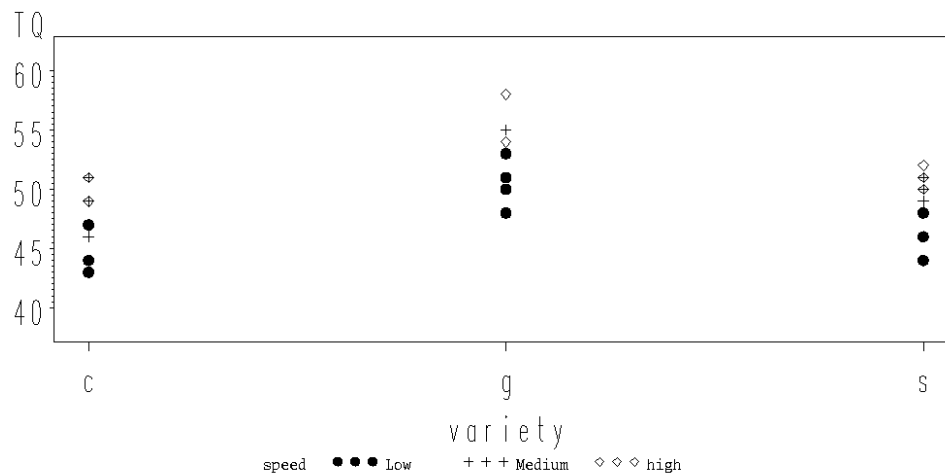
Split-plot in blocks (RCBSPD)

Researchers for an ice cream manufacturer conduct an expt. to study the effects of variety (sweetcharlie, camarosa, and gaviota) and mixing speed (slow, medium and fast) on ice cream quality. One batch of each variety of strawberries is sampled on Monday over four consecutive weeks. Each batch is divided into three parts, which are randomized to the three mixing speeds and three quarts of iced cream are produced, stored for one month, then tested for texture quality, on a scale from 1-100.

Data:

Obs	week	variety	Low	Medium	High
1	1	c	47	49	49
2	1	g	48	51	54
3	1	s	46	49	51
4	2	c	43	46	49
5	2	g	51	55	53
6	2	s	48	49	48
7	3	c	47	51	51
8	3	g	50	53	53
9	3	s	44	50	50
10	4	c	44	49	51
11	4	g	53	53	58
12	4	s	48	51	52

Ice cream texture quality



Model

$$Y_{ijk} = \mu_{ij} + B_k + (\alpha B)_{ik} + E_{ijk}$$

where

$$i = 1, 2, 3 = a(\text{ variety })$$

$$j = 1, 2, 3 = b(\text{ speed })$$

$$k = 1, 2, 3, 4 = r(\text{ week })$$

ANOVA sketch

Source	<i>df</i>	Expected MS
Variety	$a - 1 = 2$	
Block	$r - 1 = 3$	
V × B	$(a - 1)(r - 1) = 6$	
Speed	$b - 1 = 2$	
V × S	$(a - 1)(b - 1) = 4$	
Error	$(b - 1)(r - 1)a = 18$	
Total	$abr - 1 = 35$	

```
/* SAS code for split-plot in blocks */

data one;
  infile "strawberryice.dat" firstobs=3;
  input week variety $ speed $ tq;
run;
*goptions dev=ps;
axis1 offset=(1 cm,1 cm) label=(height=2 "variety")
  value=(height=2);
axis2 offset=(1 cm,1 cm) label=(height=2 "TQ")
  value=(height=2);

symbol1 value=dot c=black h=1.2;
symbol2 value=plus c=black h=1.2;
symbol3 value=diamond c=black h=1.2;

proc gplot data=one;
  title "Ice cream texture quality";
  plot tq*variety=speed/haxis=axis1 vaxis=axis2;
  *plot tq*speed=variety;
run; quit;
proc mixed data=one method=type3;
  class week variety speed;
  model tq=variety|speed;
  random week week*variety;
  lsmeans variety speed/diff;
  lsmeans variety*speed;
run;
```

The SAS System
The Mixed Procedure
Model Information

1

Data Set WORK.ONE
Dependent Variable tq
Covariance Structure Variance Components
Estimation Method Type 3
Residual Variance Method Factor
Fixed Effects SE Method Model-Based
Degrees of Freedom Method Containment

Class	Levels	Values
week	4	1 2 3 4
variety	3	c g s
speed	3	Low Medium high

Dimensions

Covariance Parameters	3
Columns in X	16
Columns in Z	16
Subjects	1
Max Obs Per Subject	36

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square
variety	2	148.666667	74.333333
speed	2	112.166667	56.083333
variety*speed	4	2.166667	0.541667
week	3	19.222222	6.407407
week*variety	6	33.111111	5.518519
Residual	18	37.666667	2.092593

Type 3 Analysis of Variance

Source	Expected Mean Square	Error Term
variety	Var(Residual) + 3 Var(week*variety) + Q(variety,variety*speed)	MS(week*variety)
speed	Var(Residual) + Q(speed,variety*speed)	MS(Residual)
variety*speed	Var(Residual) + Q(variety*speed)	MS(Residual)
week	Var(Residual) + 3 Var(week*variety) + 9 Var(week)	MS(week*variety)
week*variety	Var(Residual) + 3 Var(week*variety)	MS(Residual)
Residual	Var(Residual)	.

Type 3 Analysis of Variance

Source	Error DF	F Value	Pr > F
variety	6	13.47	0.0060
speed	18	26.80	<.0001
variety*speed	18	0.26	0.9004
week	6	1.16	0.3990
week*variety	18	2.64	0.0516
Residual	.	.	.

Covariance Parameter
Estimates

Cov Parm	Estimate
week	0.09877
week*variety	1.1420
Residual	2.0926

Fit Statistics

-2 Res Log Likelihood	118.2
AIC (smaller is better)	124.2
AICC (smaller is better)	125.3
BIC (smaller is better)	122.4

Least Squares Means

Effect	variety	speed	Estimate	Standard Error	DF	t Value	Pr > t
variety	c		48.0000	0.6961	6	68.95	<.0001
variety	g		52.6667	0.6961	6	75.66	<.0001
variety	s		48.8333	0.6961	6	70.15	<.0001
speed		Low	47.4167	0.5424	18	87.41	<.0001
speed		Medium	50.5000	0.5424	18	93.10	<.0001
speed		high	51.5833	0.5424	18	95.10	<.0001
variety*speed	c	Low	45.2500	0.9129	18	49.57	<.0001
variety*speed	c	Medium	48.7500	0.9129	18	53.40	<.0001
variety*speed	c	high	50.0000	0.9129	18	54.77	<.0001
variety*speed	g	Low	50.5000	0.9129	18	55.32	<.0001
variety*speed	g	Medium	53.0000	0.9129	18	58.06	<.0001
variety*speed	g	high	54.5000	0.9129	18	59.70	<.0001
variety*speed	s	Low	46.5000	0.9129	18	50.94	<.0001
variety*speed	s	Medium	49.7500	0.9129	18	54.50	<.0001
variety*speed	s	high	50.2500	0.9129	18	55.05	<.0001

Differences of Least Squares Means

Effect	variety	speed	_variety	_speed	Estimate	Standard Error	DF
variety	c		g		-4.6667	0.9590	6
variety	c		s		-0.8333	0.9590	6
variety	g		s		3.8333	0.9590	6
speed		Low		Medium	-3.0833	0.5906	18
speed		Low		high	-4.1667	0.5906	18
speed		Medium		high	-1.0833	0.5906	18

Differences of Least Squares Means

Effect	variety	speed	_variety	_speed	t Value	Pr > t
variety	c		g		-4.87	0.0028
variety	c		s		-0.87	0.4183
variety	g		s		4.00	0.0071
speed		Low		Medium	-5.22	<.0001
speed		Low		high	-7.06	<.0001
speed		Medium		high	-1.83	0.0832