

REGRESSION AND ANALYSIS OF VARIANCE FOR GENETICISTS



REGRESSION

(image from <http://www.geneart.org/genome-title.htm>)

MATRICES

1. Rectangular array of numbers

$$\begin{bmatrix} 3 & 5 & 7 & 8 \\ 1 & 2 & 3 & 7 \end{bmatrix} = \mathbf{A}_{2 \times 4}$$

Symbol - cap letter **A, B, C**

2. Vectors

3. Elements a_{ij}

4. Operations

(a) Addition or subtraction

Element by element

(b) Multiplication

$$\text{Vector } (1, 3, -5, 1) \begin{bmatrix} 2 \\ 0 \\ 3 \\ -2 \end{bmatrix} = 2 + 0 - 15 - 2 \\ = -15$$

Matrices

$$\begin{bmatrix} 1 & 3 & 5 \\ -2 & -1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 0 & 5 \\ 1 & -2 \end{bmatrix} = \begin{bmatrix} 7 & 8 \\ -2 & -15 \end{bmatrix}$$

Note: In general, **BA** does not equal **AB**
(may have different dimensions)

(c) Transpose A'

Note: $(AB)' = (B')(A')$

(d) Scalar Multiplication

$$3 \begin{bmatrix} 1 & 2 \\ -2 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ -6 & 9 \end{bmatrix}$$

5. Identity Matrix **I**

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{array}{l} \mathbf{IA} = \mathbf{AI} = \mathbf{A} \text{ for } \mathbf{A}_{3 \times 3} \\ \mathbf{IB} = \mathbf{B} \text{ for } \mathbf{B}_{3 \times C} \\ \mathbf{CI} = \mathbf{C} \text{ for } \mathbf{C}_{r \times 3} \end{array}$$

6. Rank and dependence

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 3 & 1 & 5 \\ 2 & 3 & 1 \end{bmatrix} \text{ columns} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}, \text{ and } \begin{bmatrix} 1 \\ 5 \\ 1 \end{bmatrix}$$

Note: $2\mathbf{C}_1 - \mathbf{C}_2 - \mathbf{C}_3 = \mathbf{\Phi}$ where $\mathbf{\Phi}$ is a column of 0's.

We say that \mathbf{C}_1 , \mathbf{C}_2 , and \mathbf{C}_3 are linearly dependent.

In general: k columns $\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_k$ are dependent if there exist scalars $\lambda_1, \lambda_2, \dots, \lambda_k$ such that

$$(1) \lambda_1 \mathbf{C}_1 + \lambda_2 \mathbf{C}_2 + \dots + \lambda_k \mathbf{C}_k = \mathbf{\Phi}$$

and

$$(2) \text{ At least one of the } \lambda\text{'s is } \underline{\text{not}} \text{ 0.}$$

If the k columns are not dependent we call them linearly independent. The combination of columns in (1) is called a linear combination, a phrase we will often use. Thus, k columns are linearly independent if the only linear combination of them which will produce the zero vector is the linear combination with all λ 's 0. Often we will collect the λ 's together in a vector $\mathbf{\Lambda}$.

The rank of a matrix is the maximum number of linearly independent columns which can be selected from the columns of the matrix. Thus the rank of \mathbf{A} is two. Notice that if the rank

of a matrix is 1, then there is one column such that all other columns are direct multiples.

For any matrix \mathbf{X} , the rank of \mathbf{X} is the same as the rank of $\mathbf{X}'\mathbf{X}$. The row rank of any matrix is always equal to the column rank.

7. Inverse of a matrix.

Symbol: \mathbf{A}^{-1}

The inverse of an $n \times n$ matrix \mathbf{A} is an $n \times n$ matrix \mathbf{B} such that $\mathbf{AB} = \mathbf{I}$. Such a matrix \mathbf{B} will exist only if \mathbf{A} is of rank n . In this case it is also true that $\mathbf{BA} = \mathbf{I}$.

Example:

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & 0 \\ 3 & 1 & 1 \end{bmatrix} \quad \mathbf{A}^{-1} = \begin{bmatrix} +0.5 & 1 & -0.5 \\ -1.0 & -1 & 1.0 \\ -0.5 & -2 & 1.5 \end{bmatrix}$$

(Check by multiplication.)

Solving equations

$$\begin{aligned} b_0 - b_1 + b_2 &= 2 \\ 2b_0 + b_1 &= 7 \\ 3b_0 + b_1 + b_2 &= -5 \end{aligned}$$

$$\begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & 0 \\ 3 & 1 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ -5 \end{bmatrix}$$

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} +0.5 & 1 & -0.5 \\ -1.0 & -1 & 1.0 \\ -0.5 & -2 & 1.5 \end{bmatrix} \begin{bmatrix} 2 \\ 7 \\ -5 \end{bmatrix} = \begin{bmatrix} 10.5 \\ -14.0 \\ -22.5 \end{bmatrix}$$

$$\mathbf{A}\mathbf{b} = \mathbf{c} \Rightarrow \mathbf{b} = \mathbf{A}^{-1}\mathbf{c}$$

For a 2×2 matrix only we have the formula:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Note: if \mathbf{A} and \mathbf{B} are square matrices then $\mathbf{A}^{-1}\mathbf{B}^{-1} = (\mathbf{BA})^{-1}$

RANDOM VECTORS

We have described a univariate random variable like weight W by writing

$$W \sim N(150, 100)$$

We might also measure height in inches H and have

$$H \sim N(68, 16)$$

Now the above tells us nothing about how the two variables height and weight covary. The covariance between H and W tells us how weight depends on height. If we know an individual is taller than the mean 68, would we predict that his weight will exceed the mean 150? If so we are claiming a positive covariance between height and weight. Formally, recall that the variance of weights is defined as an expected value, namely

$$\text{variance}(W) = E\left\{ (W - 150)^2 \right\}.$$

The covariance between W and H is defined as

$$\text{cov}(W, H) = E\left\{ (W - 150)(H - 68) \right\}.$$

Suppose the covariance is $\text{cov}(W, H) = 30$. We put this all together as

$$\begin{bmatrix} W \\ H \end{bmatrix} \sim \text{MVN} \left[\begin{bmatrix} 150 \\ 68 \end{bmatrix}, \begin{bmatrix} 100 & 30 \\ 30 & 16 \end{bmatrix} \right]$$

In general we write

$$\mathbf{Y} \sim \text{MVN}(\boldsymbol{\mu}, \mathbf{V})$$

where \mathbf{Y} is a vector with i^{th} element Y_i , $\boldsymbol{\mu}$ is a vector with i^{th} element μ_i and \mathbf{V} is a matrix whose ij^{th} element is the covariance between Y_i and Y_j .

***** Fact: If \mathbf{A} and \mathbf{B} are matrices of constants and \mathbf{Y} is a random vector with

$$\mathbf{Y} \sim \text{MVN}(\boldsymbol{\mu}, \mathbf{V})$$

then

$$\mathbf{A}\mathbf{Y} + \mathbf{B} \sim \text{MVN}(\mathbf{A}\boldsymbol{\mu} + \mathbf{B}, \mathbf{A}\mathbf{V}\mathbf{A}')$$

Example:

$$\mathbf{Y} \sim \text{MVN} \left[\begin{bmatrix} 4 \\ 6 \\ 10 \end{bmatrix}, \begin{bmatrix} 8 & 5 & 0 \\ 5 & 12 & 4 \\ 0 & 4 & 9 \end{bmatrix} \right]$$

(1) Find the distribution of $Z = Y_1 - Y_2 + Y_3$.

(2) Let $W = Y_1 - 3Y_2 + 2Y_3$. Find the joint distribution of Z and W .

Flower example

Multivariate Normal Computations

I grow flowers with certain characteristics:

$Y_1 = \text{Head diameter} \sim N(4,8)$

$Y_2 = \text{Stem length} \sim N(6,12)$

$Y_3 = \text{Root length} \sim N(10,9)$

What about covariances? To fill in the picture, here is the complete distribution:

$$\text{vector } Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \sim \text{MVN} \left(\begin{pmatrix} 4 \\ 6 \\ 10 \end{pmatrix}, \begin{pmatrix} 8 & 5 & 0 \\ 5 & 12 & 4 \\ 0 & 4 & 9 \end{pmatrix} \right)$$

Now 2 people compete to buy these flowers from me. They offer prices for each flower related to its dimensions. The prices, Z_1 and Z_2 , offered by the two buyers are:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} 1 Y_1 - 1 Y_2 + 1 Y_3 \\ 1 Y_1 - 3 Y_2 + 2 Y_3 \end{pmatrix}$$

Which is the better offer in terms of mean?

Which has the smaller variance? How often will Z_2 exceed Z_1 ?

All of these questions and more can be answered if we know the joint distribution of (Z_1, Z_2) . Calling this vector \mathbf{Z} , we want to relate it to the vector $(Y_1, Y_2, Y_3)'$ by expressing it as

$$\mathbf{Z} = \mathbf{A} \mathbf{Y} + \mathbf{b}$$

as in our notes. Multiply this out to check that it

works:

$$\text{vector } Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -3 & 2 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

This equation tells us, for example, what prices would be offered for a flower with $Y_1 = 5$, $Y_2 = 4$, and $Y_3 = 14$.

We find

$$\text{vector } Z = Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -3 & 2 \end{pmatrix} \begin{pmatrix} 5 \\ 4 \\ 14 \end{pmatrix} = \begin{pmatrix} 15 \\ 21 \end{pmatrix}$$

So for this flower with its longer than average roots and shorter than average stem, we would be better off going with the second person.

What happens on average?

Our formula for the mean of vector Z tells us that

$$\text{mean of } Z = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -3 & 2 \end{pmatrix} \begin{pmatrix} 4 \\ 6 \\ 10 \end{pmatrix} = \begin{pmatrix} 8 \\ 6 \end{pmatrix}$$

We are better off going with the first buyer who will give us 8 cents on average. Which buyer has the smaller variance in price? What is the covariance between the prices? To answer this, we compute the variance-covariance matrix of the price vector Z from the variance-covariance matrix V that goes with the Y vector (we compute AVA'):

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & -3 & 2 \end{pmatrix} \begin{pmatrix} 8 & 5 & 0 \\ 5 & 12 & 4 \\ 0 & 4 & 9 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & -3 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 11 & 22 \\ 22 & 74 \end{pmatrix}$$

Not only is the first buyer offering more on average, his price will vary less than that of the second buyer given the kind of variation I have in my population of flowers. I also see that the two prices have a correlation $.7711 = 22/\sqrt{(11*74)}$.

Finally, what is the distribution of the price difference $D = Z_2 - Z_1$ and what is the probability that D will be positive, that is, what proportion of the time will the flower dimensions be such that I regret my decision to deal with the first buyer?

We see that $D = (-1 \ 1)(Z_1, Z_2)'$ so the mean of D is $(-1 \ 1)(8 \ 6)' = -2$ and the variance is

$$(-1 \ 1) \begin{pmatrix} 11 & 22 \\ 22 & 74 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = 41$$

So the probability that D is greater than 0 is the probability that a standard normal Z variable exceeds $(0 - (-2))/6.4031 = 0.31$. Note that since the variance is 41, then 6.4031 is the standard deviation (the square root of 41). From the normal table

$$\Pr\{Z > 0.31\} = .3783$$

so we regret our decision about 38% of the time.

One more note: Usually we do not KNOW these parameters. From a sample, we would estimate them.

REGRESSION - HAND CALCULATION

Review of "hand" calculations:

DATA X 4 7 5 4
Y 5 7 6 6

7							*
6			*	*			
5			*				
4							
3							
2							
1	2	3	4	5	6	7	

$\bar{X} = 5, \bar{Y} = 6$

$X - \bar{X}$	$(X - \bar{X})^2$	$Y - \bar{Y}$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
-1	1	-1	1	1
2	4	1	1	2
0	0	0	0	0
-1	1	0	0	0
===	===	===	===	===
0	6	0	2	3
	$\Sigma(X_i - \bar{X})^2$		$\Sigma(Y_i - \bar{Y})^2$	$\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$

slope = $b = \Sigma(X_i - \bar{X})(Y_i - \bar{Y}) / \Sigma(X_i - \bar{X})^2 = 3/6 = 0.5$

intercept = $a = \bar{Y} - b\bar{X} = 6 - (0.5)(5) = 3.5$

True slope β . Recall that $b \sim N(\beta, \sigma^2 / (\Sigma(X_i - \bar{X})^2))$.

True intercept α .

Recall that $a \sim N\left(\alpha, \sigma^2 \left[1/n + \bar{X}^2 / \Sigma(X_i - \bar{X})^2\right]\right)$

Estimate σ^2 by $s^2 = (\text{total SSq} - \text{regn SSq})/\text{df}$

$$\text{Regn. SSq} = [\Sigma(X_i - \bar{X})(Y_i - \bar{Y})]^2 / \Sigma(X_i - \bar{X})^2 = 3 * 3/6 = 1.5$$

$$\text{Total SSq} = 2$$

$$\text{df} = n - 2 = 4 - 2 = 2$$

$$\text{so } s^2 = \text{MSE} = (2 - 1.5)/2 = 0.25$$

Estimated variance of a is $0.25(1/4 + 25/6) = 1.1042$

Estimated variance of b is $0.25/6 = 0.041667$

Notice that hand computations gave no indication of covariance.

EXAMPLES:

Test for no relationship between Y and X.

$$H_0: \beta=0, \quad t = \frac{b-0}{\text{std. err. of } b} = .5 / \sqrt{.041667}$$

Give 95% confidence interval for mean Y at X = 6.

$$3.5 + .5 (6) = 6.5 = \text{prediction}$$

add and subtract t times std. err. of prediction

$$\text{std. err. of prediction} = \sqrt{[1/n + (6 - \bar{X})^2 / \Sigma(X_i - \bar{X})^2] 0.25}$$

Give 95% prediction interval for individual Y at X = 6.

same prediction, 6.5

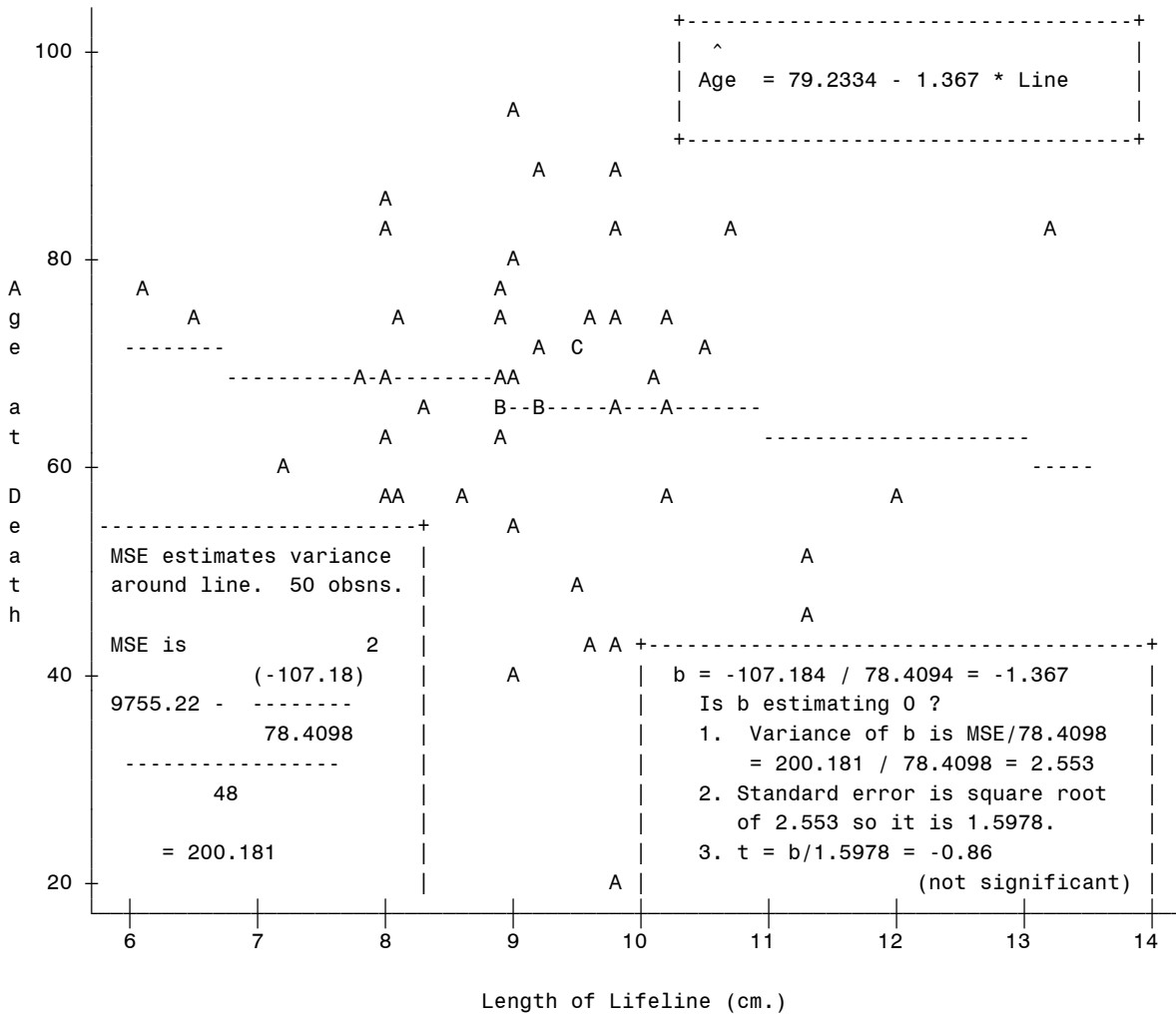
$$\text{std. err. of prediction} = \sqrt{[1 + 1/n + (6 - \bar{X})^2 / \Sigma(X_i - \bar{X})^2] 0.25}$$

Example of a regression:

Wilson and Mather, JAMA 229 (1994)						
LINE	X	XSQ	XY	AGE	Y	YSQ
9.75	0.552	0.3047	-26.3083	19	-47.66	2271.48
9.00	-0.198	0.0392	5.2787	40	-26.66	710.76
9.75	0.552	0.3047	11.780	88	21.34	455.40
9.00	-0.198	0.0392	-5.413	94	27.34	747.48
=====	=====	=====	=====	=====	=====	=====
	0	78.4098	-107.18		0.00	9755.22

Wilson and Mather, JAMA 229 (1994)

Plot of AGE*LINE. Legend: A = 1 obs, B = 2 obs, etc.
 Plot of YHAT*LINE. Symbol used is '-'.



Discussion of correlation coefficients:

We define the (population) correlation between two variables as the covariance divided by the square root of the product of the variances. For height and weight as just given, $\rho = 30/\sqrt{100*16} = 0.75$. Now suppose we take $n = 103$ people and measure their height H_i and weight W_i . As usual, we can estimate variances as

$$S_W^2 = \Sigma(W_i - \bar{W})^2/(n - 1) \text{ and } S_H^2 = \Sigma(H_i - \bar{H})^2/(n - 1).$$

The covariance is estimated as

$$S_{WH} = \Sigma(W_i - \bar{W})(H_i - \bar{H})/(n - 1).$$

Now these are just estimates of the true values so let us assume we get, say,

$$S_W^2 = 125, S_H^2 = 20, \text{ and } S_{WH} = 40.$$

Our estimated covariance matrix is then

$$\hat{V} = \begin{bmatrix} 125 & 40 \\ 40 & 20 \end{bmatrix}$$

and we compute a sample estimate, r , of ρ as

$$r = 40/\sqrt{125*20} = 0.8.$$

Notice that

$$r^2 = \frac{[\sum(W_i - \bar{W})(H_i - \bar{H})]^2}{\sum(W_i - \bar{W})^2 \sum(H_i - \bar{H})^2} = \text{regression SS/total SS}$$

from a regression of either H on W or of W on H. Thus we have explained 64% of the variability in W by regressing it on H.

A test that $\rho = 0$ is just the t-test on the coefficient in the regression of W on H (or equivalently H on W). To test any other hypothesis like $H_0: \rho = 0.5$ or to put a confidence interval around r, Fisher's transformation to Z is used. We define

$$Z_r = 0.5 \ln((1 + r)/(1 - r))$$

and define Z_ρ similarly. Now approximately we have

$$Z_r \sim N(Z_\rho, 1/(n - 3)).$$

Thus we get $Z_{0.8} = 1.09861$ so

$$1.09861 - 0.196$$

and

$$1.09861 + 0.196$$

are the lower and upper confidence bounds for Z_ρ .

Converting from Z_ρ to ρ we get $0.71 \leq \rho \leq 0.86$. Tables can be used to do the conversions or simply do them on a hand calculator.

REGRESSION IN MATRIX FRAMEWORK

The equations: (Henceforth intercept is β_0 , slope is β_1)

$$\begin{aligned} 5 &= \beta_0 + 4\beta_1 + e_1 \\ 7 &= \beta_0 + 7\beta_1 + e_2 \\ 6 &= \beta_0 + 5\beta_1 + e_3 \\ 6 &= \beta_0 + 4\beta_1 + e_4 \end{aligned} \quad \Rightarrow \quad Y_i = \beta_0 + \beta_1 X_i + e_i; \quad i = 1, 2, 3, 4$$

Matrix form:

$$\begin{bmatrix} 5 \\ 7 \\ 6 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 1 & 7 \\ 1 & 5 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} \quad \Rightarrow \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Estimate $\boldsymbol{\beta}$ by \mathbf{b} such that $(\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})$ is minimized.

Note: this is the sum of squares of residuals $\mathbf{Y} - \mathbf{X}\mathbf{b}$.

Using calculus we can show the sum of squares is minimized by solving the following "normal equations."

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y} \quad \text{i.e.} \quad \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

THIS FORMULA IS IMPORTANT

Our little example:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 20 \\ 20 & 106 \end{bmatrix} \Rightarrow (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{424 - 400} \begin{bmatrix} 106 & -20 \\ -20 & 4 \end{bmatrix}$$

We find the solution

$$\mathbf{b} = \begin{bmatrix} 4.4167 & -0.8333 \\ -0.8333 & 0.1667 \end{bmatrix} \begin{bmatrix} 24 \\ 123 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 0.5 \end{bmatrix}$$

Now let's relate the vector of estimated parameters (\mathbf{b}) to the vector of actual parameters (β). We have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'(\mathbf{X}\beta + \mathbf{e})) = \beta + (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{e}).$$

The difference between \mathbf{b} and β is $\mathbf{b} - \beta = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{e})$.

Results:

- (1) \mathbf{b} is an unbiased estimate of β .

(2) The variance-covariance matrix of \mathbf{b} (\mathbf{V}_b) is related to that of \mathbf{e} (\mathbf{V}_e) by the formula (using our $\mathbf{A}\mathbf{V}\mathbf{A}'$ with $\mathbf{A}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$)

$$\mathbf{V}_b = ((\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{V}_e \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})$$

(3) Let us assume that all the e's have the same variance, have mean 0, and are uncorrelated. That means

$$\mathbf{V}_e = \sigma^2 \mathbf{I}$$

and after all the smoke clears, we obtain the crucial formula

$$\mathbf{V}_b = \text{var}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1} \sigma^2$$

ESTIMATION OF VARIANCE OF e

- (1) Sum of squared residuals
(sum of squared errors, SSE)

$\mathbf{Y} - \mathbf{X}\mathbf{b}$ vector of residuals

For any vector \mathbf{a} , $\mathbf{a}'\mathbf{a}$ is sum of squared elements

$$\begin{aligned}
 \text{SSE} &= (\mathbf{Y} - \mathbf{Xb})' (\mathbf{Y} - \mathbf{Xb}) \\
 &= \begin{array}{c} \mathbf{Y}'\mathbf{Y} \\ \text{Uncorr.} \\ \text{Total SSq} \end{array} - \begin{array}{c} \mathbf{b}'\mathbf{X}'\mathbf{Y} \\ \text{Uncorr.} \\ \text{Regn. SSq} \end{array} \\
 &= \begin{array}{c} (\mathbf{Y}'\mathbf{Y} - n\bar{y}^2) \\ \text{Corrected} \\ \text{total SSq} \end{array} - \begin{array}{c} (\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{y}^2) \\ \text{Corrected} \\ \text{Regn. SSq} \end{array}
 \end{aligned}$$

(2) Error df = $n - 2$ (2 = one intercept + one slope)

$$\text{MSE} = \text{SSE}/\text{df}$$

(3) For our little example, check this against the previous computation.

(4) Error degrees of freedom will always give degrees of freedom for t statistics.

Example (Continued.)

Variance-covariance matrix of parameter estimates

$$\mathbf{V}_b = \begin{bmatrix} 4.4167 & -0.8333 \\ -0.8333 & 0.1667 \end{bmatrix} (0.25) = \begin{bmatrix} 1.1042 & -0.2083 \\ -0.2083 & 0.0417 \end{bmatrix}$$

(1) Test that slope is 0: $t = 0.5 / \sqrt{0.0417}$
 (2 degrees of freedom)
 $= 0.5 / 0.2041 = 2.45$

(2) Test that intercept is 0: $t = 3.5 / \sqrt{1.1042} = 3.33$
(2 degrees of freedom)

(3) Estimate mean value of Y at $X = 6$ and give 95% confidence interval.

Estimate is $3.5 + 0.5 * 6 = 6.5$

We are estimating $\beta_0 + 6 \beta_1 = (1, 6) \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$

Our estimate is $(1, 6)\mathbf{b} = b_0 + 6b_1 = 3.5 + (6)(0.5) = 6.5$

Letting $\Lambda' = (1, 6)$ we now want the variance of $\Lambda'\mathbf{b}$ but we know how to get that: ($\mathbf{A}\mathbf{V}\mathbf{A}'$ result where \mathbf{A} is now the row Λ')

$$\text{Var}(\Lambda'\mathbf{b}) = \Lambda' \mathbf{V}_b \Lambda = (1, 6) \begin{bmatrix} 1.1042 & -0.2083 \\ -0.2083 & 0.0417 \end{bmatrix} \begin{bmatrix} 1 \\ 6 \end{bmatrix} = 0.1042$$

95% confidence interval : $6.5 \pm t * \sqrt{0.1042}$

(4) Predict future individual value at $X = 6$

Individual will differ from mean by a deviation with variance estimated by $\text{MSE} = 0.25$. Any future individual value at $X = 6$ is equal to mean at $X = 6$ plus deviation. Thus the variance is the sum of the variances of these two parts, namely, $0.25 + 0.1042 = 0.3542$. Notice that the prediction interval

for an individual future value is wider than the corresponding confidence interval for the mean. We get, for our little example,

$$6.5 \pm t^* \sqrt{0.3542}$$

Since t has 2 d.f. we obtain $t = 4.30$ and thus

Confidence interval (5.11, 7.89)

Prediction interval (3.94, 9.06)

Example: You are in charge of predicting wheat yields from rainfall through July for your country so that you can place import quotas in early August. You have historic data on rain and yields. Which of the above formulas do you use and why?

Example: 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?)

SUMMARY OF REGRESSION FORMULAS

Model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{e} \sim \text{MVN}(0, \sigma^2 \mathbf{I})$

Normal Equations: $\mathbf{X}'\mathbf{X} \mathbf{b} = \mathbf{X}'\mathbf{Y}$

Solution: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{Y})$ provided $\mathbf{X}'\mathbf{X}$ is full rank

Estimate of variance-covariance matrix of \mathbf{b} : $(\mathbf{X}'\mathbf{X})^{-1}$ MSE

Predictions for observed Y's is: vector \mathbf{Xb} .

We write $\hat{Y} = \mathbf{Xb}$ (^ denotes predictor).

Residuals: $= \mathbf{Y} - \mathbf{Xb}$

SSE = $(\mathbf{Y} - \mathbf{Xb})' (\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} =$
total SSq – regn. SSq.

df = n minus rank of \mathbf{X} matrix.

(Usually, rank = number of columns in \mathbf{X} , i.e. full rank.)

Prediction of future Y's at some configuration of X's:

Prediction written as $\hat{Y} = \boldsymbol{\Lambda}'\mathbf{b}$ where $\boldsymbol{\Lambda}' = (1, \text{X-values})$. In our example $\boldsymbol{\Lambda}' = (1, 6)$.

Variances of predictions:

for mean at configuration of X's in Λ' :

$$(\Lambda'(\mathbf{X}'\mathbf{X})^{-1} \Lambda) \text{ MSE}$$

for individual at configuration of X's in Λ :

$$(\Lambda'(\mathbf{X}'\mathbf{X})^{-1} \Lambda + 1) \text{ MSE}$$

ANOVA (Column of 1's and k other columns.)

SOURCE	DF	SSq
model	k	$\mathbf{b}'\mathbf{X}'\mathbf{Y} - \text{CT}$ Note: CT= correction term = $n\bar{y}^2$
error	$n - k - 1$	$\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$
total	$n - 1$	$\mathbf{Y}'\mathbf{Y} - \text{CT}$

$$\begin{aligned} R^2 &= \text{coefficient of determination} \\ &= \text{corrected regn. SSq/corrected total SSq.} \\ &= 1 - \text{SSq(error)/SSq(corrected total)} \end{aligned}$$

$$F_{n-k-1}^k = \text{MS(MODEL)/MS(ERROR)}$$

$$\text{tests } H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

TYPE I (sequential) and TYPE II (partial) SUMS OF SQUARES

DATA

Y	X0	X1	X2
-2	1	-2	1
-3	1	-1	0
0	1	2	0
2	1	1	2

X'X **(X'X)⁻¹** **X'Y** **b**

REGRESS Y ON X0 ONLY

4 1/4 -3 -3/4
 (SSR = 9/4 = 2.25 = CT)

REGRESS Y ON X0, X1

$\begin{bmatrix} 4 & 0 \\ 0 & 10 \end{bmatrix}$ $\begin{bmatrix} 1/4 & 0 \\ 0 & 1/10 \end{bmatrix}$ $\begin{bmatrix} -3 \\ 9 \end{bmatrix}$ $\begin{bmatrix} -3/4 \\ 9/10 \end{bmatrix}$
 (SSR = 2.25 + 8.1 = 10.35)

REGRESS Y ON X0, X1, X2

$\begin{bmatrix} 4 & 0 & 3 \\ 0 & 10 & 0 \\ 3 & 0 & 5 \end{bmatrix}$ $\begin{bmatrix} 10/22 & 0 & -3/11 \\ 0 & 1/10 & 0 \\ -3/11 & 0 & 4/11 \end{bmatrix}$ $\begin{bmatrix} -3 \\ 9 \\ 2 \end{bmatrix}$ $\begin{bmatrix} -21/11 \\ 9/10 \\ 17/11 \end{bmatrix}$
 (SSR = 16.92)

Notes: No change in b_0 from 1st to 2nd regression (orthogonality). Two b 's change from 2nd to 3rd (not orthogonal).

Adding X_2 to regression 2 increases the regression sum of squares from 10.35 to 16.92. We write

$$R(X_0, X_1, X_2) = 16.92$$

$$R(X_0, X_1) = 10.35$$

$$R(X_2 | X_0, X_1) = 6.57 = R(X_0, X_1, X_2) - R(X_0, X_1)$$

	TYPE I SSq (sequential)	TYPE II SSq (partial)
SOURCE		
X1	$R(X_1 X_0) = 8.1$	$R(X_1 X_0, X_2) = 8.1$
	(NOT usually equal)	
X2	$R(X_2 X_0, X_1) = 6.57$	$R(X_2 X_0, X_1) = 6.57$
	(always equal)	

Note: The only reason type I and type II are equal for X_1 is orthogonality. Generally they are not equal. Obviously type I = type II for the last X you have (X_2 in our case).

EXAMPLE 2:

DATA

Y	X0	X1	X2
-2	1	-2	1
-3	1	1	0
0	1	2	0
2	1	1	3

$Y'Y = 17$ $SS(\text{total}) = 17 - 9/4 = 14.75$

X'X **(X'X)⁻¹** **X'Y** **b**

REGRESS Y ON X0 ONLY

4	1/4	- 3	- 3/4
---	-----	-----	-------

(SSR = 9/4 = 2.25 = CT)
(SSR is "uncorrected")

REGRESS Y ON X0, X1

$\begin{bmatrix} 4 & 2 \\ 2 & 10 \end{bmatrix}$	$\begin{bmatrix} 10/36 & -2/36 \\ -2/36 & 4/36 \end{bmatrix}$	$\begin{bmatrix} -3 \\ 3 \end{bmatrix}$	$\begin{bmatrix} -1.0 \\ 0.5 \end{bmatrix}$
---	---	---	---

(SSR = 3 + 1.5 = 4.5)

$R(X1 | X0) = 4.5 - 2.25 = 2.25$

REGRESS Y ON X0, X1, X2

$$\begin{bmatrix} 4 & 2 & 4 \\ 2 & 10 & 1 \\ 4 & 1 & 10 \end{bmatrix} \begin{bmatrix} 0.4670 & -0.0755 & -0.1792 \\ -0.0755 & 0.1132 & 0.0189 \\ -0.1792 & 0.0189 & 0.1698 \end{bmatrix} \begin{bmatrix} -3 \\ 3 \\ 4 \end{bmatrix} \begin{bmatrix} -2.3443 \\ 0.6415 \\ 1.2736 \end{bmatrix}$$

$$(SSR = 14.052)$$

$$R(X0, X1, X2) = 14.052$$

$$R(X0, X1) = 4.500$$

$$R(X2 | X0, X1) = 9.552 = R(X0, X1, X2) - R(X0, X1)$$

SOURCE	TYPE I SSq (sequential)	TYPE II SSq (partial)
X1	$R(X1 X0) = 2.25$	$R(X1 X0, X2) = \underline{\hspace{2cm}}$
X2	$R(X2 X0, X1) = 9.552$	$R(X2 X0, X1) = 9.552$

EXERCISE: Fill in the blank above by regressing Y on X0 and X2, from which you will get $R(X0, X2) = 10.417$. Are type I and type II equal for X1 in this example?

Summary: Type I - Adjust for variables that came before.

Type II - Adjust for all other variables.

(ANS: 3.635, no.)

GRADE – IQ EXAMPLE

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

ANOVA (Grade on IQ)

Source	df	SSq	Mn Sq	F
IQ	1	15.9393	15.9393	0.153
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

ANOVA (Grade on IQ, Study Time)

Source	df	SSq	Mn Sq	F
Model	2	596.12	298.06	32.57
Error	5	45.76	9.15	

SOURCE	df	TYPE I (sequential)	TYPE II (partial)
IQ	1	15.94	121.24
STUDY	1	580.18	580.18

Parameter	Estimate	t	Pr > t	Std. Err.	
INTERCEPT	0.74	0.05	0.9656	16.26	
IQ	0.47	3.64	0.0149	0.13	0.9851
STUDY	2.10	7.96	0.0005	0.26	- 3.64 3.64

From this regression we also can get

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 28.8985 & -0.2261 & -0.2242 \\ -0.2261 & 0.0018 & 0.0011 \\ -0.2242 & 0.0011 & 0.0076 \end{bmatrix}$$

1. To test H0: Coefficient on IQ is 0 (Note: calculations done with extra decimal accuracy.)

(a) Using t-test $t = 0.47 / \sqrt{0.0018 * 9.15} = 3.64$

(b) Using type II F-test, $F = 121.24 / 9.15 = 13.25 = t^2$.

Note: The type II sum of squares is defined by setting $t^2 = F$. This means that type II SSq = $b*b/c$ where b is the coefficient being tested and c is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ which corresponds to b. We have $(0.47)(0.47)/0.0018 = 121.24$.

2. Estimate the mean grade for the population of all potential students with IQ = 113 and study time = 14 hours.

(a) Write this estimate as $\Lambda' \mathbf{b}$ where $\Lambda' = (1, 113, 14)$.

(b) Variance of this is $\Lambda'(\mathbf{X}'\mathbf{X})^{-1} \Lambda * \text{MSE} = 1.303$.

(c) Prediction is $\Lambda' \mathbf{b} = 83.64$.

(d) To get confidence interval,

$$83.64 \pm 2.571 \sqrt{1.303}$$

(e) Interval (80.71, 86.57)

3. Estimate grade for individual with 113 IQ and 14 hours study time.

$$83.64 \pm 2.571 \sqrt{1.303 + 9.15}$$

$$(75.33, 91.95)$$

4. What percent of grade variability is explained by IQ, STUDY?

$$R^2 = (\text{corrected regn. SSQ})/(\text{corrected total SSq}) = 596.12/641.88 = 93\%$$

5. Notes: When a new column is added to a regression, all the coefficients and their t-statistics can change. The t's could go from significance to insignificance or vice-versa.

The exception to the above case is when the added column of \mathbf{X} is orthogonal to the original columns. This means

that the new $X'X$ has the old $X'X$ in the upper left corner, the sum of squares of the new column as the bottom right element, and all other elements 0.

Suggested exercise: Regress GRADE on TIME IQ IQ_ST where $IQ_ST = IQ * STUDY$. The IQ_ST variable could be created in your data step. For the regression of GRADE on TIME and IQ, use the option /I in PROC REG. This will output the $(X'X)^{-1}$ matrix.

" Missing Y trick ": Rerun this example adding a row 113 14 . at the end of the dataset. The dot implies a missing value. Use the statement MODEL GRADE = IQ STUDY / P CLM; Compare to part 2 above. Rerun again with CLI instead of CLM. Compare to part 3 above. Was the extra data row used in computing the regression coefficients? Let's try some of this:

```

OPTIONS LS = 80 NODATE;
DATA GRADES; INPUT IQ STUDY GRADE @@;
  IQ_ST = IQ*STUDY; DATALINES;
105 10 75 110 12 79 120 6 68 116 13 85
122 16 91 130 8 79 114 20 98 102 15 76
;
PROC REG;
  MODEL GRADE = IQ STUDY IQ_ST/SS1 SS2;
  TITLE "GRADE AND STUDY TIME EXAMPLE FROM NOTES";
PROC PLOT;
  PLOT STUDY*IQ = '*' / VPOS = 35;
DATA EXTRA;
  INPUT IQ STUDY GRADE;
DATALINES;
113 14 .

```

```

;
DATA BOTH;
  SET GRADES EXTRA;
PROC REG;
  MODEL GRADE = IQ STUDY/P CLM;
RUN;

```

GRADE AND STUDY TIME EXAMPLE FROM NOTES

DEP VARIABLE: GRADE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	3	610.810	203.603	26.217	0.0043
ERROR	4	31.064674	7.766169		
C TOTAL	7	641.875			

ROOT MSE	2.786785	R-SQUARE	0.9516
DEP MEAN	81.375000	ADJ R-SQ	0.9153
C.V.	3.42462		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T For H0: PARAMETER = 0	PROB> T	TYPE I SS
INTERCEP	1	72.206076	54.072776	1.335	0.2527	52975.125
IQ	1	- 0.131170	0.455300	- 0.288	0.7876	15.939299
STUDY	1	- 4.111072	4.524301	- 0.909	0.4149	580.176
IQ_ST	1	0.053071	0.038581	1.376	0.2410	14.695210

VARIABLE	DF	TYPE II SS
INTERCEP	1	13.848316
IQ	1	0.644589
STUDY	1	6.412303
IQ_ST	1	14.695210

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

$$\hat{G} = 72.21 - 0.13 IQ - 4.11 S + 0.0531 IQ * S.$$

Now if $IQ = 100$ we get

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

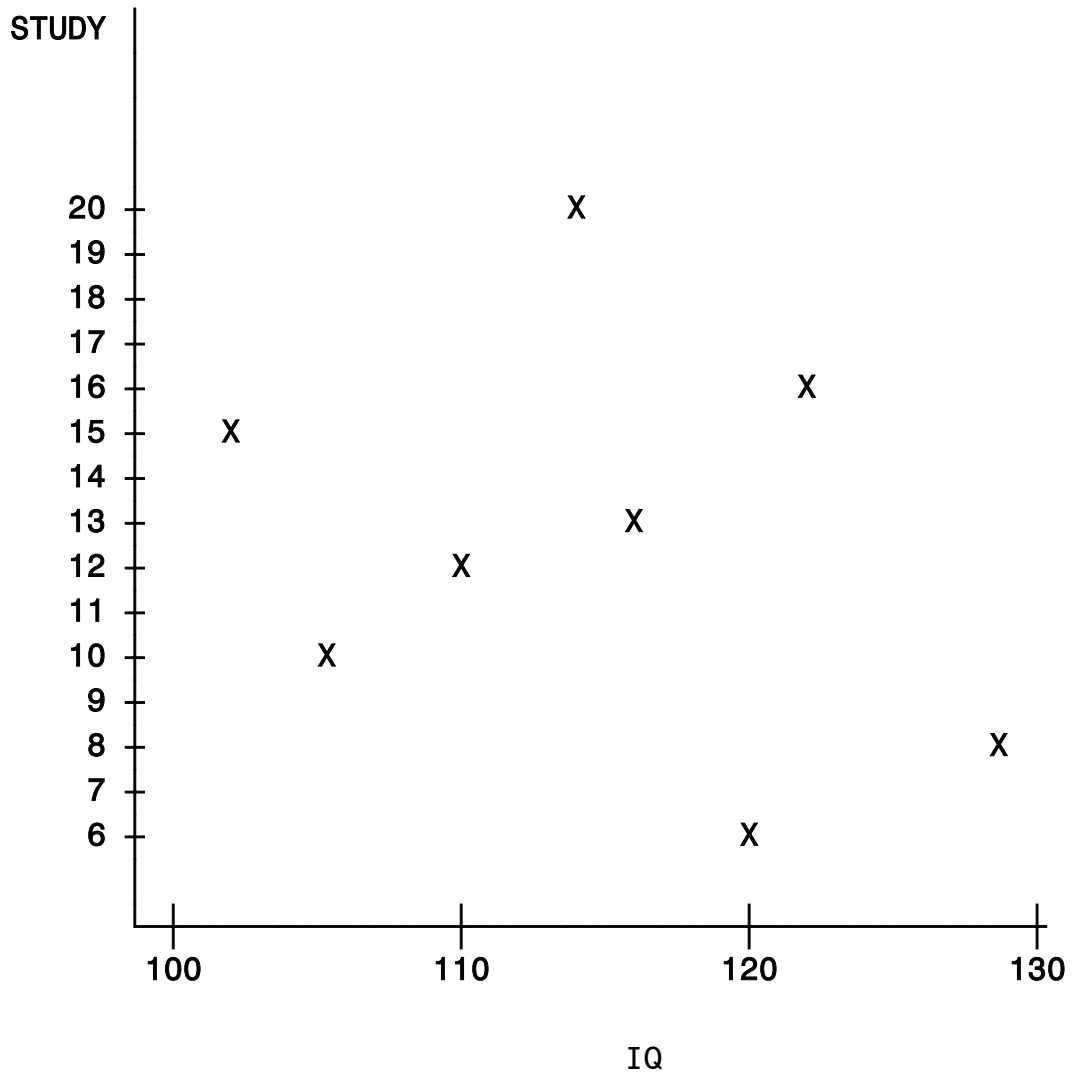
and if $IQ = 120$ we get

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

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.

Suppose we measure IQ in deviations from 100 and $STUDY$ in deviations from 8. What happens to the coefficients and t-tests in the interaction model? How about the main effects model?

GRADE AND STUDY TIME EXAMPLE FROM CLASS NOTES
Plot of STUDY*IQ. Symbol used is 'X'.



GRADE AND STUDY TIME EXAMPLE FROM NOTES

DEP VARIABLE: GRADE

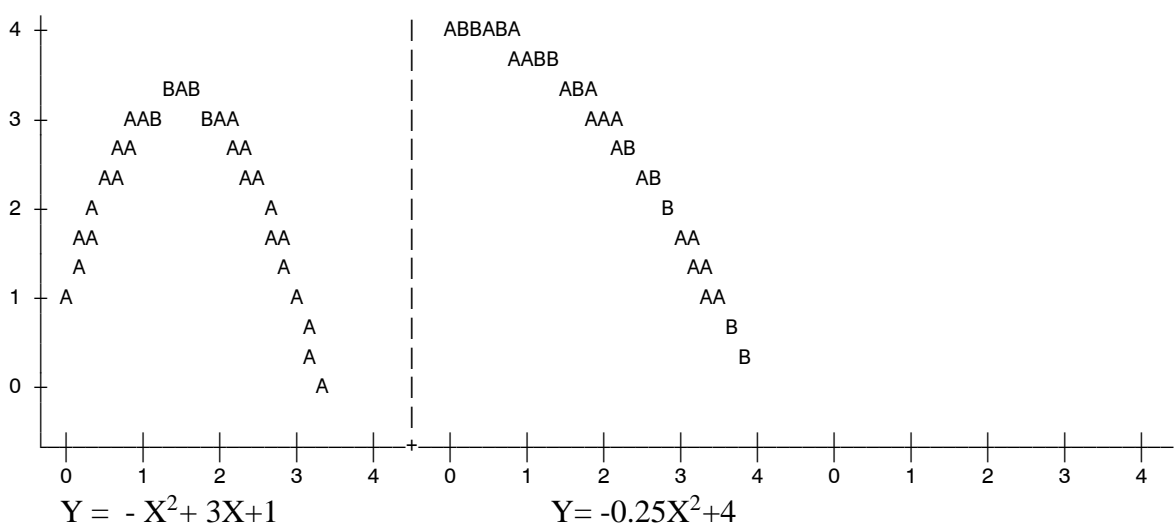
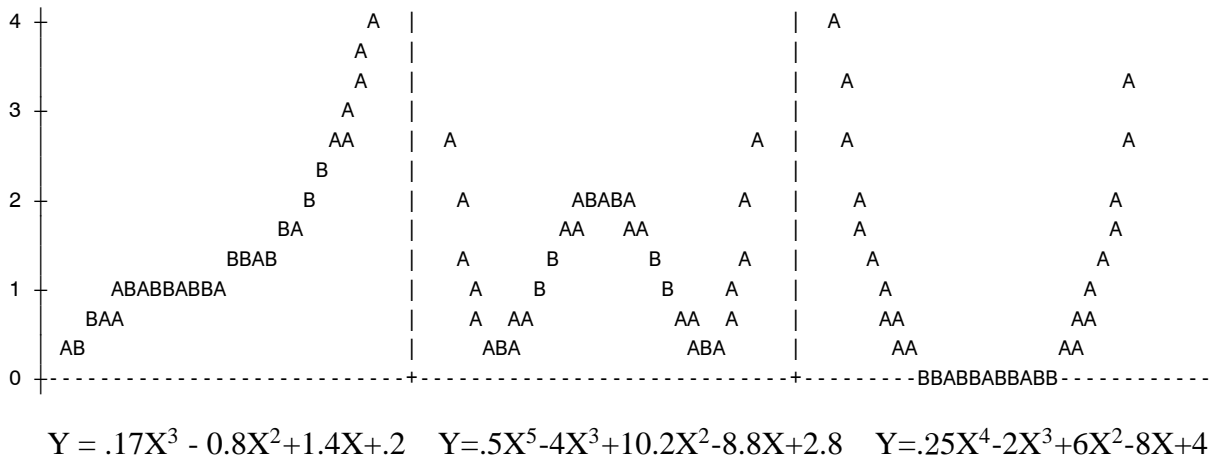
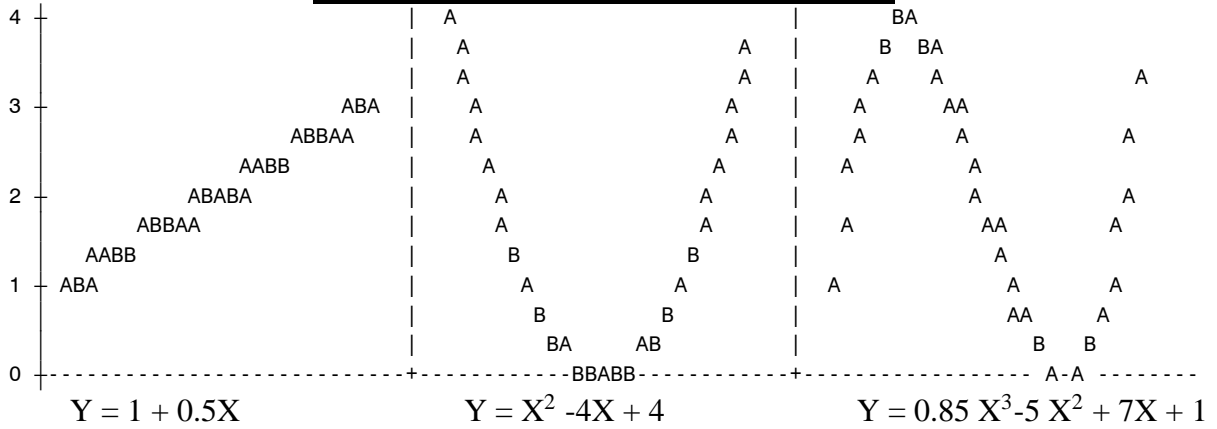
SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	2	596.115	298.058	32.568	0.0014
ERROR	5	45.759885	9.151977		
C TOTAL	7	641.875			
ROOT MSE		3.025223	R-SQUARE	0.9287	
DEP MEAN		81.375000	ADJ R-SQ	0.9002	
C.V.		3.717633			

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER = 0	PROB> T
INTERCEP	1	0.736555	16.262800	0.045	0.9656
IQ	1	0.473084	0.129980	3.640	0.0149
STUDY	1	2.103436	0.264184	7.962	0.0005

OBS	ACTUAL	PREDICT VALUE	STD ERR PREDICT	LOWER 95% MEAN	UPPER 95% MEAN	RESIDUAL
1	75.000	71.445	1.933	66.477	76.412	3.555
2	79.000	78.017	1.270	74.752	81.282	0.983001
3	68.000	70.127	1.963	65.082	75.173	- 2.127
4	85.000	82.959	1.093	80.150	85.768	2.041
5	91.000	92.108	1.835	87.390	96.826	- 1.108
6	79.000	79.065	2.242	73.303	84.827	- .064928
7	98.000	96.737	2.224	91.019	102.455	1.263
8	76.000	80.543	1.929	75.585	85.500	- 4.543
9	.	83.643	1.141	80.709	86.577	.

SUM OF RESIDUALS 7.10543E-15
 SUM OF SQUARED RESIDUALS 45.75988

POLYNOMIAL REGRESSION



POLYNOMIAL REGRESSION

Polynomial means form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k$$

EXAMPLE: We expose cells to radiation X times (dose) and observe Y=number of mutations. We see that up to a point, increasing the dose increases mutations but beyond that point, mutations begin to decrease. Our goal is to model mutations as a function of dose.

DATA (Y=MUTATIONS, X=DOSE)

Y	3	5	7	12	8	10	10	5	6	4	mean = 7
X	2	2	3	4	4	4	5	5	5	6	mean = 4

		Mean	SSq	df
2	3, 5	4	2	1
3	7	7	0	0
4	12, 8, 10	10	8	2
5	10, 5, 6	7	14	2
6	4	4	0	0
		sum = 24		5

Note: In the second half of this course you will learn a technique called the "analysis of variance" or ANOVA. The ANOVA essentially fits a mutation mean for each dose. The ANOVA error sum of squares, 24, is just the pooled (summed) SSq from within each treatment group. The predictions (group

means) are not constrained in any way. The mean Y is 7 and the sum of squares of the Y deviations, the "total sum of squares" is $(3-7)^2+(5-7)^2+\dots+(4-7)^2 = 78$ with $10-1 = 9$ degrees of freedom (DF). ANOVA breaks this 78 into the unexplained error sum of squares 24 plus a part $78-24 = 54$ that is attributed to the variation among the "treatment" means (i.e. to dose):

ANOVA

Source	df	SSq	Mn Sq	F	Pr>F
Dose	4	54	13.5	2.81	0.1436
Error	5	24	4.8		

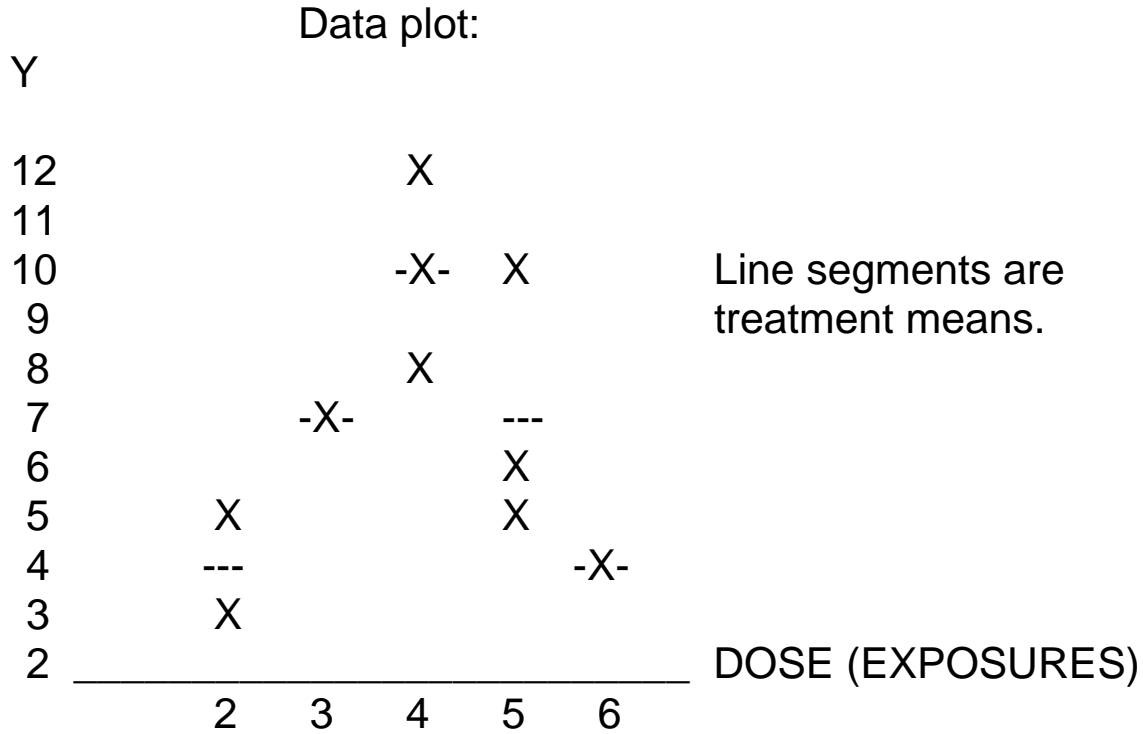
The job can be done easily in PROC GLM.

```
PROC GLM; CLASS X; MODEL Y = X;
```

Try this and you will get a printout like this:
(this is the computerized way of getting the ANOVA table).

GENERAL LINEAR MODELS PROCEDURE
CLASS LEVEL INFORMATION

		CLASS	LEVELS	VALUES				
		DOSE	5	2	3	4	5	6
SOURCE	DF	SSq	Mn Sq	F				
MODEL	4	54	13.5	2.81				
ERROR	5	24	4.8					



Next we will fit a quadratic to the data (forcing predicted values to lie on a parabola) and observe how much the fit deteriorates.

```
DATA A; INPUT Y X XSQ;
CARDS;
```

```

    3      2      4
    5      2      4
    7      3      9
   10      4     16
    8      4     16
   12      4     16
    5      5     25
   10      5     25
    6      5     25
    4      6     36
```

```
PROC REG; MODEL Y = X XSQ;
```

The output contains the following information:

SOURCE	DF	SSq	Mn Sq	F	Pr>F
MODEL	2	49.42	24.71	6.05	0.0298
ERROR	7	28.58	4.08		

PARAMETER	ESTIMATE	T	PR> T
INTERCEPT	-12.5280	-2.19	.0644
X	11.0311	3.47	.0104
XSQ	-1.3975	-3.40	.0115

The ANOVA, which basically fits group means to the 5 treatment groups, has increased the regression sum of squares over that of the quadratic, from 49.4161 (2df) to 54 (4df). Later, we will show that the ANOVA (means) model is like a "full model" and the quadratic like a "reduced model."

The test which asks if the regression does as well as the unconstrained means ANOVA in fitting the data is called a "lack of fit F test." We compute

$$F_5^2 = [(54 - 49.4161)/2] / (4.8) = 0.4775$$

Since F is insignificant (2 and 5 df) we say there is no significant lack of fit. We conclude that a model forcing mutation to be a quadratic in dose seems to explain mutation as well as a model in which unconstrained means are fit to the data.

To show that the F test is a full versus reduced model F test, I will show that the ANOVA approach is the same as fitting the highest degree polynomial possible to the data. Since there are $m = 5$ values of X , the highest possible degree is $m - 1 = 4$. Thus we issue the commands:

```
PROC GLM; MODEL Y = X X*X X*X*X X*X*X*X;
```

Notice that no CLASS statement was used and that PROC GLM will actually compute the powers of X for you. We get:

SOURCE	DF	SSq	Mn Sq	F
MODEL	4	54	13.5	2.81
ERROR	5	24	4.8	

SOURCE	TYPE I SS	TYPE II SS
X	2.25	2.93
X*X	47.17	3.57
X*X*X	0.45	3.98
X*X*X*X	4.13	4.13

PARAMETER	ESTIMATE	T	PR > T
INTERCEPT	82	0.73	0.4962
X	-100	-0.78	0.4698
X*X	44.5	0.86	0.4273
X*X*X	-8	-0.91	0.4041
X*X*X*X	0.5	0.93	0.5388

Using the TYPE I SSq we compute the lack of fit F as:

$$F = [(0.45 + 4.13)/2] / 4.8 = 0.4771$$

the same as before and we thus see that the lack of fit statistic is testing for the powers of X up to the highest possible power you can fit to the data. The only reason that there is an error term left to test this against is the fact that some X's had repeated Y's with them and so the highest possible degree, $m - 1 = 4$, is less than the total degrees of freedom $n - 1 = 9$ leaving 5 degrees of freedom (with sum of squares 24) for "pure error."

As before, we see that it is incorrect and dangerous to make a conclusion about the joint significance of all the coefficients taken together if we look only at the t statistics.

RESPONSE SURFACE METHODOLOGY

In a response surface model, a response is some function (usually quadratic) of one or more control variables. Here is an example (of yields in a chemical reaction) analyzed on the computer:

```
DATA REACT; INPUT YIELD PH TEMP@@;
PSQ = PH**2; TSQ = TEMP**2; PT = PH*TEMP;
CARDS;
  90 5 60 100 5 80 95 5 100 105 5.5 80
100 6 60 130 6 80 125 6 100 140 6.5 80
135 7 60 142 7 80 126 7 100
;
```

```

PROC PRINT;
PROC REG; MODEL YIELD = PH TEMP PSQ TSQ PT/P;
PROC RSREG; MODEL YIELD = PH TEMP;

```

Note the use of @@ to keep SAS from going to a new line for each observation read. If we omitted @@ we would get only 2 observations in our data set. You can use the "missing Y trick" to get SAS to compute a 95% prediction interval for the yield of a future reaction at PH 6.3 and temperature 92 degrees. This involves inputting the X values 6.3 and 92 and a missing value "." for Y. SAS cannot use this observation to determine the regression coefficients (it does not know the Y value) but once these are determined, it can predict since it knows the X values.

CHEMICAL PROCESS YIELDS

OBS	YIELD	PH	TEMP	PSQ	TSQ	PT
1	90	5.0	60	25.00	3600	300
2	100	5.0	80	25.00	6400	400
3	95	5.0	100	25.00	10000	500
4	105	5.5	80	30.25	6400	440
5	100	6.0	60	36.00	3600	360
6	130	6.0	80	36.00	6400	480
7	125	6.0	100	36.00	10000	600
8	140	6.5	80	42.25	6400	520
9	135	7.0	60	49.00	3600	420
10	142	7.0	80	49.00	6400	560
11	126	7.0	100	49.00	10000	700

CHEMICAL PROCESS YIELDS

Model: MODEL1

Dependent Variable: YIELD

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	3331.65539	666.33108	8.429	0.0177
Error	5	395.25370	79.05074		
C Total	10	3726.90909			

Root MSE	8.89105	R-square	0.8939
Dep Mean	117.09091	Adj R-sq	0.7879
C.V.	7.59329		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob> T
INTERCEP	1	-382.624093	239.89941787	-1.595	0.1716
PH	1	70.619739	74.04775138	0.954	0.3840
TEMP	1	5.652925	2.57066503	2.199	0.0792
PSQ	1	-2.981132	5.98302278	-0.498	0.6394
TSQ	1	-0.027675	0.01368841	-2.022	0.0991
PT	1	-0.175000	0.22227621	-0.787	0.4668

Obs	Dep Var YIELD	Predict Value	Residual
1	90.0	83.0	7.0065
2	100.0	101.1	-1.0633
3	95.0	97.0	-1.9935
4	105.0	113.7	-8.7222
5	100.0	110.3	-10.3208
6	130.0	124.9	5.1094
7	125.0	117.3	7.6792
8	140.0	134.6	5.4316
9	135.0	131.7	3.3142
10	142.0	142.8	-0.7556
11	126.0	131.7	-5.6858

Sum of Residuals 0
 Sum of Squared Residuals 395.2537
 Predicted Resid SS (Press) 3155.9162

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
PH	6.000000	1.000000
TEMP	80.000000	20.000000

Response Surface for Variable YIELD

Response Mean 117.090909
 Root MSE 8.891048
 R-Square 0.8939
 Coef. of Variation 7.5933

Regression	Degrees of Freedom	Type I SSq	R-Square	F-Ratio	Prob > F
Linear	2	2898.15	0.7776	18.331	0.0050
Quadratic	2	384.50	0.1032	2.432	0.1829
Crossproduct	1	49.00	0.0131	0.620	0.4668
Total Regress	5	3331.66	0.8939	8.429	0.0177

Residual	Degrees of Freedom	Sum of Squares	Mean Square
Total Error	5	395.253701	79.050740

Parameter	Degrees of Freedom	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEPT	1	-382.6241	239.8994	-1.595	0.1716
PH	1	70.6197	74.0478	0.954	0.3840
TEMP	1	5.6529	2.5707	2.199	0.0792
PH*PH	1	-2.9811	5.9830	-0.498	0.6394
TEMP*PH	1	-0.1750	0.2223	-0.787	0.4668
TEMP*TEMP	1	-0.0277	0.0137	-2.022	0.0991

Factor	Degrees of Freedom	Sum of Squares	Mean Square	F-Ratio	Prob > F
PH	3	2893.2796	964.426544	12.200	0.0098
TEMP	3	445.6173	148.539109	1.879	0.2508

Canonical Analysis of Response Surface
(based on coded data)

Factor	Critical Value	
	Coded	Uncoded
PH	3.751711	9.751711
TEMP	-0.435011	71.299771

Predicted value at stationary point 163.233672

Eigenvalues	Eigenvectors	
	PH	TEMP
-2.618751	0.979226	-0.202773
-11.432192	0.202773	0.979226

Stationary point is a maximum.

=====

Example:

$$\hat{Y} = -382 + 70.62 P + 5.65 T - 2.98 P^2 - 0.028 T^2 - 0.175 PT$$

- Critical point: P = 9.7517, T = 71.2998

- $\hat{Y} = 163.23 + (P-9.7517 \quad T-71.2998) \begin{pmatrix} -2.9800 & -0.0875 \\ -0.0875 & -0.0280 \end{pmatrix} \begin{pmatrix} P - 9.7517 \\ T - 71.2998 \end{pmatrix} = 163.23 + \mathbf{X'AX}$

$$\bullet \begin{pmatrix} -2.98 & -0.0875 \\ -0.0875 & -0.028 \end{pmatrix} = \mathbf{A} = \mathbf{Z L Z}' =$$

$$\begin{pmatrix} -.0296 & .9996 \\ .9996 & .0296 \end{pmatrix} \begin{pmatrix} -.0251 & 0 \\ 0 & -2.9837 \end{pmatrix} \begin{pmatrix} -.0296 & .9996 \\ .9996 & .0296 \end{pmatrix}$$

$$\bullet \hat{Y} = 163.23 + \mathbf{X}'\mathbf{Z L Z}'\mathbf{X} = 163.23 + \mathbf{W}'\mathbf{L W} =$$

$$163.23 + (-.0251) w_1^2 + (-2.984) w_2^2$$

- $(w_1, w_2) = (0,0)$ = critical point, response is 163.23. Any movement away from critical point *reduces* response.

Additional Points: Critical point may be max, min, or saddle point. It may be nowhere near experimental region. Ridge Analysis (not ridge regression) takes spheres of ever increasing radius around some point in the experimental region. On each sphere, coordinates of response maximizer (minimizer) are computed resulting in a path of maximum increase (decrease). PROC RSREG has a RIDGE statement to do this. Again eigenvalues are involved.

DUMMY VARIABLES

Y = time to get my results from the lab

X = 1 if it's lab A, 0 if it's lab B:

X	0	1	1	0	1	0	0	1
Y	18	25	24	17	27	22	20	23

$X-\bar{X}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$
$Y-22$	-4	3	2	-5	5	0	-2	1

Total sum of squares $\sum_1^8 (Y_i - \bar{Y})^2 = 16 + 9 + 4 + \dots + 1 = 84$

X variance $\sum_1^8 (X_i - \bar{X})^2 / 7 = 2 / 7$

Covariance: $[(-\frac{1}{2})(-4) + \dots + (\frac{1}{2})(1)] / 7 = 11 / 7$
 $b = 11/2 = 5.5$ $a = 22 - 5.5(1/2) = 19.25$

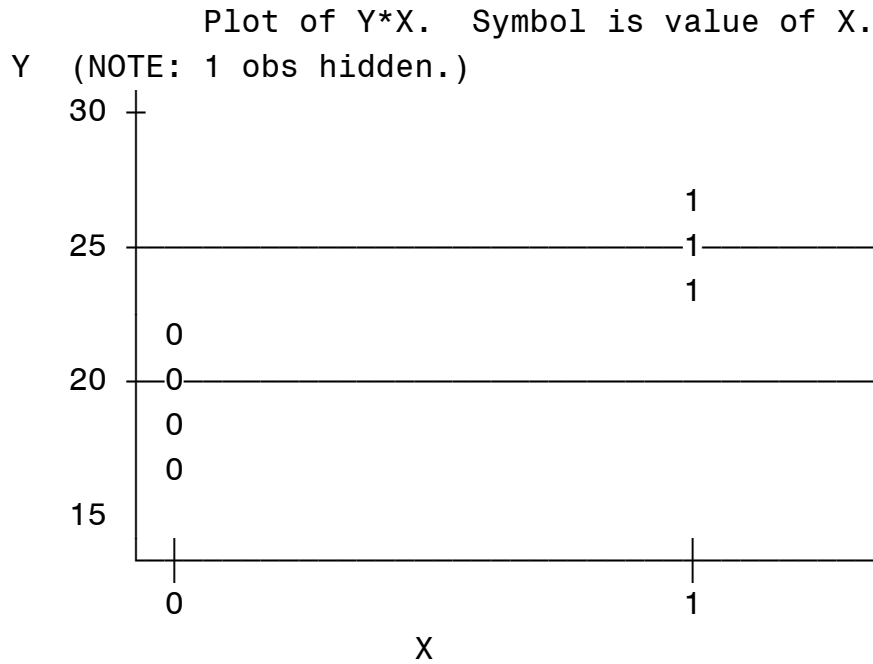
$$\hat{Y} = 19.25 + 5.5X = \begin{cases} 19.25 & \text{if } X=0 \text{ (lab B)} \\ 24.75 & \text{if } X=1 \text{ (lab A)} \end{cases}$$

slope = 5.5 is "shift" = difference of 2 levels.

```

Data labs; input X Y NUM @@; cards;
  0 18 11   1 25 14   1 24 10   0 17 11
  1 27 17   0 22 13   0 20 9    1 23 11
;
proc plot; plot Y*X=X/vpos=10 hpos=40 vref=19.25 24.75;
proc reg; model Y=X/p; run;

```



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	60.50000	60.50000	15.45	0.0077
Error	6	23.50000	3.91667		
Corrected Total	7	84.00000			

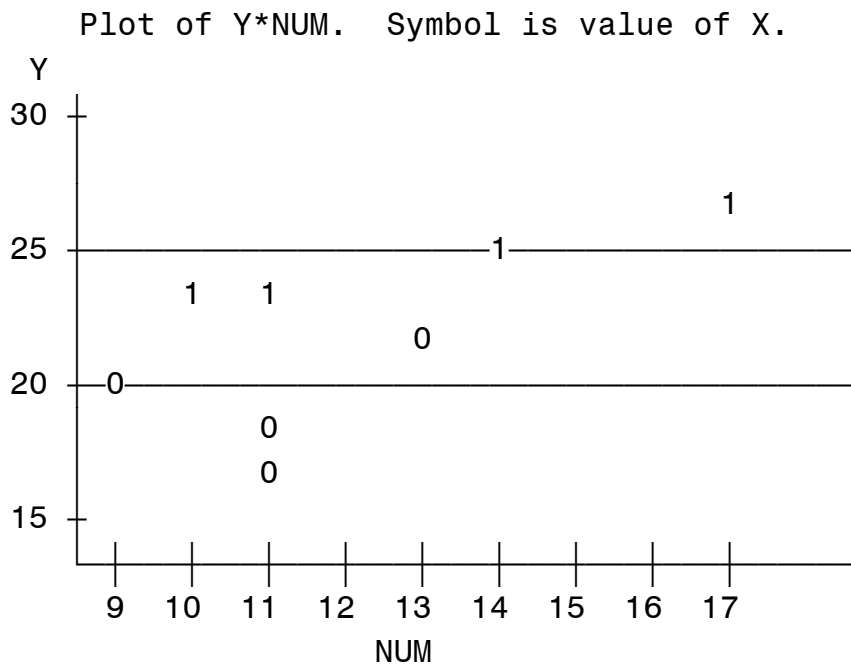
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	19.25000	0.98953	19.45	<.0001
X	1	5.50000	1.39940	3.93	0.0077

Obs	Dep Var Y	Predicted Value	Residual
1	18.0000	19.2500	-1.2500
2	25.0000	24.7500	0.2500
3	24.0000	24.7500	-0.7500
4	17.0000	19.2500	-2.2500
5	27.0000	24.7500	2.2500
6	22.0000	19.2500	2.7500
7	20.0000	19.2500	0.7500
8	23.0000	24.7500	-1.7500

Notes: predicted values are MEANS, slope is difference of means, t-test is same as usual 2-sample t.

Number of samples analyzed (NUM) may also affect turnaround.

```
proc plot;
plot Y*num=X/vpos=10 hpos=40 vref=19.25 24.75;
```



```
proc reg; model Y = X num;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	70.00000	35.00000	12.50	0.0113
Error	5	14.00000	2.80000		
Corrected Total	7	84.00000			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	13.75000	3.10093	4.43	0.0068
X	1	4.50000	1.30182	3.46	0.0181
NUM	1	0.50000	0.27145	1.84	0.1248

$$\text{Lab A: } \hat{Y} = (13.75 + 1(4.5)) + .5 * \text{Num}$$

$$\text{Lab B: } \hat{Y} = (13.75 + 0(4.5)) + .5 * \text{Num}$$

COVARIANCE ANALYSIS

The example above results in two parallel lines. This kind of analysis is often referred to as analysis of covariance or ANCOVA. The ANOVA model is written:

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

where α_1 is the lab A effect and α_2 is the lab B effect.

Now write a model which incorporates both the treatment effect and the linear effect of number of samples analyzed, N , as displayed in the graph. We write

$$Y_{ij} = \mu + \alpha_i + \beta^*(N_{ij} - \bar{N}_{..}) + e_{ij}$$

where $\bar{N}_{..}$ is the sample mean of the "covariate" (number of samples analyzed in our case). You can fit this model using either the covariate N_{ij} or the deviations $(N_{ij} - \bar{N}_{..})$ of the covariate from its sample mean.

Finally, we see that at the average number of samples $\bar{N}_{..}=22$ (or $N - \bar{N}_{..}=0$ if you used deviations as your covariate), the predicted values are $13.75 + 0 + 0.5(22) = 24.75$ and $13.75 + 4.5 + 0.5(22) = 29.25$. This illustrates the fact that a covariance analysis simply adjusts all the analysis times to the levels they would have had if both labs had analyzed $\bar{N}_{..}=22$ samples. These are often called adjusted treatment means .

A covariance example:

3 technicians. Each decodes 6 strands.

Decoding time is response, depends on strand length.

Want to compare technicians.

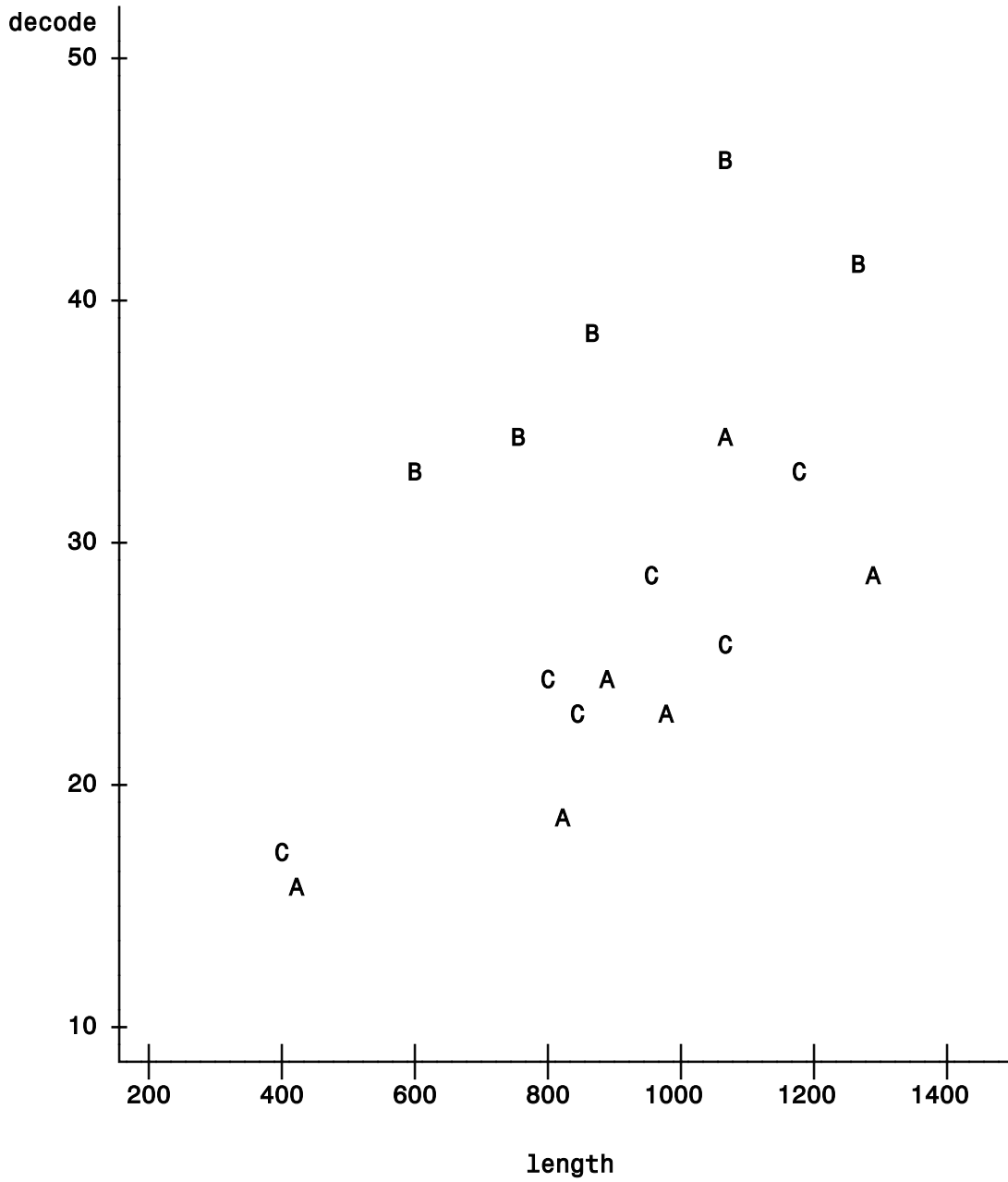
```

data gen;
input tech $ decode length tech1 tech2 tech3
           ltech1   ltech2   ltech3;
cards;
A   15.3   422   1   0   0   422   0   0
A   19.0   815   1   0   0   815   0   0
A   28.2  1279   1   0   0  1279   0   0
A   34.7  1067   1   0   0  1067   0   0
A   24.1   883   1   0   0   883   0   0
A   22.9   988   1   0   0   988   0   0
B   38.0   876   0   1   0     0   876   0
B   34.5   761   0   1   0     0   761   0
B   32.2   594   0   1   0     0   594   0
B   41.2  1276   0   1   0     0  1276   0
B   33.8  1069   0   1   0     0  1069   0
B   46.1  1071   0   1   0     0  1071   0
C   25.1  1072   0   0   1     0     0  1072
C   22.6   849   0   0   1     0     0   849
C   33.1  1173   0   0   1     0     0  1173
C   28.1   946   0   0   1     0     0   946
C   16.6   389   0   0   1     0     0   389
C   24.8   810   0   0   1     0     0   810
;

proc plot; plot decode*length=tech/vpos=30 hpos=60;

```

Plot of decode*length. Symbol is value of tech.



Fit three parallel lines:

```
proc reg; model decode = length tech1 tech2 tech3;
```

Dependent Variable: decode

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1016.21466	338.73822	24.12	<.0001
Error	14	196.63479	14.04534		
Corrected Total	17	1212.84944			

NOTE: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

NOTE: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

$$\text{tech3} = \text{Intercept} - \text{tech1} - \text{tech2}$$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	B	9.92102	3.48649	2.85	0.0130
length	1	0.01733	0.00359	4.83	0.0003
tech1	B	-1.63754	2.16756	-0.76	0.4625
tech2	B	11.40513	2.17745	5.24	0.0001
tech3	0	0	.	.	.

```
9.92 -1.64 +0.0173*Length tech 1
9.92 +11.405 +0.0173*Length tech 2
9.92 + 0 +0.0173*Length tech 3
```

Fit three arbitrary lines:

```
proc reg; model decode = length tech1 tech2 ltech1 ltech2/ss1;
run;
```

Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	5	1020.99951	204.19990	12.77	0.0002	
Error	12	191.84994	15.98749			
Corrected Total	17	1212.84944				

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	9.13489	5.93965	1.54	0.1500	15040
length	1	0.01823	0.00654	2.79	0.0164	415.15078
tech1	1	-2.33106	8.35417	-0.28	0.7850	215.73232
tech2	1	15.33839	9.20613	1.67	0.1216	385.33156
ltech1	1	0.00072746	0.00902	0.08	0.9370	1.77315
ltech2	1	-0.00424	0.00978	-0.43	0.6720	3.01170

9.13 - 2.33 +(0.0182 + .0007)*Length tech 1
 9.13 + 15.33 +(0.0182 - .0042)*Length tech 2
 9.13 + 0 +(0.0182 + 0)*Length tech 3

Are lines parallel? <=> Do we need Ltech1, Ltech2?

$F = [(1.773+3.012)/2] / 15.987 = 0.15$
 Not significant, parallel lines OK.

Or...

Parallel Lines Model SSE = 196.63
 Arbitrary Lines Model SSE = 191.85
 $F = [(196.63-191.85)/2] / 15.987 = 0.15$

Because the lines can be taken to be parallel, computing adjusted treatment means for the three technicians makes sense. These are estimates of the average decoding times for the three technicians, if all strands had been the same length (namely the average observed strand length $\bar{L} = 907.8$)

9.92 - .64	+0.0173*907.8 = 24.98	tech 1
9.92 +11.405	+0.0173*907.8 = 37.03	tech 2
9.92 + 0	+0.0173*907.8 = 25.62	tech 3

Count data

1. Estimating, testing proportions

100 seeds, 45 germinate. We estimate probability p that a plant will germinate to be 0.45 for this population. Is a 50% germination rate a reasonable possibility?

$$\Pr\{45 \text{ or less germinate in } 100 \text{ trials if } p=0.5\} = ???$$

Binomial:

n independent trials

Each trial success or failure

p =probability of success same on every trial

X = observed number of successes in n trials

$$\Pr\{X=r\} = n!/[r!(n-r)!] p^r (1-p)^{(n-r)}$$

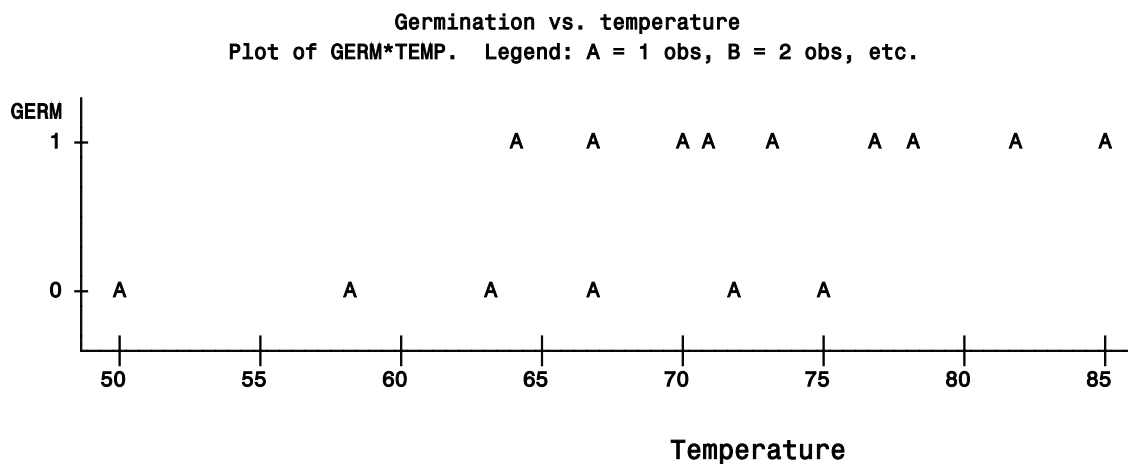
$$0!=1, \quad 1!=1, \quad 2!=2*1=2, \quad 3!=3*2*1=6, \quad 4!=4*3*2*1=24 \text{ etc.}$$

Logistic Regression

Idea: p =probability of germinating = function of some variables (maybe temperature, moisture, or both).

Example:

Temperatures									
Germinating	70	73	78	64	67	71	77	85	82
Not germ.	50	63	58	72	67	75			



Idea: Regress Germ (0 or 1) on Temperature:

Germination vs. temperature

Model: MODEL1
Dependent Variable: GERM

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	1.09755	1.09755	5.702	0.0328
Error	13	2.50245	0.19250		
C Total	14	3.60000			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	-1.550126	0.90755721	-1.708	0.1114
TEMP	1	0.030658	0.01283925	2.388	0.0328

Germination vs. temperature

Obs	TEMP	Dep Var GERM	Predict Value	Residual	
1	70	1.0000	0.5959	0.4041	
2	73	1.0000	0.6879	0.3121	
3	78	1.0000	0.8412	0.1588	
4	64	1.0000	0.4120	0.5880	
5	67	1.0000	0.5039	0.4961	
6	71	1.0000	0.6266	0.3734	
7	77	1.0000	0.8105	0.1895	
8	85	1.0000	1.0558	-0.0558	<--
9	82	1.0000	0.9638	0.0362	
10	50	0	-0.0172	0.0172	<--
11	63	0	0.3813	-0.3813	
12	58	0	0.2280	-0.2280	
13	72	0	0.6572	-0.6572	
14	67	0	0.5039	-0.5039	
15	75	0	0.7492	-0.7492	

* Normal residuals ?

* Reasonable predicted probabilities? 1.0558? -0.0172 ?

Better idea: Map $0 < p < 1$ into $L = \ln\left(\frac{p}{1-p}\right)$ then model

$$L = \alpha + \beta(\text{temperature}) + e$$

or

$$L = \alpha + \beta(\text{temperature}-70)$$

"Likelihood" = probability of sample = $p(1-p)p(1-p)p \dots p$ Use p for germinated, $1-p$ for not germinated.

Substitute $p = e^L / (1 + e^L)$, $1 - p = 1 / (1 + e^L)$ and $L = \alpha + \beta X$

"Maximum Likelihood Estimates"

Likelihood =

$$\left[\frac{e^{\alpha + \beta(70-70)}}{1 + e^{\alpha + \beta(70-70)}} \right] \left[\frac{e^{\alpha + \beta(73-70)}}{1 + e^{\alpha + \beta(73-70)}} \right] \dots \left[\frac{1}{1 + e^{\alpha + \beta(75-70)}} \right] = f(\alpha, \beta).$$

Graph $f(\alpha, \beta)$ vs. (α, β) and find values of (α, β) that maximize.

Theory also gives standard errors (large sample approximations) . Use PROC LOGISTIC, PROC GENMOD, or PROC CATMOD in SAS. We get

$$\text{Pr}\{\text{Germinate}\} = e^{-4961 + 0.1821 \cdot X} / (1 + e^{-4961 + 0.1821 \cdot X})$$

where $X = \text{temperature} - 70$.

```
Data seeds; Input Germ $ 1-3 n @; Y=(Germ="Yes");
  If Germ=" " then Y=.;
  do i=1 to n; input temp @; output; end;
cards;
Yes 9          64 67 70 71  73  77 78 82 85
No  6          50 58 63  67      72  75
    23  46 48 50 52 54 56 58 60 62 64 66 68
    70 72 74 76 78 80 82 84 86 88 90
PROC LOGISTIC data=seeds order=data;
  model germ=temp / itprint ctable pprob=.6923;
  output out=out1 predicted=p xbeta=logit;
proc plot; plot p*temp Y*temp=y/vpos=20 overlay;
run;
```

The LOGISTIC Procedure

Data Set: WORK.SEEDS
 Response Variable: GERM
 Response Levels: 2
 Number of Observations: 15
 Link Function: Logit

Response Profile

Ordered Value	GERM	Count
1	Yes	9
2	No	6

WARNING: 23 observation(s) were deleted due to missing values for the response or explanatory variables.

Maximum Likelihood Iterative Phase

Iter	Step	-2 Log L	INTERCPT	TEMP
0	INITIAL	20.190350	0.405465	0
1	IRLS	15.205626	-8.553392	0.127740
2	IRLS	14.878609	-11.501730	0.171150
3	IRLS	14.866742	-12.219688	0.181644
4	IRLS	14.866718	-12.253782	0.182141
5	IRLS	14.866718	-12.253854	0.182142

Model Fitting Information and Testing Global Null Hypothesis
BETA=0

Criterion	Intercept and Covariates			Chi-Square for Covariates
	Intercept Only	Intercept and Covariates	Intercept and Covariates	
AIC	22.190	18.867	.	
SC	22.898	20.283	.	
-2 LOG L	20.190	14.867		5.324 with 1 DF (p=0.0210)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Std Error	Wald Chi-Sq	Pr > Chi-Sq	Stdrdized Estimate	Odds Ratio
INTERCPT	1	-12.2539	7.194	2.901	0.0885	.	.
TEMP	1	0.1821	0.103	3.103	0.0782	0.917127	1.200

Association of Predicted Probabilities and Observed Responses

Concordant = 79.6%	Somers' D = 0.611
Discordant = 18.5%	Gamma = 0.623
Tied = 1.9%	Tau-a = 0.314
(54 pairs)	c = 0.806

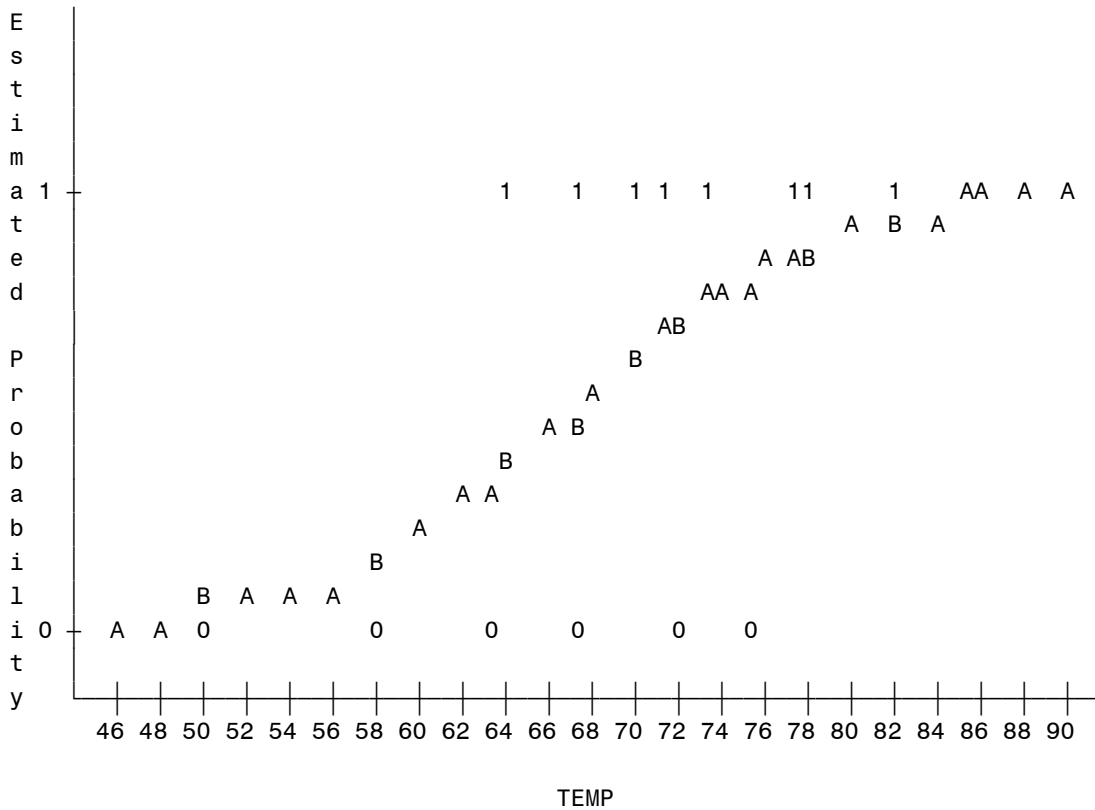
Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non- Event	Event	Non- Event	Sensi- Correct	Speci- tivity	Speci- ficity	False POS	False NEG
0.692	5	4	2	4	60.0	55.6	66.7	28.6	50.0

explanation:

	Actual		
	Event	Non-Event	
Decision			% Correct: 9/15 = 60%
Event	5	2 (7)	% Sensitivity: 5/9
Non	4	4 (8)	% Specificity: 4/6
	(9)	(6)	% False POS: 2/7
			% False NEG: 4/8

Plot of P*TEMP. Legend: A = 1 obs, B = 2 obs, etc.
 Plot of Y*TEMP. Symbol is value of Y.



Likelihood Ratio Chi-Square

Small contingency table

$$\begin{array}{ccc} 3 & 3 & (6) \\ 5 & 2 & (7) \\ (8) & (5) & [13] \end{array} \leftarrow \text{Probabilities are } \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$$

Likelihood is some constant times

$$(p_{11})^{n_{11}}(p_{12})^{n_{12}}(p_{21})^{n_{21}}(p_{22})^{n_{22}}$$

where we must have these ps summing to 1 so that

$$p_{22} = 1 - p_{11} - p_{12} - p_{21}.$$

The values of these ps that maximize the likelihood are the same values that maximize the logarithm of the likelihood, namely

$$n_{11} \ln(p_{11}) + \dots + n_{22} \ln(1 - p_{11} - p_{12} - p_{21})$$

and taking the derivatives with respect to each of the three unconstrained ps we have

$$n_{ij}/p_{ij} = n_{22}/(1 - p_{11} - p_{12} - p_{21})$$

and if we then solve these 3 equations

((i,j) = (1,1), (1,2), (2,1)) we get estimates

$$\hat{p}_{ij} = n_{ij} / n_{..}$$

so we have 3/13, 3/13, 5/13, and 2/13 which we then plug into the log likelihood function to get

$$-2 \log(\text{Likelihood}) = C - 2[3 \ln(3/13) + 3 \ln(3/13) + 5 \ln(5/13) + 2 \ln(2/13)]$$

which is $C + 34.638368$ where C is some constant.

Suppose $p_{11} = p_r p_c$ etc. where p_r and p_c are probabilities of being in the first row and of being in the first column respectively. This would be suggested by the independence hypothesis. Then the likelihood is proportional to

$$(p_r p_c)^{n_{11}} (p_r (1-p_c))^{n_{12}} ((1-p_r) p_c)^{n_{21}} ((1-p_r)(1-p_c))^{n_{22}}$$

Taking logs and differentiating we have

$$\hat{p}_r = 6/13 \text{ and } \hat{p}_c = 8/13$$

and $-2 \log(\text{Likelihood}) = C + 35.268067$. The difference in $-2 \log(\text{Likelihood})$ from the full and reduced models has approximately a Chi-square distribution with degrees of freedom equal to the difference in the number of unrestricted parameters. The difference, 0.6297, has 1 df and is the likelihood ratio Chi-square on the printout.

DATA LRT;

Input Altered \$ Frost \$ n;

datalines;

Yes No 5

Yes Yes 2

No No 3

No Yes 3

;

```
proc freq;
  table Altered*Frost/chisq norow nocol;
  weight n; run;
```

Frost damage, genetically altered and unaltered plants

The FREQ Procedure
Table of Altered by Frost

Altered	Frost		Total
	No	Yes	
Frequency			
Percent			
No	3 23.08	3 23.08	6 46.15
Yes	5 38.46	2 15.38	7 53.85
Total	8 61.54	5 38.46	13 100.00

Statistics for Table of Altered by Frost

Statistic	DF	Value	Prob
Chi-Square	1	0.6268	0.4285
Likelihood Ratio Chi-Square	1	0.6297	0.4275
Continuity Adj. Chi-Square	1	0.0484	0.8259
Mantel-Haenszel Chi-Square	1	0.5786	0.4469
Phi Coefficient		-0.2196	
Contingency Coefficient		0.2145	
Cramer's V		-0.2196	

WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Statistics for Table of Altered by Frost

Fisher's Exact Test

Cell (1,1) Frequency (F)	3
Left-sided Pr \leq F	0.4126
Right-sided Pr \geq F	0.9138
Table Probability (P)	0.3263
Two-sided Pr \leq P	0.5921

Sample Size = 13

Notice the warning. The cell counts are not high enough for our usual Chi-square or the likelihood ratio Chi-square test statistics to have close to a χ^2 distribution (both are only **approximately** Chi-square in **large** samples).

One approach to this is to use Fisher's exact test. How many tables are more extreme than this one? First, what do we mean by "extreme"? We expect $48/13 = 3.7$ unaltered plants to have no damage but we observe less (3). If we insist on preserving the row and column totals, what other tables could we get with even less unaltered plants that show no damage?

2	4	(P=.08158)	and	1	5	(P=.004662)
6	1			7	0	

are even more extreme. Fisher suggested assigning hypergeometric probabilities (as shown) to these tables. Using n_{ij} to denote the count in row i and column j , $n_{i\bullet}$ to denote the total row i count, $n_{\bullet j}$ to denote the total column j count and $n_{\bullet\bullet}$ for

the total count (13) the hypergeometric probability for the original table, for example, is

$$P = \frac{n_{1\bullet}! n_{2\bullet}! n_{\bullet 1}! n_{\bullet 2}!}{n_{11}! n_{12}! n_{21}! n_{22}! n_{\bullet\bullet}!} = \frac{6! 7! 8! 5!}{3! 3! 5! 2! 13!} = .3263 \text{ for } \begin{array}{|c|c|} \hline 3 & 3 \\ \hline 5 & 2 \\ \hline \end{array}$$

so that .3263+.08158+.00466 = .4126 = left sided Fisher exact P-value on printout.

Note: The "usual Chi-Square" is $\sum_{\text{all cells}} \frac{(O_i - E_i)^2}{E_i}$ and its degrees of freedom number is (r-1)(c-1) for a table with r rows and c columns. E_i is the expected number for cell i and O_i the observed. E_i is (row total)(column total)/(grand total) so for upper left cell E_i is (6)(8)/13.

	Altered	Frost	
Frequency Percent	No	Yes	Total
No	3 48/13	3 30/13	6
Yes	5 56/13	2 35/13	7
Total	8	5	13