

1. Consider the 928 bivariate measurements of height (y) and midparent height (x).

- (a) Use SAS to compute \bar{y} . Obtain a 95% confidence interval for the population mean height, or the expected height of a randomly sampled person from the population, $E(Y)$. The expression we need is $\bar{y} \pm t(.025, 927)\sqrt{s_y^2/928}$. One could use SAS with PROC MEANS or PROC REG to get the answer.

```
proc reg data=heights;
  model y=/clb;
run;
(Output below)
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	68.08847	0.08266	823.76	<.0001

Variable	DF	Parameter Estimates	
		95% Confidence Limits	Limits
Intercept	1	67.92626	68.25068

(b) Report the following summary statistics:

- i. $\bar{y} = 1/n \sum y_i = 1/n(y_1 + y_2 + \dots + y_n) \bar{y} = 68.09$
- ii. $\bar{x} = 1/n \sum x_i = 1/n(x_1 + x_2 + \dots + x_n) \bar{x} = 68.31$
- iii. $\sum (x_i - \bar{x})(y_i - \bar{y}) S_{xy} = 1911.7$
- iv. $\sum (x_i - \bar{x})y_i S_{xy} = 1911.7$
- v. $\sum x_i(y_i - \bar{y}) S_{xy} = 1911.7$
- vi. $S_{xx} = \sum (x_i - \bar{x})^2 = 2169.4$
- vii. $s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = 3.19$
- viii. $S_{yy} = \sum (y_i - \bar{y})^2 = 5877.2$
- ix. $s_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2 = 6.34$
- x. $r = 0.46$

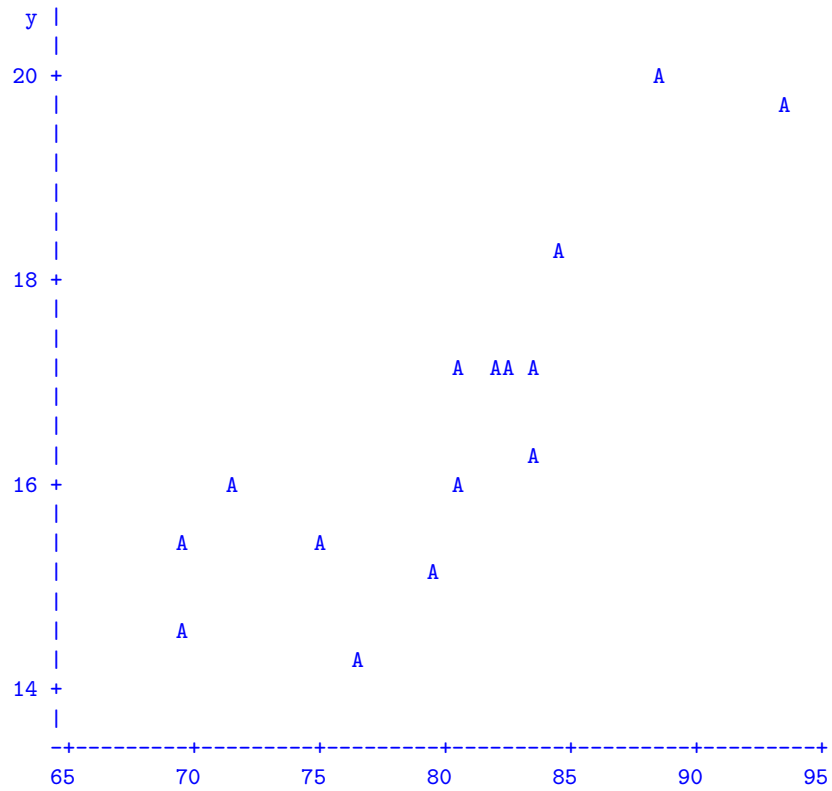
(c) Express the slope from the least squares regression line as a function of r , s_x and s_y above.

$$\hat{\beta}_1 = r \frac{s_y}{s_x} = 0.45 \frac{2.52}{1.79} = 0.65$$

2. Consider the chirp frequency data from lecture notes. These are $n = 15$ bivariate measurements on striped ground crickets.

(a) Obtain a scatterplot of these measurements (sketched or using software).

Plot of y^*x . Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 1 obs had missing values.

- (b) Specify the simple linear regression model for these data. Identify all parameters in the model, providing the interpretation of each.

$$Y_i = \beta_0 + \beta_1 x_i + E_i$$

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

β_0 - intercept

β_1 - slope, or mean increase in chirp frequency per unit increase in temperature

$\sigma^2 = \text{Var}(Y|x)$ - population variance for fixed temperature

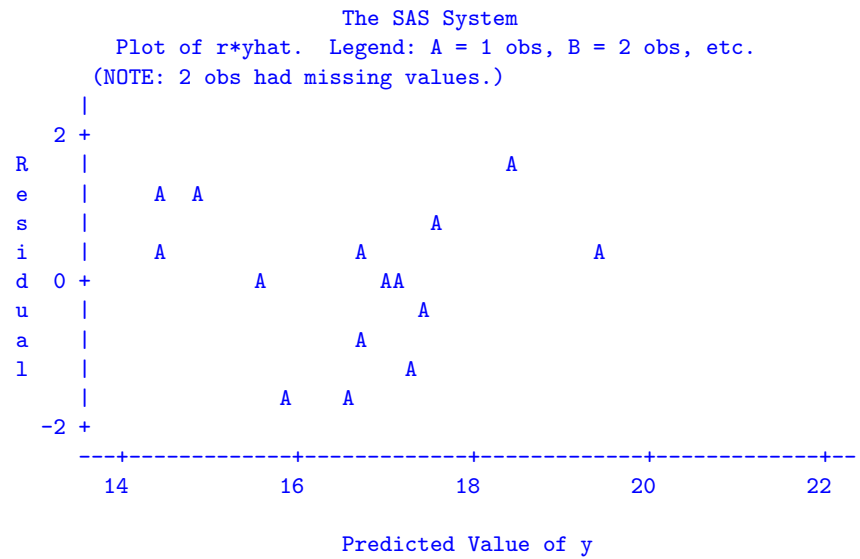
- (c) Explain how the interpretation (and the estimate) of the slope parameter changes if temperature is expressed in Celcius. Since $C = \frac{5}{9}(F - 32)$, the slope will be attenuated by a factor of $5/9$ ($\beta_{1,new} = \frac{5}{9}\beta_1$) and the intercept will also change: $\beta_{0,new} = \beta_0 - \frac{5}{9}(32)$. Same goes for the estimate $\hat{\beta}_{new}$.
- (d) Estimate the mean chirp frequency among crickets in a temperature of $80^\circ F$. Estimate the standard deviation among chirp frequency measurements made at this fixed temperature.

$$\hat{\beta}_0 + \hat{\beta}_1(80) = 16.6 \text{ chirps/second.}$$

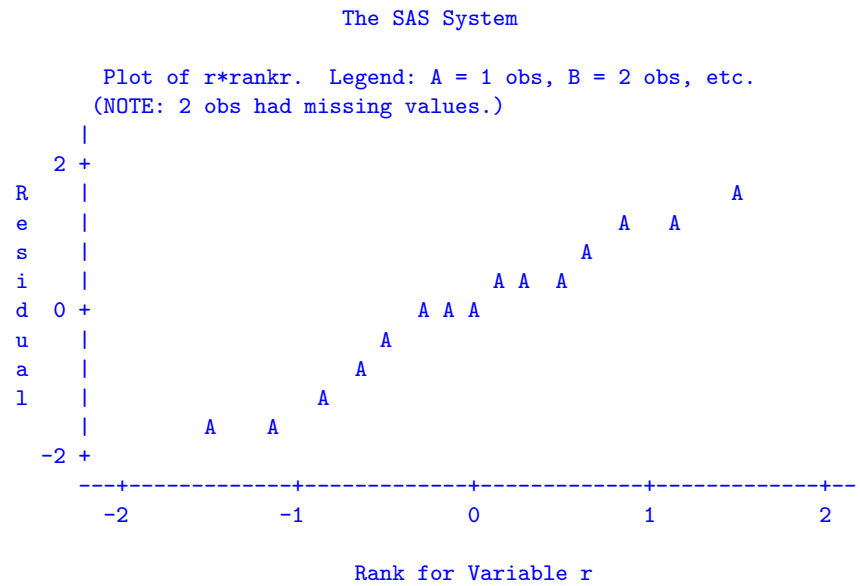
Population s.d. estimated by $\sqrt{MS[E]} = 0.97 \text{ chirps/second.}$

- (e) Estimate the mean chirp frequency among crickets in a temperature of $105^\circ F$. $\hat{\beta}_0 + \hat{\beta}_1(105) = 21.9 \text{ chirps/second.}$ (not a good idea though, since no data in temps this high)
- (f) Report the sum of squared deviations between the fitted values and the average chirp frequency, \bar{y} ? $SS[R] = 28.3$
- (g) What proportion of variance in chirp frequencies is explained by the linear regression model? $r^2 = .70$

(h) Obtain a plot of the residuals against the fitted values.



(i) Obtain a plot of the ordered residuals against the corresponding quantiles from the standard normal distribution.



3. Do the exercises on page 15 of the lecture notes.

- (a) To test $H_0 : \rho = 0$ against $H_0 : \rho \neq 0$ at level $\alpha = 0.05$, we observe $r = -0.45$ and

$$z_{obs}(\rho = 0) = \frac{1}{2}\sqrt{20 - 3} \log \frac{0.55}{1.45} = -1.99$$

which exceeds $z_{0.025} = 1.96$ in absolute value, so that $H_0 : \rho = 0$ is rejected. A 95% confidence interval (which we know won't contain 0 from the test) works out to $(-0.74, -0.009)$.

- (b) (2nd exercise from p.6 packet #2)

$$\begin{aligned} & \Pr(R > 0.7; \rho = 0.6) \\ &= \Pr\left\{\frac{1}{2}\sqrt{n-3} \left(\log \frac{1+R}{1-R} - \log \frac{1+\rho}{1-\rho}\right) > \frac{1}{2}\sqrt{n-3} \left(\log \frac{1+0.7}{1-0.7} - \log \frac{1+\rho}{1-\rho}\right)\right\} \\ &= \Pr(Z > 0.9) \\ &= 0.1828 \end{aligned}$$

4. Resolution 5k Run, Jan, 2004.

- (a) Regarding the bivariate data as a random sample from the population of all local runners, obtain the least squares regression for a model which takes the mean 5k pace to be quadratic in age:

$$\mu(x) = \beta_0 + \beta_1x + \beta_2x^2.$$

PROC REG output:

Parameter Estimates					
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

So, the LS regression is $\hat{\mu}(x) = 11.785 - 0.197x + 0.0029x^2$.

- (b) Report a standard error and 95% confidence interval for the mean pace for runners aged $x = 34$.

$$\hat{\mu}(x = 34) = (1, 34, 34^2)\hat{\beta} = 11.785 - 0.197(34) + 0.0029(34)^2 = 8.48 \text{ minutes/mile}$$

SE given by $\sqrt{(1, 34, 34^2)\hat{\Sigma}(1, 34, 34^2)'} which works out to $\widehat{SE}[\hat{\mu}(x = 34)] = 0.206$. Since $t(0.025, 126) = 1.98$, the 95% confidence interval is$

$$8.48 \pm (1.98)(0.206) \text{ or } (8.08, 8.89) \text{ or } (8 : 05, 8 : 54).$$

- (c) Report the multiple coefficient of determination for the quadratic model.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	113.64513	56.82257	13.23	<.0001
Error	157	674.45060	4.29586		
Corrected Total	159	788.09573			

Root MSE 2.07265 R-Square 0.1442

The output above indicates that the multiple coefficient of determination for this model is $R^2 = 113.6/788.1 = 0.14$.

- (d) The age where the mean pace $\mu(x)$ attains its minimum under the model is given by

$$\theta = \frac{-\beta_1}{2\beta_2}.$$

Propose an estimator of θ . Report the observed value of the estimate. (Check a scatterplot to make sure your estimate is sensible.) Solving $\mu'(x) = 0$ for x yields $x = \frac{-\beta_1}{2\beta_2}$ which can be estimated using

$$\hat{\theta} = \frac{-\hat{\beta}_1}{2\hat{\beta}_2} = \frac{0.197}{2(0.00294)} = 33.5 \text{ (yrs old)} .$$

- (e) Consider reparameterizing the quadratic model in terms of θ and two other parameters rather than $(\beta_0, \beta_1, \beta_2)$. Use PROC NLIN to obtain the least squares estimate of θ along with an approximate standard error.

The NLIN Procedure
Method: Gauss-Newton

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	2	113.6	56.8225	13.23	<.0001
Error	157	674.4	4.2959		
Corrected Total	159	788.1			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
c0	8.4818	0.2057	8.0756	8.8880
c1	0.00294	0.000574	0.00180	0.00407
peak	33.5366	1.9110	29.7620	37.3112

(So, the approximate standard error is 1.91 years.)

5. 11.3bc

- (a) 11.3b: β_1 is the partial slope for the independent variable a , representing the change in $E(y|a, b, c, d, f)$ per unit increase in a , with b, c, d, f held fixed. β_2 is interpreted similarly.
- (b) 11.3c: $\beta_1 + \beta_2 + \beta_3$ represents the change in $E(y|a, b, c, d, f)$ when each of a, b, c is simultaneously increased by one unit with d, f held fixed.

6. 11.11 abdeh

- (a) a) β_1 and β_2 are partial slopes, and their interpretations depend on what other independent variables are held fixed.
- (b) b) $\hat{y} = -0.86 + 0.8x_1 - 0.08x_2$
- (c) d) $F_{obs} = 34.6$, p -value is 0.0012, reject H_0 , conclude there is some linear association between y and either x_1 or x_2 or both.
- (d) e) $\hat{y} = -0.60 + 0.74x_1 - 0.12x_2 + 0.01x_1x_2$
- (e) h) $MS[E]_f = 0.079$, $MS[E]_r = 0.064$

7. 11.12 bdfg

- (a) 11.12b)

$$\begin{aligned}\hat{y} &= 426 - 6.3x_1 + 3.4x_2 \text{ (model 1)} \\ \hat{y} &= -2441 + 41.5x_1 + 30.9x_2 - 0.46x_1x_2 \text{ (model 2)}\end{aligned}$$

- (b) 11.12d) β_3 is tough. Consider the two cases where uncertainty is fixed at either $x_2 = 0$ or $x_2 = 1$. The partial slopes for x_1 , or mean change in stress, y per unit increase in age x_1 in these cases are β_1 and $\beta_1 + \beta_3$ respectively. So, β_3 is the difference in partial slopes for age as uncertainty increases by one unit.
- (c) 11.12f) $F_{obs} = 4.665$, $p = 0.0205$, reject $H_0 : \beta_1 = \beta_2 = 0$ and conclude that at least one of the partial slopes is not plausibly 0. There is an association between y and at least one of the ind. var.s. Model 1 is better than a constant mean model.
- (d) 11.12g) $F_{obs} = 5.26$, $p = 0.0073$, reject $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ and conclude that at least one of the three partial slopes is not plausibly 0. Model 2 is better than no model at all.

- 8. 11.22: (refer to example 11.17, p.507): Which of models 2-6 is nested in model 1? Models 2,3,4, and 5 are all nested in model 1.

9. 11.25

(a) abcdefg

- a) $SS[R]_f = SS[Tot] - SS[E]_f = 608995 - 339972 = 269023$

- b)

Source	df	SS	MS
Regression	5	269023	53805
Error	19	339972	17893
Total	24	608995	

- c) $F_{obs} = 53805/17893 = 3.0$. At level $\alpha = 0.05$, $F(0.05, 5, 19) = 2.74$, so we reject H_0 ($p = 0.036$)

- d)

$$R(\beta_3, \beta_4, \beta_5 | \beta_0, \beta_1, \beta_2) = SS[R]_f - SS[R]_r = 269023 - 181369 = 87654$$

This is a measure of variation in y explained by the quadratic model that is not already accounted for by (additive) model 1. It is a difference in regression sums of squares.

- e) $F_{obs} = \frac{87654/3}{17893} = 1.63$, not significant at level $\alpha = 0.05$ since $F(0.95, 3, 19) = 3.13$.

- f)

$$R(\beta_4, \beta_5 | \beta_0, \beta_1, \beta_2, \beta_3) = SS[R]_f - SS[R]_r = 269023 - 261164 = 7859$$

This is a measure of variation in y explained by quadratic model that is not already accounted for by (non-additive) model 2. It is a difference in regression sums of squares.

- g) $F_{obs} = \frac{7859/2}{17893} = 0.2$, not significant, conclude that the quadratic model doesn't improve models 1 or 2. Model 2 is the best one we've investigated.

(b) Report the $MS[E]$ and the coefficients of determination for models 1,2 and the full quadratic model.

Model	$MS[E]$	mult. coef. of determ.
1	19438	$r^2_{y-1,2} = 0.30$
2	16563	$r^2_{y-1,2,3} = 0.43$
quadratic	17893	$r^2_{y-1,2,3,4,5} = 0.44$

(c) True or false: the coefficient of determination for any of these regression models is the square of the correlation coefficient computed from observed (y) and predicted (\hat{y}) values of the response variable. True, coefficient of determination is the squared correlation coefficient for the bivariate (y_i, \hat{y}_i)

10. See Example 11.18 (p. 511)

- (a) Using simple linear regression, report the p -value for a test of no linear association between $\log(\text{leaf burning time})$, y , and potassium percentage, x_3 . The p -value for a test that the slope $\beta_3 = 0$ is $p = .3429$ (from fitting the reduced model).
- (b) Using multiple linear regression, report the p -value for a test of no linear association between $\log(\text{leaf burning time})$, y , and potassium percentage, x_3 , after adjusting for dependence on nitrogen (x_1) and chlorine (x_2) percentage. The p -value for a test that the partial slope $\beta_2 = 0$ is $p < .0001$ (from fitting the full model).
- (c) Estimate the mean difference in $\log(\text{leaf burning time})$ between two populations; both have fixed x_1 and x_2 , but potassium is increased from $x_3 = 6.0$ to $x_3 = 7.0$. Report a standard error for this difference. $\hat{\beta}_3 = 0.21$ ($SE = 0.04$)
- (d) Consider the difference in the slopes for x_1 and x_2 , for fixed x_3 . Estimate $\beta_1 - \beta_2$ and report a 95% confidence interval.

$$\begin{aligned} (0, 1, -1, 0)\hat{\beta} &= -0.09 \\ SE^2 &= (0, 1, -1, 0)\hat{\Sigma} \begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \end{pmatrix} \\ &= (0.11)^2 \quad \text{on } df = 26 \end{aligned}$$

since $t(.025, 26) = 2.06$, this yields a 95% c.i. $-0.09 \pm 2.06(.11)$ or $(-.32, .13)$.

- (e) Estimate the mean leaf-burning time (on the log scale) among tobacco leaves with $x_1 = 3\%$ nitrogen, $x_2 = 1\%$ chlorine and $x_3 = 7\%$ potassium. Report a standard error and confidence interval for this mean.

$$\begin{aligned} (1, 3, 1, 7)\hat{\beta} &= 1.24 \\ SE^2 &= (1, 3, 1, 7)\hat{\Sigma} \begin{pmatrix} 1 \\ 3 \\ 1 \\ 7 \end{pmatrix} \\ &= (.10)^2 \end{aligned}$$

Confidence limits given by $\pm t(.025, 26) \times .10$ or $(1.03, 1.45)$.

- (f) Estimate the standard deviation of this population of leaves. $\sqrt{MS[E]} = 0.21$
- (g) Consider a single leaf from this subpopulation with $(x_1, x_2, x_3) = (3, 1, 7)$. Estimate the leaf-burning time on the log-scale. Use a

procedure for reporting an interval that will cover the individual leaf's burning time (log-scale) with 95% confidence. Prediction is $(1, 3, 1, 7)\hat{\beta} = 1.24$ and prediction interval is

$$(1, 3, 1, 7)\hat{\beta} \pm t(0.025, 26)\sqrt{MS[E] + SE((1, 3, 1, 7)\hat{\beta})}$$

or $1.24 \pm 2.06\sqrt{0.046 + (.10)^2}$ or $(0.75, 1.73)$ (much wider than confidence limits for mean)

(h) 11.27, parts a,b and d

- (a) $s_{obs}^2 = SS[Total]/(n - 1) = 6.68952/29 = 0.23$ and $s_{pred}^2 = \sum(\hat{y}_i - \bar{y})^2/(n - 1) = 5.50473/29 = 0.1898$. (By inspection of ANOVA table, p. 523.)
- (b) $s_{pred}^2/s_{obs}^2 = 0.1898/0.23 = 0.823 = R_{y \cdot 1,2,3}^2$.
- (d) Recall that the LS estimate of β minimizes $SS[E]$ and hence maximizes $SS[R]$ with $SS[Total]$ fixed. So, $r^2 = SS[R]/SS[Total]$ is maximized by the least squares estimate $\hat{\beta}$ and r^2 will be smaller if any other fit (like $\hat{\beta}' = (3, -0.3, -0.6, 0.5)$) is used.