

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, $E(Y)$, of a randomly sampled person from the population.
 - (b) Report the following summary statistics:
 - i. $\bar{y} = (1/n) \sum y_i = (1/n)(y_1 + y_2 + \cdots + y_n)$
 - ii. $\bar{x} = (1/n) \sum x_i = (1/n)(x_1 + x_2 + \cdots + x_n)$
 - iii. $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$
 - iv. $s_{xy} = (1/(n-1)) \sum (x_i - \bar{x})(y_i - \bar{y})$
 - v. $(1/(n-1)) \sum (x_i - \bar{x})y_i$
 - vi. $(1/(n-1)) \sum x_i(y_i - \bar{y})$
 - vii. $S_{xx} = \sum (x_i - \bar{x})^2$
 - viii. $s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$
 - ix. $S_{yy} = \sum (y_i - \bar{y})^2$
 - x. $s_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$
 - xi. r
 - (c) Express the slope from the least squares regression line as a function of r , s_x and s_y above.
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).
 - (b) Specify the simple linear regression model for these data. Identify all parameters in the model, providing the interpretation of each.
 - (c) Explain how the interpretation (and the estimate) of the slope parameter changes if temperature is expressed in Celcius.
 - (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.
 - (e) Estimate the mean chirp frequency among crickets in a temperature of $105^\circ F$.
 - (f) Report the sum of squared deviations between the fitted values and the average chirp frequency, \bar{y} ?
 - (g) What proportion of variance in chirp frequencies is explained by the linear regression model?

- (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 like those on page 15 of the lecture notes:
- (a) Obtain an approximate 95% confidence interval for the population correlation coefficient ρ when a bivariate random sample of size $n = 20$ results in a sample correlation coefficient of $r_{xy} = -0.45$. Also, conduct a test of $H_0 : \rho = 0$.
 - (b) Suppose that two random variables X and Y have correlation $\rho = 0.6$. (That is, the correlation among two quantities in an entire population is $E[(X - \mu_x)(Y - \mu_y)] = 0.6$.) What is the probability that a random sample of $n = 30$ bivariate observations will yield a sample correlation coefficient that exceeds 0.7. Find $\Pr(R > 0.7; \rho = 0.6)$.

4. Resolution 5k Run, Jan, 2004. (“resrun04.data”)

- (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:

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

- (b) Report a standard error and 95% confidence interval for the mean pace for runners aged $x = 34$.
- (c) Report the multiple coefficient of determination for the quadratic model.
- (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.)

- (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.

5. 11.3bc

6. 11.11 abdeh

7. 11.12 bdfg

8. 11.22: (refer to example 11.17, p.507): Which of models 2-6 is nested in model 1?

9. 11.25

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

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 .
- (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.
- (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.
- (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.
- (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.
- (f) Estimate the standard deviation of this population of leaves.
- (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.
- (h) 11.27, parts a,b and d