

ST512, Summer Session II, 2008
Quiz 1 - solutions

1. (20 points) Four scatterplots appear below.
- (a) For each sample correlation coefficient, circle the letter of the corresponding plot (one letter per correlation coefficient).
- $r = 0.58$ (A B C **D**)
 - $r = 0.95$ (A **B** C D)
 - $r = -0.41$ (**A** B C D)
 - $r = 0.13$ (A B **C** D)
- (b) For plot *B*, what proportion of variation on the vertical (y) axis would be explained by a linear regression on the variable on the horizontal (x) axis (a linear regression of y on x)? $r^2 = 0.95^2 = 0.9$

2. (40 points) The crop yield of grapes (y , in tons/acre) harvested in August can be predicted using counts of berry clusters (x), measured in July. Data and output (from PROC REG) for a simple linear regression analysis are given at the end of the problem.
- (a) Give the simple linear model being considered here, establishing appropriate notation and specifying distributional assumptions.

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \quad i = 1, \dots, 7$$

where Y_i and x_i denote yield and cluster count, respectively, for year i . E_i denote i.i.d. $N(0, \sigma^2)$ errors.

- (b) Is there evidence that cluster count makes for a useful predictor? Find in the output a p -value for a test of the hypothesis that the mean yield does not depend on cluster count. Draw a brief conclusion.
The F -ratio for testing $\beta_1 = 0$ is given by

$$F = \frac{MS(Reg)}{MS(E)} = \frac{2.7}{0.18} = 15$$

and the p -value is .0116. (Both are given directly in the output.) One could also use the test statistic $Z = 0.5\sqrt{n-3} \log \frac{1+r}{1-r}$.

- (c) As the number of clusters increases by one unit, how much does the yield increase, on average? Report a 95% confidence interval for this quantity.
 $\hat{\beta}_1 = 0.0468$ and its standard error $SE = 0.012$ are given directly in the output. The appropriate t multiplier is $t(.025, 5) = 2.57$ leading to the 95% confidence interval of $.0468 \pm .031$.
- (d) As the number of clusters increases by 10 units, how much does the yield increase, on average? Report an estimate.

$$10\hat{\beta}_1 = .468$$

(e) The residual for the observation in 1975 is hidden. Deduce its value.

$$e_4 = y_4 - \hat{y}_4 = 4.2 - 4.0797 = .1203$$

(f) A hurricane wiped out the 1972 crop. For insurance purposes, an estimate of what the yield would have been is needed. Given that the cluster count was $x = 125.0$, provide a 95% prediction interval for what the yield would have been.

$$\hat{y} \pm t(.025, 5) \sqrt{MSE \left(1 + \frac{1}{n} + \frac{(125 - \bar{x})^2}{S_{xx}} \right)}$$

or

$$5.37 \pm 2.57 \sqrt{.18 \left(1 + \frac{1}{7} + \frac{(125 - 109.2)^2}{1239.2} \right)}$$

or

$$5.37 \pm 2.57(.49)$$

or

$$5.37 \pm 1.26$$

year	cluscount	yield
1971	116.37	5.6
1973	82.77	3.2
1974	110.68	4.5
1975	97.50	4.2
1976	115.88	5.2
1978	125.24	4.8
1979	116.15	4.9

The REG Procedure
Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	7.00000	1.00000	7.00000	0	0
cluscount	764.59000	109.22714	84753	206.53926	14.37147
yield	32.40000	4.62857	153.58000	0.60238	0.77613

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.71454	2.71454	15.08	0.0116
Error	5	0.89975	0.17995		
Corrected Total	6	3.61429			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.48355	1.32595	-0.36	0.7303
cluscount	1	0.04680	0.01205	3.88	0.0116

Output Statistics

Obs	Dependent Variable	Predicted Value	Residual
1	5.6000	4.9629	0.6371
2	3.2000	3.3903	-0.1903
3	4.5000	4.6966	-0.1966
4	4.2000	4.0797	xxxxxxx
5	5.2000	4.9399	0.2601
6	4.8000	5.3780	-0.5780
7	4.9000	4.9526	-0.0526

3. (a) Using the linear regression, estimate the mean weight among all kids who have a height of $x = 158.9 \text{ cm}$ along with a standard error.

$$\hat{\mu} = -57.67 + 0.66(158.9) \approx 47.2$$

$$S_{xx} = (n - 1)s_x^2 = 39(10.78)^2 = 4532$$

$$SE = \sqrt{MS(E)\left(\frac{1}{40} + \frac{(158.9 - 158.9)^2}{S_{xx}}\right)} = 1.14$$

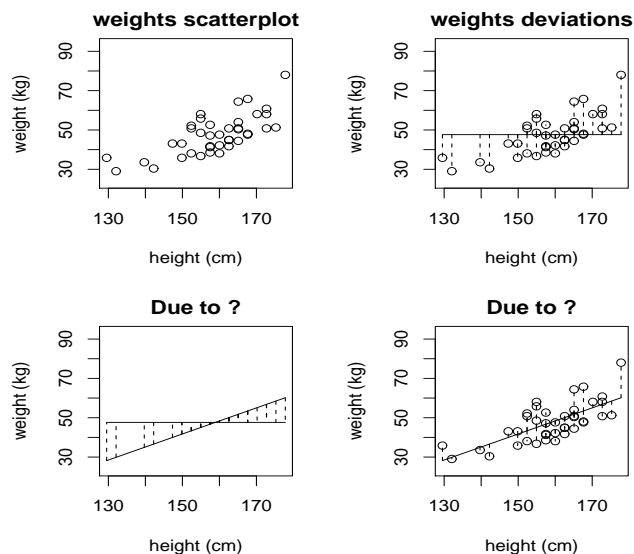
- (b) Report the standard error associated with the slope estimate, $\hat{\beta}_1 = 0.66$.

$$SE(\hat{\beta}_1) = \sqrt{MSE/S_{xx}} = \sqrt{51.7/4532} = 0.11$$

- (c) Deduce the value of the sample correlation coefficient, r .

$$\hat{\beta}_1 = r \frac{s_y}{s_x} \implies r = \hat{\beta}_1 \frac{s_x}{s_y} = 0.66 \frac{10.78}{10.07} = 0.71$$

- (d) In the matrix of graphs below, the titles for the two plots at the bottom are incomplete. Complete each with a single word that describes the vertical distances depicted in these bottom two plots. (Plot has been fixed.) Graph at left due to regression, graph at right due to error.



- (e) Consider the population of shorter kids with a height of $x = 150 \text{ cm}$. Estimate the standard deviation of weights in this population.

$$\sqrt{MS(E)} = \sqrt{51.7} = 7.2$$