STAT 3A03 Applied Regression With SAS

Assignment 1 Due at **5:00pm** on Thursday September 28, 2017.

Dropboxes for assignment submission are outside HH-105. Your assignment **MUST** be deposited in the appropriate dropbox for your lab section.

N.B. Late assignments will not be accepted

Q. 1 Suppose that $(y_1, x_1), \ldots, (y_n, x_n)$ is a dataset to which we fit a simple linear regression. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least squares estimates of the intercept and slope and let r be the sample correlation coefficient

a) Show that

$$\text{se}(\hat{\beta}_1) \;=\; \sqrt{\frac{S_{yy}(1-r^2)}{(n-2)S_{xx}}}$$

b) Hence show that

$$\frac{\hat{\beta}_1}{\text{s.e}(\hat{\beta}_1)} \;=\; \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

c) Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ $(i = 1, \ldots, n)$ be the fitted values from the simple linear regression. Now consider the pairs $(y_1, \hat{y}_1), \ldots, (y_n, \hat{y}_n)$. Suppose we wish to fit the model

$$y \;=\; \alpha_0 + \alpha_1 \hat{y} + \varepsilon$$

Show that the least squares line for this model has intercept 0 and slope 1. *Hint: First find the average of the fitted values and then write $S_{y\hat{y}}$ and $S_{\hat{y}\hat{y}}$ in terms of the original $\hat{\beta}_1, S_{xy}$ and S_{xx}. The result follows.*

Q. 2 An experiment was conducted at a U.S. university to determine how blood alcohol content (BAC) changes with consumption of regular bottled beers. 24 male students were randomly assigned to drink a certain number of standard beers (between 1 and 8) after which their blood alcohol level was tested by a police officer. Let y be the final BAC recorded and x be the number of beers consumed. Some summary statistics are:

$$\begin{array}{lll} \sum x_i \;=\; 108 & \sum y_i \;=\; 1.84 & \\ \sum x_i^2 \;=\; 612 & \sum y_i^2 \;=\; 0.1839 & \sum x_i y_i \;=\; 10.47 \end{array}$$

a) Find the least squares estimates of the intercept and slope for the simple linear regression model $y = \beta_0 + \beta_1 x_i + \varepsilon$.

b) Show that the sum of squared residuals from the fitted model can be written as

$$\text{SSE} \;=\; S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

and hence calculate an unbiased estimate of the error variance, σ^2

c) Give the standard errors for your estimates in (a).

d) Construct 95% confidence intervals for the true intercept and true slope.

e) What conclusions would you draw from your results?

Q. 3 Can early season snowfall from September 1 until December 31 predict snowfall in the remainder of the season from January 1 to June 30? To look at this question, records of snowfall from Ft Collins, Colorado were examined and a linear regression was fit in SAS. Here is the output of `PROC REG`.

```
                The REG Procedure
                  Model: MODEL1
            Dependent Variable: Late

            Number of Observations Read          93
            Number of Observations Used          93

            Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F
Model                     1      453.57591      453.57591       2.41    0.1239
Error                    91          17119      188.11903
Corrected Total          92          17572

Root MSE             13.71565    R-Square     0.0258
Dependent Mean       32.04301    Adj R-Sq     0.0151
Coeff Var            42.80387

            Parameter Estimates

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1       28.63580        2.61488      10.95      <.0001
Early         1        0.20349        0.13105       1.55      0.1239
```

a) What is the fitted regression line?

b) What is the unbiased estimate of the error variance σ^2?

c) Give confidence intervals for the true intercept and slope.

d) What percentage of the variability in the late snowfall can be explained by early snowfall?

e) What is the correlation coefficient between the two variables Early and Late?

f) Do you think, from this data, that early season snowfall is a good predictor of late season snow in Ft. Collins. Explain your answer giving statistical reasons based on the above output.

Q. 4 The dataset Anscombe.txt contains 8 variables which were constructed by the statistician Frank Anscombe to demonstrate that fitting a linear regression should not be done without examining the data. Use SAS to answer the following questions.

a) Find the values of the sample correlation between

(i) Y1 and X1

(ii) Y2 and X2

(iii) Y3 and X3

(iv) Y4 and X4

b) For each of the pairs of variables in Part (a) fit a linear regression to predict the Y value from the X value and give the values of the estimated intercepts and slopes.

c) Draw a scatterplot of each of the pairs of variables in Part (a) and add the fitted line to each plot.

d) For which of the pairs is linear regression appropriate? For those in which it is not appropriate, explain why.