Regression Analysis
Sample Midterm Questions

Instructions: These sample questions are provided to help you study for the midterm, and anticipate the kind of questions that will be asked. These are not the only kind of questions that may be asked. You should study all the course material. Answers to these sample questions will not be provided. All topics were covered in the lecture, slides or lab handouts.

1. Write out the equation for predicting on outcome on Y using the sample regression coefficients

   Y'_i = ?

2. What are the two main uses of regression?

3. Name, and describe, the four types of confidence intervals we have learned in class

4. Write out the normal errors regression model. Name, and describe, the five major assumptions of this model, and identify the portion of the model equation that embodies each assumption.

5. Write "True" or "False" for each of the following statements:
   a. Regression analysis is appropriate whenever you have two con,nuous variables.
   b. For the hypothesis H_0: $\beta_1 = 0$, p-values based on t and F statistics are the same.
   c. The total sum of squares in Y can be decomposed into two parts: regression sum of squares and residual sum of squares.
   d. If the hypothesis H_0: $\beta_1 = 0$ is rejected, then a causal relationship exists between X and Y.
   e. In the regression equation $Y = \beta_0 + \beta_1 X + \varepsilon$, β_0 is the predicted value of Y when X = 0.
   f. The confidence interval for predicting an average value of Y is the same as the confidence interval for predicting a new observation of Y.
   g. The predicted value of Y depends on whether you are predicting the mean response or an individual response.
   h. The assumption of normal distribution of errors is not necessary for the least squares estimators (b_0 and b_1) to be valid.
   i. The Sum of Squares Regression (SSReg) can be separated into two parts: SS Total and SS Error.
   j. You can use ANOVA to test the hypothesis that the slope equals zero.

6. A researcher is interested in whether the number of pieces of candy a child receives on Halloween is related to the age of the child (in years). He collects data from 13 children, and finds the following summary values:

$\bar{X} = 9,\ \bar{Y} = 18.8,\ \sum(X - \bar{X})^2 = 182,\ \ \sum(Y - \bar{Y})^2 = 358,\ \sum(X - \bar{X})(Y - \bar{Y}) = 169$

$MSE = 18.307, se(b_1) = 0.3172$

   a. (10 pts) Calculate the least squares estimates of β_0 and β_1.

   b. (5 pts) Interpret, in words, the values you obtained for the intercept and slope.

   c. (5 pts) Use the regression equation to predict the average amount of candy that children who are 8 years old will receive. Calculate

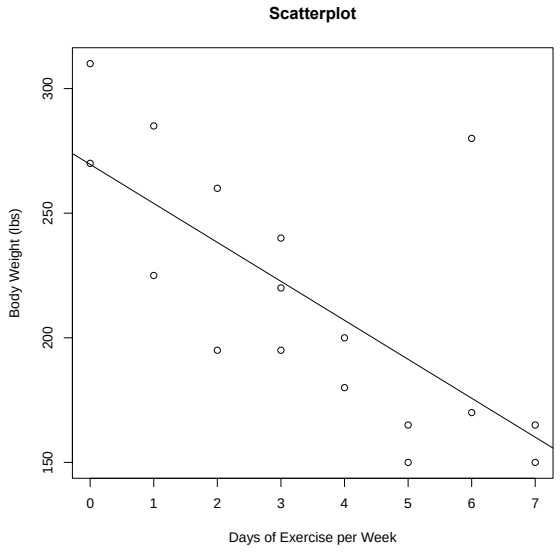
   d. (20 pts) Test whether the slope is zero against the alternative that the slope is not zero. Use $t_{critical} = 2.20$. State the null and alternative hypotheses, calculate the test statistic, compare it to the critical value, and state your conclusion.

7. A researcher is interested in knowing if individuals who exercise tend to have lower body weight. She studied 17 overweight individuals and measured the number of days each week they exercised, and their body weight (in lbs).

   Here is the R code and output analyzing these data.

```
> days=c(0, 0, 1, 1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7)
> weight=c(310, 270, 285, 225, 260, 195, 240, 220, 195, 180, 200, 165, 150, 170,
280, 165, 150)
> model2<-lm(weight~days)
> summary(model2)

Call:
lm(formula = weight ~ days)

Residuals:
    Min      1Q  Median      3Q     Max
-43.293 -27.015  -5.737  17.346 104.263

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  269.571     16.838   16.01 7.71e-11 ***
days         -15.639      4.084   -3.83  0.00164 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.48 on 15 degrees of freedom
Multiple R-squared:  0.4944,  Adjusted R-squared:  0.4607
F-statistic: 14.67 on 1 and 15 DF,  p-value: 0.001641

> anova(model2)
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value   Pr(>F)
days       1  20602 20602.0  14.666 0.001641 **
Residuals 15  21072  1404.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> confint(model2,level=.95)
                2.5 %     97.5 %
(Intercept) 233.68210 305.458962
days        -24.34319  -6.934743
> predict(model2,data.frame(days=4),interval="prediction",level=.9)
       fit      lwr      upr
1 207.0147 139.2989 274.7304
```

```
> plot(days,weight,xlab="Days of Exercise per Week",ylab="Body Weight
(lbs)",main="Scatterplot")
> abline(model2)
```



a. (5 pts) Write out the regression equation obtained from these data, and interpret it in words.

b. (10 pts) State the hypotheses for the t and F statistics, and interpret the p-values. What can you conclude from these test results?

c. (5 pts) Find and interpret the value for the coefficient of determination.

d. (5 pts) Find the 95% confidence interval for the slope. Write a statement that interprets this interval.

e. (5 pts) Find the predicted value for a person who exercises 4 days per week, along with a 90% prediction interval. Write a statement that interprets both the prediction and the interval.