Regression Background Test **Name:** _____
SELF ADMINISTERED
CLOSED BOOK/NOTE
7/24/2017
Time Limit: 120 Minutes Grade _____

This exam contains 5 pages (including this cover page) and 6 questions. Total of points is 100.

Question 1 and 2 require you to code in R and are open-Internet, but you are allowed to search for coding-related issues only (in other words, you may not search for the solutions of the answers). Other questions are closed-book, closed-note.

Good Luck!

Grade Table

Question	Points	Score
1	20	
2	12	
3	20	
4	20	
5	8	
6	20	
Total:	100	

1. (20 points) A forester is interested in relating the height of a tree (in feet), `Height` , to the stem diameter at breast height (in inches) `DBH` . Observations on height and stem diameter were recorded for 198 48-year-old trees. The data are in the file `tree.txt` at the course web site. The first column consists of a plot `id` number, the second column is the tree number, the third column is height, and the fourth column is diameter at breast height (in inches).
   (a) (2 points) Does a simple linear regression model fit well in estimating Height from DBH? Explain why.
   (b) (5 points) Consider fitting a linear model for the subset of trees for which the diameter at breast height exceeds 15 inches.

   ```
   > trees = read.table(file.choose(),header=TRUE)
   > treeBig15.lm = lm(Height ~ DBH, subset = (DBH > 15), data = trees)
   ```

   (c) (8 points) Perform checks of residual plots to assess the model.
   (d) (2 points) Report 95% confidence intervals for the model slope and intercept.
   (e) (3 points) Use the regression model to predict heights of trees with DBH of measures of 5.5 and 7.5 inches with a 95% prediction intervals.

(f) (5 points) Find the heights of the trees in the data with DBH of 5.5 and 7.5. Do the heights fall into the prediction intervals? Briefly summarize your observations and discuss the reasoning behind what you find.

2. (12 points) The file `glue.txt` contains a data set with the results of an experiment on the dry sheer strength (in pounds per square inch) of birch plywood, bonded with 5 different resin glues A, B, C, D, and E. Eight pieces of plywood were tested with each glue type. Let $\mu_A, ..., \mu_E$ be the unknown true population mean strengths for the corresponding treatments. Analyze the data with a linear model. Summarize the linear model using both the summary function in `R` and the `ANOVA` function.

   (a) (7 points) The `summary` function provides a p-value for each of several regression parameters. In each case, state the hypothesis that is being tested and provide an interpretation of the regression parameter in terms of the unknown population means.

   (b) (5 points) The ANOVA table has a single p-value. State the hypothesis that is being tested here. How does this hypothesis differ form the hypotheses in part (a)?

3. (20 points) **This is a concept-based question, answer each question with no more than THREE (3) sentences**.

   (a) (6 points) Mathematically, give the sampling covariance for OLS estimates. What does sampling covariance mean?

   (b) (4 points) Now write out the OLS regression slope estimates $\hat{\beta}_1$, ending with two cross-products, make sure to show your work.

   (c) (4 points) Explain what the following equation means.

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2$$

   (d) (6 points) Assuming the errors are uncorrelated and have zero means and common variance σ^2. Give an unbiased estimate of σ^2. Explain briefly the degree of freedom for the following regression. Why?

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3^2 + e$$

4. (20 points) Given the estimated slope and its standard error for Seattle snowfall data over 85 years with the impact of early season snowfall on the amount of late season snowfall.

$$\hat{\beta}_1 = 0.2485, SE(\hat{\beta}_1) = 0.1198$$

   (a) (4 points) Write a complete null hypothesis and alternative hypothesis.

   (b) (6 points) Calculate the appropriate statistics for this hypothesis test, carefully and clearly reason your conclusions.

   (c) (10 points) Write an ANOVA table including the df, SS, MS, F and p-value for this test. Information: SXX = 10954.069, SXY = 2229.014, SYY = 17572.048. Show your work.

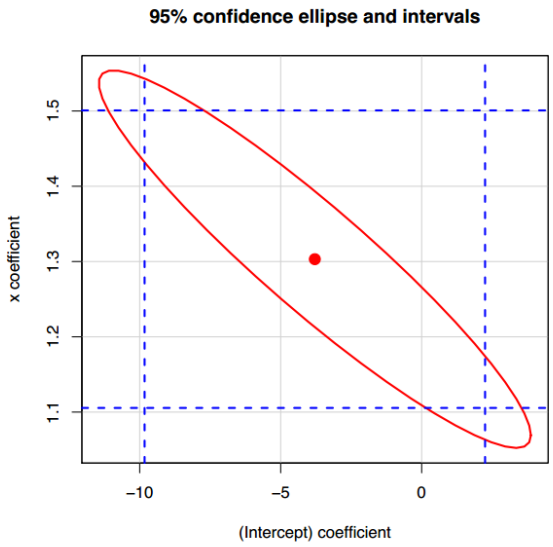

Figure 1: Output1

5. (8 points) Explain this 95% confidence ellipse and intervals. Make sure to discuss what does the region, dashed lines, and the lines do not enclose the ellipse exactly?

6. (20 points) Consider the following dataset: Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

   Consider variables: fert – standardized fertility measure (Y)

   ex – % of draftees receiving highest mark on army examination (X1)

   agr – % of males involved in agriculture as occupation (X2)

   Using ex as a measure of education, We are mainly interested in investigating the effects of the education on fertility.

   Figure 2-4 are the outputs you need for this question.

   (a) (9 points) Assess the effects of 'agr' on education, and fertility. What is the role here of education on both fertility and agriculture? Support your answers with evidence from the output.

   (b) (6 points) Can we interpret that agriculture have effects on fertility? After adjusting the 'agr for the $\hat{\beta}_1$. It changed from -1.01 to -1.20. What could you conclude based on this change.

   (c) (5 points) If the correlation between education and agriculture is -0.872. Should 'agr' be placed in the model? Is agriculture a causally prior variable to education? Why?

Regressing *fert* on *ex*, we obtain:

```
Call:
lm(formula = fert ~ ex)

Residuals:
     Min       1Q   Median       3Q      Max
-25.9375  -6.0044  -0.3393   7.9239  19.7399

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.8185     3.2576  26.651  < 2e-16
ex           -1.0113     0.1782  -5.675 9.45e-07

Residual standard error: 9.642 on 45 degrees of freedom
Multiple R-squared:  0.4172,        Adjusted R-squared:  0.4042
F-statistic: 32.21 on 1 and 45 DF,  p-value: 9.45e-07
```

Figure 2: Output1

```
Call:
lm(formula = fert ~ ex + agr)

Residuals:
     Min       1Q   Median       3Q      Max
-26.4089  -6.3234   0.0577   6.3134  20.8937

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  94.6097     7.8271  12.087 1.41e-15
ex           -1.1950     0.2445  -4.887 1.40e-05
agr          -0.0940     0.0859  -1.094     0.28

Residual standard error: 9.621 on 44 degrees of freedom
Multiple R-squared:  0.4326,        Adjusted R-squared:  0.4068
F-statistic: 16.77 on 2 and 44 DF,  p-value: 3.85e-06
```

Figure 3: Output2

```
Call:
lm(formula = fert ~ agr)

Residuals:
     Min       1Q   Median       3Q      Max
-25.5374  -7.8685  -0.6362   9.0464  24.4858

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 60.30438    4.25126  14.185   <2e-16
agr          0.19420    0.07671   2.532   0.0149

Residual standard error: 11.82 on 45 degrees of freedom
Multiple R-squared:  0.1247,        Adjusted R-squared:  0.1052
F-statistic: 6.409 on 1 and 45 DF,  p-value: 0.01492
```

Figure 4: Output3