Regression Background Test **Name:** _____
Self-Administered
Solution
7/3/2017
Time Limit: 120 Minutes Grade _____

This exam contains 7 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) <2 extra pts> Does a simple linear regression model fit well in estimating Height from DBH? Explain why.

   After fitting linear model Height=62.5+1.5(DBH), we draw the residual and Q-Q plots. The residual plot is curvature, which indicates some nonlinearity. [2 pt (this would give you bonus 1 pt)] For the Q-Q plot, residuals do not follow the line approximately, hence the residuals are non-normality. [2pt (this would give you bonus 1 pt)] Other answers are possible fine, running residual or Q-Q plot is not mandatory [or, 2 pts].

   (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)
```

For those trees with 15 inches or larger diameter, the linear model is: Height= 83.15+0.46(DBH). [2 points]

```
>trees15.lm = lm(Height~DBH, subset=(DBH>15),data=tree
s)
>summary(trees15.lm)
Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.1549   2.3162   35.90  < 2e-16 ***
DBH          0.4607  0.1160      3.97 0.000111 ***
Residual standard error: 4.173 on 151 degrees of freedom
Multiple R-Squared: 0.09453,
Adjusted R-squared: 0.08853
F-statistic: 15.76 on 1 and 151 DF,  p-value: 0.0001107
```

[3 points]

(c) (8 points) Perform checks of residual plots to assess the model.

The residual plot does not have any pattern except there are three points (99, 174 and 177) with big residuals. [1 point] The Q-Q plot has a long-tailed error. [1 point] Hence it is possible that these three points are outliers. [1 point] Generally, the linear regression assumptions are fulfilled well. [2 point] For this question, as long as you have two plots to make inference, it will be fine [3 points]

(d) (2 points) Report 95% confidence intervals for the model slope and intercept.

The confidencd interval for intercept is: (78.58, 87.73) [1 pt] The confidencd interval for slope is: (0.23,0.69) [1 pt]

(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.

[2 points, 1pt each] For DBH=5.5, 95% prediction interval is: (76.8, 94.6) For DBH=7.5, 95% prediction interval is: (77.9, 95.3)

R code: (1 point)

```
>predict(trees15.lm, data.frame(DBH=c(5.5,7.5)), interval="prediction")
    fit       lwr      upr
1 85.68897 76.79571 94.58224
2 86.61047 77.87522 95.34571
```

(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.

here is one tree (ID=589) with DBH=5.5 and one tree (ID=507) with DBH=7.5. [2 points] However, the heights do not fall into the prediction intervals. One possible reason is the model is fitted by considering the subset of trees for which the diameter at breast height exceeds 15 inches. While DBH =5.5 and 7.5 are less than 15. Using the regression equation to predict values of the dependent variable outside the range of the independent variable is not recommended since we have no evidence that the same linear relationship exists outside the observed range [3 pts]

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.

The summary table can be seen below:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)  482.250      6.210  77.659  < 2e-16 ***
GlueB         -7.500      8.782  -0.854 0.398904
GlueC         28.000      8.782   3.188 0.003011 **
GlueD         35.625      8.782   4.057 0.000265 ***
GlueE         23.750      8.782   2.704 0.010494 *    [3 points]
```

According to this summary, each t-value is testing whether that particular glue type is significantly different than glue A. (NOTE: we are comparing all to glue A since alphabetically glue A appears first. If we want to test a different comparison, we must recode the dummy variables.) [2 pts] It is found that glues C through E differ significantly than glue A in the positive direction (since the coefficients are positive) while glue B does not appear to increase or decrease the bonding strength significantly as compared to glue A. [2 points]

(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)?

```
Analysis of Variance Table
Response: Strength
Df  Sum Sq Mean Sq F value   Pr(>F)
Glue       4 11179.6  2794.9  9.0597 3.93e-05 ***
Residuals 35 10797.4   308.5                          [2points]
```

According to the F-statistic here, we are testing whether or not the mean bonding strength of the 5 types of glues are all equal. With a p-value of 3.93e-05, we are confident that there is a statistically significant difference in the bonding strength of the 5 glue types. This hypothesis differs from that in part (a) because here we are testing whether all are equal, whereas in part (a) we were testing if glues B through E are significantly different than glue A.[3 points]

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?

$$s_{XY} = \frac{SXY}{n-1}$$

(3 points) The sample covariance matrix is a square matrix whose i, j element is the sample covariance (an estimate of the population covariance) between the sets of ob-

served values of two of the variables and whose i, i element is the sample variance of the observed values of one of the variables. (3 points)

(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.

$$\hat{\beta}_1 = \frac{SXY}{SXX}$$

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

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

One possible way to estimate β_0 and β_1 is to find such values of β_0 and β_1 that minimize the residual sum of squares

(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?

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

$$\hat{\sigma^2} = \frac{RSS}{d.f.}$$

(3 points). $d.f. = n - 4$. Four parameters (include α) (3 points)

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.

The answer to this question may vary, you can set any null hypothesis: $H_0 : \beta_1 = 0; H_1 : \beta_1 \neq 0$, etc... as long as it makes sense

(b) (6 points) Calculate the appropriate statistics for this hypothesis test, carefully and clearly reason your conclusions. The answer to this question may vary, but the appropriate test will be t-test (3 points). The t-statistic follows $t(83)$ distribution under the null hypothesis (3 points)

(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.

Residual: df=85-1 = 84. $SSR = MSR = \frac{SXY^2}{SXX} = 453.57, SSE = SYY - SSR = 17118.478, MSE = \frac{SSE}{df} = 203.79, F = \frac{MSR}{MSE} = 2.226$ For p-value, check the f-table.

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?

   Red dot means the fitted b_0, b_1 coefficient, but the dashed lines show confidence intervals for each parameter. (4 points) For each pair of b_0 and b_1 coefficient, the 95% confidence interval is bounded by the ellipse. Notice that these lines do not enclose the ellipse exactly (if they did, they would be jointly correct confidence intervals). (4 points)

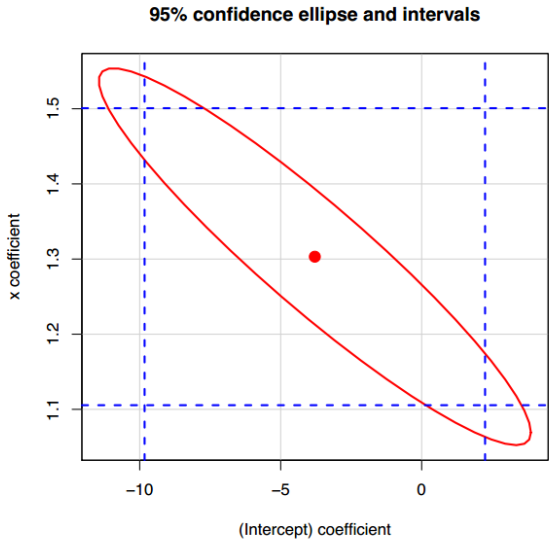

Figure 1: Output1

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. It might be a mediating variable to education on fertility. According to Output 2, agr was not significant in the multivariate regression. We assume that agr influences education (i.e., ex), and that they both potentially influence fertility. We found that agr was correlated with fert, but no longer so when ex was included.

   (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.

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

No, but on the other hand, you cannot say agriculture has NO effects on fertility. According to the correlation coefficient change, we can say that by ignoring agr, we underestimate the effect of ex on fertility: the absolute value of the estimate is smaller than when we adjust for agr.

(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?

No. Agriculture is no longer significant when education is included in the model. The entire effect of agriculture on fertility is mediated by education, after controlling for education. Agriculture has no additional effect of its own. No. Based on this information, education is actually a prior of agriculture, NOT vice versa.

```
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