Name _____

Spring 2018 **STAT 401 Eng** **Final exam**
(100 points)

Instructions:

- Full credit will be given only if you show your work.
- The questions are not necessarily ordered from easiest to hardest.
- You are allowed to use any resource except aid from another individual.
- Aid from another individual will automatically earn you a 0.
- Feel free to tear off the last page. There is no need to turn it in.

Regression assumptions State the 4 simple linear regression assumptions and describe one way that each of those assumptions could be violated. Just saying "This assumption is not true." will earn you no points. (5 pts each)

-
-
-
-

Material hardness

Researchers at Iowa State University are attempting to make a material as hard as diamond. Using a cubic boron nitride composite, the researchers use a laser followed by a water beam to etch the composite and thereby inscrease its hardness. The researchers ran a randomized block design experiment to determine the effect of water distance on hardness (in gigaPascals [GPa]). They used 3 different composite *pucks* (the material looks like a small hockey puck) with the water jet shooting at 4 different distances (in millimeters [mm]) behind the laser. Use the file `hardness.csv`, to answer the following questions.

1. Is this experiment complete? Explain why or why not. (2 pts)

2. Is this experiment replicated? Explain why or why not. (2 pts)

3. Fit a regression model with hardness as the response, puck as a categorical explanatory variable and water distance and water distance squared as continuous explanatory variables. Write the code that you used here. (6 pts)

4. Conduct an F-test to determine whether there are differences due to puck. Provide the F statistic, pvalue, and an interpretation. (6 pts)

5. Provide an estimate for the water distance that provides the maximum hardness. (4 pts)

Simple linear regression

The following table contains summary statistics for a response variable (y) and explanatory variable (x). Assume the model $y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$. Using these summary statistics and the

Variable	N	Mean	SD
x	102	5.45	2.05
y	102	-8.21	3.94

estimated correlation between x and y is -0.68, calculate the following (4 pts each):

1. Maximum likelihood estimate (MLE) for β_1

2. MLE for β_0

3. Coefficient of variation R^2

4. MLE for σ^2

5. Standard error for $\hat{\beta}_1$

Hard Drive Failure

Backblaze, a company that provides computer backups, provides data on hard drive failures. On the `Hard Drive Failure - R Code` page, there is an analysis of failure times (in years) for hard drives of various brands at capacities of 2 TB (terabytes), 4 TB, and 8 TB.

1. Write down the model that was used in this analysis. Make sure to define any notation you introduce. (20 pts)

2. Provide an interpretation for the following quantities. You may transform these quantities if it makes interpretation easier. (4 pts each)

   (a) 0.308224

   (b) 0.02185

   (c) -0.099064

3. Construct a 95% confidence interval for the multiplicative effect of brand WDC compared to brand Seagate (while holding capacity constant) on the median failure time. (4 pts)

4. Describe the null hypothesis for the ANOVA line the begins with `brand`. (4 pts)

(intentionally left blank - scratch paper)

Hard Drive Failure - R Code (Feel free to remove this page)

```
> table(hd_data$brand, hd_data$capacity)

           2  4  8
  Seagate 10  5  9
  Toshiba  8 11 14
  HGST    14  9  8
  WDC     11 11 13
> m <- lm(log(failure_time) ~ I(capacity-2) + brand, data = hd_data)
>
> summary(m)

Call:
lm(formula = log(failure_time) ~ I(capacity - 2) + brand, data = hd_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44276 -0.12799  0.02158  0.10753  0.49993

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.308224   0.042427   7.265 4.36e-11 ***
I(capacity - 2)  0.021850   0.006704   3.259  0.00146 **
brandToshiba    -0.099064   0.050700  -1.954  0.05308 .
brandHGST       -0.034235   0.051377  -0.666  0.50649
brandWDC         0.062923   0.049972   1.259  0.21046
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1885 on 118 degrees of freedom
Multiple R-squared:  0.1636,Adjusted R-squared:  0.1353
F-statistic:  5.77 on 4 and 118 DF,  p-value: 0.0002816
> anova(m)
Analysis of Variance Table

Response: log(failure_time)
                 Df Sum Sq Mean Sq F value   Pr(>F)
I(capacity - 2)   1 0.3587 0.35871 10.0957 0.001897 **
brand             3 0.4614 0.15380  4.3288 0.006212 **
Residuals       118 4.1926 0.03553
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```