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)

- Answer: Assumption: Normality of the errors

  Violations: right-skewed errors, left-skewed errors, heavy-tailed errors

- Answer: Assumption: Constant variance of the errors

  Violations: Variance is increasing with fitted value

- Answer: Assumption: Independence of the errors

  Violations: Clustering, temporal data, spatial data

- Answer: Assumption: Linearity

  Violations: Quadratic relationship

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)

   Answer: Yes, every combination of puck and water distance exists.

```
table(hardness_data$puck, hardness_data$water_distance)

##
##     1 2 3 4
##   1 2 2 2 2
##   2 2 2 2 2
##   3 2 2 2 2
```

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

   Answer: Yes, every combination of puck and water distance exists twice.

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)

   Answer:

```
m <- lm(hardness ~ factor(puck) + water_distance + I(water_distance^2),
        data = hardness_data)
```

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

   Answer:

```
drop1(m, test="F")[2,]
```

```
## Single term deletions
##
## Model:
## hardness ~ factor(puck) + water_distance + I(water_distance^2)
##              Df Sum of Sq    RSS    AIC F value  Pr(>F)
## factor(puck)  2    72.962 214.83 58.603  4.8858 0.01941 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   The small p-value indicates the data are incompatible with a regression model that is quadratic in water distance, i.e. when puck is not included in the model that was fit.

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

   Answer:

```
as.numeric(-coef(m)[4]/(2*coef(m)[5]))
```

```
## [1] 2.357895
```

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

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

Answer: We are given the following quantities:

```
n; m_X; s_X; m_Y; s_Y; rxy

## [1] 102
## [1] 5.45
## [1] 2.05
## [1] -8.21
## [1] 3.94
## [1] -0.68
```

1. Maximum likelihood estimate (MLE) for β_1

   Answer:

   $$\beta_1 = SXY/SXX$$

   ```
   SXY <- rxy*s_X*s_Y*(n-1)
   SXX <- s_X^2*(n-1)
   (beta1 = SXY/SXX)

   ## [1] -1.306927
   ```

2. MLE for β_0

   Answer:

   $$\beta_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

   ```
   (beta0 = m_Y - beta1 * m_X)

   ## [1] -1.087249
   ```

3. Coefficient of variation R^2

   Answer:

   $$R^2 = r_{XY}^2$$

```
(R2 <- rxy^2)

## [1] 0.4624
```

4. MLE for σ^2

   Answer:

   $$SSE = SST(1 - R^2) = SYY(1 - R^2)$$

   and

   $$\hat{\sigma}^2 = SSE/(n - 2)$$

```
SYY = s_Y^2*(n-1)
SSE = SYY*(1-R2)
(sigma2 = SSE/(n-2))

## [1] 8.428942
```

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

   Answer:

   $$SE(\beta_1) = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}}$$

```
sqrt(sigma2)*sqrt(1/((n-1)*s_X^2))

## [1] 0.1409198
```

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)

   Answer: For hard drive $i = 1, \ldots, n$, let

   - Y_i be the log failure time (years),
   - C_i be the capacity minus 2TB,
   - T_i be an indicator of the brand Toshiba,
   - H_i be an indicator of the brand HGST, and
   - W_i be an indicator of the brand WDC.

   The model is

   $$Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$$

   with

   $$\mu_i = \beta_0 + \beta_1 C_i + \beta_2 T_i + \beta_3 H_i + \beta_4 W_i.$$

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

   (a) 0.308224

   Answer: The median failure time in years for 2TB Seagate hard drives is $e^{0.308224} \approx 1.36$.

   (b) 0.02185

   Answer: For each 1TB increase in hard drive capacity while keeping brand constant, the median hard failure time increases by $100(e^{0.02185} - 1) \approx 2.21\%$.

   (c) -0.099064

   Answer: While keeping capcity constant, median failure time on Toshiba hard drives is $100(e^{-0.099064} - 1) \approx 9.43\%$ sooner than on Seagate drives.

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)

   Answer:

```
exp(0.062923 + c(-1,1)*qt(.975,118)*0.049972)

## [1] 0.9646064 1.1757205
```

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

   Answer: The null hypothesis here is the model where $\beta_2 = \beta_3 = \beta_4 = 0$. That is, the model without brand, but with capacity.

(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
```