Name ____

Spring 2021 **STAT 587-3** **Final exam**

(53 points)

Instructions:

1. Full credit will be given only if you show your work.
2. The questions are not necessarily ordered from easiest to hardest.
3. You are allowed to use any resource except aid from another individual.
4. Aid from another individual will automatically earn you a 0.

1. Short calculations

   (a) If the correlation is -0.8, what is the coefficient of determination?

   Answer:

```
rho = -0.8
rho^2
## [1] 0.64
```

   (b) If the sample average of the response variable is 1, the sample average of the explanatory variable is -1, and the slope of the simple linear regression line is 2, what is the intercept?

   Answer:

```
1 - (2)*(-1)
## [1] 3
```

   (c) If there are 20 observations, 3 explanatory variables (not including the intercept), and the residual sum of squares is 48, what is the MLE for the residual variance?

   Answer:

```
48/(20-4)
## [1] 3
```

   (d) If a regression equation is

$$E[Y] = -3 + 2X_1 + 4X_2$$

   what is the expected response when $X_1 = -2$ and $X_2 = 1$?

   Answer:

```
-3 + 2*-2 + 4*1
## [1] -3
```

   (e) In simple linear regression with 10 observations, what is the *t*-critical value associated with a 95% confidence interval?

   Answer:

```
qt(1-.05/2, df = 8)
## [1] 2.306004
```

2. TRUE/FALSE questions

   (a) If the coefficient of determination is 0.81, then the correlation is 0.9.

   Answer: FALSE

   (b) If I run a regression of X on Y and find the slope to be 1, then the slope for the regression of Y on X is 1.

   Answer: FALSE

   (c) If the correlation between X and Y is 0.7, then the correlation between Y and X is 0.7.

   Answer: TRUE

   (d) The coefficient of determination never decreases as additional explanatory variables are added to the model.

   Answer: TRUE

   (e) A high correlation means a there is a strong causal relationship between two variables.

   Answer: FALSE

   (f) In simple linear regression if the p-value for the slope is close to zero, then there is high correlation between the response variable and the explanatory variable. (The null hypothesis for this p-value is that the slope is equal to zero.)

   Answer: FALSE

   (g) Assume the same model, data, predictive location, and confidence level, the prediction interval is always wider than the confidence interval.

   Answer: TRUE

3. An experiment was performed to understand the effect of building material and span length on bridge strength. Answer the following questions based on the output below.

```
## Analysis of Variance Table
##
## Response: strength
##                          Df  Sum Sq Mean Sq F value  Pr(>F)
## material                  2  5.0792  2.5396  2.4708 0.12994
## factor(length)            3  6.8473  2.2824  2.2206 0.14304
## material:factor(length)  5 17.3723  3.4745  3.3804 0.04294 *
## Residuals                11 11.3062  1.0278
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) How many different materials were used?

Answer:

```
length(unique(bridge$material))
## [1] 3
```

(b) How many different lengths were used?

Answer:

```
length(unique(bridge$length))
## [1] 4
```

(c) Is the design complete?

Answer: No, the interaction degrees of freedom is less than the product of the degrees of freedom of the underlying factors.

(d) What is the estimate for σ^2 in this model?

Answer:

```
anova(m)$`Mean Sq`[4]
## [1] 1.027832
```

(e) For the `factor(length)` line,

i. what is the reduced model?
Answer: The model that only has material.

ii. what is the full model?
Answer: The model that has both material and length (as a categorical variable).

iii. what model is used for the estimate of σ^2?
Answer: The model with material, length, and their interaction.

iv. using a significance level of 0.05, what decision would you make?
Answer: Fail to reject the null hypothesis

v. using a significance level of 0.05, what conclusion would you draw?
Answer: There is insignificant evidence to conclude that the data are incompatible with the model that only includes material.

4. The file `imports.csv` contains the total value of imports (in billions) into the United States from Canada and Mexico by Air and Ship for each month from 2018 to the current. The variable `covid` indicates the months from March 2020 until now. Answer the following questions based on the model fit in this code.

```
months = c("January","February","March","April","May","June","July","August",
           "September","October","November","December")
d = read.csv("imports.csv")
d$month = factor(d$month, levels = months)
m = lm(log(value) ~ I(year-2021) + month + covid, data = d)
```

(a) Provide the point estimate, 95% confidence interval, and interpretation for the following quantities:

i. Intercept

Answer:

```
coef(m)[1]
## (Intercept)
##    1.843897
confint(m)[1,]
##    2.5 %   97.5 %
## 1.659051 2.028744
```

The median imports in January 2021 in non-COVID times is estimated to be $6.3 billion.

ii. Coefficient for ‘I(year-2021)‘

Answer:

```
coef(m)[2]
## I(year - 2021)
##     0.07975201
confint(m)[2,]
##       2.5 %      97.5 %
## 0.004288731 0.155215288
```

Every year (aside from COVID), median import value is estimated to increase by 8.3%.

iii. Coefficient for the indicator of a COVID month

Answer:

```
coef(m)[14]
##  covidyes
## -0.306147
confint(m)[14,]
##      2.5 %     97.5 %
## -0.4507822 -0.1615119
```

COVID is estimated to have decreased import value by 26%.

(b) What proportion (not percentage) of the variability in the logarithm of value is explained by this model? (1 pt)

Answer:

```
summary(m)$r.squared
## [1] 0.6276271
```

(c) Provide an estimate and 95% **prediction** interval for the value (in billions of dollars) of imports in March 2021 (assuming we are still in COVID times).

Answer:

```
nd = data.frame(year = 2021, month = "March", covid = "yes")
exp(predict(m, nd, interval = "prediction"))
##        fit      lwr      upr
## 1 5.605248 4.235118 7.418639
```

(d) Conduct an F-test for comparing the additive model to the model that does not include `Month`. Provide the F-statistic, numerator and denominator degrees of freedom, and p-value.

Answer:

```
as.data.frame(drop1(m, test = "F"))[3,]
##       Df Sum of Sq       RSS       AIC  F value    Pr(>F)
## month 11  0.236809 0.5353986 -155.9685 1.730383 0.1264302
m$df.residual # denominator degrees of freedom
## [1] 24
```

(e) Conduct an F-test for comparing the additive model to the model that includes an interaction between month and covid. Provide the F-statistic, numerator and denominator degrees of freedom, and p-value. In addition, describe the support in the data for including this interaction.

Answer:

```
mI = lm(log(value) ~ I(year-2021) + month*covid, data = d)
anova(m, mI)
## Analysis of Variance Table
##
## Model 1: log(value) ~ I(year - 2021) + month + covid
## Model 2: log(value) ~ I(year - 2021) + month * covid
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1     24 0.298590
## 2     13 0.052724 11   0.24587 5.5111 0.00247 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This small p-value indicates evidence against the model with no interaction, but we still need to check model assumptions before determining that the interaction should be included.