Carson Slater STA 5381 Homework #1

1. A researcher claims the length and weight of low birth rate infants is positively linearly related. He selects a SRS of 20 low birth weight infants and measures the length (in cm) and weight (in grams) of each. Below is some of the output from his analysis.

```
Coefficients:
              Estimate   Std. Error t value   Pr(>|t|)
(Intercept)  -1793.347    321.153   -5.584   < 0.000
Length          77.680      8.576    9.058   < 0.000
```

   Which of the following is the best interpretation of the estimated slope?

   (a) As the length of low birth weight babies increases by 1 cm, the mean weight of the babies increases by 77.68 grams.

   (b) As the length of low birth weight babies increases by 1 cm, the mean weight of the babies decreases by 1,793.35 grams.

   (c) The length of low birth weight babies increases by 77.68 cm for each 1 gram increase in weight.

   (d) As the length of low birth weight babies increases by 1 cm, the mean weight of the babies decreases by 77.68 grams.

   (e) Two of the above are correct. If so, mark which two are correct.

2. A teacher is interested in determining whether the number of hours of sleep a student had the night before an exam and how many hours the student studied for the exam are related to the student's score on the exam. She determines how many hours of sleep each student got along with how many hours each student studied and records these values with the student's exam score. What type of analysis should she consider to analyze this data?

   (a) Simple linear regression

   (b) Multiple linear regression

   (c) One-way ANOVA

   (d) Two-way ANOVA

   (e) ANCOVA

3. Woodpeckers are a valuable forest asset because they provide nest and roost holes for other birds, and they prey on many forest insect species. The article "Artificial Trees as a Cavity Substrate for Woodpeckers" (*Journal of Wildlife Management*, 1983: 790–798) reported a study of how woodpeckers behaved when provided with a plastic cylinder. The ambient temperature in Celsius was used an explanatory variable for the cavity depth that the woodpeckers excavated in the cylinders. After fitting a least-squares regression line to the data (with both a slope and an intercept), the R^2 value was 76.7%. Interpret this value.

   (a) There is a positive correlation between ambient air temperature and cavity depth.

   (b) 76.7% of the variability in cavity depth can be explained by the ambient temperature.

   (c) 76.7% of the variability in ambient air temperature can be explained by cavity depth.

(d) Not all of the variability in cavity depth can be explained by the ambient air temperature.

(e) Two of the above are true. If so, mark which two are correct.

4. The assumption that the data are independent will be made for most of this course. Which of the following in the list below represents a dataset in which knowing the response of one individual will give you information about another individual (i.e., the data are NOT independent)? Circle every choice that describes dependent data.

   (a) The wind speed (in miles/hour) at a tower 10 meters above ground level collected every hour for 5 days.

   (b) The blood pressure (systolic) of 100 randomly selected patients with Type II Diabetes.

   (c) The temperature (in °F) collected at 500 randomly selected cities in Colorado at 5 am on December 10th.

   (d) The heights (in inches) of 50 2-year Quaking Aspen saplings planted in separate buckets but all exposed to the same light, temperature, water, and soil conditions.

   (e) The number of text messages sent by 400 randomly selected high school students in one week.

5. A research institution is studying the risk of lung cancer in people who have been exposed to second-hand smoke, claiming that being exposed to second-hand smoke puts a person at greater risk for lung cancer. They took 100 random people exposed to heavy second-hand smoke, 100 random people exposed to light second-hand smoke, and 100 random people exposed to no second-hand smoke. They followed these people for 25 years and determined how many developed lung cancer. They found no significant association between these two variables. They submitted their findings to a well-known journal, but the editors of the journal rejected their paper because "the p-value is too large, and therefore the results are not interesting." What is wrong with the reasoning of the editors?

   (a) Only studies in which large p-values are obtained are interesting.

   (b) Even though the researchers did not find evidence for the alternative hypothesis, it doesn't mean that the null hypothesis is actually true. And if the null hypothesis is true, then this would be interesting to know as well.

   (c) They are concentrating on the results of the outcome when it is the set-up of the study that is fundamentally flawed.

   (d) The editors should really be concerned with the sample size of the study; it is too small to be conclusive.

   (e) Nothing is wrong with the editors' reasoning; it is only ever of interest to prove the research hypothesis is true.

Multiple Choice Answers

1. A
2. B
3. B
4. A
5. B

6. The following data are downloaded from the Colorado Department of Public Health and Environment `https://cdphe.maps.arcgis.com/apps/opsdashboard/index.html#/d79cf93c3938470ca4bcc4823328946b`. It contains four main variables:

   - The date
   - The particular utility
   - SARS CoV2 copies of RNA (measured as RNA/liter of water) in wastewater
   - The number of new Covid-19 cases

   SARS-CoV-2 is the virus that causes COVID-19, and RNA is the genetic material in each copy of the virus. SARS-CoV-2 copies per liter is one measure of how much of the virus is in the wastewater, expressed as a concentration. Studies have shown that individuals who develop COVID-19 often shed detectable SARS-CoV-2 RNA from their systems before, during, and after their infection, so higher levels of SARS-CoV-2 RNA in wastewater can indicate a rise in cases in a community. Many universities have used this method to monitor the wastewater from residence halls to obtain an early warning of a disease outbreak.

   (a) Make a scatter plot with the log of the number of cases and the log of the number of SARS CoV-2 RNA copies. Use good labels. Which do you think should be the independent variable, and which should be the dependent variable? Explain your answer.

```
plot(covid$log_sars_rna, covid$log_cases,
  pch = 19,
  xlab = "Log COVID-19 Copies per Liter",
  ylab = "Log New COVID-19 Cases (per Day)",
  col = alpha("black", 0.20))
  title("Log COVID-19 Cases per Log COVID-19 Copies", cex.main = 1.2)
```

   See Figure 1 for the scatter plot. Here the dependent variable would be the number of new cases in a given day, as studies have shown that people shed detectable COVID-19 copies before, during, and after. It appears that because people can shed it before they contract or even show symptoms, the number of copies being shed in wastewater could be a good predictor of how many new cases are being detected each day.

   (b) Fit a linear model using 'lm()' for the same data you just plotted. Overlay the model's fitted regression line.

```
mod1 <- lm(log_cases ~ log_sars_rna, data = covid)
pred <- predict(mod1, covid)
plot(covid$log_sars_rna, covid$log_cases,
  pch = 19,
  xlab = "Log COVID-19 Copies per Liter",
  ylab = "Log New COVID-19 Cases (per Day)",
  col = alpha("black", 0.20))
  lines(covid$log_sars_rna, pred, lwd = 2, col = "blue")
  title("Log COVID-19 Cases per Log COVID-19 Copies", cex.main = 1.2)
```

   See Figure 2 for scatter plot with fitted model.

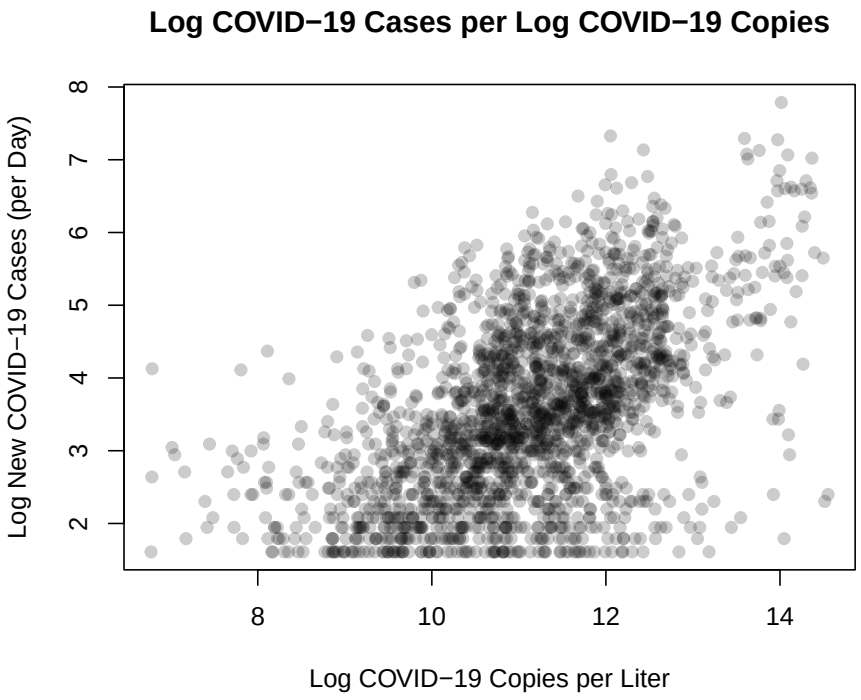

Figure 1: Scatter plot of the number of new COVID-19 cases in a given day compared to the number of COVID-19 copies per liter in wastewater.

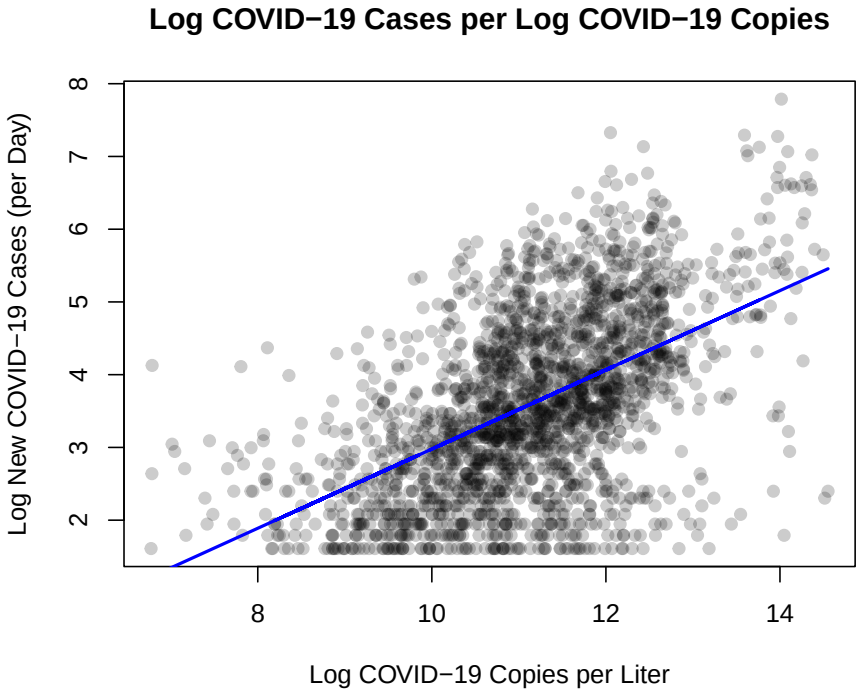

Figure 2: The sample plot in Figure 1, but with the fitted regression line in it.

(c) Investigate the fit of the model by answering the following questions:

Does it *look* like a good fit?
This plot model does not look like it is a good fit, mostly because these data appear to have a lot of variability, and it does not appear to be constant throughout the domain of COVID-19 copies per liter. A large cluster in the lower middle part of the training data appears to be pulling the line lower than it should be.

What is the value of R^2? Is that good or bad (or can you tell)?
The $R^2 = 0.2871$. This is *not* a good thing. It basically means that roughly 29% of the variability in the new COVID-19 case counts is accounted for by the COVID-19 copies per liter.

Is the slope coefficient of the model significant?

```
Call:
lm(formula = log_cases ~ log_sars_rna, data = covid)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3895 -0.6388 -0.0686  0.6959  3.2322

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.46707    0.20891  -11.81   <2e-16 ***
log_sars_rna  0.54445    0.01864   29.21   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.036 on 2115 degrees of freedom
Multiple R-squared:  0.2875,Adjusted R-squared:  0.2871
F-statistic: 853.4 on 1 and 2115 DF,  p-value: < 2.2e-16
```

The slope of the line is significant for all reasonable values of α with a p-value $\leq 2 \times 10^{-16}$.

Explain what the two coefficients mean (including their sign – positive or negative). In our context, the intercept represents the estimated value of the dependent variable when the independent variable is equal to 1 after log transformation. The coefficient for `log_sars_rna` implies that every 1% change in the COVID-19 copies per liter is associated with approximately a 0.544% change in the number of new cases discovered on that day.

Do the following problems from the Sheather (2009) textbook's chapter 2:

7. (a) Here we can find a 95% confidence interval for β_1.

```
playbill <- read.delim("playbill.txt",
                       sep =",", header = TRUE, dec =".")

# Fit the model
playbill_mod <- lm(CurrentWeek ~ LastWeek,
                   data = playbill)

# Standard Error
n <- nrow(playbill)
b1hat <- playbill_mod$coef[2]
Sxx   <- sum((playbill$LastWeek - mean(playbill$LastWeek))^2)

sigma2_hat <- sum(summary(playbill_mod)$residuals^2)/(n-2)

# 95% CI
lower <- b1hat - qt(0.975, n - 2)*sqrt(sigma2_hat/Sxx)
upper <- b1hat + qt(0.975, n - 2)*sqrt(sigma2_hat/Sxx)

cat("(", lower, " , ", upper, ")")

## ( 0.9514971  ,  1.012666 )
```

Our 95% confidence interval is from $(0.951, 1.013)$. Since our 95% confidence interval includes the value of 1, we can consider it plausible that the true value of $\beta_1 = 1$.

(b) To test $H_0 : \beta_0 = 10000$ where $H_a : \beta_0 \neq 10000$, we can find the test statistic and compute a p-value.

```
# (b) Hypothesis Test
b0hat <- playbill_mod$coef[1]
Sxx   <- sum((playbill$LastWeek - mean(playbill$LastWeek))^2)

TS <- (b0hat - 10000)/sqrt(sigma2_hat*(1 / n + mean(playbill$LastWeek)^2/Sxx))

(pval_2sided <- 2*(pt(TS, n - 2, lower.tail = TRUE)) |> unname())

## [1] 0.7517807
```

Obtaining a p-value of 0.7518 means for all reasonable values for α, we would fail to reject H_0 in favor of H_a.

(c) We form a 95% prediction interval for gross box office results.

```
# (c)
pred <- playbill_mod$coef[1] + playbill_mod$coef[2]*400000
piME <- qt(0.975, n - 2)*sqrt(sigma2_hat)*
  sqrt(1 + 1/n + (400000 - mean(playbill$LastWeek))^2/Sxx)
```

```r
# 95 % PI
lower <- round((pred - piME), 3) |> unname(); upper <- round((pred + piME), 3) |
lower; upper
```

```
## [1] 359832.8
## [1] 439442.2
```

Hence, the 95% prediction interval for gross box office results would be (359832.8, 439442.2). We would argue that \$450,000 would not be a feasible value for the gross box office whent he prior week was \$400,000. This is because a \$450,000 is not contained within the 95% prediction interval, which considers the most probably ranges for the $(n+1)^{\text{th}}$ observation.

(d) This particular rule does not consider the steady, yet slow growth in gross box office results, but honestly the slope of the model is small enough that the week before would be a good estimate. Just know the linear model suggests a statistically significant positive (yet very small) association.

8. (a) We can minimize the residual sum of squares to find $\hat{\beta}$.

$$
\begin{aligned}
RSS &= \sum_{i=1}^{n} \epsilon_i^2 \\
&= \sum_{i=1}^{n} (y_i - \beta x_i)^2 \\
\implies 0 &= \frac{\partial}{\partial \beta} \sum_{i=1}^{n} (y_i - \beta x_i)^2 \\
&= -2 \sum_{i=1}^{n} x_i (y_i - \beta x_i) \\
\implies 0 &= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \beta x_i \\
\sum_{i=1}^{n} \beta x_i &= \sum_{i=1}^{n} x_i y_i \\
\beta &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}
\end{aligned}
$$

(b) We can show the unbiasedness of β, find its variance, and derive its sampling distribution.

$$\begin{aligned}
E[\hat{\beta}|X] &= E\left[\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right] \\
&= \sum_{i=1}^n E\left[\frac{y_i}{x_i}\right] \\
&= \sum_{i=1}^n E\left[\frac{\beta x_i + \epsilon}{x_i}\right] \\
&= \sum_{i=1}^n E\left[\frac{\beta x_i}{x_i}\right] + E[\epsilon] \\
&= \beta.
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\beta}|X) &= \text{Var}\left(\sum_{i=1}^n \frac{y_i}{x_i}\right) \\
&= \sum_{i=1}^n \frac{1}{x_i^2}\text{Var}\left(y_i\right) \\
&= \sigma \sum_{i=1}^n \frac{1}{x_i^2}.
\end{aligned}$$

Because β is a linear combination of normally distributed R.V.'s we have that $\hat{\beta}|X \sim \mathcal{N}\left(\beta, \sigma \sum_{i=1}^n \frac{1}{x_i^2}\right)$.

9. In this particular case, the residual sum of squares (RSS) for Model 1 would be less than the RSS of Model 2 because the magnitude of the residuals are smaller in model 1. Likewise, the sum of squares regression (SSReg) for Model 1 would be greater than the SSReg for Model 2, as the predicted Y values look further from the average Y values from the data in Model 1 than in Model 2. So the correct answer in the book would be (d).

10. (a) To show the SST = SSReg + SSE, we can write

$$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}) = (y_i - \bar{y}) - (\hat{\beta}_{1i} - \hat{\beta}_1 \bar{x}) = (y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x}).$$

(b) Then, we can also show

$$\hat{y}_i - \bar{y} = \hat{\beta}_1 x_i - \hat{\beta}\bar{x} = \hat{\beta}_1(x_i - \bar{x}).$$

(c) It follows that

$$\begin{aligned}
\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \beta_1 x_i)\hat{\beta}_1\hat{(}x_i - \bar{x}) \\
&= \hat{\beta}_1 \left(\sum_{i=1}^n y_i(x_i - \bar{x}) - \underbrace{\beta_0 \sum_{i=1}^n (x_i - \bar{x})}_{=0} - \hat{\beta}_1 \sum_{i=1}^n x_i(x_i - \bar{x}) \right).
\end{aligned}$$

We can then show

$$
\begin{aligned}
&= \hat{\beta}_1 \left(\sum_{i=1}^{n} y_i (x_i - \bar{x}) - \hat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) \right) \\
&= \hat{\beta}_1 (S_{XY} - \hat{\beta}_1 S_{XX}) \\
&= \hat{\beta}_1 (S_{XY} - S_{XY}) = 0.
\end{aligned}
$$