Homework 4 Solutions

Due: Thursday 2/20/20 by 8:30am

Rubric:

- Maximum of 2 points each for 1. and 2., determined as follows:
  - 0 points for no solutions whatsoever or incomplete solutions;
  - 1 point for solutions provided for each part, but at least one incorrect solution;
  - 2 points for correct solutions to each part;
- Maximum of 3 points for 3. and 4., determined as follows:
  - 0 points for no solutions whatsoever or `R` output only;
  - 1 point for an honest effort but very few correct answers or `R` output only plus a figure;
  - 2 points for mostly correct answers but at least one substantial issue;
  - 3 points for nearly/exactly correct.

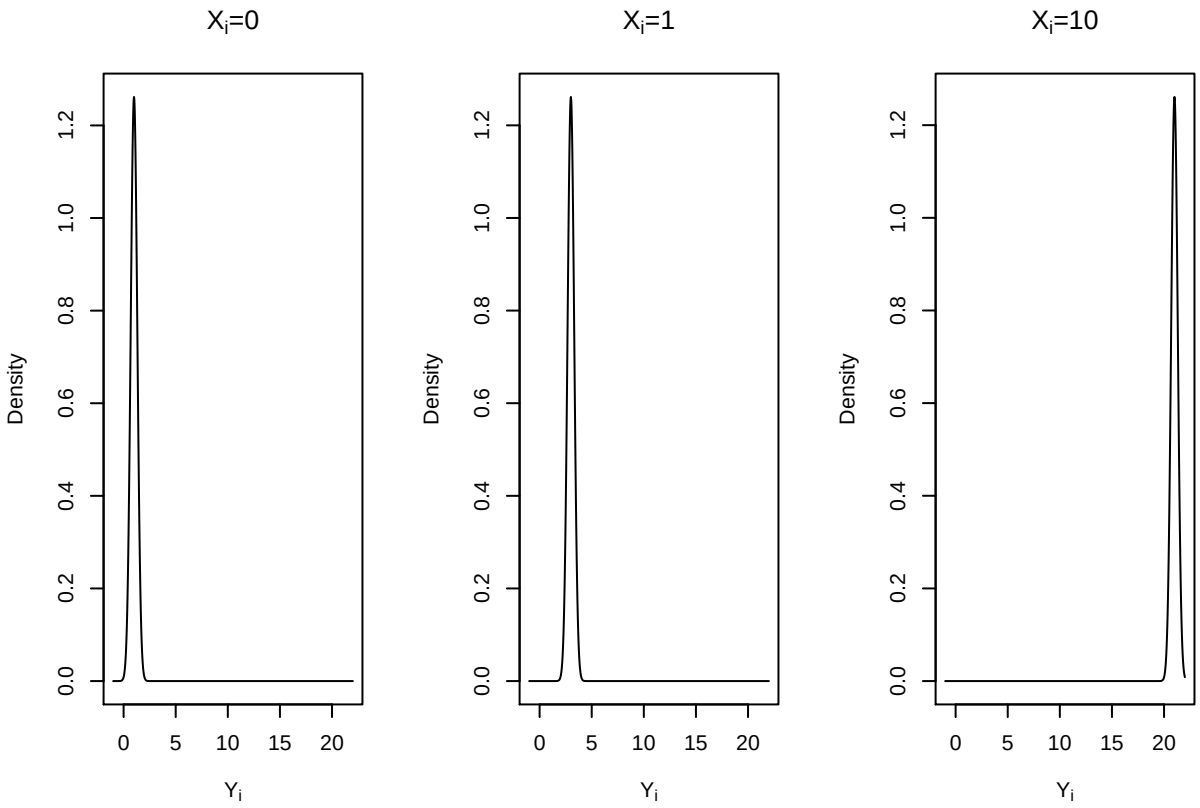
1. Think back to Problem 1 from Homeworks 2 and 3. Suppose Instagram magically knew that every time the number of times user i purchases a product, denoted by Y_i, is related to the number of times the product has been advertised to user i, denoted by X_i, as follows:

$$Y_i = 1 + 2X_i + \epsilon_i$$

where ϵ_i is a *normal* random error term with mean $E\{\epsilon_i\} = 0$ and variance $\sigma^2\{\epsilon_i\} = 0.1$; ϵ_i and ϵ_j are uncorrelated so that their covariance is zero (i.e., $\sigma\{\epsilon_i, \epsilon_j\} = 0$ for all $i \neq j$) for $i = 1, \ldots n$.

   (a) Using `R`, make a plot with three panels. You can make a single plot with three panels by typing `par(mfrow = c(1, 3))` before running any lines of code that create plots. Plot the density of the Y_i for $X_i = 0$, $X_i = 1$, and $X_i = 10$, using a separate panel for each value of X_i. Ensure that the axes are the same across all three plots.

```
par(mfrow = c(1, 3))
y.vals <- seq(-1, 22, length.out = 1000)
dens.vals1 <- dens.vals2 <- dens.vals3 <- rep(NA, length(y.vals))
for (i in 1:length(y.vals)) {
  dens.vals1[i] <- dnorm(y.vals[i], mean = 1 + 2*0, sd = sqrt(0.1))
  dens.vals2[i] <- dnorm(y.vals[i], mean = 1 + 2*1, sd = sqrt(0.1))
  dens.vals3[i] <- dnorm(y.vals[i], mean = 1 + 2*10, sd = sqrt(0.1))
}
plot(y.vals, dens.vals1, type = "l", xlab = expression(Y[i]), ylab = "Density",
     main = expression(paste(X[i], "=0", sep = "")))
plot(y.vals, dens.vals2, type = "l", xlab = expression(Y[i]), ylab = "Density",
     main = expression(paste(X[i], "=1", sep = "")))
plot(y.vals, dens.vals3, type = "l", xlab = expression(Y[i]), ylab = "Density",
     main = expression(paste(X[i], "=10", sep = "")))
```



Suppose we were actually able to work with Instagram and conduct an experiment and we randomly selected $n = 10$ of the students in our class. We sent X_i ads for coffee to student i over the course of one day, and recorded Y_i, the number of ounces of coffee student i purchased the following week.

(b) Suppose that after the experiment concluded, I told you that I fit a simple linear regression model to the data, modeling Y_i as a linear function of X_i. Imagine that I told you that when I tested the null hypothesis $H_0 : \beta_1 \geq 0$ versus the alternative $H_a : \beta_1 < 0$, I failed to reject H_0. Would you conclude that there is no linear association between X and Y?

I would not concude that there is *no* linear association between X and Y. Failing to reject this null means failing to reject that $\beta_1 \geq 0$, which includes the possibility that there is no linear association between X and Y as well as the possibility that X and Y are linearly related with $\beta_1 > 0$.

Suppose that I decided to share the results of the regression with you all.

```
summary(lm(Y~X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.940  -9.299   2.344   7.347  35.104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.489     10.729   1.257    0.244
## X             -1.001      1.788  -0.560    0.591
##
## Residual standard error: 18.75 on 8 degrees of freedom
```

```
## Multiple R-squared:  0.03769,    Adjusted R-squared:  -0.0826
## F-statistic: 0.3133 on 1 and 8 DF,  p-value: 0.591
```

(c) Imagine that an Instagram executive came to class, saw the regression results, and stated "The message I got here is that the more coffee advertisements we send, the less coffee people drink!" Would you agree or disagree? Explain.

I would disagree, because although the estimate of b_1 is negative, suggesting that one additional advertisement would correspond to average consumption of one fewer ounces consumed per day, the probability of observing a value of b_1 this extreme if β_1 were really equal to 0 would be 0.591.

2. Refer back to the Toluca Company example we have discussed in class. The `toluca` data has been posted on the Homework page.

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.876 -34.088  -5.982  38.826 103.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.366     26.177   2.382   0.0259 *
## X              3.570      0.347  10.290 4.45e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.82 on 23 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
## F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

(a) Label b_0, $s\{b_0\}$, $\frac{b_0}{s\{b_0\}}$, b_1, $s\{b_1\}$, $\frac{b_1}{s\{b_1\}}$ and s on the `R` output given above.

- From the line `(Intercept)  62.366  26.177  2.382  0.0259 *`, we have $b_0 = 62.366$, $s\{b_0\} = 26.177$, and $\frac{b_0}{s\{b_0\}} = 2.382$.
- From the line `X  3.570  0.347  10.290 4.45e-10 ***`, we have $b_1 = 3.570$, $s\{b_1\} = 0.347$, and $\frac{b_1}{s\{b_1\}} = 10.290$.
- From the line `Residual standard error: 48.82 on 23 degrees of freedom`, we have $s = 48.82$.

Simulations can be very helpful for building an intuition to interpreting intervals and test results. Using the values of b_0 and s from (a) and `rnorm`, let's simulate $k = 1, \ldots, 1,000$ synthetic datasets $Y_1^{(k)}, \ldots, Y_{25}^{(k)}$ according to:

$$Y_i^{(k)} = b_0 + \epsilon_i^{(k)}, \quad \epsilon_i \overset{i.i.d.}{\sim} \text{normal}\left(0, s^2\right) \quad \text{for } i = 1, \ldots, 25$$

We can think of this as the null model when we are considering the null hypothesis $\beta_1 = 0$, plugging in our estimates of the remaining parameters β_0 and σ^2 which are unknown. The simulated values $Y_1^{(k)}, \ldots, Y_{25}^{(k)}$ represent alternative realizations of the response that we might have observed if $\beta_1 = 0$, e.g. responses we might have observed if the Toluca data were collected in another universe.

For each of the $k = 1, \ldots, 1,000$ synthetic datasets, we will perform a level-$\alpha = 0.05$ test of the null hypothesis that $\beta_1 = 0$, and record whether or not we reject the null hypothesis.

```
set.seed(1)
reject <- rep(NA, 1000)
for (i in 1:length(reject)) {
```

```
  Y.sim <- rnorm(n = 25, mean = 62.366, sd = 48.82)
  linmod <- lm(Y.sim~X)
  b1 <- summary(linmod)$coef["X", "Estimate"]
  sb1 <- summary(linmod)$coef["X", "Std. Error"]
  reject[i] <- !((b1 + qt(0.05/2, df = 23)*sb1) <= 0 &
                   (b1 + qt(1 - 0.05/2, df = 23)*sb1) >= 0)
}
```

If we look at the proportion of times we reject the null hypothesis that $\beta_1 = 0$, which can be obtained by evaluating `mean(reject)`, we get 0.043. This just about matches α, which makes sense, because α conveys how often we would expect to reject the null that $\beta_1 = 0$ across different realizations of data generated according to the null model with $\beta_1 = 0$

(b) Simulate $k = 1, \ldots, 1,000$ synthetic datasets $Y_1^{(k)}, \ldots, Y_{25}^{(k)}$ according to:

$$Y_i^{(k)} = b_0 + 3X_i + \epsilon_i^{(k)}, \quad \epsilon_i \overset{i.i.d.}{\sim} \text{normal}\left(0, s^2\right) \quad \text{for } i = 1, \ldots, 25$$

For each of the $k = 1, \ldots, 1,000$ synthetic datasets, perform a level-$\alpha = 0.05$ test of the null hypothesis that $\beta_1 = 3$, and record whether or not we reject the null hypothesis. In what proportion/percent of simulations do you reject the null? We call this the *level* of the test, it tells us how often a level-$\alpha = 0.05$ test would lead us to reject the null hypothesis that $\beta_1 = 3$ if the true value of β_1 were in fact 3.

```
reject <- rep(NA, 1000)
for (i in 1:length(reject)) {
  Y.sim <- rnorm(n = 25, mean = 62.366 + 3*X, sd = 48.82)
  linmod <- lm(Y.sim~X)
  b1 <- summary(linmod)$coef["X", "Estimate"]
  sb1 <- summary(linmod)$coef["X", "Std. Error"]
  reject[i] <- !((b1 + qt(0.05/2, df = 23)*sb1) <= 3 &
                   (b1 + qt(1 - 0.05/2, df = 23)*sb1) >= 3)
}
```

We reject the null in 6.5% of simulations.

(c) Simulate $k = 1, \ldots, 1,000$ synthetic datasets $Y_1^{(k)}, \ldots, Y_{25}^{(k)}$ according to:

$$Y_i^{(k)} = b_0 + 3.5X_i + \epsilon_i^{(k)}, \quad \epsilon_i \overset{i.i.d.}{\sim} \text{normal}\left(0, s^2\right) \quad \text{for } i = 1, \ldots, 25$$

For each of the $k = 1, \ldots, 1,000$ synthetic datasets, perform a level-$\alpha = 0.05$ test of the null hypothesis that $\beta_1 = 3$, and record whether or not we reject the null hypothesis. In what proportion/percent of simulations do you reject the null? We call proportion this the *power* of the test for $\beta_1 = 3.5$, it tells us how often a level-$\alpha = 0.05$ test would lead us to reject the null hypothesis that $\beta_1 = 3$ if the true value of β_1 were in fact 3.5.

```
reject <- rep(NA, 1000)
for (i in 1:length(reject)) {
  Y.sim <- rnorm(n = 25, mean = 62.366 + 3.5*X, sd = 48.82)
  linmod <- lm(Y.sim~X)
  b1 <- summary(linmod)$coef["X", "Estimate"]
  sb1 <- summary(linmod)$coef["X", "Std. Error"]
  reject[i] <- !((b1 + qt(0.05/2, df = 23)*sb1) <= 3 &
                   (b1 + qt(1 - 0.05/2, df = 23)*sb1) >= 3)
}
```

We reject the null in 29% of simulations.

(d) Simulate $k = 1, \ldots, 1,000$ synthetic datasets $Y_1^{(k)}, \ldots, Y_{25}^{(k)}$ according to:

$$Y_i^{(k)} = b_0 + 6X_i + \epsilon_i^{(k)}, \quad \epsilon_i \overset{i.i.d.}{\sim} \text{normal}\left(0, s^2\right) \quad \text{for } i = 1, \ldots, 25$$

For each of the $k = 1, \ldots, 1,000$ synthetic datasets, perform a level-$\alpha = 0.05$ test of the null hypothesis that $\beta_1 = 3$, and record whether or not we reject the null hypothesis. In what proportion/percent of simulations do you reject the null? We call proportion this a *power* of a test for $\beta_1 = 6$, it tells us how often a level-$\alpha = 0.05$ test would lead us to reject the null hypothesis that $\beta_1 = 3$ if the true value of β_1 were in fact 6.

```
reject <- rep(NA, 1000)
for (i in 1:length(reject)) {
  Y.sim <- rnorm(n = 25, mean = 62.366 + 6*X, sd = 48.82)
  linmod <- lm(Y.sim~X)
  b1 <- summary(linmod)$coef["X", "Estimate"]
  sb1 <- summary(linmod)$coef["X", "Std. Error"]
  reject[i] <- !((b1 + qt(0.05/2, df = 23)*sb1) <= 3 &
                   (b1 + qt(1 - 0.05/2, df = 23)*sb1) >= 3)
}
```

We reject the null in 100% of simulations.

(e) Explain how the results of (b) , (c) and (d) differ, and comment on how the power of the test of the null hypothesis that $\beta_1 = 3$ depends on the true value of β_1.

We reject the null hypothesis that $\beta_1 = 3$ in a greater proportion of simulations as β_1 increases. This is what we would expect - the further away the true value of β_1 is from 3, the more often we reject the null hypothesis that $\beta_1 = 3$.

3. Problem 2.27 from the `.pdf` version of the textbook. Requires use of the `muscle` data that has been posted on the Homework page.

(a)

To decide whether or not there is a negative linear association between amount of muscle mass and age, we would conduct a test of the null hypothesis $H_0 : \beta_1 \geq 0$ against the alternative $H_a : \beta_1 < 0$. The decision rule for a level-$\alpha = 0.05$ test based on the test statistic $t^* = b_1 / s\{b_1\}$ would be:

- If $t^* \geq t_{58}(0.05)$, conclude H_0
- If $t^* < t_{58}(0.05)$, conclude H_a.

Because $t^* = -13.1932568$ and $t_{58}(0.05) = -1.671553$, we would reject H_0 and conclude H_a.

The p-value of the test is $P(t < t^*) = 2.0619935 \times 10^{-19}$, the probability that a t-random variable with 58 degrees of freedom is less than t^*.

```
t.star <- summary(linmod)$coef["X", "t value"]
t.quantile <- qt(0.05, df = length(Y) - 2)
p.val <- pt(t.star, df = length(Y) - 2)
```

(b)

Even if the two-sided p-value for the test of whether $\beta_0 = 0$ is 0+, it should not be concluded that b_0 provides relevant information on the amount of muscle mass at birth for a female child because $X_i = 0$ is not within the scope of the model - the smallest value of X_i in the data used to fit the model was $X_i = 41$.

(c)

The expected muscle mass for a women whose age is given by X_i is $E\{Y_i\} = \beta_0 + \beta_1 X_i$. It follows that the difference in expected muscle mass for women whose ages differ by one year is given by

$$\beta_0 + \beta_1 (X_i + 1) - (\beta_0 + \beta_1 X_i) = \beta_0 - \beta_0 + \beta_1 X_i - \beta_1 X_i + \beta_1 = \beta_1.$$

The terms that depend on X_i cancel because we have assumed the expected muscle mass for a women is linear in age X_i, so we do not need to know the specific ages to make an estimate of the difference in expected muscle mass for women whose ages differ by one year.

The estimate of the difference in expected muscle mass for women whose ages differ by one year is $b_1 = -1.1899955$, and a 95 percent confidence interval is given by $\left(b_1 + t_{58}\left(0.025\right) s\left\{b_1\right\}, b_1 + t_{58}\left(0.975\right) s\left\{b_1\right\}\right) = (-1.3705449, -1.0094461)$.

```
b1 <- summary(linmod)$coef["X", "Estimate"]
sb1 <- summary(linmod)$coef["X", "Std. Error"]
t.quantile.025 <- qt(0.025, 58)
t.quantile.975 <- qt(0.975, 58)
```

4. Problem 2.30 from the `.pdf` version of the textbook. Requires use of the `crime` data that has been posted on the Homework page.

(a)

To decide whether or not there is a linear association between amount of muscle mass and age, we would conduct a test of the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_a : \beta_1 \neq 0$. The decision rule for a level-$\alpha = 0.01$ test based on the test statistic $t^* = b_1 / s\left\{b_1\right\}$ would be:

- If $t_{82}\left(0.005\right) \leq t^* \leq t_{82}\left(0.995\right)$, conclude H_0
- If $t^* < t_{82}\left(0.005\right)$ or $t^* > t_{82}\left(0.995\right)$, conclude H_a.

Because $t^* = -4.1028971$, $t_{82}\left(0.005\right) = -2.637123$ and $t_{82}\left(0.995\right) = 2.637123$, we would reject H_0 and conclude H_a.

The p-value of the test is $P\left(|t| > |t^*|\right) = 9.5713958 \times 10^{-5}$, the probability that a t-random variable with 82 degrees of freedom is greater than $|t^*|$ in absolute value.

```
t.star <- summary(linmod)$coef["X", "t value"]
t.quantile.005 <- qt(0.005, df = length(Y) - 2)
t.quantile.995 <- qt(0.995, df = length(Y) - 2)
p.val <- 2*pt(abs(t.star), df = length(Y) - 2,
              lower.tail = FALSE)
```

(b)

The estimate of the difference in expected crime rate for areas whose percentage of high school graduates differ by one percent is $b_1 = -170.5751886$, and a 99 percent confidence interval is given by $\left(b_1 + t_{82}\left(0.005\right) s\left\{b_1\right\}, b_1 + t_{58}\left(0.995\right) s\left\{b_1\right\}\right) = (-280.2118218, -60.9385555)$.

```
b1 <- summary(linmod)$coef["X", "Estimate"]
sb1 <- summary(linmod)$coef["X", "Std. Error"]
t.quantile.005 <- qt(0.005, 82)
t.quantile.995 <- qt(0.995, 82)
```