36-617: Applied Linear Models
Fall 2021
HW01 – Solutions

1. Sheather, Ch 2, p. 38, #1

   First, we read in the data, take a look at it, and plot it (Figure 1) to make sure it looks like the data in the book.

```
> data <- read.csv("playbill.csv")
> str(data,width=72,strict.width = "cut")

'data.frame':        18 obs. of  3 variables:
 $ Production : chr  "42nd Street" "Avenue Q" "Beauty and Beast" "Bom"..
 $ CurrentWeek: int  684966 502367 594474 529298 570254 319959 579126 ..
 $ LastWeek   : int  695437 498969 598576 528994 562964 282778 583177 ..

> plot(CurrentWeek ~ LastWeek, xlab="Gross Box Office Receipts Previous Week ($)",
+      ylab="Gross Box Office Receipts Current Week ($)",data=data)
```

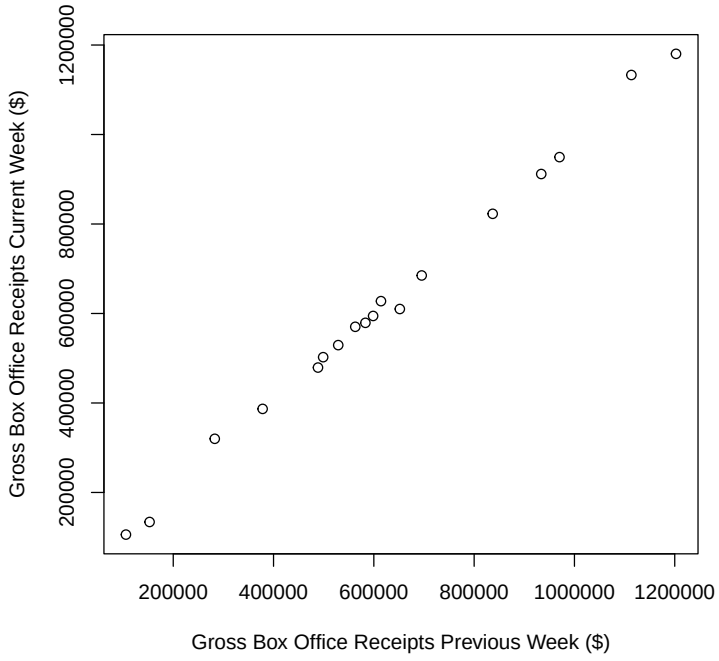

Figure 1: Scatter Plot of Gross Box Office Receipts

And, let's go ahead and run the regression of `CurrentWeek` on `LastWeek`:

```
> boxoffice.reg <- lm(CurrentWeek ~ LastWeek, data=data)
> summary(boxoffice.reg)

Call:
lm(formula = CurrentWeek ~ LastWeek, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-36926  -7525  -2581   7782  35443

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.805e+03  9.929e+03   0.685    0.503
LastWeek    9.821e-01  1.443e-02  68.071   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18010 on 16 degrees of freedom
Multiple R-squared:  0.9966,        Adjusted R-squared:  0.9963
F-statistic:  4634 on 1 and 16 DF,  p-value: < 2.2e-16
```

Now, on to the parts of the exercise:

(a) *Find a 95% confidence interval for the slope of the regression model, β_1. Is 1 a plausible value for β_1? Give a reason to support your answer.*

A reasonable "back of the envelope" 95% interval is just "Estimate ± 2SE"... We can record the necessary estimates of $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ in a small data frame `ests` as follows. Note that $\hat{\beta}_0$ and $SE(\hat{\beta}_0)$ are in the first row, first two columns, and $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ are in the second row.

```
> print(ests <- coefficients(summary(boxoffice.reg)))

                Estimate   Std. Error    t value     Pr(>|t|)
(Intercept) 6804.8860355 9.929318e+03  0.6853327 5.029432e-01
LastWeek       0.9820815 1.442723e-02 68.0714024 3.866450e-21
```

This leads to the "back of the envelope" CI

```
> ests[2,1] + c(-1,1)*2*ests[2,2]

[1] 0.953227 1.010936
```

An "exact" CI would use the exact cutoff for the appropriate t-distribution. In this case, there are $n = 18$ observations and 2 β's, so $df = 18 - 2 = 16$... We have to get the appropriate upper cutoff of the t distribution for a 95% CI

```
> (tscore <- qt(1-0.025,16))
```

```
[1] 2.119905
```

which leads to the "exact" CI:

```
> ests[2,1] + c(-1,1)*tscore*ests[2,2]
[1] 0.9514971 1.0126658
```

("exact" in quotes here because it assumes we know that the regression errors (ϵ_i's) really are iid $N(0, \sigma^2)$...)

Finally, 1 is a plausible value for β_1 since it is in the confidence interval (either the crude or the "exact" one!). Equivalently, one could construct a hyothesis test for $H_0 : \beta_1 = 1$ vs a two-sided alternative. The t statistic would be

$$\frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} = \frac{0.9821 - 1}{0.0144} = -1.24$$

and the two-sided p-value would be `2*(1-pt(abs(-1.24),16))`, which is `0.23`. Since this is greater than a "usual" level like 0.01 or 0.05, it suggests there is not enough evidence to reject $H_0 : \beta_1 = 1$.

(b) *Test the null hypothesis $H_0 : \beta_0 = 10000$ against a two-sided alternative. Interpret your result.*

This is pretty easy, given our previous work. The t-statistic we want is

$$\frac{\hat{\beta}_0 - 10000}{SE(\hat{\beta}_0)} = \frac{6804.886 - 10000}{9929.3178} = -0.32$$

The two-sided p-value is `2*(1-pt(abs(-0.32,16)))`, which is `0.75`, and again we do not have enough evidence to reject $H_0 : \beta_0 = 10000$.

Interpretation: The regression model we have fitted is

$$\texttt{CurrentWeek} = \beta_0 + \beta_1 \texttt{LastWeek} + \epsilon$$

so that $\hat{\beta}_0$ is the estimated value of the current week's receipts, given that last week's receipts were \$0; maybe we can think of this as the receipts for a play's opening week. If so, then the hypothesis test says that \$10,000 is a plausible level of receipts for the opening week of a play.

(But on the other hand \$0 is also a plausible value, as we can see from the `ests` table above [the p-value for the test of $H_0 : \beta_0 = 0$ from the table is 0.5029]). The problem, in some sense, is that $SE(\hat{\beta}_0)$ is so large than very many different values are plausible here.)

(c) *Use the fitted regression model to estimate the gross box office results for the current week (in \$) for a production with \$400,000 in gross box office the previous week. Find a 95% prediction interval for the gross box office results for the current week (in \$) for*

a production with \$400,000 in gross box office the previous week. Is \$450,000 a feasible value for the gross box office results in the current week, for a production with \$400,000 in gross box office the previous week? Give a reason to support your answer.

This is easy to get with the `predict` command in R:

```
> predict(boxoffice.reg,
+         newdata=data.frame(Production="",CurrentWeek=0,LastWeek=400000),
+         interval="prediction")
       fit      lwr      upr
1 399637.5 359832.8 439442.2
```

so that the interval is `(359832.8, 439442.2)`. Since \$450,000 is *not* in this interval, it is *not* likely to be feasible to expect \$450,000 in gross receipts for the current week, for a production whose previous week's gross receipts were \$400,000.

(d) *Some promoters of Broadway plays use the prediction rule that next week's gross box office results will be equal to this week's gross box office results. Comment on the appropriateness of this rule.*

Since we could not reject $H_0 : \beta_1 = 1$ in part (1a), and we could not reject $H_0 : \beta_0 = 0$ in the table `ests` above, it is at least plausible that the true relationship has $\beta_0 = 0$ and $\beta_1 = 1$, i.e.

$$\texttt{CurrentWeek} = 0 + 1 \cdot \texttt{LastWeek} + \epsilon \ .$$

This would make the rule of thumb that next week's gross receipts approximately equal this week's gross receipts at least plausible.

(Some caveats: (1) Obviously this doesn't work in the first week of a Broadway show; and (2) This rule of thumb could work in the middle of the run of the show, but toward the end eventually you run out of people intertested in seeing the show, and the receipts have to go down again.)

2. Sheather, Ch 2, pp. 41–42, #5. *In Figure 2 we have plotted Y vs x_1 on the left, and Y vs x_2 on the right. Fitted regression lines for regressing Y on x_1 (Model 1) and regressing Y on x_2 (Model 2) are also shown. For which model is `RSS` greater, and for which model is `SSreg` greater?*

$RSS = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ measures the sum of the squared vertical distances between the regression line and the actual data points. Since the vertical scale in the two plots in Figure 2 is the same, and the number of data points is the same (same Y) we can visually see that `RSS` for Model 1 will be *smaller* than `RSS` for Model 2.

Next we observe that $SYY = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the same for both scatter plots, and

$$SYY = SSReg + RSS \qquad (*)$$

Therefore, since `RSS` for Model 1 is *smaller* than `RSS` for Model 2, it follows that `SSReg` for Model 1 will be *larger* than `SSReg` for Model 2.

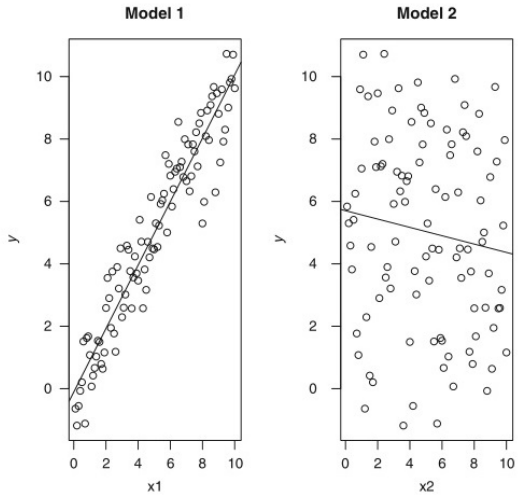

Figure 2: Scatter plots and least-squares lines

Among the 4 options given in Sheather, we would choose

(d) RSS for model 1 is less than RSS for model 2, while SSreg for model 1 is greater than SSreg for model 2.

(Another way to see this is to see that R^2 is larger for Model 1 than for Model 2, and since $R^2 = SSReg/SYY$ then `SSReg` must be larger for Model 1; using $(*)$ again, it follows that `RSS` is smaller for Model 1.)

3. Sheather, Ch 2, p. 42 #6

Show that $SYY = SSReg + RSS$ *(note:* $SYY = SST$*; BJ uses* SYY *and Sheather uses* SST*). To do this, Sheather suggests first showing that* $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 0$.

(a) *Show that* $(y_i - \hat{y}_i) = (y_i - \overline{y}) - \hat{\beta}_1(x_i - \overline{x})$.

$$\begin{aligned}(y_i - \hat{y}_i) &= (y_i - \overline{y}) + (\overline{y} - \hat{y}_i) \\ &= (y_i - \overline{y}) + (\overline{y} - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) \\ &= (y_i - \overline{y}) + (\overline{y} - ((\overline{y} - \hat{\beta}_1\overline{x}) + \hat{\beta}_1 x_i)) \quad (\text{since } \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}) \\ &= (y_i - \overline{y}) - \hat{\beta}_1(x_i - \overline{x}) \quad (\text{after regrouping and cancelling terms})\end{aligned}$$

(b) *Show that* $(\hat{y}_i - \overline{y}) = \hat{\beta}_1(x_i - \overline{x})$.

$$\begin{aligned}(\hat{y}_i - \overline{y}) &= (\hat{y}_i - y_i) + (y_i - \overline{y}) \\ &= -((y_i - \overline{y}) - \hat{\beta}_1(x_i - \overline{x})) + (y_i - \overline{y}) \quad (\text{from part (a)}) \\ &= \hat{\beta}_1(x_i - \overline{x})) \quad (\text{after rearranging and cancelling terms})\end{aligned}$$

(c) *Using the fact that* $\hat{\beta}_1 = \frac{SXY}{SXX}$*, show that* $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$.
Substituting in our identities from (a) and (b) for $(y_i - \hat{y}_i)$ and $(\hat{y}_i - \bar{y})$, we have

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}\left[(y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})\right]\left[\hat{\beta}_1(x_i - \bar{x})\right] \\
&= \hat{\beta}_1 \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - (\hat{\beta}_1)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \left(\frac{SXY}{SXX}\right)SXY - \left(\frac{SXY}{SXX}\right)^2 SXX \\
&= \frac{SXY^2}{SXX} - \frac{SXY^2}{SXX} \\
&= 0
\end{aligned}$$

Finally, we show that $SYY = SSReg + RSS$:

$$\begin{aligned}
SYY &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}\left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right]^2 \\
&= \sum_{i=1}^{n}\left[(y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\right] \\
&= RSS + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + SSReg \\
&= RSS + SSReg
\end{aligned}$$

since by part (c), $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$. Thus, $SYY = RSS + SSReg = SSReg + RSS$.

4. [Gelman & Hill (2007), Ch 3, #3] *In this exercise you will simulate two variables that are statistically independent of each other to see what happens when we run a regression of one on the other.*

   (a) *First generate 1000 data points from a normal distribution with mean 0 and standard deviation 1 by typing* `var1 <- rnorm(1000,0,1)` *in R. Generate another variable in the same way (call it* `var2`*). Run a regression of one variable on the other. Is the slope coefficient statistically significant?*

```
> var1 <- rnorm(1000,0,1)
> var2 <- rnorm(1000,0,1)
> bozo <- lm(var2 ~ var1)
> summary(bozo)
```

```
Call:
lm(formula = var2 ~ var1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8559 -0.7363 -0.0107  0.7402  3.6052

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.008008   0.032224  -0.249    0.804
var1         0.004314   0.032501   0.133    0.894

Residual standard error: 1.019 on 998 degrees of freedom
Multiple R-squared:  1.765e-05,        Adjusted R-squared:  -0.0009843
F-statistic: 0.01762 on 1 and 998 DF,  p-value: 0.8944
```

We can see from the coefficient table above that $\hat{\beta}_1 = 0.004$, $SE(\hat{\beta}_1) = 0.033$, the t-statistic for testing $H_0 : \beta_1 = 0$ is $t = \hat{\beta}_1 / SE(\hat{\beta}_1) = 0.133$, and the p-value is $p = 0.894$. With such a large p-value, we do not have enough evidence to reject $H_0 : \beta_1 = 0$, and so the slope is not statistically significant.

(b) *Now run a simulation repeating this process 100 times. This can be done using a loop. From each simulation, save the z-score (the estimated coefficient of* ***var1*** *divided by its standard error). If the absolute value of the z-score exceeds 2, the estimate is statistically significant. Here is code to perform the simulation*[1]*:*

```
z.scores <- rep (NA, 100)
for (k in 1:100) {
  var1 <- rnorm (1000,0,1)
  var2 <- rnorm (1000,0,1)
  fit <- lm (var2 ~ var1)
  z.scores[k] <- coef(fit)[2]/se.coef(fit)[2]
}
```

How many of these 100 z-scores are statistically significant?

Note that the function `se.coef` used in the code above depends on the `arm` library, which we haven't talked about in class yet. You could use this exact code if you first install the `arm` package and run the command `library(arm)` before running this code.

However, we can also just use `coefficients(summary())` to get the pieces we need, as we have done in earlier exercises in this assignment. So the code we will run is this:

```
> z.scores <- rep (NA, 100)
> for (k in 1:100) {
```

[1] We have initialized the vector of z-scores with missing values (`NA`s). Another approach is to start with `z.scores <- numeric(length=100)`, which would initialize with a vector of zeroes. In general, however, we prefer to initialize with `NA`s, because then when there is a bug in the code, it sometimes shows up as `NA`s in the final results, alerting us to the problem

```
+   var1 <- rnorm (1000,0,1)
+   var2 <- rnorm (1000,0,1)
+   fit <- lm (var2 ~ var1)
+   ests <- coefficients(summary(fit))
+   z.scores[k] <- ests[2,1]/ests[2,2]
+ }
```

The degrees of freedom for the t-statistic would be $1000 - 2 = 998$, so the t-distribution is essentially indistinguishable from the normal. So we can use the back-of-the-envelope calculation that whenever the z score calculated in the code above is greater that 2 in absolute value, it is significant at the 0.05 level. So to count the significant z scores we just calculate

```
> sum(abs(z.scores)>=2)
[1] 4
```

If we want to use the exact normal cutoff for a level 0.05 two-sided test of $H_0 : \beta_1 = 0$, we should use 1.96 instead of 2:

```
> sum(abs(z.scores)>=1.96)
[1] 5
```

(c) *Is your answer to (b) what you expected? Why or why not?*
Either answer in part (b) is about what we'd expect. For a level 0.05 test, we expect 5% false positives. 5% of 100 is 5, and we are getting 4 or 5 false positives (depending on whether we use the back-of-the-envelope cutoff or the "exact" cutoff).

5. Sheather, Ch 2, pp 42–43, #7

A statistics professor has been involved in a collaborative research project with two entomologists. The statistics part of the project involves fitting regression models to large data sets. Together they have written and submitted a manuscript to an entomology journal. The manuscript contains a number of scatter plots with each showing an estimated regression line (based on a valid model) and associated individual 95% confidence intervals for the regression function at each x value, as well as the observed data. A referee has asked the following question:

> I don't understand how 95% of the observations fall outside the 95% CI as depicted in the figures.

Briefly explain how it is entirely possible that 95% of the observations fall outside the 95% CI as depicted in the figures.

[Note: this is not as "brief" an answer as expected by the problem statement, since I want to explain more carefully what is going on. The main thing that I want you to be able to say comes at the end below.]

There is a difference between a 95% interval *for values from the population* and a 95% interval *for estimates of the mean.* For example, here I have simulated 100 values from a normal distribution $N(0, 25)$ with mean $\mu = 0$, variance $\sigma^2 = 25$

```
> x.values <- rnorm(100,0,5)
```

and we can expect that 95% of the values will be between -1.96σ and $+1.96\sigma$. Indeed,

```
> sum(abs(x.values)<=1.96*5)

[1] 94
```

which is about right (we expected 95 of the 100 values to be in this interval).

A 95% interval for the mean, however, is much narrower:

```
> est <- mean(x.values)
> se <- sd(x.values)/sqrt(100-1)
> est + c(-1,1)*1.96*se

[1] -1.098594  1.044443
```

The interval $(-1.96 \cdot \sigma, 1.96 \cdot \sigma) = (-9.8, 9.8)$ expresses how much variation we expect to see in a sample from the population of x values distributed $N(0, \sigma^2)$. The interval $(-1.96 \cdot \hat{\sigma}/\sqrt{n-1}, 1.96 \cdot \hat{\sigma}/\sqrt{n-1}) = (-1.0986, 1.0444)$ expresses how much variation we expect to see in a sample of estimates $\overline{x}$ of the mean μ, which are distributed $N(0, \sigma^2/100)$. Smaller variance, smaller confidence interval; the number of values in the sample inside this smaller interval is only

```
> sum(abs(x.values)<=1.96*se)

[1] 12
```

88% of the sample values in this particular simulation are outside the interval for the mean (which is fine, because they are not means!).

[Here's the main thing...]

The same thing is going on with 95% CI's for the regression function at each x.

The CI's for the regression are for the means $E[y]x] = \beta_0 + \beta_1 x$ at each x, not for the individual observations y associated with that x. We know that

$$y \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

while the CI for $E[y|x]$ is based on

$$\hat{y} \sim N\left(\beta_0 + \beta_1 x, \left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{SXX}\right]\sigma^2\right)$$

which has smaller variance and so will produce a narrower CI. The CI expresses variation we expect to see in estimates of $E[y|x]$, which is less than the variation we would see in individual observations y associated with that x, as shown in Figure 3 on page 10.

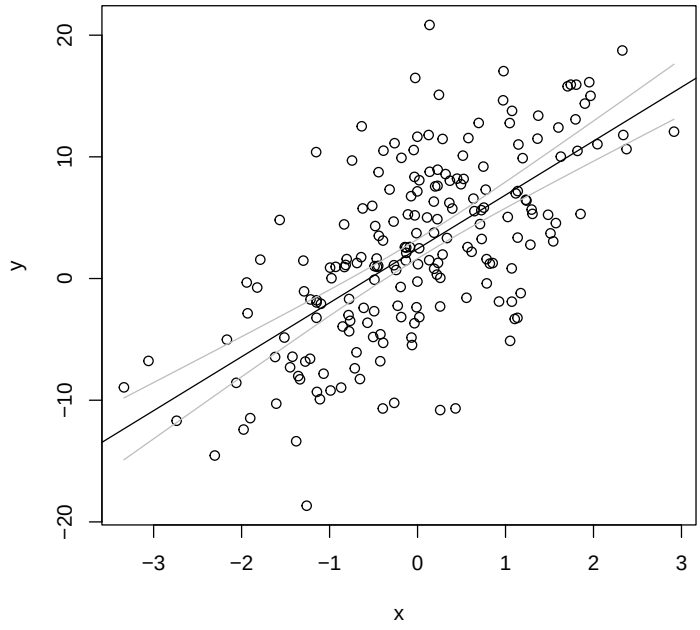

Figure 3: Regression with 95% CI's for the regression line at each x shown in grey.