ST 518 Homework 3

Eric Warren

September 15, 2023

Contents

1 Problem 1

1.1 Part A

Calculate the following matrix products.

- X
- $X'X$
- $X'Y$

1.1.1 Part i

First we are going to calculate X. As we know this matrix is a (n by $(p+1)$). Thus since $n = 4$ and $p = 2$, we have the X matrix being a (4 by $(2+1)$) or (4 by 3) matrix. We also know that the rows of this matrix are the observations with the columns having the first one just be the value of 1, the second one the value of the first predictor for the observation, and the third column being thee value of the second predictor for the observation. By looking at our chart we know that the X matrix is $X = \begin{pmatrix} 1 & 1 & -2 \\ 1 & 0 & -1 \\ 1 & 0 & 2 \\ 1 & 2 & 1 \end{pmatrix}$.

1.1.2 Part ii

Now we are going to calculate $X'X$. In this case we know what X is from **Part i**, so now we are going to find X' from *transposing* X. This is just moving the i^{th} rows in X to the j^{th} columns in X' and vice vsera. Thus by doing this flip, we can easily get the $X' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 2 \\ -2 & -1 & 2 & 1 \end{pmatrix}$. Now we need to find $X'X$. This is done by summing all the products of the i^{th} row in X' by the j^{th} column in X. For example to find $X'X_{11}$ we would multiply the first element in the first row of the X' matrix by the first element in the first column in X and add that to the second element in the first row of the X' matrix by the second element in the first column in X and continue to add and do that multiplication until we get to the last element in the first row of the X' matrix and multiply by the last element in the first column in X. This completes the first row and the first column of $X'X$ and we continue to do that until completing the matrix. Since we know that $X'X$ is a ($p+1$ by $p+1$) matrix and p+1 = 3, we should get a (3 x 3) matrix for $X'X$.

By doing our matrix multiplication we can see that $X'X$

$$= \begin{pmatrix} 1*1+1*1+1*1+1*1 & 1*1+1*0+1*0+1*2 & 1*-2+1*-1+1*2+1*1 \\ 1*1+0*1+0*1+2*1 & 1*1+0*0+0*0+2*2 & 1*-2+0*-1+0*2+2*1 \\ -2*1+-1*1+2*1+1*1 & -2*1+-1*0+2*0+1*2 & -2*-2+-1*-1+2*2+1*1 \end{pmatrix}$$

$$= \begin{pmatrix} 1+1+1+1 & 1+0+0+2 & -2-1+2+1 \\ 1+0+0+2 & 1+0+0+4 & -2+0+0+2 \\ -2-1+2+1 & -2+0+0+2 & 4+1+1+4 \end{pmatrix}$$

$$= \begin{pmatrix} 4 & 3 & 0 \\ 3 & 5 & 0 \\ 0 & 0 & 10 \end{pmatrix}$$

Therefore, we have found that $X'X = \begin{pmatrix} 4 & 3 & 0 \\ 3 & 5 & 0 \\ 0 & 0 & 10 \end{pmatrix}$.

1.1.3 Part iii

Now we are going to calculate $X'Y$. We know the Y matrix is a (n x 1) matrix and since $n = 4$, we know that Y is a (4 x 1) matrix. To find the Y matrix, all we will do is put in the values from the y_i column in order. Therefore we can see that our Y matrix is $Y = \begin{pmatrix} 68 \\ 57 \\ 90 \\ 112 \end{pmatrix}$. Now we will use the same type of matrix multiplication as we did for $X'X$ to get $X'Y$. In this case we will only have to worry about one column. We also know that the $X'Y$ matrix should be a ($p+1$ x 1) matrix form. Since $p+1 = 3$, then $X'Y$ is a (3 x 1) matrix. Now to solve $X'Y$

$$= \begin{pmatrix} 1*68+1*57+1*90+1*112 \\ 1*68+0*57+0*90+2*112 \\ -2*68+-1*57+2*90+1*112 \end{pmatrix}$$

$$= \begin{pmatrix} 68+57+90+112 \\ 68+0+0+224 \\ -136-57+180+112 \end{pmatrix}$$

$$= \begin{pmatrix} 327 \\ 292 \\ 99 \end{pmatrix}$$

Therefore, $X'Y = \begin{pmatrix} 327 \\ 292 \\ 99 \end{pmatrix}$.

1.2 Part B

We are now going to find $(X'X)^{-1}$.

To find this, first we are going to augment $X'X$ with the identity matrix to get $\left(\begin{array}{ccc|ccc} 4 & 3 & 0 & 1 & 0 & 0 \\ 3 & 5 & 0 & 0 & 1 & 0 \\ 0 & 0 & 10 & 0 & 0 & 1 \end{array}\right)$.
Now we are going to find the rows by trying to get the identity matrix on the left with our $(X'X)^{-1}$ on the right.

For purpose of notation we are going to do row numbers as R_i where i is the row number and R_i' as i being the row number and the **“”’** being the new matrix we should get to get to $(X'X)^{-1}$ eventually being on the right and the identity matrix on the left.

To start off with this, we are going to get the identity matrix on the last row by doing $R_3' = R_3/10$. This now gets us the matrix $\left(\begin{array}{ccc|ccc} 4 & 3 & 0 & 1 & 0 & 0 \\ 3 & 5 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{10} \end{array}\right)$.

Now we are going to start getting the second row in order by first doing $R_2' = 3R_1 - 4R_2$. This will now get us the matrix of $\left(\begin{array}{ccc|ccc} 4 & 3 & 0 & 1 & 0 & 0 \\ 0 & -11 & 0 & 3 & -4 & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{10} \end{array}\right)$. To complete the second row, we need to get the second element in the row to be 1 by doing $R_2' = \frac{R_2'}{-11}$. Thus, this new matrix will now give us $\left(\begin{array}{ccc|ccc} 4 & 3 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & \frac{-3}{11} & \frac{4}{11} & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{10} \end{array}\right)$.

Lastly, we need to get the first row in order by first doing $R_1' = R_1 - 3R_2$. This now gets us the matrix $\left(\begin{array}{ccc|ccc} 4 & 0 & 0 & \frac{20}{11} & \frac{-12}{11} & 0 \\ 0 & 1 & 0 & \frac{-3}{11} & \frac{4}{11} & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{10} \end{array}\right)$. To complete the first row, we need to get the first element in the row to equal 1 which we will do by saying $R_1' = \frac{R_1'}{4}$. Thus, our new matrix is now $\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{5}{11} & \frac{-3}{11} & 0 \\ 0 & 1 & 0 & \frac{-3}{11} & \frac{4}{11} & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{10} \end{array}\right)$.

Since our identity matrix is on the left and completed, we know that $(X'X)^{-1}$ is on the right side. Therefore, we have found $(X'X)^{-1} = \begin{pmatrix} \frac{5}{11} & \frac{-3}{11} & 0 \\ \frac{-3}{11} & \frac{4}{11} & 0 \\ 0 & 0 & \frac{1}{10} \end{pmatrix}$.

1.3 Part C

We are going to obtain the product of $(X'X)^{-1}X'Y$. From **Part B**, we can see that $(X'X)^{-1} = \begin{pmatrix} \frac{5}{11} & \frac{-3}{11} & 0 \\ \frac{-3}{11} & \frac{4}{11} & 0 \\ 0 & 0 & \frac{1}{10} \end{pmatrix}$ and from **Part A, Part iii**, we can see that $X'Y = \begin{pmatrix} 327 \\ 292 \\ 99 \end{pmatrix}$. Since $(X'X)^{-1}$ is a (3 x 3) matrix and $X'Y$ is a (3 x 1) matrix, we know that $(X'X)^{-1}X'Y$ is going to be a (3 x 1) matrix as well (this is know since you take the row value of the first matrix and the column value of the second matrix if the column value of the first matrix matches up to the row value of the second matrix). Now we can solve $(X'X)^{-1}X'Y$

$$= \begin{pmatrix} \frac{5}{11} & \frac{-3}{11} & 0 \\ \frac{-3}{11} & \frac{4}{11} & 0 \\ 0 & 0 & \frac{1}{10} \end{pmatrix} * \begin{pmatrix} 327 \\ 292 \\ 99 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{5}{11} * 327 + \frac{-3}{11} * 292 + 0 * 99 \\ \frac{-3}{11} * 327 + \frac{4}{11} * 292 + 0 * 99 \\ 0 * 327 + 0 * 292 + \frac{1}{10} * 99 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1635}{11} - \frac{876}{11} + 0 \\ \frac{-981}{11} + \frac{1168}{11} + 0 \\ 0 + 0 + \frac{99}{10} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{759}{11} \\ \frac{187}{11} \\ \frac{99}{10} \end{pmatrix}$$

$$= \begin{pmatrix} 69 \\ 17 \\ \frac{99}{10} \end{pmatrix}$$

.

Therefore, we can see that $(X'X)^{-1}X'Y = \begin{pmatrix} 69 \\ 17 \\ \frac{99}{10} \end{pmatrix}$. We call this vector the **vector of regression parameters** which the first element gives us our intercept, the second element gives us our first predictor's partial slope, and the third element gives us the partial slope for the second predictor.

2 Problem 2

Load in the `trees` data set and the `tidyverse` and `ppcor` packages. This will be needed to solve some questions about this data. We are then going to find an additive model of `Volume` on `Girth` and `Height`.

```
# Load in packages
library(tidyverse)
library(ppcor)

# Load in data
trees <- as_tibble(trees)

# Make the additive model
treeVolumeAdditiveModel <- lm(Volume ~ Girth + Height, trees)
treeVolumeAdditiveModel
```

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = trees)
##
## Coefficients:
## (Intercept)        Girth       Height
##    -57.9877       4.7082       0.3393
```

For **Parts A and B**, we are going to use the `pcor()` function to find these partial correlation coefficient values. Here we are going to show the table below and these will be used to answer our questions.

```
pcor(trees)
```

```
## $estimate
##             Girth     Height    Volume
## Girth   1.0000000 -0.2909777 0.9586123
## Height -0.2909777  1.0000000 0.4418950
## Volume  0.9586123  0.4418950 1.0000000
##
## $p.value
##                Girth     Height       Volume
## Girth  0.000000e+00 0.11875911 8.223304e-17
## Height 1.187591e-01 0.00000000 1.449097e-02
## Volume 8.223304e-17 0.01449097 0.000000e+00
##
## $statistic
##            Girth    Height    Volume
## Girth   0.000000 -1.609346 17.816084
## Height -1.609346  0.000000  2.606594
## Volume 17.816084  2.606594  0.000000
##
## $n
## [1] 31
##
## $gp
## [1] 1
##
```

```
## $method
## [1] "pearson"
```

2.1 Part A

Here we are going to want to obtain the partial correlation coefficient between `Volume` and `Girth` after adjusting for linear dependence on `Height`. As we can see from our table above showing the partial correlation coefficient values, we can see that the partial correlation coefficient between `Volume` and `Girth` after adjusting for linear dependence on `Height` is 0.9586123. We can find this value by looking at the `$estimate` part of the table and then finding where `Volume` and `Girth` match up in the table. That is how we find our value to be 0.9586123.

2.2 Part B

Here we are going to want to obtain the partial correlation coefficient between `Volume` and `Height` after adjusting for linear dependence on `Girth`. As we can see from our table above showing the partial correlation coefficient values, we can see that the partial correlation coefficient between `Volume` and `Height` after adjusting for linear dependence on `Girth` is 0.441895. We can find this value by looking at the `$estimate` part of the table and then finding where `Volume` and `Height` match up in the table. That is how we find our value to be 0.441895.

2.3 Part C

We are going to try to find which two models the partial correlation coefficient is calculated as the correlation from in **Part A**. We know that it is defined by the correlation coefficient between the residuals computed from the two regressions below:

- $Y = \beta_0 + \beta_2 * x_2 + ... + \beta_p * x_p$
- $X_1 = \beta_0 + \beta_2 * x_2 + ... + \beta_p * x_p$

Since we know that the variables in question are `Volume` (which is our Y variable) and `Girth` (our X_1 variable), we can plug that in for our models in this case with only two predictors. Thus the two models we are comparing which is used for our calculation are:

- $Y = \beta_0 + \beta_2 * x_2$ which in words is saying `Volume` $= \beta_0 + \beta_2$ * `Height`
- $X_1 = \beta_0 + \beta_2 * x_2$ which in words is saying `Girth` $= \beta_0 + \beta_2$ * `Height`

We will then get the residuals from both models and find the partial correlation by finding the correlation between e_{y*2} and e_{1*2}.

3 Problem 3

We know that there are 2 groups for supplements and $n = 20$ students who are doing a test.

3.1 Part A

Consider a t-test comparing posttest scores (y) with and without the supplement without including pretest scores (z) in the analysis. Report the absolute value of the t-statistic and associated degrees of freedom from such a test.

Here we are trying to find the appropriate t-statistic to perform our test. What we should do is create an ANOVA table of the simple linear regression case and use this to help find our t-statistic. We know that our total degrees of freedom is $n - 1 = 20 - 1 = 19$. We also know that our total sums of squares is 2166.2 from our chart. We can also find our `supp` degrees of freedom to be equal to $p = 1$ since it is one predictor (p) and thus our `Error` degrees of freedom is $n - p - 1 = 20 - 1 - 1 = 18$.

We also know that our Sum of Squares for `supp` is equal to the Type I SS in our chart for `supp`. As we can see this is equal to 24.2. Our Mean Square value for `supp` is equal to the sum of squares for `supp` divided by the degrees of freedom for `supp` which is equal to $\frac{24.2}{1} = 24.2$.

Now for our sum of squares for `Error`. This is equal to the Total sum of squares minus the sum of squares for `supp` which is equal to $2166.2 - 24.2 = 2142$. Our mean square error is just taking 2142 (or the sum of squares error value) and dividing it by the degrees of freedom for `Error`. Therefore we can see that our mean squared error is equal to $\frac{2166.2-24.2}{18} = 119$.

Lastly we need to get our F-value and p-value to complete the ANOVA table. Our F-value is just the mean squared value of `supp` divided by the mean square `Error`. Therefore we can see that the F-value is equal to $F = \frac{24.2}{\frac{2166.2-24.2}{18}} = 0.2033613$. To find our p-value we are going to use the F-statistic value of 0.2033613, with the numerator degrees of freedom being 1 (the number of predictors) and the denominator degrees of freedom being 18 (the `Error` degrees of freedom). Therefore, we can find the p-value by using the `pf()` function in R by plugging in `pf(24.2 / ((2166.2-24.2) / 18), 1, 18, lower.tail = FALSE)`, which gives us the p-value of 0.6574066. Now we have completed our ANOVA table for the simple linear regression case for post-test score (`y`) being modeled by if the supplement was given (`supp`). The ANOVA table is below.

Source	DF	SS	MS	F-value	p-value
supp	1	24.2	24.2	0.2033613	0.6574066
Error	18	2142	119		
Corrected Total	19	2166.2			

Now we were asked to find the t-statistic with its degrees of freedom to do the test. The degrees of freedom in a t-statistic is the same as the denominator degrees of freedom for the F-test. Therefore, we know the degrees of freedom for the t-statistic is 18. We also know that the absolute t-statistic for a two sided test is just the positive square root of the F-statistic ($t = \sqrt{F}$). So we take the square root of our f-statistic value of 0.2033613 to get 0.450956. We can check to make sure we get the same p-value with the t-statistic from the ANOVA table in which we do which is checked by `2 * pt(sqrt(24.2 / ((2166.2-24.2) / 18)), 18, lower.tail = FALSE)` which gives us this same p-value of 0.6574066.

Therefore, we have found the absolute t-statistic value of 0.450956, which has 18 degrees of freedom in the test. Lastly, we can say that if this test was conducted, we would fail to reject the H_0, as there is not statistically significant evidence to prove that `supp` is a needed predictor in our model (or its slope is not zero).

3.2 Part B

Consider a simple linear regression of posttest scores on pretest scores (z). Construct the ANOVA table from such a model. Report the coefficient of determination.

Here we can make the ANOVA table by looking at the output. We know that our only predictor is `z` in this case. So our degrees of freedom is just 1 as our only predictor in the model. Thus, the rest of our model

comes from `Error` which has our degrees of freedom as $n - p - 1 = n - 1 - 1 = 20 - 1 - 1 = 18$. Our sum of squares for `z` in the output is 2015.30067 (which can be found by looking at the Type I and Type II SS categories). We know our total sum of squares by looking at the output is 2166.20000 so our `Error` sum of squares is the total of 2166.20000 minus the `z` sum of squares which is 2015.30067 which gives us an `Error` sum of squares of 150.89933. Now our Mean Squared value for is just the sum of squares divided by its corresponding degrees of freedom. Thus, for the Mean Square of `z`, we can show that the sum of squares divided by degrees of freedom is just 2015.30067 / 1 = 2015.30067. For the Mean Square of `Error` we would show that the sum of squares divided by degrees of freedom, which is just 150.89933 / 18 = 8.3832961. We can find our F-value in our ANOVA table by taking the mean square value of `z` dividing it by the mean square value of `Error`, which gives us 2015.30067 / 8.3832961 in which our F-value is 240.3947855. Lastly we can get our p-value by taking our F-value with 1 degree of freedom in the numerator (the model or `z` degrees of freedom) and 18 degrees of freedom in the denominator (the `Error` degrees of freedom). Limiting ourselves to four decimal points, we can find our p-value of our F-value which is F(240.3947855, 1, 18) by using the `pf()` function in R and plugging in the values into it of pf(240.3947855, 1, 18, lower.tail = FALSE) to get a p-value of <.0001. Now we have all the values, we can fill in our ANOVA table.

Source	DF	SS	MS	F-value	p-value
z	1	2015.30067	2015.30067	240.3947855	<.0001
Error	18	150.89933	8.3832961		
Corrected Total	19	2166.20000			

Now we are going to find the R^2 values. We can see from our ANOVA table where we can say that SS(R) = SS(`z`) and the SS(Tot) = SS(`Corrected Total`). Thus, we can get the R^2 by saying that $R^2 = \frac{SS(R)}{SS(Tot)} = \frac{2015.30067}{2166.20000} = 0.9303392$. We can see that our R^2 value is equal to 0.9303392.

4 Problem 4

Here we are going to load in the data and see how it looks.

```
corn_data <- read_csv("cornwyear.csv")
corn_data
```

4.1 Part A

Here we are going to load in each model and make the least squares regression equation by getting the coefficients.

4.1.1 Part i – Model 1

Our first model is $E(Y|x_1) = \beta_0 + \beta_1 * x_1$. We can find the coefficients below.

```
model1_corn <- lm(yield ~ rain, corn_data)
summary(model1_corn)
```

```
##
## Call:
## lm(formula = yield ~ rain, data = corn_data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2014  -2.3530  -0.2577   3.8929   5.7515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.5521     3.2365   7.277 1.43e-08 ***
## rain          0.7755     0.2939   2.639   0.0122 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.049 on 36 degrees of freedom
## Multiple R-squared:  0.1621, Adjusted R-squared:  0.1388
## F-statistic: 6.965 on 1 and 36 DF,  p-value: 0.01221
```

From our output we can find the intercept (β_0) and our slope (β_1). Thus our regression line is $E(Y|x_1) = 23.552102 + 0.7755493 * x_1$.

4.1.2 Part ii – Model 2

Our second model is $E(Y|x_1, x_2) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$. We can find the coefficients below. Note $x_2 = x_1^2$

```
model2_corn <- lm(yield ~ rain + rain2, corn_data)
summary(model2_corn)
```

```
##
## Call:
## lm(formula = yield ~ rain + rain2, data = corn_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4642 -2.3236 -0.1265  3.5151  7.1597
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.01467   11.44158  -0.438  0.66387
## rain         6.00428    2.03895   2.945  0.00571 **
## rain2       -0.22936    0.08864  -2.588  0.01397 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.763 on 35 degrees of freedom
## Multiple R-squared:  0.2967, Adjusted R-squared:  0.2565
## F-statistic: 7.382 on 2 and 35 DF,  p-value: 0.002115
```

From our output we can find the intercept (β_0) and our partial slopes (β_1 and β_2). Thus our regression line is $E(Y|x_1) =$ -252.9588506 + 0.7566133 * x_1 + 0.1449909 * x_2.

4.1.3 Part iii – Model 3

Our third model is $E(Y|x_1, x_3) = \beta_0 + \beta_1 * x_1 + \beta_3 * x_3$. We can find the coefficients below.

```
model3_corn <- lm(yield ~ rain + year, corn_data)
summary(model3_corn)
```

```
##
## Call:
## lm(formula = yield ~ rain + year, data = corn_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7206 -2.2916  0.2468  2.8607  5.5850
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -252.95885  106.10725  -2.384  0.02268 *
## rain           0.75661    0.27282   2.773  0.00884 **
## year           0.14499    0.05562   2.607  0.01333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.758 on 35 degrees of freedom
## Multiple R-squared:  0.2984, Adjusted R-squared:  0.2583
## F-statistic: 7.441 on 2 and 35 DF,  p-value: 0.002028
```

From our output we can find the intercept (β_0) and our partial slopes (β_1 and β_3). Thus our regression line is $E(Y|x_1)$ = -252.9588506 + 0.7566133 * x_1 + 0.1449909 * x_3.

4.1.4 Part iv – Model 4

Our fourth model is $E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_1 * x_3 + \beta_5 * x_2 * x_3$. We can find the coefficients below.

```
model4_corn <- lm(yield ~ rain + rain2 + year + rain_year + rain2_year, corn_data)
summary(model4_corn)
```

```
##
## Call:
## lm(formula = yield ~ rain + rain2 + year + rain_year + rain2_year,
##     data = corn_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3429 -2.5570  0.6387  1.9513  5.0426
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.593e+03  1.938e+03  -0.822    0.417
## rain         9.923e+01  3.554e+02   0.279    0.782
## rain2        2.529e+00  1.606e+01   0.158    0.876
## year         8.351e-01  1.016e+00   0.822    0.417
## rain_year   -4.940e-02  1.863e-01  -0.265    0.793
## rain2_year  -1.423e-03  8.414e-03  -0.169    0.867
```

```
##
## Residual standard error: 3.073 on 32 degrees of freedom
## Multiple R-squared:  0.571,  Adjusted R-squared:  0.5039
## F-statistic: 8.517 on 5 and 32 DF,  p-value: 3.272e-05
```

From our output we can find the intercept (β_0) and our partial slopes (β_1, β_2, β_3, β_4, and β_5). Thus our regression line is $E(Y|x_1)$ = -1592.5855226 + 99.2321766 * x_1 + 2.5290995 * x_2 + 0.8350969 * x_3 + -0.0494003 * $x_1 * x_3$ + -0.001423 * $x_2 * x_3$.

4.1.5 Part v – Model 5

Our fifth model is $E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 * x_1 + \beta_4 * x_1 * x_3 + \beta_5 * x_2 * x_3$. We can find the coefficients below.

```
model5_corn <- lm(yield ~ rain + rain_year + rain2_year, corn_data)
summary(model5_corn)
```

```
##
## Call:
## lm(formula = yield ~ rain + rain_year + rain2_year, data = corn_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3566 -1.9724 -0.1338  2.8176  6.0257
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.592e+00  1.092e+01  -0.421   0.6767
## rain        -1.539e+01  9.436e+00  -1.631   0.1122
## rain_year    1.117e-02  4.899e-03   2.280   0.0290 *
## rain2_year  -1.187e-04  4.432e-05  -2.678   0.0113 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.595 on 34 degrees of freedom
## Multiple R-squared:  0.3762, Adjusted R-squared:  0.3211
## F-statistic: 6.834 on 3 and 34 DF,  p-value: 0.0009992
```

From our output we can find the intercept (β_0) and our partial slopes (β_1, β_4, and β_5). Thus our regression line is $E(Y|x_1)$ = -4.5923964 + -15.3873269 * x_1 + 0.0111709 * $x_1 * x_3$ + $-1.1867866 \times 10^{-4}$ * $x_2 * x_3$.

4.1.6 Part iv – Model 6

Our sixth and final model is $E(Y|x_1, x_3) = \beta_0 + \beta_1 * x_1 + \beta_3 * x_3 + \beta_4 * x_1 * x_3$. We can find the coefficients below.

```
model6_corn <- lm(yield ~ rain + year + rain_year, corn_data)
summary(model6_corn)
```

```
##
## Call:
```

```
## lm(formula = yield ~ rain + year + rain_year, data = corn_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5196 -2.1314 -0.0681  2.2965  5.6912
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.047e+03  5.222e+02  -3.920 0.000407 ***
## rain         1.683e+02  4.800e+01   3.506 0.001299 **
## year         1.085e+00  2.738e-01   3.965 0.000358 ***
## rain_year   -8.781e-02  2.516e-02  -3.490 0.001357 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.272 on 34 degrees of freedom
## Multiple R-squared:  0.4834, Adjusted R-squared:  0.4379
## F-statistic: 10.61 on 3 and 34 DF,  p-value: 4.515e-05
```

From our output we can find the intercept (β_0) and our partial slopes (β_1, β_3, and β_4). Thus our regression line is $E(Y|x_1)$ = -2047.0128548 + 168.2723591 * x_1 + 1.085449 * x_3 + -0.0878101 * $x_1 * x_3$.

4.2 Part B

Conduct F-tests comparing the following pairs of models. For each comparison, state the implicit null hypothesis (H_0) being tested and conduct the test at level $\alpha = .05$. Additionally, report the p-value associated with the test/comparison. Using a policy that adopts the reduced/nested model unless there is "significant" evidence against H_0, specify the model you'd choose for each comparison.

4.2.1 Part i – Model 1 vs. Model 2

Here we are testing to see if we should add the quadratic term of rainfall into our model. In this case, we can say that $H_0 : \beta_2 = 0$ with the $H_A : \beta_2 \neq 0$. We are going to complete an ANOVA test below to see if we should keep this variable.

```
alpha <- 0.05
print(anova(model1_corn, model2_corn))
```

```
## Analysis of Variance Table
##
## Model 1: yield ~ rain
## Model 2: yield ~ rain + rain2
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     36 590.34
## 2     35 495.53  1    94.807 6.6964 0.01397 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, our p-value is 0.0139719. This is less than our alpha level at 0.05. Therefore, we should reject our null hypothesis as we have statistically significant evidence to conclude that $\beta_2 \neq 0$. Therefore, we should use Model 2 when doing our analysis due to β_2 having an effect on our model.

4.2.2 Part ii – Model 3 vs. Model 4

Here we are testing to see if we should add the quadratic term of rainfall into our model. In this case, we can say that $H_0 : \beta_2 = \beta_4 = \beta_5 = 0$ with the $H_A : \beta_2 \neq \beta_4 \neq \beta_5 \neq 0$. We are going to complete an ANOVA test below to see if we should keep these variables.

```
alpha <- 0.05
print(anova(model3_corn, model4_corn))
```

```
## Analysis of Variance Table
##
## Model 1: yield ~ rain + year
## Model 2: yield ~ rain + rain2 + year + rain_year + rain2_year
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     35 494.34
## 2     32 302.28  3    192.06 6.7771 0.001149 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, our p-value is 0.0011489. This is less than our alpha level at 0.05. Therefore, we should reject our null hypothesis as we have statistically significant evidence to conclude that $\beta_2 \neq \beta_4 \neq \beta_5 \neq 0$. Therefore, we should use Model 4 when doing our analysis due to β_2, β_4, and/or β_5 having an effect on our model.

4.2.3 Part iii – Model 4 vs. Model 5

Here we are testing to see if we should add the quadratic term of rainfall into our model. In this case, we can say that $H_0 : \beta_2 = \beta_3 = 0$ with the $H_A : \beta_2 \neq \beta_3 \neq 0$. We are going to complete an ANOVA test below to see if we should keep these variables.

```
alpha <- 0.05
print(anova(model4_corn, model5_corn))
```

```
## Analysis of Variance Table
##
## Model 1: yield ~ rain + rain2 + year + rain_year + rain2_year
## Model 2: yield ~ rain + rain_year + rain2_year
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     32 302.28
## 2     34 439.52 -2   -137.23 7.2637 0.002506 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, our p-value is 0.0025064. This is less than our alpha level at 0.05. Therefore, we should reject our null hypothesis as we have statistically significant evidence to conclude that $\beta_2 \neq \beta_3 \neq 0$. Therefore, we should use Model 5 when doing our analysis due to β_2 and/or β_3 having an effect on our model.

4.2.4 Part iv – Model 1 vs. Model 3

Here we are testing to see if we should add the quadratic term of rainfall into our model. In this case, we can say that $H_0 : \beta_3 = 0$ with the $H_A : \beta_3 \neq 0$. We are going to complete an ANOVA test below to see if we should keep this variable.

```
alpha <- 0.05
print(anova(model1_corn, model3_corn))
```

```
## Analysis of Variance Table
##
## Model 1: yield ~ rain
## Model 2: yield ~ rain + year
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     36 590.34
## 2     35 494.34  1    95.994 6.7965 0.01333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, our p-value is 0.0133325. This is less than our alpha level at 0.05. Therefore, we should reject our null hypothesis as we have statistically significant evidence to conclude that $\beta_3 \neq 0$. Therefore, we should use Model 3 when doing our analysis due to β_3 having an effect on our model.

4.2.5 Part v – Model 2 vs. Model 4

Here we are testing to see if we should add the quadratic term of rainfall into our model. In this case, we can say that $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ with the $H_A : \beta_3 \neq \beta_4 \neq \beta_5 \neq 0$. We are going to complete an ANOVA test below to see if we should keep these variables.

```
alpha <- 0.05
print(anova(model2_corn, model4_corn))
```

```
## Analysis of Variance Table
##
## Model 1: yield ~ rain + rain2
## Model 2: yield ~ rain + rain2 + year + rain_year + rain2_year
##   Res.Df    RSS Df Sum of Sq     F   Pr(>F)
## 1     35 495.53
## 2     32 302.28  3    193.25 6.819 0.001107 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, our p-value is 0.0011075. This is less than our alpha level at 0.05. Therefore, we should reject our null hypothesis as we have statistically significant evidence to conclude that $\beta_3 \neq \beta_4 \neq \beta_5 \neq 0$. Therefore, we should use Model 4 when doing our analysis due to β_3, β_4, and/or β_5 having an effect on our model.

4.2.6 Part vi – Model 6 vs. Model 4

Here we are testing to see if we should add the quadratic term of rainfall into our model. In this case, we can say that $H_0 : \beta_2 = \beta_5 = 0$ with the $H_A : \beta_2 \neq \beta_5 \neq 0$. We are going to complete an ANOVA test below to see if we should keep these variables.

```
alpha <- 0.05
print(anova(model6_corn, model4_corn))
```

```
## Analysis of Variance Table
##
## Model 1: yield ~ rain + year + rain_year
## Model 2: yield ~ rain + rain2 + year + rain_year + rain2_year
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     34 363.94
## 2     32 302.28  2    61.658 3.2636 0.0513 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, our p-value is 0.0513015. This is greater than our alpha level at 0.05. Therefore, we should fail to reject our null hypothesis as we do not have statistically significant evidence to conclude that $\beta_2 \neq \beta_5 \neq 0$. Therefore, we should use Model 6 when doing our analysis due to β_2 and/or β_5 not having an effect on our model.

4.3 Part C

We are looking to see if Model 2 is nested in Model 3. And while we can put the linear restriction of $\beta_2 = 0$ in Model 3, this only gets to $E(Y|x_1) = \beta_0 + \beta_1 * x_1$, which is Model 1 not Model 2. Model 2 has the extra $\beta_2 * x_1^2$ term that can not be obtained from Model 3, no matter what linear restrictions we place on it. This makes sense as well since Model 2 is a quadratic model while Model 3 is linear and we cannot nest a quadratic model within a linear model. For these reasons, Model 2 is not nested within Model 3.

4.4 Part D

Use Model 1 and Model 6 to separately estimate the increase in yield when rainfall increases by 1 inch in the year 1900.

- Model 1: Here we know that $E(Y|x_1) = \beta_0 + \beta_1 * x_1$ where x_1 is rainfall. We know that β_0 remains constant, so we are only looking at β_1 which is telling us that we expect the yield to increase by β_1 amount per one increase in units for x_1 (which in this case is one inch of rainfall). Thus, our estimated increase in yield in 1900 for each increase of one inch of rainfall is just β_1 which is 0.7755493.
- Model 6: Here we know that $E(Y|x_1, x_3) = \beta_0+\beta_1*x_1+\beta_3*x_3+\beta_4*x_1*x_3$, where x_1 is rainfall and x_3 is year. Since year is constant, we can plug in 1900 for each of the x_3's to get our new model. Now we can see that $E(Y|x_1, x_3 = 1900) = \beta_0+\beta_1*x_1+\beta_3*1900+\beta_4*x_1*1900 = (\beta_0+1900*\beta_3)+(\beta_1+1900*\beta_4)*x_1$. The left side of the equation $(\beta_0 + 1900 * \beta_3)$ is a constant so we only need to look at the right side of $(\beta_1 + 1900 * \beta_4) * x_1$ to find out what we would expect our estimated increase in yield in 1900 for each increase of inch in rainfall is just $(\beta_1 + 1900 * \beta_4)$ (where $\beta_1 = 168.2723591$ and $\beta_4 = -0.0878101$) which is 1.4331823. Therefore, we would expect our estimated increase in yield in 1900 for each increase of one inch in rainfall is 1.4331823.

4.5 Part E

Consider Model 3. Is the model well-determined in the sense that both estimated regression coefficients differ significantly from 0? How much do you estimate that mean corn yield increases each year from 1890-1927, controlling for rain? Report the standard error for your estimate. Provide some explanation ("narrative") as to why corn yields appear to be increasing over time even after controlling for rainfall.

First let us take a look at Model 3 again by looking at some of its summary statistics.

```
summary(model3_corn)
```

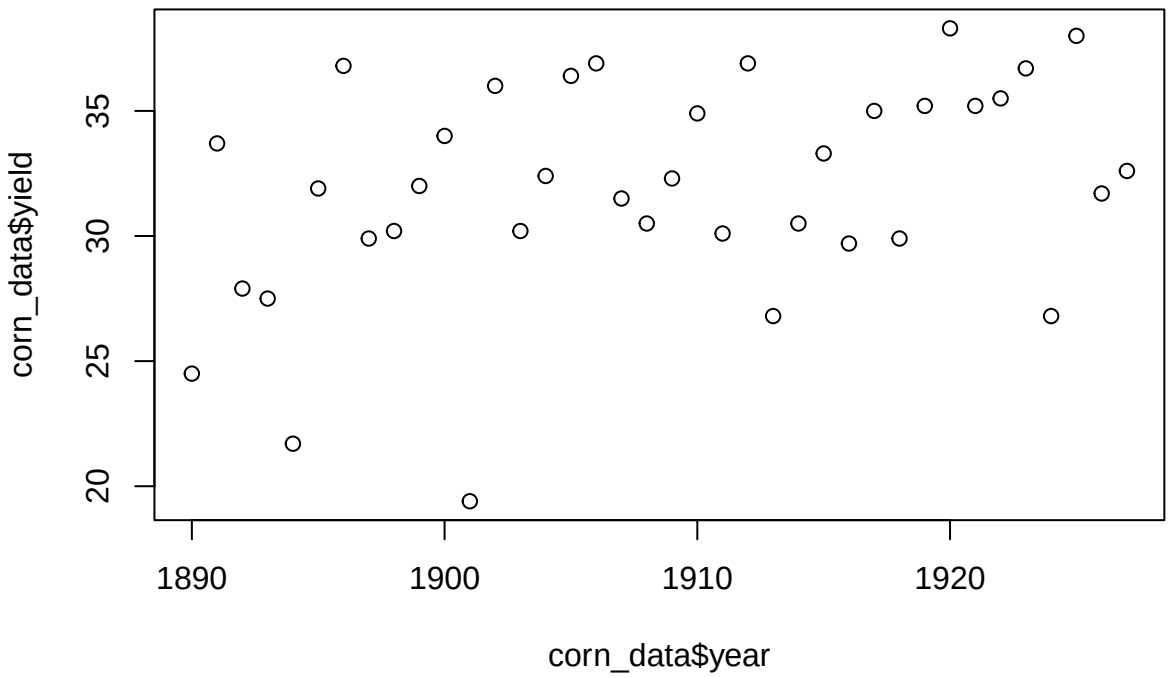
```
##
## Call:
## lm(formula = yield ~ rain + year, data = corn_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7206 -2.2916  0.2468  2.8607  5.5850
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -252.95885  106.10725  -2.384  0.02268 *
## rain           0.75661    0.27282   2.773  0.00884 **
## year           0.14499    0.05562   2.607  0.01333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.758 on 35 degrees of freedom
## Multiple R-squared:  0.2984, Adjusted R-squared:  0.2583
## F-statistic: 7.441 on 2 and 35 DF,  p-value: 0.002028
```

Using the t-statistics for partial slopes we can see that both offer p-values less than 0.05, which means we have statistically significant evidence to say that both estimated regression coefficients differ significantly from 0.

If we control for rain, we can look at our output to see that our partial slope for year is 0.1449909, which means that we expect the mean corn yield to increase by 0.1449909 for each increase of one year from 1890-1927, if we control for rainfall. We can also see that the standard error for `year` is 0.055616, so our standard error of our estimate here is 0.055616.

Corn yield can increase over time, despite looking at rainfall for a number of reasons. First from a data standpoint, we can look at the following in which we can see that some of the lower outliers occur in earlier years.

```
plot(corn_data$year, corn_data$yield)
```



As we can see, there are some dips in the later 1800s to early 1900s. Since we are performing a line of best fit, we would assume that `year` would produce an increasing amount of yield as those select years have a large impact, given the smallish sample size. We can also see in later years that a couple have some spikes further making our model believe that year is a factor (and causing this increasing trend), despite most of the time `year` having a steady amount.

From a non-statistics standpoint, there could also be the confounding variable that technology is getting better over time; as a result, this is increasing corn yield. Cropping techniques might be getting better as well as things tend to evolve over time.

Due to what has been described, this is why that `year` appears to cause an increasing corn yield, even after controlling the rainfall.