STAT 5605 Homework 3

February 19, 2026

Jack Bienvenue

Contents

1 Problem 1

A large, national grocery retailer tracks productivity and costs of its facilities closely. A subset of the data, HW3PR1.txt, was obtained from a single distribution center for a one-year period. The variables included are X_1 = the number of cases shipped, X_2 = the indirect costs of the total labor hours as a percentage, and $X_3 = 1$ if the week has a holiday and 0 otherwise. The response variable Y is the total labor hours.

```
HW3PR1 <- read.table("./HW3PR1.txt", header=TRUE)
str(HW3PR1)
```

```
## 'data.frame':    50 obs. of  5 variables:
##  $ Y   : int  4264 4496 4317 4292 4945 4325 4110 4111 4161 4560 ...
##  $ X1  : int  305657 328476 317164 366745 265518 301995 269334 267631 296350 277223 .
##  $ X2  : num  7.17 6.2 4.61 7.02 8.61 6.88 7.23 6.27 6.49 6.37 ...
##  $ X3  : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ case: int  1 2 3 4 5 6 7 8 9 10 ...
```

```
nrow(HW3PR1)
```

```
## [1] 50
```

```
PR1 <- lm(formula = Y ~ X1+X2+X3, data=HW3PR1)
summary(PR1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = HW3PR1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -267.84 -101.77  -22.09   80.93  297.79
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.100e+03  1.994e+02  20.563  < 2e-16 ***
## X1           9.343e-04  3.794e-04   2.463   0.0176 *
## X2          -1.200e+01  2.319e+01  -0.518   0.6073
## X3           6.189e+02  6.289e+01   9.842 6.73e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.2 on 46 degrees of freedom
## Multiple R-squared:  0.6985, Adjusted R-squared:  0.6788
## F-statistic: 35.52 on 3 and 46 DF,  p-value: 4.907e-12
```

```
new_x_value <- data.frame(X1 = 400000, X2=7.2, X3=0)
predicted_value1 <- predict(object = PR1, newdata = new_x_value,
```

```
                              interval="confidence")
print(predicted_value1)
```

```
##        fit      lwr      upr
## 1 4387.318 4298.682 4475.954
```

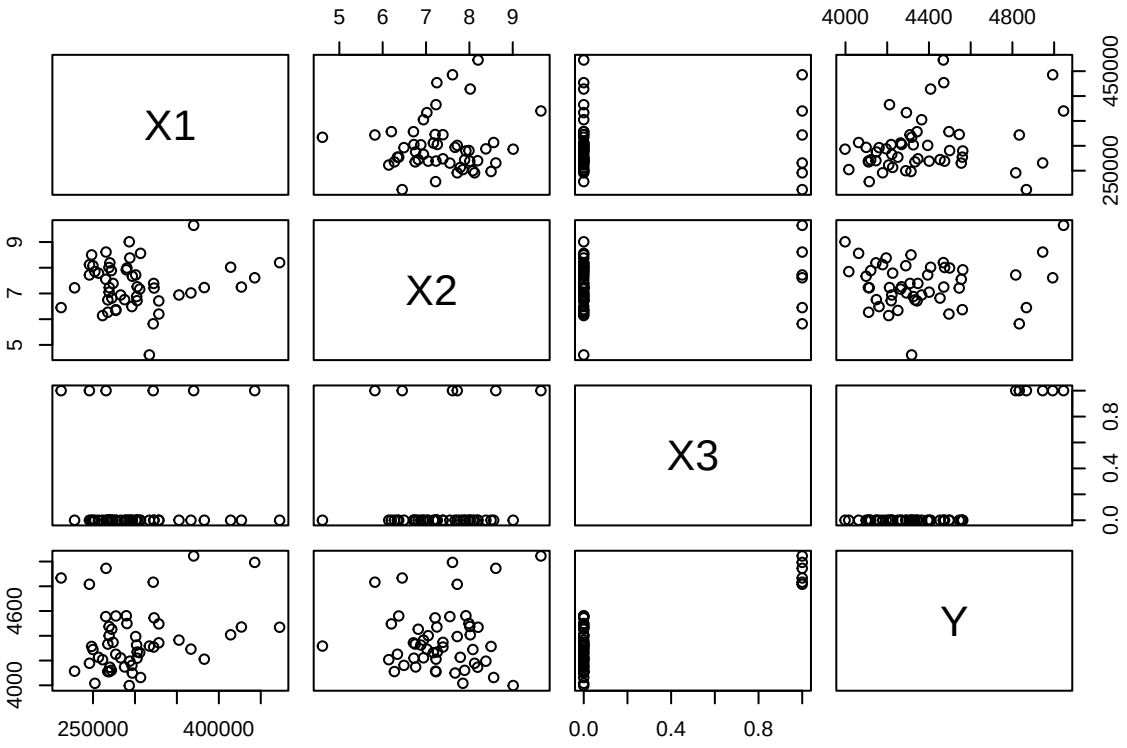
```
predicted_value2 <- predict(object = PR1, newdata = new_x_value,
                              interval="prediction")
print(predicted_value2)
```

```
##        fit      lwr      upr
## 1 4387.318 4085.697 4688.938
```

1.1 (a)

Create a multivariate scatter plot matrix for X_1,X_2, X_3, and Y. What information does this plot provide?

```
HW3PR1a <- select(HW3PR1, X1,X2,X3,Y)
pairs(HW3PR1a)
```



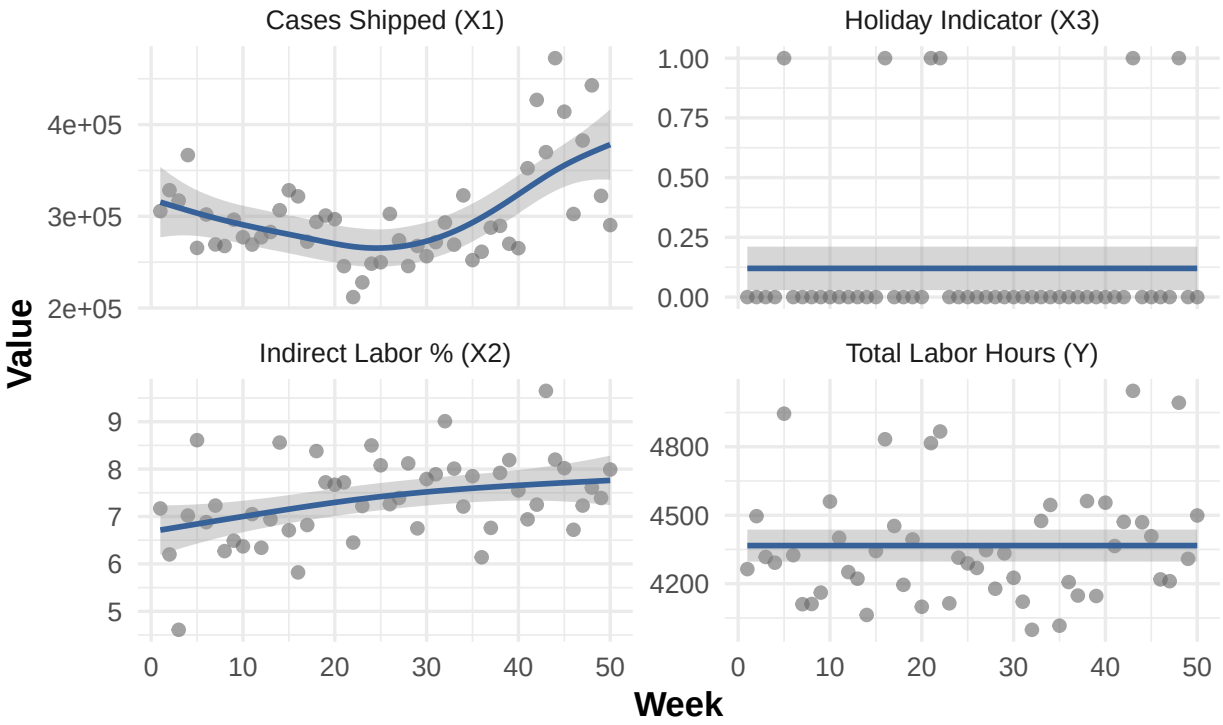
The scatter plot matrix given by this problem allows us to see scatter plots for each possible pair combination of covariates and the response variable. These scatterplots are useful to gauge whether we may have a linear association between variables. This helps in three ways: first, we can observe whether or not it seems apparent that there is a linear association between a covariate and response (which helps us determine whether regression on a variable may explain some variation in Y), second we can observe whether we may have a reason to believe that there is collinearity in our covariates, and third, we can assess the direction of the linear association if there appears to be an association. Further, the scatterplot matrix can help us view the structure of the data for each variable, as we may be interested in the range of the data or whether the data has factor levels. This particular scatterplot matrix has some interesting information. A clear linear association does not appear to exist between X_1 and X_2. The coincidence of the binary variable X_3 does not show clear clustering of X_3 in relation to values of X_1 and X_2. For relationships with Y, we do not see very clear signs of linear association for X_1 and X_2 with Y. We do see a clear grouping of the holidays in X_3 related to Y: it appears that all the holidays have higher associated values of Y. We can determine whether or not linear associations exist for X_1 and X_2 and Y and whether using a dummy variable for X_3 helps with explaining variation in Y more formally using hypothesis testing techniques. This graphical method is good for exploring data initially but we may not obtain the full story of the variables' relationships from this scatter plot matrix.

1.2 (b)

The cases in HW3PR1.txt are given in consecutive weeks. Prepare a time plot for each predictor variable as well as the response variable Y. What do these plots show?

```
# Data prep for nice time series plot:
HW3PR1_long <- HW3PR1 %>% mutate(Week = case) %>%
  select(Week, Y, X1, X2, X3) %>%
  pivot_longer( cols = c(Y, X1, X2, X3),
                names_to = "Variable",
                  values_to = "Value")
HW3PR1_long$Variable <- recode(HW3PR1_long$Variable,
  Y  = "Total Labor Hours (Y)",
  X1 = "Cases Shipped (X1)",
  X2 = "Indirect Labor % (X2)",
  X3 = "Holiday Indicator (X3)"
)

# 2x2 Scatterplot of the time series for the variables:
ggplot(HW3PR1_long, aes(x = Week, y = Value)) +
  geom_point(alpha = 0.6, color = "gray40") +
  geom_smooth(
    method = "gam",
```

```r
  formula = y ~ s(x, bs = "cs"),  # cubic spline fit
  se = TRUE,
  color = "#376199",
  linewidth = 1
) +

facet_wrap(~ Variable, ncol = 2, scales = "free_y") +

labs(
  title = "Weekly Time Series for Grocery Store Data Variables",
  x = "Week",
  y = "Value",
  caption = "Cubic Spline Fit Used for Time Series | Jack Bienvenue 2026"
) +

theme_minimal(base_size = 13) +
theme(
plot.title = element_text(
  face = "bold",
  hjust = 0.5),
axis.title = element_text(
  face = "bold"),
plot.caption = element_text(
  color = "gray60"))
```


Weekly Time Series for Grocery Store Data Variables
Cases Shipped (X1)
Holiday Indicator (X3)
Indirect Labor % (X2)
Total Labor Hours (Y)
Value
Week
4e+05
3e+05
2e+05
1.00
0.75
0.50
0.25
0.00
9
8
7
6
5
4800
4500
4200
0
10
20
30
40
50
Cubic Spline Fit Used for Time Series | Jack Bienvenue 2026

In this part of the problem, we would like to assess whether there is clear seasonality in the trends of any of the predictors or the response variable. We should realize that, if we were to create predictive models or make inferences based upon this data, that inferences or predictions may be inaccurate because of time-dependence issues.
These plots provide several insights into how time dependence plays a role in this data. For the top-left plot, we observe a potentially meaningful time series signal of cases shipped increasing on average towards the end of the year, potentially associated with the holiday season assuming that Week 0 indicates the beginning of the calendar year. Following this assumption, we see shipments gradually decrease on average as the winter and spring progress, bottoming in the summer and increasing again on average as the fall progresses.
The signal for indirect labor percentages is less clear, but suggests that the percentage may gradually increase on average as the year progresses.
The holiday indicator's plot is not helpful in its own right, but provides clarity on where the holidays occur during the year. We can observe that 12% of the weeks in the year are considered holidays.
When reviewing the outcome variable, we can see that the total labor hours are on average relatively stagnant throughout the year, not exhibiting obvious seasonality. However, in combination with the holiday indicator plot, we can observe that in this data set, the highest total labor hours week coincide with the holiday weeks.
We should note when reviewing these time series plots that we are observing the time series over a single year and that the trends we observe over a single year may not be fully representative of the true time series for these variables.

1.3 (c)

Test whether there is a regression relation. State the hypotheses and conclusion. What does your test result imply about regression coefficients (slopes) associated with X_1, X_2, and X_3? What is the p-value of the test?

Before performing our test, let's set the hypotheses for our testing:

$$H_0 : \beta_1 = \beta_2 = \beta_2 = 0, \text{ versus}$$

$$H_A : \text{At least one of } \beta_1, \beta_2, \beta_3 \neq 0.$$

We select this particular pair of hypotheses because we testing simply for whether a regression relation exists in our linear model, regardless of whether every single regressor exhibits a regression relation with the outcome variable. Furthermore, we will decide to use an $\alpha = 0.05$ significance level for our test.
Our regression output provides information for this test automatically, with a p-value of 4.907×10^{-12}, far below the designated threshold. Therefore, we will reject the null hypothesis which stated that none of the regression coefficients $\beta_1, \beta_2, \beta_3$ were nonzero. We have reason to believe that the model overall has a linear regression relationship with the outcome variable. In essence, an MLR model based upon the number of cases shipped, the indirect costs of total labor hours (%) and holiday indicator helps to explain some of the variation in total labor hours at the supermarket.

1.4 (d)

Calculate the coefficient multiple determination R^2. How is this measure interpreted here.

The multiple R^2 is included in our regression output and is equal to 0.6985. This means that the linear relationship between Y and X_1, X_2, and X_3 explains about 69.85% of the variability in Y, the total labor hours in a week at the grocery store.

1.5 (e)

For the data, HW3PR1.txt, on which the regression fit is based, would you consider a shipment of 400,000 cases with an indirect percentage of 7.2 on a non-holiday week to be within the scope of the model? What about a shipment of 400,000 cases with an indirect percentage of 9.9 on a non-holiday week? Support your answers by preparing or referring to a relevant plot.

```
# Check sample range of data

## X_1
summary(HW3PR1$X1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  211944  268035  290060  300724  320621  472476
```

```
## X_2
summary(HW3PR1$X2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.610   6.775   7.255   7.353   7.973   9.650
```

```
## Y
summary(HW3PR1$Y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3998    4198    4316    4367    4474    5045
```

Let's review the sample range of the variables from which we built our regression model.
SAMPLE RANGES:

1. X_1: $[211944, 472476]$ (# Cases Shipped),
2. X_2: $[4.610, 9.650]$ (Indirect Labor %),
3. X_3:
   $0, 1$ (Holiday Indicator $1 =$ Holiday),
4. Y: $[3998, 5045]$ (Total Labor Hours)

So, for an observation $\begin{pmatrix} 400000 \\ 7.2 \\ 0 \\ Y_j \end{pmatrix}$, every $X_i, i = 1, 2, 3$ element is within the range of data used to generate our least squares fit. Therefore, we can consider this observation to be within the scope of the model.

For an observation $\begin{pmatrix} 400000 \\ 9.9 \\ 0 \\ Y_{j'} \end{pmatrix}$, the value of X_2 exceeds the range of the data used to construct the model, and therefore the observation is outside the scope of the model.
In the figure above, we can see where the two observations' feature values lay in relation to the range of data used to construct the model.

1.6 (f)

Assume that multiple regression model for three predictor variables with independent normal error terms is appropriate. A new shipment is to be received with $x_{h1} = 292,087$, $x_{h2} = 7.77$, and $x_{h3} = 0$.

1.6.1 (f1)

Obtain a 95% confidence interval for the expected total labor hours for this shipment.

```
# Find relevant t critical value for computing intervals:
t_0.975_46 <- qt(0.975, 46)
cat("Critical t:", t_0.975_46)
```

```
## Critical t: 2.012896
```

```
print('')
```

```
## [1] ""
```

```
# Verify fitted value using model:
new_obs <- data.frame(X1 = 292087, X2 = 7.77, X3 = 0)
predict(PR1, new_obs, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 4279.658 4230.776 4328.539
```

We can start by building a 95% confidence interval for the expected total labor hours for this shipment using the construction:

$$95\%\ \text{Conf. Int.} = \hat{Y}_h \pm t_{0.975,n-4} \times \text{SE}(\hat{Y}_h) = \hat{Y}_h \pm t_{0.975,n-4} \times \sqrt{\hat{\sigma^2}\boldsymbol{x}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_h}},$$

which is equal to $(4230.776, 4328.539)$.
To calculate our confidence interval bounds, we must obtain the fitted value $\hat{Y}_h = \boldsymbol{x}_h'\hat{\boldsymbol{\beta}}$. Using our fitted models and known value for the observed $X_{ih}, i-1,2,3$, we can calculate the fitted value as $\begin{pmatrix}1 & 292087 & 7.77 & 0\end{pmatrix}\begin{pmatrix}4100\\0.0009343\\-12.00\\613.9\end{pmatrix} = 279.658$
The critical value for creating a 95% confidence interval in this scenario is $t_{0.975,n-4} = t_{0.975,46} = 2.012896$. The standard error for the fitted mean is $\text{SE}(\hat{Y}_h) = \sqrt{\hat{\sigma^2}\boldsymbol{x}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_h}}$.

1.6.2 (f2)

Compute a 95% prediction interval for the total labor hours for this new shipment.

```
# Use model to predict interval:
predict(PR1, new_obs, interval = "prediction", level = 0.95)
```

```
##        fit     lwr      upr
## 1 4279.658 3987.24 4572.075
```

We can continue by building a 95% prediction interval for the total labors hours for this new shipment using the construction:
Here, what changes for a prediction interval as opposed to the confidence interval for the expected total hours is an expansion of the interval by inflating the standard error. For the prediction interval, our standard error formulation is changed to $\text{SE}_{\text{pred}}(\hat{Y}_h) = \sqrt{\hat{\sigma^2}(1 + \boldsymbol{x}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_h})}$. So, our interval is:

$$95\%\ \text{Pred. Int.} = \hat{Y}_h \pm t_{0.95,n-4} \times \sqrt{\hat{\sigma^2}(1 + \boldsymbol{x}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_h})} = (3987.24, 4572.075).$$

2 Problem 2

An analyst wanted to fit a regression model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$, $i = 1, 2, \ldots, n$, by the method of least squares when it is known that $\beta_2 = 5$. How can the analyst obtain the desired fit by using a multiple regression computer program? What if it is known that $\beta_3 = 0$?

First Case: To generate a least squares fit for MLR where one coefficient (e.g., $\beta_2 = 0$) is already known, we can adjust our response variable by subtracting the true coefficient multiplied by its input covariate X_{i2}, changing our output variable for our MLR to be $Y_i^\star = Y_i - 5X_{i2}$. Then, we can use the usual least squares method to regress the adjusted response on the remaining covariates and unknown parameters,

$$Y_i^\star = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \epsilon_i$$

Second Case: I'm going to assume that the problem statement indicates that β_2 is now unknown again and that β_3 alone is known, and is equal to zero. In this circumstance, the model can be reduced to $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$ and fit using least squares as a MLR model regressing on X_{i1} and X_{i2} with an intercept term included. If we do actually know that $\beta_2 = 5$, then our regression model could become:

$$Y_i^\star = \beta_0 + \beta_1 X_{i1} + \epsilon_i.$$

So, if the analyst is clever about how they implement the known values, they can proceed with building regression models with no issue.

3 Problem 3

Consider the multiple linear regression model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$. Which of the following statements would be correct? Provide a necessary explanation.

3.1 (a)

The number of predictors is 2.

INCORRECT* - We have three predictors that we are using for our model: X_{i1},X_{i2}, and $X_{i1}X_{i2}$.
However, if we are considering predictors to be unique features, we have two unique features here: X_1 and X_2. We also have an interaction term X_1X_2 formed from X_1 and X_2. Under this circumstance, we could claim that this statement is true.

3.2 (b)

The number of regressors is 3.

INCORRECT* - We are using four regressors here, $1, X_{i1}, X_{i2},$ and $X_{i1}X_{i2}$, multiplied by $\beta_0, \beta_1, \beta_2, \beta_3$, respectively.
If we do not consider the intercept term to be a regressor, then we could say we have three distinct regressors for Y: $\beta_1 X_{i1}$, $\beta_2 X_{i2}$, $\beta_3 X_{i1} X_{i2}$.

3.3 (c)

The number of model parameters is 3.

INCORRECT - We have five distinct parameters for our model: $\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2$.

3.4 (d)

The number of model parameters is 4.

INCORRECT - We have five distinct parameters for our model: $\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2$.

3.5 (e)

The number of model parameters is 5.

CORRECT - the model parameters are $\beta_0, \beta_1, \beta_2, \beta_3$ and the variance of the error term, σ^2. This makes for a total of 5 parameters in the model.

3.6 (f)

The number of model parameters is ∞.

INCORRECT - our model has closed form $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$ with 5 uniquely identifiable parameters as we observed for the previous statement. The number of model parameters therefore is finite and equal to 5, the number of model parameters is not ∞.

3.7 (g)

None of the above.

INCORRECT - We found at least **part e** to be true.

4 Problem 4

An analysis is performed to study the relationship between three explanatory variables, X_1, X_2 and X_3, and a response variable Y. Consider the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$. The observed data is denoted by $\{(Y_i, X_{1i}, X_{2i}, X_{3i}),\ i = 1, 2, \ldots, n\}$ and we further assume that the Y_i's are independent. We fit the above model to this data set. The resulting ANOVA table is given below and the coefficient of determination is 0.637538.

The ANOVA Table

Analysis of Variance					
Source	DF	Sum of Square	Mean Square	F Stat	Prob > F
Model	*	*	*	*	*
Error	117	17.90761	0.15306		
C Total	*	*			

Answer the following questions:

4.1 (i)

Fill in the missing values (denoted by "*'') in the ANOVA table.

```
prob4_1_quantity <- pf(65.72, 3, 117, lower.tail = FALSE)
print(prob4_1_quantity)
```

```
## [1] 5.563281e-25
```

In this problem, we are given a small amount of information, but we should be able to work backwards to get all of the quantities we need to fill our table.
Besides the information about the error degrees of freedom, associated SS, and MS, we know that the coefficient of determination $R^2 = 0.627538 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$. Right away, we can actually find $SSTO$ and SSR:

$$\frac{SSE}{SSTO} = 1 - 0.627538,$$

Leading to:

$$\frac{17.90761}{SSTO} = 0.372462 \implies SSTO = \frac{17.90761}{0.372462} \approx 48.07903.$$

Therefore our SSR is:

$$SSR = 0.627538 \times 48.07903 \approx 30.17142.$$

Moving on, we know in MLR that the degrees of freedom for the error term is $n - p$, with p representing the number of parameters to be estimated. Here, we have four parameters to be estimated, so using our given error degrees of freedom 117, we can find that our sample size is $n = 121$. To find the model and total degrees of freedom, we recognize that the MLR model degrees of freedom is equal to the number of predictors, 3, and the corrected total degrees of freedom is equal to $n - 1$, so here the corrected total degrees of freedom would be 120.
To find the relevant mean square error for the regression model, we simply need to divide the SSR by its corresponding degrees of freedom, $\frac{30.17142}{3} = 10.05714$.
Finally, we can round out our table by finding the F Statistic and its corresponding p-value by first identifying the F Statistic's value, $F = \frac{MSR}{MSE} = \frac{10.05714}{0.15306} \approx 65.70717$. This means that we can find the p-value through calculating the probability $P(F_{3,117} > 65.72)$. This probability will be extremely close to zero. Comfortably we can claim that the p-value would end up being $\ll 0.0001$.
So, our completed table is:

The ANOVA Table

Analysis of Variance

Source	DF	Sum of Square	Mean Square	F Stat	Prob > F
Model	3	30.17142	10.05714	65.72	$\ll 0.0001$
Error	117	17.90761	0.15306		
C Total	120	48.07903			

4.2 (ii)

State the null and alternative hypotheses (H_0 and H_1) for the F test in the ANOVA table.

For the type of F test employed by our table setup, we are testing whether the linear regression model overall explains any variation in Y. Therefore we are testing the hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0, \text{ against}$$

$$H_A : \text{ At least one of } \beta_1, \beta_2, \beta_3 \neq 0$$

We found that there was a linear association between at least one of the coefficient-covariate pairs and the response variable using our F-test.

4.3 (iii)

What is an estimated value of σ^2 based on the results shown in the above ANOVA table?

The value of σ^2 can be estimated by $\hat{\sigma^2} = MSE = 0.15306$.

4.4 (iv)

Under the null hypothesis H_0 specified in Part (ii), find the distribution of R^2 and then compute $P(R^2 \geq 0.637538|H_0)$. What does the value of $P(R^2 \geq 0.637538|H_0)$ imply?

```
prob4_4_quantity <- pbeta(0.637538, (3/2), (117/2), lower.tail = FALSE)
print(prob4_4_quantity)
```

```
## [1] 1.148074e-25
```

Under the null hypothesis, our model would reduce to $Y = \beta_0 + \epsilon_i$. In multiple linear regression, under the null hypothesis the coefficient of determination is Beta-distributed with parameters that are equal to half of the degrees of freedom associated with the model and error. So, under the null hypothesis for this problem, $R^2 \sim \text{Beta}\left(\frac{3}{2}, \frac{117}{2}\right)$. Now since we have a defined distribution for R^2 under the null hypothesis, we can compute the probability that R^2 exceeds 0.637538 under the null hypothesis. We can generate this particular value using R, which turns out to be incredibly small. This extremely low probability indicates that it is incredibly improbable to observe a value of $R^2 \geq 0.637538$ given that the true model is $Y = \beta_0 + \epsilon_i$.

5 Problem 5

Consider the multiple regression model:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where the ϵ_i are uncorrelated, with $E[\epsilon_i] = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$, and X_{1i} and X_{2i} are two covariates. Let

$$X = \begin{pmatrix} X_{11} & X_{21} & \ldots & X_{n1} \\ X_{12} & X_{22} & \ldots & X_{n2} \end{pmatrix}',$$

which is an $n \times 2$ matrix. Assume that $(X_{1j}, X_{2j}, \ldots, X_{nj})' \neq (1, 1, \ldots, 1)'$ for $j = 1, 2$ and $X'X$ is of full rank, i.e., $|X'X| \neq 0$.

5.1 (a)

Derive the normal equations using the LS criterion.

Using the LS criterion means minimizing the score function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \beta_1 X_{i1} - \beta_2 X_{i2})^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

To get the normal equations, we can differentiate this function w.r.t. $\boldsymbol{\beta}$ and subsequently set the derivative equal to 0:

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \implies \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$$

5.2 (b)

Derive the LS estimators of β_1 and β_2.

The LS estimators of β_1 and β_2 are the selections of the elements of $\boldsymbol{\beta}$ that solve the normal equations and are $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$.
The Gram matrix $\boldsymbol{X'X}$ is a matrix in $\mathbb{R}^{2\times 2}$ which we can calculate to be:

$$\begin{pmatrix} X_{11} & X_{21} & \dots & X_{n1} \\ X_{12} & X_{22} & \dots & X_{n2} \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \\ \vdots & \vdots \\ X_{n1} & X_{n2} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n X_{i1}^2 & \sum_{i=1}^n X_{i1}X_{i2} \\ \sum_{i=1}^n X_{i1}X_{i2} & \sum_{i=1}^n X_{i2}^2 \end{pmatrix}$$

The inverse of this matrix can be expressed by first finding the determinant, $|\boldsymbol{X'X}| = (\sum_{i=1}^n X_{i1}^2)(\sum_{i=1}^n X_{i2}^2) - (\sum_{i=1}^n X_{i1}X_{i2})^2$, and then expressing the matrix as:

$$\frac{1}{(\sum_{i=1}^n X_{i1}^2)(\sum_{i=1}^n X_{i2}^2) - (\sum_{i=1}^n X_{i1}X_{i2})^2} \begin{pmatrix} \sum_{i=1}^n X_{i2}^2 & -\sum_{i=1}^n X_{i1}X_{i2} \\ -\sum_{i=1}^n X_{i1}X_{i2} & \sum_{i=1}^n X_{i1}^2 \end{pmatrix}$$

This inverse Gram matrix is then multiplied by:

$$\boldsymbol{X'Y} = \begin{pmatrix} X_{11} & X_{21} & \dots & X_{n1} \\ X_{12} & X_{22} & \dots & X_{n2} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n Y_i X_{i1} \\ \sum_{i=1}^n Y_i X_{i2} \end{pmatrix}$$

So, together, our expression for the LS estimators of β_1 and β_2 is:

$$\hat{\boldsymbol{\beta}} = \frac{1}{(\sum_{i=1}^n X_{i1}^2)(\sum_{i=1}^n X_{i2}^2) - (\sum_{i=1}^n X_{i1}X_{i2})^2} \begin{pmatrix} \sum_{i=1}^n X_{i2}^2 & -\sum_{i=1}^n X_{i1}X_{i2} \\ -\sum_{i=1}^n X_{i1}X_{i2} & \sum_{i=1}^n X_{i1}^2 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n Y_i X_{i1} \\ \sum_{i=1}^n Y_i X_{i2} \end{pmatrix} =$$

$$= \frac{1}{(\sum_{i=1}^n X_{i1}^2)(\sum_{i=1}^n X_{i2}^2) - (\sum_{i=1}^n X_{i1}X_{i2})^2} \begin{pmatrix} \sum_{i=1}^n X_{i2}^2 \sum_{i=1}^n Y_i X_{i1} - \sum_{i=1}^n X_{i1}X_{i2} \sum_{i=1}^n Y_i X_{i2} \\ -\sum_{i=1}^n X_{i1}X_{i2} \sum_{i=1}^n Y_i X_{i1} + \sum_{i=1}^n X_{i1}^2 \sum_{i=1}^n Y_i X_{i2} \end{pmatrix}$$

$$\therefore \hat{\boldsymbol{\beta}} = \begin{pmatrix} \frac{\sum_{i=1}^n X_{i2}^2 \sum_{i=1}^n Y_i X_{i1} - \sum_{i=1}^n X_{i1}X_{i2} \sum_{i=1}^n Y_i X_{i2}}{(\sum_{i=1}^n X_{i1}^2)(\sum_{i=1}^n X_{i2}^2) - \left(\sum_{i=1}^n X_{i1}X_{i2}\right)^2} \\ \frac{-\sum_{i=1}^n X_{i1}X_{i2} \sum_{i=1}^n Y_i X_{i1} + \sum_{i=1}^n X_{i1}^2 \sum_{i=1}^n Y_i X_{i2}}{(\sum_{i=1}^n X_{i1}^2)(\sum_{i=1}^n X_{i2}^2) - \left(\sum_{i=1}^n X_{i1}X_{i2}\right)^2} \end{pmatrix}$$

5.3 (c)

Let e_i denote the residual and also let $\hat{Y}_i$ denote the fitted value for $i = 1, 2, \ldots, n$. Which of the following statements are true?

5.3.1 (c1) $\sum_{i=1}^{n} e_i = 0$ always.

FALSE: This is true in SLR and MLR cases when an intercept is included, because a normal equation would give $\sum e_i = 0$, but for our particular intercept-free model, this constraint is not forced by our normal equations.

5.3.2 (c2) $\sum_{i=1}^{n} e_i \hat{Y}_i = 0$ always.

TRUE: Under the least squares method, residuals are always orthogonal to fitted values. Least squares gives us $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and so $\boldsymbol{e}'\hat{\boldsymbol{Y}} = \boldsymbol{e}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{e})'\hat{\boldsymbol{\beta}} = 0$ because $(\boldsymbol{X}'\boldsymbol{e} = 0)$ due to the normal equations dictating orthogonality of the predictors and residuals.

5.3.3 (c3) $\sum_{i=1}^{n} e_i X_{ij} = 0$ always for $j = 1, 2$.

TRUE: This is true, and is how we proved the prior statement was true. For SLR and MLR regression models using predictors $X_{\cdot j}$, each predictor will generate a normal equation where $\sum_{i=1}^{n} e_i X_{ij} = 0$.

5.3.4 (c4) $\sum_{i=1}^{n} e_i Y_i = 0$ always.

FALSE: We know that we can represent the true $\boldsymbol{Y}$ values as $\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{e}$, and pre-multiplying by $\boldsymbol{e}$ we get $\boldsymbol{e}'\boldsymbol{Y} = \boldsymbol{e}'\hat{\boldsymbol{Y}} + \boldsymbol{e}'\boldsymbol{e}$. In part c2, we showed that $\boldsymbol{e}'\hat{\boldsymbol{Y}} = 0$. Following that realization we get $\boldsymbol{e}'\boldsymbol{Y} = \sum e_i^2$. This is only zero if all residuals are 0, which does not happen always (or nearly ever in practice). Therefore, the statement is false.

6 Problem 6

In a data set with 2 covariates and 100 observations, the sample variance for the responses is 8.158. When fitting a linear regression on the data, the F statistics for testing the usefulness of the overall model is 41.26, the residual for the first observation is -2.393, and its standard error is $\sqrt{4.343}$. Find the standard error of the fitted value for the first observation.

To get the standard error of a fitted value, we can use a formula that related this standard error to the mean squared error and leverage of the observation: $\text{SE}(\hat{Y}_1) = \sqrt{\text{MSE} \times h_{11}}$. Let's see if the information provided can help us get to the leverage of the first observation and the mean squared error.

Leverage appears in the formula for the variance of the error term: $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$. The variance of the error can be replaced by $\text{SE}(e_i)^2 = (\sqrt{4.343})^2 = 4.343$ and we can substitute unknown σ^2 with its appropriate estimator MSE. So, we can isolate leverage:

$$4.343 = \text{MSE}(1 - h_{ii}) \implies h_{ii} = 1 - \frac{4.343}{\text{MSE}}.$$

Now, we need to find MSE to proceed with our calculation of the standard error for the fitted value of the first observation.

We are given the F-statistic, and we know that the F-statistic is calculated as $F = \frac{MSR}{MSE} = \frac{\frac{SSR}{2}}{\frac{SSE}{97}}$. (These degrees of freedom represent the number of predictors, and $n - p$ where p is the number of coefficients in the model).

The sample variance of Y is $s_Y^2 = 8.158$, which can be related to the total sum of squares SSTO which can help us to figure out the decomposition values of SSE and SSR. The total sum of squares, $\text{SSTO} = (n-1)s_Y^2 = 99 \times 8.158 = 807.642$. While we do not how the total sum of squares is distributed to SSR and SSE, we can try to find these values, starting indirectly.

The regression sum of squares can alternatively be expressed as $\text{SSR} = F \times k \times \text{MSE}$, and the error sum of squares can alternatively be expressed as $\text{SSE} = (n-p)\text{MSE}$. So, we can express the total sum of squares in terms of its decomposition:

$$\text{SSTO} = Fk\text{MSE} + (n-p)\text{MSE} \implies 807.642 = (41.26)(2)\text{MSE} + (97)\text{MSE},$$

Which can help us yield MSE as:

$$\text{MSE} = \frac{807.642}{(41.26)(2) + 97} \approx 4.499$$

Now we can finally use our attained MSE in order to calculate the standard error of the fitted value for the first observation:

$$\text{SE}(\hat{Y}_1) = \sqrt{\text{MSE} \times \left(1 - \frac{4.343}{\text{MSE}}\right)} = \sqrt{4.499 \times \left(1 - \frac{4.343}{4.499}\right)} \approx 0.395.$$