# C1M2_peer_reviewed

March 4, 2025

## 1 Module 2: Peer Reviewed Assignment

### 1.0.1 Outline:

The objectives for this assignment:

1. Mathematically derive the values of $\hat{\beta}_0$ and $\hat{\beta}_1$
2. Enhance our skills with linear regression modeling.
3. Learn the uses and limitations of RSS, ESS, TSS and $R^2$.
4. Analyze and interpret nonidentifiability.

General tips:

1. Read the questions carefully to understand what is being asked.
2. This work will be reviewed by another human, so make sure that you are clear and concise in what your explanations and answers.

[3]:

```
# Load Required Packages
library(RCurl) #a package that includes the function getURL(), which allows for␣
 ↪reading data from github.
library(tidyverse)
```

```
  Attaching packages                                   tidyverse
1.3.0

  ggplot2 3.3.0       purrr   0.3.4
  tibble  3.0.1       dplyr   0.8.5
  tidyr   1.0.2       stringr 1.4.0
  readr   1.3.1       forcats 0.5.0

  Conflicts
tidyverse_conflicts()
  tidyr::complete() masks
RCurl::complete()
  dplyr::filter()   masks
stats::filter()
  dplyr::lag()      masks stats::lag()
```

## 1.1 Problem 1: Maximum Likelihood Estimates (MLEs)

Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, ..., n,\ \ \varepsilon_i \sim N(0, \sigma^2)$. In the videos, we showed that the least squares estimator in matrix-vector form is $\widehat{\beta} = (\beta_0, \beta_1)^T = (X^T X)^{-1} X^T \mathbf{Y}$. In this problem, you will derive the least squares estimators for simple linear regression without (explicitly) using linear algebra.

Least squares requires that we minimize

$$f(\mathbf{x}; \beta_0, \beta_1) = \sum_{i=1}^{n} \Big( Y_i - [\beta_0 + \beta_1 x_i] \Big)^2$$

over $\beta_0$ and $\beta_1$.

**1. (a) Taking Derivatives** Find the partial derivative of $f(\mathbf{x}; \beta_0, \beta_1)$ with respect to $\beta_0$, and the partial derivative of $f(\mathbf{x}; \beta_0, \beta_1)$ with respect to $\beta_1$. Recall that the partial derivative with respect to $x$ of a multivariate function $h(x, y)$ is calculated by taking the derivative of $h$ with respect to $x$ while treating $y$ constant.

[ ]:

```
X <- b0
Y <- b1

X_bar <- mean(b0)
Y_bar <- mean(b1)
```

**1. (b) Solving for $\hat{\beta}_0$ and $\hat{\beta}_1$** Use **1. (a)** to find the minimizers, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, of $f$. That is, set each partial derivative to zero and solve for $\beta_0$ and $\beta_1$. In particular, show

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \text{and} \qquad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}$$

[ ]:

```
beta1 <- sum((X - X_bar) * (Y - Y_bar))/sum((X - X_bar)^2)
beta0 <- Y_bar - beta1 * X_bar

model <- lm(Y ~ X)
summary(model)$coefficients
```

## 1.2 Problem 2: Oh My Goodness of Fit!

In the US, public schools have been slowly increasing class sizes over the last 15 years [https://stats.oecd.org/Index.aspx?DataSetCode=EDU_CLASS]. The general cause for this is because it saves money to have more kids per teacher. But how much money does it save? Let's use some of our new regression skills to try and figure this out. Below is an explanation of the variables in the dataset.

Variables/Columns:
School
Per-Pupil Cost (Dollars)
Average daily Attendance
Average Monthly Teacher Salary (Dollars)
Percent Attendance
Pupil/Teacher ratio

Data Source: E.R. Enlow (1938). “Do Small Schools Mean Large Costs?,” Peabody Journal of Educaltion, Vol. 16, #1, pp. 1-11

[13]:

```
school.data = read_table("school.dat")
names(school.data) = c("school", "cost", "avg.attendance", "avg.salary", "pct.
 ↪attendance", "pup.tch.ratio")
head(school.data)
dim(school.data)
```

```
Parsed with column specification:
cols(
  Adair = col_character(),
  `66.90` = col_double(),
  `451.4` = col_double(),
  `160.22` = col_double(),
  `90.77` = col_double(),
  `33.8` = col_double()
)
```

A tibble: 6 × 6

| school<br>&lt;chr&gt; | cost<br>&lt;dbl&gt; | avg.attendance<br>&lt;dbl&gt; | avg.salary<br>&lt;dbl&gt; | pct.attendance<br>&lt;dbl&gt; | pup.tch.ratio<br>&lt;dbl&gt; |
|---|---|---|---|---|---|
| Calhoun | 108.57 | 219.1 | 161.79 | 89.86 | 23.0 |
| Capitol View | 70.00 | 268.9 | 136.37 | 92.44 | 29.4 |
| Connally | 49.04 | 161.7 | 106.86 | 92.01 | 29.4 |
| Couch | 71.51 | 422.1 | 147.17 | 91.60 | 29.2 |
| Crew | 61.08 | 440.6 | 146.24 | 89.32 | 36.3 |
| Davis | 105.21 | 139.4 | 159.79 | 86.51 | 22.6 |

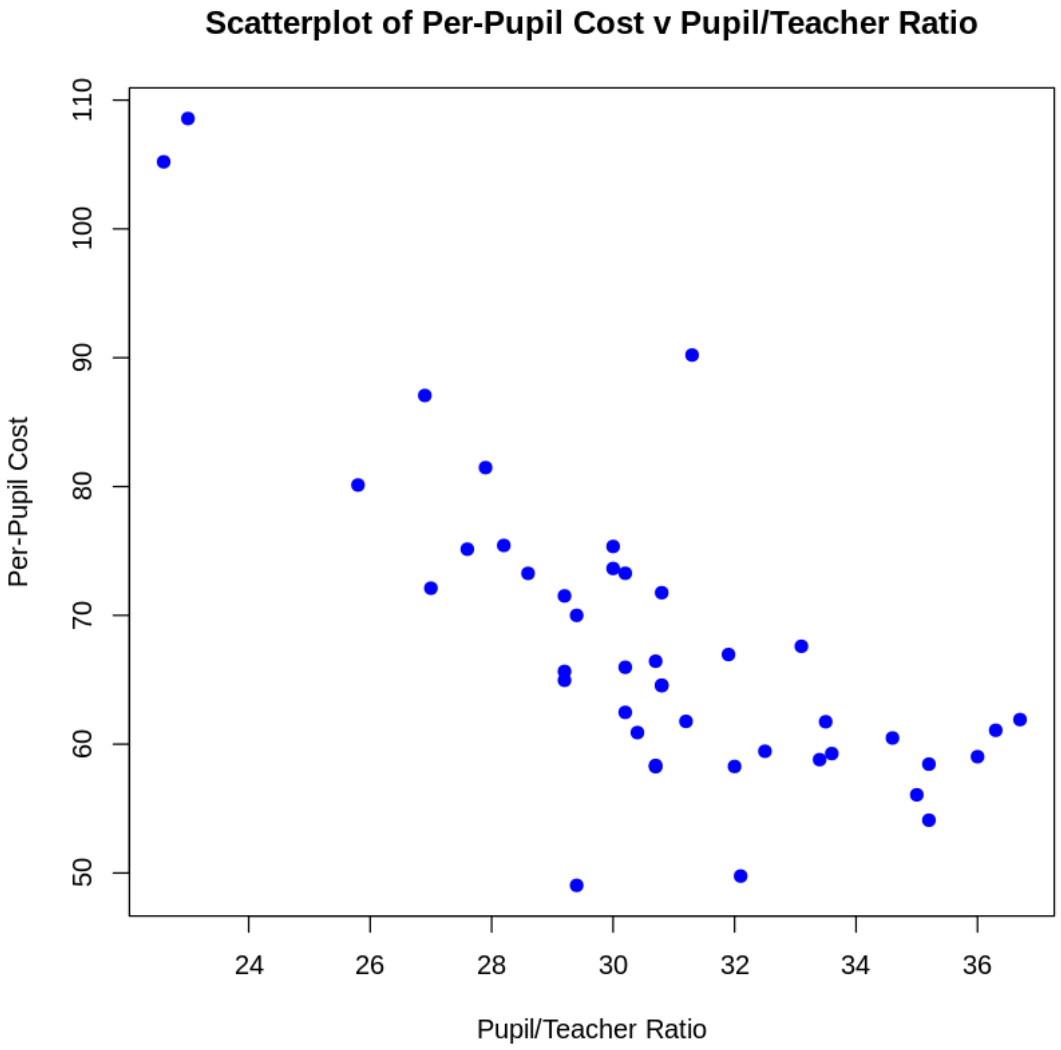
1. 43 2. 6

**2. (a) Create a model** Begin by creating two figures for your model. The first with `pup.tch.ratio` on the x-axis and `cost` on the y-axis. The second with `avg.salary` on the x-axis and `cost` on the y-axis. Does there appear to be a relation between these two predictors and the response.

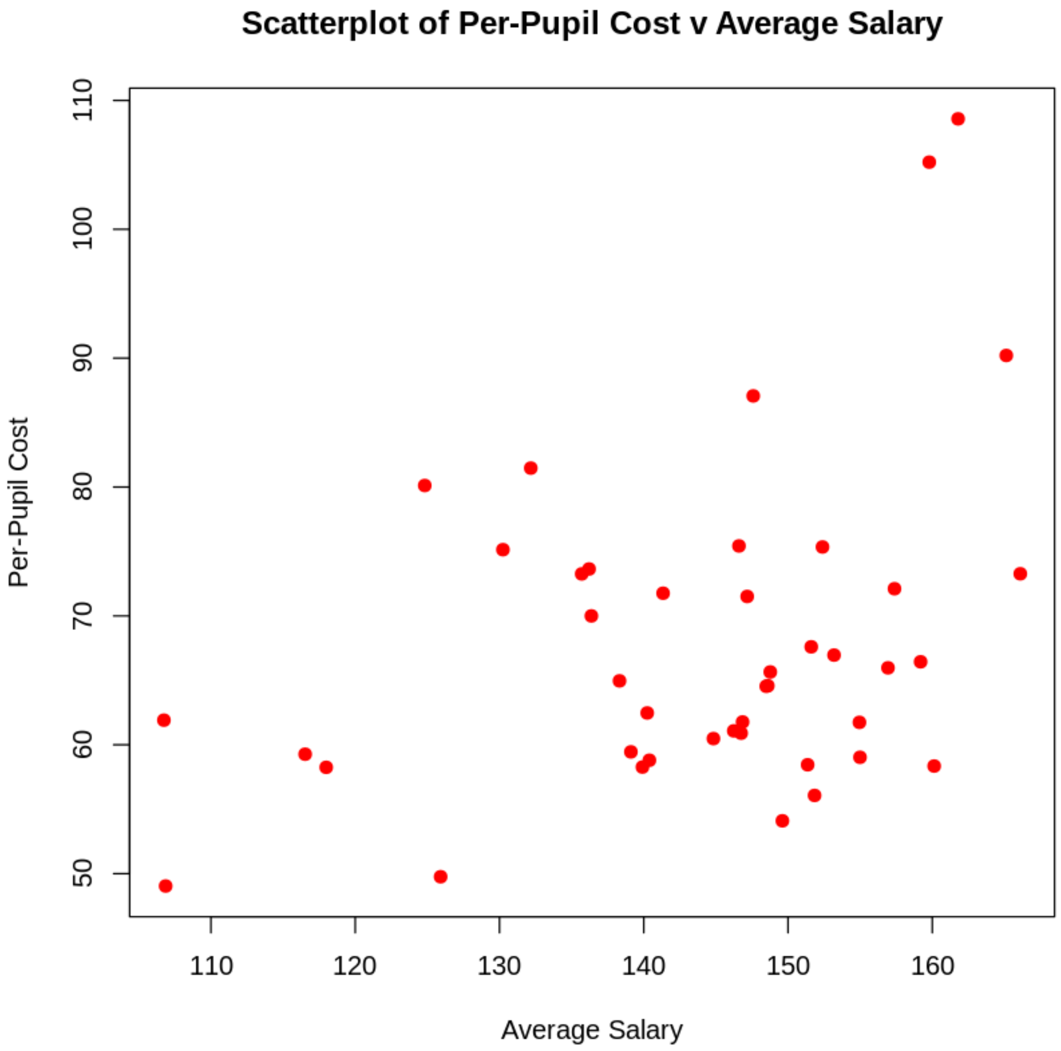
Then fit a multiple linear regression model with `cost` as the response and `pup.tch.ratio` and `avg.salary` as predictors.

[16]:

```
plot(school.data$pup.tch.ratio, school.data$cost,
    xlab = "Pupil/Teacher Ratio", ylab = "Per-Pupil Cost",
    main = "Scatterplot of Per-Pupil Cost v Pupil/Teacher Ratio",
    pch = 19, col = "blue")

plot(school.data$avg.salary, school.data$cost,
    xlab = "Average Salary", ylab = "Per-Pupil Cost",
    main = "Scatterplot of Per-Pupil Cost v Average Salary",
    pch = 19, col = "red")


lm_cost = lm(cost ~ pup.tch.ratio + avg.salary, data = school.data)

summary(lm_cost)

# Your Code Here
```



```
Call:
lm(formula = cost ~ pup.tch.ratio + avg.salary, data = school.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.8290  -5.2752  -0.8332   3.8253  19.6986 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   120.23756   17.73230   6.781 3.79e-08 ***
pup.tch.ratio  -2.82585    0.37714  -7.493 3.90e-09 ***
avg.salary      0.24061    0.08396   2.866   0.0066 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 7.721 on 40 degrees of freedom
Multiple R-squared:  0.6372,Adjusted R-squared:  0.6191 
F-statistic: 35.13 on 2 and 40 DF,  p-value: 1.559e-09 
```



**2. (b) RSS, ESS and TSS** In the code block below, manually calculate the RSS, ESS and TSS for your MLR model. Print the results.

[17]:
```
RSS = sum(resid(lm_cost)^2)
TSS = sum((fitted(lm_cost) - mean(school.data$cost))^2)
ESS = TSS - RSS

RSS
ESS
TSS

# Your Code Here
```

2384.59693812852

1803.97138885924

4188.56832698776

**2. (c) Are you Squared?** Using the values from **2.b**, calculate the $R^2$ value for your model. Check your results with those produced from the `summary()` statement of your model.

In words, describe what this value means for your model.

[30]:

```
R2 <- 1 - (RSS/TSS)
R2

summary(lm_cost)$r.squared

print("Manual R2 is less than the value using the summary function - the model␣
 ↪is not the best fit as it is a larger value.")

# Your Code Here
```

0.430689258961327

0.637222427559589

```
[1] "Manual R2 is less than the value using the summary function - the model is
not the best fit as it is a larger value."
```

**2. (d) Conclusions** Describe at least two advantages and two disadvantages of the $R^2$ value. print("Using R2 does not indicate causality and it is more possible to overfit the model using R2. The measure of fit provides a measure as to how well the independent variables explain the variability of the model. R2 is a good comparative measure for the different models to explain variance.")

## 2 Problem 3: Identifiability

**This problem might require some outside-of-class research if you haven't taken a linear algebra/matrix methods course.**

Matrices and vectors play an important role in linear regression. Let's review some matrix theory as it might relate to linear regression.

Consider the system of linear equations

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i, \qquad (1)$$

for $i = 1, ..., n$, where $n$ is the number of data points (measurements in the sample), and $j = 1, ..., p$, where

1. $p + 1$ is the number of parameters in the model.
2. $Y_i$ is the $i^{th}$ measurement of the *response variable.*
3. $x_{i,j}$ is the $i^{th}$ measurement of the $j^{th}$ *predictor variable.*
4. $\varepsilon_i$ is the $i^{th}$ *error term* and is a random variable, often assumed to be $N(0, \sigma^2)$.
5. $\beta_j$, $j = 0, ..., p$ are *unknown parameters* of the model. We hope to estimate these, which would help us characterize the relationship between the predictors and response.

**3. (a) MLR Matrix Form** Write the equation above in matrix vector form. Call the matrix including the predictors $X$, the vector of $Y_i$s $\mathbf{Y}$, the vector of parameters $\beta$, and the vector of error terms $\varepsilon$. (This is more LaTeX practice than anything else…)** Y = XB+E

Y = vector of responses X = matrix of predictors B = vector os parameters E = vector of error terms

**3. (b) Properties of this matrix** In lecture, we will find that the OLS estimator for $\beta$ in MLR is $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$. Use this knowledge to answer the following questions:

1. What condition must be true about the columns of $X$ for the "Gram" matrix $X^T X$ to be invertible?

2. What does this condition mean in practical terms, i.e., does $X$ contain a deficiency or redundancy?

3. Suppose that the number of measurements $(n)$ is less than the number of model parameters $(p + 1)$. What does this say about the invertibility of $X^T X$? What does this mean on a practical level?

4. What is true about about $\widehat{\boldsymbol{\beta}}$ if $X^T X$ is not invertible?

1. The columns of X must be linearly independent.
2. If X contains a redundancy then there is a linear relationship in the predictor variables - cannot estimate unique coefficients for each as they are not independent.
3. Matrix XTX is not invertible because it would become a singular matrix - there would be more parameters than data points. This would lead to overfitting the model and we could not get unique estimates for the parameters.
4. If XTX is not invertible then there is not a unique solution for Beta Hat.

## 2.1 Problem 4: Downloading...

The following data were collected to see if time of day madea difference on file download speed. A researcher placed a file on a remote server and then proceeded to download it at three different time periods of the day. They downloaded the file 48 times in all, 16 times at each Time of Day (`time`), and recorded the Time in seconds (`speed`) that the download took.

**4. (a) Initial Observations** The `downloading` data is loaded in and cleaned for you. Using `ggplot`, create a boxplot of `speed` vs. `time`. Make some basic observations about the three categories.

[37]:

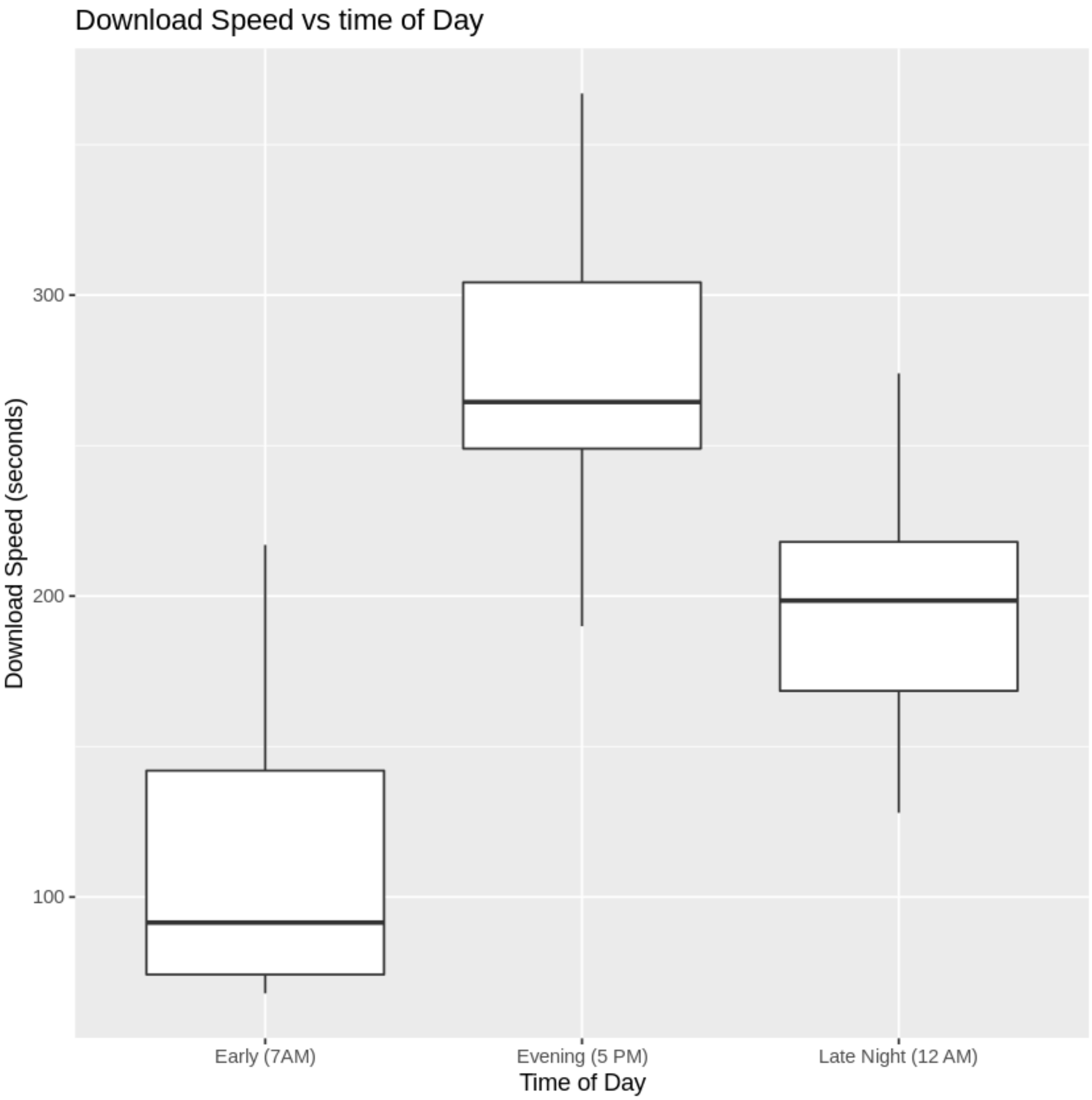
```
# Load in the data and format it
downloading = read.csv("downloading.txt", sep="\t")
names(downloading) = c("time", "speed")
# Change the types of brand and form to categories, instead of real numbers
downloading$time = as.factor(downloading$time)
summary(downloading)
```

```
                 time         speed      
 Early (7AM)       :16   Min.   : 68.0  
 Evening (5 PM)    :16   1st Qu.:129.8  
 Late Night (12 AM):16   Median :198.0  
                         Mean   :193.2  
                         3rd Qu.:253.0  
                         Max.   :367.0  
```

[40]:

```
downloading.model <- lm(speed ~ time, data = downloading)

ggplot(downloading, aes(x = time, y = speed)) +
geom_boxplot() +
labs(title = "Download Speed vs time of Day", x = "Time of Day", y = "Download␣
 ↪Speed (seconds)")
summary(downloading.model)
```

```
Call:
lm(formula = speed ~ time, data = downloading)

Residuals:
    Min      1Q  Median      3Q     Max 
-83.312 -34.328  -5.187  26.250 103.625 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)              113.37      11.79   9.619 1.73e-12 ***
timeEvening (5 PM)       159.94      16.67   9.595 1.87e-12 ***
timeLate Night (12 AM)    79.69      16.67   4.781 1.90e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.15 on 45 degrees of freedom
Multiple R-squared:  0.6717,Adjusted R-squared:  0.6571 
F-statistic: 46.03 on 2 and 45 DF,  p-value: 1.306e-11
```



**4. (b) How would we model this?** Fit a regression to these data that uses `speed` as the response and `time` as the predictor. Print the summary. Notice that the result is actually *multiple* linear regression, not simple linear regression. The model being used here is:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$$

where

1. $X_{i,1} = 1$ if the $i^{th}$ download is made in the evening (5 pm).
2. $X_{i,2} = 1$ if the $i^{th}$ download is made at night (12 am).

Note: If $X_{i,1} = 0$ and $X_{i,2} = 0$, then the $i^{th}$ download is made in the morning (7am).

**To confirm this is the model being used, write out the explicit equation for your model - using the parameter estimates from part (a) - and print out it's design matrix.**

[47]:
```
downloading.model <- lm(speed ~ time, data = downloading)
summary(downloading.model)
X = model.matrix(downloading.model)
head(X)




# Your Code Here
```

```
Call:
lm(formula = speed ~ time, data = downloading)

Residuals:
    Min      1Q  Median      3Q     Max 
-83.312 -34.328  -5.187  26.250 103.625 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)              113.37      11.79   9.619 1.73e-12 ***
timeEvening (5 PM)       159.94      16.67   9.595 1.87e-12 ***
timeLate Night (12 AM)    79.69      16.67   4.781 1.90e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.15 on 45 degrees of freedom
Multiple R-squared:  0.6717,Adjusted R-squared:  0.6571 
F-statistic: 46.03 on 2 and 45 DF,  p-value: 1.306e-11
```

A matrix: 6 × 3 of type dbl

| | (Intercept) | timeEvening (5 PM) | timeLate Night (12 AM) |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 |
| 3 | 1 | 0 | 0 |
| 4 | 1 | 0 | 0 |
| 5 | 1 | 0 | 0 |
| 6 | 1 | 0 | 0 |

**4. (c) Only two predictors?** We have three categories, but only two predictors. Why is this the case? To address this question, let's consider the following model:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_2 X_{i,3} + \varepsilon_i$$

where

1. $X_{i,1} = 1$ if the $i^{th}$ download is made in the evening (5 pm).
2. $X_{i,2} = 1$ if the $i^{th}$ download is made at night (12 am).
3. $X_{i,3} = 1$ if the $i^{th}$ download is made in the morning (7 am).

**Construct a design matrix to fit this model to the response, `speed`. Determine if something is wrong with it. Hint: Analyze the design matrix.**

[50]:

```
downloading.model
model.matrix(downloading.model)
```

```
Call:
lm(formula = speed ~ time, data = downloading)

Coefficients:
           (Intercept)      timeEvening (5 PM)  timeLate Night (12 AM)  
                113.37                  159.94                   79.69  
```

A matrix: 48 × 3 of type dbl

| | (Intercept) | timeEvening (5 PM) | timeLate Night (12 AM) |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 |
| 3 | 1 | 0 | 0 |
| 4 | 1 | 0 | 0 |
| 5 | 1 | 0 | 0 |
| 6 | 1 | 0 | 0 |
| 7 | 1 | 0 | 0 |
| 8 | 1 | 0 | 0 |
| 9 | 1 | 0 | 0 |
| 10 | 1 | 0 | 0 |
| 11 | 1 | 0 | 0 |
| 12 | 1 | 0 | 0 |
| 13 | 1 | 0 | 0 |
| 14 | 1 | 0 | 0 |
| 15 | 1 | 0 | 0 |
| 16 | 1 | 0 | 0 |
| 17 | 1 | 1 | 0 |
| 18 | 1 | 1 | 0 |
| 19 | 1 | 1 | 0 |
| 20 | 1 | 1 | 0 |
| 21 | 1 | 1 | 0 |
| 22 | 1 | 1 | 0 |
| 23 | 1 | 1 | 0 |
| 24 | 1 | 1 | 0 |
| 25 | 1 | 1 | 0 |
| 26 | 1 | 1 | 0 |
| 27 | 1 | 1 | 0 |
| 28 | 1 | 1 | 0 |
| 29 | 1 | 1 | 0 |
| 30 | 1 | 1 | 0 |
| 31 | 1 | 1 | 0 |
| 32 | 1 | 1 | 0 |
| 33 | 1 | 0 | 1 |
| 34 | 1 | 0 | 1 |
| 35 | 1 | 0 | 1 |
| 36 | 1 | 0 | 1 |
| 37 | 1 | 0 | 1 |
| 38 | 1 | 0 | 1 |
| 39 | 1 | 0 | 1 |
| 40 | 1 | 0 | 1 |
| 41 | 1 | 0 | 1 |
| 42 | 1 | 0 | 1 |
| 43 | 1 | 0 | 1 |
| 44 | 1 | 0 | 1 |
| 45 | 1 | 0 | 1 |
| 46 | 1 | 0 | 1 |
| 47 | 1 | 0 | 1 |
| 48 | 1 | 0 | 1 |

[ ]:

```

```

**4. (d) Interpretation** Interpret the coefficients in the model from **4.b**. In particular:

1. What is the difference between the mean download speed at 7am and the mean download speed at 5pm?
2. What is the mean download speed (in seconds) in the morning?
3. What is the mean download speed (in seconds) in the evening?
4. What is the mean download speed (in seconds) at night?

[59]:

```
downloading.model



print("
1 <- 1 = 159.94
2 <- 0 = 113.37
3 <- 0 + 1 = 113.37 + 159.94 = 273.3125
4 <- 0 + 2 = 113.37 + 79.69 = 193.0625
")
```

```
Call:
lm(formula = speed ~ time, data = downloading)

Coefficients:
           (Intercept)      timeEvening (5 PM)  timeLate Night (12 AM)
                113.37                  159.94                   79.69



[1] "\n1 <- 1 = 159.94\n2 <- 0 = 113.37\n3 <- 0 + 1 = 113.37 + 159.94 =
273.3125\n4 <- 0 + 2 = 113.37 + 79.69 = 193.0625\n"
```