Assignment 2: Parameter Estimation

Nicole McCarthy

9-11-25

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 β_0 and β_1.

1. (a) Taking Derivatives Find the partial derivative of $f(\mathbf{x}; \beta_0, \beta_1)$ with respect to β_0, and the partial derivative of $f(\mathbf{x}; \beta_0, \beta_1)$ with respect to β_1.

Solutions

$$\frac{\partial f}{\partial \beta_0} = -2 \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]$$

$$\frac{\partial f}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i [Y_i - (\beta_0 + \beta_1 x_i)]$$

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 β_0 and β_1. In particular, show

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

Solutions

Solve for $\widehat{\beta}_0$.

$$0 = -2 \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]$$

$$0 = \sum_{i=1}^{n} Y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i$$

$$n\beta_0 = \sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} x_i$$

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^n Y_i}{n} - \frac{\beta_1 \sum_{i=1}^n x_i}{n}$$
$$\widehat{\beta}_0 = \bar{Y} - \beta_1 \bar{x}$$

Solve for $\widehat{\beta}_1$.

$$0 = -2 \sum_{i=1}^n x_i [Y_i - (\beta_0 + \beta_1 x_i)]$$
$$0 = \sum_{i=1}^n (Y_i x_i - \beta_0 x_i - \beta_1 x_i^2)$$

Substitute $\bar{Y} - \beta_1 \bar{x}$ for $\widehat{\beta}_0$. (See above.)

$$0 = \sum_{i=1}^n [Y_i x_i - (\bar{Y} - \beta_1 \bar{x}) x_i - \beta_1 x_i^2]$$
$$0 = \sum_{i=1}^n (Y_i x_i - \bar{Y} x_i + \beta_1 \bar{x} x_i - \beta_1 x_i^2)$$
$$0 = \sum_{i=1}^n x_i (Y_i - \bar{Y}) + \beta_1 \sum_{i=1}^n x_i (\bar{x} - x_i)$$
$$-\beta_1 \sum_{i=1}^n x_i (\bar{x} - x_i) = \sum_{i=1}^n x_i (Y_i - \bar{Y})$$
$$\beta_1 \sum_{i=1}^n x_i (x_i - \bar{x}) = \sum_{i=1}^n x_i (Y_i - \bar{Y})$$

Let's take a closer look at $\sum_{i=1}^n x_i (x_i - \bar{x})$.

$$\sum_{i=1}^n x_i (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x} + \bar{x})(x_i - \bar{x})$$

$\bar{x}$ is added and subtracted (no overall change) to first term in sum. Now insert parentheses to manipulate the sum.

$$\sum_{i=1}^n (x_i - \bar{x} + \bar{x})(x_i - \bar{x}) = \sum_{i=1}^n ([x_i - \bar{x}] + \bar{x})(x_i - \bar{x})$$
$$\sum_{i=1}^n ([x_i - \bar{x}] + \bar{x})(x_i - \bar{x}) = \sum_{i=1}^n [(x_i - \bar{x})^2 + \bar{x}(x_i - \bar{x})]$$

Note:

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$$
$$\sum_{i=1}^n [(x_i - \bar{x})^2 + \bar{x}(x_i - \bar{x})] = \sum_{i=1}^n (x_i - \bar{x})^2$$

So far, our findings from this aside assert:

$$\sum_{i=1}^n x_i (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})^2$$

This is because $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, as mentioned earlier. For simplicity's sake, we can loosely assume:

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n}(x_i - \bar{x})$$

This is not strictly true, but substituting one for the other in the case of this problem is useful and ends up being true. Now back to our original problem. This is where we left off:

$$\beta_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i(Y_i - \bar{Y})$$

We can use our conclusion from the aside to assert this:

$$\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$$

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

Problem 2: Oh My Goodness of Fit!

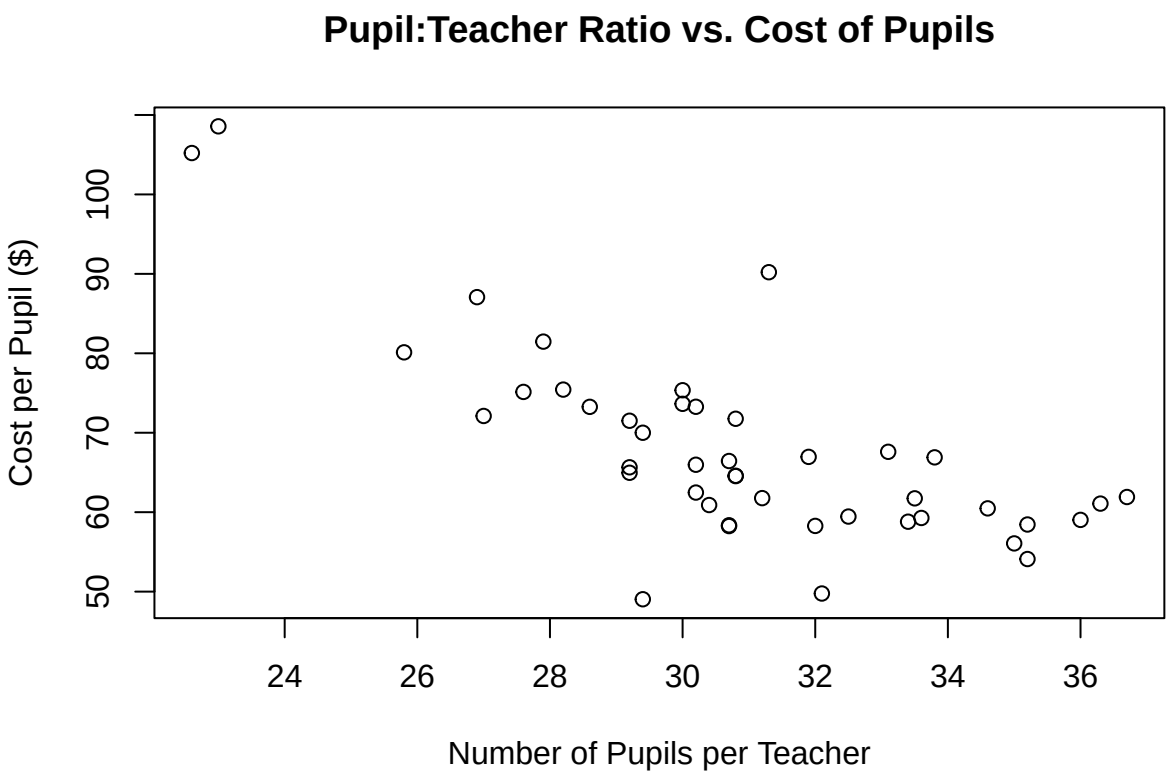
In the US, public schools have been slowly increasing class sizes over the last 15 years [https://stats.oecd.org/Index.aspx?DataSetCode=EDU_CLASS]. The general cause for this is because it saves money to have more kids per teacher. But how much money does it save? Let's use some of our new regression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

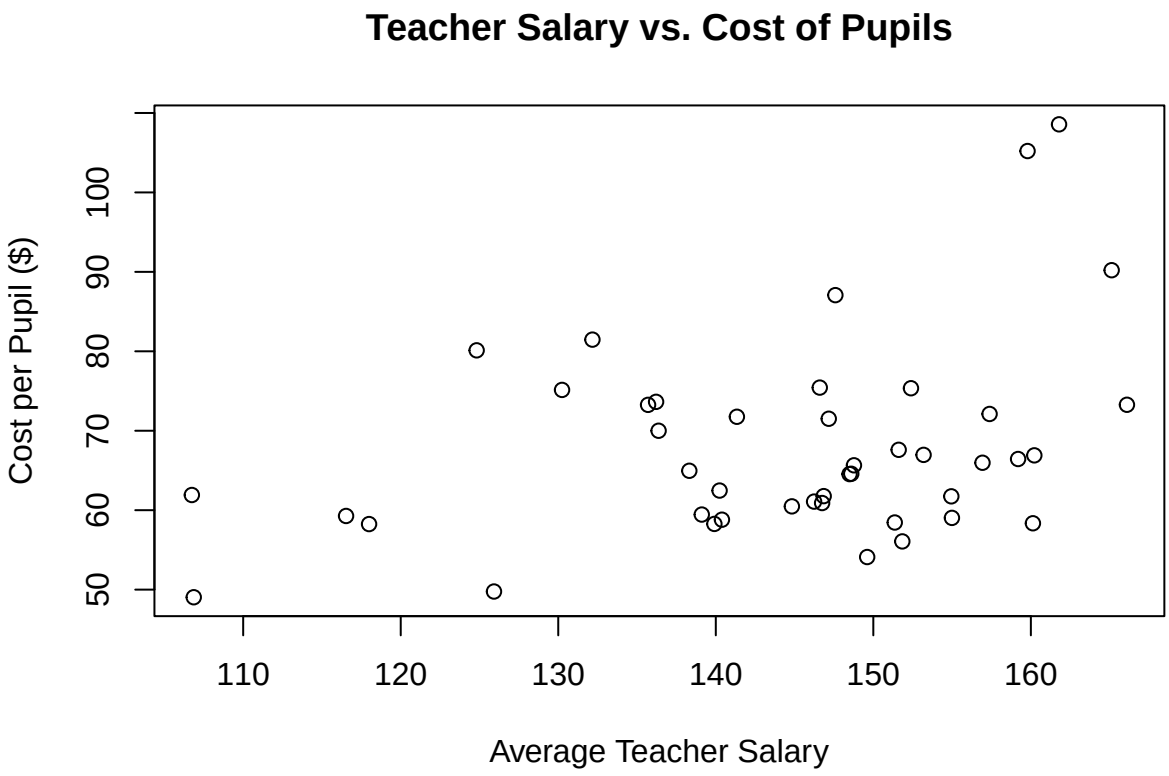
Variables/Columns:
School
Per-Pupil Cost (Dollars)
Average daily Attendance
Average Monthly Teacher Salary (Dollars)
Percent Attendance
Pupil/Teacher ratio

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

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.


Pupil:Teacher Ratio vs. Cost of Pupils
Cost per Pupil ($)
50
60
70
80
90
100
24
26
28
30
32
34
36
Number of Pupils per Teacher



There appears to be a slight negative correlation between the cost per pupil and the number of pupils per teacher, meaning that the more students a teacher has in their classroom, the lower the cost of educating each student. Additionally, there is a small positive correlation between the cost per pupil and the average teacher salary. This means that the more teachers are paid, the more it costs to educate each student. The spread of the data on the second plot is worth considering as this is not strictly true for all schools.

```
model = lm(cost ~ pup.tch.ratio + avg.salary, data = school.data)
model.summary <- summary(model)
model.summary
```

```
##
## Call:
## lm(formula = cost ~ pup.tch.ratio + avg.salary, data = school.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8538  -5.3484  -0.6884   3.5671  19.7207
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   118.44679   17.11396   6.921 2.13e-08 ***
## pup.tch.ratio  -2.79829    0.36853  -7.593 2.43e-09 ***
## avg.salary      0.24770    0.08168   3.033  0.00419 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.646 on 41 degrees of freedom
```

```
## Multiple R-squared:  0.6354, Adjusted R-squared:  0.6176
## F-statistic: 35.73 on 2 and 41 DF,  p-value: 1.039e-09
```

The linear model that explains cost per pupil with average teacher salary and the pupil:teacher ratio explains approximately 63% of the spread of the data, with each variable being statistically significant.

2. (b) RSS, ESS and TSS Manually calculate the RSS, ESS and TSS for your MLR model. Print the results.

```
# Grab just variables needed for model predictions
necessary.data <- data.frame(school.data$avg.salary, school.data$pup.tch.ratio)
colnames(necessary.data) <- c("avg.salary", "pup.tch.ratio")

# Get predictions (y_hat)
predictions <- predict(model, newdata = necessary.data)

# Calculate Residual Sum of Squares
residuals.sq <- (school.data$cost - predictions)**2
rss <- sum(residuals.sq)
print(paste("RSS:", round(rss, 2)))

# Calculate Explained Sum of Squares
ybar <- mean(school.data$cost)
explained.sq <- (predictions - ybar)**2
ess <- sum(explained.sq)
print(paste("ESS:", round(ess, 2)))

# Calculate Total Sum of Squares
total.sq <- (school.data$cost - ybar)**2
tss <- sum(total.sq)
print(paste("TSS:", round(tss, 2)))
```

```
## [1] "RSS: 2396.74"
## [1] "ESS: 4177.42"
## [1] "TSS: 6574.16"
```

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.

```
r2 <- 1 - (rss/tss)
Rsq <- model.summary$r.squared

print("R-squared Comparison")
print(paste("My Calculation:", round(r2, 4)))
print(paste("R Calculation:", round(Rsq, 4)))
```

```
## [1] "R-squared Comparison"
## [1] "My Calculation: 0.6354"
## [1] "R Calculation: 0.6354"
```

My calculation and R's calculation are exactly the same, which means my manual results were correct. This shows that about **63% of the variability in the cost per pupil is explained by my linear model** that uses average teacher salary and pupil:teacher ratio as predictors.

2. (d) Conclusions Describe at least two advantages and two disadvantages of the R^2 value.

Advantages

1. R^2 *gives a relative understanding of how the predictors affect the response.* We know that about 63% of the variability in the cost per student can be explained by the two predictor variables used in the model. This also tells us that about 37% of the variability in cost per student is explained by other factors and it might be worthwhile to look into other predictors.
2. R^2 *minimizes the amount of unexplained variance by the model.* This means it is the *best estimator* for understanding the variation explained by the model.

Disadvantages

1. *The R^2 statistic cannot be used to compare different linear models with different numbers of predictors.* This is because the R^2 value of a model will always increase with more predictors, even if those predictors are not statistically significant. That means it is possible to get a R^2 value close to 1 when the model is not useful because it is over-fitted to the training data.
2. *It is possible to get a R^2 value close to 0, even when the model is well fitted because there is a lot of variability in the data.* Therefore, it may not be possible to explain more the 63% of the variability in cost per student because there is no good explanation for it, or the reasons behind the rest of the variability are not measurable.

Problem 3: Identifiability

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

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

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. ε_i is the i^{th} *error term* and is a random variable, often assumed to be $N(0, \sigma^2)$.
5. β_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_is $\mathbf{Y}$, the vector of parameters β, and the vector of error terms ε.

$$Y_{n\times 1} = X_{n\times(p+1)}\beta_{(p+1)\times 1} + E_{n\times 1}$$

Y is the (n x 1) vector of responses. X is the (n x $[p + 1]$) matrix of observations for each predictor, where the first column is 1's. β is the ($[p + 1]$ x 1) vector of parameters. E is the (n x 1) matrix of random errors for each response.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ 1 & x_{2,1} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

3. (b) Properties of this matrix The OLS estimator for β 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?

   There must be less columns than rows in the matrix X ($n >= (p+1)$). Furthermore, the columns must be independent.

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

   If $X^T X$ IS invertible, X does NOT contain any redundant columns or columns that are linear combinations of other columns.

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?

   If $n < (p+1)$, then $X^T X$ is NOT invertible. On a practical level, this tells us that there are infinitely many solutions and therefore the model is non-identifiable.

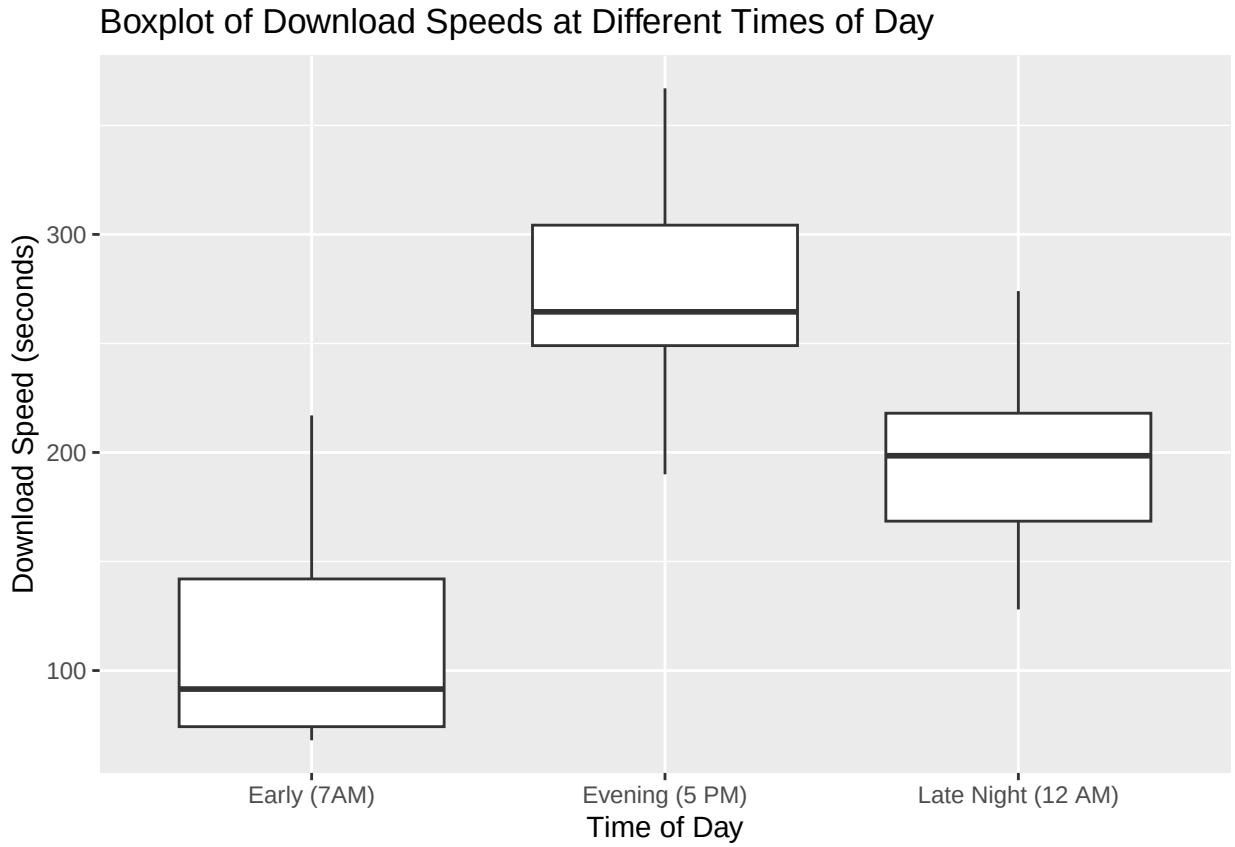
4. What is true about about $\widehat{\boldsymbol{\beta}}$ if $X^T X$ is not invertible?

   If $X^T X$ is not invertible, then $\widehat{\boldsymbol{\beta}}$ has infinitely many solutions and therefore cannot be finitely determined.

Problem 4: Downloading...

The following data were collected to see if time of day made a 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.



The fastest download speeds occur in the early morning (7AM) and the slowest occur in the evening (5PM). There seems to be a minimum download time of just under 50 seconds, but no maximum is indicated. The late night (12AM) average speed is about 200 seconds, which is equal to the upper end of the early (7AM) downloads and the lower end of the evening (5PM) downloads.

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.

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

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

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.312 -34.328  -5.188  26.250 103.625
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               113.38      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
```

```
design.matrix <- model.matrix(download.model)
design.matrix
```

```
##    (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
## attr(,"assign")
## [1] 0 1 1
## attr(,"contrasts")
## attr(,"contrasts")$time
## [1] "contr.treatment"
```

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.

There are only two predictors because the lack of either indicates the presence of the third predictor, which is not explicitly stated in the model's formula.

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?

```
morning.avg <- mean(downloading$speed[downloading$time == "Early (7AM)"])
evening.avg <- mean(downloading$speed[downloading$time == "Evening (5 PM)"])
morning.avg - evening.avg
```

```
## [1] -159.9375
```

2. What is the mean download speed (in seconds) in the morning?

```
## [1] 113.375
```

3. What is the mean download speed (in seconds) in the evening?

```
## [1] 273.3125
```

4. What is the mean download speed (in seconds) at night?

```r
mean(downloading$speed[downloading$time == "Late Night (12 AM)"])
```

```
## [1] 193.0625
```