Assignment 3-4

JARED YU

1. In a typical multiple linear regression model where x_1 and x_2 are two regressors. The expected value of the response variable y given x_1 and x_2 is denoted by $E(y|x_1, x_2)$.
   a. As the value of x_1 increases, the magnitude of change in the value of $E(y|x_1, x_2)$ will not depend on the value of x_2. Write down the multiple linear regression model with assumptions for this scenario.

Ans:

In the case when a change in x_1 does not depend on a change in x_2, then the model can appear something like as follows,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Then the resulting $E(y|x_1, x_2)$ would appear as follows,

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

The following are some assumptions:

The data is such that there are n observations and $k = 2$ for β_1 and β_2. It requires then that $n > k$. Furthermore, y_i is the observed response while x_{i1} and x_{i2} are the regressors. An assumption is that $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$, and that the errors are uncorrelated. In deriving the OLS estimators, it must be such that $(\mathbf{X}'\mathbf{X})^{-1}$ is calculatable, or in other words the inverse exists.

   b. As the value of x_1 increases, the magnitude of change in the value of $E(y|x_1, x_2)$ will depend on the value of x_2. Write down the multiple linear regression model with assumptions for this scenario.

Ans:

In the case when a change in x_1 does depend on a change in x_2, then the model can appear something like as follows,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon.$$

Then the resulting $E(y|x_1, x_2)$ would appear as follows,

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

The following are some assumptions:

The data is such that there are n observations and $k = 3$ for β_1, β_2, and β_3. It requires then that $n > k$. Furthermore, y_i is the observed response while x_{i1} and x_{i2} are the regressors. An assumption is that $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$, and that the errors are uncorrelated. In deriving the OLS estimators, it must be such that $(\mathbf{X}'\mathbf{X})^{-1}$ is calculatable, or in other words the inverse exists. Here it is interesting since β_3 is the coefficient for the interaction term $x_1 x_2$.

2. Do Problem 3.24, page 126 of Textbook

Show that an alternate computing formula for the regression sum of squares in a linear regression model is

$$SS_R = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2$$

Ans:

The “regular” way to express SS_R is as follows,

$$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n}.$$

First, begin with the simpler terms on the right-hand side, $n\bar{y}^2$ and $\frac{(\sum_{i=1}^{n} y_i)^2}{n}$. It will be shown that these two are equivalent.

$$n\bar{y}^2 = n\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2 = \frac{n(\sum_{i=1}^{n} y_i)^2}{n^2} = \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

Next, it will be shown that $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$ and $\sum_{i=1}^{n} \hat{y}_i^2$ are equivalent.

$$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}]'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{X}[(\mathbf{X}'\mathbf{X})']^{-1}\mathbf{X}'\mathbf{y}$$

$$= \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{H}\mathbf{y} = \mathbf{y}'\mathbf{H}\mathbf{H}\mathbf{y} = \mathbf{y}'\mathbf{H}'\mathbf{H}\mathbf{y} = (\mathbf{H}\mathbf{y})'\mathbf{H}\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} = \sum_{i=1}^{n} \hat{y}_i^2$$

Therefore, it follows that since $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = \sum_{i=1}^{n} \hat{y}_i^2$ and $n\bar{y}^2 = \frac{(\sum_{i=1}^{n} y_i)^2}{n}$, then

$$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2 = SS_R. \blacksquare$$

3. Do Problem 3.27, page 127 of Textbook

Show that $Var(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$.

Ans:

$$Var(\hat{\mathbf{y}}) = Var\left(\mathbf{X}\hat{\boldsymbol{\beta}}\right) = Var[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}]$$
$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'Var(\mathbf{y})[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}']'$$
$$= \sigma^2\ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}']'$$
$$= \sigma^2\ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}[(\mathbf{X}'X)^{-1}]'\mathbf{X}'$$

Note that $[(\mathbf{X}'X)^{-1}]' = [(\mathbf{X}'X)']^{-1} = (\mathbf{X}'\mathbf{X})^{-1}$.

$$\cdots = \sigma^2\ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
$$= \sigma^2\mathbf{X}\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \sigma^2\mathbf{H}\ \blacksquare$$

4. Do Problem 3.28, page 128 of Textbook

Prove that the matrices $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are idempotent, that is $\mathbf{HH} = \mathbf{H}$ and $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I} - \mathbf{H}$.

Ans:

$$\mathbf{HH} = \ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
$$= \ \mathbf{X}\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}\blacksquare$$

$$(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I}(\mathbf{I} - \mathbf{H}) - \mathbf{H}(\mathbf{I} - \mathbf{H})$$
$$= \mathbf{II} - \mathbf{IH} - \mathbf{HI} + \mathbf{HH}$$
$$= \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{HH}$$
$$= \mathbf{I} - 2\mathbf{H} + \mathbf{HH}$$
$$= \mathbf{I} - 2\mathbf{H} + \mathbf{H}$$

Note that the previous proof is used, $\mathbf{HH} = \mathbf{H}$.

$$= \mathbf{I} - \mathbf{H}\ \blacksquare$$

5. In a multiple linear regression model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where y is a response variable, x_1 and x_2 are non-random regressors, and ε is a random error, the parameter β_2 is **nonzero**. Suppose that n subjects give data on (y_1, x_1, x_2) to generate the ordinary least-squares (OLS) estimators of all three β parameters in this model. We then fit the same data to the simple linear regression model $y = \beta_0 + \beta_1 x_1 + \varepsilon$.

a. Create a hypothetical data set and perform regression analysis to compare the OLS estimate of β_1 in the regression model including x_2 with the OLS estimate of β_1 in the regression model excluding x_2. What have you learned?

Ans:
In the simulation, a total of $n = 1{,}000$ samples observations were generated based on the following model,

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

where $\beta_1 = 0.2$, $\beta_2 = 3$, and $\varepsilon \overset{\text{i.i.d.}}{\sim} N\left(\mu = 0, \sigma^2 = \sqrt{0.1}\right)$. The x_1 and x_2 are examples randomly generated from a $N(\mu = 0, \sigma^2 = 1)$ and $N\left(\mu = 3, \sigma^2 = \sqrt{0.2}\right)$ respectively. Let the model including x_2 be called the "full model," while the model excluding x_2 be called the "reduced model."

Full Model	Reduced Model
$\hat{\beta}_0 = -0.0314017$	$\hat{\beta}_0 = 8.9918760$
$\hat{\beta}_1 = 0.2049049$	$\hat{\beta}_1 = 0.2087786$
$\hat{\beta}_2 = 3.0110085$	
$SS_{Res} = 10.59006$	$SS_{Res} = 402.4067$

Table 1 The above table shows the resulting calculations from the full model (left) and the reduced model (right). It includes the different values for $\hat{\beta}$ along with SS_{Res}.

Above in Table 1, there is a side-by-side comparison of some of the coefficients and statistics from calculating the OLS estimators for the full and reduced models. It is interesting to note that both $\hat{\beta}_1's$ are quite similar in size. However, something else to take into account is the fact that x_1 itself comes from a $N(0,1)$ distribution and therefore it makes sense that the corresponding coefficient has a similarly small magnitude.

A fairly large difference can be seen in the SS_{Res}, it shows that for the full model it is 10.59006 while for the reduced model it is 402.4067. This indicates that the full model has a much better fit in comparison to the reduced model. This is logical given that the full model is a more complex model and can therefore potentially capture more information from the sample data.

Looking at the $\hat{\beta}_0$, it is apparent that the reduced model has a much larger value for this coefficient. It seems then that a possibility is that the intercept in the reduced model is trying to compensate for the lack of a $\hat{\beta}_2$ in the linear model.

A takeaway from this then is that although the SLR model is perhaps more general, it loses the ability to capture more information in comparison to an MLR model. It can be seen in the coefficients themselves, how the intercept in the SLR will try to compensate for the "missing" coefficient. This could mean then that it is in general better to choose a type of MLR model over SLR, despite the idea of generalizability regarding the SLR model. Furthermore, the reduced error seems quite large in this example and so it could be argued therefore that the increased complexity and reduced generalizability in the MLR is worthwhile. The most interesting note however is that despite the MLR model having an additional coefficient, the $\hat{\beta}_1$ coefficient in both models remained fairly similar. It is possible that with a different x_1 that has a greater variance, the same results would still be seen.

b. Discuss with mathematical arguments whether the OLS estimators of β_1 from the two model fittings are equal. If not, discuss with mathematical arguments the condition(s) under which the two OLS estimators of β_1 are equal (Note: β_1 is **nonzero**).

Ans:

6. Use any math/stat software (e.g., www.numbergenerator.org/randomnumbergenerator) of your choice to find a random number generator to randomly select 22 rows of Table B.3 (page 556) used in Problem 3.5 (page 122) of Textbook and then do (a), (b), (c), (d), (e), (f), (g).

The numbers chosen are: 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 14, 15, 16, 18, 19, 20, 21, 27, 29, 30, 32. Two numbers were manually switched (3 and 6) to avoid using rows 23 and 25 due to the existence of 'NA' values.

Consider the gasoline mileage data in Table B.3.

a. Fit a multiple linear regression model relating gasoline mileage y (miles per gallon) to engine displacement x_1 and the number of carburetor barrels x_6.

Ans:

The following model was created,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_6 + \varepsilon.$$

The resulting OLS estimators are as follows, $\hat{\beta}_0 = 33.37847405$, $\hat{\beta}_1 = -0.06122111$, and $\hat{\beta}_2 = 1.45319169$. They were calculated by using the least-squares normal equations, where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

b. Construct the analysis-of-variance table and test for significance of regression.

Ans:

The goal is to test for the significance of regression of the model created in part (a). The following are calculated and can be seen below in Table 2: SS_R, SS_{Res}, SS_T, and F_0. These are based on equations that can be seen on pp. 86-87 in the Textbook. Let the hypothesis be as follows,

$$H_0\colon \beta_1 = \beta_2 = 0 \text{ vs.} H_1\colon \text{at least one } \beta_j \neq 0 \text{ for } j = 1,2.$$

After calculating the analysis-of-variance table (Table 2), the resulting critical value is $F_0 = 46.54955$. The critical value for $F_{2,19}$ at $\alpha = 0.01$ is 5.925879. The F-statistic is much larger than the critical value and so we reject the null hypothesis at the $\alpha = 0.01$ confidence level. Furthermore, the resulting p-value based on the F-statistic is 4.7527×10^{-8}. The conclusion then is that at least one of the β_j, for $j = 1, 2$ is nonzero.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R = 726.4536$	$k = 2$	$MS_R = 363.2268$	$\frac{MS_R}{MS_{Res}} = 46.54955$
Residual	$SS_{Res} = 148.2573$	$n - k - 1 = 19$	$MS_{Res} = 7.803015$	
Total	$SS_T = 874.7109$	$n - 1 = 21$		

Table 2 The above table shows the Test for Significant of Regression for the MLR model in part (a).

c. Calculate R^2 and R^2_{Adj} for this model. Compare this to the R^2 and the R^2_{Adj} for the simple linear regression model relating mileage to engine displacement in Problem 2.4.

Ans:

The formula for R^2 is $\frac{SS_R}{SS_T}$ and the formula for R^2_{Adj} is $1-\frac{SS_{Res}/(n-p)}{SS_T/(n-1)}$. These are used to calculate the values seen below in Table 3. The left column shows the results for the model from Problem 2.4 where $y=\beta_0+\beta_1 x_1$ and the right column shows the current model for Problem 3.5 where $y=\beta_0+\beta_1 x_1+\beta_2 x_6$. It can be seen that there is an improvement in both the R^2 and R^2_{Adj} values when the model includes the second regressor, x_6. It is apparent that the R^2 and R^2_{Adj} for both models are not particularly high. It would be logical to think however that by including more regressors from the entire data set, that the R^2 and R^2_{Adj} would start to get much closer to 1 (e.g. 0.95). At this point however, by using only two of the nine features in the data set, the increase from using only one is not too large.

	Problem 2.4	Problem 3.5
R^2	0.7911184	0.8305071
R^2_{Adj}	0.7806743	0.8126658

Table 3 The table above shows the R^2 and R^2_{Adj} values for the models from Problem 2.4 (left) and Problem 3.5 (right).

d. Find a 95% CI for β_1.

Ans:

The formula for the 95% CI for β_1 is as follows,

$$\hat{\beta}_1 - t_{0.025,19}\sqrt{\hat{\sigma}^2 C_{11}} \le \beta_1 \le \hat{\beta}_1 + t_{0.025,19}\sqrt{\hat{\sigma}^2 C_{11}}.$$

Here, $\hat{\sigma}^2 = MS_{Res}$ and C_{jj} is the $j'th$ diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Plugging in those values leads to the following:

$$\rightarrow -0.06122111 - (2.093024)\sqrt{(7.803015)(6.211432\times 10^{-6})} \le \beta_1$$
$$\le -0.06122111 + (2.093024)\sqrt{(7.803015)(6.211432\times 10^{-6})}$$
$$\rightarrow \boxed{-0.07579251 \le \beta_1 \le -0.04664971}.$$

e. Compute the t statistics for testing $H_0\colon \beta_1 = 0$ and $H_0\colon \beta_6 = 0$. What conclusions can you draw?

Ans:

The first hypothesis is as follows,

$$H_0\colon \beta_1 = 0 \text{ vs.} H_1\colon \beta_1 \ne 0.$$

It is testing whether x_1 (displacement) can be deleted from the model. The test statistic for such as hypothesis is as follows,

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}.$$

Then for the first hypothesis, $t_0 = -8.793749$. Since $t_{0.025,19} = 2.093024$, we reject $H_0\colon \beta_1 = 0$ and conclude that the regressor x_1 contributes significantly to the model given that x_6 (carburetor) is also in the model at the $\alpha = 0.05$ confidence level.

In the second hypothesis,

$$H_0: \beta_6 = 0 \text{ vs.} H_1: \beta_6 \neq 0.$$

It is testing whether x_6 can be deleted from the model. Then for the second hypothesis, $t_0 = 2.101295$. Since $t_{0.025,19} = 2.093024$, we reject $H_0: \beta_6 = 0$ and conclude that the regressor x_6 contributes significantly to the model given that x_1 is also in the model at the $\alpha = 0.05$ confidence level.

It is interesting to see that the critical value along the t-distribution is quite close to the cut-off point for β_6. Although it passes the test at $\alpha = 0.05$, it would not pass at $\alpha = 0.01$. It seems then that in this model that includes only x_1 and x_6 that x_1 is more influential in the overall model.

f. Find a 95% CI on the mean gasoline mileage when $x_1 = 275$ in.^{3} and $x_6 = 2$ barrels.

Ans:

The formula for such a CI is as follows,

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \leq E(y|x_0) \leq \hat{y}_0 + t_{\frac{\alpha}{2},n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}.$$

Here, $\mathbf{x_0} = \begin{bmatrix} 1 \\ 275 \\ 2 \end{bmatrix}$. The fitted value at this point is as follows,

$$\hat{y}_0 = \boldsymbol{x}_0'\widehat{\boldsymbol{\beta}} = [1 \quad 275 \quad 2]\begin{bmatrix} 33.37847405 \\ -0.06122111 \\ 1.45319169 \end{bmatrix} = 19.44905.$$

The variance of $\hat{y}_0$ is estimated by,

$$\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = \cdots = 0.4737911.$$

Therefore, a 95% CI on the mean gasoline mileage at this point is,

$$\rightarrow 19.44905 - 2.860935\sqrt{0.4737911} \leq E(y|x_0) \leq 19.44905 + 2.860935\sqrt{0.4737911}$$

$$\rightarrow \boxed{17.47980 \leq E(y|x_0) \leq 21.41831}.$$

g. Find a 95% prediction interval for a new observation on gasoline mileage when $x_1 = 257$ in.^{3} and $x_6 = 2$ barrels.

Ans:

The formula for such a PI is as follows,

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-p}\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \leq y_0 \leq \hat{y}_0 + t_{\frac{\alpha}{2},n-p}\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}.$$

Here, $\mathbf{x_0} = \begin{bmatrix} 1 \\ 257 \\ 2 \end{bmatrix}$. The fitted value at this point is as follows,

$$\hat{y}_0 = \boldsymbol{x}_0'\widehat{\boldsymbol{\beta}} = [1 \quad 257 \quad 2]\begin{bmatrix} 33.37847405 \\ -0.06122111 \\ 1.45319169 \end{bmatrix} = 20.55103.$$

The variance of $\hat{y}_0$ is estimated by,

$$\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0) = \cdots = 8.240775.$$

Therefore, a 95% PI on the mean gasoline mileage at this point is,

$$\rightarrow 20.55103 - 2.860935\sqrt{0.4737911} \leq E(y|x_0) \leq 20.55103 + 2.860935\sqrt{0.4737911}$$

$$\rightarrow \boxed{14.54264 \leq E(y|x_0) \leq 26.55942}.$$

Supplemental Module 4 Lecture

In the regression fit, calculation of CI and prediction interval, what are the assumptions I made?

Ans:

The data is such that $n = 7$, $k = 2$, and $p = k + 1 = 3$. This matches the context of $n > k$. Furthermore, y_i is the observed response while z_{i1} and z_{i2} are the regressors. The sample regression model has the following expression,

$$y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \varepsilon_i.$$

An assumption is that $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$, and that the errors are uncorrelated. In deriving the OLS estimators, it must be such that $(\mathbf{Z}'\mathbf{Z})^{-1}$ is calculatable, or in other words the inverse exists.

Some more assumptions are that the error term, ε, follows a normal distribution, so y follows normal distribution also. Furthermore, $\hat{y}$ is a linear combination of y, so $\hat{y}$ also follows normal distribution, however, we don't know the true variance of $\hat{y}$, we use $\hat{\sigma}^2$ instead, so, both CI and prediction interval use a t-value from the Student's t distribution.

Code Appendix

```
library(MPV)
### 5
### a
set.seed(1)
n <- 1e3
x1 <- rnorm(n = n, mean = 0, sd = 1)
x2 <- rnorm(n = n, mean = 3, sd = 0.2)
error <- rnorm(n = n, mean = 0, sd = 0.1)
y <- 0.2 * x1 + 3 * x2 + error

beta_hat_calc <- function(X, y) {
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
  return(beta_hat)
}
H_calc <- function(X) {
  H <- X %*% solve(t(X) %*% X) %*% t(X)
  return(H)
}
y_hat_calc <- function(H, y) {
  y_hat <- H %*% y
  return(y_hat)
}
e_calc <- function(y, y_hat) {
  e <- y - y_hat
  return(e)
}

# full model
ones <- rep(1, n)
X_full <- cbind(ones, x1, x2)
beta_hat_full <- beta_hat_calc(X = X_full, y = y)
H_full <- H_calc(X = X_full)
y_hat_full <- y_hat_calc(H = H_full, y = y)
e_full <-  e_calc(y = y, y_hat = y_hat_full)
t(e_full) %*% e_full # 10.59006

# SLR model
ones <- rep(1, n)
X_red <- cbind(ones, x1)
beta_hat_red <- beta_hat_calc(X = X_red, y = y)
H_red <- H_calc(X = X_red)
y_hat_red <- y_hat_calc(H = H_red, y = y)
e_red <-  e_calc(y = y, y_hat = y_hat_red)
t(e_red) %*% e_red # 402.4067


### b

### 6
set.seed(1); n <- 22 # Sample rows
random_rows <- sort(sample(x = seq(1, 32), size = n, replace = FALSE))
na_rows <- c(23, 25)
random_rows <- random_rows[!random_rows %in% na_rows]
random_rows <- c(random_rows, 3, 6) # Ignore NA rows

# Create table subset
table_b3 <- MPV::table.b3; car_data <- table_b3[random_rows,]
car_data$index <- as.numeric(row.names(car_data))
car_data <- car_data[order(car_data$index),]
car_data$index <- NULL # Fix the data set

### a
# Subset data
y <- car_data[,1]; x1 <- car_data[,2]; x6 <- car_data[,7]; ones <- rep(1, n)
X <- cbind(ones, x1, x6)
beta_hat <- beta_hat_calc(X = X, y = y)
H <- H_calc(X = X)
```

```
y_hat <- y_hat_calc(H = H, y = y)
# e <- e_calc(y = y, y_hat = y_hat)
# t(e) %*% e # 148.2573

### b
SS_Res_calc <- function(y, beta_hat, X) {
  SS_Res <- (t(y) %*% y) - (t(beta_hat) %*% t(X) %*% y)
  return(SS_Res)
}
SS_T_calc <- function(y) {
  n <- length(y)
  SS_T <- (t(y) %*% y) - ((sum(y)^2) / n)
  return(SS_T)
}
SS_R_calc <- function(beta_hat, X, y) {
  n <- length(y)
  SS_R <- (t(beta_hat) %*% t(X) %*% y) - ((sum(y)^2) / n)
  return(SS_R)
}

SS_Res <- SS_Res_calc(beta_hat = beta_hat, X = X, y = y)
SS_T <- SS_T_calc(y = y)
SS_R <- SS_R_calc(beta_hat = beta_hat, X = X, y = y)

SS_Res == SS_T - SS_R
SS_Res; SS_T;  SS_R # 148.2573, 874.7109, 726.4536
k <- 2; n - k - 1; n - 1
MS_R <- SS_R / k # 363.2268
MS_Res <- SS_Res / (n - k - 1) # 7.803015
F <- MS_R / MS_Res # 46.54955

alpha <- 0.01
qf(p = (1 - alpha), df1 = k, df2 = (n - k - 1))
pf(q = F, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)

### c
p <- k + 1
r_squared_calc <- function(SS_R, SS_T) {
  r_squared <- SS_R / SS_T
  return(r_squared)
}
adj_r_squared_calc <- function(SS_Res, SS_T, n, p) {
  adj_r_squared <- 1 - ((SS_Res / (n - p)) / (SS_T / (n - 1)))
  return(adj_r_squared)
}

r_squared <- r_squared_calc(SS_R = SS_R, SS_T = SS_T) # 0.8305071
adj_r_squared <- adj_r_squared_calc(
  SS_Res = SS_Res, SS_T = SS_T, n = n, p = p) # 0.8126658

### 2.4
k <- 1; p <- k + 1
X <- cbind(ones, x1)
beta_hat <- beta_hat_calc(X = X, y = y)
H <- H_calc(X = X)
y_hat <- y_hat_calc(H = H, y = y)
SS_Res <- SS_Res_calc(beta_hat = beta_hat, X = X, y = y)
SS_T <- SS_T_calc(y = y)
SS_R <- SS_R_calc(beta_hat = beta_hat, X = X, y = y)

r_squared <- r_squared_calc(SS_R = SS_R, SS_T = SS_T) # 0.7911184
adj_r_squared <- adj_r_squared_calc(
  SS_Res = SS_Res, SS_T = SS_T, n = n, p = p) # 0.7806743

### d
beta_hat_1 <- beta_hat[2]
C <- solve(t(X) %*% X); C_11 <- C[2,2]
sigma_hat_squared <- MS_Res
```

```
alpha <- 0.05
t_stat <- qt(p = (1 - alpha / 2), df = n - p)

CI <- c(beta_hat_1 - t_stat * sqrt(sigma_hat_squared * C_11),
        beta_hat_1 + t_stat * sqrt(sigma_hat_squared * C_11))

### e
test_stat_1 <- beta_hat_1 / sqrt(sigma_hat_squared * C_11) # -8.793749

beta_hat_6 <- beta_hat[3]; C_66 <- C[3,3]
test_stat_6 <- beta_hat_6 / sqrt(sigma_hat_squared * C_66) # 2.101295

t_stat <- qt(p = (1 - alpha / 2), df = (n - k - 1))

### f
x_0 <- c(1, 275, 2)
y_hat_0 <- t(x_0) %*% beta_hat
var_y_hat_0 <- sigma_hat_squared * (t(x_0) %*% C %*% x_0)
test_stat_0 <- qt(p = (1 - alpha / 2), df = n - p)

CI_0 <- c(y_hat_0 - test_stat_0 * sqrt(var_y_hat_0),
          y_hat_0 + test_stat_0 * sqrt(var_y_hat_0))

### g
x_0 <- c(1, 257, 2)
y_hat_0 <- t(x_0) %*% beta_hat
var_y_hat_0 <- sigma_hat_squared * (1 + (t(x_0) %*% C %*% x_0))
test_stat_0 <- qt(p = (1 - alpha / 2), df = n - p)

PI_0 <- c(y_hat_0 - test_stat_0 * sqrt(var_y_hat_0),
          y_hat_0 + test_stat_0 * sqrt(var_y_hat_0))
```