Assignment 3-4

Jared Yu

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

Ans:

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

Then the resulting would appear as follows,

The following are some assumptions:

The data is such that there are observations and for and . It requires then that . Furthermore, is the observed response while and are the regressors. An assumption is that , , and that the errors are uncorrelated. In deriving the OLS estimators, it must be such that is calculatable, or in other words the inverse exists.

* 1. As the value of increases, the magnitude of change in the value of will depend on the value of . Write down the multiple linear regression model with assumptions for this scenario.

Ans:

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

Then the resulting would appear as follows,

The following are some assumptions:

The data is such that there are observations and for , , and . It requires then that . Furthermore, is the observed response while and are the regressors. An assumption is that , , and that the errors are uncorrelated. In deriving the OLS estimators, it must be such that is calculatable, or in other words the inverse exists. Here it is interesting since is the coefficient for the interaction term .

1. 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

Ans:

The “regular” way to express is as follows,

First, begin with the simpler terms on the right-hand side, and . It will be shown that these two are equivalent.

Next, it will be shown that and are equivalent.

Therefore, it follows that since and , then

1. Do Problem 3.27, page 127 of Textbook

Show that .

Ans:

Note that .

1. Do Problem 3.28, page 128 of Textbook

Prove that the matrices and are idempotent, that is and .

Ans:

Note that the previous proof is used, .

1. In a multiple linear regression model, , where is a response variable, and are non-random regressors, and is a random error, the parameter is **nonzero**. Suppose that subjects give data on to generate the ordinary least-squares (OLS) estimators of all three parameters in this model. We then fit the same data to the simple linear regression model .
   1. Create a hypothetical data set and perform regression analysis to compare the OLS estimate of in the regression model including with the OLS estimate of in the regression model excluding . What have you learned?

Ans:

In the simulation, a total of samples observations were generated based on the following model,

where , , and . The and are examples randomly generated from a and respectively. Let the model including be called the “full model,” while the model excluding be called the “reduced model.”

|  |  |
| --- | --- |
| Full Model | Reduced Model |
|  |  |
|  |  |
|  |  |
|  |  |

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 along with .

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 are quite similar in size. However, something else to take into account is the fact that itself comes from a distribution and therefore it makes sense that the corresponding coefficient has a similarly small magnitude.

A fairly large difference can be seen in the , 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 , 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 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 coefficient in both models remained fairly similar. It is possible that with a different that has a greater variance, the same results would still be seen.

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

Ans:

1. Use any math/stat software (e.g., [www.numbergenerator.org/randomnumbergenerator](http://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.

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

Ans:

The following model was created,

The resulting OLS estimators are as follows, , , and . They were calculated by using the least-squares normal equations, where

* 1. 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: , , , and . These are based on equations that can be seen on pp. 86-87 in the Textbook. Let the hypothesis be as follows,

After calculating the analysis-of-variance table (Table 2), the resulting critical value is . The critical value for at is 5.925879. The F-statistic is much larger than the critical value and so we reject the null hypothesis at the confidence level. Furthermore, the resulting p-value based on the F-statistic is . The conclusion then is that at least one of the , for is nonzero.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square |  |
| Regression |  |  |  |  |
| Residual |  |  |  |  |
| Total |  |  |  |  |

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

* 1. Calculate and for this model. Compare this to the and the for the simple linear regression model relating mileage to engine displacement in Problem 2.4.

Ans:

The formula for is and the formula for is . 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 and the right column shows the current model for Problem 3.5 where . It can be seen that there is an improvement in both the and values when the model includes the second regressor, . It is apparent that the and 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 and 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 |
|  |  |  |
|  |  |  |

Table 3 The table above shows the and values for the models from Problem 2.4 (left) and Problem 3.5 (right).

* 1. Find a 95% CI for .

Ans:

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

Here, and is the diagonal element of . Plugging in those values leads to the following:

* 1. Compute the statistics for testing and . What conclusions can you draw?

Ans:

The first hypothesis is as follows,

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

Then for the first hypothesis, . Since , we reject and conclude that the regressor contributes significantly to the model given that (carburetor) is also in the model at the confidence level.

In the second hypothesis,

It is testing whether can be deleted from the model. Then for the second hypothesis, . Since , we reject and conclude that the regressor contributes significantly to the model given that is also in the model at the confidence level.

It is interesting to see that the critical value along the -distribution is quite close to the cut-off point for . Although it passes the test at , it would not pass at . It seems then that in this model that includes only and that is more influential in the overall model.

* 1. Find a 95% CI on the mean gasoline mileage when and barrels.

Ans:

The formula for such a CI is as follows,

Here, The fitted value at this point is as follows,

The variance of is estimated by,

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

* 1. Find a 95% prediction interval for a new observation on gasoline mileage when and barrels.

Ans:

The formula for such a PI is as follows,

Here, The fitted value at this point is as follows,

The variance of is estimated by,

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

**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 , , and . This matches the context of . Furthermore, is the observed response while and are the regressors. The sample regression model has the following expression,

An assumption is that , , and that the errors are uncorrelated. In deriving the OLS estimators, it must be such that is calculatable, or in other words the inverse exists.

Some more assumptions are that the error term, , follows a normal distribution, so follows normal distribution also. Furthermore, is a linear combination of , so also follows normal distribution, however, we don’t know the true variance of , we use instead, so, both CI and prediction interval use a -value from the Student’s 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))