Assignment 1-2

Jared Yu

1. In a simple linear regression analysis, *independent* paired data are fitted to the model

where is the only non-random independent variable (or so-called regressor), is a known real number, and is the random error that has mean zero and unknown constant variance . Before the data for are available, we need to construct estimators for the parameters.

1. Construct the ordinary least squares (OLS) estimator of .

Ans:

(*Note: Let* )

where

1. Construct the variance of the OLS estimator of in a) and construct an unbiased estimator of this variance.

Ans:

The following shows the calculation of the variance of :

This is possible since the term has only the constant term and fixed , it can factor out of the variance and the covariance will zero out. The sum of the variance is the variance of the sum (where ), based on the assumption that the terms are uncorrelated.

Next, an unbiased estimator of the variance will be constructed. It requires that an estimator for , let the estimator of be denoted as .

First, is as follows:

The variance of is as follows:

Next, the expectation of is as follows:

(*Note: Let* )

To simplify the process, first let’s focus on the green section:

Next, let’s focus on the blue section:

(*Note: Let* )

Notice that:

(*Note: Let* )

From this is follows that:

It follows that the unbiased estimator of , can be rewritten as:

Therefore, the unbiased estimator of is as follows:

Since the following holds true:

1. In Problem 1, add the intercept term to the model. Then do a) and b).

Ans:

The model with the intercept added is as follows,

The least squares criterion can be written as follows,

Then the least-squares estimator of , is as follows:

The least-squares estimator of , is as follows:

Here,

Since

and

In trying to derive the variance and an unbiased estimator of the variance, it could be useful to try and re-express the formula in an alternate format:

where

Then it follows that the estimator of the slope is as follows:

Notice:

and that:

Furthermore, can be rewritten to:

where

and

Also, for , let the following be shown:

Next, let’s analyze the variance of the coefficients:

since consists of terms that are viewed as constants

The unbiased estimator of is (based on the textbook results and expanding to the version):

where

and

Then it follows that:

So, the unbiased estimator of is:

where

since

Next, let us look at the other coefficient, , consider:

Let’s examine the above in pieces:

Thus, we have:

Then, the unbiased estimator of is

where

and

1. In Problem 1, the is a real number but the value is unknown. Please do a) and b).
2. Construct the ordinary least squares (OLS) estimator of .

Ans:

*Note: It is assumed that .*

*Note: It is ignored when .*

where

Returning back to :

where

1. Construct the variance of the OLS estimator of in a) and construct an unbiased estimator of this variance.

Ans:

The following has been shown in part a):

where .

Since the term has only the fixed , it can factor out of the variance.

Thus, is has been shown that .

Next the unbiased estimator will be constructed.

From here it can be seen that none of the highlighted parts in the section above contain any reference to the variable or . It is the same as the normal for simple linear regression. Therefore,

or,

Therefore, the unbiased estimator of is as follows:

Since the following holds true:

1. Consider a regression model , where is a non-random regressor. Discuss whether the ordinary least-squares estimator of the slope is always unbiased and whether it always has the smallest variance than **any** estimator of , irrespectively of what the value of is. State assumptions in your discussion. Be careful about the word “**any**”.

Ans:

For to be unbiased, it requires that . For this to be the case, there are two important requirements to that need to be addressed, and . The first requirement needs the assumption that the error terms have a mean of . The second requirement needs that the are distributed in such a way that the term can exist. More specifically,

must be such that the denominator, is nonzero. If the data has the characteristic where , then would fail to exist since it’s not possible to divide by .

To show the steps more carefully, it will be shown how can be seen as unbiased:

To show that the OLS estimate of is unbiased, the following two equations must hold true:

The formula for are is follows:

Furthermore, the formula for can be further rewritten as follows:

where . To show the unbiasedness, we will first start with (pp. 18-19):

In equation (6), we have simply applied the expectation to .

In equation (7), we move the expectation into the summation, based on the *linearity of expectation*. Furthermore, we are treating as a constant, fixed term, since it consists entirely of . It is important to note that this step is not possible if is also a random variable.

In equation (8), the is rewritten as , since . Then the summation is distributed to each part of this new term.

In equation (9), it is shown that the term equals 0, since .

In equation (10), the formula has been expanded. To show that it is equivalent to , it must be shown that the numerator is equivalent to the denominator . This will be shown as follows:

So, it follows that and , therefore equation (8) evaluates to . Thus far it has been shown then that is an unbiased estimator. From here it has been shown that derived using OLS is always unbiased given the conditions of and .

Another important assumption in the model is that the variance of each terms are all , in other words they’re homoscedastic. It can also be said that they must be uncorrelated. Without this assumption, then the Gauss-Markov theorem for the OLS estimators does not hold. The assumptions of the Gauss-Markov theorem are that and , and uncorrelated errors.

The Gauss-Markov theorem makes the case that the OLS coefficients and , given that the assumptions are met, are unbiased and have minimum variance in comparison to other unbiased estimators that are linear combinations of the observations . This definition makes it clear that it only has the minimum variance in comparison to other estimators that are linear combinations of . This excludes certain other models which may obtain a smaller variance with a more complicated structure.

It is worth noting, that based on my entire discussion response to Module 2, that I had stated it is possible to construct OLS estimators without the constant variance assumption holding true within the data. Therefore, it is possible to construct OLS estimates without this assumption. The issue is that the OLS estimates will no longer have the property of best linear unbiased estimators (BLUE) as outlined in the Gauss-Markov theorem. They will still have the unbiased property, but they will not have the minimum variance amongst other unbiased estimators that are linear combinations of the .

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 15 rows of Table for Problem 2.18 (page 63-64) of Textbook and then do (a), (b), (c), (d).

**State assumptions for all steps in your analyses.**

Chosen rows: 1, 2, 3, 4, 5, 7, 9, 11, 13, 14, 16, 17, 18, 19, 21.

1. Fit the simple linear regression model to these data.

In the following, the amount a company spends on advertising are represented by and the retained impressions are represented by the .

Some important assumptions are that the error term has mean zero, constant variance and so are uncorrelated. Also, the error terms need to be independent. This also follows from the thinking that is nonrandom, but such an assumption is not certain and it’s possibly a random variable also. In that case, further assumptions would be required for the model to hold. For example, the conditional expectation of the error term w.r.t. also needs to have mean zero.

1. Is there a significant relationship between the amount a company spends on advertising and retained impressions? Justify your answer statistically.

To test whether there’s a significant relationship between the amount a company spends on advertising and retained impressions, a hypothesis test will be done on to see if it’s equal to . If it can be shown that statistically , then it would be concluded that there is no linear relationship between the two variables. To reject the null hypothesis, it must be shown that , where is the test statistic (shown below) and is the critical value of the Student’s *t*-distribution at the percentage point and is the degrees of freedom. *Note: The textbook uses to refer to the same critical value. The difference is versus , but this is simply differing notation for the same t-table.*

The critical value of is approximately . Since , the decision is to reject the null hypothesis at the confidence level. The conclusion then is that there’s a 95% probability that and that there’s a significant relationship between the amount a company spends on advertising and retained impressions.

*Note: Changing to 0.01 leads to the critical value being approximately 9.9248. Therefore, the conclusion would be to fail to reject the null hypothesis that at the 0.01 confidence level.*

1. Construct the 95% confidence and prediction bands for these data.

The assumption is that the errors, , are normally and independently distributed. Then the sampling distribution of is a Student’s *t*-distribution with degrees of freedom. The percent confidence interval (CI) of

First the 95% confidence bands for the data will be created. To do this, we are looking to estimate the mean response for a value from . Then let be a level from that we wish to estimate the mean response with, An assumption it that is within the range of the original data of . The following is an unbiased point estimator of ,

The point estimator is a normally distributed random variable since it’s a linear combination of the observations , where it’s assumed that the . The variance of is as follows (proof on p. 31 of the textbook),

To derive this result, it must be that .Then the sampling distribution of the standardized point estimator,

is a Student’s *t*-distribution with degrees of freedom. Therefore, the percent CI of the mean response at point is as follows,

Each of the variables in the confidence interval (i.e., , , , , , and ) have been calculated already in the above problems. The result of this calculation can be seen below in Table 1. Furthermore, a graph of the confidence band can be seen below in Figure 1.

Next, the process will be shown on constructing a 95% prediction interval. In the case of a new point that isn’t part of the original data, the point estimate is as follows,

This is the same value for as seen above with the confidence interval. Then the random variable follows a normal distribution with mean zero and the following variance,

An assumption is that the covariance zeroes out because the point is independent of . By predicting with , the standard error of can then be used for a prediction interval. The prediction interval for then is as follows,

Like before, all the variables have been calculated already in the previous work shown. The result of this calculation can be seen below in Table 1. Furthermore, a graph of the prediction band can be seen below in Figure 1.

A picture containing building

Description automatically generated

Table The table shows the upper and lower bounds for the confidence interval (CI) and prediction interval (PI) of the data. The indicates the point from which the calculations are based upon. This value is identical to the variable from the dataset. The variable is also included as a reference. (Note: the variable is the amount a company spent on advertising in millions of dollars and the variable is the retained impressions per week in millions.)

A close up of a map

Description automatically generated

Figure 1 The above figure shows a scatter plot of the retained impressions against the amount spent from the dataset. The points represent each of the companies. The solid blue line is the estimated regression function , the dotted blue line is the confidence interval for and the dotted red line is the prediction interval for .

1. Give the 95% confidence and prediction intervals for the number of retained impressions for MCI.

The company MCI has and for the two and variables in the dataset. It’s worth noting that this company wasn’t included in the original dataset since it was not sampled by the random generator. Another important note is that these values aren’t outside the range of the original data.

The 95% confidence interval for MCI is as follows:

The 95% prediction interval for MCI is as follows:

**Supplemental Exercises**

1. Create a hypothetical data set and then perform simple linear regression analysis and the corresponding inverse regression analysis.

A hypothetical dataset was generated in the following manner. A true slope was set to 1.5 and a true intercept was set to 5. Next, 15 points were generated from the normal distribution with a mean of 25 and a standard deviation of 1.5. Another 15 points were generated from the normal distribution with a mean of 0 and a standard deviation of 1.2. The first set of points represent the hypothetical values and the second set represent the corresponding terms. The set of values were generated by multiplying the values with the slope and adding the intercept and error terms.

Using simple linear regression, the following are the resulting regression coefficients:

Testing the significance of the regression, the following hypothesis test was conducted:

The critical value of is approximately . Since , the decision is to reject the null hypothesis at the confidence level. The conclusion then is that there’s a 99% probability that and that there’s a significant relationship between the and variables.

The confidence and prediction bands were also generated for each of the 15 points in the hypothetical dataset. They can be seen below in Table 2. This was done in the same style as it was done before with the textbook data. The model is also plotted in Figure 2 (further below).

A screenshot of a cell phone

Description automatically generated

Table 2 The table shows the upper and lower bounds for the confidence interval (CI) and prediction interval (PI) of the data. The indicates the point from which the calculations are based upon. This value is identical to the variable from the dataset. The variable is also included as a reference.

The process will now be somewhat repeated using inverse regression analysis. The model for inverse regression is as follows,

The term is estimated using . The term is estimated using . The following shows these estimated coefficients from inverse regression:

The critical value of is approximately . Since , the decision is to reject the null hypothesis at the confidence level. The conclusion then is that there’s a 99% probability that and that there’s a significant relationship between the and variables.

The model can be seen below in Figure 2.

A close up of a map

Description automatically generated

Figure 2 The figure shows the 15 simulated data points. The true regression function that the simulated points are based off is seen as a solid blue line. The estimated regression and estimated inverse regression functions are plotted also as dotted blue and red lines, respectively. They’re directly overlapping and so it’s difficult to discern them from each other.

1. Construct unbiased estimators for the new intercept and new slope in the regression of on .

The formula for the model is given as follows:

It will be re-expressed to follow more similarly the traditional simple linear regression formula:

where and

Some important assumptions are that has mean and constant variance, . It also follows that the error terms are uncorrelated.

In the case where is nonrandom:

where

So,

Then it follows that for that the unbiased estimator is:

Lastly, for , the unbiased estimator is:

**Code:**

set.seed(0) # Set seed  
random\_rows <- sample(1:21, 15) # Select rows  
sort(random\_rows)  
advertisements <- data.frame( # input data  
 x = c(50.1, 74.1, 19.3, 22.9, 82.4,  
 185.9, 20.4, 27, 154.9, 5,  
 26.9, 5.7, 7.6, 9.2, 6.1),  
 y = c(32.1, 99.6, 11.7, 21.9, 60.8,  
 92.4, 21.4, 40.8, 88.9, 12,  
 38, 10, 12.3, 23.4, 4.4)  
)  
  
# part a  
# Coefficients  
x\_bar <- mean(advertisements$x)  
y\_bar <- mean(advertisements$y)  
S\_xy <- sum((advertisements$x - x\_bar) \* (advertisements$y - y\_bar))  
S\_xx <- sum((advertisements$x - x\_bar)^2)  
beta\_hat\_1 <- S\_xy / S\_xx  
beta\_hat\_0 <- y\_bar - beta\_hat\_1 \* x\_bar  
  
# part b  
# Testing significance  
SS\_T <- sum((advertisements$y - y\_bar)^2) # Sum of Squares Total  
SS\_Res <- SS\_T - beta\_hat\_1 \* S\_xy # Sum of Squares Residual  
n <- nrow(advertisements) # Number of observations  
MS\_Res <- SS\_Res / (n - 2) # Mean Squared Residual  
se\_beta\_hat\_1 <- sqrt(MS\_Res / S\_xx) # Standard error of beta-hat-1  
t\_0 <- beta\_hat\_1 / se\_beta\_hat\_1 # Test statistic  
  
alpha <- 0.05  
qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)  
alpha <- 0.01  
qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)  
  
# part c  
# confidence band, prediction band  
### The band\_calculator function calculates the confidence interval  
### and prediction interval of the mean of y for a given point x\_0  
### at a given alpha level.  
band\_calculator <- function(x\_0, alpha = 0.05) {  
 mu\_hat <- beta\_hat\_0 + beta\_hat\_1 \* x\_0  
 y\_hat\_0 <- mu\_hat  
   
 denominator\_ci <- sqrt(MS\_Res \* ((1 / n) + ((x\_0 - x\_bar)^2) / S\_xx))  
 denominator\_pi <- sqrt(MS\_Res \* ((1 + 1 / n) + ((x\_0 - x\_bar)^2) / S\_xx))  
   
 critical\_value <- qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)  
   
 ci <- c(mu\_hat + critical\_value[1] \* denominator\_ci,  
 mu\_hat + critical\_value[2] \* denominator\_ci)  
   
 pi <- c(y\_hat\_0 + critical\_value[1] \* denominator\_pi,  
 y\_hat\_0 + critical\_value[2] \* denominator\_pi)  
 return(c(ci, pi))  
}  
band\_vec <- Vectorize(band\_calculator, vectorize.args = c("x\_0"))  
ci\_pi <- t(band\_vec(x\_0 = advertisements$x, alpha = 0.05))  
ci\_table <- data.frame(lower\_bound\_pi = ci\_pi[,3],  
 lower\_bound\_ci = ci\_pi[,1],  
 x\_0 = advertisements$x,  
 upper\_bound\_ci = ci\_pi[,2],  
 upper\_bound\_pi = ci\_pi[,4],  
 y = advertisements$y)  
ci\_table <- ci\_table[order(ci\_table$x\_0),]  
rownames(ci\_table) <- NULL  
colnames(ci\_table) <- c("PI Lower Bound", "CI Lower Bound", "x0",  
 "CI Upper Bound", "PI Upper Bound", "y")  
knitr::kable(x = ci\_table, "markdown")  
  
# plot confidence and prediction bands  
plot(advertisements$x, advertisements$y, pch = 20,  
 xlim = c(-10, 250), ylim = c(-20, 250),  
 main = "Retained Impressions vs. Amount Spent",  
 xlab = "Amount Spent (millions)",  
 ylab = "Returned Impressions per week (millions)")  
abline(a = beta\_hat\_0, b = beta\_hat\_1, col = 'blue')  
x\_seq <- seq(-20, 250, length.out = 1e4)  
  
# Calculate bands  
ci\_pi\_lines <- band\_vec(x\_0 = x\_seq, alpha = 0.05)  
  
# Plot lines  
lines(x\_seq, ci\_pi\_lines[1,], lty = 2, col = 'blue')  
lines(x\_seq, ci\_pi\_lines[2,], lty = 2, col = 'blue')  
lines(x\_seq, ci\_pi\_lines[3,], lty = 2, col = 'red')  
lines(x\_seq, ci\_pi\_lines[4,], lty = 2, col = 'red')  
legend("topleft",  
 legend = c("Estimated Regression Function",  
 "95% Confidence Interval", "95% Prediction Interval"),  
 lty = c(1, 2, 2), col = c("blue", "blue", "red"))  
  
# part d  
MCI\_x <- 26.9  
band\_calculator(x\_0 = MCI\_x, alpha = 0.05)

### Supplemental exercises  
# create a hypothetical dataset  
hypothetical\_intercept <- 5  
hypothetical\_slope <- 1.5  
set.seed(0)  
hypothetical\_data <- data.frame(  
 x = rnorm(n = 15, mean = 25, sd = 1.5),  
 epsilon = rnorm(n = 15, mean = 0, sd = 1.2)  
)  
hypothetical\_data$y <- hypothetical\_intercept +  
 hypothetical\_slope \* hypothetical\_data$x +  
 hypothetical\_data$epsilon  
  
# Coefficients  
x\_bar <- mean(hypothetical\_data$x)  
y\_bar <- mean(hypothetical\_data$y)  
S\_xy <- sum((hypothetical\_data$x - x\_bar) \* (hypothetical\_data$y - y\_bar))  
S\_xx <- sum((hypothetical\_data$x - x\_bar)^2)  
beta\_hat\_1 <- S\_xy / S\_xx  
beta\_hat\_0 <- y\_bar - beta\_hat\_1 \* x\_bar  
  
# Testing significance  
SS\_T <- sum((hypothetical\_data$y - y\_bar)^2) # Sum of Squares Total  
SS\_Res <- SS\_T - beta\_hat\_1 \* S\_xy # Sum of Squares Residual  
n <- nrow(hypothetical\_data) # Number of observations  
MS\_Res <- SS\_Res / (n - 2) # Mean Squared Residual  
se\_beta\_hat\_1 <- sqrt(MS\_Res / S\_xx) # Standard error of beta-hat-1  
t\_0 <- beta\_hat\_1 / se\_beta\_hat\_1 # Test statistic  
  
alpha <- 0.1  
qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)  
alpha <- 0.05  
qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)  
alpha <- 0.01  
qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)  
  
# confidence and prediction bands  
ci\_pi <- t(band\_vec(x\_0 = hypothetical\_data$x, alpha = 0.01))  
ci\_table <- data.frame(lower\_bound\_pi = ci\_pi[,3],  
 lower\_bound\_ci = ci\_pi[,1],  
 x\_0 = hypothetical\_data$x,  
 upper\_bound\_ci = ci\_pi[,2],  
 upper\_bound\_pi = ci\_pi[,4],  
 y = hypothetical\_data$y)  
ci\_table <- ci\_table[order(ci\_table$x\_0),]  
rownames(ci\_table) <- NULL  
colnames(ci\_table) <- c("PI Lower Bound", "CI Lower Bound", "x0",  
 "CI Upper Bound", "PI Upper Bound", "y")  
knitr::kable(x = ci\_table, "markdown")  
  
inverse\_coeff\_1 <- beta\_hat\_0 / beta\_hat\_1  
inverse\_coeff\_2 <- 1 / beta\_hat\_1  
est\_x <- - inverse\_coeff\_1 + inverse\_coeff\_2 \* hypothetical\_data$y  
  
# plot data and lines  
plot(hypothetical\_data$x, hypothetical\_data$y,  
 main = 'Simulated Hypothetical Data',  
 xlab = 'x', ylab = 'y')  
  
abline(a = hypothetical\_intercept, b = hypothetical\_slope, col = 'blue', lty = 1)  
abline(a = beta\_hat\_0, b = beta\_hat\_1, col = 'blue', lty = 2)  
legend("topleft", legend = c("True Regression Function",  
 "Estimated Regression Function",  
 "Estimated Inverse Regression Function"),  
 col = c('blue', 'blue', 'red'), lty = c(1, 2, 2))  
  
lines(x = est\_x, y = hypothetical\_data$y, col = 'red', lty = 2)