Test 2

Jared Yu

1. Prove that in using a regression model analysis to compare the differences of the expected values of the response variable between the levels of a categorical regressor , the sum of squares, , , , will not change regardless of how the indicators of are coded (Recall any dummy variable can be coded in many ways, e.g., or , or others). [10 points]

**State assumptions in each step of your proof.**

Ans:

In this problem we are analyzing the situation of a single regressor that is a categorical variable with levels. It uses a type of encoding that leads to dummy variables. The type of encoding, , is arbitrary. To show that the encoding doesn’t impact , , or , we can first look at the formulas (based on the Textbook) for the total sum of squares (), regression sum of squares (), and residual sum of squares ()

There are alternate forms for these sums of squares, which I have done in Discussion 4. The same reformulations will be shown below. (*Note: An assumption is that* *is calculatable.*)

:

It will be shown that an equivalent formula for is , where is a vector consisting all of identical terms.

The previous step holds since and are both scalars with the same resulting values. So, the transpose of each other is still the same scalar.

Let’s first look at . It is the inner product of the same vector which is filled with identical terms. Therefore, the result is .

Next, let’s look at . By first ignoring the constant term -2, it can be seen that . Therefore, .

From this it follows that,

:

It will be shown that an equivalent formula for is .

Starting with :

Next has already been shown in .

It can be said that and are both scalars where the same transpose logic can be applied as with and . So, it will be said first that .

Note that is an matrix consisting entirely of 1’s.

Note that the hat matrix, , is a symmetric matrix.

Note that , this will first be explained. Let be an vector consisting of 1’s only.

What is shown above is that first the matrix is expressed as a vector of vectors, . The hat matrix is then multiplied to this to get to . This is simply another way to show typical matrix multiplication. The point is that the hat matrix is a projection matrix on the column space of , the design matrix. As the design matrix, it includes in the first column a vector of 1’s, or . times **1**, i.e. project **1** onto the column space of , is equal to **1** since **1** is already in the column space of . **J** is a matrix filled with all 1’s. The result then is that multiplied by **J** is also equal to **J**.

Therefore,

:

It will be shown that an equivalent formula for is .

Now, looking at the formulas for each of these sums of squares, it can be seen first that only contains the terms. Therefore, any sort of encoding will not directly impact it. So, this sum of squares is unaffected. However, it can be seen that and both depend on in addition to the terms. It must be shown then that any type of encoding will have no impact on the resulting term. It is worth noting however, that changing the encoding will change both the term and term which when multiplied together make up .

To prove then that does not impact , it will be shown that the hat matrix , which is the projection matrix that projects the vector onto the column space of , does not change due to changes in the type of encoding, . The logic is that since , if is unimpacted, then is unchanged. From this it would be clear that , , is unimpacted by the encoding .

We can start by looking at . Given that it is the design matrix for the regressor, it can be written as follows

The meaning is such that the first column, , is simply the vector consisting only of 1’s that are used to represent the intercept term in the linear regression model. For each of the categories that are encoded as dummy variables, they are binary in the sense that they have one value in the case that the observation is true for that level and another value otherwise. In the common case, it is 1 if the observation is true for that level and 0 otherwise. It can however be changed to something like 1 if true and -1 otherwise, or any other variety of values that act in such a way.

We can generalize this to something like in the case of the first level out of the levels. We can think of and as the two possible types of encodings for a level. For example, in the case of 1, -1 for the type of encoding, then and . In the case that the encoding varies per level, then each of the and can vary for each of the levels that can take on. The and terms are such that they are vectors with binary values 1 and 0. Each of them are 1 in the case that an observation is true for that treatment and 0 otherwise.

To better understand these, we can first look at a simple example. Let’s say that there are levels and so we have 2 dummy variables. The levels represent treatment A, B, and C respectively. To use dummy variables, we drop the level C from representation and say that

This would lead to the following regression model

where . In the case that the first observation is from treatment A, and the second is from treatment B, it can be seen as following in the design matrix

where the first row corresponds to the first observation and the second corresponds with the second. The first column is all 1’s to go with the intercept term . The second and third columns correspond with and respectively. This is similar to what has been in the textbook so far. However, to do the generalized encoding we can represent the second column as

and the third column as

We can first focus on the second column, which when looking at the design matrix is . In such a case, we can represent it using a set of two binary vectors multiplied by their corresponding encodings. This can be seen as follows

which can also be extended to the third column

As can be seen above, the and correspond with the positive and negative encodings of respectively for the th treatment level of the categorical variable. The and are binary vectors that act as Boolean values to help indicate for a certain observation whether it should be positive or negative for a treatment level, after multiplying by the and values.

Next, we can try to better frame the argument for how we can compare to hat matrices. We can denote the hat matrix derived from a matrix with one arbitrary type of encoding as and the hat matrix derived from another matrix with a different arbitrary type of encoding as . Therefore, we would like to know if it’s guaranteed that in such a regression model, . (*Note: An assumption is that the encoding, , consists of real numbers only.*)

Thinking back to we have,

where the column space of , denoted can be seen as

however, this is in the general sense where hasn’t yet been given the special generalization. So, to combine the two we have

This can be thought of as the span of all the vectors , , and which can be denoted as

From this, it can be understood that for any type of encoding used for the design matrix , the span will be identical.

Then the hat matrix, , which is the projection matrix of onto , for any type of encoding will lead to the same identical . To more easily think about it, the hat matrix projects vertically onto the column space (in the simple case of only have two dimensions). Since the column space is the same for any type of encoding, the resulting projection matrix must also be identical across all encodings, . It can then be understood that for any two arbitrary types of encoding that .

Therefore, since is unimpacted by the type of encoding, then will also be unimpacted by the encoding. It makes sense that not only is unchanged, but and which are dependent on must also remain unchanged.

1. In a study, there are four treatments (labeled as 1, 2, 3, 4) to compare. Assume that there are subjects per treatment.
   1. Construct an analysis of variance model to compare the four treatments; that is, test whether there is at least one pair of treatments that differ and construct an estimator of every pair of expected treatment difference. [20 points]

Ans:

An analysis of variance model for such a study can be seen below,

where is the th observation for the th treatment, is the grand mean, is a parameter that represents the effect of the th treatment, and is an error component.

To test whether at least one pair of treatments differ, it is possible to look at it using the following hypothesis test,

The reason is that if every pair of treatments are the same, they must be all equal to zero based on the condition that within the analysis of variance model. The test statistic for such a hypothesis is as follows,

where and . If , then we reject the null hypothesis in favor of the alternative. Otherwise we fail to reject the null hypothesis at significance level . In the case of rejecting the null hypothesis in favor of the alternative, then we conclude that at least one pair of treatments are different. If we fail to reject the null hypothesis, then the conclusion is that there is no difference across the different treatment levels.

To construct an estimator for every pair of expected treatment difference, there would be total possibilities. The below Table 1 shows all such combinations. *Note: It is possible to show more combinations such as not only Treatment 1 – Treatment 2, but also Treatment 2 – Treatment 1. However, this would be repetitive and not show any new interesting information so it will be skipped.*

Table The below table shows all the different combinations for comparing the difference between treatments.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Treatment 3 – Treatment 4 |  |  |

* 1. Construct a linear regression model such that the regression analysis is equivalent to the analysis of variance in (a). [20 points]

Ans:

The same analysis of variance model in part a) can be rewritten using a regression model. Since there are four treatment levels, it would require three indicator variables defined as follows:

The corresponding regression model then becomes

where is the value of the indicator variable for observation in treatment , is the value of the indicator variable for observation in treatment , and is the value of the indicator variable for observation in treatment . Furthermore, is an error component.

Comparing the regression model to the analysis of variance model, with a total of four treatments in the study, the regression coefficients can be understood as follows

The above formulation comes from the following basic example of treatment 1, where

Applying the above example of the th observation of treatment 1, it leads to

Then, looking towards the analysis of variance model we have,

Comparing these two results we have

Likewise, for any treatment , we have

Lastly, for treatment 4, the resulting regression model is

while the analysis-of-variance model is

Therefore, for treatment 4,

From the above derivation, it follows then that

as stated previously.

From here, to test the null hypothesis, we can instead of

we can test

The reason that we can use such a hypothesis test is that if for , then for (based on the above derivation for , ). Such a hypothesis then is equivalent to testing .

To test the above hypothesis test we can use the following test statistic,

To understand the test statistic, first the coefficients will be stated:

The full model of the regression sum of squares is as follows

The reduced model regression sum of squares is as follows

From here it follows that the partial regression sum of squares is as follows

This last step will be proven below

Then for the denominator, can be seen as follows

The proof for the last step will be shown below

From this it follows that the test statistic becomes

This derived test statistic is the same as the one in the previous part. It likewise follows the distribution.

**State assumptions in each step of your proof or derivation in a) and b).**

**Do not use any math/stat software for calculation for Problem 3, except for obtaining the percentile of standard normal, t, chi-square, or F distribution, matrix operations, or basic mathematical calculations.**

1. Ten observations on the response variable associated with two regressor variables and are given in the following table. The model fitted to these observations is

where ’s are identically and independently distributed as normal random variable with mean zero and a known variance .

* 1. Test the null hypothesis “there is no difference between the -intercept for and the -intercept for and there is no difference between the slope for and the slope for ” at a statistical significance level of 0.05. [20 pts]

Ans:

In the case of :

In the case of :

Then the null hypothesis is stating that

Therefore, the null can be hypothesis stated as

Turning this into a model, we can state both the null and the alterative as,

The above is a partial test, where we are comparing a reduced model to the full model. The test statistic then can be seen as follows,

where follows a distribution.

The numerator can be understood as follows

We can first look at the reduced model

where are the set of coefficients estimates pertaining to the reduced model, and is the reduced dataset used to generate that set of coefficient estimators. The full model regression sum of squares can be seen as follows

Therefore,

From this, the test statistic can be calculated to be , which has a corresponding -value of . This -value is larger than the significance level , therefore the decision is that we fail to reject the null hypothesis. The conclusion then is that there is not enough evidence to reject the claim that there is no difference between the -intercept for and the -intercept for and there is no difference between the slope for and the slope for

* 1. Estimate the difference, , and calculate its 95% confidence interval. [10 pts]

Ans:

(*Note: Here the full model in part a) is being used, rather than the reduced model.*) We can estimate first try to estimate the difference by finding out the formula for the expected differences between the two points.

So, we can estimate the difference by using the estimated coefficients and . The resulting .

Let and be the estimated values given and respectively, where and . To find the confidence interval of this difference, we must first find the variance of the difference. This will be shown as follows

From this, we can try to normalize the difference

So, we see that the confidence interval is

Therefore, it follows that the 95% C.I. is

* 1. Predict the difference in value at between and . [10 pts]

Ans:

We can think of a predicted observation as . For a new example separate from the dataset, we can denote it as . Then for the two new observations and , we have and . We must then find out the variance for the difference between the two. (*Note: An assumption is that the random errors have the same distribution as the random errors from the sample data.*)

So, it follows that a P.I. for is as follows

Therefore, it follows that a 95% P.I. is

The question however doesn’t specifically ask for a prediction interval. It asks to predict the difference between the two predictions for the two and observations. In such a case, it would be the same as in the previous part where we are looking at the estimated difference . However, from the Textbook so far, in the situations of discussing predictions, it has only been the case that we are looking at prediction intervals rather than simply the predictions themselves.

* 1. Now fit Model (2): to the 10 observations. Calculate the residual for the observation #8 and its variance. [10 pts]

Ans:

The Model 2 can be fit using linear regression. It leads to the following estimated coefficients and . The residual for the 8th observation is . The variance for this residual can be calculated as follows

where . So, it follows that the variance for the 8th observation is the value in row 8, column 8 of . Calculating this in RStudio leads to a value of .

**State assumptions in your derivations and calculations in a), b), c), d).**

**Code Appendix**

### Problem 3  
df <- data.frame(  
 y=c(7,8,5,4,2,10,9,10,8,8),  
 x1=c(9,6,10,8,5,7,6,5,5,4),  
 x2=rep(c(1,-1), each=5)  
)  
y <- df[,1]  
n <- nrow(df)  
ones <- rep(1, n)  
x1 <- df[,2]; x2 <- df[,3]  
X <- cbind(ones, x1, x2, x1\*x2)  
  
### part (a)  
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)  
}  
  
X\_red <- cbind(ones, x1)  
beta\_hat\_full <- beta\_hat\_calc(X = X, y = y)  
beta\_hat\_red <- beta\_hat\_calc(X = X\_red, y = y)  
  
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\_calc <- function(y, beta\_hat, X) {  
 SS\_Res <- (t(y) %\*% y) - (t(beta\_hat) %\*% t(X) %\*% y)  
 return(SS\_Res)  
}  
  
SS\_Res\_full <- SS\_Res\_calc(y = y, beta\_hat = beta\_hat, X = X)  
k <- 3; p <- k + 1  
r <- 2  
MS\_Res <- SS\_Res\_full / (n - p)  
SS\_R\_full <- SS\_R\_calc(beta\_hat = beta\_hat, X = X, y = y)  
SS\_R\_red <- SS\_R\_calc(beta\_hat = beta\_hat\_red, X = X\_red, y = y)  
  
F\_0 <- ((SS\_R\_full - SS\_R\_red) / r) / MS\_Res  
alpha <- 0.05  
qf(p = (1 - alpha), df1 = k, df2 = (n - k - 1))  
pf(q = F\_0, df1 = r, df2 = (n - k - 1), lower.tail = FALSE)  
  
### part (b)  
2 \* beta\_hat[3] + 10 \* beta\_hat[4] # -4.38551  
  
z\_value <- qnorm(alpha/2, lower.tail = FALSE)  
x\_01 <- c(1, 5, 1, 5)  
x\_02 <- c(1, 5, -1, -5)  
y\_hat\_01 <- t(x\_01) %\*% beta\_hat  
y\_hat\_02 <- t(x\_02) %\*% beta\_hat  
sigma\_squared <- 2  
  
CI\_bound <- z\_value \* sqrt(sigma\_squared \* t(x\_01 - x\_02) %\*%  
 solve(t(X) %\*% X) %\*%  
 (x\_01 - x\_02))  
  
(y\_hat\_01 - y\_hat\_02) + CI\_bound  
(y\_hat\_01 - y\_hat\_02) - CI\_bound  
  
### part (c)  
PI\_bound <- z\_value \* sqrt(sigma\_squared \*  
 (2 +  
 t(x\_01 - x\_02) %\*%  
 solve(t(X) %\*% X) %\*%  
 (x\_01 - x\_02)))  
  
(y\_hat\_01 - y\_hat\_02) + PI\_bound  
(y\_hat\_01 - y\_hat\_02) - PI\_bound  
  
### part (d)  
X\_d <- X[,c(1,3)]  
beta\_hat\_d <- beta\_hat\_calc(X = X\_d, y = y)  
H\_d <- H\_calc(X = X\_d)  
y\_hat\_d <- y\_hat\_calc(H = H\_d, y = y)  
e\_d <- e\_calc(y = y, y\_hat = y\_hat\_d)  
  
e\_d[8]  
  
residual\_variance <- sigma\_squared \* (diag(n) - H\_d)  
residual\_variance[8,8]