Homework Assignment 2

2023-02-08

Instructions

Follow the instructions as for the first homework assignment. You need to submit both .Rmd and .pdf files. But I will allow you to submit them separately or as a zipped file. Make sure the submissions are made by **noon (12pm) next Wednesday, February 15, 2023.**

Simple linear regression

In this question, you will load the flights dataset and build a simple linear regression model to relate the arrival delay of flights as the DV to the departure delay as the IV. (2+2+2+2+4+4+4 points)
a. Write the regression equation in both forms (i.e., with and without error terms) as shown in class.
b. Load the flights dataset from the nycflights13 package using the library("nycflights13") function (you need to have already installed this package on your computer using the install.packages("nycflights13") command in R).
c. Run a simple linear regression model using the lm() function and save the model as M1.
d. Output the elements of M1 using the summary() function.
e. Explain the regression coefficients (intercept and slope estimates) in one sentence each. I want a "common sense" explanation, e.g., "The regression slope is AAA which denotes that the arrival delay changes by AAA hours for every 1 hour increase in departure delay" etc.
f. Explain the $Pr(>|t|)$ values for the two coefficients and how they capture the uncertainty in the regression coefficient estimates.
g. Explain what the R^2 numerical value indicates for this particular model in one sentence.

Multiple linear regression

Now we are going to build a multiple linear regression model with 3 IVs: dep_delay, sched_arr_time, and distance. Note that all 3 IVs are quantitative in this case. (2+2+6+6+4 points)
a. Write out the regression equations in both forms for this multiple linear regression model.
b. Build the model in R using the lm() function and save this model as M2.
c. Explain the regression constant and the different regression coefficients in 1 precise sentence each. Note that each regression coefficient has to be explained also in terms of keeping all the other IVs constant. Also, explain how the regression coefficient for dep_delay changed between M1 and M2.
d. Comment on the uncertainty in each of the regression coefficient estimates based on the $Pr(>|t|)$ values in the R output.
e. Comment on the change in R^2 value compared to M1.

Multiple linear regression with a qualitative variable

Now we will learn to use linear regression when the IV is a quantitative variable. First, we will build a model with a dichotomous variable, then extend it to a more general categorical variable. (2+2+4+4+4+4+10+10 points)
a. Create a new variable called carrierAA using the mutate() function. This variable is TRUE if the carrier is AA, else it is FALSE.
b. Build a model M3 that predicts arr_delay based on this dichotomous variable carrierAA.
c. Explain what the regression slope and the regression constant for this model mean by writing out the regression equation in the predictive form. Note that TRUE is evaluated as 1, and FALSE as 0 by R in a numerical context. Also, find the predicted arr_delay for flights that are AA, and flights that are not AA, based on the regression equation and the regression coefficients calculated from R.
d. Now, build a model M4 <- lm(arr_delay ~ carrier, data = flights). Run the summary(M4) and comment on what R has done in terms of the regression coefficients.
e. Dummy variable coding: When a qualitative variable is included in a regression model, R creates a series of variables called dummy variables, each of which takes values TRUE or FALSE, similar to what we did with carrierAA. Notice that there are 16 unique values for the carrier variable in the dataset, and R creates 15 dummy variables for the regression. Explain how this set of 15 dummy variables allows for all 16 values in the original dataset to be represented. Hint: First find the one carrier value that is not represented in the dummy variable set. Then, notice the logic that when we want to represent the carrier YV, we could set carrierYV = 1 (TRUE), and all other dummy variables will have to be 0 (FALSE). Now, what happens when all the dummy variables are set to 0?
f. Write out a regression equation that corresponds to the output of the lm() function using the 15 dummy variables. Remember that these variables mathematically can only take values 0 or 1.
g. Explain what the regression constant and the different regression coefficients mean in terms of predicted arr_delay in this model. Notice the regression constant corresponds to setting all variables in the regression equation to 0, and hence corresponds to a particular carrier (called the reference or the baseline value). This reference or baseline can be chosen by the user but R automatically chooses a carrier to be the baseline in this case. Similary a unit change in a dummy variable corresponds to changing that variable from 0 to 1, which means changing from the reference carrier to the carrier in question. h. Tabulate the predicted delay for each carrier based on this model using the regression equation.

Multiple regression with both quantitative and qualitative variables

Now we bring both quantitative and qualitative variables into one big model. (2+16+2)
a. Build a model M5 in R that has all the variables in M2 as well as the carrier variable.
b. Explain the regression coefficients including the regression constant. In this case, the slopes due to the quantitative variables do not depend on the carrier. The effect of each carrier is to add a fixed predicted delay when all the quantitative variables are held constant.
c. Explain the R^2 value for M5, and explain how this model compares to models M1, M2, and M4.