HW2-Trinath Sai Subhash Reddy-Pittala

Subhash

2023-03-05

Simple linear regression

A. Write the regression equation in both forms (i.e., with and without error terms) as shown in class.

$$arrivaldelay = \beta_0 + \beta_1 * depdelay + \epsilon$$
$$\hat{arrivaldelay} = \hat{\beta_0} + \hat{\beta_1} * depdelay$$

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).

```
# Load the nycflights13 library
library(nycflights13)
flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      542        540       2     923     850      33 AA
##  4  2013     1     1      544        545      -1    1004    1022     -18 B6
##  5  2013     1     1      554        600      -6     812     837     -25 DL
##  6  2013     1     1      554        558      -4     740     728      12 UA
##  7  2013     1     1      555        600      -5     913     854      19 B6
##  8  2013     1     1      557        600      -3     709     723     -14 EV
##  9  2013     1     1      557        600      -3     838     846      -8 B6
## 10  2013     1     1      558        600      -2     753     745       8 AA
## # ... with 336,766 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

C. Run a simple linear regression model using the lm() function and save the model as M1.

```
M1 <- lm(arr_delay ~ dep_delay, data = flights)
```

D. Output the elements of M1 using the summary() function.

```
summary(M1)
```

```
##
## Call:
## lm(formula = arr_delay ~ dep_delay, data = flights)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -107.587  -11.005   -1.883    8.938  201.938
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.8994935  0.0330195  -178.7   <2e-16 ***
## dep_delay    1.0190929  0.0007864  1295.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.03 on 327344 degrees of freedom
##   (9430 observations deleted due to missingness)
## Multiple R-squared:  0.8369, Adjusted R-squared:  0.8369
## F-statistic: 1.679e+06 on 1 and 327344 DF,  p-value: < 2.2e-16
```

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.

The regression slope is 1.02 which denotes that the arrival delay increases by 1.02 hours for every 1 hour increase in departure delay. Also, since intercept is -5.89, if departure delay for a flight is 0 then the arrival delay is -5.89.

F. Explain the Pr(> |t|) values for the two coefficients and how they capture the uncertainty in the regression coefficient estimates.

Pr(> |t|) are p-values for probability of observing for T-statistic and deciding on Null hypothesis. Since for both slope and intercept the values are small (<2e-16) indicates that our estimates are significant hence eliminating null hypothesis and indicating stronger corelation between arrival and departure delays.

G. Explain what the R2 numerical value indicates for this particular model in one sentence.

R2 value is the amount of data which can fit the given distribution. Here 83.69% of data is fitting the distribution i.e, there is 83.69% of variability of Arrival delay is explained by the Departure delay.

Multiple linear regression

A. Write out the regression equations in both forms for this multiple linear regression model.

$$arrivaldelay = \beta_0 + \beta_1 * depdelay + \beta_2 * schedarrtime + \beta_3 * distance + \epsilon$$

$$\hat{arrivaldelay} = \hat{\beta_0} + \hat{\beta_1} * depdelay + \hat{\beta_2} * schedarrtime + \hat{\beta_3} * distance$$

B. Build the model in R using the lm() function and save this model as M2.

```r
M2 <- lm(arr_delay ~ dep_delay + sched_arr_time + distance, data = flights)
summary(M2)
```

```
##
## Call:
## lm(formula = arr_delay ~ dep_delay + sched_arr_time + distance,
##     data = flights)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -107.983  -11.023   -2.018    8.668  205.024
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.825e+00  1.082e-01  -16.86   <2e-16 ***
## dep_delay       1.020e+00  7.926e-04 1286.88   <2e-16 ***
## sched_arr_time -9.553e-04  6.393e-05  -14.94   <2e-16 ***
## distance       -2.501e-03  4.271e-05  -58.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.92 on 327342 degrees of freedom
##   (9430 observations deleted due to missingness)
## Multiple R-squared:  0.8387, Adjusted R-squared:  0.8387
## F-statistic: 5.675e+05 on 3 and 327342 DF,  p-value: < 2.2e-16
```

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.

Regression constant (β_0) is the intercept of the given linear regression equation i.e, value for which the other 3 IV's are zeroes.

β_1 is the regression coefficient for the Independent variable Departure Delay i.e, the differential of Departure delay vs arrival delay when other IV's are constant.

β_2 is the regression coefficient for the Independent variable Scheduled Arrived Time i.e, the differential of Scheduled Arrived Time vs arrival delay when other IV's are constant.

β_3 is the regression coefficient for the Independent variable Distance i.e, the differential of Distance vs arrival delay when other IV's are constant.

The β_1 is same in M1 and M2 because addition of Scheduled Arrived Time and Distance did not influence the relation between Arrival Delay and Departure delay in this particular case.

D. Comment on the uncertainty in each of the regression coefficient estimates based on the Pr($>|$t$|$) values in the R output.

Pr($>|$t$|$) are p-values for probability of observing for T-statistic and deciding on Null hypothesis. Since for both slope and intercept the values are small (<2e-16) indicates that our estimates are significant hence eliminating null hypothesis and indicating corelation between arrival and departure delays, Scheduled Arrived Time, Distance.

E. Comment on the change in R2 value compared to M1.

R2 value increased from 83.69 percent to 83.87 percent indicating that data is fitting the M2 better than M1.

Multiple linear regression with a qualitative variable

A. Create a new variable called carrierAA using the mutate() function. This variable is TRUE if the carrier is AA, else it is FALSE.

```
flights <- mutate(flights, carrierAA = ifelse(carrier == "AA",
    TRUE, FALSE))
```

B. Build a model M3 that predicts arr_delay based on this dichotomous variable carrierAA.

```
M3 <- lm(arr_delay ~ carrierAA, data = flights)
summary(M3)
```

```
##
## Call:
## lm(formula = arr_delay ~ carrierAA, data = flights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -93.6  -23.6  -11.6    7.4 1264.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.60170    0.08203   92.67   <2e-16 ***
## carrierAATRUE -7.23741    0.26257  -27.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.58 on 327344 degrees of freedom
##   (9430 observations deleted due to missingness)
## Multiple R-squared:  0.002316,   Adjusted R-squared:  0.002313
## F-statistic: 759.8 on 1 and 327344 DF,  p-value: < 2.2e-16
```

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.

$$\hat{arrival delay} = \hat{\beta}_0 + \hat{\beta}_1 * carrierAA$$

β_0 is regression Constant which describes the arrival delay when flight is not American Airlines and β_1 is Regression Coefficient for carrierAA which describes the differential of arrival delay when flight is American Airlines

Arrival Delay estimate for CarrierAA is +7.60170-7.23741(1) = 0.36429

Arrival Delay estimate for not CarrierAA is +7.60170-7.23741(0) = 7.60170

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.

```
M4 <- lm(arr_delay ~ carrier, data = flights)
summary(M4)
```

```
##
## Call:
## lm(formula = arr_delay ~ carrier, data = flights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -87.76  -23.64  -11.46    7.44 1278.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.3797     0.3368  21.908  < 2e-16 ***
## carrierAA    -7.0154     0.4182 -16.775  < 2e-16 ***
## carrierAS   -17.3106     1.6974 -10.198  < 2e-16 ***
## carrierB6     2.0783     0.3870   5.370 7.87e-08 ***
## carrierDL    -5.7353     0.3932 -14.585  < 2e-16 ***
## carrierEV     8.4168     0.3897  21.598  < 2e-16 ***
## carrierF9    14.5410     1.7306   8.402  < 2e-16 ***
## carrierFL    12.7362     0.8553  14.891  < 2e-16 ***
## carrierHA   -14.2949     2.4189  -5.910 3.43e-09 ***
## carrierMQ     3.3951     0.4380   7.751 9.12e-15 ***
## carrierOO     4.5514     8.2328   0.553     0.58
## carrierUA    -3.8217     0.3840  -9.953  < 2e-16 ***
## carrierUS    -5.2501     0.4609 -11.391  < 2e-16 ***
## carrierVX    -5.6152     0.7050  -7.965 1.66e-15 ***
## carrierWN     2.2695     0.5257   4.317 1.58e-05 ***
## carrierYV     8.1773     1.9289   4.239 2.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.3 on 327330 degrees of freedom
##   (9430 observations deleted due to missingness)
## Multiple R-squared:  0.01502,    Adjusted R-squared:  0.01498
## F-statistic: 332.9 on 15 and 327330 DF,  p-value: < 2.2e-16
```

R has made 15 dummy variables to represent 16 variables (all carriers). We can deduce the value of the 16th category from the values of the other 15 dummy variables. If all the 15 dummy variables are set to 0 (FALSE), then it implies that the leftover is 16th variable.

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?

In the dataset, there are 16 unique values for the carrier variable, but only 15 dummy variables are created for the regression. This is because we can deduce the value of the 16th category from the values of the other 15 dummy variables. If all the 15 dummy variables are set to 0 (FALSE), then it implies that the 16th category (which is not represented in the dummy variable set) is present.

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.

$$arr\ delay = 7.3797 + (-7.0154) * carrierAA + (-17.3106) * carrierAS + ... + (8.1773) * carrierYV$$

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.

The predicted arr_delay for each carrier can be calculated using the regression equation:

$$\hat{predicted arr\ delay} = \hat{\beta_0} + \hat{\beta_1} * carrierAA + \hat{\beta_2} * carrierAS + ... + \hat{\beta_{15}} * carrierYV$$

Where: β_0 is the intercept (7.3797) and 16th carrier (9E) which is the reference carrier β_1 to β_{15} are the coefficients for each of the 15 carriers (AA, AS, B6, DL, EV, F9, FL, HA, MQ, OO, UA, US, VX, WN, and YV) where each of them are either 0 or 1

H. Tabulate the predicted delay for each carrier based on this model using the regression equation.

```
# Create a data frame
res <- summary(M4)$coefficients
rownames(res)[1] = c("carrier9E")
res[-1, 1] = res[-1, 1] + res[1, 1]
res[-1, 2] = res[-1, 2] + res[1, 2]
res = res[, 1:2]
colnames(res)[1] = c("Arrival_delay")
colnames(res)[2] = c("Standard_Error")

kable(res)
```

	Arrival_delay	Standard_Error
carrier9E	7.3796692	0.3368478
carrierAA	0.3642909	0.7550460
carrierAS	-9.9308886	2.0342434
carrierB6	9.4579733	0.7238520
carrierDL	1.6443409	0.7300918
carrierEV	15.7964311	0.7265418
carrierF9	21.9207048	2.0674396
carrierFL	20.1159055	1.1921315
carrierHA	-6.9152047	2.7557636
carrierMQ	10.7747334	0.7748456
carrierOO	11.9310345	8.5696239
carrierUA	3.5580111	0.7208096
carrierUS	2.1295951	0.7977351
carrierVX	1.7644644	1.0418483
carrierWN	9.6491199	0.8625789
carrierYV	15.5569853	2.2657375

Multiple regression with both quantitative and qualitative variables

A. Build a model M5 in R that has all the variables in M2 as well as the carrier variable.

```
M5 <- lm(arr_delay ~ dep_delay + sched_arr_time + distance +
    carrierAA, data = flights)
summary(M5)
```

```
##
## Call:
## lm(formula = arr_delay ~ dep_delay + sched_arr_time + distance +
##     carrierAA, data = flights)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.054  -11.023   -2.014    8.653  204.683
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.741e+00  1.083e-01  -16.08   <2e-16 ***
## dep_delay       1.020e+00  7.926e-04 1286.40   <2e-16 ***
## sched_arr_time -9.529e-04  6.390e-05  -14.91   <2e-16 ***
## distance       -2.398e-03  4.306e-05  -55.70   <2e-16 ***
## carrierAATRUE  -1.939e+00  1.065e-01  -18.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.91 on 327341 degrees of freedom
##   (9430 observations deleted due to missingness)
## Multiple R-squared:  0.8389, Adjusted R-squared:  0.8389
## F-statistic: 4.262e+05 on 4 and 327341 DF,  p-value: < 2.2e-16
```

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.

$$arrivaldelay = \beta_0 + \beta_1 * depdelay + \beta_2 * schedarrtime + \beta_3 * distance + \beta_4 * carrier + \epsilon$$

$$\hat{arrivaldelay} = \hat{\beta_0} + \hat{\beta_1} * depdelay + \hat{\beta_2} * schedarrtime + \hat{\beta_3} * distance + \hat{\beta_4} * carrier$$

Regression coefficients (β_i) quantifies the amount of correlation (the differential) between the Arrival delay with Departure Delay, Scheduled Arrival Time, Distance, and Carrier.

Regression Constant (β_0) is the Baseline Arrival Delay given all the other quantitative variables are constant and equal to zero.

C. Explain the R2 value for M5, and explain how this model compares to models M1, M2, and M4.

R2 for M5 is 0.8389. Its highest of M1, M2 and M4 suggesting a better fit to a Linear regression model.