HW6

Roberto Campos

3/20/2020

library(readxl)  
VACATION <- read\_excel("/Users/robertocampos/Desktop/Main\_Folder/CSUS\_classes/stat103/VACATION.xlsx")  
  
amus <- read\_excel("/Users/robertocampos/Desktop/Main\_Folder/CSUS\_classes/stat103/AMUSEMENT.xlsx")  
  
cata = read\_excel('/Users/robertocampos/Desktop/Main\_Folder/CSUS\_classes/stat103/CATALOG.xlsx')

1. A developer of vacation homes is considering purchasing a tract of land near a lake. From previous experience, she knows that the price of a lot is affected by the lot size, number of mature trees, and distance to the lake. From a nearby area, she gathers data on 60 recently sold lots. These data are stored in the file VACATION. Use the data to construct a regression model to predict the value of a lot.
2. Do the model assumptions appear to be satisfied? If not, which ones are violated?

* -model: price = 51.39 + .699 \* lot size + .6788 \* Trees - .378 \* Distance
* -The P Value is .0013 which means that it supports the assumption there is a linear relation for the variables mentioned above, but looking at the summary we can see that the independent variable ‘lot size’ could be removed.

1. What is R2? What does it tell you?

* -R^2 in this model is .2425 which means that atleast .25 of the observed prices are explained by the model. 0% indicates that the model explains none of the variability of the response data around its mean, we want a higher R^2, because it would typically indicate that the model fit’s our data better.

1. Which of the explanatory variables is linearly related to the response variable in this (the original)model?

* -If we take a look at the summary, the independent variable Trees is the only variable that is linearly related to our dependent variable.

1. If necessary, create a new model by removing insignificant variables.

-After removing the independent variable, we get the following model: Price = 75.5248 + .7671 \* Trees .4327 \* Distance. The model gives a better P Value at .0008 and both independent variables being Trees and Distance are linearly related with the dependent variable at a significance level.

1. Interpret the slopes in the new model.

-Each tree will add about .767 or 767 dollars to the average price and each unit of distance (feet) for the lake the average price will decrease by .432 or 432 dollars.

1. Predict with 95% confidence the selling price of a 40,000-square foot lot with 50 mature trees that is located 75 feet from the lake.

- With 95% confidence the lot can sell up to 157k

1. Estimate with 95% confidence the average selling price of all such lots.
   * With 95% confidence the mean price of such lots are 62K to 89k

model = lm(formula = VACATION$Price~VACATION$`Lot size`+VACATION$Trees+VACATION$Distance, data = VACATION)  
  
model2 = lm(formula = VACATION$Price~VACATION$Trees+VACATION$Distance, data = VACATION)  
  
summary(model)

##   
## Call:  
## lm(formula = VACATION$Price ~ VACATION$`Lot size` + VACATION$Trees +   
## VACATION$Distance, data = VACATION)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -66.702 -35.272 0.365 28.854 84.966   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 51.3912 23.5165 2.185 0.0331 \*   
## VACATION$`Lot size` 0.6999 0.5589 1.252 0.2156   
## VACATION$Trees 0.6788 0.2293 2.960 0.0045 \*\*  
## VACATION$Distance -0.3784 0.1952 -1.938 0.0577 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 40.24 on 56 degrees of freedom  
## Multiple R-squared: 0.2425, Adjusted R-squared: 0.2019   
## F-statistic: 5.975 on 3 and 56 DF, p-value: 0.001315

summary(model2)

##   
## Call:  
## lm(formula = VACATION$Price ~ VACATION$Trees + VACATION$Distance,   
## data = VACATION)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -73.600 -33.159 -4.829 33.828 97.281   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 75.5248 13.5464 5.575 7.06e-07 \*\*\*  
## VACATION$Trees 0.7671 0.2193 3.498 0.000917 \*\*\*  
## VACATION$Distance -0.4327 0.1913 -2.262 0.027549 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 40.44 on 57 degrees of freedom  
## Multiple R-squared: 0.2213, Adjusted R-squared: 0.1939   
## F-statistic: 8.097 on 2 and 57 DF, p-value: 0.0008031

1. The manager of an amusement park would like to be able to predict daily attendance. After some consideration, he decided that the following three factors are critical, yesterday’s attendance, whether it’s a weekday or weekend, and the predicted weather. He then took a random sample of 40 days and recorded the data in the file AMUSEMENT. Since two of the variables are qualitative, he created the following sets if dummy variables:
2. Construct a regression model to predict attendance. Is the model likely to be useful? Include all relevant computer output, organized so that I can follow it.

* -The model P Value is 5.841e-11 which means that it supports the assumption that the model is useful. Although the independent variable rain does not seem to have linear relationship with the dependent variable Attendance as we can see in the output summary.

1. Can we conclude that weather is a factor in determining attendance?

* -We cannot conclude but we can approximate that weather gives a good indication of attendance. The independent variable Rain had a bigger P Value than expected but Sunny had a good P Value which means there is a linear relationship and we can use it to predict attendance.

1. Determine the best model using the Akaike Information Criterion (AIC) and Mallow’s Cp statistics. How does this affect your choice of final model? Does it change your answer to part (b)?

* -According to AIC, I should choose my first model: amus\_model using the three independent variables.
* -According to Mallows cp, we should choose the first model: amus\_model using the three independent variables.
* -AIC and Mallows test should be the about the same.

1. Does this data provide sufficient evidence that weekend attendance is, on average, larger than weekday attendance? Support your answer.

-Yes, since the independent variable weekend is significant we can assume that on average there is a larger weekend attendance. Significant meaning there is a linear relationship between the dependent variable Attendance and the independent variable weekend.

amus\_model = lm(amus$Attendance ~ amus$Yesterday + amus$Weekend + amus$Sunny + amus$Rain)  
summary(amus\_model)

##   
## Call:  
## lm(formula = amus$Attendance ~ amus$Yesterday + amus$Weekend +   
## amus$Sunny + amus$Rain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1181.11 -494.98 41.44 487.71 1725.30   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4514.44912 570.67150 7.911 2.66e-09 \*\*\*  
## amus$Yesterday 0.20612 0.08472 2.433 0.02023 \*   
## amus$Weekend 933.90177 311.61848 2.997 0.00499 \*\*   
## amus$Sunny 1074.83737 322.76468 3.330 0.00205 \*\*   
## amus$Rain -727.09560 365.57648 -1.989 0.05457 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 699.6 on 35 degrees of freedom  
## Multiple R-squared: 0.7768, Adjusted R-squared: 0.7513   
## F-statistic: 30.45 on 4 and 35 DF, p-value: 5.841e-11

amus\_model2 = lm(amus$Attendance ~ amus$Yesterday + amus$Weekend + amus$Sunny)  
summary(amus\_model2)

##   
## Call:  
## lm(formula = amus$Attendance ~ amus$Yesterday + amus$Weekend +   
## amus$Sunny)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1268.0 -389.3 186.8 338.6 2021.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.731e+03 4.296e+02 8.686 2.33e-10 \*\*\*  
## amus$Yesterday 2.953e-01 7.479e-02 3.948 0.000351 \*\*\*  
## amus$Weekend 9.771e+02 3.234e+02 3.022 0.004609 \*\*   
## amus$Sunny 1.234e+03 3.253e+02 3.794 0.000547 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 727.7 on 36 degrees of freedom  
## Multiple R-squared: 0.7516, Adjusted R-squared: 0.7309   
## F-statistic: 36.3 on 3 and 36 DF, p-value: 5.551e-11

\*AIC

AIC(amus\_model, amus\_model2)

## df AIC  
## amus\_model 6 644.2090  
## amus\_model2 5 646.4921

\*Marllows Cp

## [1] "Model 1 vs model 2"

## [1] 2.344408

## [1] "Model 2 vs model 1"

## [1] 6.955729

3.The general manager of a chain of catalog stores wanted to determine the factors that affect how long it takes to unload a truck delivering orders. A random sample of 50 deliveries to a store was observed. The times (in minutes) to unload the truck, the total number of boxes, and the total weight (in hundreds of pounds) were recorded in the file CATALOG.

1. Determine the multiple regression equation.

- Time = -28.427 + .604 \* boxes + .374 \* weight

1. How well does the model fit the data?

\_ According to the R^2 adjusted the model fits well. It means that .80 of the time, the dependent variable is explained by the independent variables.

1. Perform diagnostics on the model and report your findings.

- The P Value of the model according to the summary is 2.2e-16 which means that there is a linear relationship between the dependent and independent variables. Both the independent variables seem to have a liner relationship and are significant to the model in this case. We can keep both of then in this case.

1. Is multicollinearity a problem? If so, propose a solution.

-There seem's to be no multicollinearity. We can do a check by using a Variance Inflation Factor test. The test showed that there is no multicollinearity. It say's that the is no correlation between one predictor and the other predictor variables and that the varience was not inflated.

1. Construct a regression model that includes the information for the time of day.

- Time = -41.4221 + .644 \* boxes + .3494 \* weight + 4.54 \* codes

1. Does the time of day affect the unloading time? Explain.

- Yes, the time of day seems to affect the loading time. It's there is a linear relationship between the independet variable time of day (which is encoded 1, 2 or 3). It appears that in the morning and late afternoon it takes less time.

cata\_model = lm(cata$Time ~ cata$Boxes + cata$Weight, data = cata)  
summary(cata\_model)

##   
## Call:  
## lm(formula = cata$Time ~ cata$Boxes + cata$Weight, data = cata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.0809 -3.9519 0.5004 3.6927 17.1160   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -28.42712 6.88696 -4.128 0.000149 \*\*\*  
## cata$Boxes 0.60411 0.05568 10.850 2.16e-14 \*\*\*  
## cata$Weight 0.37430 0.08467 4.420 5.78e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.069 on 47 degrees of freedom  
## Multiple R-squared: 0.8072, Adjusted R-squared: 0.799   
## F-statistic: 98.37 on 2 and 47 DF, p-value: < 2.2e-16

## [1] "variance inflation factor"

## cata$Boxes cata$Weight   
## 1.146707 1.146707

## [1] "Average time it takes to unload in the morning"

## [1] 55.77778

## [1] "Average time it takes to unload in the early after"

## [1] 68.66667

## [1] "Average time it takes to unload in the late after"

## [1] 55.90909