6414 RMarkdown

James Will Trawick

11/15/2019

Process:

1. Write Down Task
2. Insert R Script and write code
3. Document Code

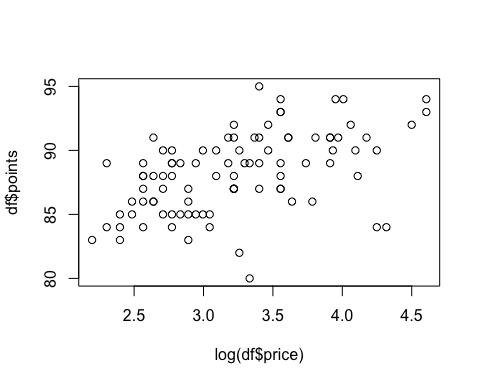
Example:

Task 1: Fit Simple Linear Model

#Set Working Directory  
#setwd("C:/Users/james/Desktop")  
  
#Read Data  
filename = "testwine.csv"  
OriginalData = read.csv(filename, header = TRUE)  
df = as.data.frame(OriginalData)  
  
#fit Linear Model   
model1 = lm(points ~ log(price), data = df)  
summary(model1)

##   
## Call:  
## lm(formula = points ~ log(price), data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.4931 -1.7207 0.1875 1.7565 6.3212   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 79.5209 1.5878 50.08 < 2e-16 \*\*\*  
## log(price) 2.6926 0.4834 5.57 2.57e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.701 on 91 degrees of freedom  
## (7 observations deleted due to missingness)  
## Multiple R-squared: 0.2542, Adjusted R-squared: 0.246   
## F-statistic: 31.02 on 1 and 91 DF, p-value: 2.568e-07

attach(df)  
plot(log(df$price), df$points)



Rich - pulling last couple of columns and will send data with 100ish rows with all columns not subsetted

–impute/delete –subset based on atleast 40 observations per

columns = c("country")

Tasks

Data Cleaning —— Subset data frame by columns

MLR HW TASKS

Question 1: Exploratory Data Analysis [6 points] (a) Create a boxplot of the qualitative predictor and the response. Does there appear to be a relationship between the predictor and the response? Interpret the boxplot.

1. Create a scatterplot matrix containing the quantitative predictors and the response, and display the correlations between each pair of variables. Interpret the correlations.
2. Is it reasonable to use a multiple linear regression model to predict Rings using all the predictor variables?

Question 2: Fitting the Multiple Linear Regression Model Build a linear regression model, model1, to predict Rings using all the predictors. Display the model1 summary.

deletemissing = function(df, columns){  
 return(df[complete.cases(df[ , columns]),])  
}  
  
#provide dataframe  
#provide columns to check for missing data  
#returns dataframe minus the rows where NA existed in columns provided

#Build a full model first  
  
allColumns = c("country", "description", "designation", "points", "price", "province", "region\_1", "title", "variety", "winery", "year", "location", "avg\_temp", "range\_temp", "stdv\_temp", "avg\_precip", "range\_precip", "stdv\_precip", "lat", "lng", "elevation", "review\_length", "abv", "bottle\_size category", "importer")  
  
  
#All columns except description, designation, and location. Log(price) and log(points)  
#modelFull = lm(log(points) ~ country + province + log(price) + region\_1 + variety + winery + year + avg\_temp + range\_temp + range\_temp + stdv\_temp + avg\_precip + range\_precip + stdv\_precip + lat + lng + elevation + abv + bottle\_size + category + importer, data = df)  
  
#Columns I think are meaninful (for discussion as a group)  
#modelFull = lm(log(points) ~ country + province + log(price) + region\_1 + year + avg\_precip + elevation + abv + bottle\_size + category, data=df)  
  
#As a tester...log(price) gives NA in confint output  
modelFull = lm(log(points) ~ price + country, data = df)  
  
summary(modelFull)

##   
## Call:  
## lm(formula = log(points) ~ price + country, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.09307 -0.01364 0.00000 0.01980 0.06509   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.4246280 0.0146015 303.025 < 2e-16 \*\*\*  
## price 0.0007112 0.0001872 3.800 0.000285 \*\*\*  
## countryAustralia 0.0371219 0.0234855 1.581 0.118010   
## countryAustria 0.0496047 0.0216106 2.295 0.024399 \*   
## countryBulgaria 0.0519177 0.0350825 1.480 0.142935   
## countryChile 0.0042826 0.0268050 0.160 0.873475   
## countryFrance 0.0428219 0.0169829 2.521 0.013725 \*   
## countryGermany 0.0284782 0.0350848 0.812 0.419437   
## countryGreece 0.0246978 0.0267935 0.922 0.359485   
## countryItaly 0.0386839 0.0203307 1.903 0.060766 .   
## countryNew Zealand 0.0319858 0.0268086 1.193 0.236440   
## countryPortugal 0.0440037 0.0193952 2.269 0.026045 \*   
## countrySouth Africa 0.0599164 0.0353174 1.697 0.093776 .   
## countrySpain 0.0320634 0.0202963 1.580 0.118207   
## countryUnited States of America 0.0305563 0.0156138 1.957 0.053924 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.03202 on 78 degrees of freedom  
## (7 observations deleted due to missingness)  
## Multiple R-squared: 0.3043, Adjusted R-squared: 0.1794   
## F-statistic: 2.437 on 14 and 78 DF, p-value: 0.006768

1. What is the coefficient estimate for Height? Interpret this coefficient.

See summary output

1. What is the coefficient estimate for the Sex category Infant? Interpret this coefficient.

See summary output

1. What is the coefficient estimate for the Sex category Male? Interpret this coefficient

See summary output

1. Give a 90% confidence interval for Length. Interpret this interval.

conf90 = function(model, columnName) {  
 return(confint(model, columnName, level=.90))  
}  
  
# provide a model  
# provide column for which you want an interval  
# returns 90% confint for that variable  
  
#conf90(modelFull, "price")  
log(confint(modelFull, "price", level=.90))

## 5 % 95 %  
## price -7.824836 -6.885253

Question 3: Checking Model Assumptions [20 points] (a) Create a scatterplot of the standardized residuals versus fitted values. What conclusions can you draw from this plot?

library(car)

## Loading required package: carData

resids = rstandard(modelFull)  
fits = modelFull$fitted  
cook = cooks.distance(modelFull)  
  
#pairs.panels(df, hist.col="#00AFBB", ellipses=FALSE, lm=TRUE)

1. Create a histogram and normal QQ plot for the residuals. What conclusions can you draw from these plots?
2. Create a plot for the Cook’s Distances. What conclusions can you draw from this plot?
3. Display the VIF of each predictor. What conclusions can you draw?

Question 4: Model Comparison and Prediction [20 points]

Create a second model, model2, that uses all the predictors except for: Whole weight, Viscera weight, and Length. Display the model summary.

1. Compare and discuss the R-squared and Adjusted R-squared of model2 with model1.
2. Conduct a partial F-test comparing model2 with model1. What can you conclude?
3. Predict Rings for the last 11 rows of data (i.e. dataTest) using both model2 and model1. Compare and discuss the point prediction and intervals of both models. Compare and discuss the mean squared prediction error and the precision measure.
4. Suppose you’ve found an adult female abalone with a diameter of 0.4mm, a height of 0.2mm, a shucked weight of 0.3 grams, and a shell weight of 0.3 grams. Using model2, predict the number of rings on this abalone with a 90% prediction interval. What is the age of this abalone?