ISE 537 Assignment 1 Smith Keenan

“Built with R Version 4.1.2”

# Section 1

1. The estimated simple linear regression coefficient 𝛽0, measures the strength of a linear relationship between the predicting and response variables.

* False

1. In simple linear regression, a negative value of 𝛽1 is consistent with an inverse relationship between the predicting variable and the response variable.

* True

1. In simple linear regression, 𝛽0 is an unbiased estimator for 𝛽0.

* True

1. If the assumptions of a simple linear regression model hold, then the estimator for the variance of the error terms, 𝜎2 is a random variable.

* False

1. Under the normality assumption, the estimated simple linear regression coefficient, 𝛽1, is a linear combination of normally distributed random variables.

* True

1. The p-value is a measure of the probability of rejecting the null hypothesis.

* False

1. In simple linear regression, the prediction interval of one member of the population will always be wider than the confidence interval of the mean response for all members of the population when using the same predicting value.

* True

1. In simple linear regression, the normality assumption states that the response variable is normally distributed.

* True

# Section 2

### Package Initiation Block

#Initiate Tidyverse since all relevant packages are located within  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.7  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

### Data Import and Function Definition

Function for Regression Model source <https://sejohnston.com/2012/08/09/a-quick-and-easy-function-to-plot-lm-results-in-r/>

# Importing machine.csv in a dplyr dataframe  
machine <- read\_csv('machine.csv', show\_col\_types = FALSE)  
# Inspecting Dataframe to ensure good import  
head(machine)

## # A tibble: 6 x 3  
## vendor chmax performance  
## <chr> <dbl> <dbl>  
## 1 adviser 128 198  
## 2 amdahl 32 269  
## 3 amdahl 32 220  
## 4 amdahl 32 172  
## 5 amdahl 16 132  
## 6 amdahl 32 318

# ggplot Regression Function for a Nicer Looking Plot  
ggplotRegression <- function (fit) {  
  
require(ggplot2)  
  
ggplot(fit$model, aes\_string(x = names(fit$model)[2], y = names(fit$model)[1])) +   
 geom\_point() +  
 stat\_smooth(method = "lm", col = "red") +  
 labs(title = paste("Adj R2 = ",signif(summary(fit)$adj.r.squared, 5),  
 "Intercept =",signif(fit$coef[[1]],5 ),  
 " Slope =",signif(fit$coef[[2]], 5),  
 " P =",signif(summary(fit)$coef[2,4], 5)))  
}

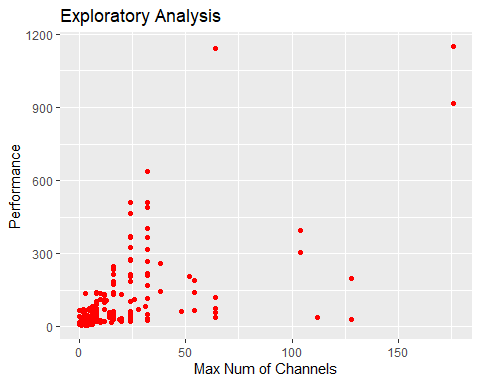
## 2.1 Exploratory Data Analysis

1. Use a scatter plot to describe the relationship between CPU performance and the maximum number of channels. Describe the general trend (direction and form). Include plots and R-code used. Exploratory Data Plotting

summary(machine)

## vendor chmax performance   
## Length:209 Min. : 0.00 Min. : 6.0   
## Class :character 1st Qu.: 5.00 1st Qu.: 27.0   
## Mode :character Median : 8.00 Median : 50.0   
## Mean : 18.27 Mean : 105.6   
## 3rd Qu.: 24.00 3rd Qu.: 113.0   
## Max. :176.00 Max. :1150.0

max\_performance <- ggplot(data = machine, aes(x = chmax, y = performance)) +  
 geom\_point(color = 'red') +  
 labs(x = 'Max Num of Channels', y = 'Performance',  
 title = 'Exploratory Analysis')  
  
max\_performance



1. What is the value of the correlation coefficient between performance and chmax? Please interpret the strength of the correlation based on the correlation coefficient. Correlation Coefficient Calculation

chmax\_vec <- machine %>%  
 pull(var = 2)  
  
performance\_vec <- machine %>%  
 pull(var = 3)  
  
correlation\_ch\_perf <- cor(x = chmax\_vec, y = performance\_vec)  
correlation\_ch\_perf

## [1] 0.6052093

1. Based on this exploratory analysis, would you recommend a simple linear regression model for the relationship?

* Based on the correlation coefficient of ~ 0.61, their is a moderate to moderate-strong positive association between the number of channels and performance. A linear model should be explored to analyze further values.

## Section 2.2 Linear Model

Fitting a Linear Model to the Data

performance\_lm <- lm(performance ~ chmax, data = machine)  
  
summary(performance\_lm)

##   
## Call:  
## lm(formula = performance ~ chmax, data = machine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -486.47 -42.20 -22.20 20.31 867.15   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.2252 10.8587 3.428 0.000733 \*\*\*  
## chmax 3.7441 0.3423 10.938 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 128.3 on 207 degrees of freedom  
## Multiple R-squared: 0.3663, Adjusted R-squared: 0.3632   
## F-statistic: 119.6 on 1 and 207 DF, p-value: < 2.2e-16

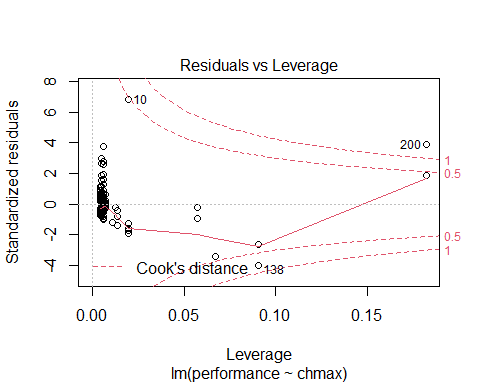
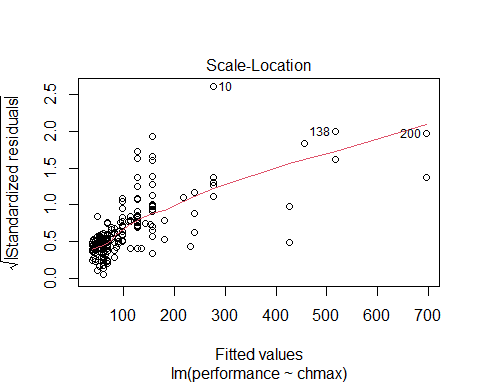
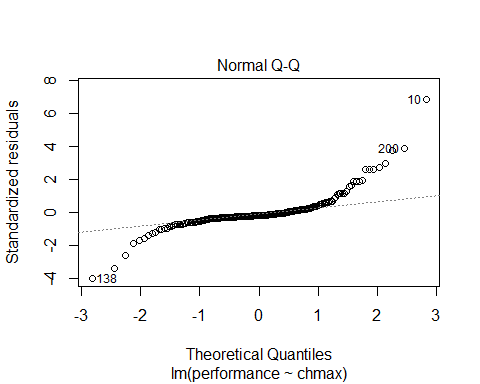
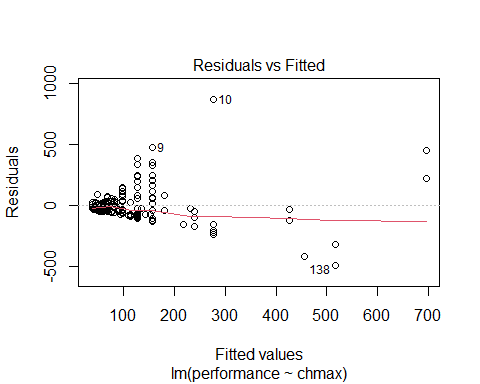
beta1\_95 <- confint(performance\_lm, level = 0.95)  
beta1\_99 <- confint(performance\_lm, level = 0.99)  
  
beta1\_95

## 2.5 % 97.5 %  
## (Intercept) 15.817392 58.633048  
## chmax 3.069251 4.418926

beta1\_99

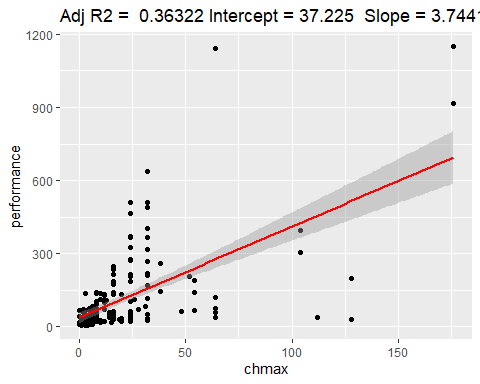
## 0.5 % 99.5 %  
## (Intercept) 8.994891 65.455549  
## chmax 2.854185 4.633991

plot(performance\_lm)



ggplotRegression(performance\_lm)

## `geom\_smooth()` using formula 'y ~ x'



1. What are the model parameters and what are their estimates?

* 𝛽1 = 3.7441 𝛽0 = 37.2235 and Sigma\_hat = 128.3 on 207 df

1. Write down the estimated simple linear regression equation.
2. Interpret the estimated value of the 𝛽1 parameter in the context of the problem.

* According to the model, for each additional 1 max channel in a processor, there is a performance increase of an average of ~3.74 within the processor.

1. Find a 95% confidence interval for the 𝛽1 parameter. Is 𝛽1 statistically significant at this level?

* (3.069251, 4.418926) The p-value is less than 2e-16 which is less than alpha .05

1. Is 𝛽1 statistically significantly positive at an α-level of 0.01? What is the approximate p-value of this test?

* The p-value is less than 2e-16. The model is statistically significant.