Install.packages (“psych”)

Library(psych)

attach(GBS_MBA)

Simple Statistics Explanation

summary()

describe()

descrieBy(GBS_MBA2, group = GBS_MBA$Gender)

Explanation:

mean + – 2sd == 95% of confidence. Is this range too high?

range: max- min

Interquartile Range: 3rd quartile – 1st quartile where the majority of values lie, can be used to find outliers. Better measure of spread than the range as it is not affected by outliers.

1st Quartile : the median of the *lower half* of the data set.

3rd quartile: the median of the *upper half* of the data set

Histogram

1. Formula: Hist(file$weight, nclass=20)
    a. Nclass means to create a number of specified bins
2. Prediction: left or right skewed? Is it normally distributed?

Scatterplot

plot(GBS_MBA$Height, GBS_MBA$Weight)

abline(lm(GBS_MBA$Weight ~ GBS_MBA$Height), col=”red”)

abline(lm(data$Y~ data$X), col=”red”)

pairs(Seawatch_C_Data[,c(3:20)])

Regression Model Equation

#Linear Model:

```
fitlinear <- lm(Weight ~ Height)
summary(fitlinear)
```

#Prediction :

```
predict.lm(fitlinear, newdata=data.frame(Height=70), interval="prediction", level=0.95)
 # Predict
 SeaWatchC_clean$predict <- predict(model_SW, newdata=SeaWatchC_clean)
 # Error
 SeaWatchC_clean$error <- resid(model_SW)
 View(SeaWatchC_clean)
 summary(SeaWatchC_clean$error)
 # Residual
 SeaWatchC_clean$resid <- SeaWatchC_clean$GROSS-SeaWatchC_clean$predict
 View(SeaWatchC_clean)
 # Correlation between Residual and COLLPR
 SeaWatchC_clean$COLLPR <- as.numeric(SeaWatchC_clean$COLLPR)
 cor(SeaWatchC_clean$resid, SeaWatchC_clean$COLLPR, use="complete.obs")
 # Look at the correlation again
 cor(SeaWatchC_clean[,c(3,4,5:15,18,19,22)])
 # We found no correlation between error and other variables
 # Add more models to get lower standard errors
 model_SW_2 <- lm(GROSS ~ CNVHRS + CART + REAG)
 summary(model_SW_2)  # RSME = 920
```

#Which significant variables are missing?

```
SeaWatch_C_data$resid <- resid(modelF)

SeaWatch_C_data$resid <- SeaWatch_C_data$predict - SeaWatch_C_data$GROSS

smaller<- SeaWatch_C_data[, c(29, 3, 5:20)]

View(smaller)

cor(smaller, use="complete.obs")
```

Analysis:

Residuals: also known as errors, used when assessing the quality of a model

Residual Standard Error: The smaller the better the model is.

R Squared: how much variance is explained by this model, the higher the better the model is

Adjusted R-squared: In a multiple regression, each additional independent variable may increase the R-squared without improving the actual fit. An adjusted R-squared is calculated that represents the more accurate fit with multiple independent variables. It is always lower than R-squared

Multiple R-squared – Adjusted R-squared = how accurate the model is

p value: <0.05 Significant, there's 1-p value percent chance that it is statistically significant with the dependent variable

t value: >|2| Significant, means it represents at least 95% of data to show Normal Distribution

Significance codes: indicate how certain we can be that the coefficient has an impact on the dependent variable ex. 0.001 means we can be 99.9% sure that it is significant

The coefficients(slopes) of x variables: the range between the value the model gives you +/- the sd of that variable in the model

Multicollinearity:

install.packages ("regclass")

library(regclass)

VIF(model_Best)