# Module 6 Assignment 2

## Graybill, Gil

### Statistical Analysis

library(tidyverse)

## -- Attaching packages ------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

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

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(ggplot2)  
library(readxl)  
Perceptions <- read\_excel("Perceptions.xlsx")  
RespiratoryExchangeSample <- read\_excel("RespiratoryExchangeSample.xlsx")  
library(readr)  
Insurance <- read\_csv("Insurance.csv")

## Parsed with column specification:  
## cols(  
## age = col\_double(),  
## sex = col\_character(),  
## bmi = col\_double(),  
## children = col\_double(),  
## smoker = col\_character(),  
## region = col\_character(),  
## charges = col\_double()  
## )

Advertising <- read\_csv("Advertising.csv")

## Parsed with column specification:  
## cols(  
## ID = col\_double(),  
## Rating = col\_double(),  
## Group = col\_double()  
## )

#### Regression and Correlation

Regression analysis is a statistical method that allows you to examine the relationship between two or more variables of interest. Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables (e.g. height and weight). This particular type of analysis is useful when a researcher wants to establish if there are possible connections between variables.

#### Insurance Costs

We would like to determine if we can accurately predict insurance costs based upon the factors included in the data. We would also like to know if there are any connections between variables (for example, is age connected or correlated to charges).

#### Correlations of bmi, age, children and cost

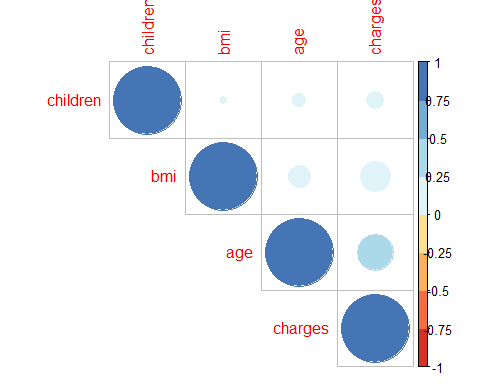
library(corrplot)

## corrplot 0.84 loaded

library(RColorBrewer)  
Insurance2 <- Insurance %>%  
 select(bmi, age, children, charges)  
cor(Insurance2)

## bmi age children charges  
## bmi 1.0000000 0.1092719 0.01275890 0.19834097  
## age 0.1092719 1.0000000 0.04246900 0.29900819  
## children 0.0127589 0.0424690 1.00000000 0.06799823  
## charges 0.1983410 0.2990082 0.06799823 1.00000000

corr\_matrix <- cor(Insurance2)  
corrplot(corr\_matrix, type="upper", order="hclust",  
 col=brewer.pal(n=8, name="RdYlBu"))



None of the values in the correlation matrix are above 0.3. The threshold for strong correlation is 0.7, so it doesn’t look like there is a strong correlation between any of these variables.

#### Regression Analysis

fit3 <- lm(charges ~ bmi + age + children, data = Insurance2)  
summary(fit3)

##   
## Call:  
## lm(formula = charges ~ bmi + age + children, data = Insurance2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13884 -6994 -5092 7125 48627   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6916.24 1757.48 -3.935 8.74e-05 \*\*\*  
## bmi 332.08 51.31 6.472 1.35e-10 \*\*\*  
## age 239.99 22.29 10.767 < 2e-16 \*\*\*  
## children 542.86 258.24 2.102 0.0357 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11370 on 1334 degrees of freedom  
## Multiple R-squared: 0.1201, Adjusted R-squared: 0.1181   
## F-statistic: 60.69 on 3 and 1334 DF, p-value: < 2.2e-16

The variables “bmi” and “age” are significant. We can determine that from looking at the P-value and seeing that it is less than 0.001. Age has the smallest P-value, so that would make it the most significant variable. “Children” has a p-value of 0.036, which is getting close to the threshold of 0.05 that is usually used to eliminate a variable from regression analysis.

Insurance <- mutate(Insurance, gender=ifelse(sex=="female",1,0))  
Insurance <- mutate(Insurance, smoker2=ifelse(smoker=="yes",1,0))  
fit3 <- lm(charges ~ bmi + age + children + gender + smoker2, data = Insurance)  
summary(fit3)

##   
## Call:  
## lm(formula = charges ~ bmi + age + children + gender + smoker2,   
## data = Insurance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11837.2 -2916.7 -994.2 1375.3 29565.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12181.10 963.90 -12.637 < 2e-16 \*\*\*  
## bmi 322.36 27.42 11.757 < 2e-16 \*\*\*  
## age 257.73 11.90 21.651 < 2e-16 \*\*\*  
## children 474.41 137.86 3.441 0.000597 \*\*\*  
## gender 128.64 333.36 0.386 0.699641   
## smoker2 23823.39 412.52 57.750 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6070 on 1332 degrees of freedom  
## Multiple R-squared: 0.7497, Adjusted R-squared: 0.7488   
## F-statistic: 798 on 5 and 1332 DF, p-value: < 2.2e-16

It looks like all variables contribute to charges except for gender, which has a p-value of 0.7, making it almost irrelevant. A change from the previous test shows that children is now farther away from elimination. Further analysis would be needed to figure out what caused that.

#### Group Comparisons with t-tests

A study of the effect of caffeine on muscle metabolism used volunteers who each underwent arm exercise tests. Half the participants were randomly selected to take a capsule containing pure caffeine one hour before the test. The other participants received a placebo capsule. During each exercise the subject’s respiratory exchange ratio (RER) was measured. (RER is the ratio of CO2 produced to O2 consumed and is an indicator of whether energy is being obtained from carbohydrates or fats).

summary(RespiratoryExchangeSample)

## Placebo Caffeine   
## Min. : 80.00 Min. :100.0   
## 1st Qu.: 85.00 1st Qu.:106.0   
## Median : 90.00 Median :110.5   
## Mean : 90.11 Mean :110.8   
## 3rd Qu.: 95.25 3rd Qu.:117.0   
## Max. :100.00 Max. :120.0

t.test(RespiratoryExchangeSample$Placebo, RespiratoryExchangeSample$Caffeine)

##   
## Welch Two Sample t-test  
##   
## data: RespiratoryExchangeSample$Placebo and RespiratoryExchangeSample$Caffeine  
## t = -33.742, df = 397.67, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -21.95369 -19.53631  
## sample estimates:  
## mean of x mean of y   
## 90.105 110.850

Performing a t-test on the two samples show that there is a difference between the Caffeine and Placebo groups. The p-value is practically 0, indicating that there is a difference between the two groups. Additionally, the 95% confidence interval is (-21.9, -19.5), indicating that there is a 95% chance that the difference in the means falls into that range.

#### Impact of Advertising

You are a marketing researcher conducting a study to understand the impact of a new marketing campaign. To test the new advertisements, you conduct a study to understand how consumers will respond based on see the new ad compared to the previous campaign. One group will see the new ad and one group will see the older ads. They will then rate the ad on a scale of 0 to 100 as a percentage of purchase likelihood based on the ad.

The question you are trying to answer is whether to roll out the new campaign or stick with the current campaign.

summary(Advertising)

## ID Rating Group   
## Min. : 1.0 Min. : 0.00 Min. :1.000   
## 1st Qu.: 250.8 1st Qu.: 25.75 1st Qu.:1.000   
## Median : 500.5 Median : 53.00 Median :1.000   
## Mean : 500.5 Mean : 51.06 Mean :1.499   
## 3rd Qu.: 750.2 3rd Qu.: 76.00 3rd Qu.:2.000   
## Max. :1000.0 Max. :100.00 Max. :2.000   
## NA's :184

t.test(Rating ~ Group, Advertising, var.equal=TRUE)

##   
## Two Sample t-test  
##   
## data: Rating by Group  
## t = 1.2509, df = 814, p-value = 0.2113  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.440198 6.501170  
## sample estimates:  
## mean in group 1 mean in group 2   
## 52.33827 49.80779

The P-value is 0.2113, which is not less than 0.05. That means there’s a a 21% chance that this data is wrong. So even though Group 1 (new ad) has a slightly higher mean rating than Group 2 (old ad), the difference may be due to random sampling. From this data, it is not recommended to roll out the new ad campaign.

#### ANOVA

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis. Basically, you’re testing groups to see if there’s a difference between them. Examples of when you might want to test different groups:

* A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
* A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
* Students from different colleges take the same exam. You want to see if one college outperforms the other.

#### Perceptions of Social Media Profiles

This study examines how certain information presented on a social media site might influence perceptions of trust, connectedness and knowledge of the profile owner. Specifically, participants were shown weak, average and strong arguments that would influence their perceptions of the above variables. Using the dataset provided, the following code runs an ANOVA with post-hoc analyses to understand argument strength impacts on perceptions.

aov1 <- aov(Trust ~ Argument, data=Perceptions)  
summary(aov1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Argument 2 26.59 13.293 16.34 2.4e-07 \*\*\*  
## Residuals 221 179.75 0.813   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(aov1)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Trust ~ Argument, data = Perceptions)  
##   
## $Argument  
## diff lwr upr p adj  
## strong-average -0.03333333 -0.3808438 0.3141771 0.9721584  
## weak-average -0.74855856 -1.0972410 -0.3998761 0.0000026  
## weak-strong -0.71522523 -1.0639077 -0.3665427 0.0000073

aov2 <- aov(Connectedness ~ Argument, data=Perceptions)  
summary(aov2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Argument 2 29.7 14.859 9.869 7.85e-05 \*\*\*  
## Residuals 221 332.7 1.506   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(aov2)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Connectedness ~ Argument, data = Perceptions)  
##   
## $Argument  
## diff lwr upr p adj  
## strong-average -0.2733333 -0.7461312 0.1994645 0.3615643  
## weak-average -0.8736637 -1.3480561 -0.3992712 0.0000628  
## weak-strong -0.6003303 -1.0747228 -0.1259378 0.0087959

aov3 <- aov(Knowledge ~ Argument, data=Perceptions)  
summary(aov3)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Argument 2 0.47 0.2333 0.315 0.73  
## Residuals 221 163.67 0.7406

Based on the small P-values, Trust and Connectedness look to be significantly affected by the argument. Knowledge does not seem to be affected by the strength of argument.

Group analysis of the AOV shows that Trust and Connectedness are affected by the difference between a weak argument and any other type of argument. The big takeaway: Weak arguments affect Trust and Connectedness when it comes to social media.