# Module 6 - Assignment 2

## Pierce, Michaela

### Statistical Analysis

# Load libraries  
library(readxl)  
library(readr)  
  
# Load datasets  
Respiratory <- read\_excel("RespiratoryExchangeSample.xlsx")  
Advertising <- read\_csv("Advertising.csv")

## Rows: 1000 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (3): ID, Rating, Group  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

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

## Rows: 1338 Columns: 7  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): sex, smoker, region  
## dbl (4): age, bmi, children, charges  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Perceptions <- read\_excel("Perceptions.xlsx")

# 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(dplyr)

##   
## Attaching package: 'dplyr'

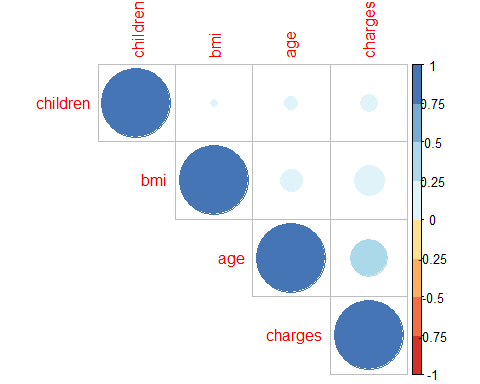
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(corrplot)

## corrplot 0.92 loaded

library(RColorBrewer)   
  
# Create new dataset insurance 2 with age, bmi, children, and charges  
insurance2 <- insurance %>%   
 select(age, bmi, children, charges)  
  
# Calculating correlation matrix  
Corr\_matrix <- cor(insurance2)  
  
# Visualizing the correlation matrix  
corrplot(Corr\_matrix, type="upper", order="hclust",   
col=brewer.pal(n=8, name="RdYlBu"))



**Based on the matrix and visuals, explain the results from your correlation matrix in a paragraph after the chunk of code. Are any of the variables highly correlated?**  
Age and Charges are highly correlated. BMI and Charges are next to being highly correlated and then BMI and Age.

#Regression Analysis

# Regression Analysis  
  
# Perform multiple linear regression  
lm\_model <- lm(charges ~ age + bmi + children, data = insurance2)  
  
# Display coefficients  
coefficients(lm\_model)

## (Intercept) age bmi children   
## -6916.2433 239.9945 332.0834 542.8647

# Display summary of the regression model  
summary(lm\_model)

##   
## Call:  
## lm(formula = charges ~ age + bmi + 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 \*\*\*  
## age 239.99 22.29 10.767 < 2e-16 \*\*\*  
## bmi 332.08 51.31 6.472 1.35e-10 \*\*\*  
## 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

**Based on the results, which variables were significant and what particular significant variable had the largest impact on charges? Provide a summary after the chunk of code.**

Based on the results, age has the largest impact on charges, with a coefficient estimate of approximately $240. This shows age plays a crucial part in predicting medical insurance charges. In addition to age, BMI and the number of children covered by health insurance are also significant predictors.

# Create new variable for gender  
insurance <- mutate(insurance, gender = ifelse(sex == "female", 1, 0))  
  
# Create new variable for smoker  
insurance <- mutate(insurance, smoker2 = ifelse(smoker == "yes", 1, 0))  
  
# Running regression with additional variables  
lm\_results <- lm(charges ~ age + bmi + children + gender + smoker2, data = insurance)  
summary(lm\_results)

##   
## Call:  
## lm(formula = charges ~ age + bmi + 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 \*\*\*  
## age 257.73 11.90 21.651 < 2e-16 \*\*\*  
## bmi 322.36 27.42 11.757 < 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

**After the code, provide an explanation of the new results. Does gender and smoking have an impact on cost?**

Smoking has a significant impact on medical insurance charges, with smokers incurring much higher costs compared to non-smokers. Gender does not appear to have that much impact on cost.

# Group Comparisons with t-tests

The t-test is used to compare the values of the means from two samples and test whether it is likely that the samples are from populations having different mean values. This is often used to compare 2 groups to see if there are any significant differences between these groups.

# Caffeine Impacts on Respiratory Exchange Ratio

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).

library(dplyr)  
  
# Summary statistics  
summary(Respiratory)

## 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

# Separate data into caffeine and placebo groups  
caffeine\_group <- Respiratory$Caffeine  
placebo\_group <- Respiratory$Placebo  
  
# Perform Welch t-test  
t\_test\_result <- t.test(caffeine\_group, placebo\_group)  
t\_test\_result

##   
## Welch Two Sample t-test  
##   
## data: caffeine\_group and placebo\_group  
## 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:  
## 19.53631 21.95369  
## sample estimates:  
## mean of x mean of y   
## 110.850 90.105

**Does caffeine affect RER during exercise?**

The Welch t-test showed a highly significant difference (p-value < 2.2e-16) in mean respiratory exchange ratio (RER) between the caffeine and placebo groups. Participants who consumed caffeine had an average RER of 110.85, while those on placebo averaged 90.11. This suggests caffeine intake affects RER during exercise, indicating potential impacts on muscle metabolism.

# Impact if 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.

colnames(Advertising)

## [1] "ID" "Rating" "Group"

# Summary statistics  
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

# Separate data into new ad and old ad groups  
new\_ad\_group <- Advertising$Rating[Advertising$Group == 1]  
old\_ad\_group <- Advertising$Rating[Advertising$Group == 2]  
  
# Perform Student's t-test  
t\_test\_result <- t.test(new\_ad\_group, old\_ad\_group, var.equal = TRUE)  
print(t\_test\_result)

##   
## Two Sample t-test  
##   
## data: new\_ad\_group and old\_ad\_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 of x mean of y   
## 52.33827 49.80779

**Should they roll out the new campaign or stick with the current campaign?**

The p-value of 0.2113 is greater than the typical significance level of 0.05, indicating no significant difference in purchase likelihood between the new and old ad groups. As a result, it’s unclear whether the new campaign effectively increases purchase likelihood compared to the old one. Further analysis is needed for a definitive conclusion.

# AVOVA

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.

# ANOVA 1: Trust across Argument  
anova\_trust <- aov(Trust ~ Argument, data = Perceptions)  
summary(anova\_trust)

## 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

# ANOVA 2: Connectedness across Argument  
anova\_connectedness <- aov(Connectedness ~ Argument, data = Perceptions)  
summary(anova\_connectedness)

## 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

# ANOVA 3: Knowledge across Argument  
anova\_knowledge <- aov(Knowledge ~ Argument, data = Perceptions)  
summary(anova\_knowledge)

## 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

Two ANOVAs are significant: ANOVA 1 (Trust across Argument) and ANOVA 2 (Connectedness across Argument) with p-values of 2.4e-07 and 7.85e-05, respectively. These indicate notable differences in Trust and Connectedness perceptions among Argument levels. ANOVA 3 (Knowledge across Argument) is not significant (p = 0.73), suggesting insignificant differences in Knowledge perception across Argument levels.

#Using the two significant ANOVAS (1 & 2)  
  
tukey\_trust <- TukeyHSD(anova\_trust)  
tukey\_connectedness <- TukeyHSD(anova\_connectedness)  
  
print(tukey\_trust)

## 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

print(tukey\_connectedness)

## 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

The post-hoc analysis shows significant differences in Trust perception between weak-average and weak-strong arguments, as well as between weak-average and strong-average arguments. However, there is no significant difference in Trust perception between strong and average arguments (p adj = 0.97).

Regarding Connectedness, significant differences exist among weak-average, weak-strong, and strong-average arguments, while the difference between strong and average arguments is not significant (p adj = 0.36).