# Module 6 - Assignment 2

## Angelo, Elizabeth

### Statistical Analyses

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(corrplot)

## corrplot 0.92 loaded

library(readxl)  
library(readr)  
Perceptions <- read\_excel("Perceptions.xlsx")  
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.

### 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

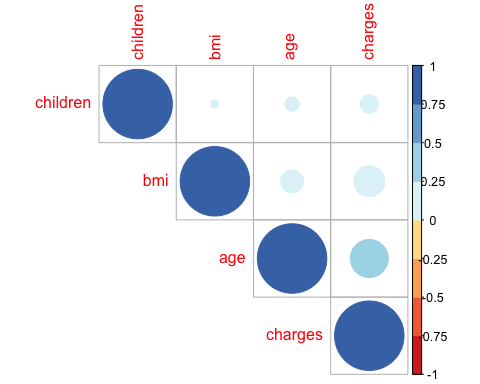
We would like to determine if we can accurately predict insurance costs based upon the factors included in the data. We would also like to know if there are any connections between variables (for example, is age connected or correlated to charges).

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

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

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

corrplot(Corr\_matrix, type = "upper", order = "hclust", col = brewer.pal(n=8, name = "RdYlBu"))



Based on the matrix and correlation plot, none of the variables appear to be highly correlated with the other variables. Age and Charges show a moderate positive correlation, while all other variables show a weak positive correlation with each other.

### Regression Analysis

lm\_model <- lm(charges ~ age + bmi + children, data = Insurance2)  
coefficients(lm\_model)

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

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

When analyzing the regression model, all predictors were statistically significant, with age and bmi having high statistical significance at the p < 0.001 level, and children having moderate statistical significance at the p < 0.05 level. Among the predictors, age and bmi have a large impact on charges, while children do not. The model explains a relatively small proportion of variance in insurance charges, indicated by the Multiple R-Squared 0.1201. The overall model is statistically significant and supported by the F-statistic and p-value < 2.2e-16.

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

## (Intercept) age bmi children gender smoker2   
## -12181.1018 257.7350 322.3642 474.4111 128.6399 23823.3925

summary(lm\_model\_updated)

##   
## 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

When analyzing the updated regression model, age, bmi, children, and smoker2 are statistically significant at the p < 0.001 level. Gender is the only predictor that did not hold statistical significance at any level, and can be assumed to have no impact on insurance costs. Smoking on the other hand did have a high level of significance, and shows a high significant impact on insurance costs. The overall model is statistically significant with a p-value of < 2.2e-16, and explains a substantial proportion of variance in insurance charges with a Multiple R-squared value of 0.7497.

### 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).  
  
The question you are trying to answer is whether caffeine impacts RER during exercise.

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

caffeine\_data <- Respiratory$Caffeine  
placebo\_data <- Respiratory$Placebo  
t\_test\_result <- t.test(caffeine\_data, placebo\_data)  
print(t\_test\_result)

##   
## Welch Two Sample t-test  
##   
## data: caffeine\_data and placebo\_data  
## 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

The p-value < 2.2e-16 for the Welch Two Sample t-test is extremely small, suggesting a significant difference between the means of the two groups. The p-value provides supporting evidence to reject the null hypothesis, while the t-value 33.742 suggests a substantial difference between the means of the two groups. This is supported by the 95% confidence interval, indicating precision of the estimated difference in means.

### 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.

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

group1\_data <- subset(Advertising, Group == 1)  
group2\_data <- subset(Advertising, Group == 2)  
t\_test\_result <- t.test(group1\_data, group2\_data)  
print(t\_test\_result)

##   
## Welch Two Sample t-test  
##   
## data: group1\_data and group2\_data  
## t = 0.30245, df = 2813.9, p-value = 0.7623  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -17.93378 24.47516  
## sample estimates:  
## mean of x mean of y   
## 194.7001 191.4294

The results of the Welch Two Sample t-test do not provide strong evidence to reject the null. The p-value 0.7623 is high, indicating a lack of statistical significance between the [p < 0.001, p < 0.05] range. The t-value 0.30245 suggest a small difference in means between the two groups, with the 95% confidence interval suggesting the true difference in means could be zero. Based on these results, there is not enough evidence to conclude that the new advertising campaign has a significant impact, and further investigation may be needed before deciding to move forward.

### 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.

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\_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\_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

tukey\_trust <- TukeyHSD(anova\_trust)  
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

tukey\_connectedness <- TukeyHSD(anova\_connectedness)  
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

Using the two statistically significant anova models Trust and Connectedness, we examine the p adj within the table to see which groups are different. We can see that the difference for both Trust and Connectedness, for strong and average arguments are not significant with a p adj 0.9721584 and p adj 0.3615643. On the other hand, weak and average arguments do show strongly significant differences for both Trust and Connectedness, and are supported with the p adj values 0.0000026 and 0.0000628. As for weak and strong arguments, statistically significant differences were found for both Trust and Connectedness. While the significance was stronger for Trust with a p adj 0.0000073, the p adj for Connectedness was still significant at 0.0087959.