Module 6 Assignment 2: Statistical Analyses

Tabitha Hagen

2022-08-14

# Module 6 - Assignment 2

## Hagen, Tabitha

### Statistical Analyses

## Part 1: Open libraries and Importing the dataset using readr

knitr::opts\_chunk$set(echo = TRUE)  
library(tidyverse)  
  
library(readr)  
Advertising <- read\_csv("Advertising.csv")  
Insurance <- read\_csv("Insurance.csv")  
  
library(readxl)  
RespiratoryExchangeSample <- read\_excel("RespiratoryExchangeSample.xlsx")  
Perceptions <- read\_excel("Perceptions.xlsx")

## Part 2:

### 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

#### Correlation Matrix

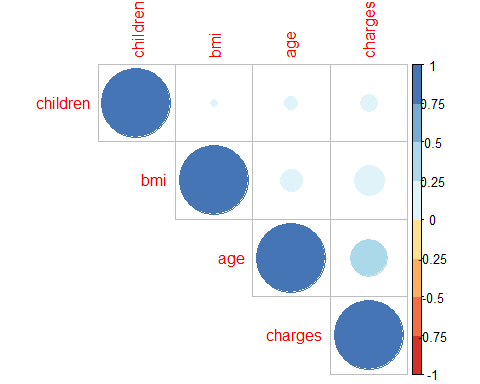
Insurance2<- Insurance %>% # piping reduces redundant commands  
 select(age, bmi, children, charges) # filter by Page, bmi, children & charges  
  
cor(Insurance2) # look at a correlation of all the variables in Insurance2

## 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

Corr\_matrix<- cor(Insurance2)  
  
library(corrplot) # Open corrplot package

## corrplot 0.92 loaded

library(RColorBrewer) # Open RColorBrewer package  
  
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?

Well, According to the class video on stats, above a 0.7 is considered good and highly correlated. In the graphic chart just drawn, it shows that the darker the blue, the more correlated the variables are. So, the category “charges” and “age” have the darkest circle (besides the big dark blue ones that show the variables correlate to themselves). But, that is actually the color that refers to 0.5 to 0.75, so whether it is highly correlated may not meet the must of being 0.7 or above. Everything else is definately below that standard and not showing correlation.

#### Regression Analysis

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

##   
## 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

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. According to the Multiple Linear Regression chart, both bmi and age have a direct correlation to insurance charges as they are both getting their Pr(>|t|) values closer to zero. This makes sense because they are more likely to have health conditions being monitored by physicians. However, the number of children having health insurance isn’t related to the charges. In other words, having a high bmi and being older can predict that there will be more insurance charges.

#### Regression Analysis

Insurance <- mutate(Insurance, gender=ifelse(sex=="male",0,1))   
# change sex (or gender) to be 1 or 0  
Insurance <- mutate(Insurance, smoker=ifelse(smoker=="yes",1,0))   
# change smoker to be 1 or 0  
  
Insurance2$gender <- Insurance$gender # add newly changed gender column to Insurance2  
Insurance2$smoker <- Insurance$smoker # add newly changed smoker column to Insurance2  
  
fit <- lm(charges ~ bmi + age + children + gender + smoker, data=Insurance2)  
summary(fit)

##   
## Call:  
## lm(formula = charges ~ bmi + age + children + gender + smoker,   
## data = Insurance2)  
##   
## 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   
## smoker 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

Provide an explanation of the new results. Does gender and smoking have an impact on cost?

According to the Multiple Linear Regression Model comparing Insurance charges to bmi, age, children, gender and sex, everything except gender is correlated to charges. Gender has the largest Pr(>|t|) value of 0.699641 which is significantly away from zero. It makes sense that if one smokes, they would have more insurance charges. Nowadays, even their premiums cost more because they have more co-morbidity issues.

## Part 3:

### 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 to answer is whether caffeine impacts RER during exercise.

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

# default is welch t-test

Does caffeine impacts RER during exercise?

According to the T-Test (Welch) on the Respiratory Exchange Sample, the answer is yes. Looking at the p-value < 2.2e-16, it is less than 0.05 and statistically significant. This makes sense as caffeine makes the heart rate increase and your lungs may not be as efficient under the stress.

#### 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 to answer is whether to roll out the new campaign or stick with the current campaign.

summary(Perceptions)

## Argument Processing Gender Age   
## Length:224 Min. :1.0 Min. :1.000 Min. :18.00   
## Class :character 1st Qu.:1.0 1st Qu.:1.000 1st Qu.:19.00   
## Mode :character Median :1.5 Median :1.000 Median :20.00   
## Mean :1.5 Mean :1.362 Mean :20.56   
## 3rd Qu.:2.0 3rd Qu.:2.000 3rd Qu.:21.00   
## Max. :2.0 Max. :2.000 Max. :42.00   
## SSC1 SSC2 SSC3 SSC4   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:4.000 1st Qu.:5.000 1st Qu.:4.000 1st Qu.:4.000   
## Median :6.000 Median :6.000 Median :6.000 Median :5.000   
## Mean :5.205 Mean :5.281 Mean :5.201 Mean :5.094   
## 3rd Qu.:6.000 3rd Qu.:6.000 3rd Qu.:6.000 3rd Qu.:6.000   
## Max. :7.000 Max. :7.000 Max. :7.000 Max. :7.000   
## SSC6 SSC7 Connectedness RTrust1   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:4.000 1st Qu.:4.000 1st Qu.:4.333 1st Qu.:4.000   
## Median :5.000 Median :5.000 Median :5.333 Median :5.000   
## Mean :5.058 Mean :4.973 Mean :5.135 Mean :4.964   
## 3rd Qu.:6.000 3rd Qu.:6.000 3rd Qu.:6.000 3rd Qu.:6.000   
## Max. :7.000 Max. :7.000 Max. :7.000 Max. :7.000   
## RTrust2 RTrust3 RTrust4R RTrust5   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:4.000 1st Qu.:4.000 1st Qu.:3.000 1st Qu.:3.000   
## Median :5.000 Median :5.000 Median :4.000 Median :4.000   
## Mean :4.902 Mean :5.054 Mean :4.018 Mean :4.286   
## 3rd Qu.:6.000 3rd Qu.:6.000 3rd Qu.:5.000 3rd Qu.:5.000   
## Max. :7.000 Max. :7.000 Max. :7.000 Max. :7.000   
## RTrust7 Trust CCC3 CCC4   
## Min. :1.000 Min. :2.167 Min. :1.000 Min. :1.000   
## 1st Qu.:4.000 1st Qu.:4.000 1st Qu.:5.000 1st Qu.:5.000   
## Median :5.000 Median :4.833 Median :6.000 Median :6.000   
## Mean :4.866 Mean :4.682 Mean :5.589 Mean :5.402   
## 3rd Qu.:6.000 3rd Qu.:5.333 3rd Qu.:6.000 3rd Qu.:6.000   
## Max. :7.000 Max. :7.000 Max. :7.000 Max. :7.000   
## CCC5 CCC6 CCC7 Knowledge   
## Min. :2.000 Min. :2.00 Min. :2.000 Min. :2.000   
## 1st Qu.:5.000 1st Qu.:5.00 1st Qu.:5.000 1st Qu.:5.000   
## Median :6.000 Median :6.00 Median :6.000 Median :5.600   
## Mean :5.491 Mean :5.58 Mean :5.384 Mean :5.489   
## 3rd Qu.:6.000 3rd Qu.:6.00 3rd Qu.:6.000 3rd Qu.:6.000   
## Max. :7.000 Max. :7.00 Max. :7.000 Max. :7.000

t.test(Perceptions$Knowledge,Perceptions$Processing, var.equal=TRUE)

##   
## Two Sample t-test  
##   
## data: Perceptions$Knowledge and Perceptions$Processing  
## t = 60.094, df = 446, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 3.858820 4.119751  
## sample estimates:  
## mean of x mean of y   
## 5.489286 1.500000

After not knowing which two variables to compare, I finally figured out the the old and new ads were contained in the Processing column, which were either a 1 or 2. The first column, Argument was of character value so it could not be compared without converting it into 3 numeric values. So, I went with the Knowledge column, though that left out the many other factors like trust, whatever the SSC’s were, and the CCC’s. Anyhow, the result was a p-value < 2.2e-16. This was not close to 0.5, so not significant. All those other columns might have been more useful, but I don’t know what those variables stood for. So, I say, NO, the add campaign should freeze and redefine the parameters of this study for the statistician to know what they are looking at.

## Part 4:

### Create a 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.

#the difference of Trust across Argument   
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

#Pr(>F) = 2.4e-07 \*\*\* close to zero

#the difference of Connectedness across Argument   
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

#Pr(>F) = 7.85e-05 \*\*\* close to zero

#the difference of Knowledge across Argument   
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

#Pr(>F) = 0.73 NOT close to zero

You should see that two of these ANOVAs are significant. However, we know the groups are different but we don’t know which specific group is different from another. For example, we know there is a difference between weak, average and strong but we don’t know the difference of weak to average, weak to strong and average to strong.

#### TurkeyHSD

TukeyHSD(aov1) #the difference of Trust across Argument

## 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

# p adj for strong-average was 0.9721584 which is NOT significant (should be 0.05 or less)  
# p adj for weak-strong was 0.0000026 which is significant (should be 0.05 or less)  
# p adj for weak-strong was 0.0000073 which is significant (should be 0.05 or less)  
  
TukeyHSD(aov2) #the difference of Connectedness across Argument

## 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

# p adj for strong-average was 0.3615643 which is NOT significant (should be 0.05 or less)  
# p adj for weak-strong was 0.0000628 which is NOT significant (should be 0.05 or less)  
# p adj for weak-strong was 0.0087959 which is NOT significant (should be 0.05 or less)  
  
TukeyHSD(aov3) #the difference of Knowledge across Argument

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Knowledge ~ Argument, data = Perceptions)  
##   
## $Argument  
## diff lwr upr p adj  
## strong-average 0.002666667 -0.3289304 0.3342637 0.9998015  
## weak-average -0.095675676 -0.4283911 0.2370398 0.7762378  
## weak-strong -0.098342342 -0.4310578 0.2343731 0.7652284

# p adj for strong-average was 0.9998015 which is NOT significant (should be 0.05 or less)  
# p adj for weak-strong was 0.7762378 which is NOT significant (should be 0.05 or less)  
# p adj for weak-strong was 0.7652284 which is NOT significant (should be 0.05 or less)

Using the TukeyHSD method, both Trust and Connectedness across Argument were significant. However, Knowledge was not significant across argument. This explains why I did not say go ahead with the ad campaign early on. The comparison I chose was not significant to each other. Maybe I should have chose trust or Connectedness.