Exploratory Project

Jillian Douglasss

4/4/2019

Code libraries used:

knitr::opts\_chunk$set(echo = TRUE)
library(ggplot2)
library(dplyr)
depression <-read.delim("/Users/jilliandouglass/Desktop/math130/data/depress\_081217.txt", header=TRUE, sep='\t')

## Introduction

The data set comes from the first set of interviews of a study on depression in adults living in Los Angeles County. The data set consists of 294 observations and 37 variables. Out of those 37 variables, I will be looking at cesd levels and health to see if there is any relationship between the two. Cesd is a variable that measures depression levels on a scale of 1-60, 1 being the lowest score. The other variable I will analyze is general health, ranked on a scale of 1-4, 1 being excellent health and 4 being poor health. I hypothesize that general health will be an influencing factor in higher cesd levels.

## Univariate Analysis

Summary Statistics Variable 1: Health Shown below is a summary table of the health variable. From this table, we can see the number of respondents in each health category from the study in LA County. From the table, we can see that 130 people with excellent health suffered from depression while only 14 people with poor health did.

depression$health\_fac <- depression$health
depression$health\_fac <- factor(depression$health, labels=c("Excellent(1)", "Good(2)", "Fair(3)", "Poor(4)"))
summary(depression$health\_fac)

## Excellent(1) Good(2) Fair(3) Poor(4)
## 130 115 35 14

Shown directly below is a graph illustrating the summary statistics from the table above.

ggplot(depression, aes(x=health\_fac, fill="depression")) + geom\_bar(color="black")



Summary Statistics Variable 2: CESD levels In the table below,is the summary statistics of the CESD varibale. We can see that the average cesd level of those who participated in the study is 8.884 and the highest score that was recieved is 47.000.

depression$cesd\_fac <- depression$cesd
summary(depression$cesd\_fac)

## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 0.000 3.000 7.000 8.884 12.000 47.000

Below is a histogram showing the distribution of CESD levels among the respondants and, as we can see, the histogram is skewed right.This graph illustrates that majority of patients scored below a CESD level of 20.

ggplot(depression, aes(x=cesd\_fac)) + geom\_histogram(fill='purple', color='black')

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## Bivariate Analysis

ggplot(depression, aes(x=health\_fac, y=cesd\_fac, fill='color')) + geom\_boxplot(color='black') + theme(legend.position = "none")



This boxplot demonstrates the comparison between the health and CESD levels of the patients. Patients with excellent and good health scored lower averages of CESD levels than patients with fair or poor health, but there are a few exceptions shown as outliers on the plot.

ggplot(depression,aes(x=cesd\_fac, fill=health\_fac)) + geom\_density() + facet\_wrap(~health\_fac)



When we look at these two variables as density distributions, we can see that for each health category, CESD scores are skewed right. These graphs further illustrates that the majority of the patients scored a CESD level of below 20. Patients’ with fair or poor health CESD levels peaked closer to a score of 10, while those with excellent or good health CESD levels peaked at scores of less than 10.

## Conclusion

After comparing health and depression levels, it’s clear that health is an influencing factor on depression levels. The boxplot illustrastes that patients observed in this study with excellent or good health scored lower CESD levels. I can conclude that health is an influencing factor on depression levels and people with better health are less likely to be depressed. The data shows that health is an influencing factor on depression, but there are other factor that contribute as well.