Week_6

Kalyani Cauwenberghs

2/14/2020

Part 1

Group 7:

Nice galaxy theme background

Somewhat confusing topic, so far I don't understand it too well

What are the transposed table's column names? The table before table 1.

Your group tells us what you all did to the columns however we are not really sure what each column means.. In addition, what does the value of each column mean in context?

I think your group did a good job trying to explain the context of stars and celestial bodies, however I think the topic is still a bit confusing with such limited background, even if the data and analysis is sufficient

Experimental design: Your parameters should not include estimates. Just name the parameters (p1, p2) and make clear that the numbers are estimates (p1 hat, p2 hat)

CI for risk ratio should contain 1 for the difference to be insignificant

Explanation for chi square distribution was really long and unnecessary. Although it was probably a good introduction for those who had background info.

It is good that you explained between Fisher and Chi-Squared

Group 8:

Good start by explaining what each column signifies and represents.. Also good job for telling us which columns you dropped and explaining the reasons for why you dropped those respective columns

The frequency of sighting plot where there are a lot of colored bars is a bit hard to look at and compare between the bars. The colors have good contrast but the plot is too busy and it is hard for the eyes to navigate through the plot

It would be a good idea to plot relative proportions rather than number of observations for your bar plots since pre-2000 always has less observations than post-2000.

Good idea to combine the "other" data as a category along with the top 5.

Why does your H0 have an approx. equal sign? It should be equal bc you are talking about parameters, which are set in stone Why use a t-test? This is for comparing 2 means. We want to compare 2 proportions, which is the homework assignment.

It is good that you explained Fisher and Chi-squared and explained why you could reject the null hypothesis for both case, however, make sure you understand the variables, the IID parameters

Write your formulas using the equation editor. I can't read them

Perhaps consider using smaller portions of datasets, take random samples of large datasets as runtime error or memory errors may occur

Part 2

Research and write about the use of regression models in the context of

a.) prediction,

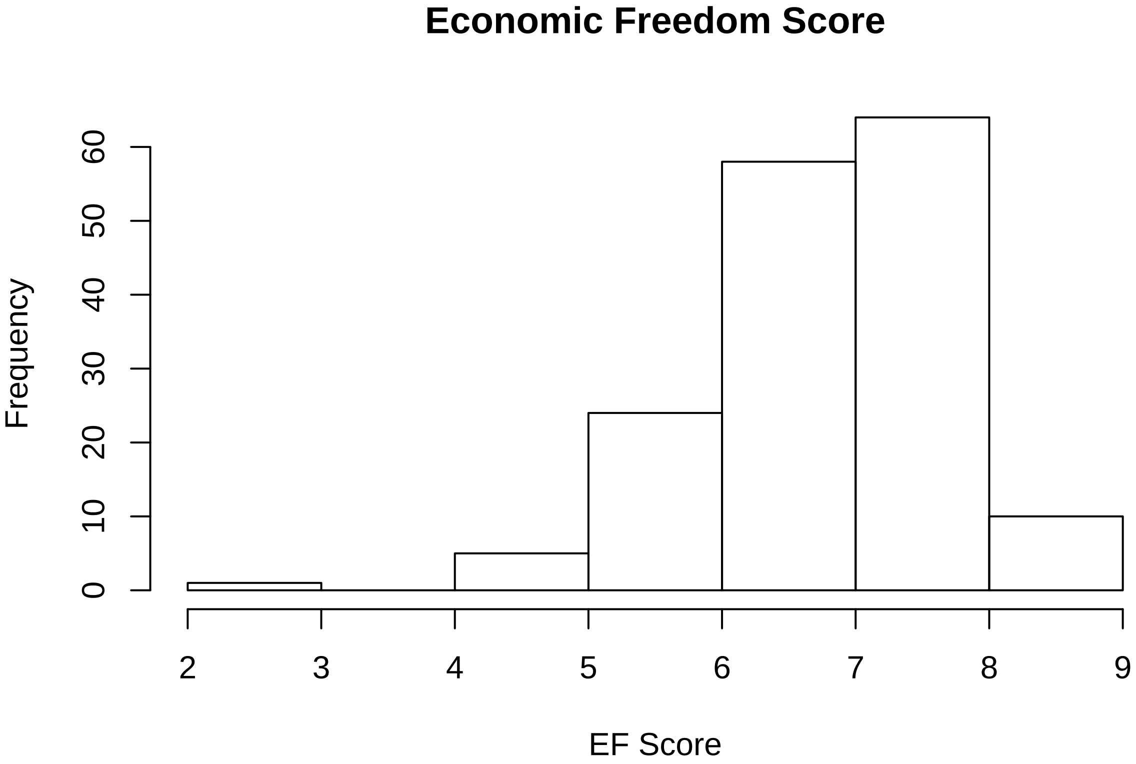
- In regards to prediction, regression models help to relate comparisons between units.
- When applying regression models for prediction, the goal is to develop a formula for making predictions about the dependent variable, based on the observed values of the independent variable
- It seems that prediction is the primary use of regression models outside of academia and research. This is especially true in the wake of big data and private industries using predictive analysis to help guide their decision making.
- Having a large R2 is more important for prediction as maximizing this value is crucial for prediction. It is important to have a large R2 for causal inference too, but it is not as crucial compared to prediction purposes.
- Multicollinearity is not as big of a concern for prediction because in prediction, we do not care about the individual coefficients so the multicollinearity can be tolerated more. Measurement errors affect prediction in biased ways as they affect the estimates of regression coefficients.

b.) causal inference on effect of a variable on the outcome.

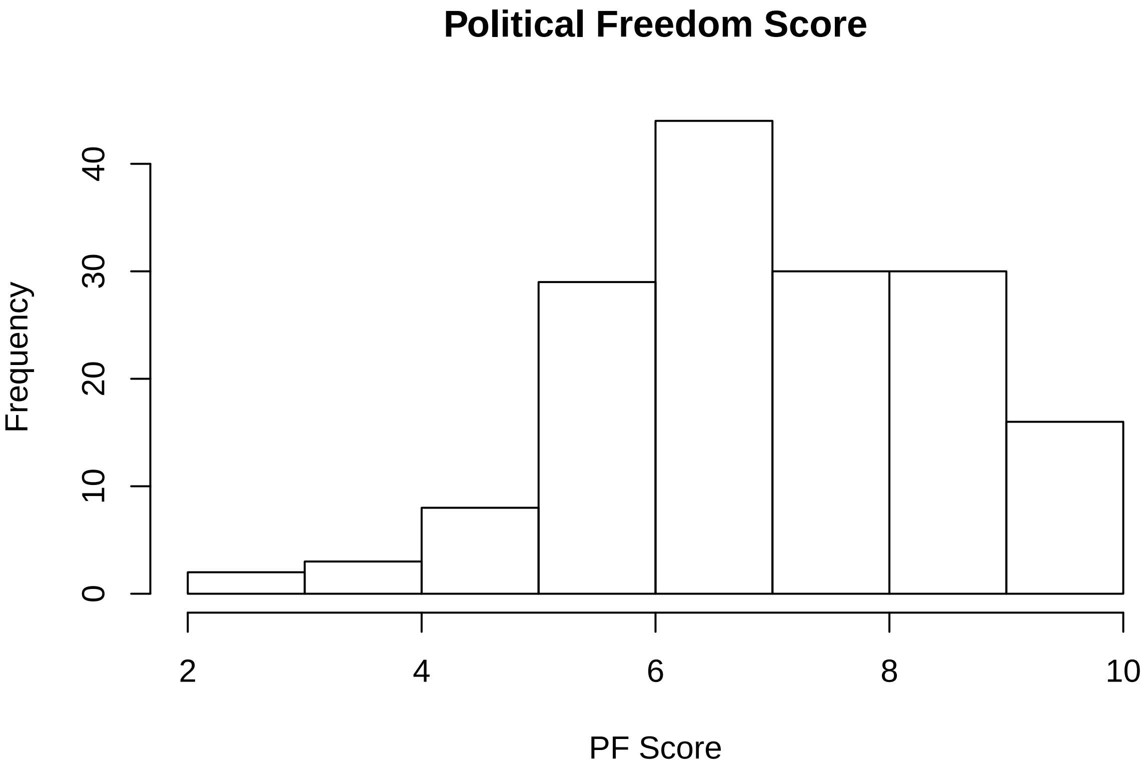
- Regression context in the context of causal inference helps to address comparisons of different treatments if applied to the same units.
- When applying regression models for causal inference, the independent variables are regarded as the causes for the dependent variables. The goal of causal inference studies is to determine whether a particular independent variable actually affects the dependent variable and to estimate the magnitude of that effect.
- A major goal is to get unbiased estimate of the regression coefficients.
- Omitted variables or missing data is significantly more detrimental towards causal inference in contrast to using regression models for prediction purposes
- Multicollinearity is a major concern for causal inference because when two variables are correlated, it can be difficult to get reliable estimates of the coefficients.

Part 3

a.) Describe the distribution of the outcome variable, identify a main predictor that you're interested in studying its effect on the outcome



Our outcome (EF score) is normal but left skewed, mean around 6.5, median around 7. Our main predictor is PF score.



b.) Identify other variables (i.e. predictors, often called covariates) that might be related to the outcome or the main predictor discuss these variables in the context of part 2 above of this assignment.

- The variables that determine political freedom can be used as covariates for economic freedom because they are not directly used in the calculation of economic freedom (pf_rol, pf_ss, pf_movement), but are correlated, which we proved in previous week's assignments.
- In this case, regression of these values will likely result in a line with a high goodness-of-fit, but with accuracy values should be lower than the regression using PF_score. This is due to the fact that PF is a more general variable than the various pf subcategories that we are using as covariables.

c.) Carry out univariate logistic regression of the outcome on each of the predictors including the main predictor, interpret the results in terms of odds ratio etc.

```
## (Intercept)      pf_rol
##   4.5847292   0.4335556
```

```
## (Intercept)       pf_ss
##   3.7070908   0.3788679
```

```
## (Intercept) pf_movement
##   5.4572220   0.1718419
```

```
## (Intercept)    pf_score
##   3.6833651   0.4465781
```

d.) Fit a multiple logistic regression model by including more than one predictors, interpret the results in terms of conditional odds ratio etc.

Coefficients:

```
## (Intercept)      pf_rol       pf_ss pf_movement
##  3.91901808  0.34819576  0.08122382  0.05619437
```

```
#read in dataset
data<-read.csv("hfi_cc_2019.csv")
data<-data[data$year=="2017",]
#predictor: political freedom
#outcome: economic freedom
#make sure both columns have no missing data
#sum(as.character(data$pf_score)=="-")==0
#sum(as.character(data$ef_score)=="-")==0

#distribution of outcome aka EF
hist(as.numeric(as.character(data$ef_score)), main = "Economic Freedom Score", xlab = "EF Score")
hist(as.numeric(as.character(data$pf_score)), main = "Political Freedom Score", xlab = "PF Score")
#make the columns into numerics
pf_rol<-as.numeric(as.character(data$pf_rol))
pf_ss<-as.numeric(as.character(data$pf_ss))
pf_movement<-as.numeric(as.character(data$pf_movement))
pf_score<-as.numeric(as.character(data$pf_score))
ef_score<-as.numeric(as.character(data$ef_score))
#predictor: pf_rol, outcome: ef_score
rol<-glm(ef_score~pf_rol)
rol$coefficients
#predictor: pf_ss, outcome: ef_score
ss<-glm(ef_score~pf_ss)
ss$coefficients
#predictor: pf_movement, outcome: ef_score
movement<-glm(ef_score~pf_movement)
movement$coefficients
#predictor: pf_score, outcome: ef_score
score<-glm(ef_score~pf_score)
score$coefficients
reg<-glm(ef_score~pf_rol+pf_ss+pf_movement)
reg$coefficients
```