Week_7

Kalyani Cauwenberghs

2/14/2020

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

## Loading required package: lattice

## Loading required package: robustbase
```

Week 7

Part 1

Group 9:

Good job on explaining the dataset, informing us on how many rows and columns there are. Also it is good that you shared the numerical, categorical, and miscellaneous variables

How did you all end up using the miscellaneous variables? In what calculations are they used? I am not very sure of how these variables are useful

The box plots were good for showing us the variables of interests, easy way to show the outliers and average values (median)

Good color contrast on the slides using a dark background followed with a light colored text, also the font was fine and not too distracting

Professor note: Include more categorical variables (language, origin, etc) → improve the R2

Great example of logistic regression that included a sigmoid curve

Good explanation of what you are trying to do. Good job condensing the information taught in lecture

Good job explaining statistical terms such as wald statistics, univariate model, etc.

Recommend that you calculate the for each observation and use that to compare in your analysis

Provide the odds ratio for univariate regressions

1.0 CL (confidence level). 1.0 = 1 star (alpha = 0.05). Specify the alphas next time.

What was the point of including interactions in the presentation without having done any interactions things in your data?

You should consider looking at the fractions for the following week

Group 10:

I would recommend not just staring at the computer or the slides and try to make more eye contact with the audience

Introduction is a bit too long and includes unnecessary stuff. Implies lack of understanding of the subject matter.

Your group should focus more on the statistical concepts covered in class for the week, such as univariate models, multivariate models, etc.

Why is the description of data cleaning necessary? Perhaps more information about how the missing ness would affect the analysis is needed.

High pb = higher than 120 systolic, higher than 80 diastolic

Missing description about colinearity calculations

You can standardize the variance to minimize the number of calculations that you need to do

The univariate model and interpretation section is informative and useful, but get to these points quicker

Overall, the data cleaning and selection of variables (?) and discussion of collinearity took way too much time

Interpretation and odds ratios have very good real life interpretations (an x increase in this would predict a y increase in that)

"One unit increase": should provide units (cm, etc.)

The visuals and images shown in the slides for multiple logistic regression models and for univariate models are quite difficult to look at. It is not the most visually appealing images to demonstrate your findings

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

Description of data set

Purpose: measures the level of human freedom ("absence of coercive constraint") Observations: 162 countries Variables: measures of human freedom. Human Freedom score is the total score for each country that is determined by political and economic freedom scores, which are each determined by their own respective subscores like rule of law or economic regulation. All these categories and subcategories are variables in this dataset, along with country/region/year. The variables we will use:

ef_score: economic freedom score

ef_score_binary: 1 if the above is above avg, 0 if not.

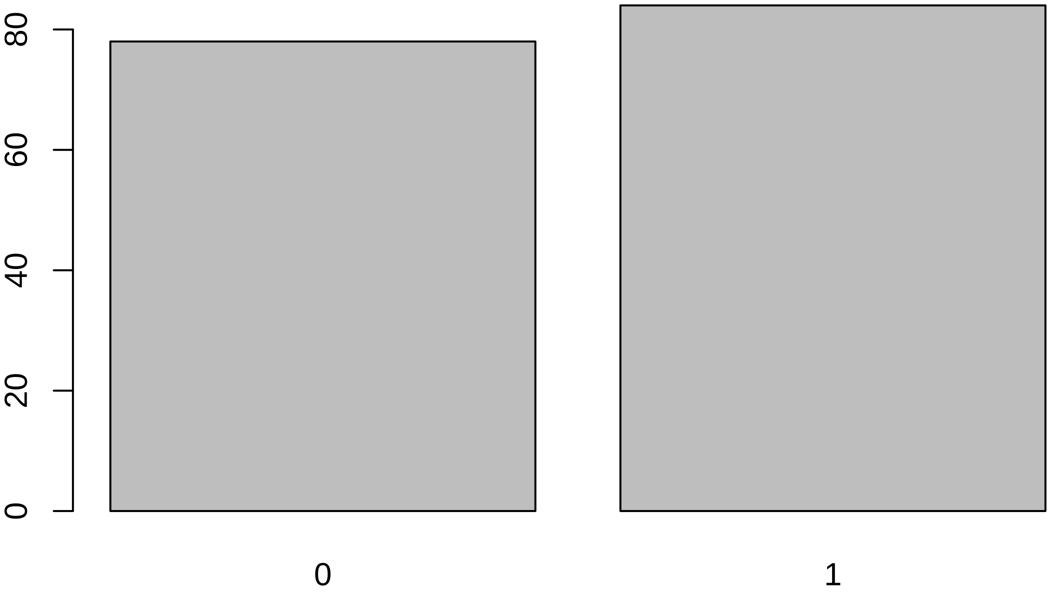
pf_score: political freesom score

pf_rol: rule of law, subscore of pf_score

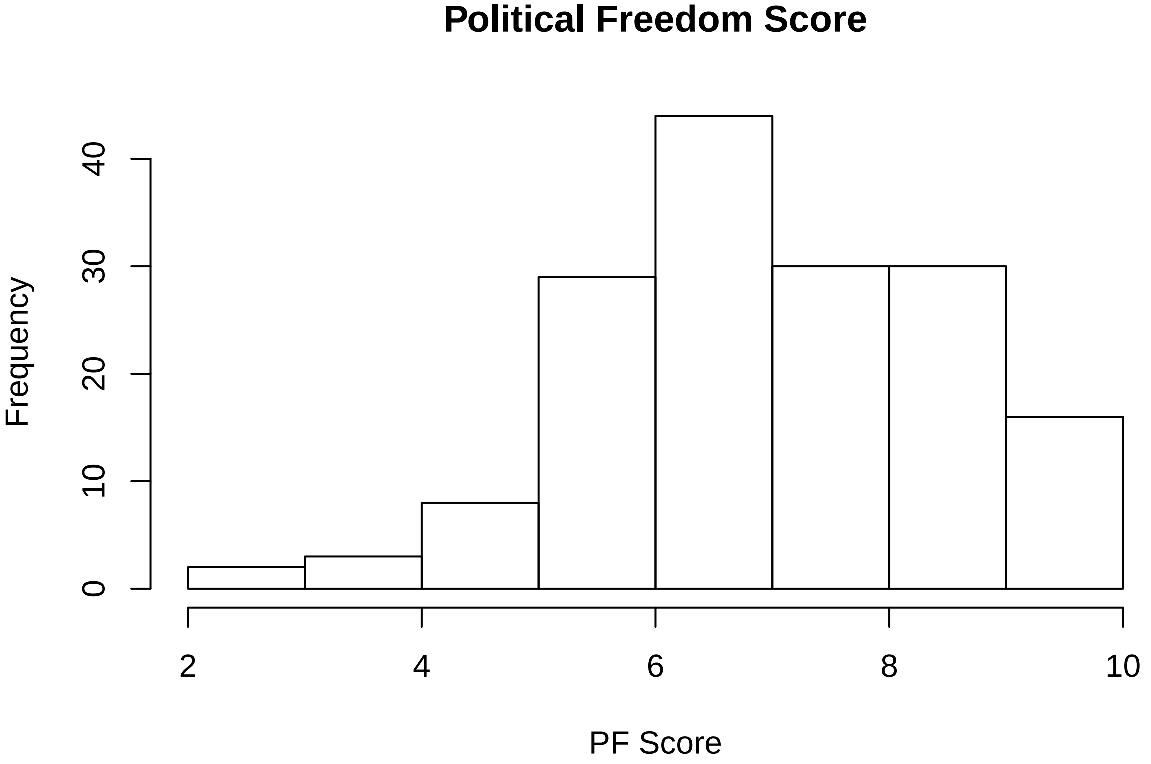
pf_ss: security and safety, subscore of pf_score

pf_movement: movement, subscore of pf_score

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 is binary economic freedom score (1 if above average and 0 if average or below). Slightly more observations are above the mean than below. Our main predictor is political freedom score.



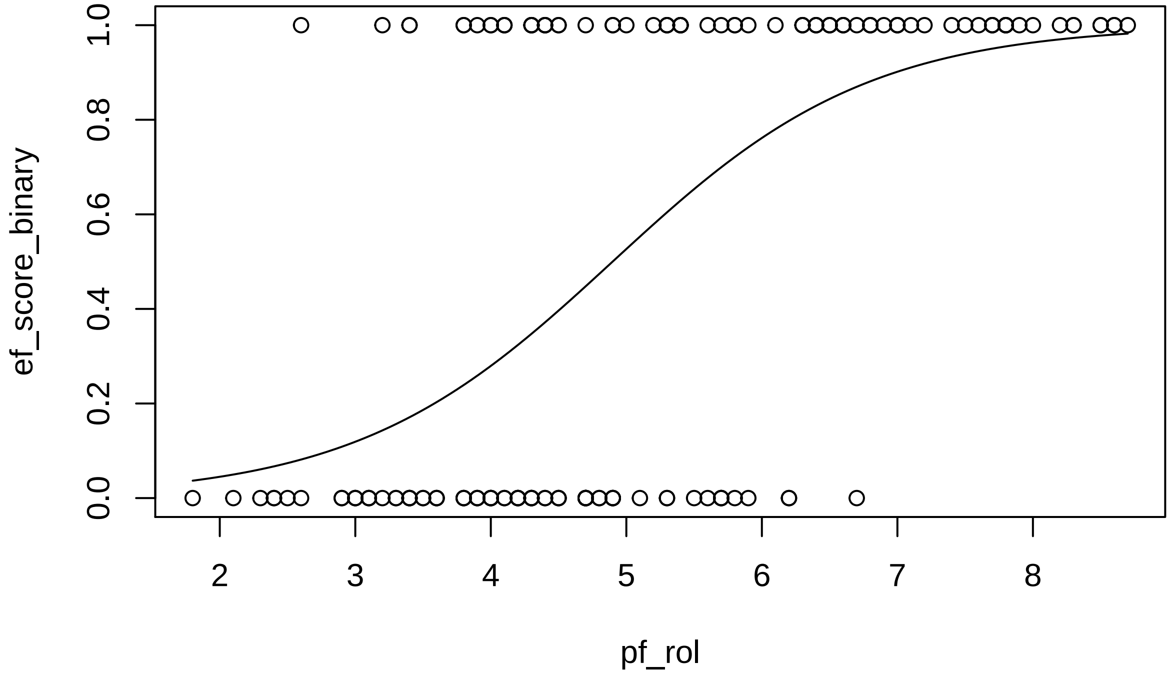
b.) Identify other variables (i.e. predictors, often called covariates) that might be related to the outcome or the main predictor discuss these variables in the context of part 2 above of this assignment.

- 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.

```
## [1] "predictor: pf_rol, outcome: ef_score_binary"
```

```
## (Intercept)      pf_rol
##   -5.162178    1.053811
```

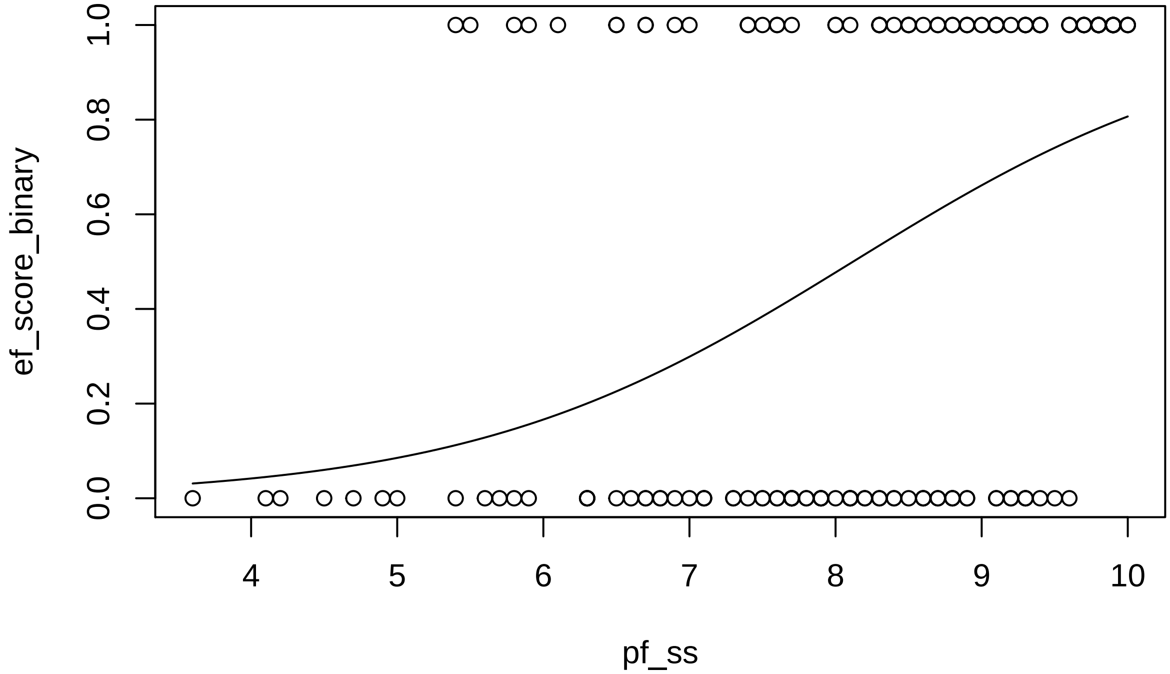


```
## [1] "odds ratio:"
```

```
##   pf_rol
## 2.868563
```

```
## [1] "predictor: pf_ss, outcome: ef_score_binary"
```

```
## (Intercept)       pf_ss
##  -6.1738924   0.7603208
```

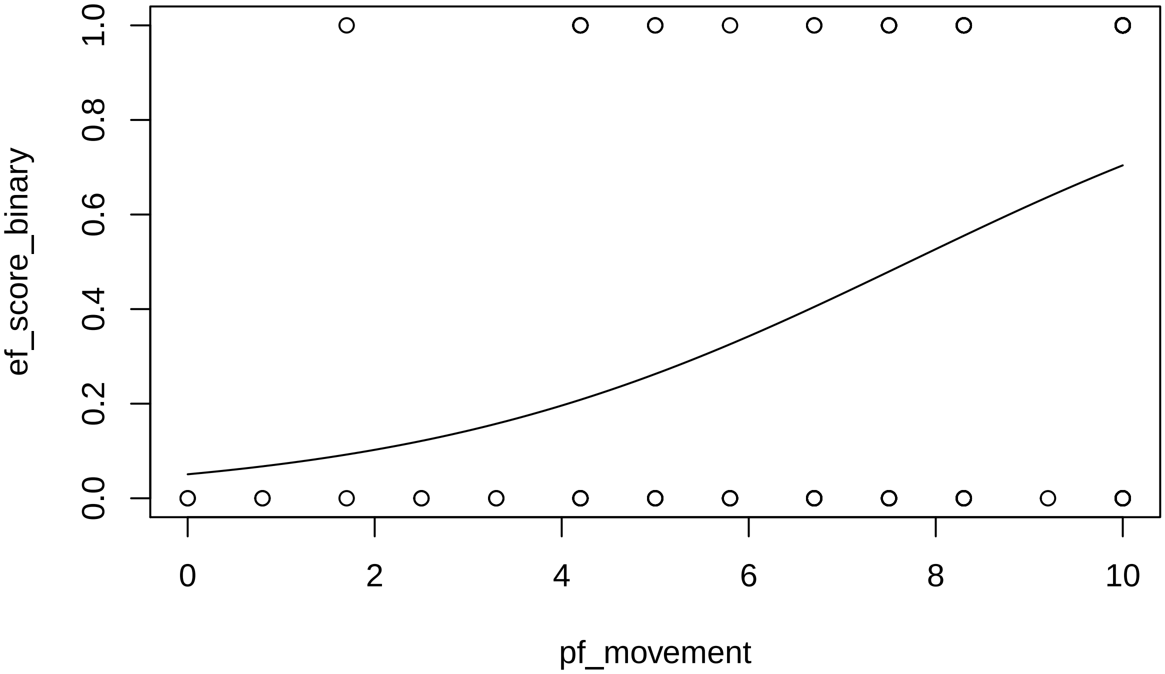


```
## [1] "odds ratio:"
```

```
##    pf_ss
## 2.138962
```

```
## [1] "predictor: pf_movement, outcome: ef_score_binary"
```

```
## (Intercept) pf_movement
##  -2.9295762   0.3796197
```

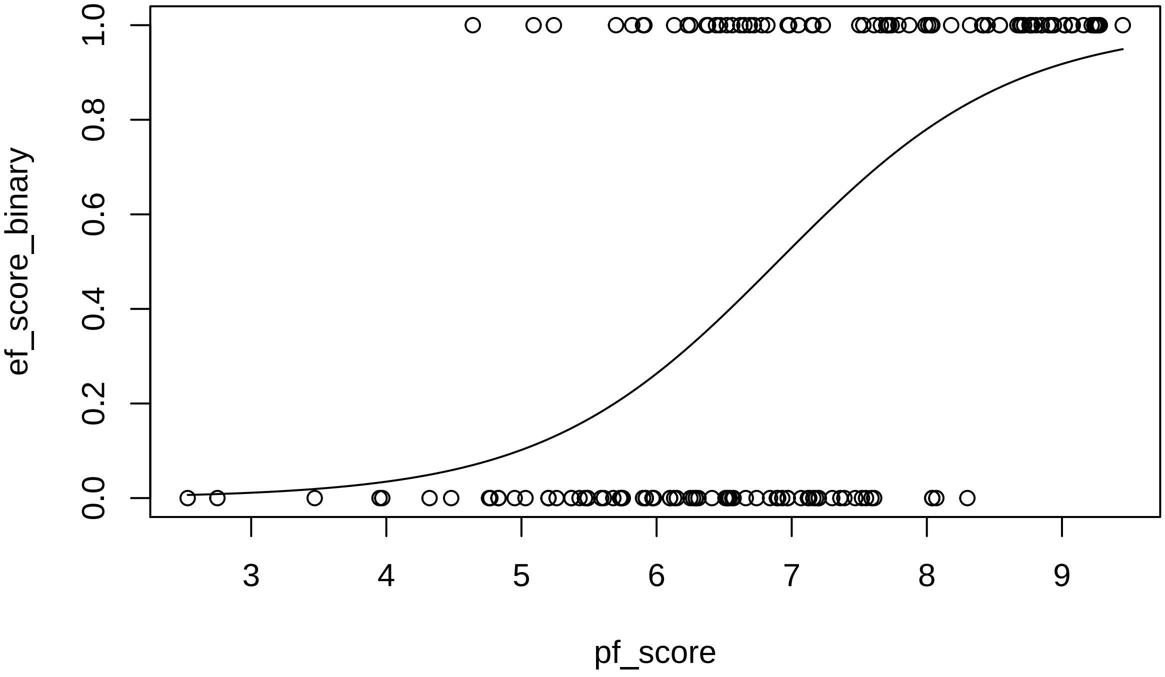


```
## [1] "odds ratio:"
```

```
## pf_movement
##    1.461729
```

```
## [1] "predictor: pf_score, outcome: ef_score_binary"
```

```
## (Intercept)    pf_score
##   -7.915437    1.147747
```



```
## [1] "odds ratio:"
```

```
## pf_score
## 3.151084
```

Interpretation of odds-ratio: a one unit increase in the predictor predicts that the odds of ef_score being 1 gets multiplied by beta 1.

d.) Use one of the stepwise procedures we talked about, together with computing the generalized R-squared and AIC for each model considered during the process, to arrive at a 'final' multiple logistic regression model. Consider interaction terms also. Interpret the results from your final model in the context of the research question that you are trying to answer.

We will use forwards stepwise selection.

No interaction terms:

```
## [1] "Already ran univariate regression on each predictor."
```

```
## [1] "Now time to compare p-values:"
```

```
## [1] "p-value rol"
```

```
##  (Intercept)       pf_rol
## 1.188909e-09 7.953086e-10
```

```
## [1] "p-value ss"
```

```
##  (Intercept)        pf_ss
## 8.247448e-07 3.506663e-07
```

```
## [1] "p-value movement"
```

```
##  (Intercept)  pf_movement
## 1.871991e-05 3.577073e-06
```

```
## [1] "They all have p-value below our threshold, 0.05."
## [1] "pf_rol has the lowest p-value among all the predictors,"

## [1] "so we will make this the first variable."

## [1] "###Model 1:"

## [1] "##coefficients: "

## (Intercept)      pf_rol
##   -5.162178    1.053811

## ##AIC:  162.9363

## ##generalized R-squared:  0.3322452

## [1] "Now we will regress ef_score_binary with pf_rol and one"

## [1] "other variable for each other variable (pf_movement and pf_ss)."

## [1] "p-value rol ss"

##  (Intercept)       pf_rol        pf_ss
## 2.340530e-06 1.102322e-06 2.596491e-01

## [1] "p-value rol movement"

##  (Intercept)       pf_rol  pf_movement
## 2.515938e-09 5.097898e-08 9.364143e-03

## [1] "The p-value for pf_ss is above threshold, so we will  eliminate it."

## [1] "###Model 2 (final model):"

## [1] "##coefficients: "

## (Intercept)      pf_rol pf_movement
##  -6.6164737   0.9915910   0.2277021

## ##AIC:  157.6976

## ##generalized R-squared:  0.3614256
```

Interaction terms: We will again use forward stepwise selection.

```
## [1] "The first model from the last time was based on pf_rol:"

## [1] "###Model 1:"

## [1] "##coefficients: "

## (Intercept)      pf_rol
##   -5.162178    1.053811

## ##AIC:  162.9363

## ##generalized R-squared:  0.3322452

## [1] "Now we will regress ef_score_binary with pf_rol and one"

## [1] "other variable for each other variable (pf_movement and pf_ss)."

## [1] "p-value rol ss"

##  (Intercept)       pf_rol        pf_ss
## 2.340530e-06 1.102322e-06 2.596491e-01

## [1] "p-value rol, ss, rol x ss"
##  (Intercept)       pf_rol        pf_ss pf_rol:pf_ss
##    0.8718737    0.7484612    0.3929201    0.2092805
```

```
## [1] "p-value rol movement"
```

```
##  (Intercept)       pf_rol  pf_movement
## 2.515938e-09 5.097898e-08 9.364143e-03
```

```
## [1] "p-value rol, movement, rol x movement"
```

```
##        (Intercept)             pf_rol        pf_movement pf_rol:pf_movement
##         0.02603582         0.07256173         0.39093085         0.75427493
```

```
## [1] "pf_rol and pf_movement are the only pair of variables"
```

```
## [1] "that pass the threshold of p<0.05. We throw the others out."
```

```
## [1] "we are left with pf_movement as the 2nd predictor."
```

```
## [1] "###Model 2 (final model):"
```

```
## [1] "##coefficients: "
```

```
## (Intercept)      pf_rol pf_movement
##  -6.6164737   0.9915910   0.2277021
```

```
## ##AIC:  157.6976
```

```
## ##generalized R-squared:  0.3614256
```

Our research question: Does high political freedom predict high economic freedom?

Answer: our final model is the following:

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -6.6164737 1.11006878 -5.960418 2.515938e-09
## pf_rol       0.9915910 0.18201472  5.447862 5.097898e-08
## pf_movement  0.2277021 0.08762946  2.598466 9.364143e-03
```

As one can see, for the coefficients preceding both variables (pf_rol and pf_movement), the estimate plus or minus the standard error does not encapsulate 0. This means it is reasonable to say that these coefficients are both positive. This means pf_rol and pf_movement positively predict ef_score_binary. Both variables are only part of political freedom, so we can say only some, not all, aspects of political freedom positively predict economic freedom. Since taking into account interacting terms increases p-values by a lot, we can reasonably say that these individual aspects of political freedom don't interact with each other.

e.) write a paragraph discussing limitations from your data source, assumptions, approaches etc. as applicable. For those that the grader marked comments about the i.i.d. assumption from week 5 homework, be sure to including discussion on those.

The data that we are using is survey data. As such there are some considerations that were made to use this dataset for analysis. Firstly, one of our data points is the average sum of count values. Due to the data not having a zero value and decimal points being not much of a problem in our dataset, we use this datapoint as a continuous variable. This is clearly not ideal, but it should not present too much of an issue. The fact that this datapoint has decimal points is not much of a problem as the decimal points add more information and detail about the dataset at hand. The other large assumption that we are making is the idea that our data is independent. Our data points are academic surveys done on the quality of freedom in the legislation made by the specific country and the execution of that legislation in different countries. The two arguments we make in regards to this data's independence are: no government can use its legislation power to directly impact the legislation of another sovereign country, and that the decisions of academics regarding the policies of one country are not impacted by decisions that they made on those of other countries. The first argument is certainly weak, as there are many influences between nations, and we are living in a globalized society. However, a very strict interpretation of the question, where we claim independence if and only if there is a direct relationship between one country's legislation and other countries. That is, if one country can issue legislation that forces the other country to issue its own legislation. The second point is more solid. Our data is not ground-truth but is rather the perception and analysis of experts and academics regarding the ground-truth data. Unless this study was run irresponsibly, holding the assumption that the experts considered each country independently is reasonable. Finally, as the data points are from a secondary source (professional opinion), the subset of data points that the study chose may not be accurately representing the true metric. For example, our study does a weighted average ~50 individual metrics to determine the political freedom score of the respective country. The actual political freedom of a country may actually consist of 200 or 2000 individual metrics.

Week 8:

Part 1:

Group 11:

First group to provide an outline for how you will cover the entire presentation. A nice way to start and set up your presentation

The slides are relatively easy to look at.. It shows what is necessary to convey the information without being too busy

Excellent, simple explanation of what insulin is

Love the color coding of the variable type (blue for discrete etc.)

What is the range of pedigree function? Of BMI values, etc.?

Wouldn't skin fold thickness be closely associated with BMI? The predictors should be independent iirc.

The age vs diabetes graph should take into account that there was probably not an equal amount of people in each age range, so the distribution would not be reflective of the likelihood of diabetes in each age group

Good thing to mention that high BMI does not = fat, so this is an instance where BMI would not predict type 2 diabetes

Glucose level x-axis is too wide, wasted space

Your explanation of using all the P-values to evaluate the variable is correct and well explained

Good walkthrough of your presentation by explaining how removing a certain variable will affect other variables (e.g AIC example)

Make sure to also compute Rˆ2 for logistic regression

Watch the timing of your presentation, the timing was informative, but a little slow so your presentation ended up extending for a long time

Group 12:

Make sure to fix the errors you had in the presentation slides

R2 is not the R2 she talked about in class. It is the linear model, not the logistic model.

It is a good thing to get typed equations and code in there to relate between the two

The presentation was fine, it was just rushed.. Not necessarily your fault because the previous group took too much time

Good work explaining different terms and concepts such as interaction terms Make sure to be clear on how to select interaction terms

There could be a lot of biased because of the multiple taco shops, so you should make sure to account for this potential bias

Multiple locations for the same taco shop → correlated ratings, not i.i.d.

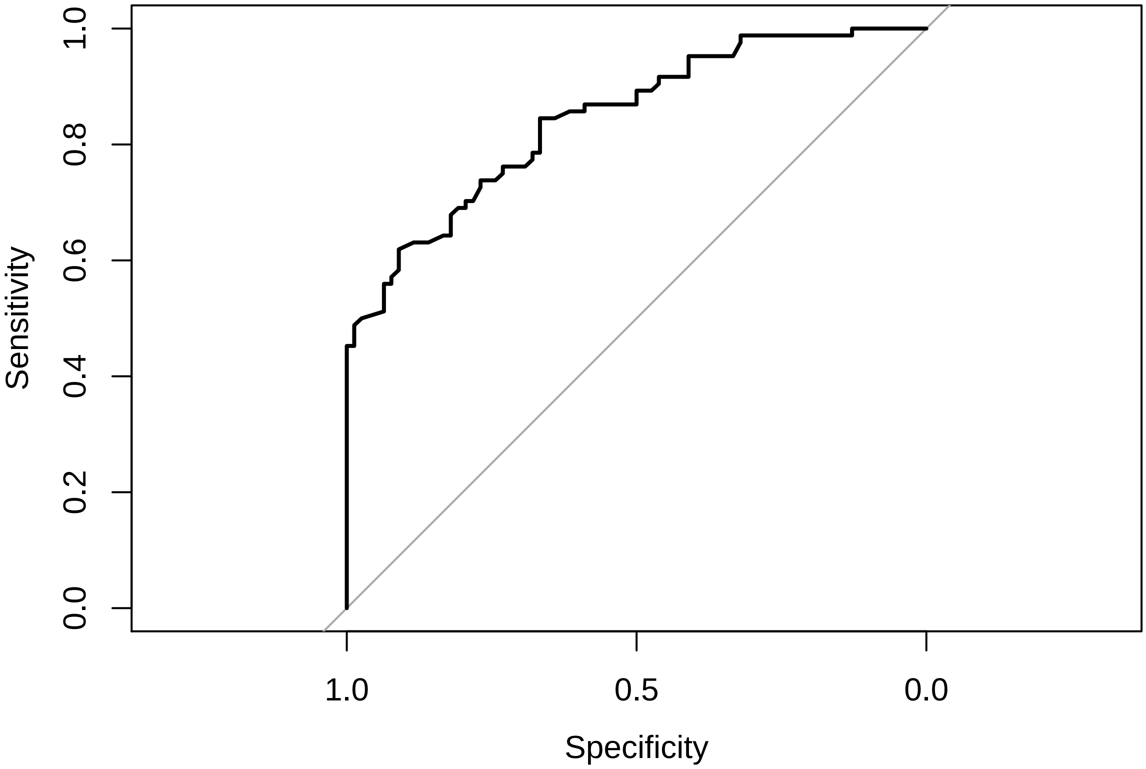
Part 2:

a) Perform prediction on the whole data set, plot the ROC curve and compute the AUC;

```
## [1] "Plotting ROC of our selected model"
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



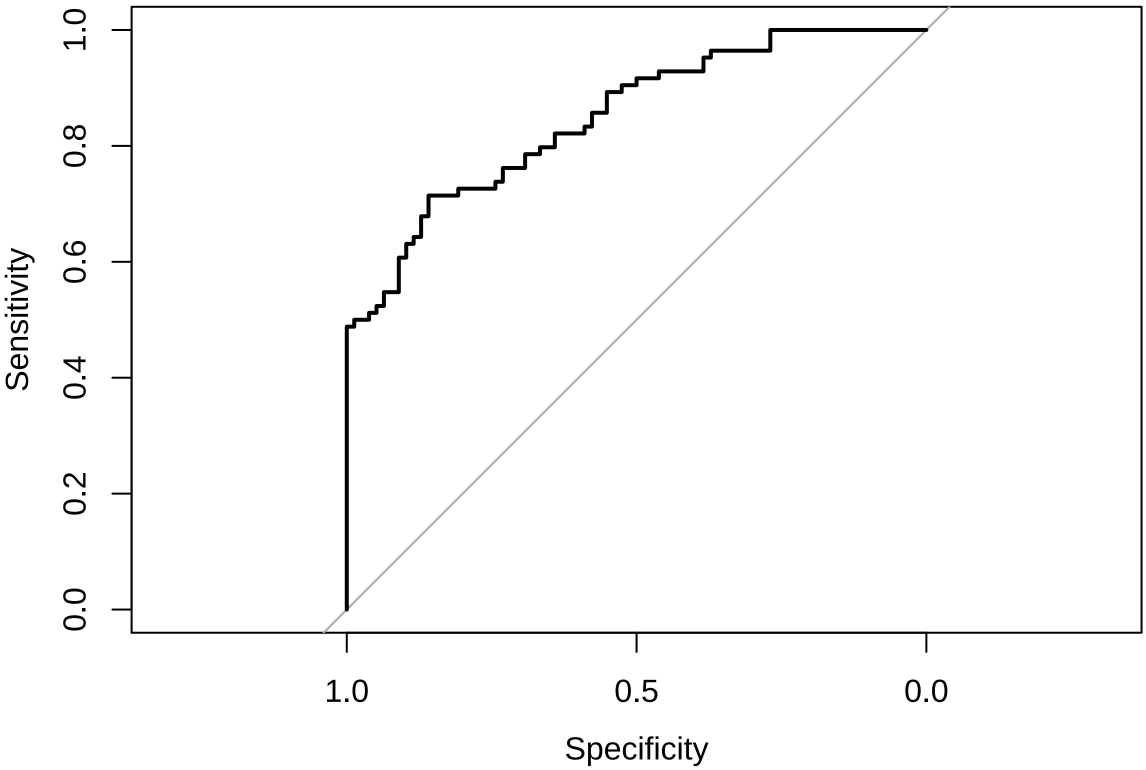
```
## [1] "AUC of our selected model"
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
## Area under the curve: 0.8452
```

```
## [1] "Plotting ROC of all variables (whole data set)"
```

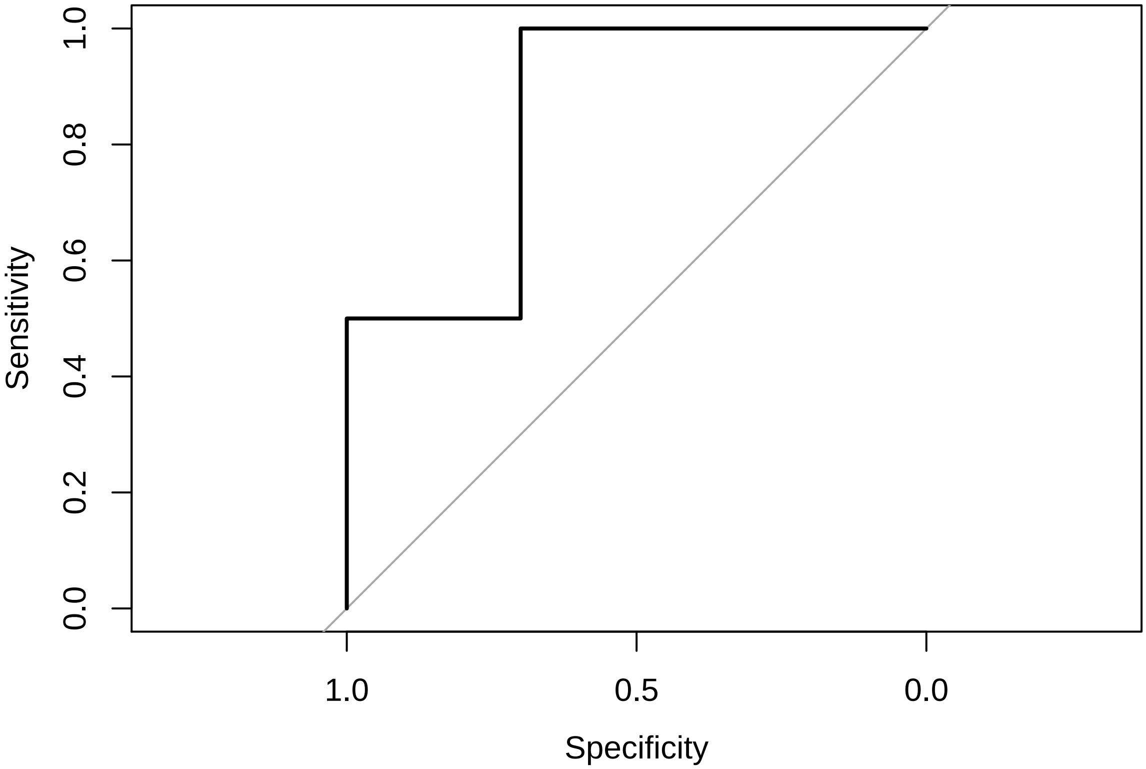
```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



```
## [1] "AUC of all variables (whole data set)"

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases

## Area under the curve: 0.8497
```

b) Use a randomly chosen 90% of your observations as training sample to fit the final model (if your data set is too small, you may reduce your final model to a smaller one this week), and use the rest 10% as test sample to compute the out-of-sample AUC;

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04738 0.32978 0.49649 0.50987 0.71644 0.97130

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```



```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
## Area under the curve: 0.85
```

c) Now instead of test-training sample, carry out 10-fold CV with 10 repetitions to estimate the out-of-sample AUC;

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## [1] "final AUC:"

## [1] 0.85342
```

d) Comment on your results.

The AUC of our selected model and our complete model from (a) as well as our cross validation AUC from (c) were all around 0.85 with not too much variance. This means that after training (the 90% of our observations), our model performs with 85% accuracy when working with testing data (the 10% of our observations).

Our AUC from part (b) was around 0.85 but with lots of variation when running it again and again. This was due to the small sample size (16) in the testing data.

For (a), our ROC curves were smooth because we didn't split our data into groups. This made our data have more observations to plot. For (b), they were jagged because of lack of observations.

For (c), since these ROC curves were just (b) repeated, they were also a bunch of jagged edges for each fold.

Code:

```
#function code
plot_logistic<-function(r, predictor){
  b0<-r$coefficients[1]
  b1<-r$coefficients[2]
  min<-min(predictor)
  max<-max(predictor)
  x<-seq(min,max,0.01)
  func<-(exp(b0+b1*x))/(1+exp(b0+b1*x))
  lines(x,func)
}

r2log<-function(glm_obj) {
  glm_null<-glm(ef_score_binary~1, family = binomial())
  #likelihood ratio test statistic
  t<-2*(logLik(glm_obj)-logLik(glm_null))/162
  r2<-1-exp(-t)
  return(r2)
}
#required package
#install.packages("pROC")
library("pROC")
#install.packages("cvTools")
library(cvTools)
#read in dataset
```

```r
data<-read.csv("hfi_cc_2019.csv")
data<-data[data$year=="2017",]
#predictor: political freedom
#outcome: economic freedom (binary)
#make sure both columns have no missing data
#sum(as.character(data$pf_score)=="-")==0
#sum(as.character(data$ef_score_binary)=="-")==0
ef_score<-as.numeric(as.character(data$ef_score))
mean_ef_score<-mean(ef_score)
ef_score_binary<-numeric(length(ef_score))
ef_score_binary[ef_score>mean_ef_score]<-1
ef_score_binary[ef_score<=mean_ef_score]<-0
#plot EF binary
barplot(table(ef_score_binary))
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$pf_score))
print("predictor: pf_rol, outcome: ef_score_binary")
rol<-glm(ef_score_binary~pf_rol, family = binomial)
rol$coefficients
plot(pf_rol,ef_score_binary)
plot_logistic(rol,pf_rol)
print("odds ratio:")
print(exp(rol$coefficients[2]))

print("predictor: pf_ss, outcome: ef_score_binary")
ss<-glm(ef_score_binary~pf_ss, family = binomial)
ss$coefficients
plot(pf_ss,ef_score_binary)
plot_logistic(ss,pf_ss)
print("odds ratio:")
print(exp(ss$coefficients[2]))

print("predictor: pf_movement, outcome: ef_score_binary")
movement<-glm(ef_score_binary~pf_movement, family = binomial)
movement$coefficients
plot(pf_movement,ef_score_binary)
plot_logistic(movement,pf_movement)
print("odds ratio:")
print(exp(movement$coefficients[2]))

print("predictor: pf_score, outcome: ef_score_binary")
score<-glm(ef_score_binary~pf_score, family = binomial)
score$coefficients
plot(pf_score,ef_score_binary)
plot_logistic(score,pf_score)
print("odds ratio:")
print(exp(score$coefficients[2]))
p_val<-0.05
print("Already ran univariate regression on each predictor.")
print("Now time to compare p-values:")
print("p-value rol")
print(coef(summary(rol))[,4])
print("p-value ss")
print(coef(summary(ss))[,4])
print("p-value movement")
print(coef(summary(movement))[,4])
print("They all have p-value below our threshold, 0.05.")
print("pf_rol has the lowest p-value among all the predictors,")
print("so we will make this the first variable.")
print("###Model 1:")
print("##coefficients: ")
print(rol$coefficients)
cat("##AIC: ", rol$aic)
cat("##generalized R-squared: ", r2log(rol))
print("Now we will regress ef_score_binary with pf_rol and one")
print("other variable for each other variable (pf_movement and pf_ss).")
print("p-value rol ss")
print(coef(summary(glm(ef_score_binary~pf_rol+pf_ss, family = binomial())))[,4])
print("p-value rol movement")
print(coef(summary(glm(ef_score_binary~pf_rol+pf_movement, family = binomial())))[,4])
print("The p-value for pf_ss is above threshold, so we will  eliminate it.")
print("###Model 2 (final model):")
m2<-glm(ef_score_binary~pf_rol+pf_movement, family = binomial())
print("##coefficients: ")
print(m2$coefficients)
cat("##AIC: ", m2$aic)
cat("##generalized R-squared: ", r2log(m2))
print("The first model from the last time was based on pf_rol:")
print("###Model 1:")
print("##coefficients: ")
print(rol$coefficients)
cat("##AIC: ", rol$aic)
cat("##generalized R-squared: ", r2log(rol))
print("Now we will regress ef_score_binary with pf_rol and one")
print("other variable for each other variable (pf_movement and pf_ss).")
print("p-value rol ss")
print(coef(summary(glm(ef_score_binary~pf_rol+pf_ss, family = binomial())))[,4])
print("p-value rol, ss, rol x ss")
print(coef(summary(glm(ef_score_binary~pf_rol+pf_ss+pf_rol*pf_ss, family = binomial())))[,4])
print("p-value rol movement")
print(coef(summary(glm(ef_score_binary~pf_rol+pf_movement, family = binomial())))[,4])
print("p-value rol, movement, rol x movement")
print(coef(summary(glm(ef_score_binary~pf_rol+
          pf_movement+pf_rol*pf_movement,
          family = binomial())))[,4])

print("pf_rol and pf_movement are the only pair of variables")
print("that pass the threshold of p<0.05. We throw the others out.")
print("we are left with pf_movement as the 2nd predictor.")
print("###Model 2 (final model):")
m2<-glm(ef_score_binary~pf_rol+pf_movement, family = binomial())
print("##coefficients: ")
print(m2$coefficients)
cat("##AIC: ", m2$aic)
cat("##generalized R-squared: ", r2log(m2))
coef(summary(m2))
#formatting issues
data$ef_score_binary<-ef_score_binary
data$pf_rol<-pf_rol
data$pf_ss<-pf_ss
data$pf_movement<-pf_movement
data$pf_score<-pf_score


print("Plotting ROC of our selected model")
plot.roc(ef_score_binary, predict(m2, type = "response"))
print("AUC of our selected model")
auc(ef_score_binary, predict(m2, type = "response"))
print("Plotting ROC of all variables (whole data set)")
all<-glm(ef_score_binary~pf_rol+pf_movement+pf_ss+pf_score, data=data)
plot.roc(ef_score_binary, predict(all, type = "response"))
print("AUC of all variables (whole data set)")
auc(ef_score_binary, predict(all, type = "response"))
#m2 = final model from week 7
train = sample(162, round(162*0.9),replace = FALSE)
fit = glm(ef_score_binary~pf_rol+pf_movement, family=binomial(),data=data[train,])
summary(predict(fit, newdata = data[-train,], type = "response"))
plot.roc(data[-train,]$ef_score_binary, predict(fit,newdata = data[-train,], type = "response"))
auc(data[-train,]$ef_score_binary, predict(fit, newdata = data[-train,], type = "response"))
folds <- cvFolds(n = 162, K = 10, R = 1)
auc = numeric(10)
for( i in 1:10){
  train = folds$subsets[folds$which != i]
  fit = glm(ef_score_binary~pf_rol+pf_movement, family=binomial(),data=data[train,])
  auc[i]<-auc(data[-train,]$ef_score_binary, predict(fit, newdata = data[-train,], type = "response"))
}
print("final AUC:")
print(mean(auc))
```