Project 2: Modeling, Testing, and Predicting

Josh Goh SDS348

05/01/2020

Table of Contents

1. Introduction:

One of the leading causes of morbidity and mortality in the world today, heart disease is an umbrella term that covers all different types of specific cardiovascular disease such as strokes, heart attacks, arrhythmia, angina, coronary artery disease, and so on. From the raw Cleveland Heart Disease dataset there have been 1,025 observations on patients since 1988 describing information regarding 14 prediction attributable variables. After cleaning the dataset, omitting NAs, and dropping variables that were not clearly elucidated, the dataset is left with 944 observations with 11 variables: age (range from 29-77 years of age), sex (Male or Female), resting blood pressure (range from 94-192 mmHg), serum cholesterol (range from 126-564 mg/dL), fasting blood sugar (above or below 120 mg/dL), resting electrocardiographic results (normal, ST-T wave abnormality, or left ventricular hypertrophy), maximum attainable heart rate (71-202 bpm), exercise induced angina (yes or no), ST peak slope (flat, upsloping, or downsloping), thalassemia status (normal, fixed defect, or reversible defect), and heart disease diagnosis (‘1’ = Disease or ‘0’ = No Disease).

#Library

library(tidyverse)  
library(lmtest)  
library(plotROC)  
library(glmnet)  
library(MASS)  
library(nnet)  
library(ggplot2)  
library(dplyr)  
library(plotly)  
library(ggExtra)  
library(tidyverse)  
library(mvtnorm)  
library(vegan)  
library(sandwich)

#Preparing Dataset

#importing dataset  
Heart\_Disease\_Raw <- read.csv("heart.csv")  
  
#cleaning dataset  
Heart\_Disease <- Heart\_Disease\_Raw %>% select(-cp, -ca, -oldpeak) %>% rename(Age=age, Sex=sex, Rest\_Blood\_Pressure = trestbps, Serum\_Cholesterol = chol, Fasting\_Blood\_Sugar = fbs, Rest\_ECG = restecg, Max\_HR = thalach, Exercise\_Induced\_Angina = exang, ST\_Slope = slope, Thalassemia = thal, Diagnosis = target) %>% mutate(Sex=recode(Sex, "1" = "Male", "0" = "Female")) %>% mutate(Fasting\_Blood\_Sugar=recode(Fasting\_Blood\_Sugar, "1" = "> 120 mg/dL", "0" = "< 120 mg/dL")) %>% mutate(Rest\_ECG=recode(Rest\_ECG, "2" = "left ventricular hypertrophy", "1" = "ST-T wave abnormality", "0" = "normal")) %>% mutate(Exercise\_Induced\_Angina=recode(Exercise\_Induced\_Angina, "1" = "Yes", "0" = "No")) %>% mutate(ST\_Slope=recode(ST\_Slope, "3" = "downsloping", "2" = "flat", "1" = "upsloping")) %>% mutate(Thalassemia=recode(Thalassemia, "3" = "reversible defect", "2" = "fixed defect", "1" = "normal")) %>% na.omit()  
head(Heart\_Disease)

## Age Sex Rest\_Blood\_Pressure Serum\_Cholesterol  
Fasting\_Blood\_Sugar Rest\_ECG Max\_HR  
## 1 52 Male 125 212 < 120 mg/dL ST-T wave abnormality 168  
## 4 61 Male 148 203 < 120 mg/dL ST-T wave abnormality 161  
## 5 62 Female 138 294 > 120 mg/dL ST-T wave abnormality  
106  
## 6 58 Female 100 248 < 120 mg/dL normal 122  
## 8 55 Male 160 289 < 120 mg/dL normal 145  
## 9 46 Male 120 249 < 120 mg/dL normal 144  
## Exercise\_Induced\_Angina ST\_Slope Thalassemia Diagnosis  
## 1 No flat reversible defect 0  
## 4 No flat reversible defect 0  
## 5 No upsloping fixed defect 0  
## 6 No upsloping fixed defect 1  
## 8 Yes upsloping reversible defect 0  
## 9 No flat reversible defect 0

#dataset information  
Heart\_Disease %>% summarize\_all(n\_distinct)

## Age Sex Rest\_Blood\_Pressure Serum\_Cholesterol  
Fasting\_Blood\_Sugar Rest\_ECG Max\_HR  
## 1 41 2 47 147 2 3 89  
## Exercise\_Induced\_Angina ST\_Slope Thalassemia Diagnosis  
## 1 2 2 3 2

Heart\_Disease %>% summarize\_all(min)

## Age Sex Rest\_Blood\_Pressure Serum\_Cholesterol  
Fasting\_Blood\_Sugar Rest\_ECG  
## 1 29 Female 94 126 < 120 mg/dL left ventricular  
hypertrophy  
## Max\_HR Exercise\_Induced\_Angina ST\_Slope Thalassemia  
Diagnosis  
## 1 71 No flat fixed defect 0

Heart\_Disease %>% summarize\_all(max)

## Age Sex Rest\_Blood\_Pressure Serum\_Cholesterol  
Fasting\_Blood\_Sugar Rest\_ECG Max\_HR  
## 1 77 Male 192 564 > 120 mg/dL ST-T wave abnormality 202  
## Exercise\_Induced\_Angina ST\_Slope Thalassemia Diagnosis  
## 1 Yes upsloping reversible defect 1

1. MANOVA Testing:

A MANOVA test was conducted on the variables age, resting blood pressure, serum cholesterol, and maximum attainable heart rate to see if any of these numeric variables accounted for a significant difference across the levels of our categorical variable thalassemia status. The MANOVA test resulted in a p-value of < 2.2e-16, which surpasses both the normal alpha-value of 0.5 and the bonferroni correction of 0.003571429. This means that significant differences were found between the Thalassemia groups for at least one of the numeric variables. Univariate ANOVAs were subsequently conducted on each numeric variable to uncover significant differences. The age variable had a resulting p-value of 0.0001388, the resting blood pressure variable had a resulting p-value of 2.194e-05, the serum cholesterol variable had a resulting p-value of 0.004149, and the maximum attainable heart rate variable had a resulting p-value of 2.2e-16. According to the Bonferroni correction of 0.003571429, all of the ANOVAs for the numeric variables were significant except for the serum cholesterol variable. Subsequent post-hoc tests on the significant numeric variables were run in accordance with the bonferroni correction of 0.003571429 to give significant differences between fixed defect and reversible defect with Age (p-value of 0.00013), between normal and fixed defect group with resting blood pressure (p-value of 3.5e-05), between normal and fixed defect group with maximum attainable heart rate (p-value of 1.4e-11), and between reversible defect and fixed defect group with maximum attainable heart rate (p-value of 5.1e-16). In total 1 MANOVA test was performed, 4 ANOVA tests, and 9 post-hoc t tests for a total of 14 hypothesis tests. Without adjustment our type I error rate would have been 51.2325%, but after the bonferroni correction that sets the new alpha level at 0.003571429 the type I error rate can be maintained at 5%. According to eyeball assumptions on multivariate normality and homogeneity of within-group covariance matrices, we can see that the dataset fails to meet the assumptions of MANOVA in multiple ways. The DVs do not visibly have multivariate normality in the multivariate plots and the covariant matrices do not show homogeneity of covariances. A PERMANOVA may have been a better test to avoid the hard-to-meet assumptions.

#MANOVA test on categorical variable 'Thalassemia'  
manova <-manova(cbind(Age, Rest\_Blood\_Pressure, Serum\_Cholesterol, Max\_HR)~Thalassemia, data=Heart\_Disease)  
summary(manova)

## Df Pillai approx F num Df den Df Pr(>F)  
## Thalassemia 2 0.12875 16.152 8 1878 < 2.2e-16 \*\*\*  
## Residuals 941  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1  
' ' 1

#ANOVA test  
summary.aov(manova)

## Response Age :  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Thalassemia 2 1486 743.20 8.967 0.0001388 \*\*\*  
## Residuals 941 77991 82.88  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1  
' ' 1  
##  
## Response Rest\_Blood\_Pressure :  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Thalassemia 2 6156 3078.2 10.85 2.194e-05 \*\*\*  
## Residuals 941 266958 283.7  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1  
' ' 1  
##  
## Response Serum\_Cholesterol :  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Thalassemia 2 29632 14816.2 5.5169 0.004149 \*\*  
## Residuals 941 2527157 2685.6  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1  
' ' 1  
##  
## Response Max\_HR :  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Thalassemia 2 45531 22765.5 48.175 < 2.2e-16 \*\*\*  
## Residuals 941 444682 472.6  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1  
' ' 1

#post-hoc t tests  
pairwise.t.test(Heart\_Disease$Age, Heart\_Disease$Thalassemia, p.adj="none")

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: Heart\_Disease$Age and Heart\_Disease$Thalassemia   
##   
## fixed defect normal   
## normal 0.01233 -   
## reversible defect 0.00013 0.49192  
##   
## P value adjustment method: none

pairwise.t.test(Heart\_Disease$Rest\_Blood\_Pressure, Heart\_Disease$Thalassemia, p.adj="none")

##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data: Heart\_Disease$Rest\_Blood\_Pressure and  
Heart\_Disease$Thalassemia  
##  
## fixed defect normal  
## normal 3.5e-05 -  
## reversible defect 0.0040 0.0061  
##  
## P value adjustment method: none

pairwise.t.test(Heart\_Disease$Max\_HR, Heart\_Disease$Thalassemia, p.adj="none")

##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data: Heart\_Disease$Max\_HR and Heart\_Disease$Thalassemia  
##  
## fixed defect normal  
## normal 1.4e-11 -  
## reversible defect 5.1e-16 0.0038  
##  
## P value adjustment method: none

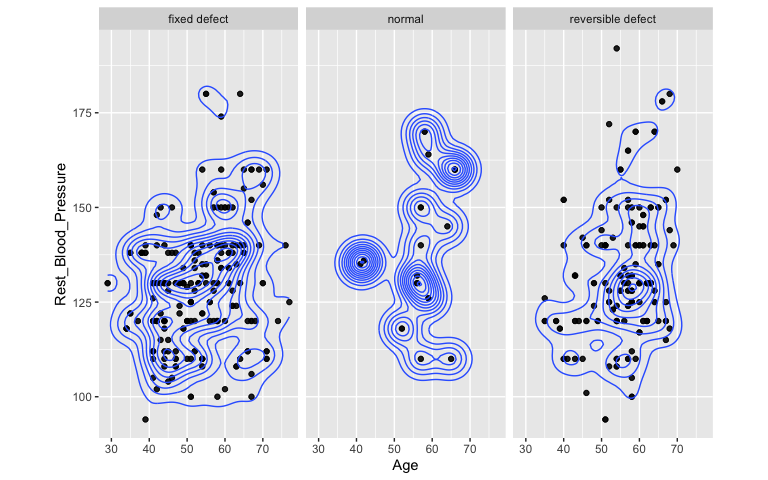
#Type 1 error rate  
print(1-0.95^14)

## [1] 0.512325

#Bonferroni Correction  
print(0.05/14)

## [1] 0.003571429

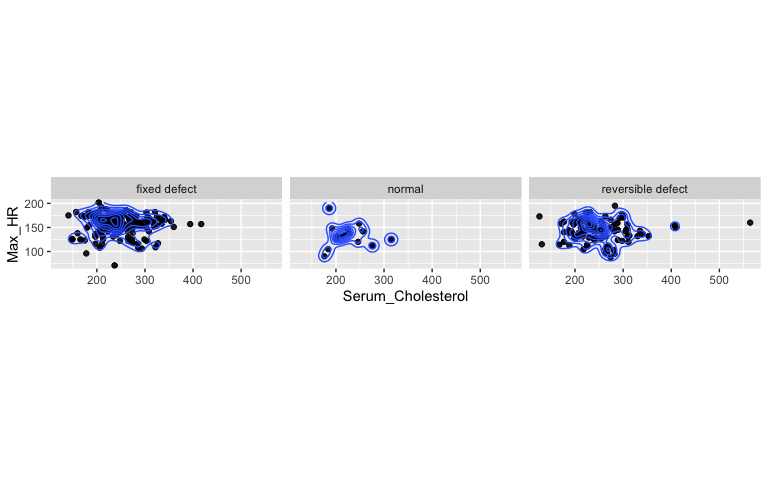
#MANOVA Multivariate Normality Assumption Tests  
ggplot(Heart\_Disease, aes(x = Age, y = Rest\_Blood\_Pressure)) + geom\_point(alpha = .5) + geom\_density\_2d(h=12.98568) + coord\_fixed() + facet\_wrap(~Thalassemia) #multivariate normality assumption



(bandwidth.nrd(Heart\_Disease$Age) + bandwidth.nrd(Heart\_Disease$Rest\_Blood\_Pressure)) / 2 #bandwidth estimate

## [1] 12.98568

ggplot(Heart\_Disease, aes(x = Serum\_Cholesterol, y = Max\_HR)) + geom\_point(alpha = .5) + geom\_density\_2d(h=38.51339) + coord\_fixed() + facet\_wrap(~Thalassemia) #multivariate normality assumption



(bandwidth.nrd(Heart\_Disease$Serum\_Cholesterol) + bandwidth.nrd(Heart\_Disease$Max\_HR)) / 2 #bandwidth estimate

## [1] 38.51339

#MANOVA Homogeneity of (Co)variances Assumption Tests  
covmats<-Heart\_Disease %>% select(Age, Rest\_Blood\_Pressure, Serum\_Cholesterol, Max\_HR, everything()) %>% group\_by(Thalassemia)%>%do(covs=cov(.[1:4]))  
for(i in 1:3){print(as.character(covmats$Thalassemia[i])); print(covmats$covs[i])}

## [1] "fixed defect"  
## [[1]]  
## Age Rest\_Blood\_Pressure Serum\_Cholesterol Max\_HR  
## Age 98.93646 39.429749 148.85076 -109.534985  
## Rest\_Blood\_Pressure 39.42975 255.951269 170.15679  
8.846254  
## Serum\_Cholesterol 148.85076 170.156786 2465.53871  
-41.558811  
## Max\_HR -109.53498 8.846254 -41.55881 479.592888  
##  
## [1] "normal"  
## [[1]]  
## Age Rest\_Blood\_Pressure Serum\_Cholesterol Max\_HR  
## Age 56.869376 35.45283 -49.44739 -5.919448  
## Rest\_Blood\_Pressure 35.452830 335.63788 48.67017  
-172.275399  
## Serum\_Cholesterol -49.447388 48.67017 1450.71626  
-99.761611  
## Max\_HR -5.919448 -172.27540 -99.76161 522.167634  
##  
## [1] "reversible defect"  
## [[1]]  
## Age Rest\_Blood\_Pressure Serum\_Cholesterol Max\_HR  
## Age 64.29760 38.43641 76.111912 -24.286017  
## Rest\_Blood\_Pressure 38.43641 314.83509 68.458510  
39.376848  
## Serum\_Cholesterol 76.11191 68.45851 3162.203273  
-6.387787  
## Max\_HR -24.28602 39.37685 -6.387787 455.922560

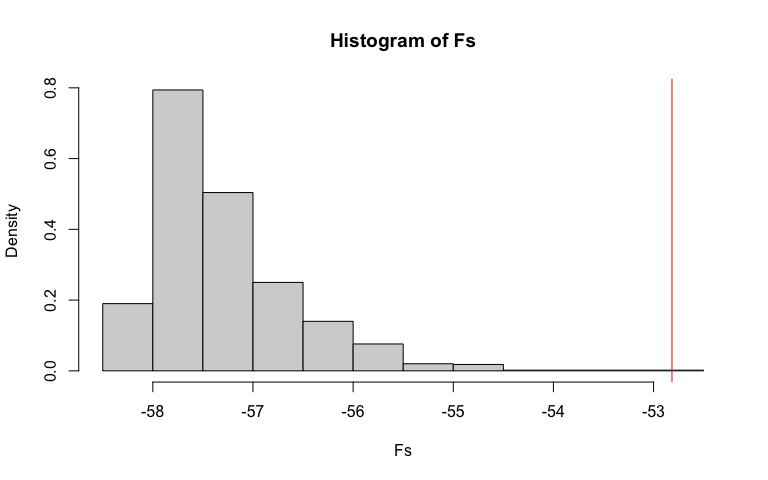
1. Randomization Test:

Considering that a MANOVA test was conducted to determine Thalassemia status based on the 4 numeric variables within the Heart\_Disease dataset led to the discovery of 4 significant differences, a PERMANOVA test should be conducted in place of the MANOVA test so as to replicate the results of the MANOVA without violating the inherent assumptions. PERMANOVA is a simple randomization-test MANOVA that does not carry the annoying and restrictive assumptions of the standard MANOVA test. Therefore, the null hypothesis for our PERMANOVA: For each response variable (Age, Rest\_Blood\_Pressure, Serum\_Cholesterol, Max\_HR), the means of the groups in the Thalassemia variable are equal. The alternative hypothesis for our PERMANOVA: For at least 1 response variable (Age, Rest\_Blood\_Pressure, Serum\_Cholesterol, Max\_HR), at least 1 group mean in the Thalassemia variable significantly differs. The results of the PERMANOVA through the adonis function in the vegan package give a p-value of 0.001, which is sufficient to be considered significant given the standard alpha-value of 0.05. This means that the null hypothesis can be rejected and we can accept the alternative hypothesis. Though further tests must be run to determine the specific significant difference in means between groups in the Thalassemia variable, we can conclude that there is at least 1 group mean in the Thalassemia variable that significantly differs. The following plot shows the visual representation of the null distribution of F-statistics and the observed F-statistic, which is not within the distribution, indicating a rejection of the null hypothesis.

#PERMANOVA Test  
dists<-Heart\_Disease%>%select(Age, Rest\_Blood\_Pressure, Serum\_Cholesterol, Max\_HR)%>%dist() #compute distances/dissimilarities  
adonis(dists~Thalassemia, data=Heart\_Disease) #perform PERMANOVA on distances/dissimilarities

##  
## Call:  
## adonis(formula = dists ~ Thalassemia, data =  
Heart\_Disease)  
##  
## Permutation: free  
## Number of permutations: 999  
##  
## Terms added sequentially (first to last)  
##  
## Df SumsOfSqs MeanSqs F.Model R2 Pr(>F)  
## Thalassemia 2 82806 41403 11.746 0.02436 0.001 \*\*\*  
## Residuals 941 3316788 3525 0.97564  
## Total 943 3399594 1.00000  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1  
' ' 1

#Plot visualizing null distribution and F test statistic  
#compute observed F  
SST<- sum(dists^2)/944  
SSW<-Heart\_Disease %>% select(Age, Rest\_Blood\_Pressure, Serum\_Cholesterol, Max\_HR, everything()) %>% group\_by(Thalassemia) %>% select(Age, Rest\_Blood\_Pressure, Serum\_Cholesterol, Max\_HR) %>% do(d=dist(.[1:4],"euclidean")) %>% ungroup() %>% summarize(sum(d[[1]]^2)/517 + sum(d[[2]]^2)/53+ sum(d[[3]]^2)/374)%>%pull  
F\_obs<-((SST-SSW)/2)/(SSW/941) #observed F statistic  
  
# compute null distribution for F  
Fs<-replicate(1000,{  
new<-Heart\_Disease%>% select(Age, Rest\_Blood\_Pressure, Serum\_Cholesterol, Max\_HR, everything()) %>%mutate(Thalassemia=sample(Thalassemia)) #permute the Thalassemia vector  
SSW<-new%>%group\_by(Thalassemia)%>%select(Age, Rest\_Blood\_Pressure, Serum\_Cholesterol, Max\_HR)%>%do(d=dist(.[1:4],"euclidean"))%>%ungroup()%>%  
 summarize(sum(d[[1]]^2)/517 + sum(d[[2]]^2)/53+ sum(d[[3]]^2)/374)%>%pull  
((SST-SSW)/2)/(SSW/941) #calculate new F on randomized data  
})  
  
#plot  
{hist(Fs,prob = T); abline(v=F\_obs, col="red", add=T)}



mean(Fs>F\_obs) #p-value: reject null hypothesis, just like MANOVA!

## [1] 0

1. Linear Regression Model:

The linear regression model created predicts the resting blood pressure of patients from the numeric response variable ‘Age’ and the categorical response variable ‘Sex’ as well as the interaction between the two variables. The coefficients for the model are 131.4553 for the intercept coefficient, -0.6229 for the coefficient of SexMale, 0.5773 for the coefficient of Age\_c, and -0.1314 for the coefficient of SexMale:Age\_c. The intercept indicates that the predicted resting blood pressure for an average aged, female is 131.4553 mmHg. The coefficient of SexMale indicates that when controlling for Age, resting blood pressure among males is 0.6229 mmHg lower than resting blood pressure among females on average. The coefficient of Age\_c indicates that when controlling for sex, for every one year increase in age resting blood pressure increases 0.5773 mmHg. The coefficient of SexMale:Age\_c indicates that the slope for Age\_c on resting blood pressure is 0.1314 lower for males compared to females.

With linear regression models we have to take into account these 4 assumptions:linear relationship between each predictor and response (linearity), independent observations/random sample, normally distributed residuals (normality), and equal variance of points/residuals along regression line (homoskedasticity). We assume that the assumption of independent observations/random sample is met by the researchers who collected the data, but linearity, normality, and homoskedasticity must be checked. Our data failed all the assumptions for linear regression as a scatter plot of our numeric response and explanatory variable indicates minimal linearity through showing a weak relationship, the Shapiro-Wilk test resulted in a p-value of 1.289e-07, leading to a rejection of the null hypothesis of normality, and the Breuch-Pagan test resulted in a p-value of 4.796e-07, also leading to a rejection of the null hypothesis of homoskedasticity.

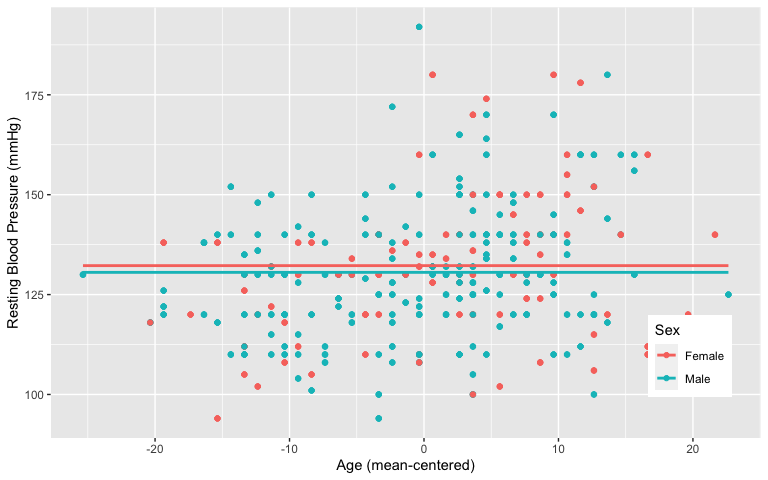
Recomputing regression results using robust standard errors addresses failure to meet heteroskedastic assumptions. This new model indicates that out of the two variables used to predict resting blood pressure, only the mean-centered ‘Age’ variable had a significant effect on resting blood pressure (p-value of 1.048e-07). The ‘Sex’ variable given by the coefficient ‘SexMale’ and the interaction between ‘SexMale’ and ‘Age\_c’ were both not significant in predicting resting blood pressure in patients (p-value of 0.5963 and p-value of 0.2993, respectively). This means that sex did not make a significant difference in resting blood pressure for patients and neither did the interaction. Their incorporation into the model is a result of noise rather than a real relationship. In terms of changes in significance of results before/after robust SEs were applied, there were no changes in significance as the mean-centered variable ‘Age’ was the only predictor that had a significant effect on resting blood pressure.

The proportion of the variation in our response variable (resting blood pressure) explained by the overall model is 7.237%, but when accounting for the random chance that random numbers may have some association with the outcome the proportion of the variation is decreased to 6.941% because R-squared includes a penalty for each extra explanatory variable.

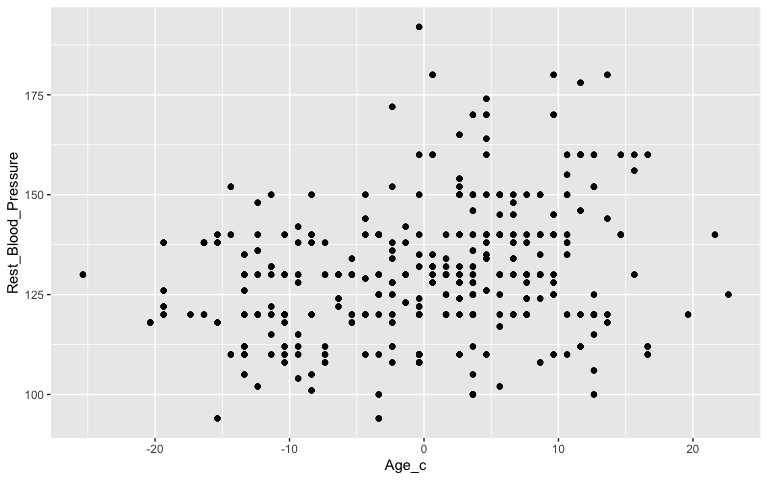
#linear regression model  
Heart\_Disease$Age\_c <- Heart\_Disease$Age - mean(Heart\_Disease$Age) #mean-centering  
linear\_regression\_model <-lm(Rest\_Blood\_Pressure ~ Sex \* Age\_c, data=Heart\_Disease)  
summary(linear\_regression\_model)

##  
## Call:  
## lm(formula = Rest\_Blood\_Pressure ~ Sex \* Age\_c, data =  
Heart\_Disease)  
##  
## Residuals:  
## Min 1Q Median 3Q Max  
## -36.467 -10.942 -0.576 10.475 61.329  
##  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 131.4553 0.9706 135.439 < 2e-16 \*\*\*  
## SexMale -0.6229 1.1650 -0.535 0.593  
## Age\_c 0.5773 0.1010 5.713 1.49e-08 \*\*\*  
## SexMale:Age\_c -0.1314 0.1239 -1.060 0.289  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1  
' ' 1  
##  
## Residual standard error: 16.42 on 940 degrees of freedom  
## Multiple R-squared: 0.07237, Adjusted R-squared: 0.06941  
## F-statistic: 24.44 on 3 and 940 DF, p-value: 3.098e-15

#ggplot  
ggplot(Heart\_Disease, aes(x=Age\_c, y=Rest\_Blood\_Pressure, group=Sex)) + geom\_point(aes(color=Sex)) + geom\_smooth(method="lm",formula=y~1,se=F,fullrange=T,aes(color=Sex)) + theme(legend.position=c(.9,.19)) + xlab("Age (mean-centered)") + ylab("Resting Blood Pressure (mmHg)")



#Assumptions  
#Linearity scatterplot test  
ggplot(Heart\_Disease, aes(x=Age\_c, y=Rest\_Blood\_Pressure))+  
 geom\_point(alpha=1)



#Normality test (Shapiro)  
resids<-lm(Rest\_Blood\_Pressure ~ Sex \* Age\_c, data=Heart\_Disease)$residual  
shapiro.test(resids)

##   
## Shapiro-Wilk normality test  
##   
## data: resids  
## W = 0.98655, p-value = 1.289e-07

#Homoskedasticity test (Breuch-Pagan test)  
bptest(linear\_regression\_model)

##   
## studentized Breusch-Pagan test  
##   
## data: linear\_regression\_model  
## BP = 32.18, df = 3, p-value = 4.796e-07

#Robust standard errors  
coeftest(linear\_regression\_model, vcov = vcovHC(linear\_regression\_model))

##  
## t test of coefficients:  
##  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 131.45527 0.98925 132.8838 < 2.2e-16 \*\*\*  
## SexMale -0.62293 1.17543 -0.5300 0.5963  
## Age\_c 0.57729 0.10770 5.3599 1.048e-07 \*\*\*  
## SexMale:Age\_c -0.13143 0.12655 -1.0385 0.2993  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1  
' ' 1

1. Regression Model w/ Boostrapped Standard Errors:

The bootstrapped standard errors when resampling rows resulted in an intercept of 0.9835272, coefficient SexMale of 1.17607, coefficient Age\_c of 0.1070851, and coefficient SexMale:Age\_c of 0.1262455, with the significant predictors being ‘Age\_c’ (95% confidence interval of [0.3780571, 0.7943142]) and ‘SexMale:Age\_c’ (95% confidence interval of [-0.3852867, 0.1082507]). The bootstrapped standard errors when resampling residuals resulted in an intercept of 0.9624771, coefficient SexMale of 1.157939, coefficient Age\_c of 0.1019327, and coefficient SexMale:Age\_c of 0.126127, with the significant predictors being ‘Age\_c’ (95% confidence interval of [0.3773724, 0.7739855]) and ‘SexMale:Age\_c’ (95% confidence interval of [-0.3738012, 0.1178722 ]). The results between the bootstrapped standard errors when resampling residuals or resampling rows are very similar with only slight differences in the values of the coefficients. However, when compared to the original SEs and robust SEs, the bootstrapped SEs differ in the magnitude of the coefficients and the bootstrapped SEs identify the interaction of ‘SexMale:Age\_c’ as a significant predictor of resting blood pressure, which the original SEs and robust SEs classify as insignificant.

In detail, the original SEs had an intercept of 131.4553, coefficient SexMale of -0.6229, coefficient Age\_c of 0.5773, and coefficient SexMale:Age\_c of -0.1314, with the only significant predictor being ‘Age\_c’ (p-value of 1.49e-08). Likewise, the robust SEs had an intercept of 131.45527, coefficient SexMale of -0.62293, coefficient Age\_c of 0.57729, and coefficient SexMale:Age\_c of -0.13143, with the only significant predictor being ‘Age\_c’ (p-value of 1.048e-07).

#Boostrap standard error (resampling rows)  
samp\_distn<-replicate(5000, {   
 boot\_dat <- sample\_frac(Heart\_Disease, replace=T) #bootstrap your data   
 fit\_rows <- lm(Rest\_Blood\_Pressure ~ Sex \* Age\_c, data=boot\_dat) #fit model   
 coef(fit\_rows) #save coefs  
 })  
#final bootstrap standard errors  
samp\_distn%>%t%>%as.data.frame%>%summarize\_all(sd)

## (Intercept) SexMale Age\_c SexMale:Age\_c  
## 1 0.9685782 1.154225 0.1067126 0.125589

#Bootstrapped 95% confidence intervals  
samp\_distn%>%t%>%as.data.frame%>%gather%>%group\_by(key)%>%  
 summarize(lower=quantile(value,.025), upper=quantile(value,.975))

## # A tibble: 4 x 3  
## key lower upper  
## <chr> <dbl> <dbl>  
## 1 (Intercept) 130. 133.   
## 2 Age\_c 0.376 0.794  
## 3 SexMale -2.85 1.68   
## 4 SexMale:Age\_c -0.380 0.108

#Bootstrap standard error (resampling residuals)  
fit\_residuals <- lm(Rest\_Blood\_Pressure ~ Sex \* Age\_c, data=Heart\_Disease) #fit model  
resids<-fit\_residuals$residuals #save residuals   
fitted<-fit\_residuals$fitted.values #save yhats   
resid\_resamp<-replicate(5000,{   
 new\_resids<-sample(resids,replace=TRUE) #resample resids w/ replacement   
 Heart\_Disease$new\_y<-fitted+new\_resids #add new resids to yhats to get new "data"   
 fit\_residuals<-lm(new\_y ~ Sex \* Age\_c, data=Heart\_Disease) #refit model   
 coef(fit\_residuals) #save coefficient estimates (b0, b1, etc)  
 })  
#final bootstrap standard errors  
resid\_resamp%>%t%>%as.data.frame%>%summarize\_all(sd)

## (Intercept) SexMale Age\_c SexMale:Age\_c  
## 1 0.9677457 1.151337 0.1003782 0.1215928

#Bootstrapped 95% confidence intervals  
resid\_resamp%>%t%>%as.data.frame%>%gather%>%group\_by(key)%>%  
 summarize(lower=quantile(value,.025), upper=quantile(value,.975))

## # A tibble: 4 x 3  
## key lower upper  
## <chr> <dbl> <dbl>  
## 1 (Intercept) 130. 133.   
## 2 Age\_c 0.377 0.773  
## 3 SexMale -2.86 1.63   
## 4 SexMale:Age\_c -0.372 0.108

1. Logistic Regression:

The logistic regression model gives the following coefficient estimates equal to the log of the odds (logit): 2.457409 for the intercept, -0.005234 for Serum\_Cholesterol (p-value of 0.000846), -1.943493 for Thalassemianormal (p-value of 4.88e-10), and -2.404526 for Thalassemiareversible defect (p-value of < 2e-16). These coefficients tell us that when controlling for thalassemia condition, a one-unit increase in serum cholesterol corresponds to a decrease in the log of the odds of diagnosis and that it has a significant negative impact on odds of diagnosis. Similarly, when controlling for serum cholesterol levels, thalassemia statsuses normal and reversible defect have significantly lower odds of diagnosis. All predictors show significant impact on the odds of diagnosis. After exponentiating coefficients, the new model has the following coefficients: 11.67451955 for the intercept, 0.99478014 for Serum\_Cholesterol, 0.14320289 for Thalassemianormal, and 0.09030831 for Thalassemiareversible defect. The intercept indicates that the odds of a positive diagnosis for patients with a thalassemia status of fixed defect and when serum cholesterol = 0 is 11.67451955. The coefficient for serum cholesterol indicates that when controlling for thalassemia, a one unit increase in serum cholesterol level corresponds to a decrease in the odds of diagnosis by a factor of 0.99478014. The coefficient for Thalassemianormal indicates that when controlling for serum cholesterol level, odds of diagnosis for a patient with a thalassemia status of normal is 0.14320289 times the odds of diagnosis for a patient with a thalassemia status of fixed defect. Similarly, the coefficient for Thalassemiareversible defect indicates that when controlling for serum cholesterol level, odds of diagnosis for a patient with a thalassemia status of reversible defect is 0.09030831 times the odds of diagnosis for a patient with a thalassemia status of fixed defect. The following confusion matrix created from this model gives an accuracy of 76.05932%, a sensitivity (TPR) of 79.39394%, specificity (TNR) of 72.38307%, and a recall (PPV) of 76.01547%. These values indicate that the performance of the model is not stellar, but also not bad. The performance is modest and the AUC value of the ROC curve should provide a better indicator of performance.

The ROC curve plot was generated to visualize the trade-off between sensitivity (true positive rate) and specificty (true negative rate) of our logistic regression model. The resulting area under the curve on the graph summarizes both sensitivity and specificity in one single value that denotes the performance of the model. With a computed AUC value of 0.7960023 that falls in the 0.7-0.8 range, the model is classified as a fair predictor of diagnosis.

After performing a 10-fold CV, the average out-of-sample accuracy was 76.04591%, sensitivity was 79.65353%, and recall was 75.97101%. All of which were similar to the original model’s values for accuracy, sensitivity, and recall. The new out-of-sample AUC was 0.7954738, which is not very different from our original AUC. This means that our original logistic regression model did not have a lot of overfitting.

#logistic regression model  
logistic\_regression\_model<-glm(Diagnosis ~ Serum\_Cholesterol + Thalassemia, data=Heart\_Disease, family = "binomial")  
summary(logistic\_regression\_model)

##  
## Call:  
## glm(formula = Diagnosis ~ Serum\_Cholesterol +  
Thalassemia, family = "binomial",  
## data = Heart\_Disease)  
##  
## Deviance Residuals:  
## Min 1Q Median 3Q Max  
## -1.9230 -0.7517 0.6228 0.7523 2.4300  
##  
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 2.457409 0.409004 6.008 1.88e-09 \*\*\*  
## Serum\_Cholesterol -0.005234 0.001568 -3.337 0.000846 \*\*\*  
## Thalassemianormal -1.943493 0.312319 -6.223 4.88e-10 \*\*\*  
## Thalassemiareversible defect -2.404526 0.162808 -14.769  
< 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1  
' ' 1  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
## Null deviance: 1306.4 on 943 degrees of freedom  
## Residual deviance: 1024.4 on 940 degrees of freedom  
## AIC: 1032.4  
##  
## Number of Fisher Scoring iterations: 4

#interpretation of coefficients  
exp(coef(logistic\_regression\_model))

## (Intercept) Serum\_Cholesterol Thalassemianormal  
## 11.67451955 0.99478014 0.14320289  
## Thalassemiareversible defect  
## 0.09030831

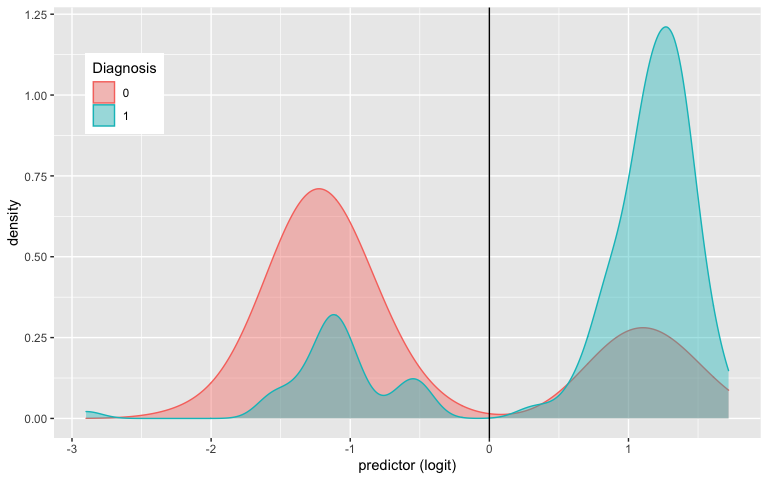
#confusion matrix  
probs <- predict(logistic\_regression\_model, type="response")  
table(predict=as.numeric(probs>.5),truth=Heart\_Disease$Diagnosis)%>%addmargins

## truth  
## predict 0 1 Sum  
## 0 325 102 427  
## 1 124 393 517  
## Sum 449 495 944

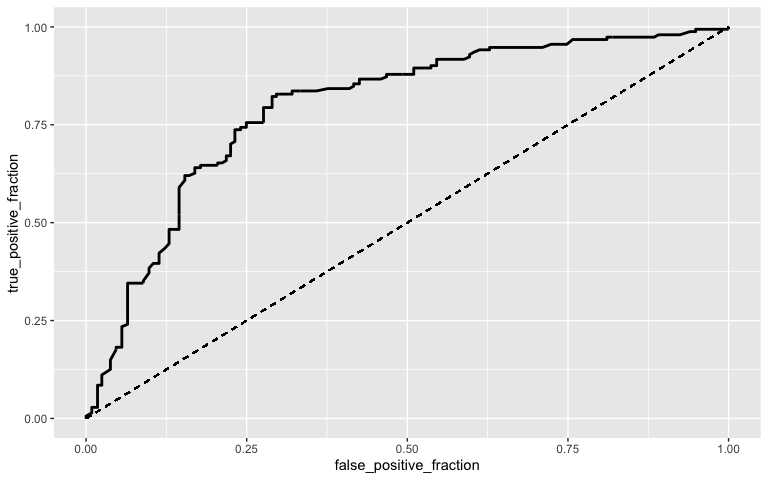
#classification diagnostics  
#class\_diag function  
class\_diag<-function(probs,truth){  
   
 tab<-table(factor(probs>.5,levels=c("FALSE","TRUE")),truth)  
 acc=sum(diag(tab))/sum(tab)  
 sens=tab[2,2]/colSums(tab)[2]  
 spec=tab[1,1]/colSums(tab)[1]  
 ppv=tab[2,2]/rowSums(tab)[2]  
   
 if(is.numeric(truth)==FALSE & is.logical(truth)==FALSE) truth<-as.numeric(truth)-1  
   
 #CALCULATE EXACT AUC  
 ord<-order(probs, decreasing=TRUE)  
 probs <- probs[ord]; truth <- truth[ord]  
   
 TPR=cumsum(truth)/max(1,sum(truth))   
 FPR=cumsum(!truth)/max(1,sum(!truth))  
   
 dup<-c(probs[-1]>=probs[-length(probs)], FALSE)  
 TPR<-c(0,TPR[!dup],1); FPR<-c(0,FPR[!dup],1)  
   
 n <- length(TPR)  
 auc<- sum( ((TPR[-1]+TPR[-n])/2) \* (FPR[-1]-FPR[-n]) )  
   
 data.frame(acc,sens,spec,ppv,auc)  
}  
class\_diag(probs, Heart\_Disease$Diagnosis)

## acc sens spec ppv auc  
## 1 0.7605932 0.7939394 0.7238307 0.7601547 0.7960023

#density plot  
Heart\_Disease$logit<-predict(logistic\_regression\_model,type="link")  
Heart\_Disease%>%mutate(Diagnosis = as.character(Diagnosis)) %>% ggplot(aes(logit,color=Diagnosis,fill=Diagnosis))+geom\_density(alpha=.4)+ theme(legend.position=c(.1,.8))+geom\_vline(xintercept=0)+xlab("predictor (logit)")



#ROC curve  
Heart\_Disease$prob<-predict(logistic\_regression\_model,type="response")  
ROCplot<-ggplot(Heart\_Disease)+geom\_roc(aes(d=Diagnosis,m=prob), n.cuts=0)+  
 geom\_segment(aes(x=0,xend=1,y=0,yend=1),lty=2)  
ROCplot



#AUC calculation  
calc\_auc(ROCplot)

## PANEL group AUC  
## 1 1 -1 0.7960023

#10-fold CV   
k=10  
data <- Heart\_Disease %>% sample\_frac #put rows of dataset in random order  
folds <- ntile(1:nrow(data),n=10) #create fold labels  
diags<-NULL  
for(i in 1:k){  
 train <- data[folds!=i,] #create training set (all but fold i)  
 test <- data[folds==i,] #create test set (just fold i)  
 truth <- test$Diagnosis #save truth labels from fold i  
   
 fit <- glm(Diagnosis~Serum\_Cholesterol + Thalassemia, data=Heart\_Disease, family="binomial")  
 probs <- predict(fit, newdata=test, type="response")  
   
 diags<-rbind(diags,class\_diag(probs,truth))  
}  
summarize\_all(diags,mean)

## acc sens spec ppv auc  
## 1 0.7607167 0.7926435 0.7248184 0.7598495 0.7975736

1. Lasso:

After running LASSO regression on our dataset to predict the binary variable ‘Diagnosis’ with lambda.1se to give the simplest model whose accuracy is near the best, 9 variables were retained. Those variables were ‘Age,’ ‘SexFemale,’ ‘Rest\_Blood\_Pressure,’ ‘Serum\_Cholesterol,’ ‘Fasting\_Blood\_Sugar> 120 mg/dL,’ ‘Rest\_ECGST-T wave abnormality,’ ‘Max\_HR, ’Exercise\_Induced\_AnginaYes,’ ‘ST\_Slopeupsloping,’ and ‘Thalassemiareversible defect.’ These 9 variables were incorporated into a new model and a 10-fold CV was run. The resulting accuracy of the new model was 81.24748%, which is increased from the 76.04591% accuracy of the out-of-sample 10-fold cv originally run with only serum cholesterol levels and thalassemia status. This means that the new model will make more accurate predictions on new data than the past one because appropriate response variables were incorporated and unnecessary ones were removed.

#lasso  
y<-as.matrix(Heart\_Disease$Diagnosis) #grab response  
x<-model.matrix(Diagnosis~-1+.,data=Heart\_Disease) #grab predictors  
cv <- cv.glmnet(x,y, family="binomial") #picks an optimal value for lambda through 10-fold CV  
lasso<-glmnet(x,y,family="binomial",lambda=cv$lambda.1se)  
coef(lasso)

## 17 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) -2.8431927396  
## Age -0.0055562711  
## SexFemale 0.6143164855  
## SexMale .   
## Rest\_Blood\_Pressure -0.0030386168  
## Serum\_Cholesterol .   
## Fasting\_Blood\_Sugar> 120 mg/dL .   
## Rest\_ECGnormal .   
## Rest\_ECGST-T wave abnormality 0.2890961721  
## Max\_HR 0.0184675505  
## Exercise\_Induced\_AnginaYes -0.8280534135  
## ST\_Slopeupsloping -0.5412844809  
## Thalassemianormal .   
## Thalassemiareversible defect -0.2418713150  
## Age\_c -0.0007646219  
## logit .   
## prob 2.2923721775

#10-fold cv with new variables  
Heart\_Disease$SexFemale <-ifelse(Heart\_Disease$Sex=="Female",1,0)  
Heart\_Disease$Rest\_ECG\_ST\_Twaveabnormality <-ifelse(Heart\_Disease$Rest\_ECG=="ST-Twaveabnormality",1,0)  
Heart\_Disease$Exercise\_Induced\_AnginaYes <-ifelse(Heart\_Disease$Exercise\_Induced\_Angina=="Yes",1,0)  
Heart\_Disease$ST\_Slopeupsloping <-ifelse(Heart\_Disease$ST\_Slope=="upsloping",1,0)  
Heart\_Disease$Thalassemiareversibledefect <-ifelse(Heart\_Disease$Thalassemia=="reversible defect",1,0)  
  
k=10  
data <- Heart\_Disease %>% sample\_frac #put rows of dataset in random order  
folds <- ntile(1:nrow(data),n=10) #create fold labels  
diags<-NULL  
for(i in 1:k){  
 train <- data[folds!=i,] #create training set (all but fold i)  
 test <- data[folds==i,] #create test set (just fold i)  
 truth <- test$Diagnosis #save truth labels from fold i  
   
 fit <- glm(Diagnosis~Age + SexFemale +   
 Rest\_Blood\_Pressure + Serum\_Cholesterol +   
 Rest\_ECG\_ST\_Twaveabnormality +   
 Max\_HR + Exercise\_Induced\_AnginaYes +   
 ST\_Slopeupsloping +   
 Thalassemiareversibledefect, data=Heart\_Disease, family="binomial")  
 probs <- predict(fit, newdata=test, type="response")  
   
 diags<-rbind(diags,class\_diag(probs,truth))  
}  
summarize\_all(diags,mean)

## acc sens spec ppv auc  
## 1 0.8124748 0.8174455 0.8062902 0.8247129 0.8838692

# Modeling

## Instructions

A knitted R Markdown document (as a PDF) and the raw R Markdown file (as .Rmd) should both be submitted to Canvas by 11:59pm on 5/1/2020. These two documents will be graded jointly, so they must be consistent (i.e., don’t change the R Markdown file without also updating the knitted document). Knit an html copy too, for later! In the .Rmd file for Project 2, you can copy the first code-chunk into your project .Rmd file to get better formatting. Notice that you can adjust the opts\_chunk$set(…) above to set certain parameters if necessary to make the knitting cleaner (you can globally set the size of all plots, etc). You can copy the set-up chunk in Project2.Rmd: I have gone ahead and set a few for you (such as disabling warnings and package-loading messges when knitting)!

Like before, I envision your written text forming something of a narrative structure around your code/output. All results presented must have corresponding code. Any answers/results/plots etc. given without the corresponding R code that generated the result will not be graded. Furthermore, all code contained in your final project document should work properly. Please do not include any extraneous code or code which produces error messages. (Code which produces warnings is acceptable, as long as you understand what the warnings mean).

## Find data:

Find one dataset with at least 5 variables that wish to use to build models. At least one should be categorical (with 2-5 groups) and at least two should be numeric. Ideally, one of your variables will be binary (if not, you will need to create one by discretizing a numeric, which is workable but less than ideal). You will need a minimum of 40 observations (*at least* 10 observations for every explanatory variable you have, ideally 20+ observations/variable).

It is perfectly fine to use either dataset (or the merged dataset, or a subset of your variables) from Project 1. However, you could also diversify your portfolio a bit by choosing a different dataset to work with (particularly if the variables did not reveal interesting associations in Project 1). The only requirement/restriction is that you may not use data from any examples we have done in class or lab. It would be a good idea to pick more cohesive data this time around (i.e., variables that you actually thing might have a relationship you would want to test). Think more along the lines of your Biostats project.

Again, you can use data from anywhere you want (see bottom for resources)! If you want a quick way to see whether a built-in (R) dataset has binary and/or character (i.e., categorical) variables, check out this list: <https://vincentarelbundock.github.io/Rdatasets/datasets.html>.

## Guidelines and Rubric

* **0. (5 pts)** Introduce your dataset and each of your variables (or just your main variables if you have lots) in a paragraph. What are they measuring? How many observations?
* **1. (15 pts)** Perform a MANOVA testing whether any of your numeric variables (or a subset of them, if including them all doesn’t make sense) show a mean difference across levels of one of your categorical variables (3). If they do, perform univariate ANOVAs to find response(s) showing a mean difference across groups (3), and perform post-hoc t tests to find which groups differ (3). Discuss the number of tests you have performed, calculate the probability of at least one type I error (if unadjusted), and adjust the significance level accordingly (bonferroni correction) before discussing significant differences (3). Briefly discuss assumptions and whether or not they are likely to have been met (2).
* **2. (10 pts)** Perform some kind of randomization test on your data (that makes sense). This can be anything you want! State null and alternative hypotheses, perform the test, and interpret the results (7). Create a plot visualizing the null distribution and the test statistic (3).
* **3. (35 pts)** Build a linear regression model predicting one of your response variables from at least 2 other variables, including their interaction. Mean-center any numeric variables involved in the interaction.
  + Interpret the coefficient estimates (do not discuss significance) (10)
  + Plot the regression using ggplot(). If your interaction is numeric by numeric, refer to code near the end of WS15 to make the plot. If you have 3 or more predictors, just chose two to plot for convenience. (8)
  + Check assumptions of linearity, normality, and homoskedasticity either graphically or using a hypothesis test (4)
  + Regardless, recompute regression results with robust standard errors via coeftest(..., vcov=vcovHC(...)). Discuss significance of results, including any changes from before/after robust SEs if applicable. (8)
  + What proportion of the variation in the outcome does your model explain? (4)
* **4. (5 pts)** Rerun same regression model (with interaction), but this time compute bootstrapped standard errors. Discuss any changes you observe in SEs and p-values using these SEs compared to the original SEs and the robust SEs)
* **5. (40 pts)** Perform a logistic regression predicting a binary categorical variable (if you don’t have one, make/get one) from at least two explanatory variables (interaction not necessary).
  + Interpret coefficient estimates in context (10)
  + Report a confusion matrix for your logistic regression (2)
  + Compute and discuss the Accuracy, Sensitivity (TPR), Specificity (TNR), and Recall (PPV) of your model (5)
  + Using ggplot, plot density of log-odds (logit) by your binary outcome variable (3)
  + Generate an ROC curve (plot) and calculate AUC (either manually or with a package); interpret (10)
  + Perform 10-fold (or repeated random sub-sampling) CV and report average out-of-sample Accuracy, Sensitivity, and Recall (10)
* **6. (10 pts)** Choose one variable you want to predict (can be one you used from before; either binary or continuous) and run a LASSO regression inputting all the rest of your variables as predictors. Choose lambda to give the simplest model whose accuracy is near that of the best (i.e., lambda.1se). Discuss which variables are retained. Perform 10-fold CV using this model: if response in binary, compare model’s out-of-sample accuracy to that of your logistic regression in part 5; if response is numeric, compare the residual standard error (at the bottom of the summary output, aka RMSE): lower is better fit!

## Where do I find data again?

You can choose ANY datasets you want that meet the above criteria for variables and observations. You can make it as serious as you want, or not, but keep in mind that you will be incorporating this project into a portfolio webpage for your final in this course, so choose something that really reflects who you are, or something that you feel will advance you in the direction you hope to move career-wise, or something that you think is really neat, or whatever. On the flip side, regardless of what you pick, you will be performing all the same tasks, so it doesn’t end up being that big of a deal.

If you are totally clueless and have no direction at all, log into the server and type

data(package = .packages(all.available = TRUE))

This will print out a list of **ALL datasets in ALL packages** installed on the server (a ton)! Scroll until your eyes bleed! Actually, do not scroll that much… To start with something more manageable, just run the command on your own computer, or just run data() to bring up the datasets in your current environment. To read more about a dataset, do ?packagename::datasetname.

If it is easier for you, and in case you don’t have many packages installed, a list of R datasets from a few common packages (also downloadable in CSV format) is given at the following website: <https://vincentarelbundock.github.io/Rdatasets/datasets.html>.

* A good package to download for fun/relevant data is fivethiryeight. Run install.packages("fivethirtyeight"), load the packages with library(fivethirtyeight), run data(), and then scroll down to view the datasets. Here is an online list of all 127 datasets (with links to the 538 articles). Lots of sports, politics, current events, etc.
* If you have already started to specialize (e.g., ecology, epidemiology) you might look at discipline-specific R packages (vegan, epi, respectively). We will be using some tools from these packages later in the course, but they come with lots of data too, which you can explore according to the directions above
* However, you *emphatically DO NOT* have to use datasets available via R packages! In fact, I would much prefer it if you found the data from completely separate sources and brought them together (a much more realistic experience in the real world)! You can even reuse data from your SDS328M project, provided it shares a variable in common with other data which allows you to merge the two together (e.g., if you still had the timestamp, you could look up the weather that day: <https://www.wunderground.com/history/>). If you work in a research lab or have access to old data, you could potentially merge it with new data from your lab!
* Here is a curated list of interesting datasets (read-only spreadsheet format): <https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit>
* Here is another great compilation of datasets: <https://github.com/rfordatascience/tidytuesday>
* Here is the UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/index.php>
  + See also <https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research#Biological_data>
* Here is another good general place to look: <https://www.kaggle.com/datasets>
* To help narrow your search down or to see interesting variable ideas, check out <https://www.tylervigen.com/spurious-correlations>. This is the spurious correlations website, and it is fun, but if you look at the bottom of each plot you will see sources for the data. This is a good place to find very general data (or at least get a sense of where you can scrape data together from)!
* If you are interested in medical data, check out www.countyhealthrankings.org
* If you are interested in scraping UT data, they make *loads* of data public (e.g., beyond just professor CVs and syllabi). Check out all the data that is available in the statistical handbooks: <https://reports.utexas.edu/statistical-handbook>

##### Broader data sources:

[Data.gov](www.data.gov) 186,000+ datasets!

[Social Explorer](Social%20Explorer) is a nice interface to Census and American Community Survey data (more user-friendly than the government sites). May need to sign up for a free trial.

[U.S. Bureau of Labor Statistics](www.bls.gov)

[U.S. Census Bureau](www.census.gov)

[Gapminder](www.gapminder.org/data), data about the world.

…