Project 2

Kathleen Strybos

4/20/2020

###Project 2: Modeling, Testing and Predicting on Data for Food Insecurity in Texas Counties

#Introduction

*The dataset that was chosen for this project is centered around health across counties in Texas. The dataset contains 12 different variables for different health measures across counties in Texas. A few of these variables include average life expectancy, the rate of diabetes prevalence, rates of both food insecurity across the county and the rate of limited food access in the counties, average household incomes, free/reduced lunch rates among school-aged children, urbanization status indicators of whether the counties are rural or urban, and the border status of the counties, which indicates whether or not they are on the border. This data was acquired at the County Health Rankings and Roadmaps website and funded by the Robert Wood Johnson Foundation Program. Overall there are 254 observations, to signify how many counties there are in Texas.*

#MANOVA

library(tidyverse)  
library(dplyr)  
library(ggplot2)  
library(lmtest)  
library(sandwich)  
library(plotROC)  
joindata<-read\_csv("joindata.csv")  
  
#MANOVA  
  
man1<-manova(cbind(lifeexpectancy,foodinsecurerate, diabeticrate,limitedaccessfoodrate,   
 Household.Income, free\_reduced\_lunch\_percent,  
 costpermeal)~urbanization, data=joindata)  
summary(man1)

## Df Pillai approx F num Df den Df Pr(>F)   
## urbanization 1 0.32767 15.874 7 228 < 2.2e-16 \*\*\*  
## Residuals 234   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Homogeneity of Variances  
  
covmats<-joindata%>%group\_by(urbanization)%>%drop\_na()%>%do(covs=cov(.[2:8]))  
for(i in 1:3){print(as.character(covmats$urbanization[i])); print(covmats$covs[i])}

## [1] "Rural"  
## [[1]]  
## lifeexpectancy diabeticrate foodinsecurerate  
## lifeexpectancy 6.35630750 -6.536322e-01 -4.9811102  
## diabeticrate -0.65363217 2.261165e+00 4.0666946  
## foodinsecurerate -4.98111018 4.066695e+00 18.8661919  
## limitedaccessfoodrate 2.63329284 -2.157059e+00 -4.2066192  
## free\_reduced\_lunch\_percent -0.02870968 -1.945706e+00 -5.9684541  
## costpermeal 0.01756598 2.879765e-02 0.1211542  
## Household.Income 2513.97620444 -1.113480e+03 -3231.8512778  
## limitedaccessfoodrate free\_reduced\_lunch\_percent  
## lifeexpectancy 2.633293e+00 -2.870968e-02  
## diabeticrate -2.157059e+00 -1.945706e+00  
## foodinsecurerate -4.206619e+00 -5.968454e+00  
## limitedaccessfoodrate 6.638014e+01 2.449074e+01  
## free\_reduced\_lunch\_percent 2.449074e+01 9.200310e+01  
## costpermeal -3.053687e-01 -4.528550e-01  
## Household.Income -1.819561e+04 -4.767283e+04  
## costpermeal Household.Income  
## lifeexpectancy 0.01756598 2513.9762  
## diabeticrate 0.02879765 -1113.4801  
## foodinsecurerate 0.12115417 -3231.8513  
## limitedaccessfoodrate -0.30536866 -18195.6123  
## free\_reduced\_lunch\_percent -0.45285505 -47672.8320  
## costpermeal 0.04267863 539.1542  
## Household.Income 539.15419564 55909302.1399  
##   
## [1] "Urban"  
## [[1]]  
## lifeexpectancy diabeticrate foodinsecurerate  
## lifeexpectancy 5.690944e+00 -1.501944e+00 -3.262778e+00  
## diabeticrate -1.501944e+00 2.294444e+00 2.027778e+00  
## foodinsecurerate -3.262778e+00 2.027778e+00 1.286111e+01  
## limitedaccessfoodrate 5.791806e+00 1.431944e+00 -8.847222e-01  
## free\_reduced\_lunch\_percent -5.079444e+00 5.156944e+00 2.415278e+00  
## costpermeal 6.335231e-02 -5.189815e-02 6.832407e-02  
## Household.Income 1.363101e+04 -7.869583e+03 -9.214233e+03  
## limitedaccessfoodrate free\_reduced\_lunch\_percent  
## lifeexpectancy 5.791806e+00 -5.079444e+00  
## diabeticrate 1.431944e+00 5.156944e+00  
## foodinsecurerate -8.847222e-01 2.415278e+00  
## limitedaccessfoodrate 6.959444e+01 4.611944e+01  
## free\_reduced\_lunch\_percent 4.611944e+01 2.117944e+02  
## costpermeal -2.170648e-01 -1.500023e+00  
## Household.Income -4.476632e+04 -1.697131e+05  
## costpermeal Household.Income  
## lifeexpectancy 0.06335231 13631.010  
## diabeticrate -0.05189815 -7869.583  
## foodinsecurerate 0.06832407 -9214.233  
## limitedaccessfoodrate -0.21706481 -44766.321  
## free\_reduced\_lunch\_percent -1.50002315 -169713.133  
## costpermeal 0.03638994 1321.786  
## Household.Income 1321.78619444 195357053.075  
##   
## [1] NA  
## [[1]]  
## NULL

#Multivariate Normality  
nrow(filter(joindata, urbanization=="Rural"))

## [1] 172

nrow(filter(joindata, urbanization=="Urban"))

## [1] 82

#ANOVA  
summary.aov(man1)

## Response lifeexpectancy :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## urbanization 1 52.16 52.164 8.5113 0.003873 \*\*  
## Residuals 234 1434.15 6.129   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response foodinsecurerate :  
## Df Sum Sq Mean Sq F value Pr(>F)  
## urbanization 1 13.4 13.446 0.7998 0.3721  
## Residuals 234 3934.3 16.813   
##   
## Response diabeticrate :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## urbanization 1 23.42 23.4242 10.307 0.001511 \*\*  
## Residuals 234 531.77 2.2725   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response limitedaccessfoodrate :  
## Df Sum Sq Mean Sq F value Pr(>F)  
## urbanization 1 21.8 21.784 0.3228 0.5705  
## Residuals 234 15790.1 67.479   
##   
## Response Household.Income :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## urbanization 1 8.9697e+09 8969747764 86.594 < 2.2e-16 \*\*\*  
## Residuals 234 2.4239e+10 103583747   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response free\_reduced\_lunch\_percent :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## urbanization 1 4954.7 4954.7 37.265 4.248e-09 \*\*\*  
## Residuals 234 31112.0 133.0   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response costpermeal :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## urbanization 1 0.1439 0.143912 3.5509 0.06075 .  
## Residuals 234 9.4837 0.040529   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## 18 observations deleted due to missingness

#Probability of at least one Type 1 Error  
type1overall<-1-(1-0.05)^8  
type1overall

## [1] 0.3365796

#Bonferroni Correction  
bonferroni<-0.05/8  
bonferroni

## [1] 0.00625

*There is a significant effect, with a p-value of 2.2e-16, in urbanization when the MANOVA is performed, meaning there is a mean difference across urbanization status for one or some of the numeric response variables. However, we cannot decipher which variable is showing a mean difference with a MANOVA, so univariate ANOVAs were used to narrow this down. The assumptions for MANOVA that were likely met include the assumption for multivariate normality because there were 172 samples for Rural counties taken and 82 samples for Urban counties that were taken, which is more than 25 in each category. Homogeneity of covariances does not appear to be met because the covariances between urban and rural counties differ largely in a few areas, with some univariate and multivariate outliers, but this might be passable for the MANOVA, as the assumptions are difficult to meet.*

*Univariate ANOVAs were performed to find responses showing a mean difference across groups, and life expectancy (p-value of 0.003873), diabetic rate (p-value of 0.001511), household income (p-value of 2.2e-16) and free and reduced lunch percent (p-value of 0.000000004248) all showed significance, meaning that the mean differences by urbanization status differed. For an ANOVA, we can assume that the assumptions of random sampling, normal distribution of the groups and equal variance of each group have been met, as the sample sizes contain independent responses for every county in Texas, and each variable contains more than 25 samples within each group.*

*There was one MANOVA performed, and seven ANOVAs so overall 8 tests were performed.The probability that I made at least one Type-1 error is 0.3365796. The boneferroni adjusted significance level that I will use to keep the overall type I error rate at .05 is 0.00625. All four previously significant tests showed a significant difference even after Bonferroni adjustment for life expectancy, diabetic rate, household income, and free/reduced lunch percentage, as their p-values were much smaller than even the Bonferroni adjusted p-value. No Post-Hoc tests were done because there were only two categories within “urbanization,” so it was visible after doing ANOVAs where the differences were between groups.*

## Randomization Test

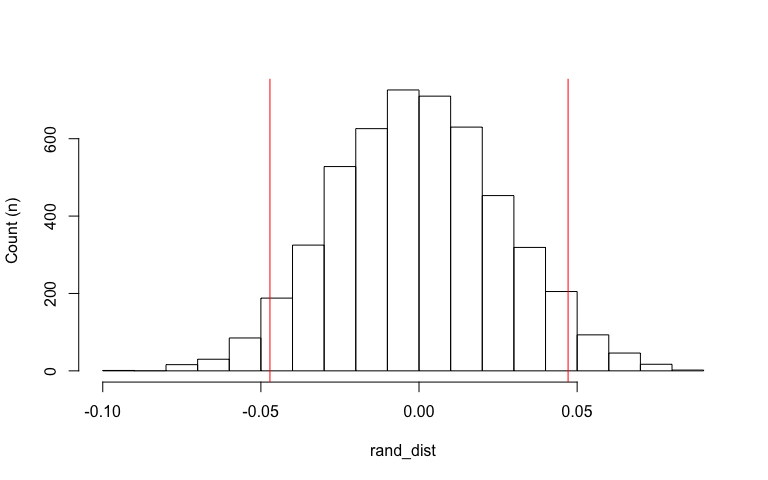
#Mean Difference in Cost Per Meal  
joindata%>%group\_by(urbanization)%>%  
 summarize(means=mean(costpermeal))%>%summarize(`mean\_diff:`=diff(means))

## # A tibble: 1 x 1  
## `mean\_diff:`  
## <dbl>  
## 1 0.0471

#Permutations  
set.seed(1234)  
rand\_dist<-vector()  
for(i in 1:5000){  
new<-data.frame(costpermeal=sample(joindata$costpermeal),  
 urbanization=joindata$urbanization)  
rand\_dist[i]<-mean(new[new$urbanization=="Urban",]$costpermeal)-  
 mean(new[new$urbanization=="Rural",]$costpermeal)}  
mean(rand\_dist>0.04713556| rand\_dist< -0.04713556)

## [1] 0.0774

#Histogram of the Null Distribution and Test Statistic  
{hist(rand\_dist,main="", ylab = "Count (n)",); abline(v=-0.04713556, col="red");   
 abline(v=0.04713556, col="red")}



#Comparison to t-test  
t.test(data=joindata, costpermeal~urbanization)

##   
## Welch Two Sample t-test  
##   
## data: costpermeal by urbanization  
## t = -1.8102, df = 168.89, p-value = 0.07205  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.098540332 0.004269203  
## sample estimates:  
## mean in group Rural mean in group Urban   
## 2.851279 2.898415

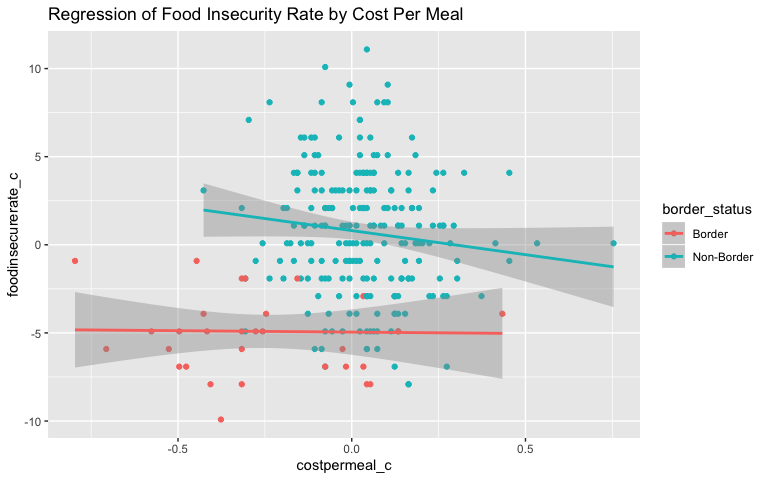
*The null hypothesis of this randomization test is that the mean cost per meal is the same for urban and rural counties. The alternative hypothesis is that the mean cost per meal is not the same for urban and rural counties. The two tailed p-value of the randomization test came out to be 0.0754, which is not significant in comparison to a reference of 0.05. Moreover, when a Two Sample t-test was done, the p-value was 0.07205, which is also not significant, indicating that the randomization test achieved similar results. We can then fail to reject the null hypothesis that the mean cost per meal is the same for urban and rural counties, due to a large p-value. The plot of the null distribution shows the red line at the positive and negative test statistics, which was 0.04713556.*

#Linear Regression

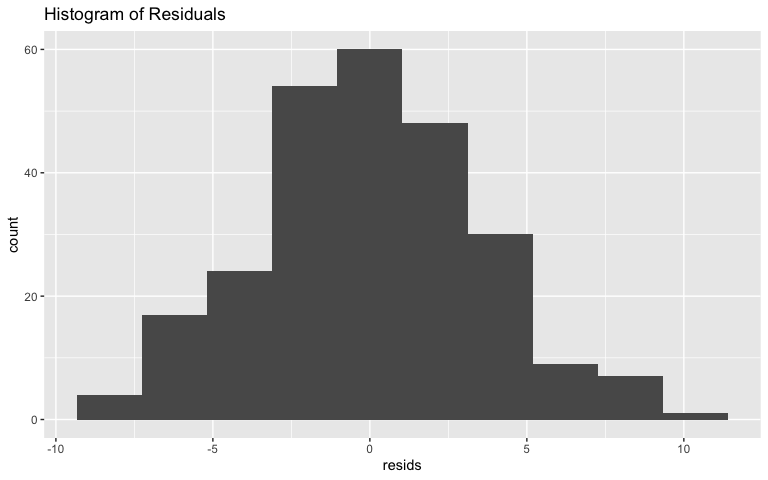
#Linear Regression Model  
joindata$foodinsecurerate\_c<-joindata$foodinsecurerate-mean(joindata$foodinsecurerate,   
 na.rm=T)  
joindata$costpermeal\_c<-joindata$costpermeal-mean(joindata$costpermeal, na.rm=T)  
fit<-lm(foodinsecurerate\_c ~ costpermeal\_c \* border\_status, data=joindata)  
summary(fit)

##   
## Call:  
## lm(formula = foodinsecurerate\_c ~ costpermeal\_c \* border\_status,   
## data = joindata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.2732 -2.4593 -0.0027 2.3783 10.3988   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.9538 0.8647 -5.729 2.90e-08 \*\*\*  
## costpermeal\_c -0.1617 2.3624 -0.068 0.945   
## border\_statusNon-Border 5.7605 0.8991 6.407 7.32e-10 \*\*\*  
## costpermeal\_c:border\_statusNon-Border -2.5713 2.8136 -0.914 0.362   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.574 on 250 degrees of freedom  
## Multiple R-squared: 0.2246, Adjusted R-squared: 0.2153   
## F-statistic: 24.13 on 3 and 250 DF, p-value: 9.478e-14

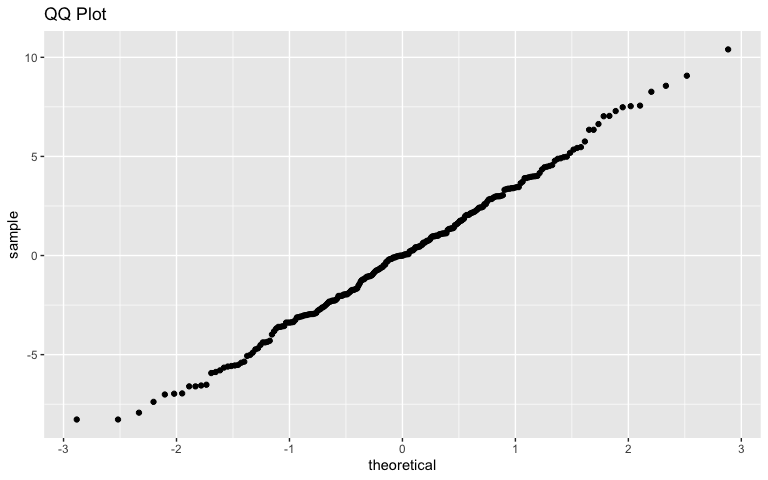
#Regression Plot  
fit%>%ggplot(aes(costpermeal\_c, foodinsecurerate\_c, color=border\_status))+geom\_point()+  
 geom\_smooth(method="lm")+  
 ggtitle("Regression of Food Insecurity Rate by Cost Per Meal")



#Normality Assumption: Data is Normal  
resids<-lm(foodinsecurerate\_c~costpermeal\_c\*border\_status, data=joindata)$residuals  
ggplot()+geom\_histogram(aes(resids),bins=10)+  
 ggtitle("Histogram of Residuals")



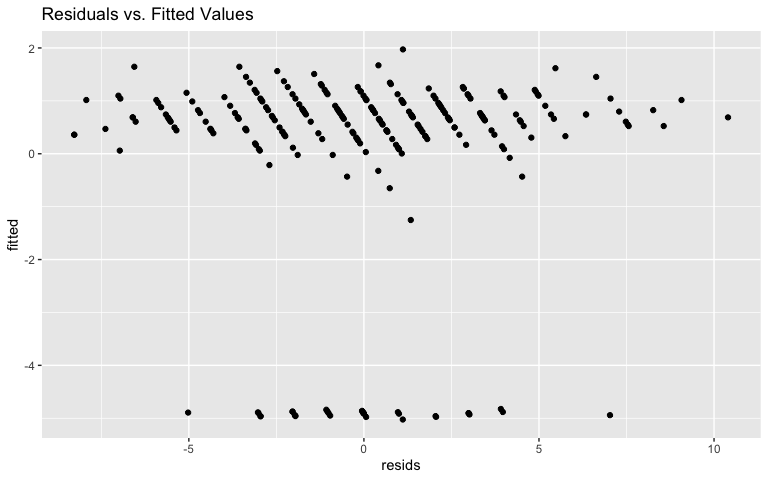
ggplot()+geom\_qq(aes(sample=resids))+geom\_qq()+  
 ggtitle("QQ Plot")



shapiro.test(resids)

##   
## Shapiro-Wilk normality test  
##   
## data: resids  
## W = 0.9957, p-value = 0.7085

#Linearity Assumption  
fitted<-lm(foodinsecurerate\_c~costpermeal\_c\*border\_status, data=joindata)$fitted.values  
ggplot()+geom\_point(aes(resids,fitted))+  
 ggtitle("Residuals vs. Fitted Values")



#Homoskedasticity  
bptest(fit) #Fail to reject the null hypothesis that data is homoskedastic

##   
## studentized Breusch-Pagan test  
##   
## data: fit  
## BP = 5.3442, df = 3, p-value = 0.1483

#Robust Standard Errors: Redo Regression  
coeftest(fit, vcov = vcovHC(fit))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.95379 0.63103 -7.8503 1.219e-13 \*\*\*  
## costpermeal\_c -0.16166 1.70938 -0.0946 0.9247   
## border\_statusNon-Border 5.76048 0.68115 8.4569 2.315e-15 \*\*\*  
## costpermeal\_c:border\_statusNon-Border -2.57129 2.17218 -1.1837 0.2376   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Proportion of the Variation Explained  
summary(fit)$r.sq

## [1] 0.2245676

*Predicted food insecurity rate for an average cost per meal at a border county is 4.9538 percent below the centered mean of the food insecurity rate. Controlling for border status, for every one dollar increase in the centered cost per meal, the food insecurity rate decreases by 0.16166 percent. At an average cost per meal, non-border counties have a 5.76048 percent higher food insecurity rate as compared to border counties. The slope for cost per meal on food insecurity rate is 2.57129 units lower for non-border counties compared to border counties.*

*The data meets the assumption of normality and homoskedasticity, and therefore linearity is also met through homoskedasticity. A Shapiro\_Wilk Normality test was run, along with a histogram and QQ plot to check for normality, and all results confirmed normality of the data. The null hypothesis for homoskedasticity fails to be rejected after running the Breusch-Pagan test, with a p-value of 0.1483, which is greater than 0.05.*

*After redoing the regression using homoskedastic standard errors, significance was maintained for the effect of border status when the cost per meal is controlled for. The standard errors for border status decreased by about two-tenths from 0.8991 to 0.68115, and the standard error for cost per meal decreased from 2.3624 to 1.70938. The standard error for the interaction between border status and cost per meal decreased slightly from 2.8136 to 2.17218 with robust standard errors.*

*The proportion of the variation in the outcome that the model explains, or the R-squared value, is 0.2246, or 22.46%.*

#Linear Regression with Bootstrapped Standard Errors

fit<-lm(foodinsecurerate\_c ~ costpermeal\_c\*border\_status, data=joindata)  
summary(fit)

##   
## Call:  
## lm(formula = foodinsecurerate\_c ~ costpermeal\_c \* border\_status,   
## data = joindata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.2732 -2.4593 -0.0027 2.3783 10.3988   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.9538 0.8647 -5.729 2.90e-08 \*\*\*  
## costpermeal\_c -0.1617 2.3624 -0.068 0.945   
## border\_statusNon-Border 5.7605 0.8991 6.407 7.32e-10 \*\*\*  
## costpermeal\_c:border\_statusNon-Border -2.5713 2.8136 -0.914 0.362   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.574 on 250 degrees of freedom  
## Multiple R-squared: 0.2246, Adjusted R-squared: 0.2153   
## F-statistic: 24.13 on 3 and 250 DF, p-value: 9.478e-14

#Bootstrapping  
set.seed(1234)  
samp\_distn<-replicate(5000, {  
boot\_dat <- sample\_frac(joindata, replace=T)   
fit <- lm(foodinsecurerate\_c ~ costpermeal\_c\*border\_status, data=boot\_dat)  
coef(fit)  
})  
samp\_distn %>% t %>% as.data.frame %>% summarize\_all(sd)

## (Intercept) costpermeal\_c border\_statusNon-Border  
## 1 0.6358113 1.635783 0.6828607  
## costpermeal\_c:border\_statusNon-Border  
## 1 2.110855

*The bootstrapped standard errors shows similar standard errors that were seen in the robust standard errors of the linear regression, however they are very different than the original standard errors from the linear regression. The bootstrapped SE for costpermeal\_c is 1.673076, which is similar to the robust SE of 1.70938, but different from the original SE of 2.3624. Because the bootstrapped and robust SEs went down, this means the p-values got smaller as well. The bootstrapped SE for border status is 0.6910803, which is similar to the robust SE of 0.68115, and smaller than the original SE of 0.8991, meaning the p-value is smaller for the adjusted SEs. The bootstrapped SE for the interaction is 2.158202, which is similar to the robust SE of 2.17218, but smaller than the original SE of 2.8136, meaning the p-value is smaller for the adjusted SEs too.*

#Logistic Regression

#Logistic Regression  
joindata\_log<- joindata%>%mutate(Urb\_binary=ifelse(urbanization=="Urban",1,0))  
fit<-glm(Urb\_binary ~ limitedaccessfoodrate+Household.Income, data=joindata\_log,  
 family=binomial(link="logit"))  
coeftest(fit)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.1719e+00 9.1239e-01 -6.7645 1.337e-11 \*\*\*  
## limitedaccessfoodrate -6.2741e-03 1.6491e-02 -0.3805 0.7036   
## Household.Income 1.0923e-04 1.7054e-05 6.4047 1.507e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

exp(coef(fit))

## (Intercept) limitedaccessfoodrate Household.Income   
## 0.002087343 0.993745502 1.000109233

#Confusion Matrix  
join\_conf<-joindata\_log%>%drop\_na(limitedaccessfoodrate)%>%mutate(prob=predict(fit,   
 type="response"),   
 prediction=ifelse(prob>0.5,1,0))  
classify<-join\_conf%>%transmute(prob,prediction,truth=Urb\_binary)  
classify

## # A tibble: 253 x 3  
## prob prediction truth  
## <dbl> <dbl> <dbl>  
## 1 0.163 0 0  
## 2 0.665 1 0  
## 3 0.216 0 0  
## 4 0.237 0 1  
## 5 0.541 1 1  
## 6 0.437 0 1  
## 7 0.287 0 1  
## 8 0.619 1 1  
## 9 0.167 0 0  
## 10 0.454 0 1  
## # … with 243 more rows

table(prediction= classify$prediction, truth=classify$truth)%>%addmargins()

## truth  
## prediction 0 1 Sum  
## 0 154 48 202  
## 1 17 34 51  
## Sum 171 82 253

#Accuracy  
(154+34)/253

## [1] 0.743083

#Sensitivity (TPR)  
34/82

## [1] 0.4146341

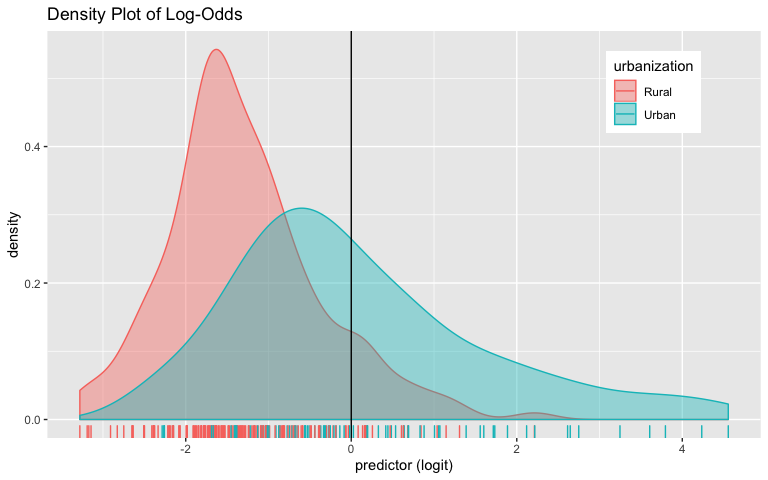
#Specificity (FPR)  
154/171

## [1] 0.9005848

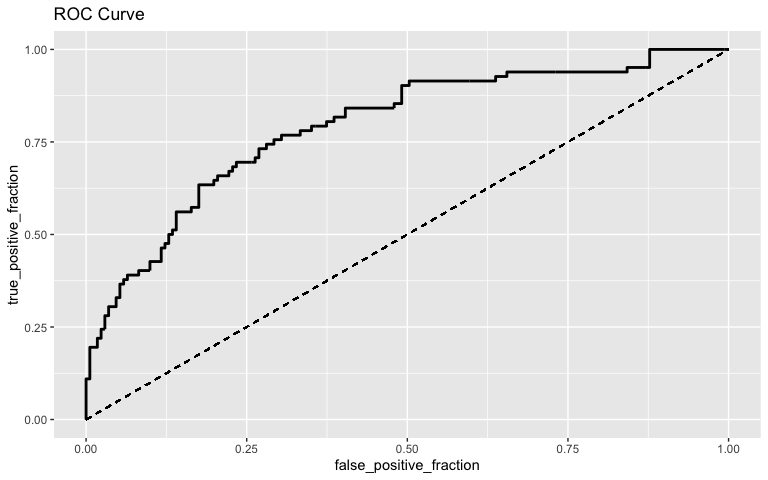
#Precision (PPV)  
34/51

## [1] 0.6666667

#Density Plot of Log-Odds  
join\_conf$logit<-predict(fit,type="link")  
join\_conf%>%ggplot()+geom\_density(aes(logit,color=urbanization,fill=urbanization),   
 alpha=.4)+  
 theme(legend.position=c(.85,.85))+geom\_vline(xintercept=0)+xlab("predictor (logit)")+  
 geom\_rug(aes(logit,color=urbanization)) + ggtitle("Density Plot of Log-Odds")



#ROC Plot  
library(plotROC)  
ROCplot<-ggplot(classify)+geom\_roc(aes(d=truth,m=prob), n.cuts=0)+geom\_segment(aes(x=0,  
 xend=1, y=0, yend=1), lty=2)+ggtitle("ROC Curve")  
ROCplot



#AUC Calculation  
calc\_auc(ROCplot)

## PANEL group AUC  
## 1 1 -1 0.7925403

#10-Fold CV  
set.seed(1234)  
k=10  
  
data1<-joindata\_log[sample(nrow(joindata\_log)),]   
folds<-cut(seq(1:nrow(joindata\_log)),breaks=k,labels=F)   
  
diags<-NULL  
for(i in 1:k){   
 train<-data1[folds!=i,]   
 test<-data1[folds==i,]   
   
 truth<-test$Urb\_binary  
   
 log\_reg<- glm(Urb\_binary ~ limitedaccessfoodrate+Household.Income,  
 data=train,   
 family="binomial")  
 probs<- predict(log\_reg, type = "response", newdata=test)  
   
 diags<-rbind(diags,class\_diag(probs,truth))  
}  
summarize\_all(diags,mean,na.rm=TRUE)

## acc sens spec ppv auc  
## 1 0.7312308 0.4194805 0.8875317 0.6244048 0.7852403

*The odds of urbanization with a limited food access rate of zero and a household income of zero is 0.002087343. Controlling for household income, for every one additional unit increase in the rate of limited access to food, the odds of urbanization decrease by a factor of 0.993745502, but this does not have a significant negative impact. Controlling for limited access to food rate, household income has a significant positive impact on the odds of urbanization. Controlling for limited access to food rate, for every one additional unit of household income,the odds of urbanization increase by a factor of 1.000109233.*

*The accuracy of the model was reported to be 0.743083, which means that true results of classification, either positive or negative are reported 74.31% of the time for the classification of a county as urban or rural. The sensitivity was reported to be 0.4146341, which means that the true positive rate of counties being identified as urban happened 41.46% of the time. The specificity was reported to be 0.9005848, meaning that the true negative rate of counties being identified as rural happened 90.06% of the time. The precision of the model was 0.6666667, meaning that this is the proportion of counties classified as urban that actually are urban in reality.*

*The ROC curve above shows the tradeoff between sensitivity and specificity, and the AUC is a reflection of the summarization of those values. We can see that as the true positive rate increases, the false positive rate increases as well. The calculation of the AUC from the ROC plot gave a value of 0.7925403, which falls within the “fair” range for how well the model of the rate of limited access to food and household income predicts the data of urbanization classification.*

*A 10-fold cross validation was done, and the average out-of-sample accuracy was reported to be 0.7274744, which means that true results, either positive or negative are reported 72.75% of the time for the classification of a county as urban or rural. The sensitivity was reported to be 0.4075536, which means that the true positive rate of counties being identified as urban happened only 40.76% of the time. The specificty was reported to be 0.8796805, which means that correct identification of rural counties happened 87.97% of the time. The PPV was reported to be 0.6665476, meaning that this is the proportion of counties classified as urban that actually are urban in reality. The AUC of the model is 0.7722236, which falls in the classification of “fair” for how well the model predicts the data.*

#LASSO and CV

library(glmnet)  
joindata\_nona<-joindata\_log%>%drop\_na()%>%select(-c(County,   
 foodinsecurerate\_c, costpermeal\_c, urbanization))  
y<-as.matrix(joindata\_nona$Urb\_binary)  
x<-model.matrix(Urb\_binary~.,data=joindata\_nona)[,-1]  
head(x)

## lifeexpectancy diabeticrate foodinsecurerate limitedaccessfoodrate  
## 1 73.8 11 19 15  
## 2 76.8 9 9 12  
## 3 76.3 14 20 11  
## 4 76.9 14 16 20  
## 5 77.0 11 15 5  
## 6 72.4 11 14 20  
## free\_reduced\_lunch\_percent costpermeal Household.Income uninsuredchildrenrate  
## 1 61 2.75 42412 11  
## 2 42 2.94 63451 12  
## 3 66 2.94 45318 10  
## 4 56 2.95 46970 13  
## 5 32 2.83 58311 13  
## 6 51 3.06 55337 12  
## border\_statusNon-Border childfoodinsecurityrate  
## 1 1 0.234  
## 2 1 0.184  
## 3 1 0.253  
## 4 1 0.263  
## 5 1 0.201  
## 6 1 0.203

cv<-cv.glmnet(x,y, family="binomial")  
lasso<-glmnet(x,y,lambda=cv$lambda.1se)  
coef(lasso)

## 11 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) 0.0690545319  
## lifeexpectancy .   
## diabeticrate .   
## foodinsecurerate .   
## limitedaccessfoodrate .   
## free\_reduced\_lunch\_percent .   
## costpermeal .   
## Household.Income 0.0000142346  
## uninsuredchildrenrate -0.0359355608  
## border\_statusNon-Border .   
## childfoodinsecurityrate .

#10-fold CV  
set.seed(1234)  
k=10  
data <- joindata\_nona %>% sample\_frac  
folds <- ntile(1:nrow(data),n=10)  
diags<-NULL  
for(i in 1:k){  
train <- data[folds!=i,]  
test <- data[folds==i,]   
truth <- test$Urb\_binary  
fit <- glm(Urb\_binary~uninsuredchildrenrate+Household.Income,data=train,  
 family="binomial")  
probs <- predict(fit, newdata=test, type="response")  
diags<-rbind(diags,class\_diag(probs,truth))  
}  
diags%>%summarize\_all(mean)

## acc sens spec ppv auc  
## 1 0.8179348 0.6327922 0.9250668 0.825873 0.8799389

*The LASSO regression retained Household Income and Rate of Uninsured Children as the variables that were the best predictors for Urban classification. The 10-fold CV performed on the binary Urban classification variable gave an out-of-sample accuracy of 0.8269928, which is greater than the out-of-sample accuracy in the logisitic regression, which was 0.7274744. The AUC from the 10-fold CV for the LASSO was 0.874748 which is really good, as opposed to an AUC of 0.7722236 for the logistic regression, which is fair. This means that the LASSO regression performs better than a logisitic regression at predicting true results, either positive or negative for the classification of a county as urban or rural, by reducing overfitting of the model.*