Objective2\_Logistic\_Regression

Christina Kuang

4/19/2022

**Question 2) What characteristics of nursing homes in New Mexico help predict if a nursing home is rural or non-rural?**

***Rural patients suffer from a lack of locally available nursing home beds. Understanding the relationships of these characteristics and how they define rural vs. non-rural nursing homes is helpful to know how to make rural nursing homes financially viable.***

**Objective 2: Use characteristics of nursing homes variables to develop a logistic regression model that helps predict whether a nursing home is rural or non-rural**

set.seed(10) ##for reproducibility to get the same split  
sample<-sample.int(nrow(Data), floor(.8\*nrow(Data)), replace = F)  
train<-Data[sample, ] ##training data frame  
test<-Data[-sample, ] ##test data frame

## Data Exploration

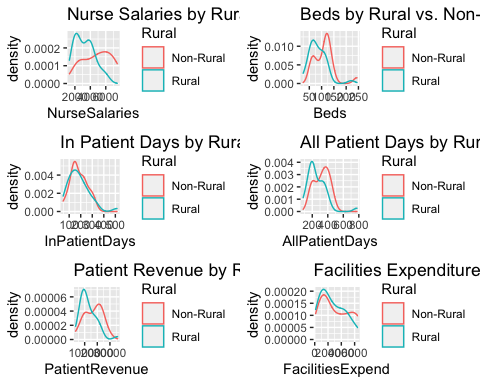
The density plots of nurse salaries for rural is right skewed, which means a higher proportion of rural nursing homes have lower annual nurse salaries. In contrast, the density plots of nurse salaries for non-rural is left skewed, which means a higher proportion of non-rural nursing homes have higher annual nurse salaries.

Similarly, the density plots of patient revenue for rural is right skewed, which means a higher proportion of rural nursing homes have lower annual patient revenues. In contrast, the density plots of nurse salaries for non-rural is left skewed, which means a higher proportion of non-rural nursing homes have higher annual patient revenues.

These two variables, nurse salaries and patient revenue, may be good predictors, because the density plots for rural and non-rural are not very similar.

In comparison, the density plots of beds, patient days, all patient days, and facilities expenditure are similar for rural and non-rural nursing home facilities. As a result, these variables are less likely to be good predictors for whether a nursing home is rural or non-rural.

g1<-ggplot(Data, aes(x=NurseSalaries, color=Rural))+  
 geom\_density()+  
 labs(title="Nurse Salaries by Rural vs. Non-rural")  
  
g2<-ggplot(Data, aes(x=Beds, color=Rural))+  
 geom\_density()+  
 labs(title="Beds by Rural vs. Non-rural")  
g3<-ggplot(Data, aes(x=InPatientDays, color=Rural))+  
 geom\_density()+  
 labs(title="In Patient Days by Rural vs. Non-rural")  
  
g4<-ggplot(Data, aes(x=AllPatientDays, color=Rural))+  
 geom\_density()+  
 labs(title="All Patient Days by Rural vs. Non-rural")  
  
g5<-ggplot(Data, aes(x=PatientRevenue, color=Rural))+  
 geom\_density()+  
 labs(title="Patient Revenue by Rural vs. Non-rural")  
  
g6<-ggplot(Data, aes(x=FacilitiesExpend, color=Rural))+  
 geom\_density()+  
 labs(title="Facilities Expenditure by Rural vs. Non-rural")  
  
grid.arrange(g1,g2,g3,g4,g5,g6, ncol = 2, nrow = 3)



## Full Logistic Regression Model with all predictors

result <- glm(Rural~NurseSalaries+FacilitiesExpend+Beds+PatientRevenue+AllPatientDays+InPatientDays, family = "binomial", data=train)  
summary(result)

##   
## Call:  
## glm(formula = Rural ~ NurseSalaries + FacilitiesExpend + Beds +   
## PatientRevenue + AllPatientDays + InPatientDays, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.9895 -0.6264 0.4284 0.7897 1.5208   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.45546582 1.49436274 2.312 0.0208 \*  
## NurseSalaries -0.00070947 0.00034640 -2.048 0.0406 \*  
## FacilitiesExpend 0.00033007 0.00027613 1.195 0.2319   
## Beds -0.03715240 0.02909603 -1.277 0.2016   
## PatientRevenue 0.00002393 0.00012707 0.188 0.8506   
## AllPatientDays -0.00323788 0.00993208 -0.326 0.7444   
## InPatientDays 0.01613626 0.01076389 1.499 0.1338   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 53.850 on 40 degrees of freedom  
## Residual deviance: 39.768 on 34 degrees of freedom  
## AIC: 53.768  
##   
## Number of Fisher Scoring iterations: 5

## Wald Test for $$1

With a p-value of 0.04, we can reject the null hypothesis. There appears to be a significant relationship between and the response variable.

## 95% onfidence interval for

The 95% confidence interval for is (-0.001413439498, -0.000005500502). In other words, we are 95% confident the odds of a nursing home being rural is between (exp -0.001413439498, exp -0.000005500502) = (0.9985876, 0.9999945) times the odds of a nursing home being non-rural, for given value of other predictors. Since 0 does not lie within the confidence interval, there appears to be a significant effect of nurse salaries on whether a nursing home facility is rural or non-rural, for given values of other predictors. This is consistent with our Wald Test results .

n <- 41  
p <- 7  
se\_beta1 <- 0.00034640  
beta1 <- -0.00070947  
critical <-qt((1-(0.05/2)),(n-p))  
multiplier <- critical\*se\_beta1  
lower\_bound <- beta1 - multiplier  
upper\_bound <- beta1 + multiplier  
print(c(lower\_bound,upper\_bound))

## [1] -0.001413439498 -0.000005500502

## Delta G-Squared Test to see if we can drop predictors (Beds, InPatientDays, AllPatientDays, PatientRevenue, and FacilitiesExpend) that have an insignificant Wald test.

H0: predictors in   
Ha: at least one of the coefficients in is nonzero

With a p-value of 0.35, we fail to reject the null. Since none of the subset predictors appear to be significant, we can drop them from our model.

reduced <- glm(Rural~NurseSalaries, family = "binomial", data=train)  
summary(reduced)

##   
## Call:  
## glm(formula = Rural ~ NurseSalaries, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.9928 -0.8887 0.5521 0.8964 1.5295   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.0935426 1.0668428 2.900 0.00373 \*\*  
## NurseSalaries -0.0006423 0.0002477 -2.593 0.00952 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 53.850 on 40 degrees of freedom  
## Residual deviance: 45.317 on 39 degrees of freedom  
## AIC: 49.317  
##   
## Number of Fisher Scoring iterations: 3

TS <- reduced$deviance-result$deviance  
TS

## [1] 5.54864

1-pchisq(TS,5)

## [1] 0.3526418

## Delta G-squared Test to see if our model is better than the intercept-only model

:   
:

The test statistic is 8.53 with a p-value of 0.014. Since the p-value is less than 0.05, we can reject the null hypothesis. The model is better at prediction than the intercept-only model.

#delta G^2 test to see if coefficients for all predictors are 0. Null deviance in output minus residual deviance in output.   
TS\_2 <- reduced$null.deviance-reduced$deviance  
TS\_2

## [1] 8.533789

1-pchisq(TS\_2,2)

## [1] 0.01402527

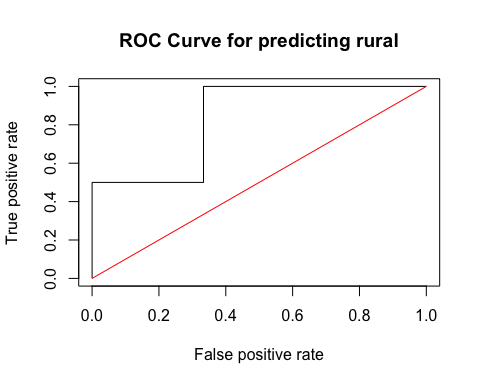
## Final Model

= 3.0935426 -0.0006423(NurseSalaries)

## Validate model using testing data

Since the ROC curve is above the diagonal line and AUC value is greater than 0.5, it tells us that the logistic regression model performs better than random guessing.

##predicted survival rate for test data based on training data  
preds<-predict(reduced,newdata=test, type="response")  
  
##produce the numbers associated with classification table  
rates<-prediction(preds, test$Rural)  
  
##store the true positive and false postive rates  
roc\_result<-performance(rates,measure="tpr", x.measure="fpr")  
  
##plot ROC curve and overlay the diagonal line for random guessing  
plot(roc\_result, main="ROC Curve for predicting rural")  
lines(x = c(0,1), y = c(0,1), col="red")



##compute the AUC  
auc<-performance(rates, measure = "auc")  
auc@y.values

## [[1]]  
## [1] 0.8333333

## Confusion Matrix

With a threshold of 0.4, the error rate is 0.09 and the accuracy rate is 0.91. In addition, the false positive rate is 0.33 and the false negative rate is 0. We decided to set the threshold to 0.4, because we are more concerned with the false negative rate. We want a lower false negative rate, because we don’t want to classify a nursing home as non-rural when it is indeed rural. If we incorrectly classify a rural nursing home facility as non-rural, it may affect the facilities’ overall public funding. It is also important to note that since the dataset is small with an unblanaced number of rural and non-rural nursing homes.

##confusion matrix. Actual values in the rows, predicted classification in cols  
confusion <- table(test$Rural, preds>0.4)  
confusion

##   
## FALSE TRUE  
## Non-Rural 2 1  
## Rural 0 8

tp <- confusion[2,2]  
tn <- confusion[1,1]  
fp <- confusion[1,2]  
fn <- confusion[2,1]  
fpr <- fp/(tn+fp)  
fpr

## [1] 0.3333333

fnr <- fn/(fn+tp)  
fnr

## [1] 0

error\_rate <- (fp+fn)/(fp+fn+tn+tp)  
error\_rate

## [1] 0.09090909

accuracy\_rate <- (tp+tn)/(fp+fn+tn+tp)  
accuracy\_rate

## [1] 0.9090909

## Conclusion

**Based on our results, we conclude that nursing salaries appear to be the most important factor in determining if a nursing home is rural or non-rural. This information can be used by policymakers to close the financial gap between rural and non-rural nursing homes.** **For each additional $100 in annual nurse salary, the log odds of being a rural nursing home facility decreases by 0.00064.**