STAT 6021: Project 2

Group 2: Gregory Madden, Christina Kuang, Chi Do, Trey Hamilton

Table of Contents

[Executive Summary 2](#_Toc102835119)

[Description of Data and Variables 3](#_Toc102835120)

[Multiple Linear Regression 3](#_Toc102835121)

[Exploratory Data Analysis: 3](#_Toc102835122)

[Model Selection Using Automated Search Procedures: 6](#_Toc102835123)

[Identifying Outliers 8](#_Toc102835124)

[Checking linear regression assumptions: 11](#_Toc102835125)

[Press Statistic 14](#_Toc102835126)

[Multiple Linear Regression Conclusion 14](#_Toc102835127)

[Logistic Regression Model 14](#_Toc102835128)

[Exploratory Data Analysis 14](#_Toc102835129)

[Full Logistic Regression Model with all predictors 15](#_Toc102835130)

[Dropping Predictors 16](#_Toc102835131)

[Testing Model Usefulness 16](#_Toc102835132)

[Model Validation 16](#_Toc102835133)

[Logistic Regression Conclusion 17](#_Toc102835134)

# 

# Executive Summary

The goal of this project was to explore the relationship between the different characteristics of nursing homes in New Mexico. By understanding the relationship between the different characteristics of nursing homes, we can help the government better allocate subsidy funds to nursing homes. For this project, we focused on two research questions: What characteristics of nursing homes in New Mexico dictate annual nurse salaries at those institutions? What characteristics of nursing homes in New Mexico help predict if a nursing home is rural or non-rural?

Our main objective for our first research question was to create a multiple linear regression model with cumulative annual nurse salaries for individual nursing homes using the available financial characteristics for each institution. We want to develop a model that reliably predicts institutions with low annual nursing salaries among the larger group of nursing homes across the state. The practical implication of our linear model is policymakers can use the model to predict cumulative annual nurse salaries for a given nursing home to rationally distribute subsidy funds to institutions that are expected to contribute the lowest salaries to a particular area. Based on our findings, 66.15% of the variability in annual nursing salaries can be explained by annual in patient stays, annual total patient days, and annual total facilities expenditure – as long as they are not rural facilities with more than 200 beds.

Our goal for our second research question was to use characteristics of nursing homes variables to develop a logistic regression model that helps predict whether a nursing home is rural or non-rural. Rural patients often suffer from a lack of locally available nursing home beds. By understanding the relationships of these characteristics and how they define rural vs. non-rural nursing homes, we can learn how to make rural nursing homes financially viable. Our final analysis showed that annual nursing salaries is the most important factor in determining whether a nursing home facility is rural or non-rural. The predicted log odds of a nursing facility being rural decrease by 0.00064 for each additional $100 in annual nurse salary. This information may be useful in helping policymakers close the nursing salary gap between rural and non-rural nursing homes.

# Description of Data and Variables

For this project, we used the Nursing dataset from the Stat2Data package. The dataset contains characteristics of nursing homes in New Mexico, with 52 observations and 7 variables. Each row in the dataset represents an individual nursing home. The dataset includes the following variables:

|  |  |
| --- | --- |
| Variable | Description |
| Beds: | Number of beds in the nursing home |
| InPatientDays: | Annual medical in-patient days (in hundreds) |
| AllPatientDays: | Annual total patient days (in hundreds) |
| PatientRevenue: | Annual patient care revenue (in hundreds of dollars) |
| NurseSalaries: | Annual nursing salaries (in hundreds of dollars) |
| FacilitiesExpend: | Annual facilities expenditure (in hundreds of dollars) |
| Rural: | 1=rural or 0=non-rural |

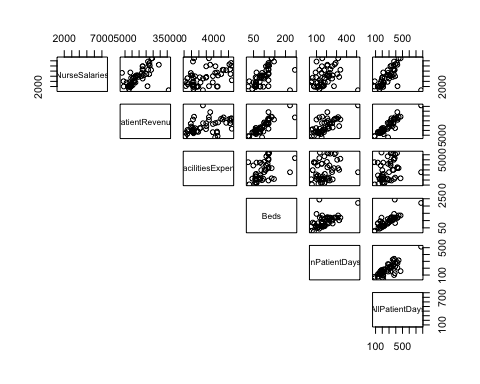
# Multiple Linear Regression

## Exploratory Data Analysis:

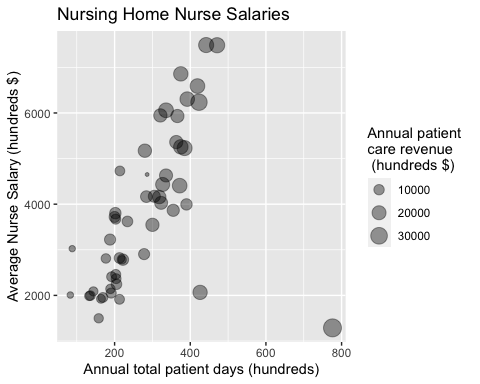
Based on the scatter plots and correlation table, it appears that nurse salaries has a moderate correlation with beds, all patient days, and patient revenue. There also appears to be a strong linear relationship between beds and all patient days, beds and patient revenue, in-patient days and all patient days, in patient days and patient revenue, and all patient days and patient revenue.

cor(Data[1:6])

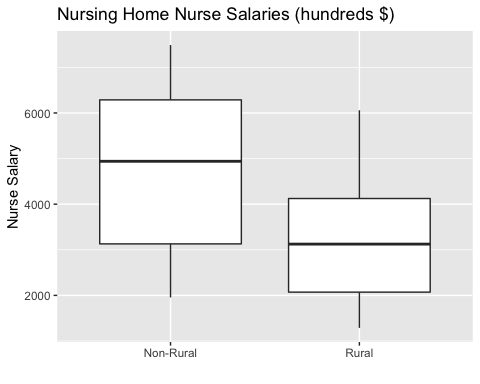
## Beds InPatientDays AllPatientDays PatientRevenue  
## Beds 1.0000000 0.5680006 0.8182959 0.8437752  
## InPatientDays 0.5680006 1.0000000 0.8116225 0.7070754  
## AllPatientDays 0.8182959 0.8116225 1.0000000 0.9030608  
## PatientRevenue 0.8437752 0.7070754 0.9030608 1.0000000  
## NurseSalaries 0.5094241 0.2541355 0.5153965 0.5894065  
## FacilitiesExpend 0.4602559 0.2583959 0.3047354 0.4337859  
## NurseSalaries FacilitiesExpend  
## Beds 0.5094241 0.4602559  
## InPatientDays 0.2541355 0.2583959  
## AllPatientDays 0.5153965 0.3047354  
## PatientRevenue 0.5894065 0.4337859  
## NurseSalaries 1.0000000 0.4550656  
## FacilitiesExpend 0.4550656 1.0000000



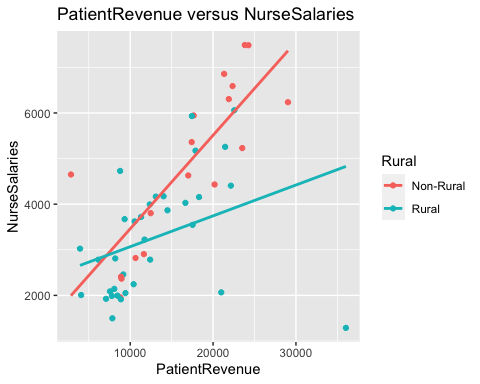
We also examined the relationship between annual total patient days and annual nurse salaries by annual patient revenues. From our scatterplot, we can see there appears to be a moderately strong correlation between annual total patient days and average nurse salary, with at least one apparent outlier in terms of high patient days and low salary – observation #26.



We also created a box plot to demonstrate the differences in institutional nurse salaries in New Mexico for rural and non-rural areas. Based on the box plot, it appears there is a greater variability for non-rural nurse salary. The nurses in non-rural regions also have a higher median salary.



We then examined the relationship between patient revenue and nurse salaries for rural and non-rural facilities using a scatter plot. The plot shows the slopes are not parallel, which indicates there is an interaction effect between patient revenue and nurse salaries. Once again, we noticed observation #26 appears to be an outlying observation.



## Model Selection Using Automated Search Procedures:

For our initial multiple linear regression model, we used backward selection to find the model with the lowest AIC value. The model selected consists of the following predictors: patient revenue, facilities expenditure, and rural.

## Start: AIC=744.15  
## NurseSalaries ~ Beds + InPatientDays + AllPatientDays + PatientRevenue +   
## FacilitiesExpend + Rural  
##   
## Df Sum of Sq RSS AIC  
## - Beds 1 2343232 67522373 743.99  
## <none> 65179141 744.15  
## - AllPatientDays 1 2846474 68025615 744.38  
## - PatientRevenue 1 3971831 69150972 745.23  
## - InPatientDays 1 4876882 70056023 745.91  
## - Rural 1 7838332 73017473 748.06  
## - FacilitiesExpend 1 9086310 74265451 748.94  
##   
## Step: AIC=743.99  
## NurseSalaries ~ InPatientDays + AllPatientDays + PatientRevenue +   
## FacilitiesExpend + Rural  
##   
## Df Sum of Sq RSS AIC  
## - AllPatientDays 1 1414904 68937277 743.07  
## - PatientRevenue 1 2585846 70108219 743.94  
## <none> 67522373 743.99  
## - InPatientDays 1 3560726 71083098 744.66  
## - Rural 1 7195730 74718102 747.26  
## - FacilitiesExpend 1 7207894 74730267 747.26  
##   
## Step: AIC=743.07  
## NurseSalaries ~ InPatientDays + PatientRevenue + FacilitiesExpend +   
## Rural  
##   
## Df Sum of Sq RSS AIC  
## - InPatientDays 1 2149798 71087074 742.66  
## <none> 68937277 743.07  
## - FacilitiesExpend 1 6158014 75095291 745.52  
## - Rural 1 9421115 78358391 747.73  
## - PatientRevenue 1 15130303 84067580 751.39  
##   
## Step: AIC=742.66  
## NurseSalaries ~ PatientRevenue + FacilitiesExpend + Rural  
##   
## Df Sum of Sq RSS AIC  
## <none> 71087074 742.66  
## - FacilitiesExpend 1 6780636 77867711 745.40  
## - Rural 1 13691261 84778336 749.82  
## - PatientRevenue 1 16637023 87724097 751.60

##   
## Call:  
## lm(formula = NurseSalaries ~ PatientRevenue + FacilitiesExpend +   
## Rural, data = Data)  
##   
## Coefficients:  
## (Intercept) PatientRevenue FacilitiesExpend RuralRural   
## 2621.88996 0.09382 0.20764 -1121.87554

**Reduced model summary:**

reduced <- lm(NurseSalaries~PatientRevenue + FacilitiesExpend +   
 Rural,data=Data)  
summary(reduced)

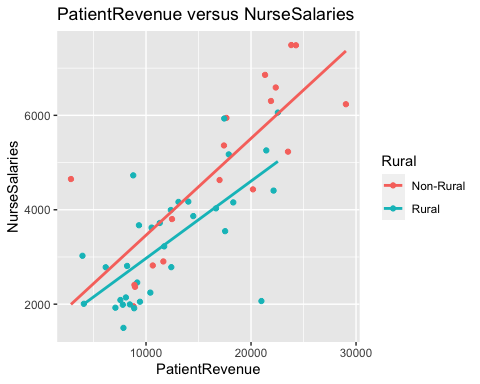
##   
## Call:  
## lm(formula = NurseSalaries ~ PatientRevenue + FacilitiesExpend +   
## Rural, data = Data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4120.3 -708.2 -71.7 833.6 2306.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2621.88996 526.64925 4.978 0.00000868 \*\*\*  
## PatientRevenue 0.09382 0.02799 3.352 0.00157 \*\*   
## FacilitiesExpend 0.20764 0.09704 2.140 0.03749 \*   
## RuralRural -1121.87554 368.97560 -3.041 0.00382 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1217 on 48 degrees of freedom  
## Multiple R-squared: 0.4939, Adjusted R-squared: 0.4622   
## F-statistic: 15.61 on 3 and 48 DF, p-value: 0.0000003216

## Identifying Outliers

Once we had our initial model , we calculated the externally studentized residuals to find outlying observations. Our results show that observation is an outlier with an externally residual value of -5.04. Next, we calculated the leverage values of each observation to find observations with high leverages. Using a threshold of 2p/n, we identified observations 26 and 31 as records with high leverages. The results from both analyses indicate observation # 26 is both outlying in the predictors and appears to have high leverage.

Unfortunately, this dataset nor the primary reference [Smith et al. “A Comparison of Financial Performance, Organizational Characteristics, and Management Strategy Among Rural and Urban Nursing Facilities,” Journal of Rural Health, 1992, pp 27-40.] provide identifying information for individual Nursing Home facilities, so we cannot make specific conclusions about the facility for observation #26. What we can say about observation 26 is that it is a relatively large facility with 221 beds, especially among other rural facilities. For example, the number of beds for this facility is 79% higher than the next largest rural nursing home (123 beds). Also, despite very high patient revenue and Patient census, the nurses salary is the lowest of the entire dataset. Therefore, we suspect this facility may not be comparable with other institutions, perhaps owing to it’s unique combination of large facility and rural classification. Since our primary objective is to identify low nursing salaries particularly in rural areas, we proposed re-running the regression while excluding observation #26 on the basis that is is an extraordinarily large (>200 beds) rural facility. Future predictions for such institutions will not be made using this model unless rural facilities are under 200 beds. Separate policy considerations will then be made for rare large/rural institutions.

Interestingly, after re-running the EDA plots, we noticed that the slope and the intercept of the relationship between PatientRevenue and NurseSalaries no longer appear to change depending on rural vs. non-rural status. This interaction effect appears gone after excluding observation #26.



After we removed observation #26 from our dataset, we re-ran the backward selection automated search procedure. This time it selected the following model:

NurseSalaries ~ InPatientDays + AllPatientDays + FacilitiesExpend

We noticed all predictors have a significant t-test, with a p-value less than the alpha value of 0.05. Once we determined all predictors are significant, we calculated the VIF values to help detect multicollinearity. Since all three predictors have VIF values below four, there appears to be no evidence of multicollinearity in the model. In the end, we decided to use this as our final multiple linear regression model.

## Start: AIC=700.24  
## NurseSalaries ~ Beds + InPatientDays + AllPatientDays + PatientRevenue +   
## FacilitiesExpend + Rural  
##   
## Df Sum of Sq RSS AIC  
## - Beds 1 387718 35976230 698.79  
## - PatientRevenue 1 1096139 36684650 699.79  
## - Rural 1 1171291 36759803 699.89  
## <none> 35588512 700.24  
## - FacilitiesExpend 1 2797910 38386422 702.10  
## - InPatientDays 1 3594969 39183481 703.15  
## - AllPatientDays 1 10398211 45986723 711.31  
##   
## Step: AIC=698.79  
## NurseSalaries ~ InPatientDays + AllPatientDays + PatientRevenue +   
## FacilitiesExpend + Rural  
##   
## Df Sum of Sq RSS AIC  
## - PatientRevenue 1 801202 36777432 697.92  
## - Rural 1 1024185 37000415 698.23  
## <none> 35976230 698.79  
## - FacilitiesExpend 1 2419510 38395740 700.11  
## - InPatientDays 1 3228413 39204643 701.18  
## - AllPatientDays 1 10180142 46156373 709.50  
##   
## Step: AIC=697.92  
## NurseSalaries ~ InPatientDays + AllPatientDays + FacilitiesExpend +   
## Rural  
##   
## Df Sum of Sq RSS AIC  
## - Rural 1 1054507 37831939 697.36  
## <none> 36777432 697.92  
## - InPatientDays 1 3542237 40319669 700.61  
## - FacilitiesExpend 1 4014212 40791644 701.20  
## - AllPatientDays 1 33957926 70735359 729.27  
##   
## Step: AIC=697.36  
## NurseSalaries ~ InPatientDays + AllPatientDays + FacilitiesExpend  
##   
## Df Sum of Sq RSS AIC  
## <none> 37831939 697.36  
## - FacilitiesExpend 1 3958107 41790046 700.43  
## - InPatientDays 1 5922091 43754029 702.78  
## - AllPatientDays 1 54134669 91966608 740.66

##   
## Call:  
## lm(formula = NurseSalaries ~ InPatientDays + AllPatientDays +   
## FacilitiesExpend, data = Data)  
##   
## Coefficients:  
## (Intercept) InPatientDays AllPatientDays FacilitiesExpend   
## 366.0890 -6.7782 15.7325 0.1555

**Final reduced model summary:**

## Call:  
## lm(formula = NurseSalaries ~ InPatientDays + AllPatientDays +   
## FacilitiesExpend, data = Data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3667.4 -442.8 -126.8 598.5 1789.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 366.08897 378.34122 0.968 0.3382   
## InPatientDays -6.77821 2.49895 -2.712 0.0093 \*\*   
## AllPatientDays 15.73249 1.91840 8.201 0.000000000128 \*\*\*  
## FacilitiesExpend 0.15552 0.07013 2.217 0.0315 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 897.2 on 47 degrees of freedom  
## Multiple R-squared: 0.7176, Adjusted R-squared: 0.6995   
## F-statistic: 39.8 on 3 and 47 DF, p-value: 0.0000000000005928

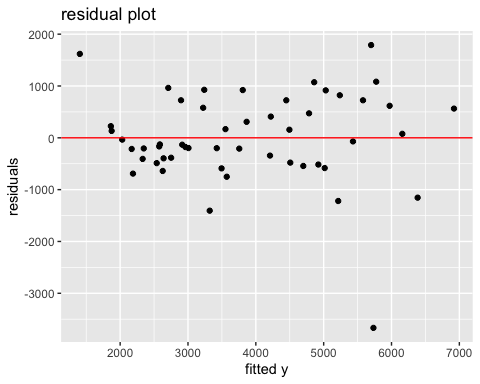
**VIF output:**

## InPatientDays AllPatientDays FacilitiesExpend   
## 2.134690 2.259279 1.183270

## 

## Checking linear regression assumptions:

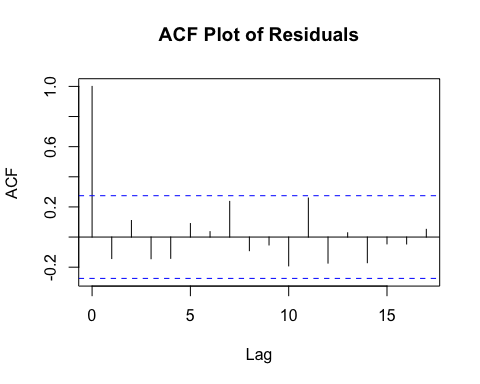
Next, we continued with our analysis by checking to see if the model meets all linear regression assumptions. First, we generated a residual plot to show the model’s residuals against the fitted values. The plot showed that residuals were evenly distributed along the y-axis. The error terms appear to have a mean of zero and the a constant variance. In addition to the residual plot, we also created a box cox plot to see if it is necessary to transform the response variable. Since the plot shows 1 lies within the 95% CI for lambda, we decided we do not need to transform the y variable. The first two assumptions are met and we do not have to perform any data transformations.



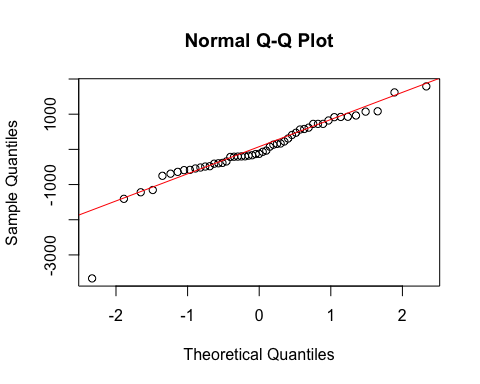
Chart, diagram

Description automatically generated

We continued with our analysis by generating an autocorrelation plot (ACF) to determine if the error terms are correlated. The output of the ACF plot indicates that the observations are independent, because all lag points are statistically insignificant. In other words, our third assumption is also met.



Finally, we created a normal probability plot to check if the errors follow a normal distribution. Since majority of the points fall closely to the line, we decided that the errors do follow a normal distribution and the last assumption is met.



## Press Statistic

Since we will be using this model to predict nursing salaries for new data, we are interested in the . Based on our calculated value of 0.6615, the final model might be able to explain 66.15% of the variability in the new observations (as long as they are not rural facilities >200 beds). The R2 is 0.7176. Both values are fairly close to each other, so overfitting is not a major concern.

## Multiple Linear Regression Conclusion

Our final model is:

NurseSalaries = 366.09 – (6.78\*InpatientDays) + (15.73\*AllPatientDays)

+ (0.16\*FacilitiesExpenditure)

, where bed size for rural institutions is <200, NursingSalary = Estimated Annual nursing salaries (in hundreds of dollars), InpatientDays represents annual medical in-patient days (in hundreds), AllPatientDays represents annual total in-patient days (in hundreds), and FacilitiesExpenditure = hundreds of $.

According to our analysis, total inpatient days, total medical inpatient days, and total facilities expenditures appear to be the most important factors in determining nursing salaries among all predictors we looked at. Based on the , our final model summarized above might be able to explain 66% of the variability in nursing salaries of future nursing homes, as long as they are not rural facilities with more than 200 beds.

# Logistic Regression Model

## Exploratory Data Analysis

We created density plots for each variable to explore the relationship between the variables and whether a nursing home is rural or non-rural. We can see the density plots of nurse salaries for rural are 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.

A picture containing chart

Description automatically generated

## Full Logistic Regression Model with all predictors

First, we split our dataset into a training dataset and a testing set using set.seed(10). Since our dataset is small, we decided to do a 80-20 split, so we have enough records for our training set. Next, we fitted a logistic regression model with all six predictors. According to the Wald test results in our summary output, only nurse salaries is significant.

##   
## 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

In addition to conducting a Wald test, we also calculated the 95% confidence interval for nurse salaries, which 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 zero 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 result for nurse salaries.

## Dropping Predictors

Since the predictors facilities expenditure, beds, patient revenue, all patient days, and in patient days do not have significant Wald test results, we performed a delta G-squared test to see if we can drop the predictors from the model. With a p-value of 0.35, we failed to reject the null. Since none of the subset predictors appear to be significant, we dropped them from our model.

## Testing Model Usefulness

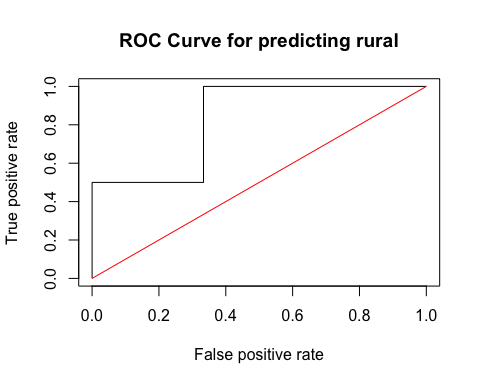
With only one predictor remaining, we performed a second delta G-squared test to see if the one predictor model is better than the intercept-only mode. The results revealed the test statistic is 8.53 with a p-value of 0.014. Since the p-value is less than 0.05, we rejected the null hypothesis. The results show that the one predictor model is better at prediction than the intercept-only model. As a result, our final model is:

= 3.0935426 -0.0006423(NurseSalaries)

## 

## Model Validation

After finalizing the model, we generated an ROC curve and calculated the AUC value to help evaluate the model. Based on the ROC curve, our logistic regression model performs better than random guessing, because the curve is above the diagonal line. In addition, our AUC value of 0.83 is greater than 0.5, which supports the claim that the model performs better than random guessing.



Lastly, we created a confusion matrix to help determine the predictive ability of our model. We initially used a threshold of 0.5, but we later 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. Therefore, 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. It is important to note that since the dataset is small with an unblanaced number of rural and non-rural nursing homes, it leads to a higher false positive rate, because there are more rural nursing facilities in the dataset.

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

## Logistic Regression 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.