STAT 6021: Project 2

Group 2

2022-05-06

Group 2 Members:

Gregory Madden, Christina Kuang, Chi Do, Trey Hamilton

Description of Dataset (provided by Stat2Data Package)

Our dataset contains characteristics of nursing homes in New Mexico, with 52 observations and 7 variables. Each row represents an individual nursing home.

Variables/Columns:

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

```
# reading in data
data(Nursing)
Data <- Nursing
Data$Rural <- factor(Data$Rural)
levels(Data$Rural) <- c("Non-Rural", "Rural")
# checking categorical variable classifiers
contrasts(Data$Rural)
```

```
##           Rural
## Non-Rural     0
## Rural         1
```

Glimpse of data overview:

```
glimpse(Data)
```

```
## Rows: 52
## Columns: 7
## $ Beds             <int> 244, 59, 120, 120, 120, 65, 120, 90, 96, 120, 62, 120~
## $ InPatientDays    <int> 128, 155, 281, 291, 238, 180, 306, 214, 155, 133, 148~
```

```
## $ AllPatientDays   <int> 385, 203, 392, 419, 363, 234, 372, 305, 169, 188, 192~
## $ PatientRevenue   <int> 23521, 9160, 21900, 22354, 17421, 10531, 22147, 14025~
## $ NurseSalaries    <int> 5230, 2459, 6304, 6590, 5362, 3622, 4406, 4173, 1955,~
## $ FacilitiesExpend <int> 5334, 493, 6115, 6346, 6225, 449, 4998, 966, 1260, 64~
## $ Rural            <fct> Non-Rural, Rural, Non-Rural, Non-Rural, Non-Rural, Ru~
```

Question 1. What characteristics of nursing homes in New Mexico dictate annual nurse salaries at those institutions?

Practical implications of a linear model for predicting cumulative annual nurse salaries for a given nursing home could be used by policymakers to rationally distribute subsidy funds to institutions that are expected to contribute the lowest salaries to a particular area.

Objective 1: Fit a multiple linear regression model with cumulative annual nurse salaries for individual nursing homes using the available financial characteristics for each institution. The goal is to develop a model using these available data to reliably predict institutions with low annual nursing salaries among the larger group of nursing homes across the state.

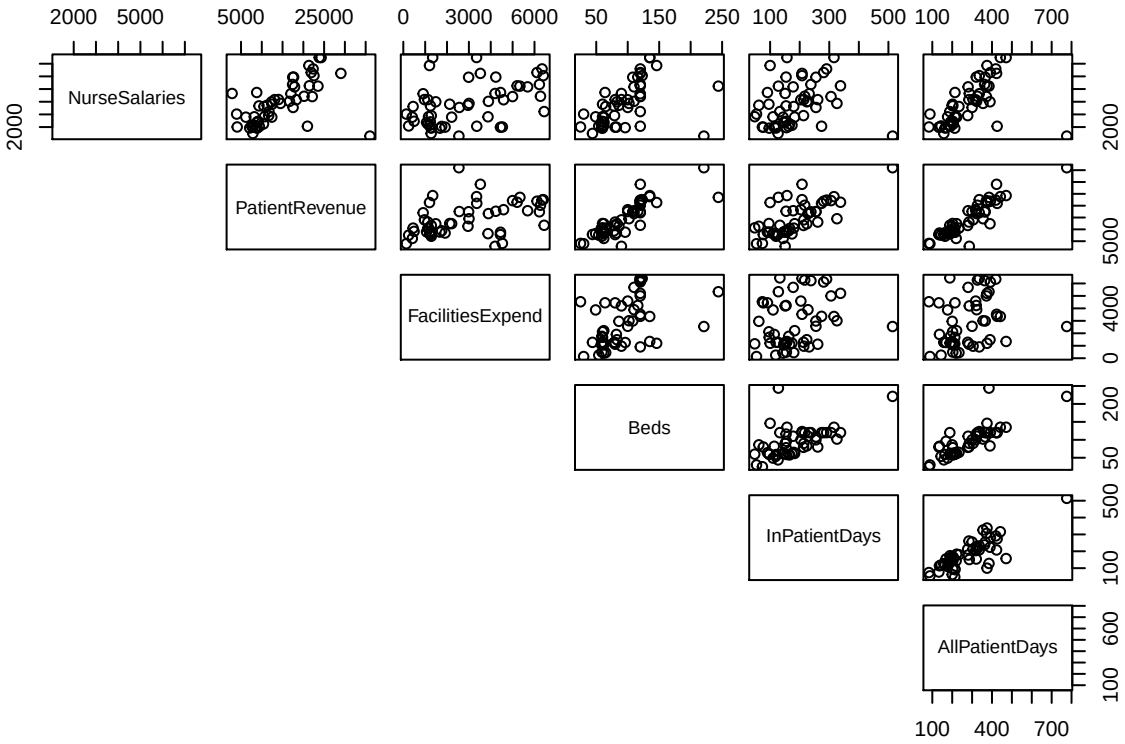
Exploratory Data Analysis:

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 AllPatientDays, 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])
```

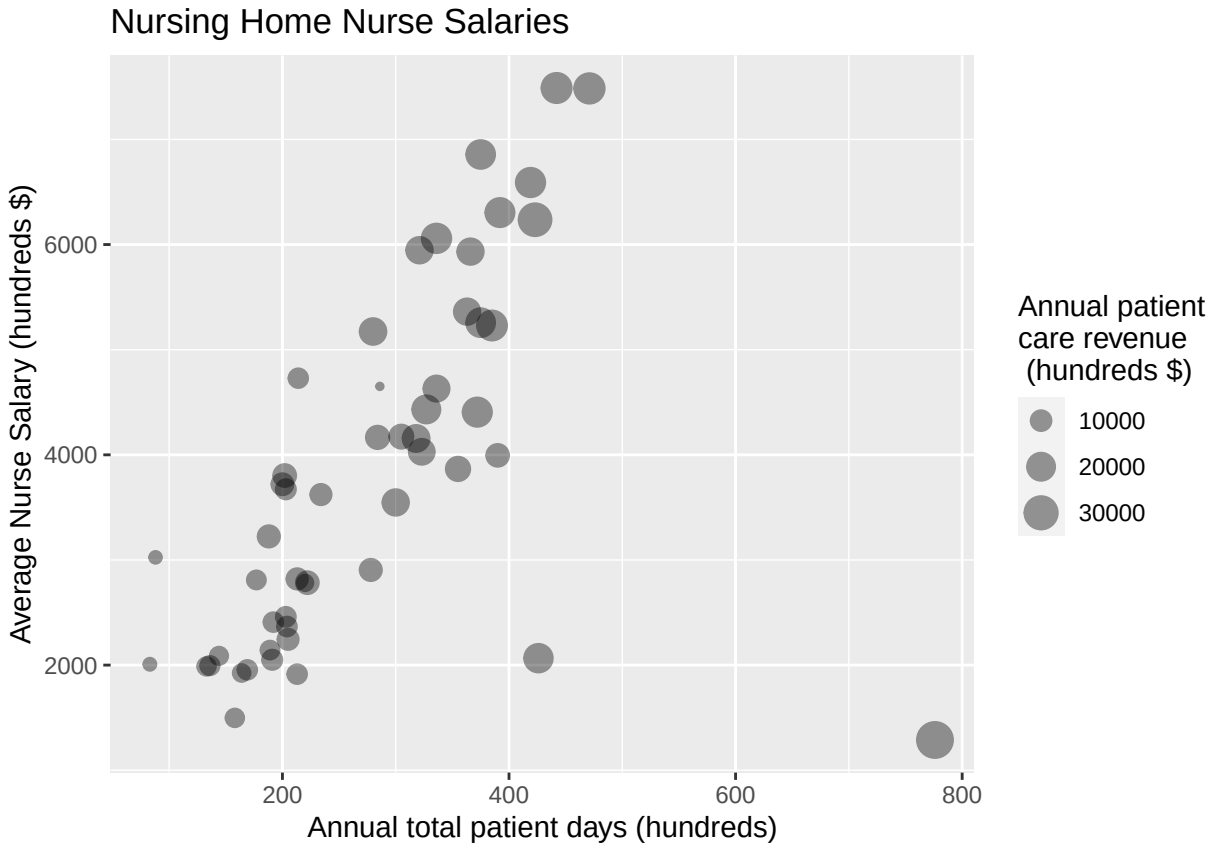
```
##                     Beds InPatientDays AllPatientDays PatientRevenue
## Beds             1.0000000     0.5680006      0.8182959      0.8437752
## InPatientDays    0.5680006     1.0000000      0.8116225      0.7070754
## AllPatientDays   0.8182959     0.8116225      1.0000000      0.9030608
## 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
```

```
pairs(~NurseSalaries + PatientRevenue + FacilitiesExpend + Beds + InPatientDays +
    AllPatientDays, data = Data, lower.panel = NULL)
```



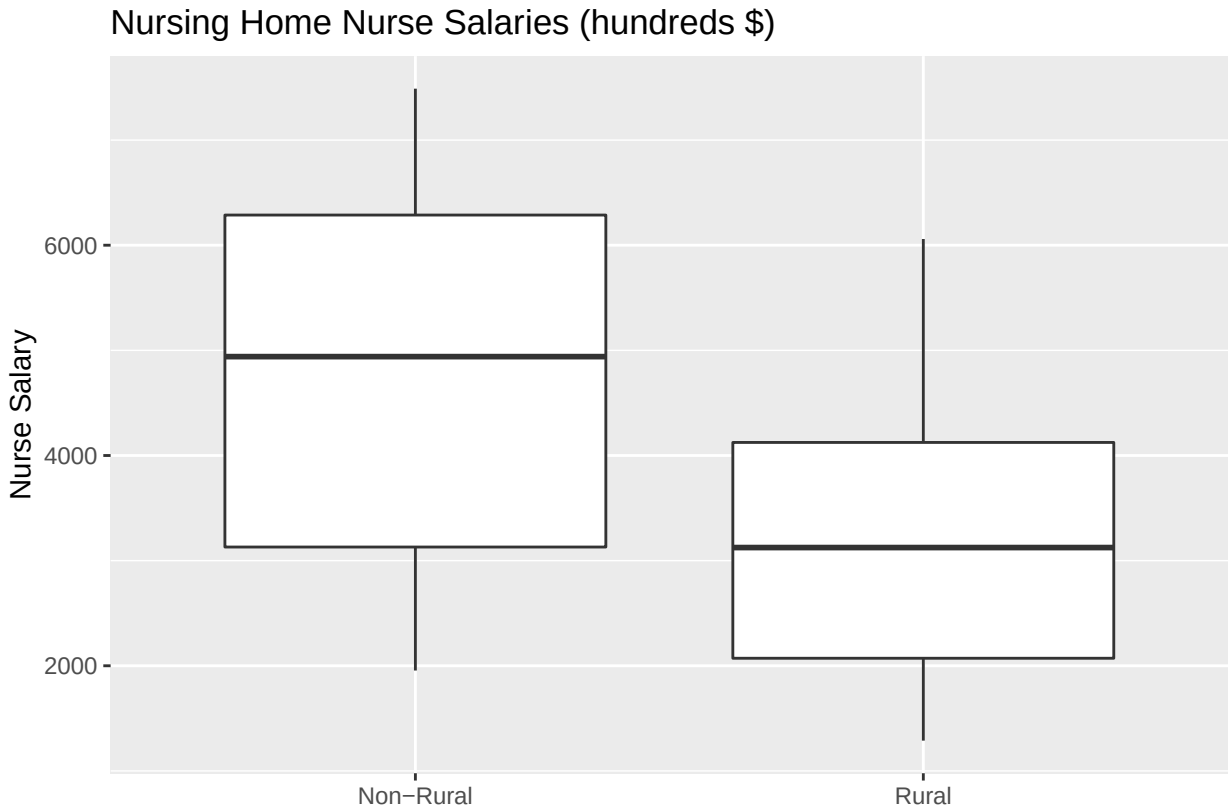
Nursing home salaries by census in Annual total patient days (in hundreds) and Nursing home size (by number of beds). There appears to be a moderately strong correlation between annual total patient days and average nurse salary, with at least one apparent outlier in terms of high patient days and low salary, observation #26.

```
Data %>%
    ggplot(aes(x = AllPatientDays, y = NurseSalaries, size = PatientRevenue)) + geom_point(alpha = 0.4)
    labs(x = "Annual total patient days (hundreds)", y = "Average Nurse Salary (hundreds $)",
        title = "Nursing Home Nurse Salaries") + guides(size = guide_legend(title = "Annual patient \nca
```



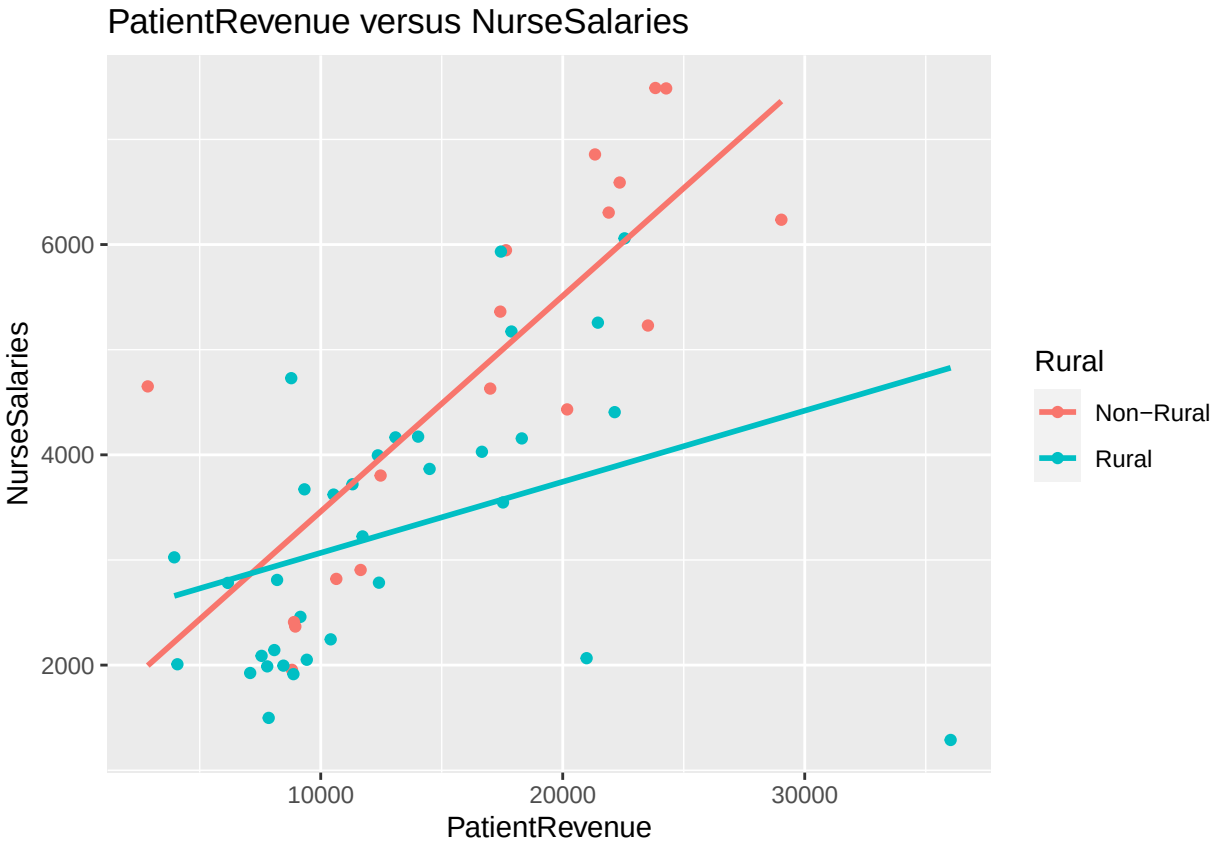
Boxplot demonstrating differences in Institutional nurse's salaries in New Mexico for Rural Areas compared with Non-Rural: Based on the box plot, it appears there is a greater variability for non-rural nurse salary. The nurses in non-rural regions also have a higher median salary.

```
Data$Rural <- factor(Data$Rural)
levels(Data$Rural) <- c("Non-Rural", "Rural")

Data %>%
    ggplot(aes(x = Rural, y = NurseSalaries)) + geom_boxplot() + labs(x = "", y = "Nurse Salary",
    title = "Nursing Home Nurse Salaries (hundreds $)")
```



Scatter plot of Patient Revenue versus Nurse Salaries: The slopes are not parallel, which indicates there is an interaction effect between Patient Revenue and Nurse Salaries. Again seen is outlying observation #26.

```
ggplot(Data, aes(x = PatientRevenue, y = NurseSalaries, color = Rural)) + geom_point() +
    geom_smooth(method = "lm", se = FALSE) + labs(x = "PatientRevenue", y = "NurseSalaries",
    title = "PatientRevenue versus NurseSalaries")
```



Carrying out initial automated search procedures:

Using backward selection to find the best model according to AIC. Start with the first-order model with all the predictors.

The model selected is: NurseSalaries ~ PatientRevenue + FacilitiesExpend + 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 outlying observations.

Observation #26 is outlying.

```
## externally studentized residuals, t_i
ext.student.res <- rstudent(reduced)
## identify outliers with t_i## critical value using Bonferroni procedure
n <- dim(Data)[1]
# p is the number of predictors + 1 for the intercept
p <- 4
crit <- qt(1 - 0.05/(2 * n), n - 1 - p)
## identify
ext.student.res[abs(ext.student.res) > crit]
```

```
##        26
## -5.038319
```

Identifying observations that have high leverage.

Calculating the leverage values h_{ii} below and identifying ones that are >2*p/n.

Observations 26 and 31 have high leverage.

```
lev <- lm.influence(reduced)$hat
lev[lev > 2 * p/n]
```

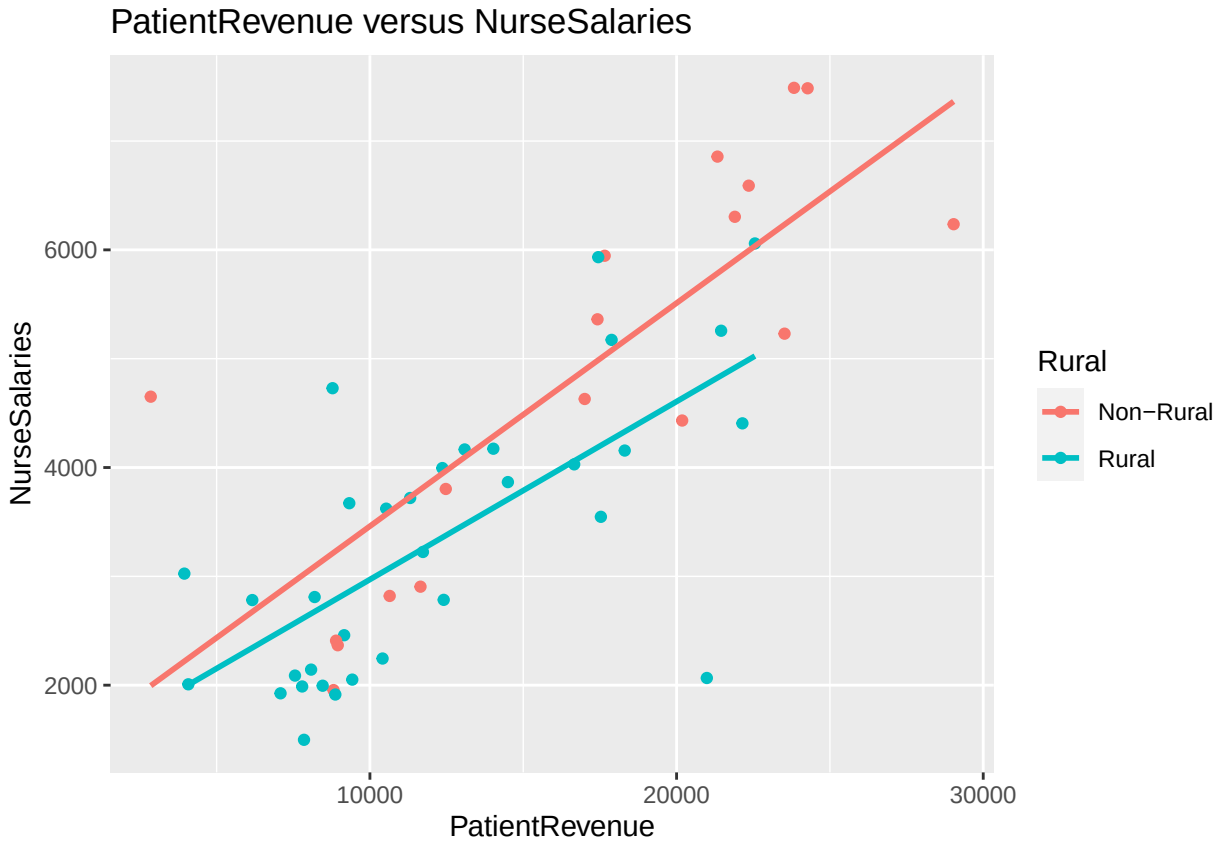
```
##        26        31
## 0.3189973 0.1873532
```

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.

```
Data <- Data[-26, ]
```

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.

```
Data %>%
    ggplot(aes(x = PatientRevenue, y = NurseSalaries, color = Rural)) + geom_point() +
    geom_smooth(method = "lm", se = FALSE) + labs(x = "PatientRevenue", y = "NurseSalaries",
    title = "PatientRevenue versus NurseSalaries")
```



Next, we repeated the automated search procedures with backward selection again after outlier observation #26 was removed.

The model selected is: NurseSalaries ~ InPatientDays + AllPatientDays + FacilitiesExpend

```
# intercept only
regnull <- lm(NurseSalaries ~ 1, data = Data)
# full
regfull <- lm(NurseSalaries ~ ., data = Data)
# backward elimination
step(regfull, scope = list(lower = regnull, upper = regfull), direction = "backward")
```

```
## 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

```
reduced <- lm(NurseSalaries ~ InPatientDays + AllPatientDays + FacilitiesExpend,
    data = Data)
summary(reduced)
```

```
##
## 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
```

All VIFs below 4 so multicollinearity does not appear to be an issue.

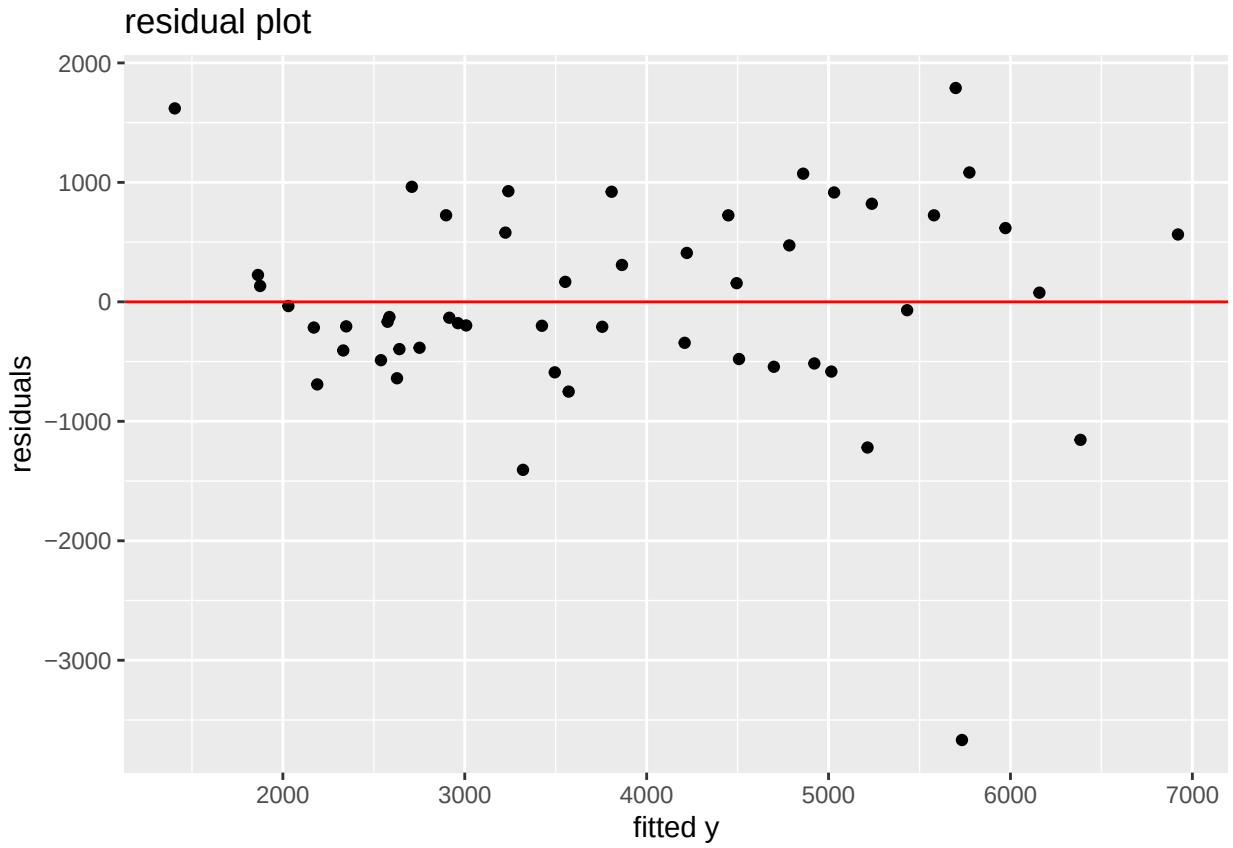
```
# requires package faraway
vif(reduced)
```

```
##    InPatientDays   AllPatientDays FacilitiesExpend
##         2.134690         2.259279         1.183270
```

Checking linear regression assumptions:

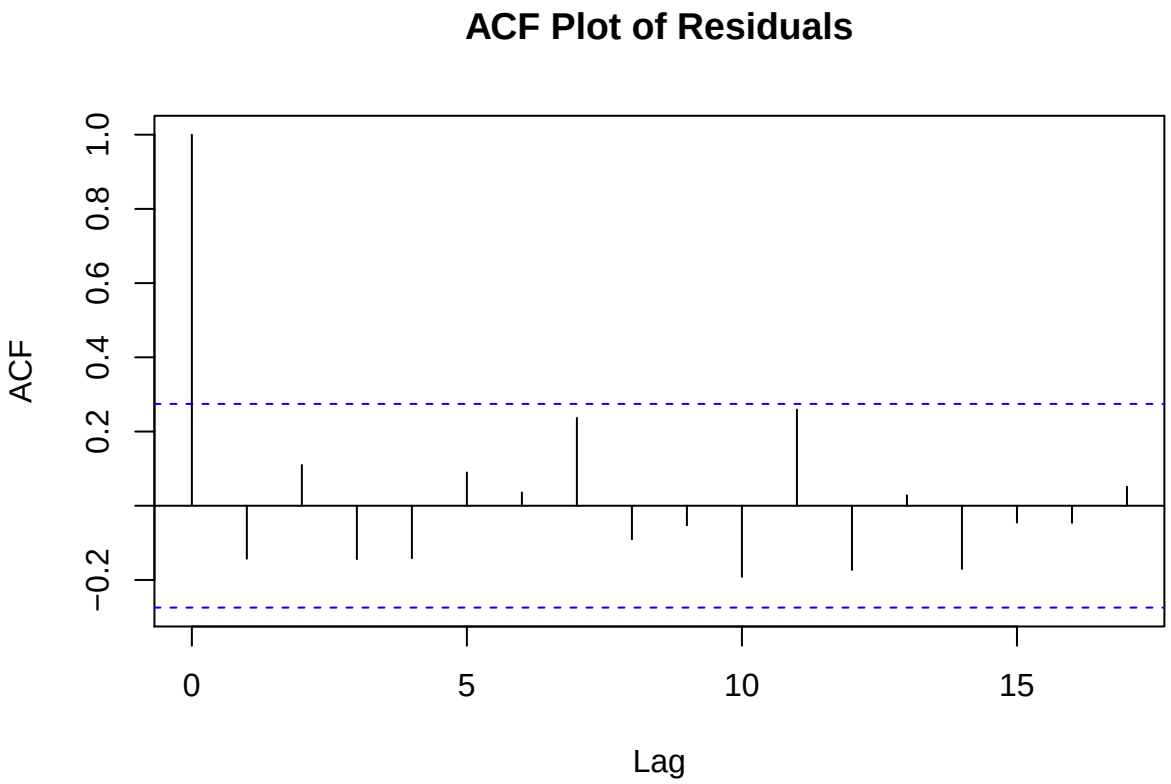
Assumptions 1/2 appear to be generally met: Mean of Errors = 0, constant variance.

```
# storing fitted y residuals
yhat <- reduced$fitted.values
res <- reduced$residuals
# add to data frame

Data <- data.frame(Data, yhat, res)

# residual plot
Data %>%
    ggplot(aes(x = yhat, y = res)) + geom_point() + geom_hline(yintercept = 0, color = "red") +
    labs(x = "fitted y", y = "residuals", title = "residual plot")
```



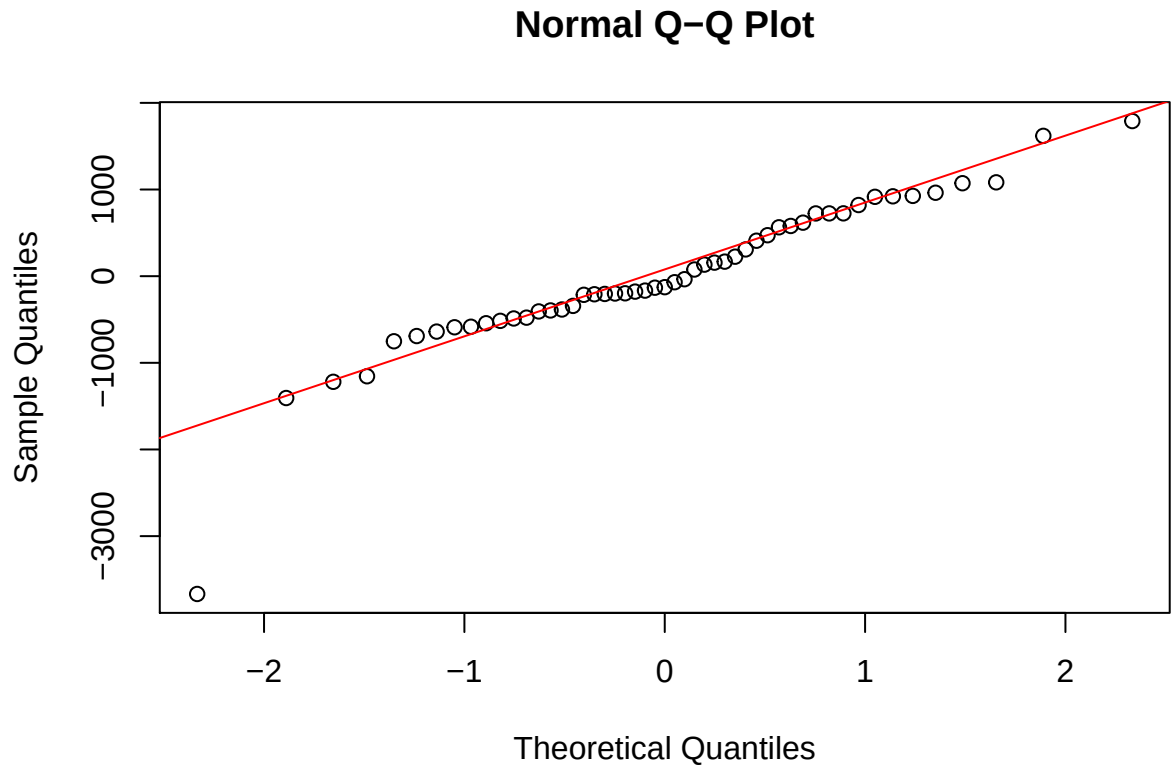
ACF Plot: no autocorrelation seen.

```
acf(res, main = "ACF Plot of Residuals")
```



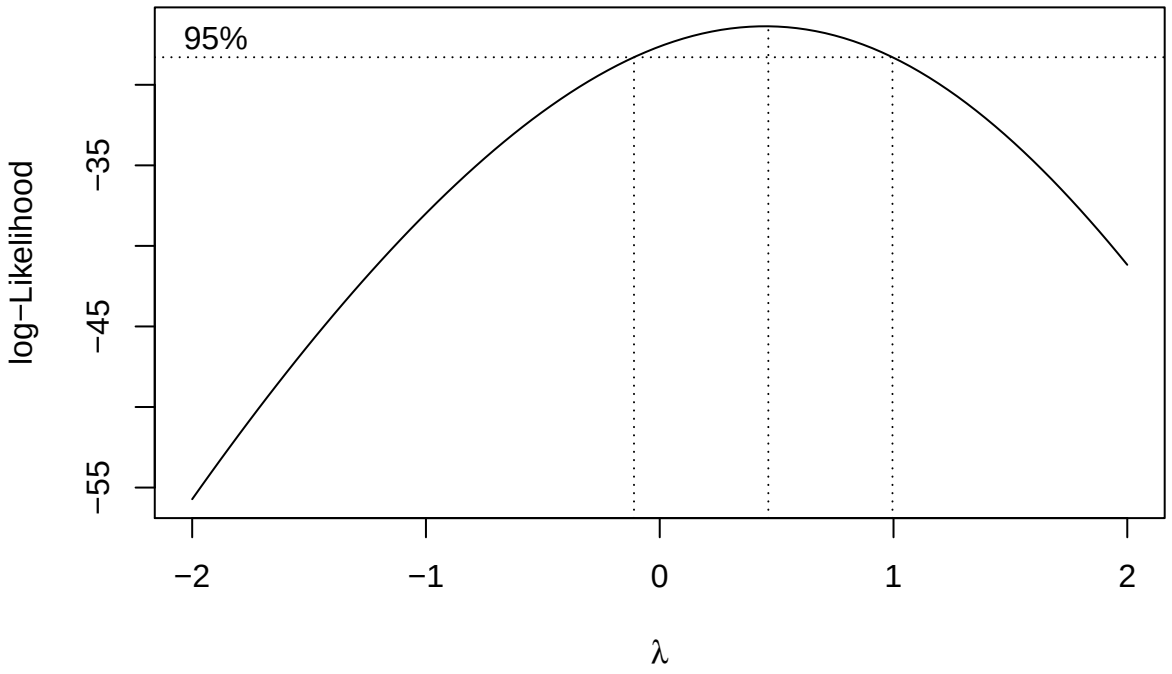
Errors are fairly normally distributed.

```
qqnorm(res)
qqline(res, col = "red")
```



Using boxcox method, we see that 1 lies just within the 95% CI for lambda so we did not transform the y variable.

```
boxcox(reduced)
```



Since we will be using this model to predict nursing salaries for new data, we are interested in the $R^2_{prediction}$.

Based on this value, 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.

```
# creating function to calculate PRESS statistic
PRESS <- function(model) {
    i <- residuals(model)/(1 - lm.influence(model)$hat)
    sum(i^2)
}

# PRESS(reduced) Find SST
anova_result <- anova(reduced)
SST <- sum(anova_result$"Sum Sq")
## R2 pred
Rsq_pred <- 1 - PRESS(reduced)/SST
Rsq_pred
```

```
## [1] 0.6615208
```

So, our final multiple linear regression equation is:

$NursingSalary = 366.08897 - 6.77821 * InpatientDays + 15.73249 * AllPatientDays + 0.15552 * FacilitiesExpenditure$

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 $.

Objective 1 Conclusions:

1) Patient census parameters (both total inpatient days as well as medical inpatient days) and total facilities expenditures appear to be the most important factors in determining nursing salaries among the predictors we looked at (others included PatientRevenue, Rural vs. Non-Rural).

2) Based on the $R^2_{prediction}$, 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 >200 beds.

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(0.8 * nrow(Data)), replace = F)
train <- Data[sample, ]  ##training data frame
test <- Data[-sample, ]  ##test data frame
```

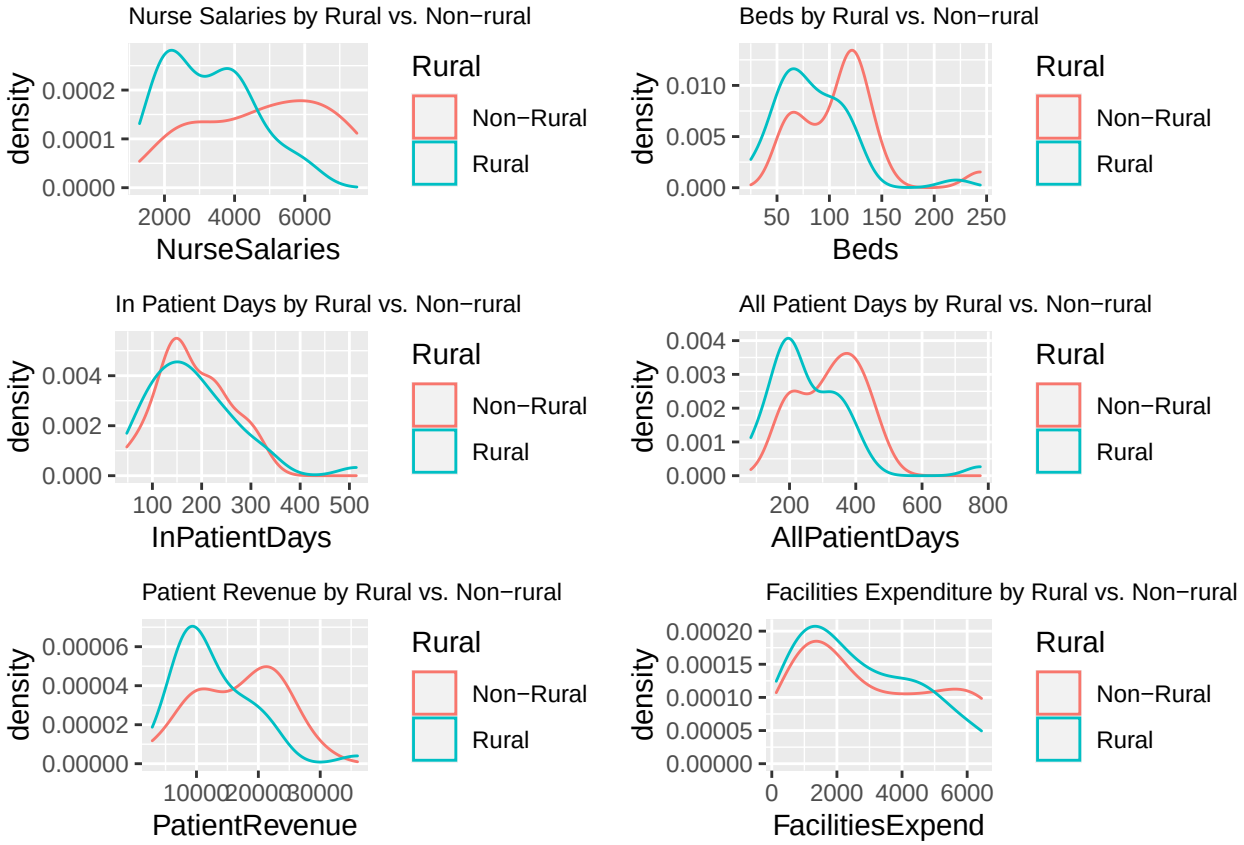
Data Exploration

The density plots of nurse salaries for rural 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.



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 β_1

$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

With a p-value of 0.04, we can reject the null hypothesis. There appears to be a significant relationship between β_1 and the response variable.

95% onfidence interval for β_1

The 95% confidence interval for β_1 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 β_1.

```
n <- 41
p <- 7
se_beta1 <- 0.0003464
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.

$\beta_1 = NurseSalaries$
$\beta_2 = Beds, InPatientDays, AllPatientDays, PatientRevenue, FacilitiesExpend$

H_0 : predictors in $\beta_2 = 0$

H_a : at least one of the coefficients in β_2 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

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

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
```

```r
1 - pchisq(TS_2, 2)
```

```
## [1] 0.01402527
```

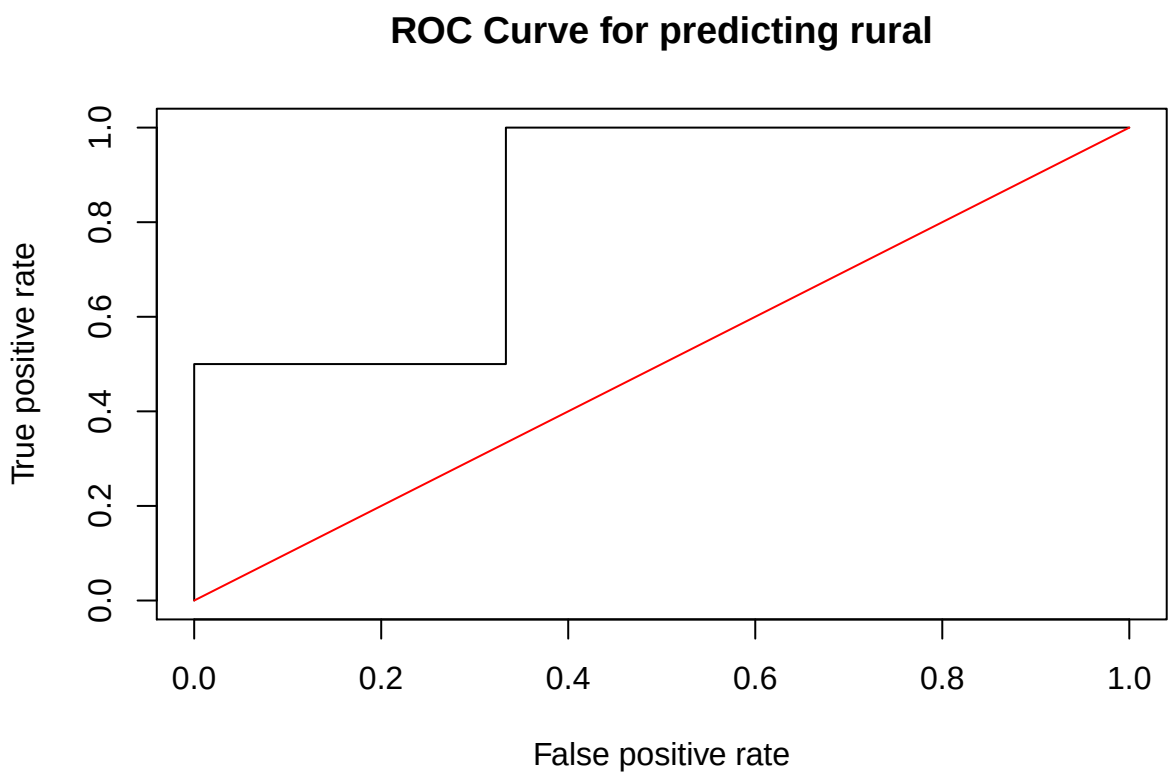
Final Model

$log(\pi/(1-\pi))$= 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.

```r
## 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
```

Objective 2 Conclusions:

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.