Objective 1 MLR Final Version

Greg Madden

2022-05-05

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

Our dataset contains characteristics of nursing homes in New Mexico.

Description of Dataset (provided by Stat2Data Package) Format: A dataset with 52 observations on the following 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

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.

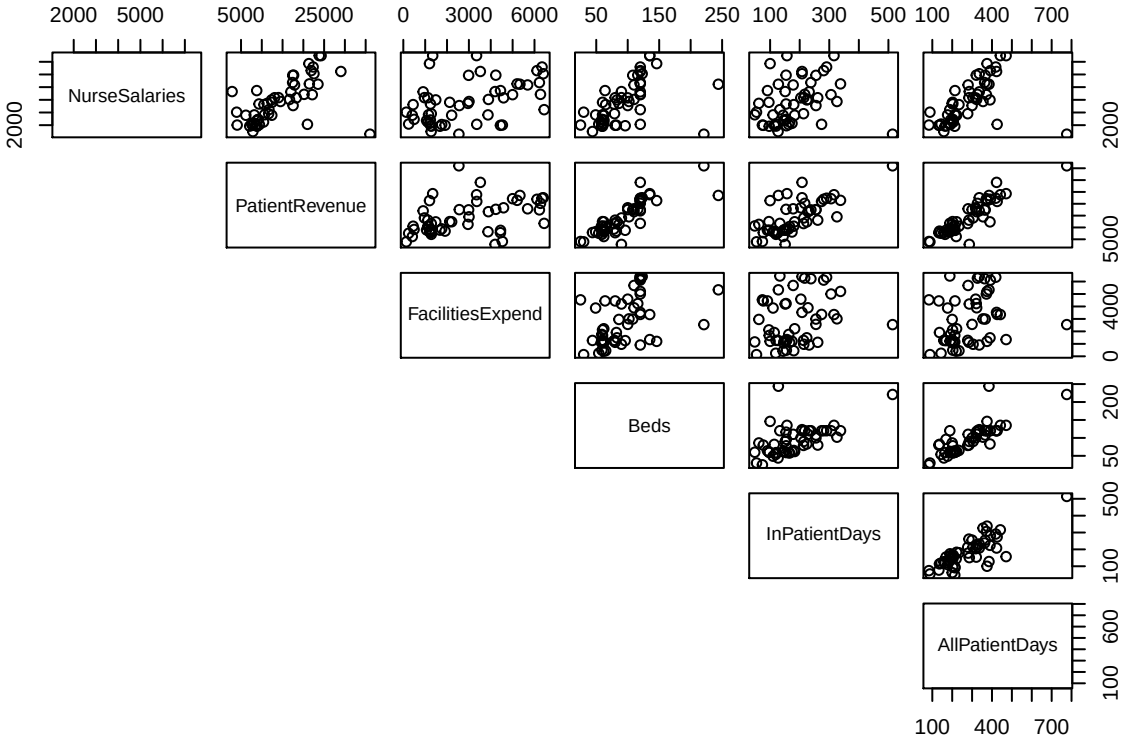
##Exploratory Data Analysis:

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])
```

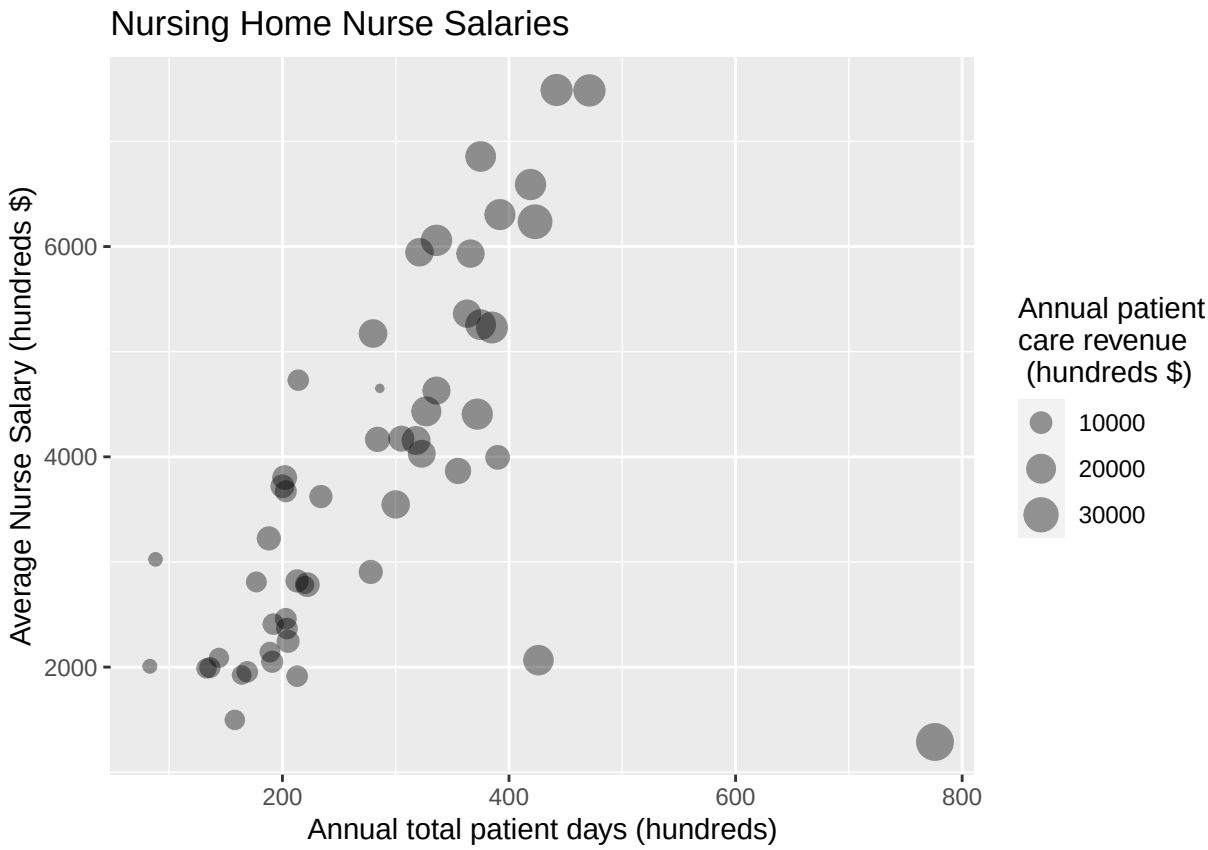
```
##                       Beds InPatientDays AllPatientDays PatientRevenue
## Beds             1.0000000     0.5680006      0.8182959      0.8437752
## InPatientDays    0.5680006     1.0000000      0.8116225      0.7070754
## AllPatientDays   0.8182959     0.8116225      1.0000000      0.9030608
##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



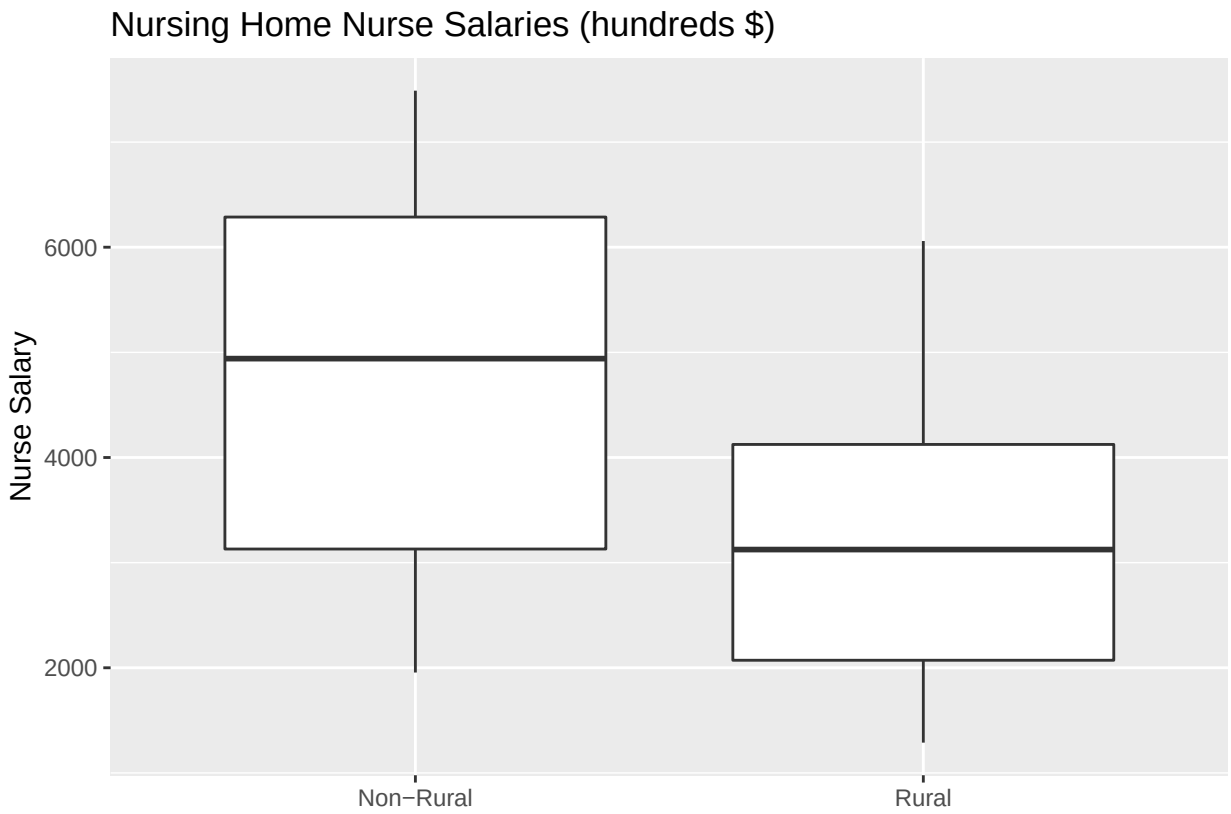
Nursing home salaries by census in Annual total patient days (in hundreds) and Nursing home size (by number of beds). There appears to be a moderately strong correlation between annual total patient days and average nurse salary, with at least one apparent outlier in terms of high patient days and low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



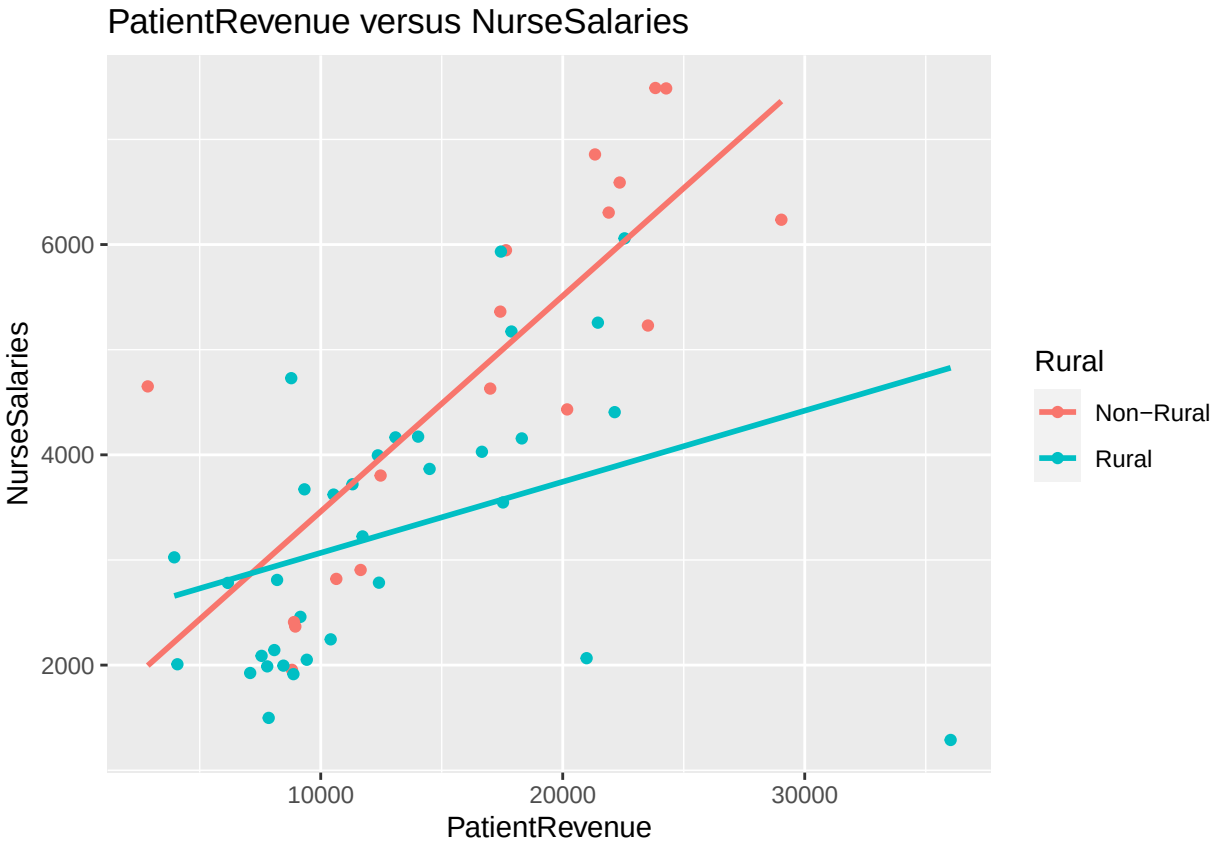
Boxplot demonstrating differences in Institutional nurse's salaries in New Mexico for Rural Areas compared with Non-Rural: Based on the box plot, it appears there is a greater variability for non-rural nurse salary. The nurses in non-rural regions also have a higher median salary.

```
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

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]
```

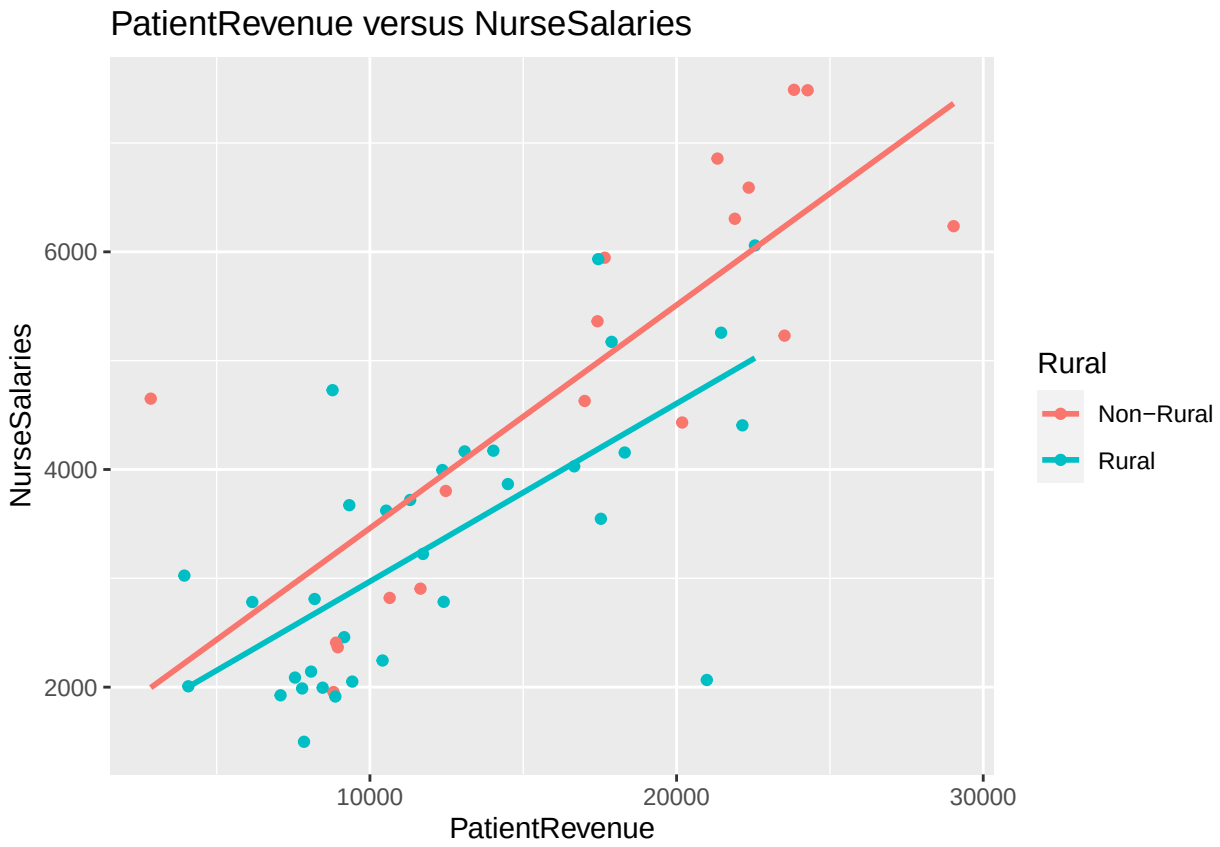
```
##        26        31
## 0.3189973 0.1873532
```

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.] do not 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
```

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.

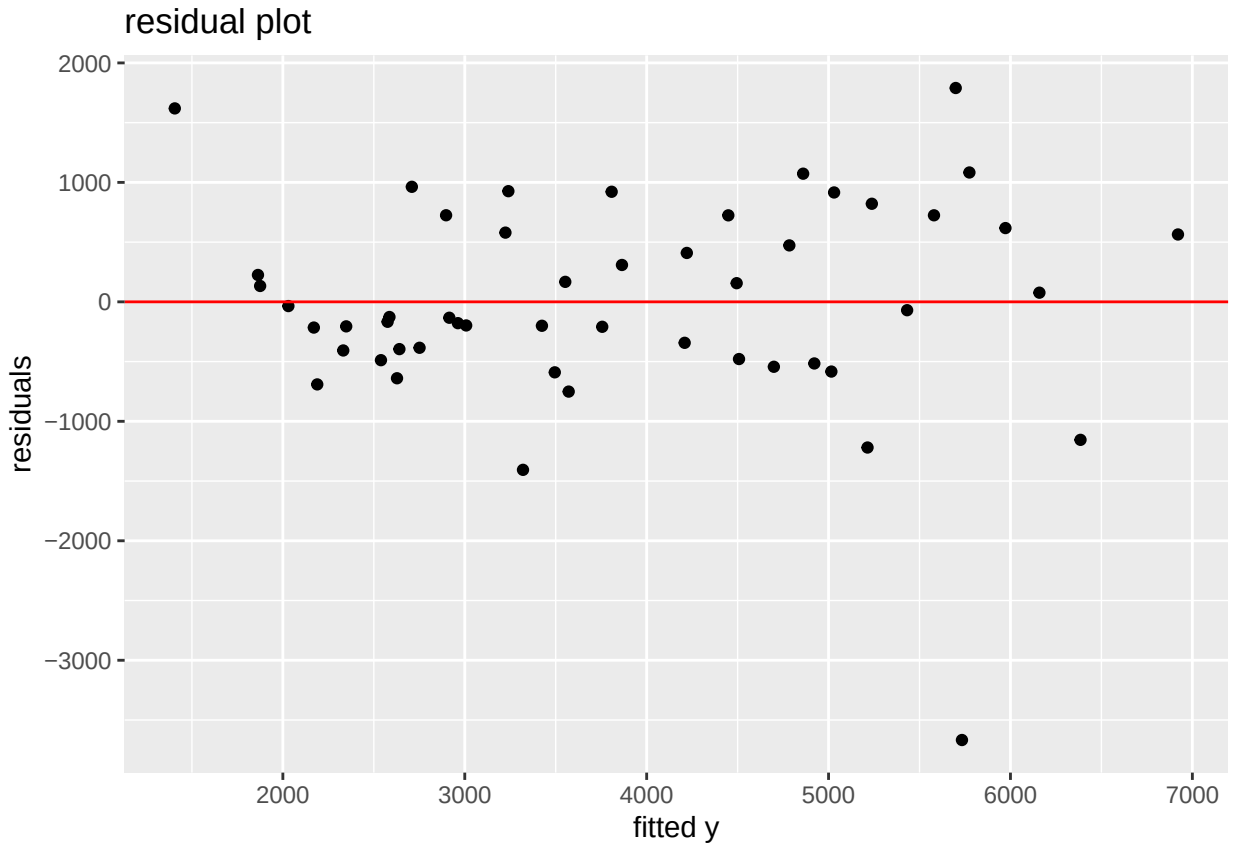
```
# requires package faraway
vif(reduced)
```

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

Checking linear regression assumptions:

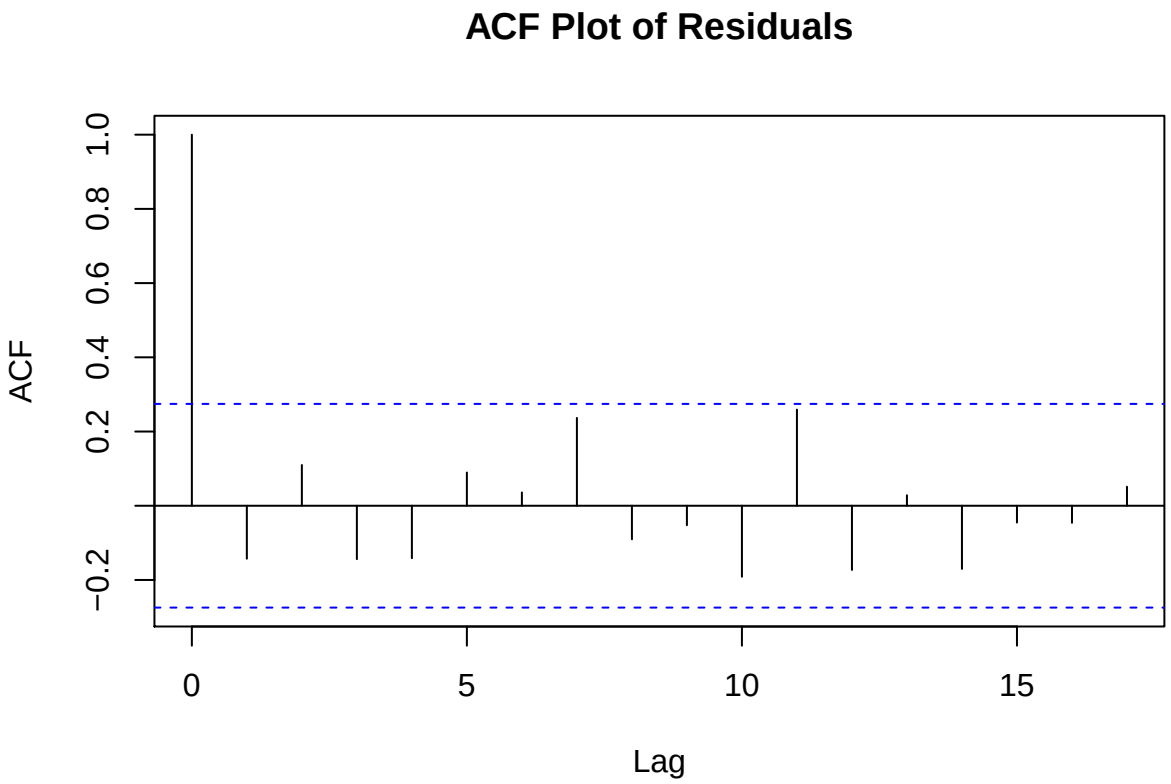
Assumptions 1/2 appear to be generally met: Mean of Errors = 0, constant variance.

```
# 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")
```



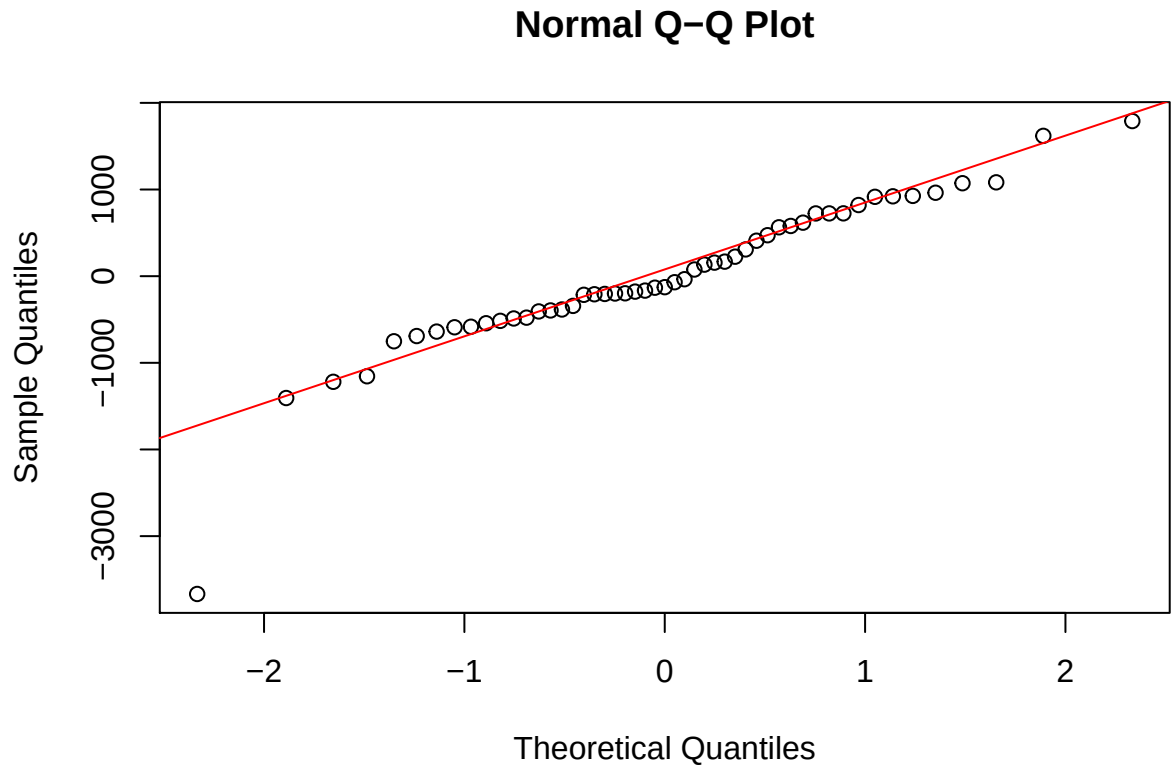
ACF Plot: no autocorrelation seen.

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



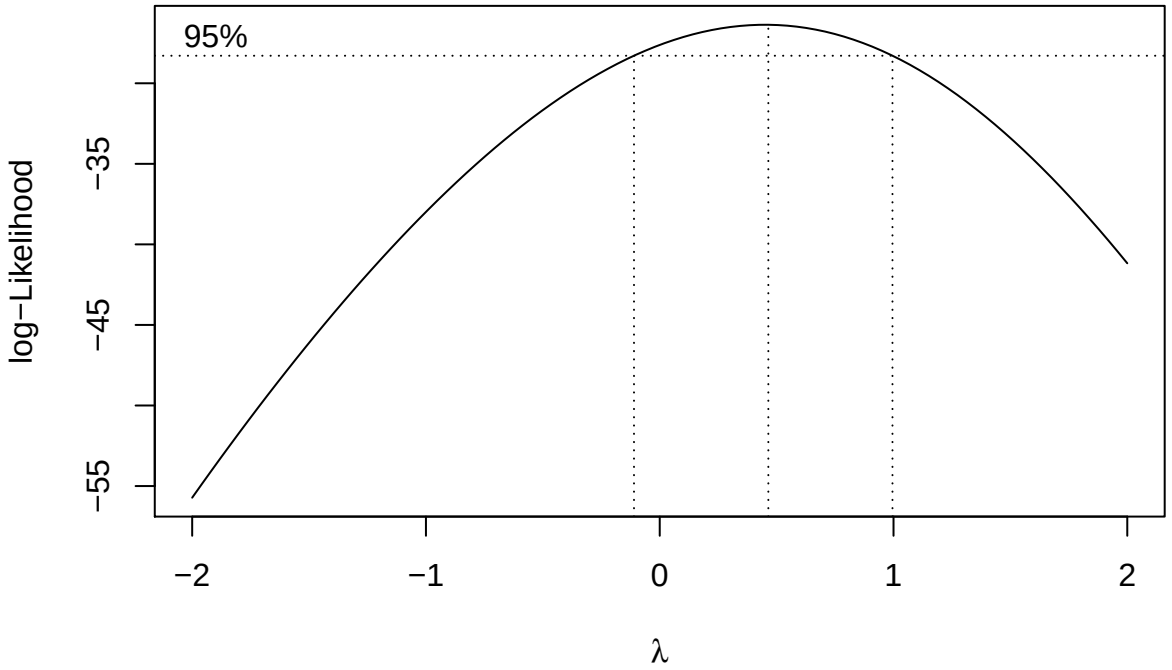
Errors are fairly normally distributed.

```
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 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), , FacilitiesExpenditure = hundreds of $.

Conclusions:

1) Patient 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.