Group 2: Objective 1 Multiple Linear Regression

Greg Madden

2022-04-19

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

```
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:

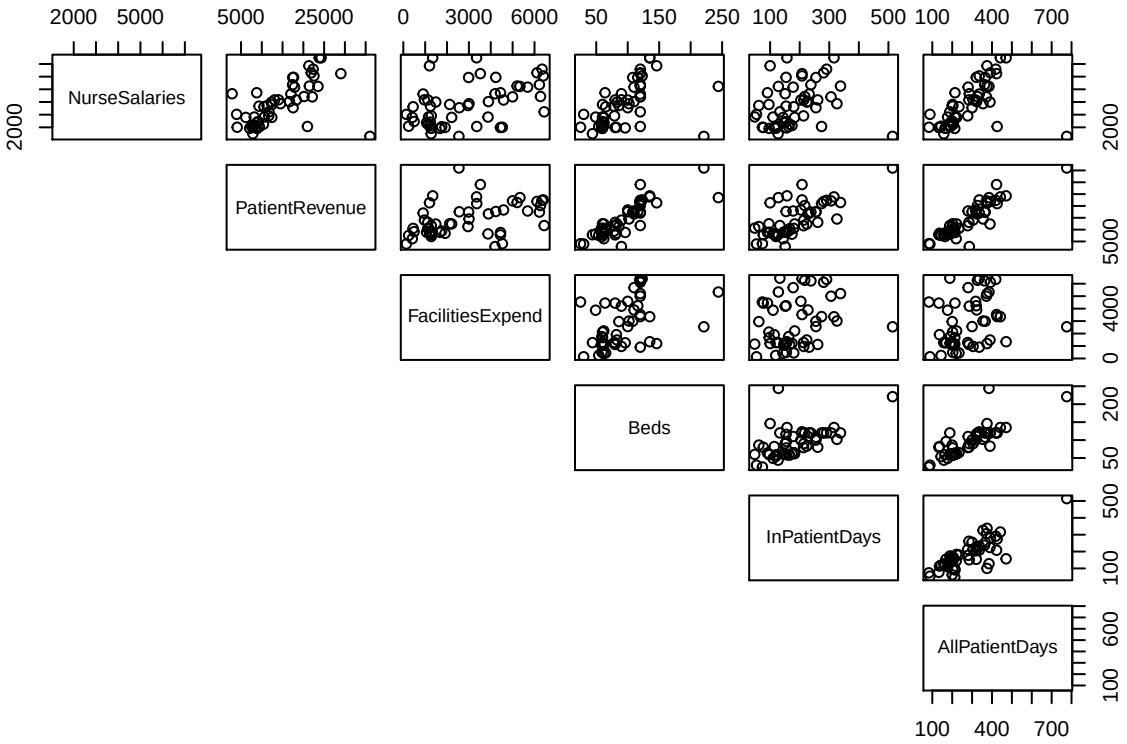
Based on the scatter plots and correlation table, it appears that Nurse Salaries has a moderate correlation with Beds, All Patient Days, and Patient Revenue. There also appears to be a strong linear relationship between Beds and AllPatientDays, Beds and Patient Revenue, In Patient Days and All Patient Days, In Patient Days and Patient Revenue, and All Patient Days and Patient Revenue. Further analysis is needed to determine if there is multicollinearity among the predictors.

```
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
```

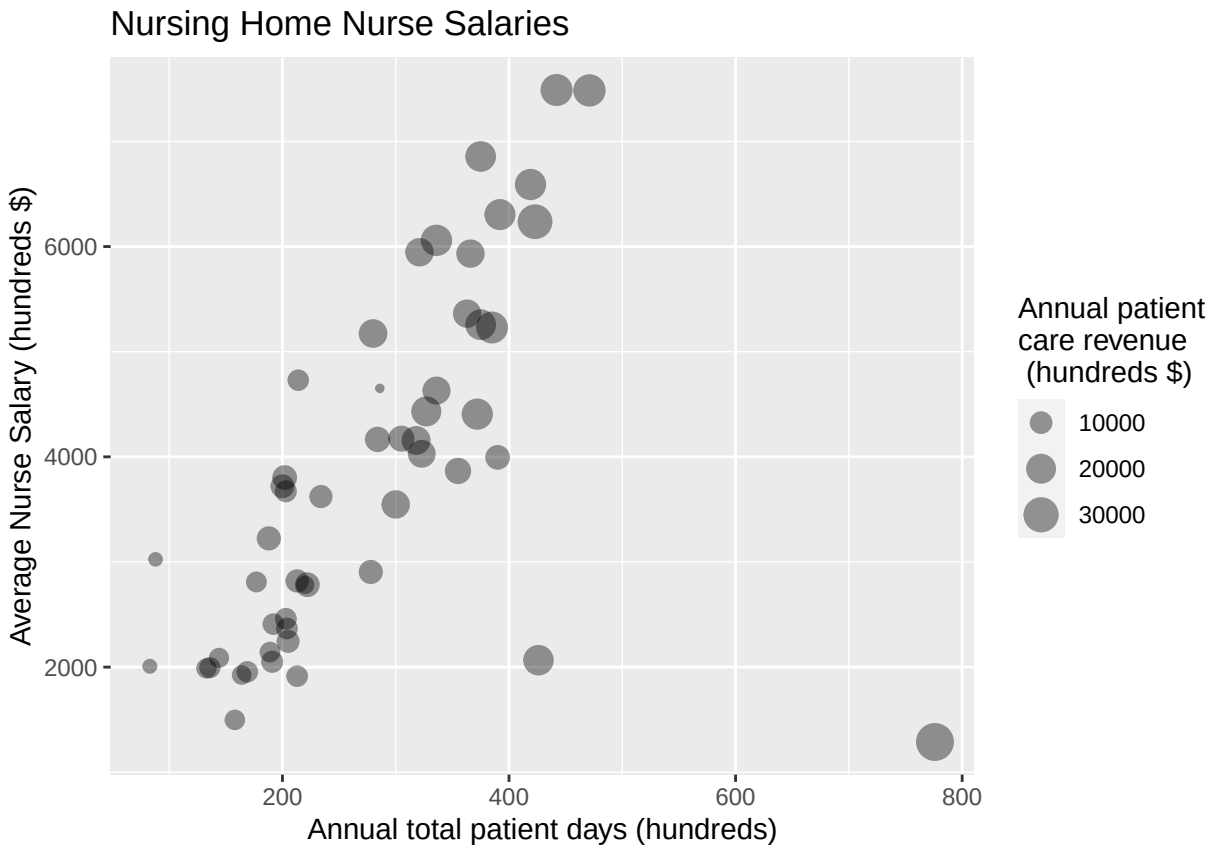
```
## FacilitiesExpend 0.4602559     0.2583959      0.3047354      0.4337859
##                  NurseSalaries FacilitiesExpend
## Beds                 0.5094241        0.4602559
## InPatientDays        0.2541355        0.2583959
## AllPatientDays       0.5153965        0.3047354
## PatientRevenue       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)
```



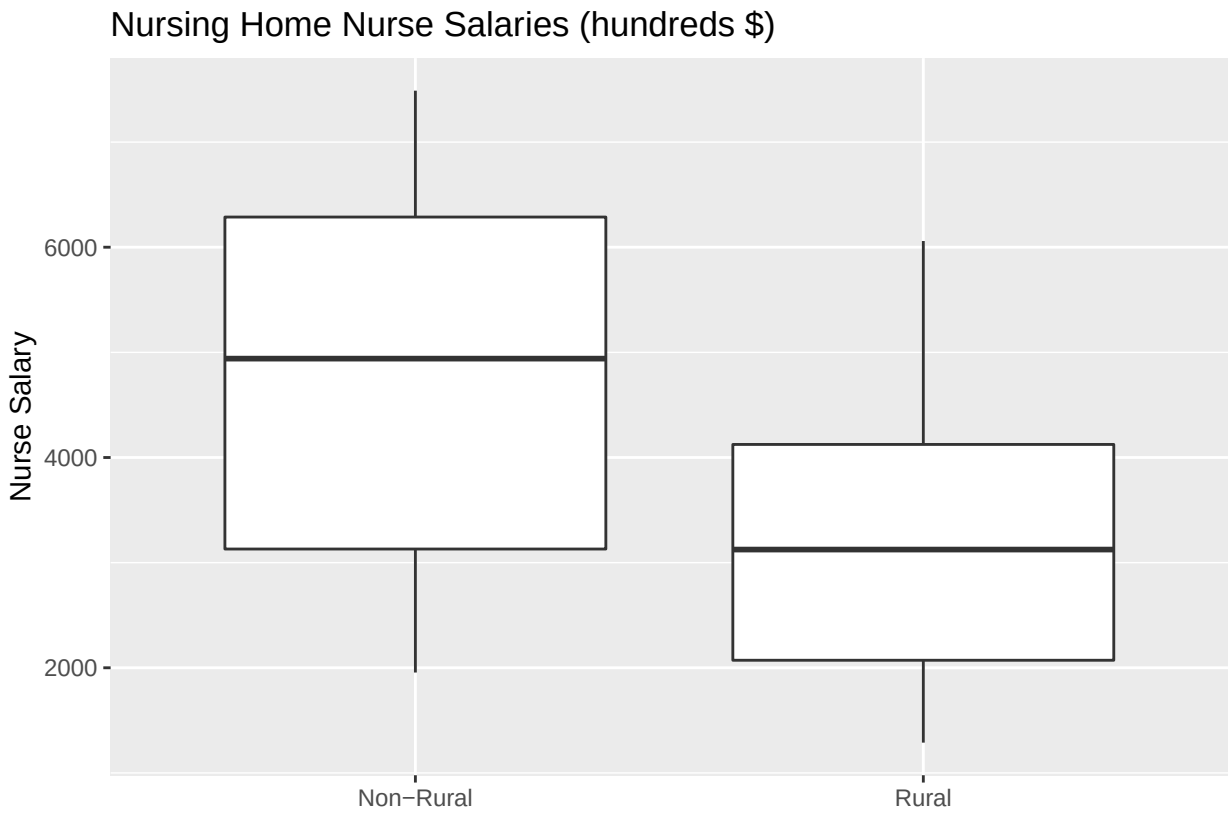
Nursing home salaries by census in Annual total patient days (in hundreds) and Nursing home size (by number of beds). There appears to be a moderately strong correlation between annual total patient days and average nurse salary, with at least one apparent outlier in terms of high patient days and low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

```
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 \nc
```



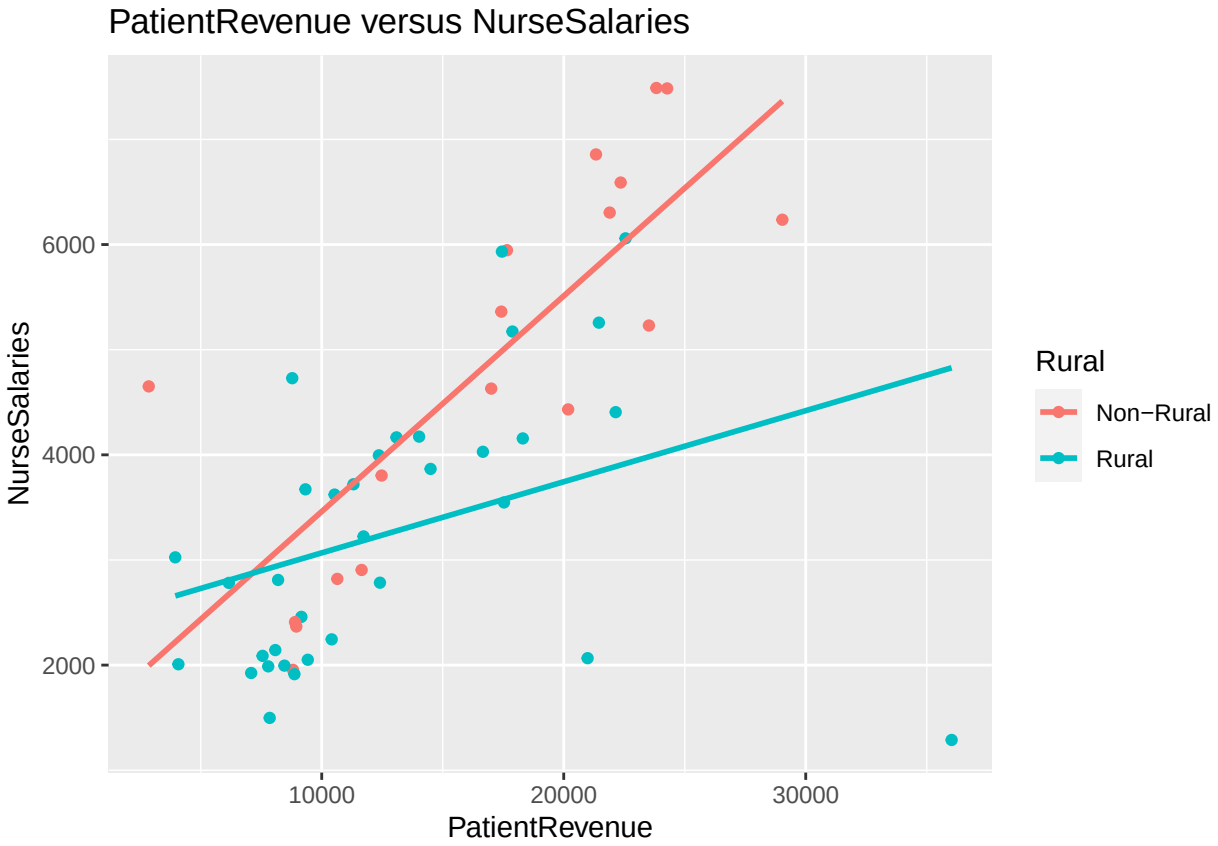
Boxplot demonstrating differences in Institutional nurse's salaries in New Mexico for Rural Areas compared with Non-Rural: Based on the box plot, it appears there is a greater variability for non-rural nurse salary. The nurses in non-rural regions also have a higher median salary.

```
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.

```
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 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

```
# 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=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)
```

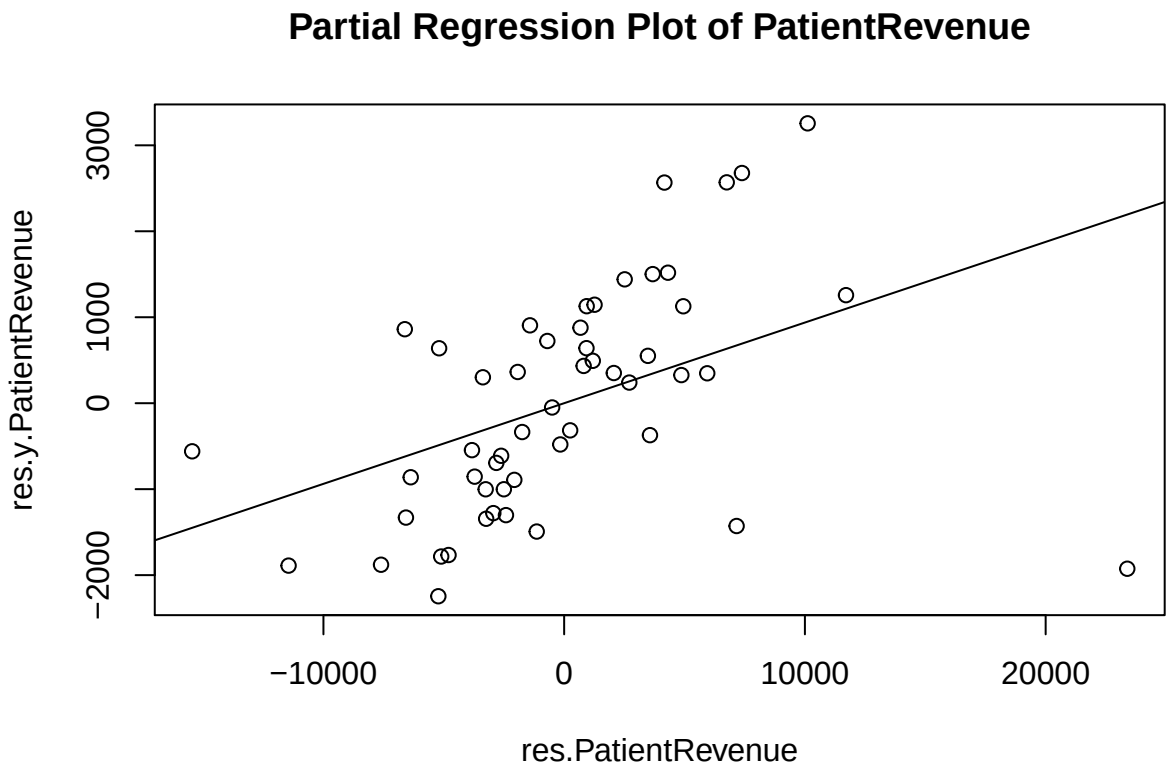
```
##
## Call:
## lm(formula = NurseSalaries ~ PatientRevenue + FacilitiesExpend +
##     Rural, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4120.3  -708.2   -71.7   833.6  2306.6
```

```
##
## 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
```

Partial residual plot for PatientRevenue

Since the plots are evenly scattered across the regression line, PatientRevenue should be added as a linear term.

```
## Create partial regression plot for passing yards, x2, to see if a non linear
## terms should be used.
result.y.PatientRevenue <- lm(NurseSalaries ~ FacilitiesExpend + Rural, data = Data)  ##fit y against o
result.PatientRevenue <- lm(PatientRevenue ~ FacilitiesExpend + Rural, data = Data)  ##fit PatientReven
res.y.PatientRevenue <- result.y.PatientRevenue$residuals  #store the residuals. info in y not explaine
res.PatientRevenue <- result.PatientRevenue$residuals  ##store residuals. info in PatientRevenue not ex

# partial regression plot for x2 creating regression line
regPatientRevenue = lm(PatientRevenue ~ FacilitiesExpend + Rural, Data)
regyPatientRevenue = lm(NurseSalaries ~ FacilitiesExpend + Rural, Data)

plot(res.PatientRevenue, res.y.PatientRevenue, main = "Partial Regression Plot of PatientRevenue")
lmx2 = lm(regyPatientRevenue$residuals ~ regPatientRevenue$residuals)
abline(lmx2)
```

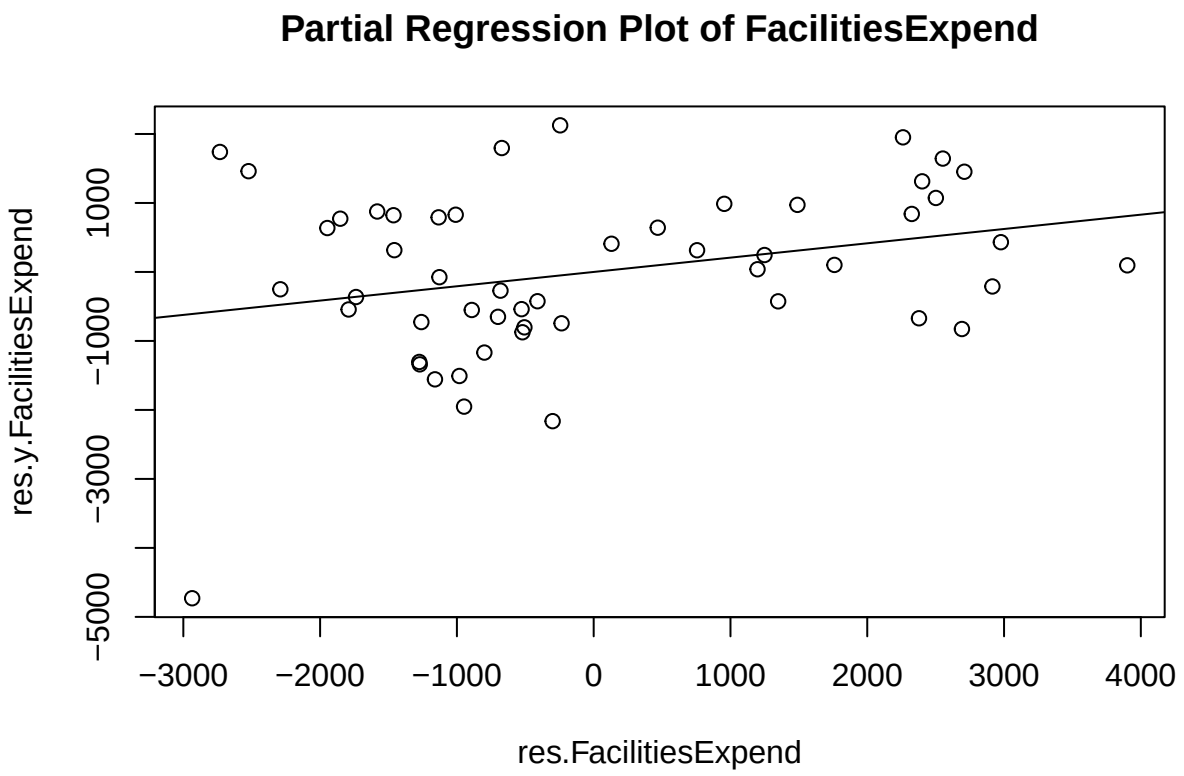



```
# abline(h=0)
```

Partial residual plot for FacilitiesExpend

Since the plots are evenly scattered across the regression line, FacilitiesExpend should be added as a linear term.

```
## Create partial regression plot for passing yards, x2, to see if a non linear
## terms should be used.
result.y.FacilitiesExpend <- lm(NurseSalaries ~ PatientRevenue + Rural, data = Data)  ##fit y against o
result.FacilitiesExpend <- lm(FacilitiesExpend ~ PatientRevenue + Rural, data = Data)  ##fit Facilities
res.y.FacilitiesExpend <- result.y.FacilitiesExpend$residuals  #store the residuals. info in y not expl
res.FacilitiesExpend <- result.FacilitiesExpend$residuals  ##store residuals. info in FacilitiesExpend

# partial regression plot for x2 creating regression line
regFacilitiesExpend = lm(FacilitiesExpend ~ PatientRevenue + Rural, Data)
regyFacilitiesExpend = lm(NurseSalaries ~ PatientRevenue + Rural, Data)
# plot
plot(res.FacilitiesExpend, res.y.FacilitiesExpend, main = "Partial Regression Plot of FacilitiesExpend"
lmx2 = lm(regyFacilitiesExpend$residuals ~ regFacilitiesExpend$residuals)
abline(lmx2)
```



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

The PRESS Statistic is 100651740. The model might be able to explain 28.335% of the variability in the new observations (not great). The R2 is 0.4939. Both values are fairly close to each other, so overfitting is not a major concern.

```
# function to calculate PRESS Statistic
PRESS <- function(model) {
    ## get the residuals from the linear.model.  extract hat from lm.influence
    ## to obtain the leverages
    i <- residuals(model)/(1 - lm.influence(model)$hat)
    ## calculate the PRESS by squaring each term and adding them up
    sum(i^2)
}
# summary(lm.y)
PRESS(reduced)
```

```
## [1] 100651740
```

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

```
## [1] 0.283351
```

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
```

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.small.rural <- Data[-26, ]
reduced2 <- lm(NurseSalaries ~ PatientRevenue + FacilitiesExpend + Rural, data = Data.small.rural)
summary(reduced2)
```

```
##
## Call:
## lm(formula = NurseSalaries ~ PatientRevenue + FacilitiesExpend +
##     Rural, data = Data.small.rural)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2812.2  -720.7  -152.4   690.0  2051.3
```

```
##
## Coefficients:
##                     Estimate Std. Error t value    Pr(>|t|)
## (Intercept)      1720.87508  464.65549   3.704    0.000558 ***
## PatientRevenue      0.16870    0.02721   6.200 0.000000134 ***
## FacilitiesExpend    0.09472    0.08214   1.153    0.254674
## RuralRural       -701.52213  311.83402  -2.250    0.029193 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 991 on 47 degrees of freedom
## Multiple R-squared:  0.6554, Adjusted R-squared:  0.6334
## F-statistic:  29.8 on 3 and 47 DF,  p-value: 0.00000000006086
```

Based on the t-test above, we will drop FacilitesExpend.

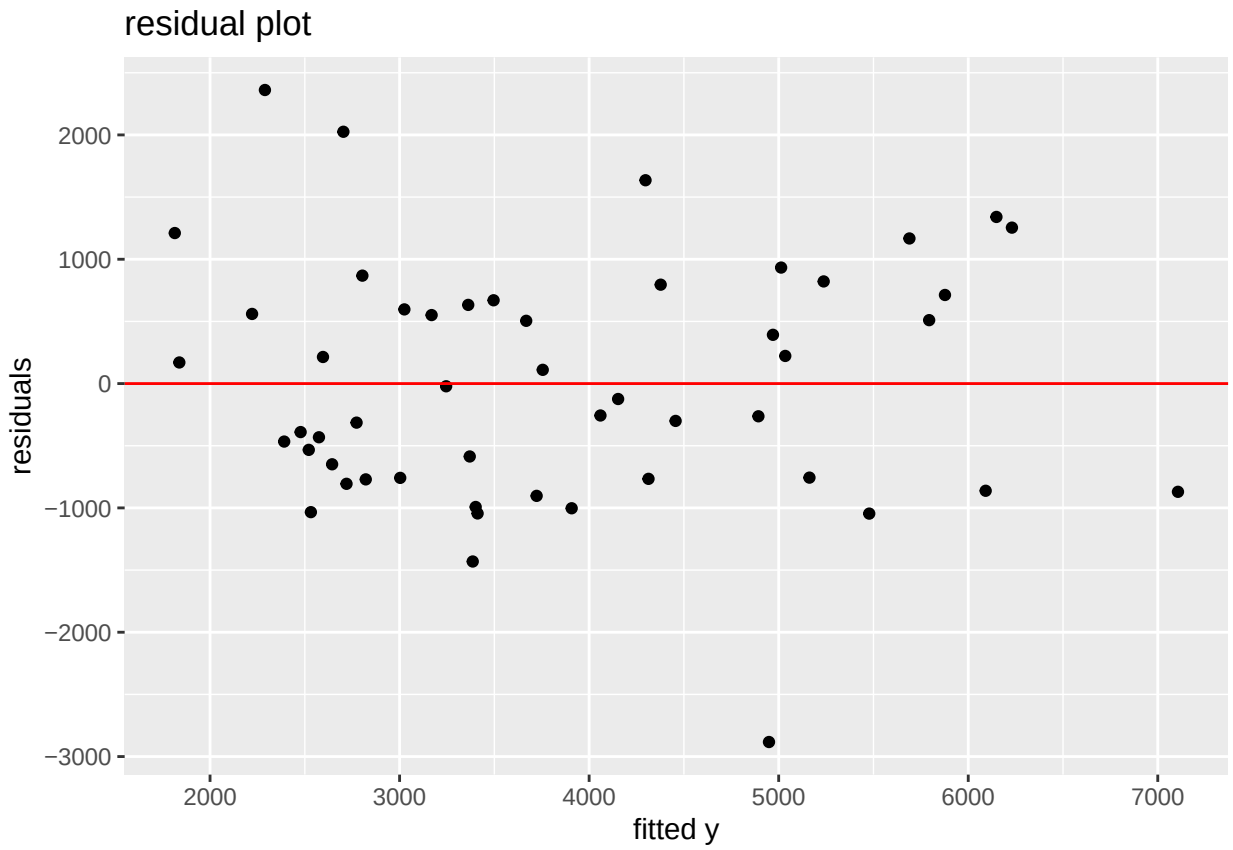
```
reduced3 <- lm(NurseSalaries ~ PatientRevenue + Rural, data = Data.small.rural)
summary(reduced3)
```

```
##
## Call:
## lm(formula = NurseSalaries ~ PatientRevenue + Rural, data = Data.small.rural)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2882.9  -762.3  -124.0   651.5  2361.4
##
## Coefficients:
##                   Estimate Std. Error t value     Pr(>|t|)
## (Intercept)    1764.75448  464.68295   3.798     0.000411 ***
## PatientRevenue    0.18397    0.02385   7.713 0.0000000006 ***
## RuralRural     -676.94337  312.17164  -2.168     0.035107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 994.4 on 48 degrees of freedom
## Multiple R-squared:  0.6457, Adjusted R-squared:  0.6309
## F-statistic: 43.73 on 2 and 48 DF,  p-value: 0.00000000001535
```

Checking other assumptions:

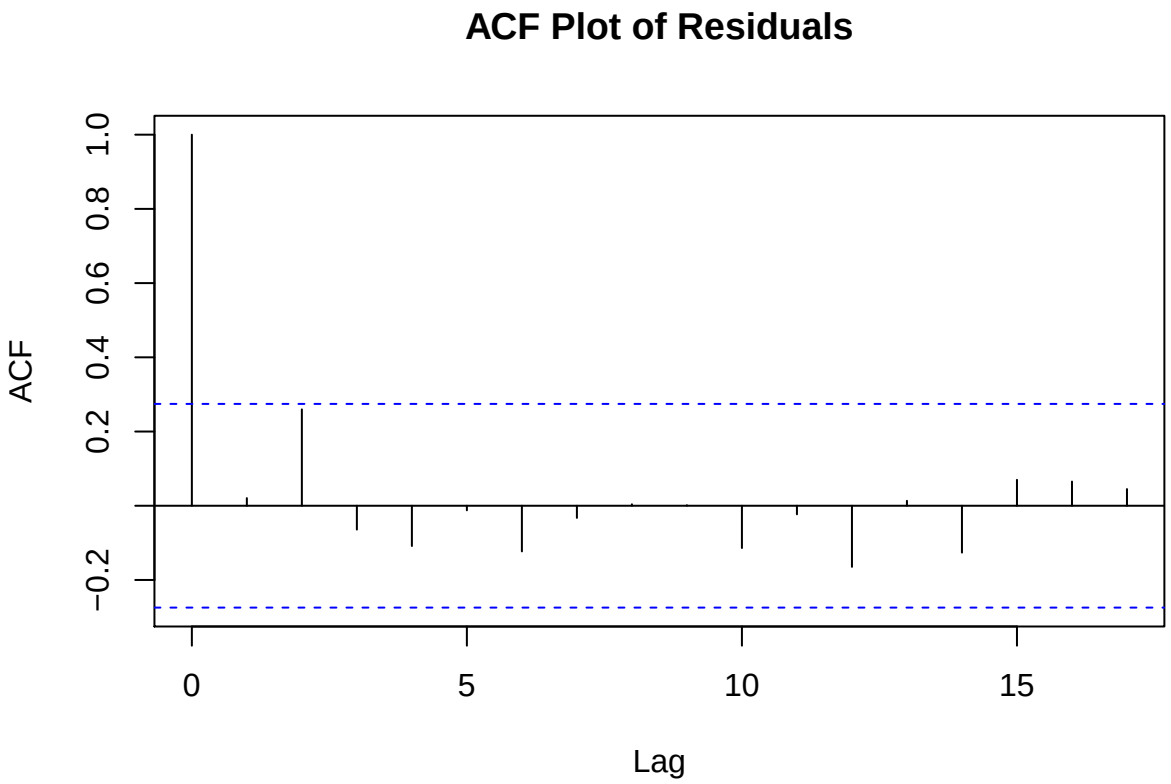
Assumptions 1/2 appear to be met: Mean of Errors = 0, constant variance.

```
# storing fitted y residuals
yhat <- reduced3$fitted.values
res <- reduced3$residuals
# add to data frame

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

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



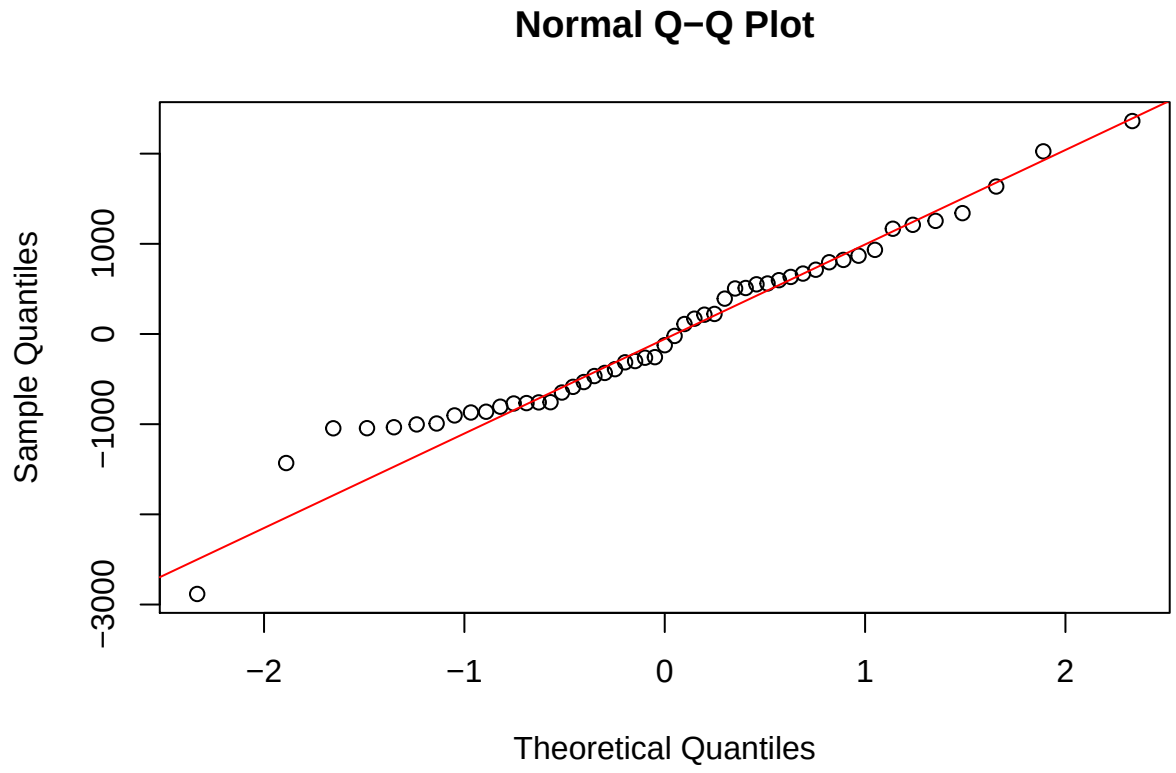
ACF Plot: no autocorrelation seen.

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



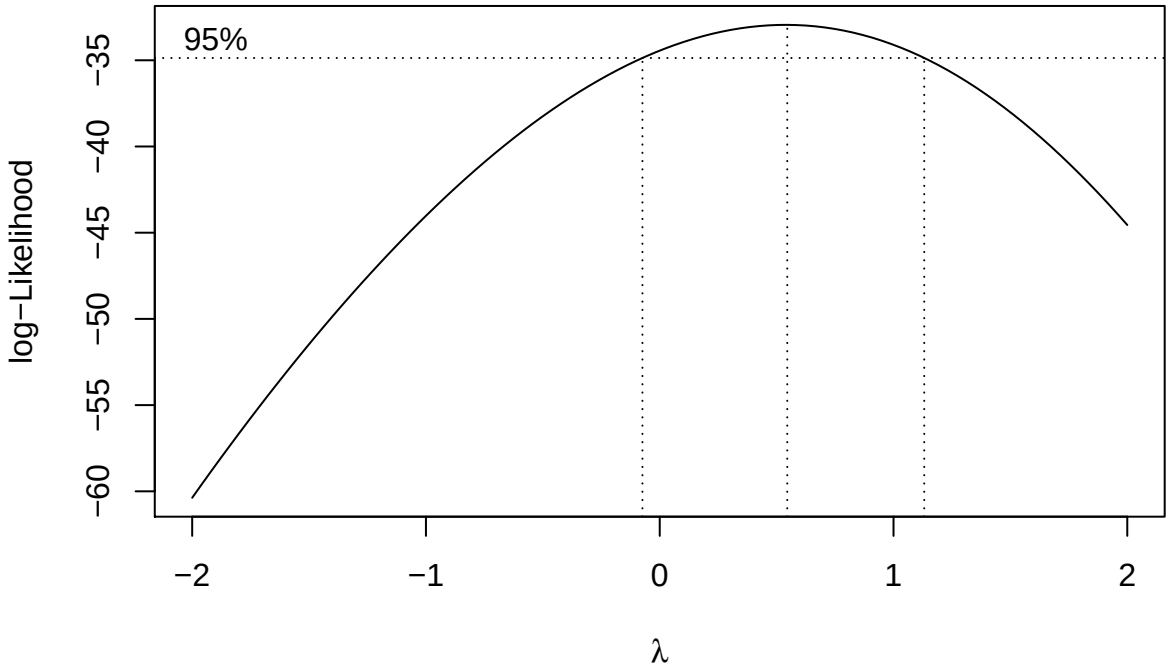
Errors are fairly normally distributed.

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



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

```
boxcox(reduced3)
```



Calculating the $R^2_{prediction}$, the final model might be able to explain 58.18% of the variability in the new observations (as long as they are not rural facilities >200 beds). The R2 is 0.6457. Both values are fairly close to each other, so overfitting is not a major concern.

```
reduced3 <- lm(NurseSalaries ~ PatientRevenue + Rural, data = Data.small.rural)
PRESS(reduced3)
```

```
## [1] 56017052
```

```
## Find SST
anova_result <- anova(reduced3)
SST <- sum(anova_result$"Sum Sq")
## R2 pred
Rsq_pred <- 1 - PRESS(reduced3)/SST
Rsq_pred
```

```
## [1] 0.5818
```

So, our final regression equation is:

$\hat{y} = 1764.75448 + 0.18397 * PatientRevenue(hundreds\ of\ \$) - 676.94337 * Rural$

Where bed size for rural institutions is <200, $\hat{y}$***=Predicted Nursing Salary*** $(in\ hundreds\ of\ \$)$***, Rural Nursing Home = 1, and Non-Rural = 0.***

Conclusions:

-Patient revenue and rural versus non-rural status appear to be the most important factors in determining nursing salaries. -Based on the $R^2_{prediction}$, our final model summarized above might be able to explain

58.18% of the variability in nursing salaries, as long as they are not rural facilities >200 beds. -The above model suggests that while holding patient revenue constant, rural nursing homes pay their nurses in total $67,694 less per year.