STAT 425: Case Study

Afnan Dzaharudin (afnanfd2), Linfeng Wang (linfeng7), Bowen Liu (bowenl5), Qingzhe Lin (qingzhe2)

Preface

We have been given a dataset of 440 rows and 17 columns of data. Our goal is to build a predictive model for the number of active physicians in a given area, based on our choice of predictor variables from this dataset. To achieve this, we utilize the methods we have learned from the STAT 425 lectures up to the time of submission of this report.

Data Analysis & Cleaning

After spending some time building models in a more haphazard manner for the sake of testing, we realized that it was wiser to take a look at the nature of the data we were handling and using as predictors, before actually constructing a predictive model.

First, we realized that `id` was an indexing number and nothing more, so we removed it from our dataset as it held no relation to other data. We also noticed that `county`, `state`, and `region` are all indicators of location, though at varying generality. We chose to remove `county` and `state` on the grounds that they were too specific for a predictive model, and kept `region` since we believed it was general enough.

This leaves us with `pop_total`, `land_area`, `pct_18_24`, `pct_65_older`, `beds`, `crime`, `pct_highschool`, `pct_bachelor`, `pct_below_poverty`, `pct_unemployed`, `per_capita_income`, `total_personal_income`, and `region`.

For these remaining variables, We went on to look at their distributions and correlations. A normal-looking distribution of variables are preferable, as skewness tends to mess with the nature of a model's residuals.

We found that `pop_total`, `land_area`, `beds`, `crime`, and `total_personal_income` **were highly left-skewed** (so is our response variable, `n_physicians`, which we note to handle later). As for correlations, we again found `pop_total`, `beds`, `crime`, and `total_personal_income` to be highly correlated (≥ 0.9 in absolute value). We infer that this is due to them being **measured in totals**, so yes, a larger population would generally lead to more beds, crime, etc. in that county. To remedy both the skewness and high correlation of these data, we opt to `mutate` new variables that represent proportions of values instead; "Percentage Beds", `pct_beds` is the ratio of hospital beds to population count; "Percentage Crime", `pct_crime`, is the ratio of occurrences of crime to population count; `per_capita_income` is already a per-person representation of total income, so we discard `total_personal_income`; Finally, "Population Density", `pop_density` is the ratio of land area to population count. We find that `pop_density` is still skewed, but at least uncorrelated to other predictors, so we transform it further by taking its `log()`, forming `log_pop_density`.

This gives us a set of predictor variables that are no longer highly correlated, and we are confident in proceeding to build our predictive model.

Model Selection

Full Model

At this stage we do not know much about our predictor variables, nor do we know which would be significant predictors for our model. Therefore, we will begin with a full model that uses all predictor variables, and work our way down.

Our observations are that not many predictors pass a **single-predictor test** on $H_0 : \hat{\beta}_i = 0, H_a : \hat{\beta}_i \neq 0$ at $\alpha = 0.05$. Only `per_capita_income`, `log(pop_density)`, `region`, `pct_18_24` and `pct_below_poverty` pass these as they have a `Pr(>|t|)` value less than `0.05`. The **overall F-test** *does* pass at $\alpha = 0.01$, meaning H_a: at least one predictor is nonzero. Overall, this tells us that it is likely that *many* of the coefficients are not statistically significant in our regression model, but at least *some* are.

Our plots however show that the residuals are not distributed equally around the fitted values, nor do they fit the QQ-plot well. For the Kolmogorov-Smirnov test (chosen over the Shapiro-Wilk test since $n = 440 > 50$) we obtain an extremely small p-value close to 0, therefore we must assume that **the model residuals are not normally distributed**. However, the studentized Breusch-Pagan test for constant variance outputs a p-value of approximately 0.03761. Therefore at $\alpha = 0.01$ confidence we can conclude that **the residuals of the model are at least with constant variance**.

Our next step is to work towards the model of best fit by reducing this model.

Reduction by Permutation Test

Because the current model's residuals are likely not normally distributed, but with constant variance, we perform a **Permutation Test** in place of the partial F-test with a chosen full model and null model.

If we fail to reject the null hypothesis, we can conclude that the null model (which is a smaller subset of the full model) is sufficient.

The null model we chose below is our final choice after many trials, aiming to reduce the model so that the remaining coefficients have extremely low single-test p-values.

More specifically, now, our test will be, at $\alpha = 0.01$:

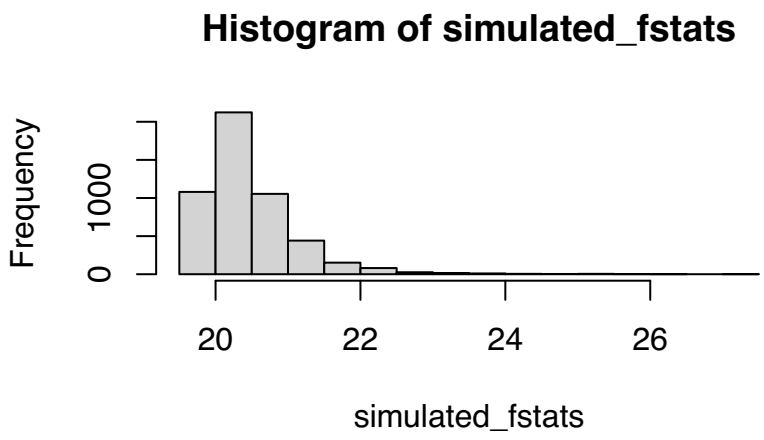
$$H_0 : \hat{\beta}_{\text{pct\_beds}} = \hat{\beta}_{\text{pct\_crime}} = \hat{\beta}_{\text{pct\_18\_24}} = \hat{\beta}_{\text{pct\_65\_older}} = \hat{\beta}_{\text{pct\_highschool}}$$
$$= \hat{\beta}_{\text{pct\_bachelor}} = \hat{\beta}_{\text{pct\_unemployed}} = 0$$

H_a : At least one of these $\hat{\beta}$ are nonzero

In other words, our null and full models are:

$$H_0 : \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_{\text{pct\_below\_poverty}} x_{\text{pct\_below\_poverty}} + \hat{\beta}_{\text{per\_capita\_income}} x_{\text{per\_capita\_income}}$$
$$+ \hat{\beta}_{\text{log(pop\_density)}} x_{\text{log(pop\_density)}} + \hat{\beta}_{\text{region}} x_{\text{region}}$$
$$H_a : \hat{y}_1 = \hat{y}_0 + \hat{\beta}_{\text{pct\_crime}} x_{\text{pct\_crime}} + \hat{\beta}_{\text{pct\_18\_24}} x_{\text{pct\_18\_24}} + \hat{\beta}_{\text{pct\_65\_older}} x_{\text{pct\_65\_older}}$$
$$+ \hat{\beta}_{\text{pct\_highschool}} x_{\text{pct\_highschool}} + \hat{\beta}_{\text{pct\_bachelor}} x_{\text{pct\_bachelor}} + \hat{\beta}_{\text{pct\_unemployed}} x_{\text{pct\_unemployed}}$$

We perform the Permutation Test with 5000 iterations. We assign F as the F-value of our orignal full model. We recreate the full model with permutations of our predictor variables that are *not* kept, obtaining 5000 F-values as $F_{Mi}, 1 < i < 5000$. Our p-value is $P(F_M > F) = \frac{\text{Number of } F_M < F}{5000}$ and our test is at $\alpha = 0.01$




```
## Full model F-statistic =  20.4398931022169 ; p-value =  0.4036
```

Therefore the probability we would observe an F-value as extreme as our full model's ($F = 20.44$) is 0.4036. We fail to reject the null hypothesis at $\alpha = 0.01$, and therefore conclude that our null model is sufficient.

We can move forward using the following reduced model:

```
n_physicians ~ pct_below_poverty + per_capita_income + log_pop_density + region
```

First Reduced Model

Now, all our coefficients are significant with extremely small p-values for a single predictor test (except `region`, which only passes 0.05 confidence, but we continue to keep this variable). However, we still have some problems: the residuals are still not evenly-distributed around the fitted values, and have some points that do not fit well around the QQ-plot line. We also must conclude from the low p-value in the studentized Breush-Pagan test for this model that **the residual variance is not constant**

We're happy with our predictor variables at least, so we proceed to treat the response variable. As was noted earlier, the `n_physicians` data is not well-distributed; it is highly left-skewed like some other predictors we saw. Therefore we apply a Box-Cox transformation to the variable.

```
## Lambda =  -0.105263157894737
```

We find this value of λ to be closer to 0 than -0.5, so we choose to set $\lambda = 0$. According to the rules of the Box-Cox transformation, we should be taking the log of the response, i.e. `log(n_physicians)`

Final Model

We move on to the following model:

```
log(n_physicians) ~ pct_below_poverty + per_capita_income + log_pop_density + region
```

We find much more appealing results. However, we notice that the intercept coefficient $\hat{\beta}_0$ is valued `-0.039` with a poor p-value. Because of this we set the intercept to 0 and so our final model is:

```
log(n_physicians) ~ 0 + pct_below_poverty + per_capita_income + log_pop_density + region
```

We observe that all the remaining coefficients are statistically significant according to single parameter tests. The model has a mutiple R^2 value of `0.9864`, which raised some suspicion for being so high, but we choose to accept it as the predictors are not highly correlated (though one pair has `0.7` correlation). The residuals of this model are much better-distributed around the fitted values, resembling the football shape we seek. Most of the residuals are close to the QQ-plot line, though there is some departure towards the end. The

distribution of the residuals seems bimodal, and still does not pass the Kolmogorov-Smirnov test at $\alpha = 0.01$ confidence. However, the studentized Breusch-Pagan test returns a p-value of `0.2414`, which means we can conclude at $\alpha = 0.01$ that the residuals are with constant variance.

We can conclude the model's residuals are **not normally distributed** and have **constant variance**. We choose this to be our final model as it has an R^2 value we are sufficiently confident with.

Diagnostics

We choose to do some further diagnostics to check for unusual observations by finding outliers, highly-influential points, high-leverage points.

Outliers We calculate the studentized residuals, and test them at 0.01 confidence using a Bonferroni-corrected confidence value of $\alpha = \frac{0.01}{4} = 0.0025$. We notice a few outliers that pass $t_{0.9975}(n - p - 1)$. The data points with studentized residuals that outlie come from Los Angeles, CA; Maricopa, AZ; San Bernardino, CA; Pima, AZ; Arlington County, VA; and Alexandria City, VA. We do not remove these points for we do not know why they deviate; further study of these counties is necessary to make that decision.

Highly-influential Points There are no highly-influential points; none of the Cook's Distance values exceed 1.

High-leverage Points We found the model to have a 37 high-leverage points, i.e. 37 leverage points out of 440 that are larger than $\frac{2p}{n}$, where p: number of coefficients and n: sample size. We use a `halfnorm` plot to show us the pattern of leverage points, then calculate a bound to identify 'bad' leverage points. According to this bound the data points considered to have 'bad leverage' are Nassau, NY, and Philadelphia, PA. Again we do not remove these as we do not know why exactly they deviate so much.

Conclusion

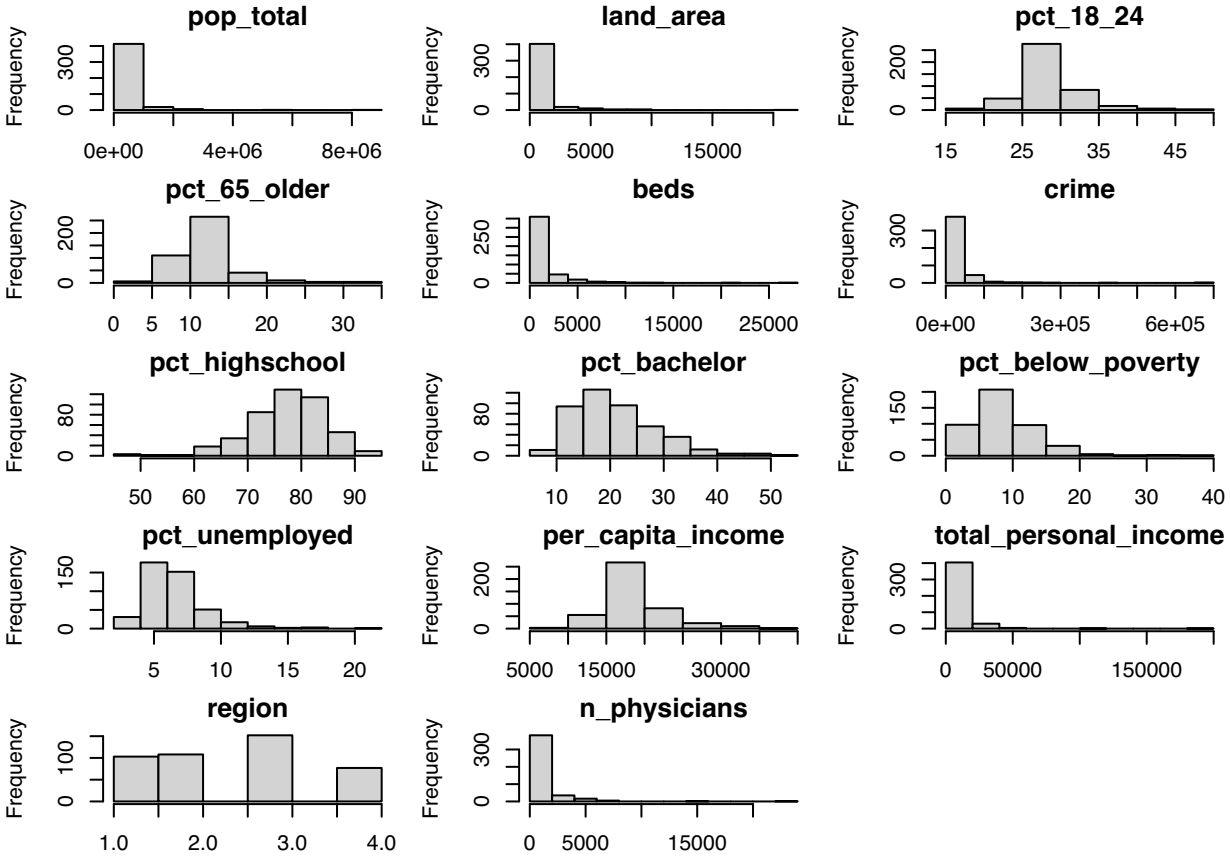
We believe the following linear regression formula to be sufficiently good for predicting the of the number of active physicians in a county:

$$\begin{aligned}[\text{Number of Physicians}] = \exp\Bigg(&0.067[\%\text{ Population Below Poverty}] + 1.1 \times 10^{-4}[\text{Per Capita Income}] \\ &+ 0.52 \log\left[\frac{[\text{Population Total}]}{[\text{Land Area}]}\right] + 0.16[\text{Region}]\Bigg)\end{aligned}$$

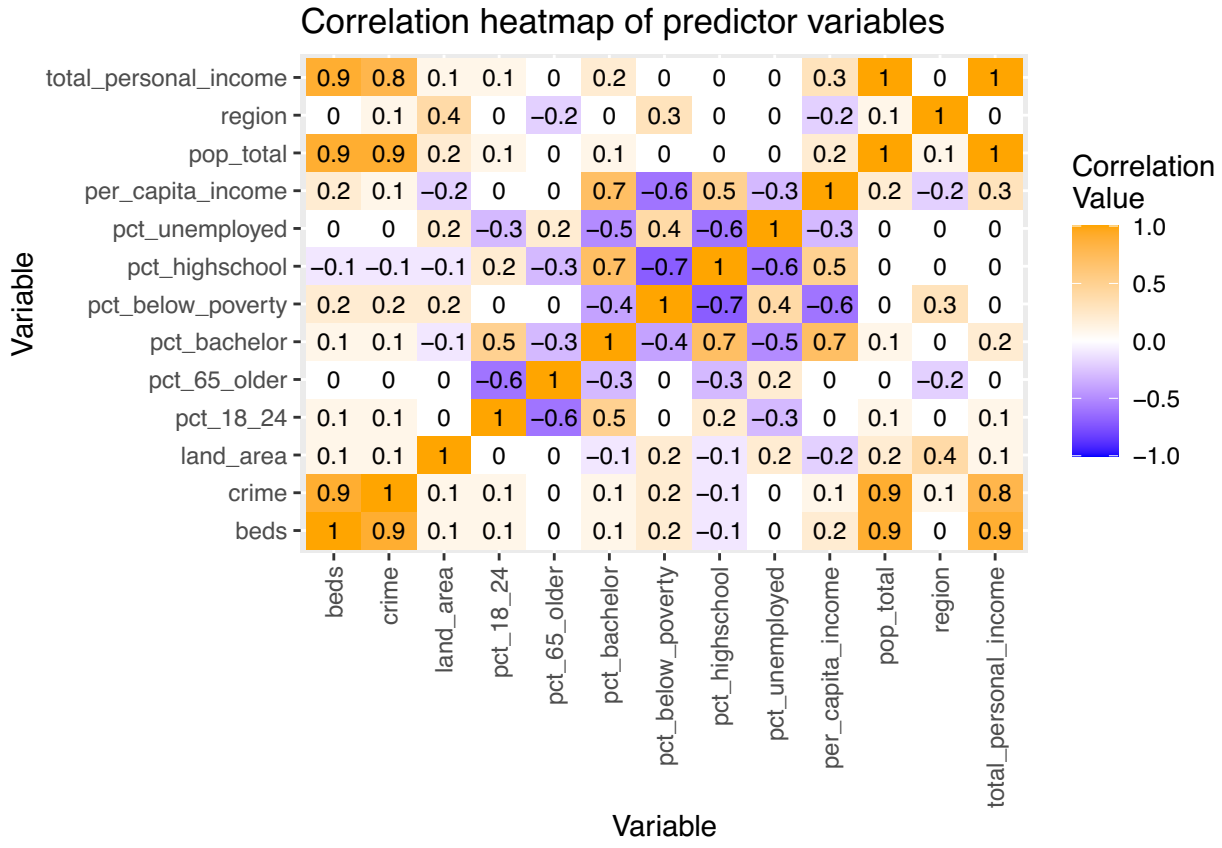
This model was obtained by first reinterpreting certain data that initially seemed too problematic to use in our model, then using tests to attempt to reduce the number of predictors in the model to a fewer number. We then also applied a treatment to our response, the number of active physicians, to improve the fit of this model. We have found certain points of data coming from certain counties that may have impacted our model's accuracy in a negative way, which could be grounds for those interested to study further.

Appendix

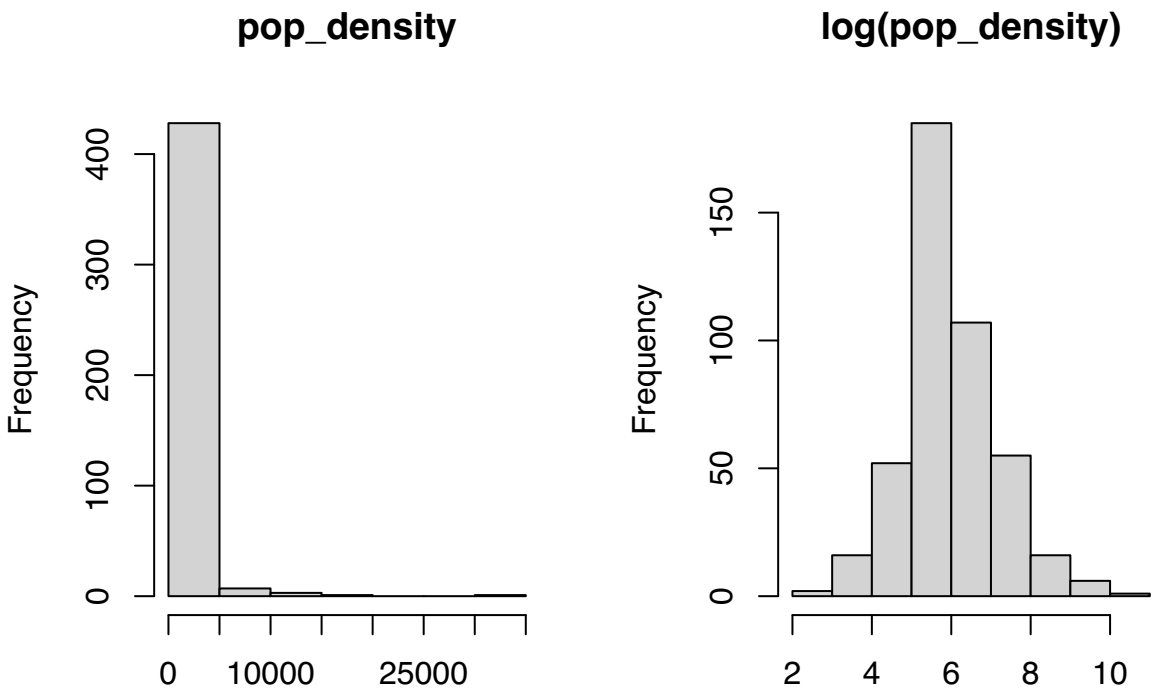
Distributions of Uncleaned, Initial Predictor Variables



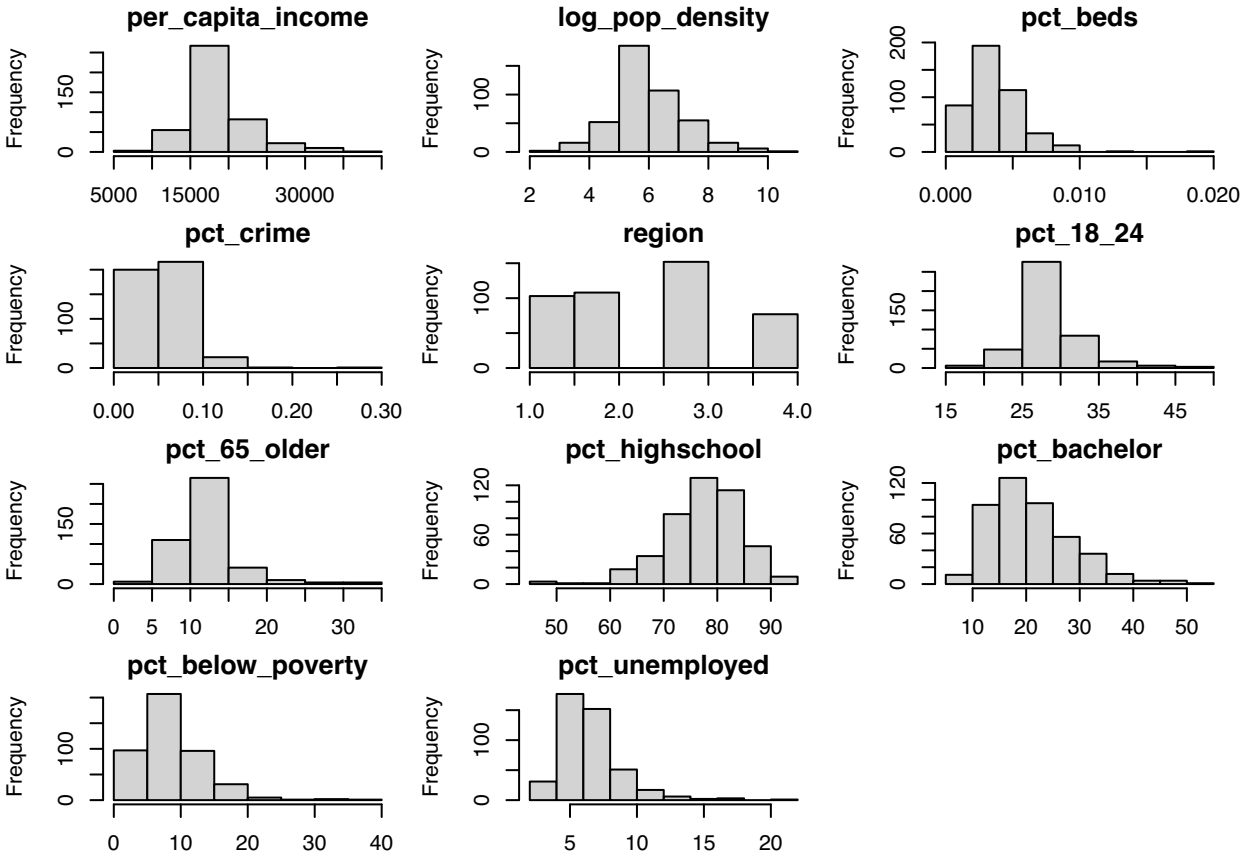
Correlation Heatmap of Uncleaned, Initial Predictor Variables



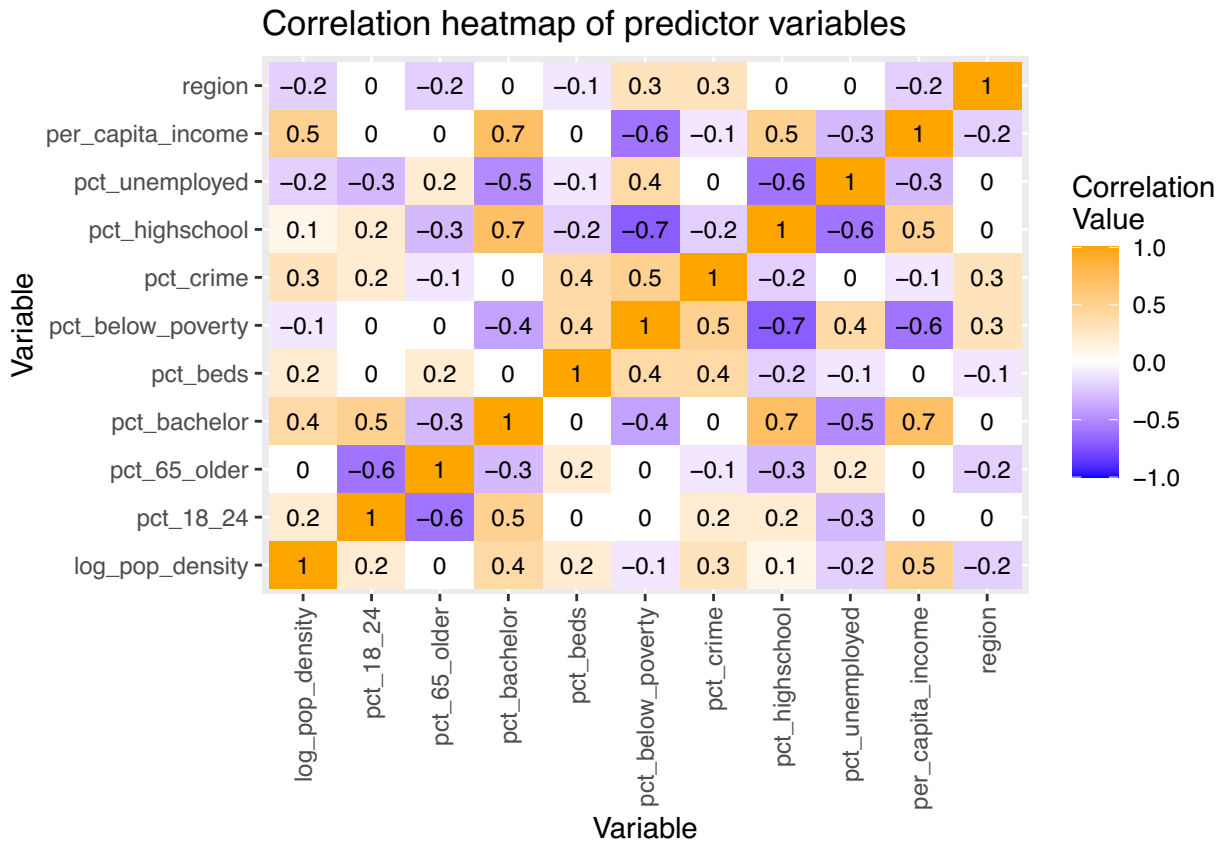
Distributions of `pop_density` and `log(pop_density)`



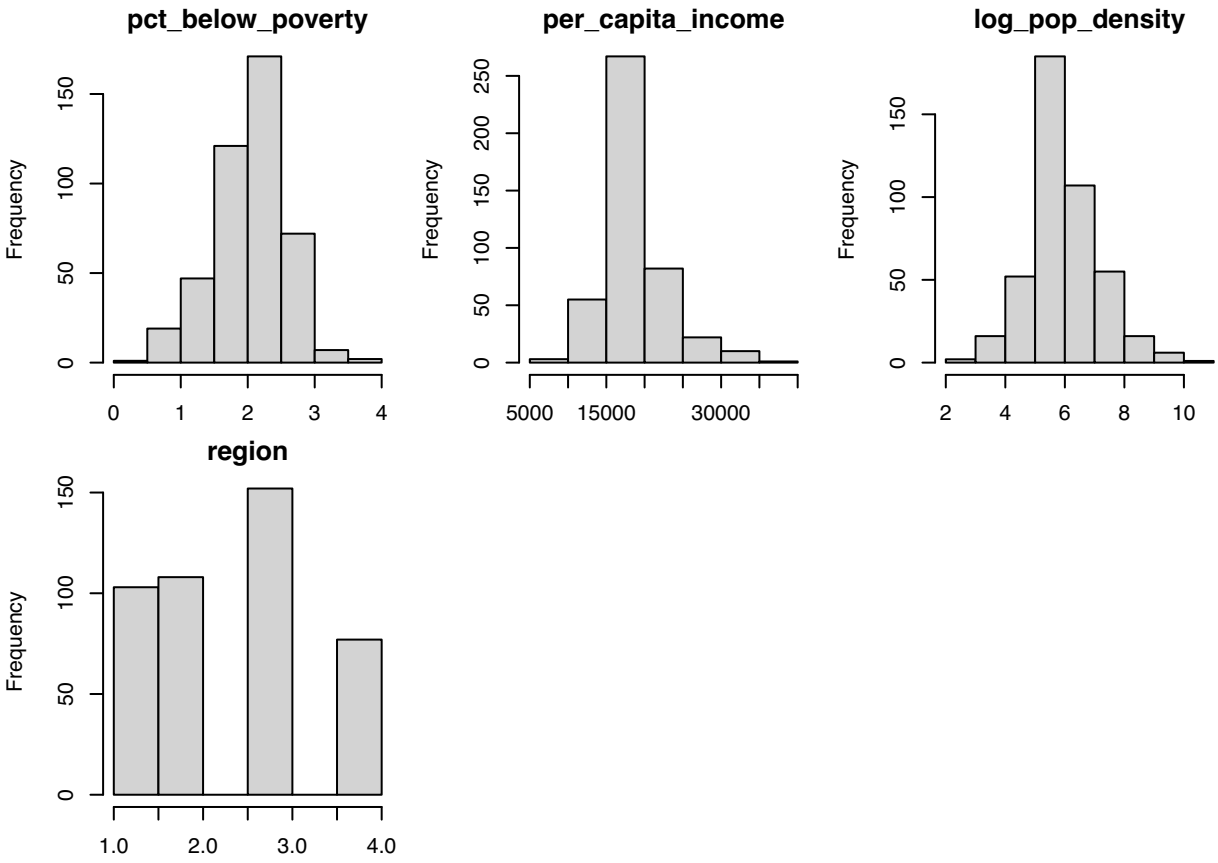
Distributions of Cleaned Data



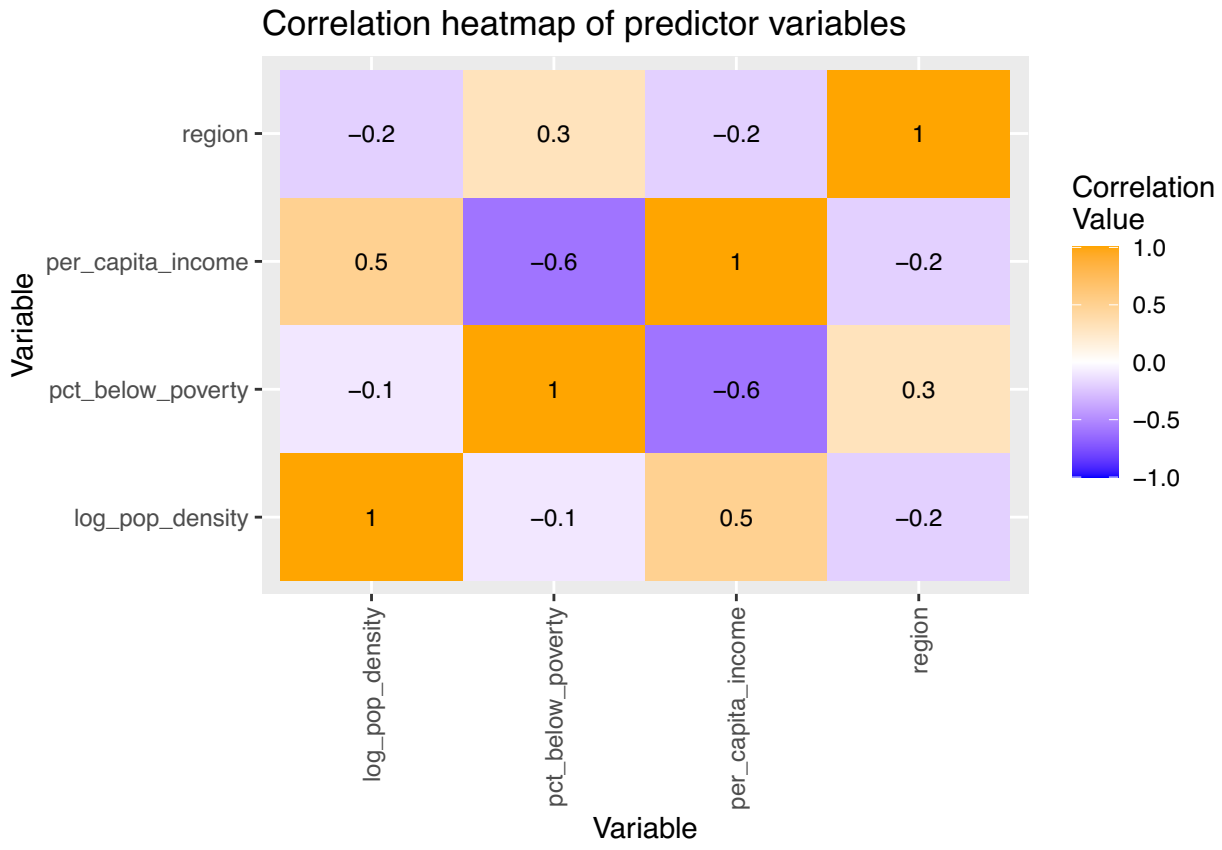
Correlation Heatmap of Cleaned Data



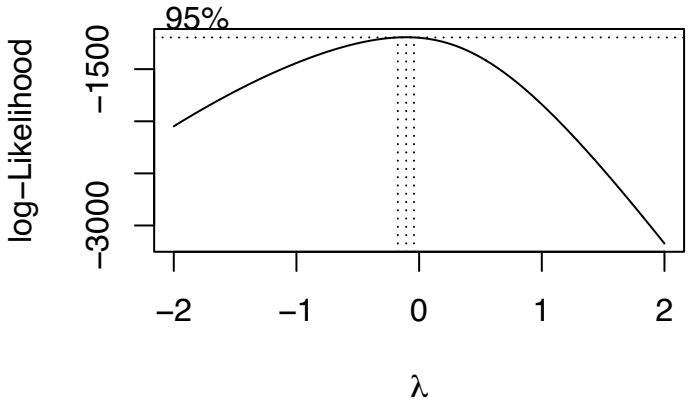
Distributions of Cleaned Data used in Final Model


pct_below_poverty
Frequency
0
50
100
150
0
1
2
3
4
per_capita_income
Frequency
0
50
100
150
200
250
5000
15000
30000
log_pop_density
Frequency
0
50
100
150
2
4
6
8
10
region
Frequency
0
50
100
150
1.0
2.0
3.0
4.0

Correlation Heatmap of Cleaned Data used in Final Model

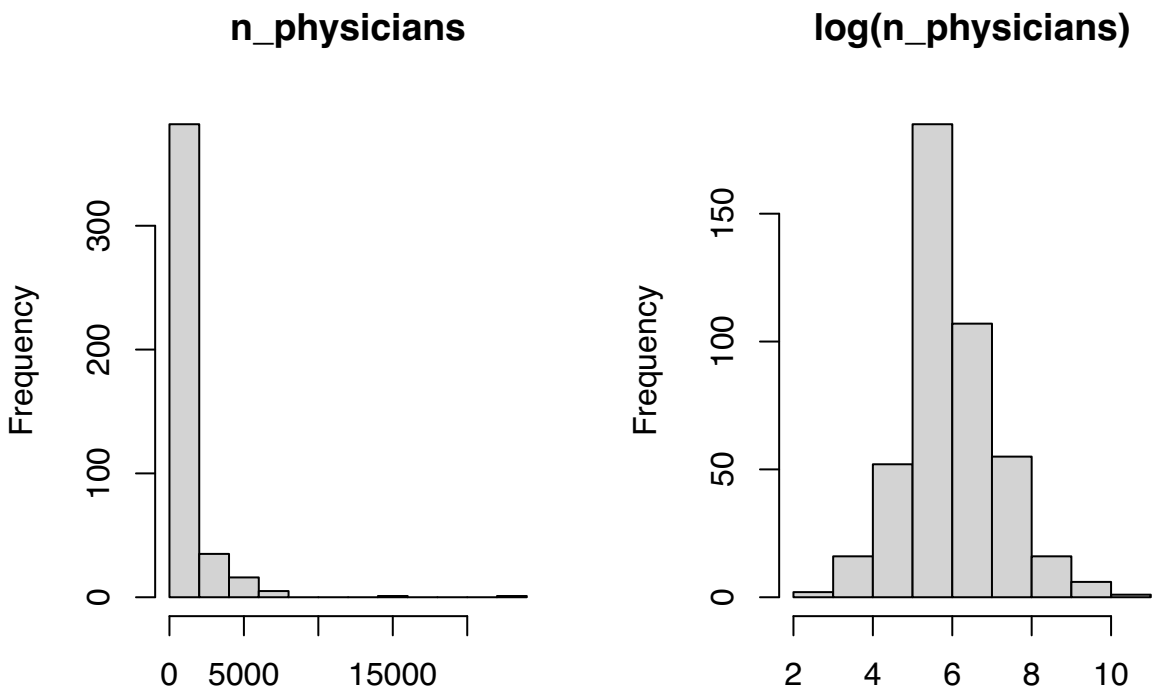


Lambda result for a Box-Cox Transformation




```
## Lambda =  -0.105263157894737
```

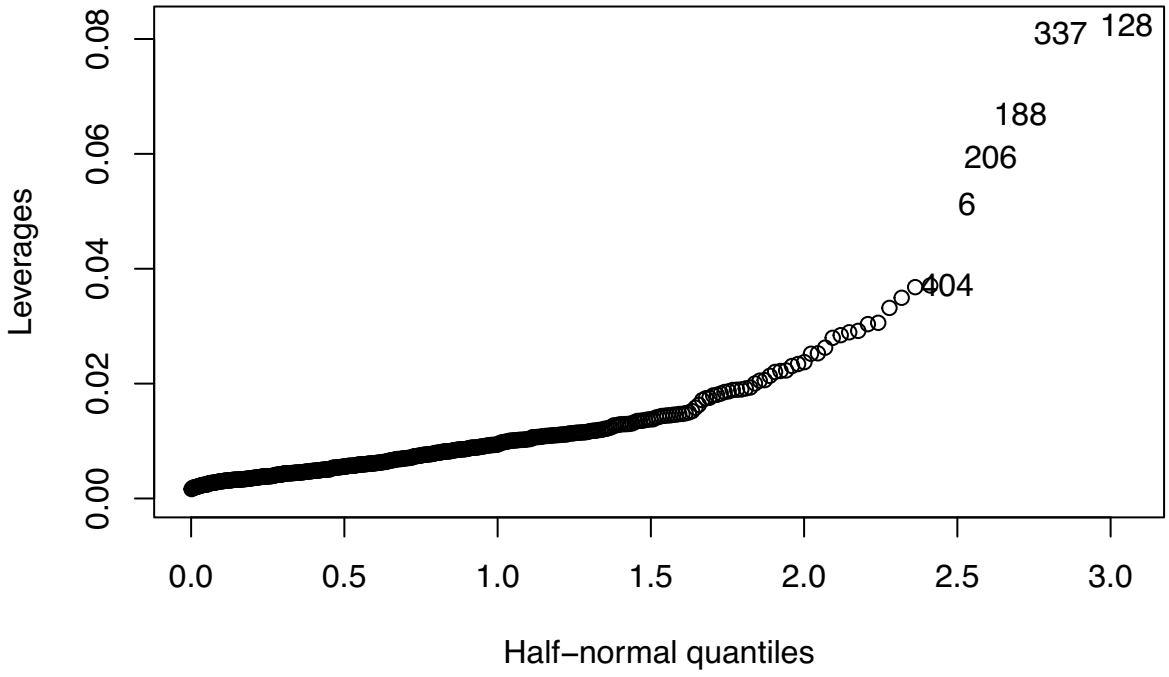
Distributions of `n_physicians` and `log(n_physicians)` (our Box-Cox transform)



Outliers of the Final Model

```
## # A tibble: 6 x 21
##      id county        state land_~1 pop_t~2 pct_1~3 pct_6~4 n_phy~5  beds  crime
##   <dbl> <chr>         <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>  <dbl>
## 1     1 Los_Angeles   CA       4060 8863164    32.1     9.7   23677 27700 688936
## 2     7 Maricopa      AZ       9204 2122101    29.2    12.5    4320  6104 177593
## 3    14 San_Bernardi~ CA      20062 1418380    30.1     8.8    2463  3349  83110
## 4    65 Pima          AZ       9187  666880    28.9    13.7    1841  2016  57051
## 5   272 Arlington_Co~ VA         26  170936    37.6    11.3     615   781  12526
## 6   396 Alexandria_C~ VA         15  111183    38.3    10.3     652   662   8537
## # ... with 11 more variables: pct_highschool <dbl>, pct_bachelor <dbl>,
## #   pct_below_poverty <dbl>, pct_unemployed <dbl>, per_capita_income <dbl>,
## #   total_personal_income <dbl>, region <dbl>, pct_beds <dbl>, pct_crime <dbl>,
## #   pop_density <dbl>, log_pop_density <dbl>, and abbreviated variable names
## #   1: land_area, 2: pop_total, 3: pct_18_24, 4: pct_65_older, 5: n_physicians
```

Leverage Points of the Final Model




```
## Bound for good leverage points: ( 3.4731101499811 ,  8.67812855923859 )

## high leverage points outside this bound are bad leverage points

## Table shows the data points considered 'bad high leverage' according to the bound

## # A tibble: 2 x 21
##      id county       state land_a~1 pop_t~2 pct_1~3 pct_6~4 n_phy~5  beds  crime
##   <dbl> <chr>        <chr>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>  <dbl>
## 1    11 Philadelphia PA         135 1585577    29.1    15.2    6641 10494 109148
## 2    19 Nassau       NY         287 1287348    25.7    14.2    6147  5200  43203
## # ... with 11 more variables: pct_highschool <dbl>, pct_bachelor <dbl>,
## #   pct_below_poverty <dbl>, pct_unemployed <dbl>, per_capita_income <dbl>,
## #   total_personal_income <dbl>, region <dbl>, pct_beds <dbl>, pct_crime <dbl>,
## #   pop_density <dbl>, log_pop_density <dbl>, and abbreviated variable names
## #   1: land_area, 2: pop_total, 3: pct_18_24, 4: pct_65_older, 5: n_physicians
```

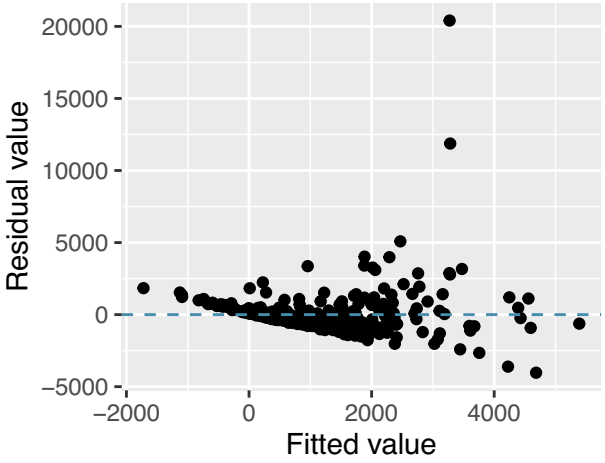
Model Summary: Full Model

```
n_physicians ~ per_capita_income + log_pop_density + pct_beds + pct_crime + region +
pct_18_24 + pct_65_older + pct_highschool + pct_bachelor + pct_below_poverty + pct_unemployed
```

```
## MODEL SUMMARY:
## Call:
## lm(formula = "n_physicians ~ per_capita_income + log_pop_density + pct_beds + pct_crime + region + p
##     data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4031.9  -538.5  -186.8   231.9 20404.9
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.709e+03  2.181e+03  -3.535 0.000452 ***
## per_capita_income  1.439e-01  4.017e-02   3.584 0.000377 ***
## log_pop_density    6.389e+02  8.676e+01   7.364 9.24e-13 ***
## pct_beds           3.271e+04  4.661e+04   0.702 0.483234
## pct_crime          9.551e+02  3.688e+03   0.259 0.795790
## region             2.787e+02  8.739e+01   3.189 0.001531 **
## pct_18_24          5.490e+01  2.727e+01   2.013 0.044729 *
## pct_65_older       2.574e+01  2.463e+01   1.045 0.296559
## pct_highschool    -1.227e+01  2.017e+01  -0.608 0.543329
## pct_bachelor      -1.411e+01  2.294e+01  -0.615 0.538721
## pct_below_poverty  6.562e+01  3.251e+01   2.019 0.044146 *
## pct_unemployed     2.322e+01  4.230e+01   0.549 0.583344
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1468 on 428 degrees of freedom
## Multiple R-squared:  0.3444, Adjusted R-squared:  0.3276
## F-statistic: 20.44 on 11 and 428 DF,  p-value: < 2.2e-16
##
## TEST FOR NORMALITY:
##  One-sample Kolmogorov-Smirnov test
##
## data:  model$residuals
## D = 0.60904, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
## TEST FOR CONSTANT VARIANCE:
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 20.613, df = 11, p-value = 0.03761
```

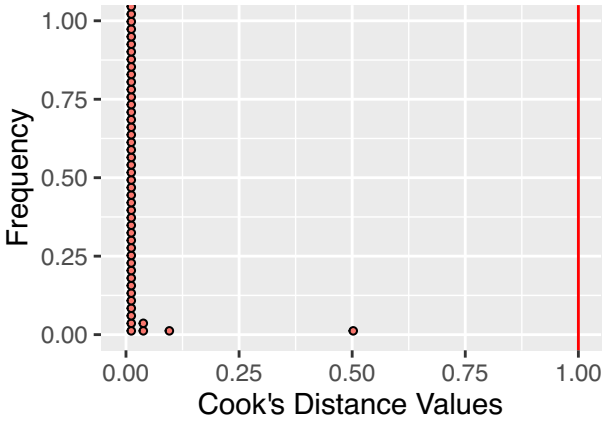
Diagnostic Plots for: n_physicians ~ per_capita_income + log_pop_density + pct_beds + pct_crime + region + pct_18_24 + pct_65_older + pct_highschool + pct_bachelor + pct_below_poverty + pct_unemployed

Distribution of Residuals

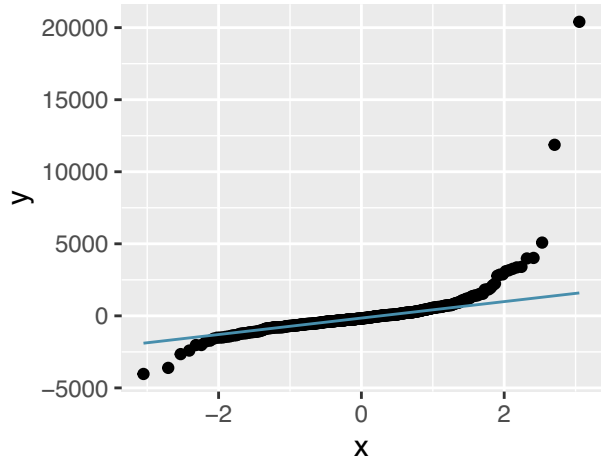


Plot of Cook's Distances

points past line are high influential

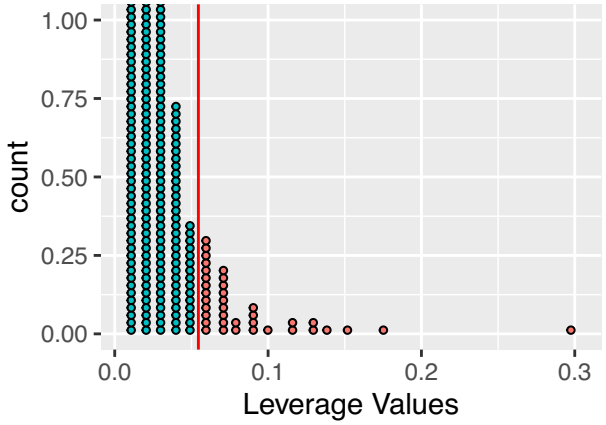


QQ−Plot of Model Residuals

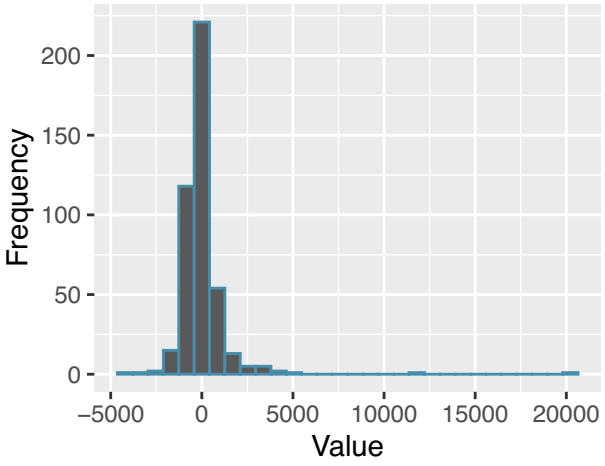


Plot of Leverage Points

points past line are high leverage

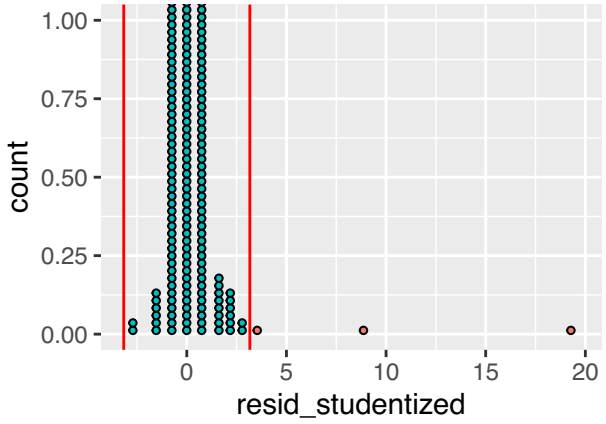


Histogram of Model Residuals



Studentized Residuals

points past line are outliers

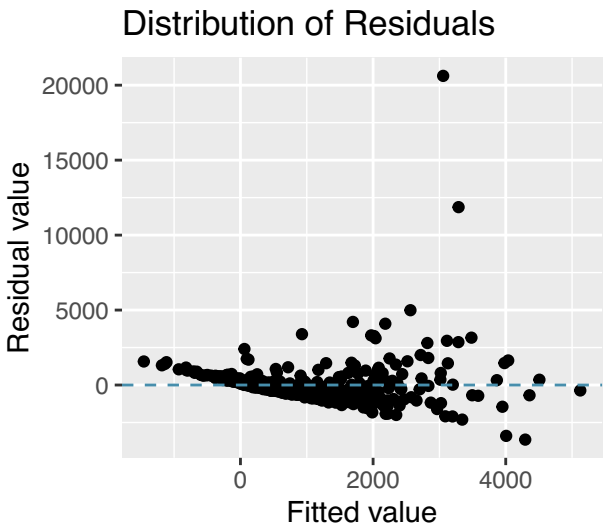


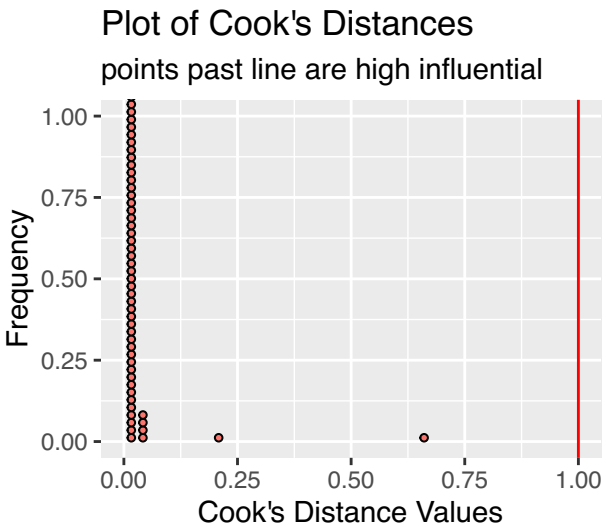
Model Summary: Reduced Model (Reduced via Permutation Test in place of Partial F-Test)

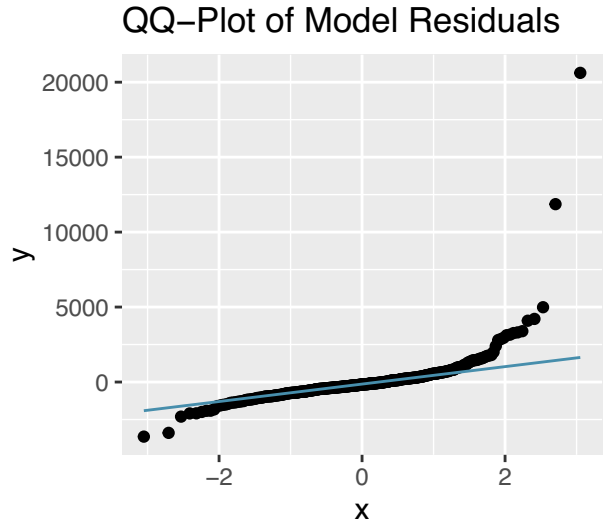
```
n_physicians ~ pct_below_poverty + per_capita_income + log_pop_density + region
```

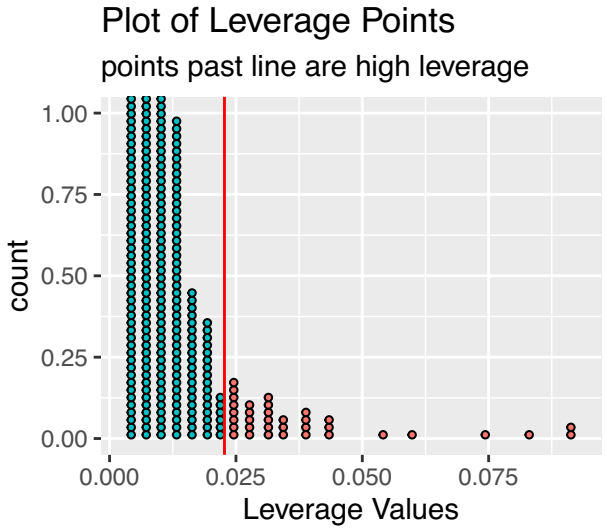
```
## MODEL SUMMARY:
## Call:
## lm(formula = "n_physicians ~ pct_below_poverty + per_capita_income + log_pop_density + region",
##     data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3642.0  -529.2  -165.5   256.0 20620.8
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6.634e+03  5.750e+02 -11.537  < 2e-16 ***
## pct_below_poverty  9.222e+01  2.011e+01   4.585 5.93e-06 ***
## per_capita_income  1.152e-01  2.594e-02   4.441 1.14e-05 ***
## log_pop_density    6.913e+02  7.479e+01   9.243  < 2e-16 ***
## region             2.278e+02  7.247e+01   3.144  0.00178 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1467 on 435 degrees of freedom
## Multiple R-squared:  0.3339, Adjusted R-squared:  0.3278
## F-statistic: 54.52 on 4 and 435 DF,  p-value: < 2.2e-16
##
## TEST FOR NORMALITY:
##  One-sample Kolmogorov-Smirnov test
##
## data:  model$residuals
## D = 0.625, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
## TEST FOR CONSTANT VARIANCE:
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 13.411, df = 4, p-value = 0.009434
```

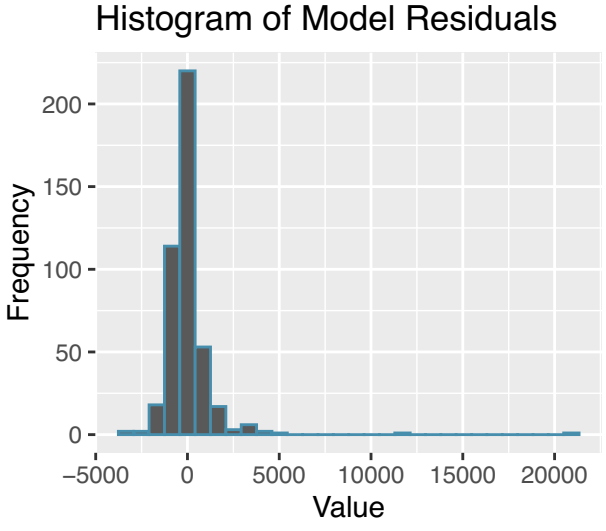
Diagnostic Plots for: n_physicians ~ pct_below_poverty + per_capita_income + log_pop_density + region

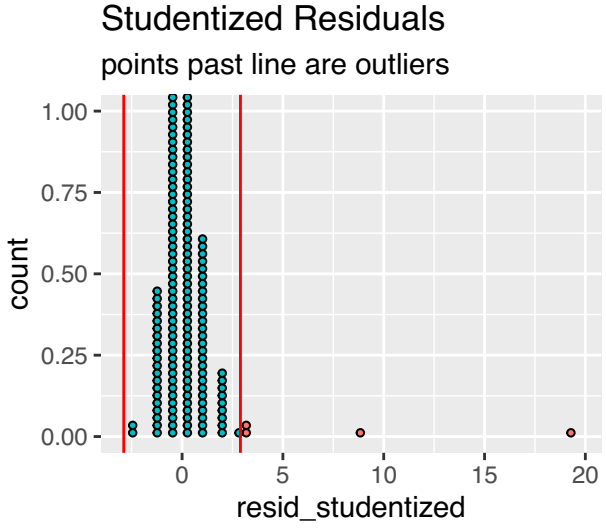










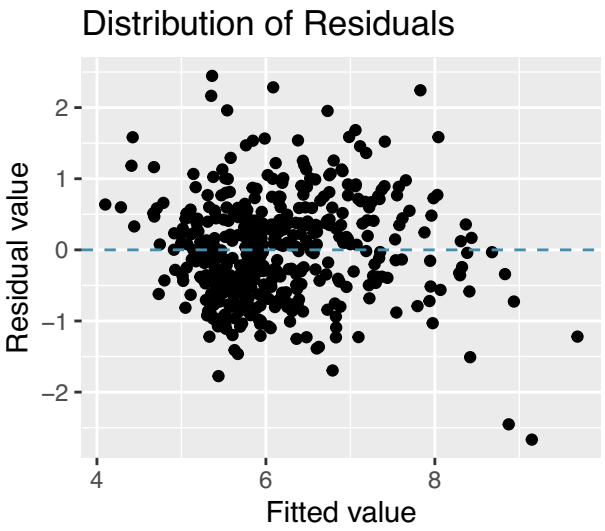


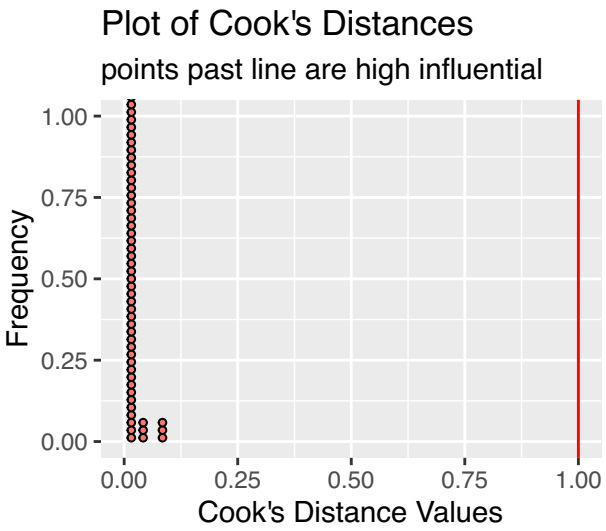
Model Summary: Pre-Final Model (Applied Boxcox Transformation ONLY)

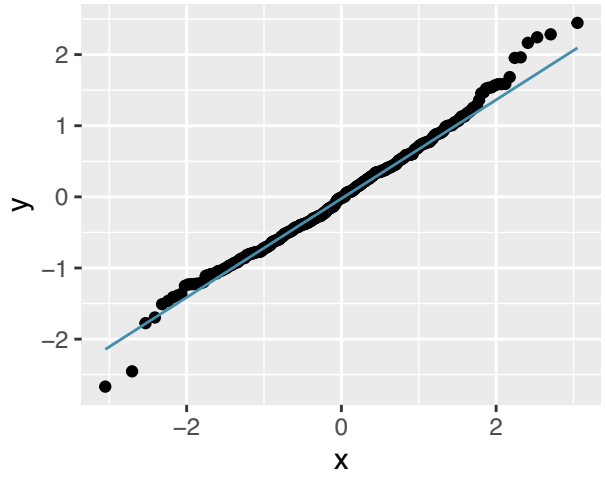
```
log(n_physicians) ~ pct_below_poverty + per_capita_income + log_pop_density + region
```

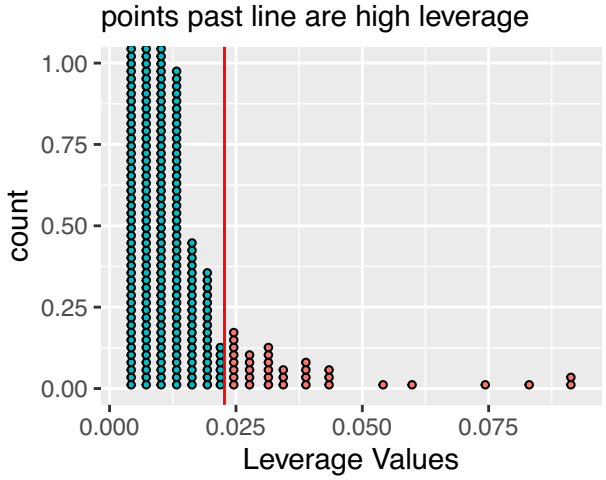
```
## MODEL SUMMARY:
## Call:
## lm(formula = "log(n_physicians) ~ pct_below_poverty + per_capita_income + log_pop_density + region",
##     data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.66781 -0.49328 -0.01732  0.44346  2.44584
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.884e-01  2.865e-01  -1.356    0.176
## pct_below_poverty  7.342e-02  1.002e-02   7.326 1.16e-12 ***
## per_capita_income  1.236e-04  1.292e-05   9.562  < 2e-16 ***
## log_pop_density    5.330e-01  3.727e-02  14.301  < 2e-16 ***
## region             1.748e-01  3.611e-02   4.841 1.80e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7312 on 435 degrees of freedom
## Multiple R-squared:  0.5953, Adjusted R-squared:  0.5916
## F-statistic:   160 on 4 and 435 DF,  p-value: < 2.2e-16
##
## TEST FOR NORMALITY:
##  One-sample Kolmogorov-Smirnov test
##
## data:  model$residuals
## D = 0.096592, p-value = 0.0005436
## alternative hypothesis: two-sided
##
## TEST FOR CONSTANT VARIANCE:
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 19.161, df = 4, p-value = 0.0007308
```

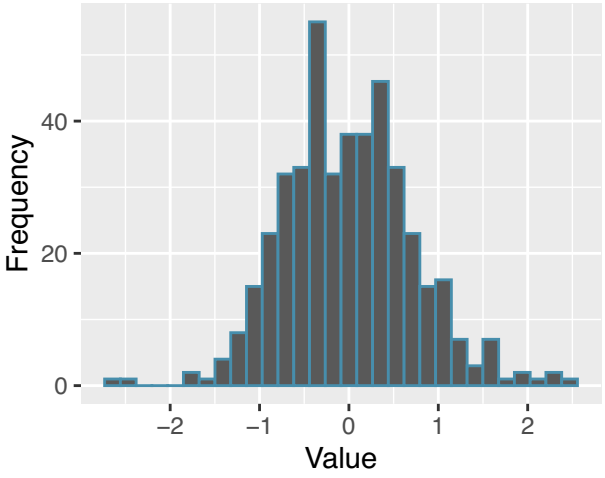
Diagnostic Plots for: log(n_physicians) ~ pct_below_poverty + per_capita_income + log_pop_density + region

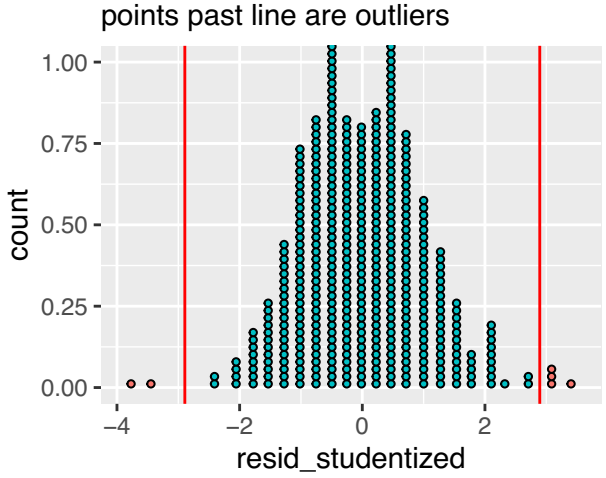












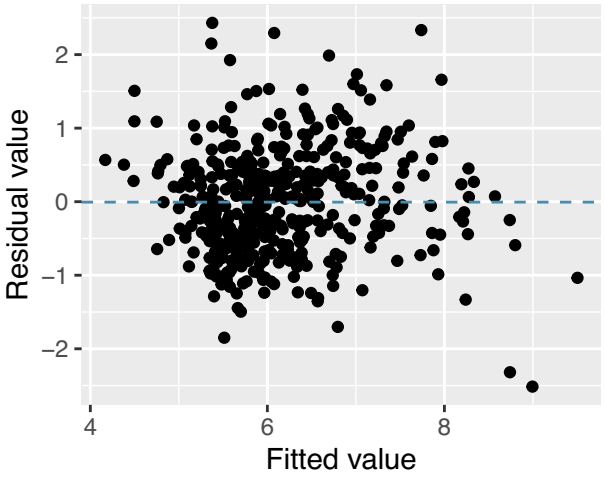
Model Summary: Final Model (Applied Boxcox Transformation AND intercept = 0)

```
log(n_physicians) ~ 0 + pct_below_poverty + per_capita_income + log_pop_density + region
```

```
## MODEL SUMMARY:
## Call:
## lm(formula = "log(n_physicians) ~ 0 + pct_below_poverty + per_capita_income + log_pop_density + regi
##     data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.51463 -0.49216 -0.01285  0.42630  2.43148
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## pct_below_poverty 0.0663618  0.0085725   7.741 6.93e-14 ***
## per_capita_income 0.0001134  0.0000105  10.791  < 2e-16 ***
## log_pop_density   0.5182723  0.0356893  14.522  < 2e-16 ***
## region            0.1570192  0.0336759   4.663 4.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7319 on 436 degrees of freedom
## Multiple R-squared:  0.9864, Adjusted R-squared:  0.9863
## F-statistic:  7931 on 4 and 436 DF,  p-value: < 2.2e-16
##
## TEST FOR NORMALITY:
##  One-sample Kolmogorov-Smirnov test
##
## data:  model$residuals
## D = 0.095922, p-value = 0.000609
## alternative hypothesis: two-sided
##
## TEST FOR CONSTANT VARIANCE:
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 4.1924, df = 3, p-value = 0.2414
```

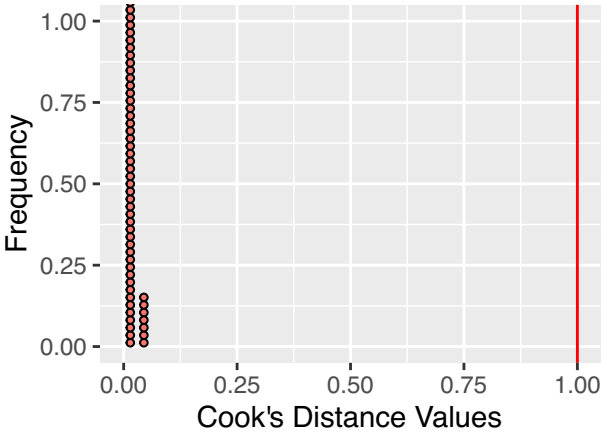
Diagnostic Plots for: log(n_physicians) ~ 0 + pct_below_poverty + per_capita_income + log_pop_density + region

Distribution of Residuals

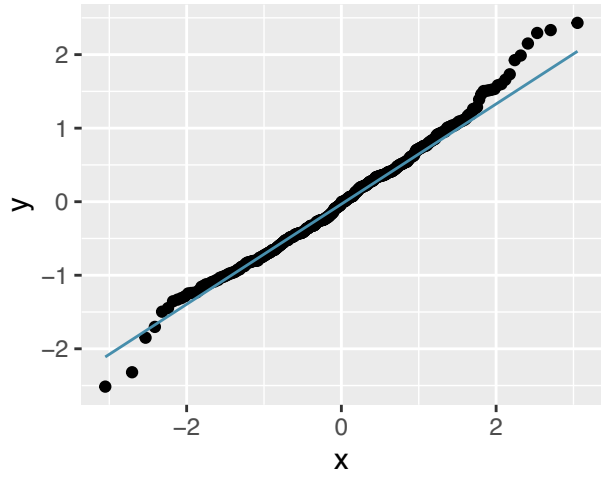


Plot of Cook's Distances

points past line are high influential

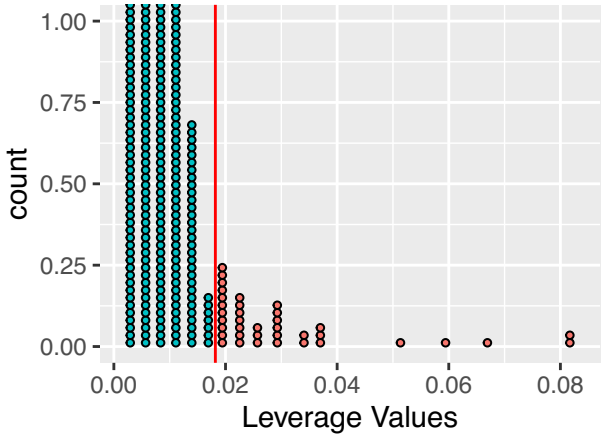


QQ−Plot of Model Residuals

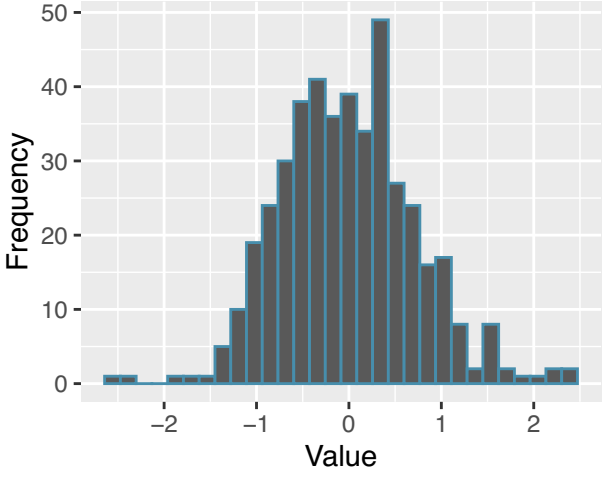


Plot of Leverage Points

points past line are high leverage



Histogram of Model Residuals



Studentized Residuals

points past line are outliers