Project Summary

Financial and Academic Factors' Impact on U.S. Undergraduate Program Completion Rates

Aragorn Wang - Brock Sowa - Eric Flores - Mikias Mengistu

The full code can be found at:

https://github.com/aragorn-w/math4387-project-report/blob/main/GroupA_Project_Report.Rmd

Introduction

Question of interest.

Describe your question of interest. What specific question are your trying to answer?

We are hoping to determine what factors positively or negatively affect undergraduate degree completion rate. Some example factors we'd like to consider are as follows: - How do student and faculty expenses affect student completion rates for four-year degree programs? - Do certain SAT section scores have a stronger correlation with students' four-year degree completion rates than other SAT sections' scores?

Importance

Explain why this topic is important (why should we care)?

We hope to answer some questions so that students or institutions focus school choice and/or policy decisions to increase the probability of completing a 4-year undergraduate degree.

Background

Provide background information to put your analysis into context.

We retrieved our data from the U.S. Department of Education's College Scorecard:

- Primary Website: https://collegescorecard.ed.gov/data/
- Primary Dataset: https://ed-public-download.app.cloud.gov/downloads/Most-Recent-Cohorts-Institution_09012022.zip
- Additional Dataset: https://ed-public-download.app.cloud.gov/downloads/Most-Recent-Cohorts-Field-of-Study_09012022.zip

The data set has 6,681 observations. Each observation is an institution of higher learning, differentiated by an Institution Name and/or Office of Postsecondary Education Identifier (OPEID) code.

Variables

Provide a table that includes the variables you will use in your analysis. The table should have 3 columns: `name`, `type`, and `description`. The `name` column is the name of the variable, `type` is the type of data (`numeric (continous)`, `numeric (discrete)`, `factor`, `date`, `time`, etc.), while the `description` column summarizes what the variable measures (make sure to include units!) ***Your response variable should be the first row of the table****.*

cleaned_name	original_name	type	description
`completion_rate_200`	`C200_4_POOLED`	`numeric` (continuous)	Completion rate for first-time, full-time bachelor's-degree-seeking students at four-year institutions (200% of expected time to completion), pooled for two year rolling averages
`institution_type`	`CONTROL`	`factor`	Control of institution with 3-levels: (1 = `public`, 2 = `private_non_profit`, 3 = `private_for_profit`)
`region`	`REGION`	`factor`	Integrated Postsecondary Education Data System (IPEDS) Region with 10-levels: (0 = `us_service_schools`, 1 = `new_england` (CT, ME, MA, NH, RI, VT), 2 = `mid_east` (DE, DC, MD, NJ, NY, PA), 3 = `great_lakes` (IL, IN, MI, OH, WI), 4 = `plains` (IA, KS, MN, MO, NE, ND, SD), 5 = `southeast` (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV), 6 = `southwest` (AZ, NM, OK, TX), 7 =

cleaned_name	original_name	type	description
			rocky_mountains (CO, ID, MT, UT, WY), 8 = far_west (AK, CA, HI, NV, OR, WA), 9 = outlying_areas (AS, FM, GU, MH, MP, PR, PW, VI))
admission_rate	ADM_RATE	numeric (continuous)	Admission rate
faculty_fulltime	PFTFAC	numeric (continuous)	Proportion of faculty that is full-time
faculty_salary_avg	AVGFACSAL	numeric (continuous)	Average faculty salary
expenditure_per_student	INEXPFTE	numeric (discrete)	Instructional expenditures per full-time equivalent student
family_income	FAMINC	numeric (continuous)	Average family income
cost_attendance_per_year	COSTT4_A	numeric (discrete)	Average cost of attendance (academic year institutions)
tuition_revenue_per_student	TUITFTE	numeric (discrete)	Net tuition revenue per full-time equivalent student
debt_median_all	DEBT_MDN	numeric (discrete)	The median original amount of the loan principal upon entering repayment

Data exploration

Numeric Summaries

Provide a numeric summary (think 5- or 6-number summary) of the response variable and at least 3 predictors.

Summaries of completion_rate_200

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.4505  0.5704  0.5735  0.6955  1.0000
```

Separated by institution_type variable

```
## inst_data$institution_type: public
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1776  0.4348  0.5324  0.5461  0.6539  0.9475
## ------------------------------------------------------------
## inst_data$institution_type: private_non_profit
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.4753  0.5987  0.5962  0.7178  1.0000
## ------------------------------------------------------------
## inst_data$institution_type: private_for_profit
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2587  0.3903  0.3788  0.4494  0.8572
```

Summaries of `faculty_salary_average`

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1085    6587    7820    8262    9577   21143
```

Separated by `institution_type` variable

```
## inst_data$institution_type: public
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3499    7415    8700    9031   10277   19286
## ------------------------------------------------------------
## inst_data$institution_type: private_non_profit
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1085    6172    7335    7903    8924   21143
## ------------------------------------------------------------
## inst_data$institution_type: private_for_profit
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2543    6066    7458    7101    8316   10239
```

Summaries of `cost_attendance_per_year`

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9565   24481   37584   39692   51383   81531
```

Separated by `institution_type` variable

```
## inst_data$institution_type: public
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9565   19794   23016   23050   25747   39595
## ------------------------------------------------------------
## inst_data$institution_type: private_non_profit
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11238   38780   47672   48705   58804   81531
## ------------------------------------------------------------
## inst_data$institution_type: private_for_profit
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24323   29954   34317   35151   37118   65272
```

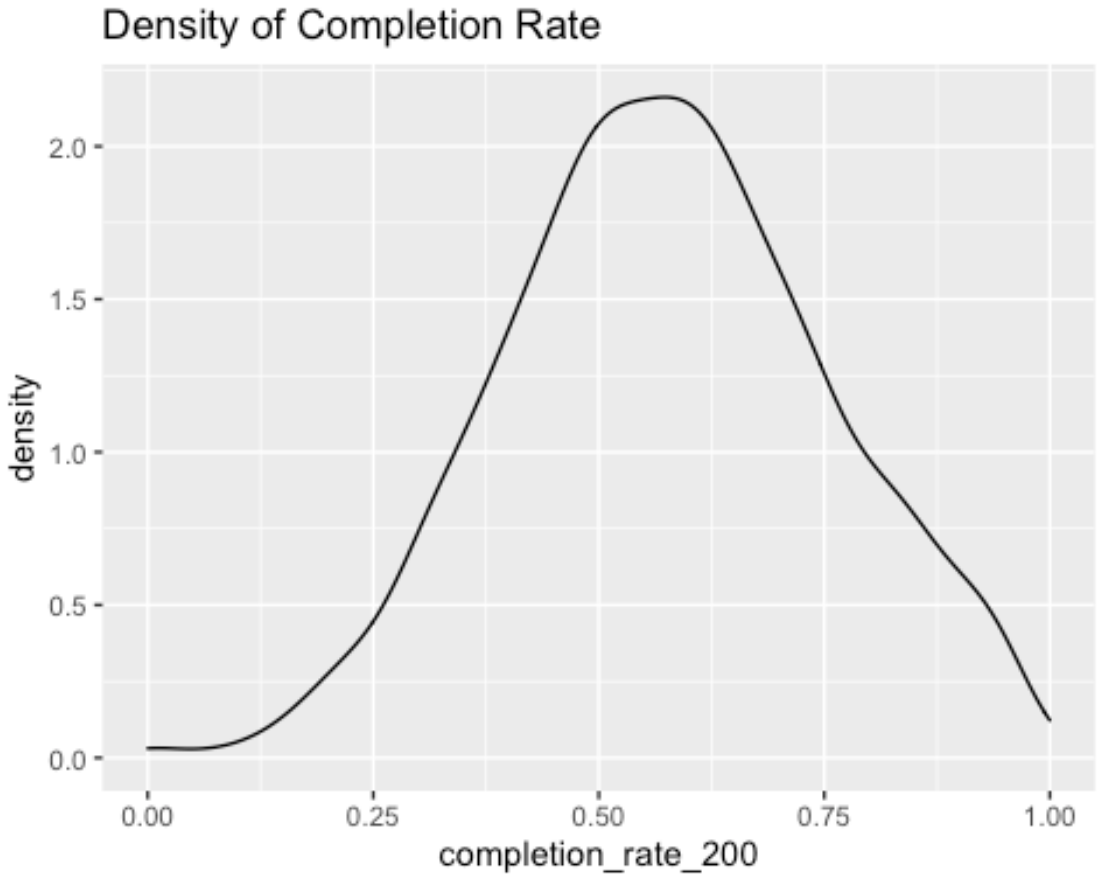
We do observe a large amount of `NA` values in our dataset, which will reduce the number of rows that our regression model will calculate with. Altogether, the data set contains 6,681 observations. The variables in our final model have 1,187 rows with no `NA` values. We still

feel this is an adequately representative sample, as our scatter plots appear to be consistent with our regression lines computed by both the parallel lines and separate lines models.

Univariate graphics

Provide a density plot of the response. Provide a univariate plot for 3 predictors. If the predictor is continuous, use a density plot. If the predictor is discrete, use a histogram. If the predictor is categorical, use a bar plot. Provide a brief interpretation of each graphic (unimodal, bimodal, skewness, unusual observations, etc.). I would focus on the 3 predictors that are most related to the response. A univariate graphic only includes a single variable.

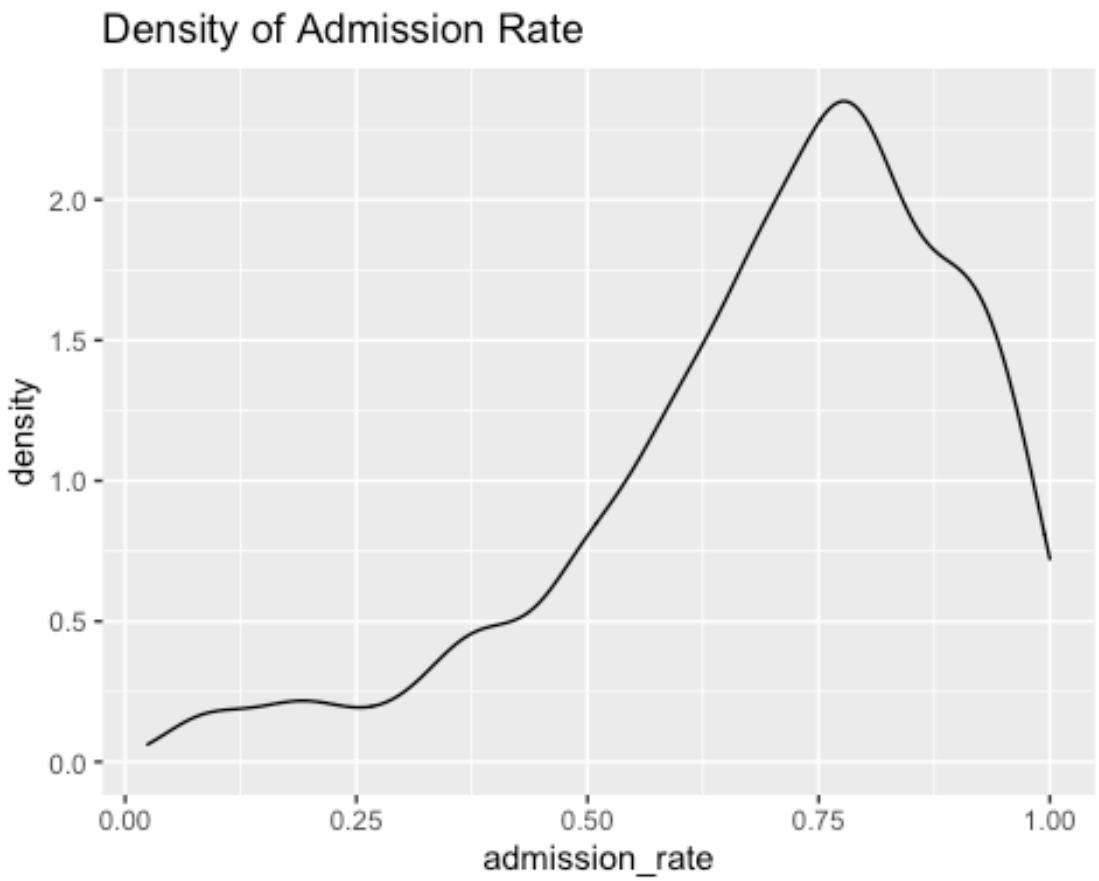
Density plot of `completion_rate_200`



Interpretation of `completion_rate_200`

The unimodal distribution of the completion rate appears to be just a little left-skewed. The main bulk of admission rates appear to reside near 0.55.

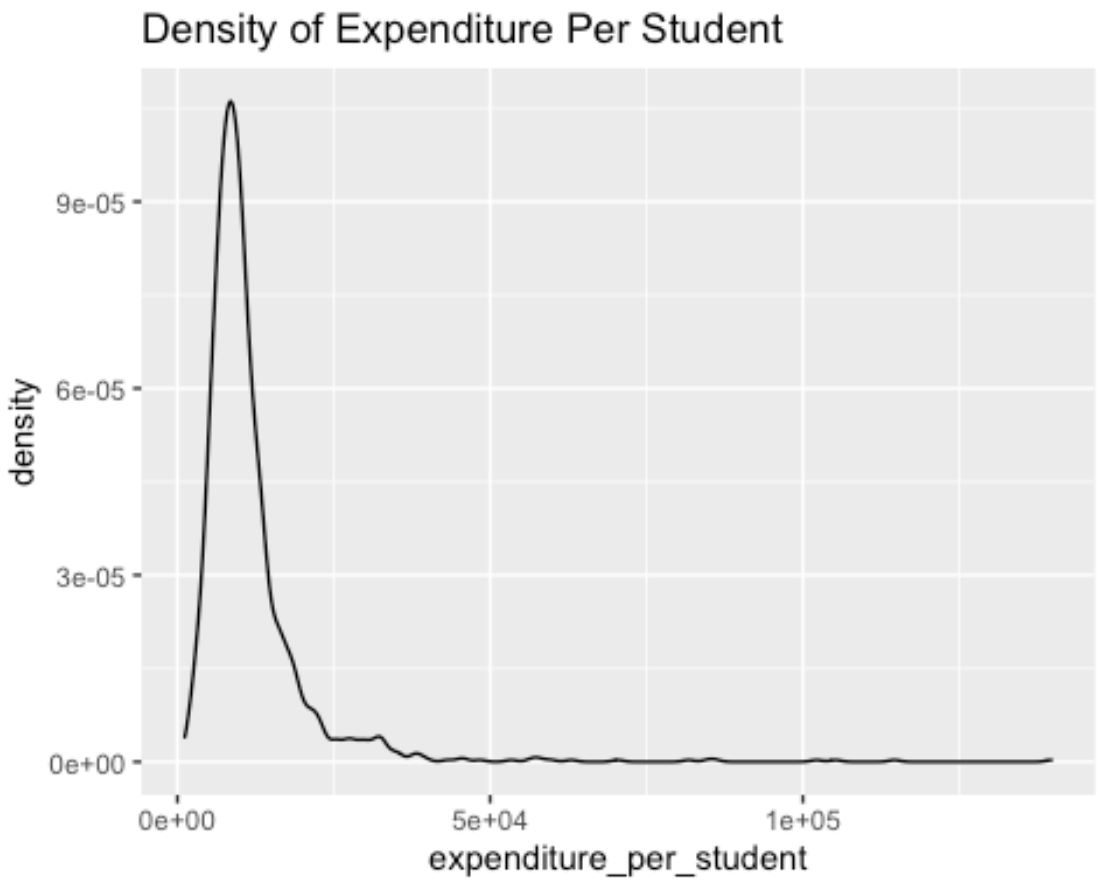
Density plot of `admission_rate`



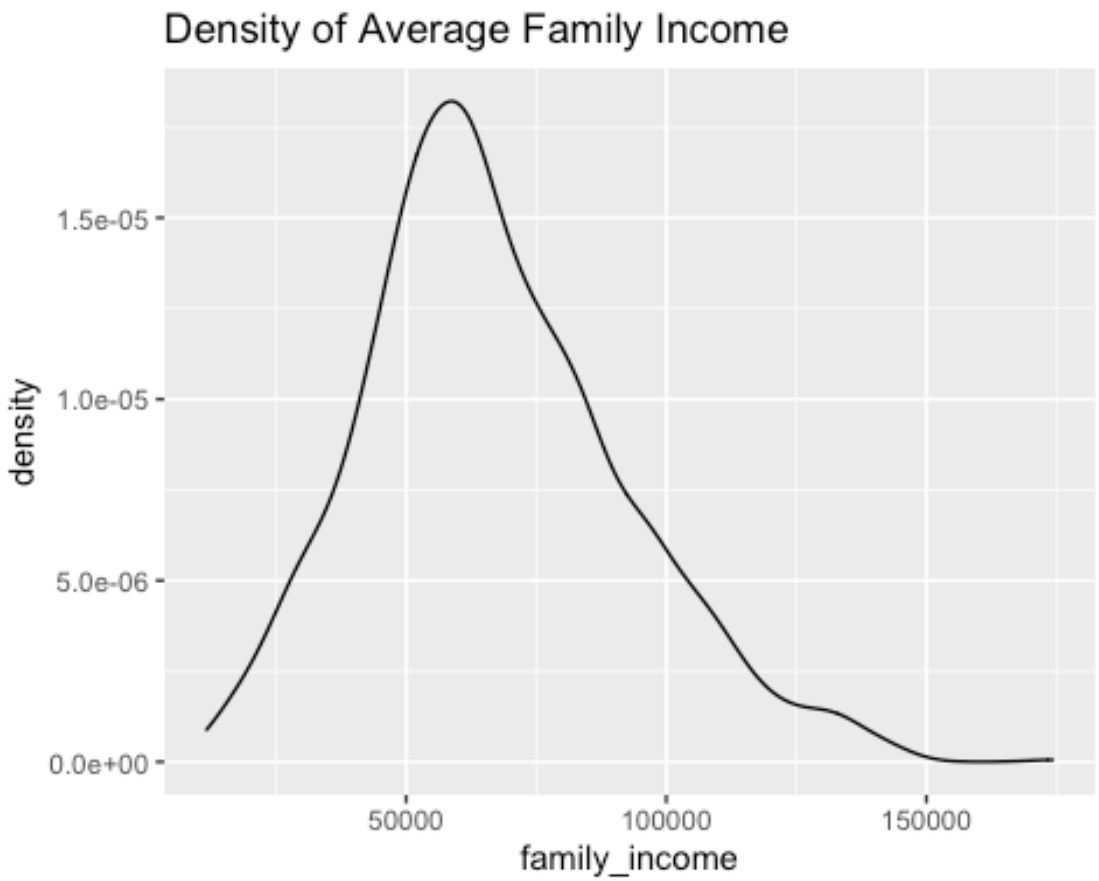
Interpretation of `admission_rate`

The unimodal distribution of the admission rate appears to be left-skewed, indicating that a squaring transformation might be useful. The main bulk of admission rates appear to reside near 0.75.

Density plot of `expenditure_per_student`



Density plot of `family_income`



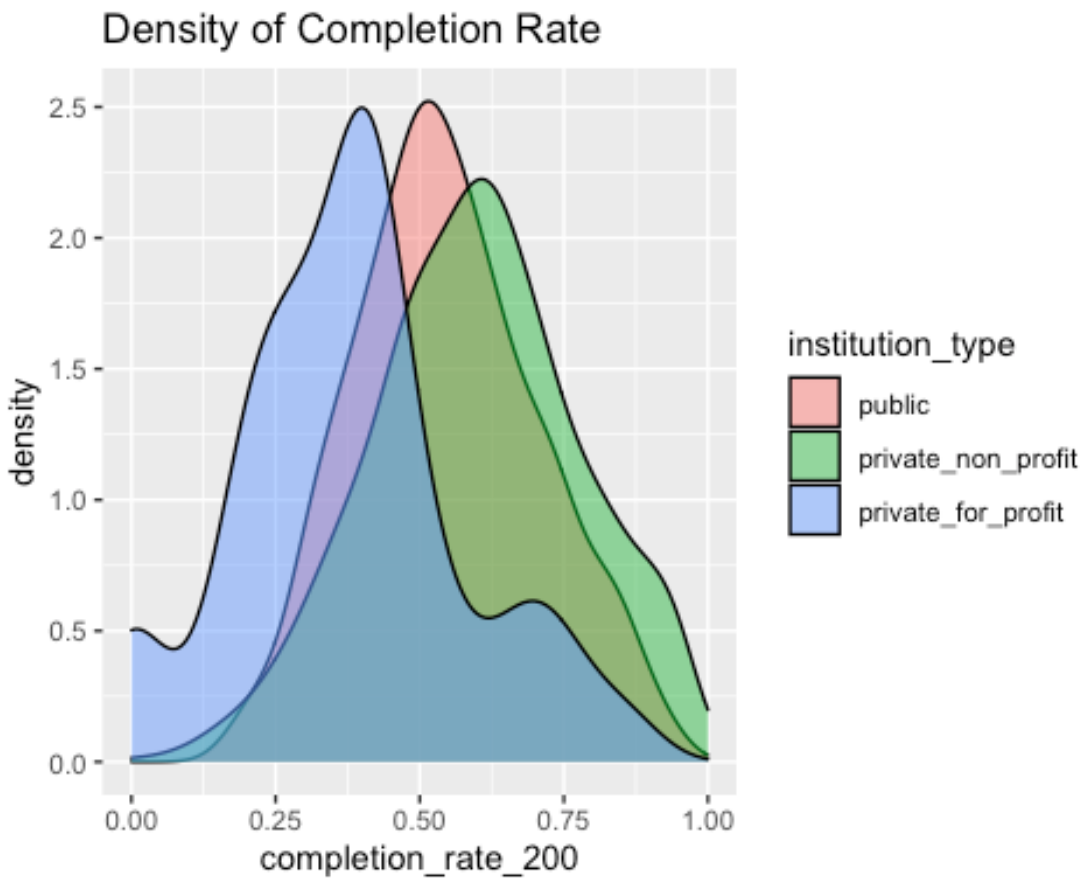
Interpretation of `family_income`

The unimodal distribution of the average family income appears to be right-skewed, indicating that a square root transformation might be useful. The bulk of the average family income appears to be near $62,000.

Multivariate graphic

Provide a bivariate plot of the response versus a predictor or a grouped scatterplot of the response variable versus a predictor with coloring based on a third graphic. Note: include the graphic that is most interesting and useful so you can support your later conclusions. Interpret the plot.

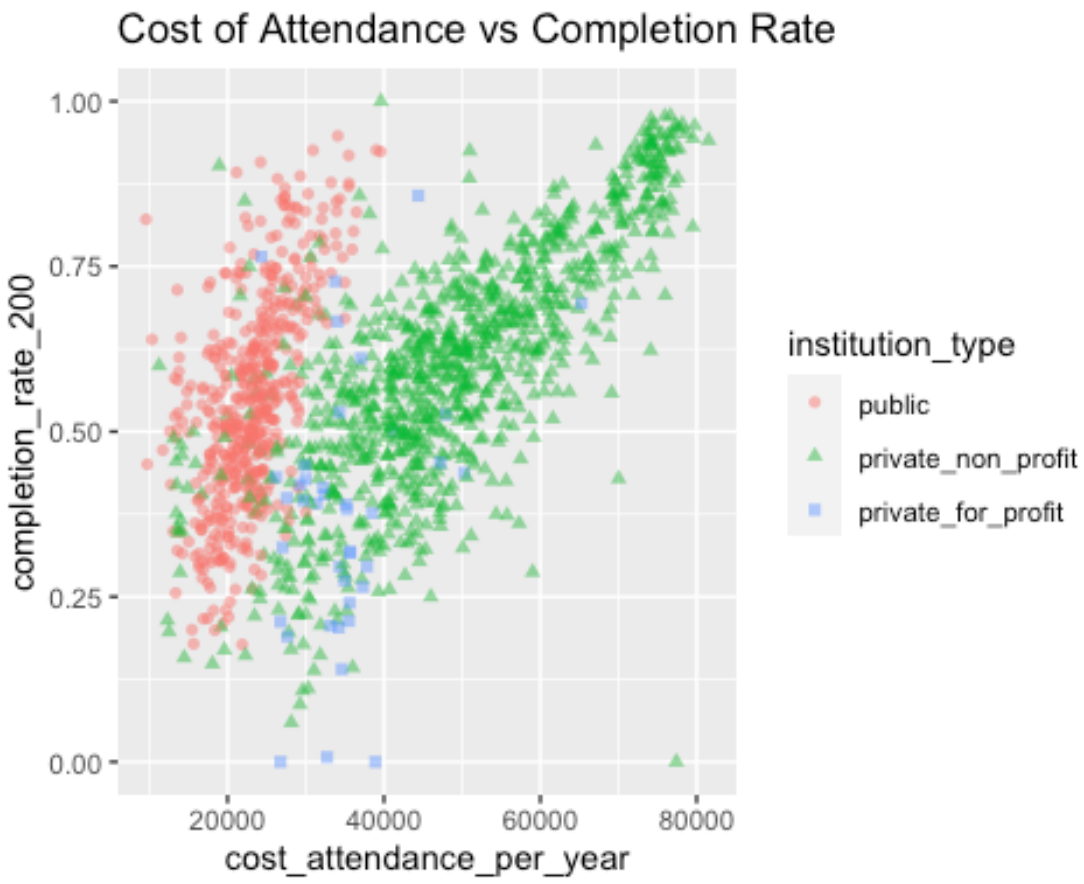
Density Plots of Completion Rate by Institution Type



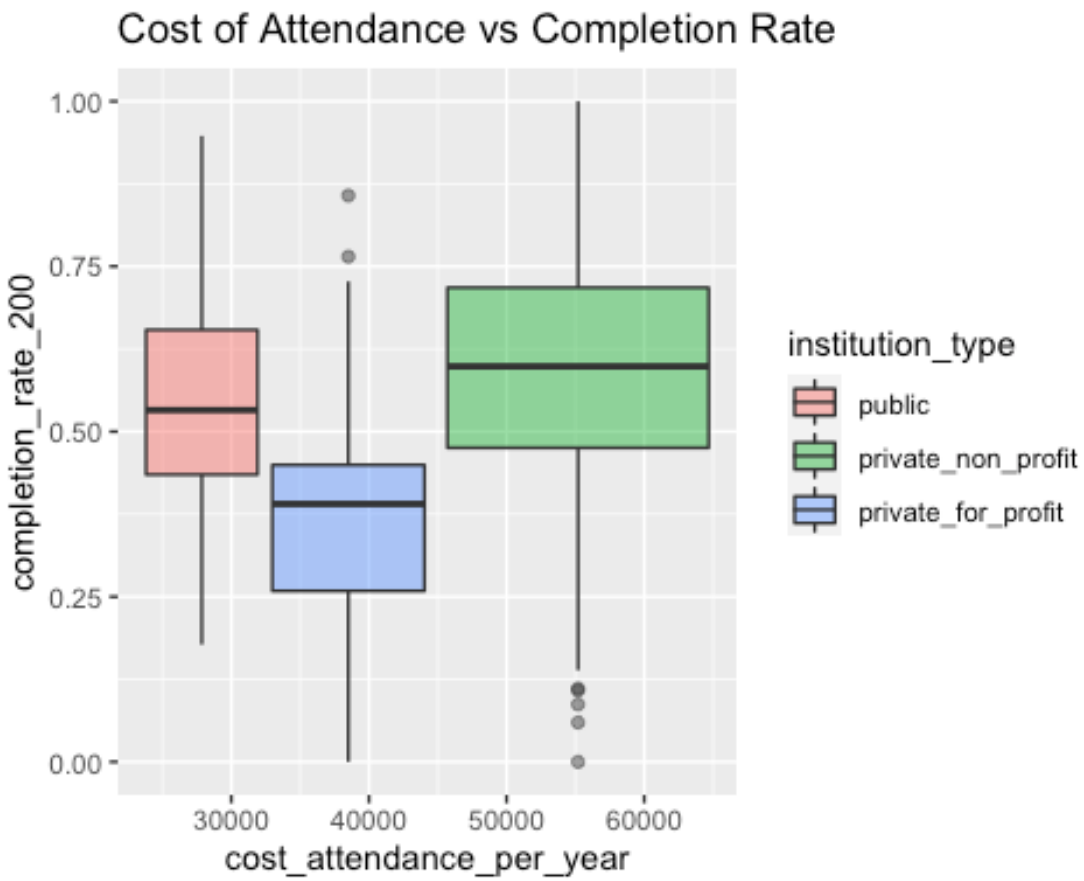
Interpretation of `completion_rate_200`

The density plot utilizes the `factor` variable `institution_type`. We observe completion rates for **private for-profit** institutions is noticeably smaller than completion rates for **public** and **private non-profit** institutions.

Grouped Scatter Plots of Yearly Cost of Attendance vs. Completion Rate by Institution Type



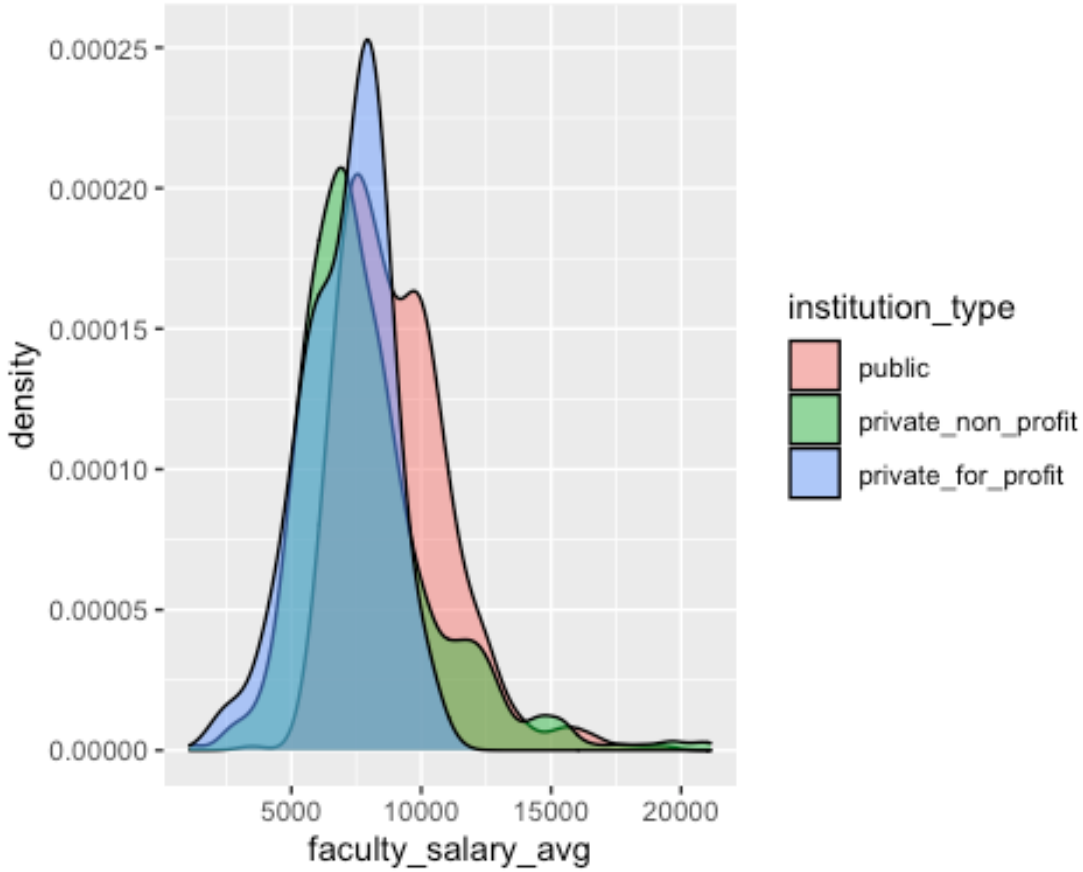
Parallel Box Plots of Yearly Cost of Attendance vs. Completion Rate by Institution Type



Interpretation of `cost_attendance_per_year` **vs** `completion_rate_200`

Provided a grouped scatterplot and a boxplot for the `cost_attendance_per_year` versus the `completion_rate_200` variables, separated by the `institution_type` variable. We note the similarities between completion rate with respect to public and private non-profit institutions. We also not that there does not appear to be a significant relationship between cost of attendance and completion rate. That is to say, there appears to be a marginal benefit, but we feel that could be attributed to selection bias for students who choose or have the resources to attend private non-profit institutions. Our assessment is that attending institutions where cost is higher is of less importance as it relates to the completion rate. This is not an assessment of perceived quality of education or prospective employment opportunities post graduation.

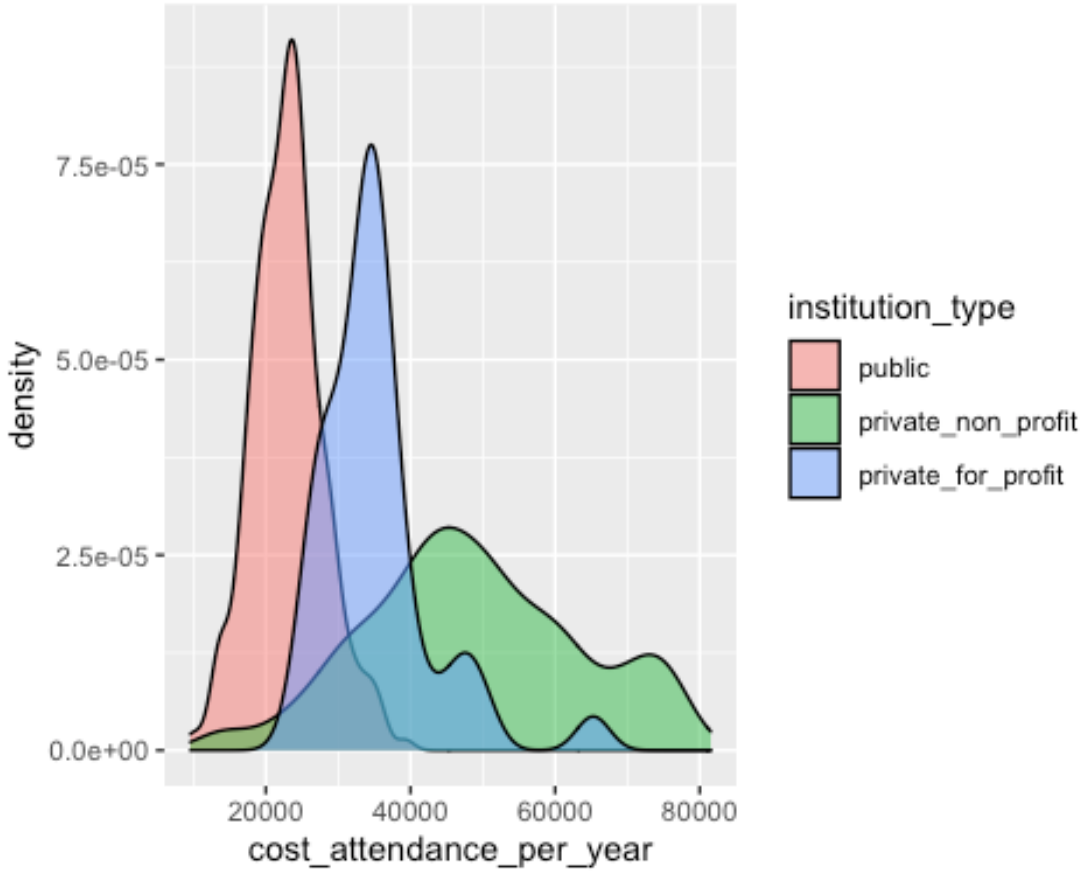
Density Plots of Average Faculty Salary by Institution Type



Interpretation of `faculty_salary_avg`

The average faculty salary for public and private non-profit institutions appear to be similar. Average faculty salary for private for-profit institutions appears to be lower.

Density Plots of Yearly Cost of Attendance by Institution Type



Interpretation of `cost_attendance_per_year`

Cost of attendance varies widely depending on the institution type. Cost of attendance at public institutions is lower than for private for-profit institutions. Cost of attendance at private non-profit institutions varies widely; the density plot is much flatter than with the other two institutions types.

Variable selection

Perform variable selection using at least two selection criteria.

Full Model Fit

```
lmod_full <- lm(completion_rate_200 ~ . + I(admission_rate^2) +
sqrt(faculty_salary_avg) + log(expenditure_per_student) +
sqrt(family_income), data = inst_data)
(summary_full <- summary(lmod_full))

##
## Call:
## lm(formula = completion_rate_200 ~ . + I(admission_rate^2) +
```

```
##     sqrt(faculty_salary_avg) + log(expenditure_per_student) +
##     sqrt(family_income), data = inst_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60106 -0.05002  0.00036  0.05665  0.50287
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -4.408e-01  1.573e-01  -2.803 0.005137
**
## institution_typeprivate_non_profit  5.999e-03  1.144e-02   0.524 0.600139
## institution_typeprivate_for_profit -2.085e-02  2.031e-02  -1.027 0.304754
## regionnew_england                  -1.576e-01  9.860e-02  -1.599 0.110087
## regionmid_east                     -1.309e-01  9.842e-02  -1.330 0.183598
## regiongreat_lakes                  -1.237e-01  9.837e-02  -1.258 0.208656
## regionplains                       -1.234e-01  9.833e-02  -1.255 0.209803
## regionsoutheast                    -1.371e-01  9.825e-02  -1.395 0.163177
## regionsouthwest                    -1.587e-01  9.856e-02  -1.610 0.107507
## regionrocky_mountains              -1.158e-01  9.934e-02  -1.166 0.243931
## regionfar_west                     -9.361e-02  9.869e-02  -0.949 0.343024
## regionoutlying_areas                7.889e-02  1.009e-01   0.782 0.434588
## admission_rate                     -2.688e-01  7.609e-02  -3.532 0.000424
***
## faculty_fulltime                    1.214e-02  1.114e-02   1.090 0.275911
## faculty_salary_avg                  4.728e-05  9.662e-06   4.893 1.10e-06
***
## expenditure_per_student            -2.298e-06  6.476e-07  -3.548 0.000400
***
## family_income                      -8.904e-07  9.096e-07  -0.979 0.327779
## cost_attendance_per_year            1.184e-06  3.963e-07   2.987 0.002866
**
## tuition_revenue_per_student        -2.331e-06  6.027e-07  -3.868 0.000115
***
## debt_median_all                     5.340e-06  7.927e-07   6.736 2.32e-11
***
## I(admission_rate^2)                 1.473e-01  5.680e-02   2.593 0.009595
**
## sqrt(faculty_salary_avg)           -4.835e-03  1.781e-03  -2.715 0.006697
**
## log(expenditure_per_student)        8.086e-02  1.123e-02   7.203 9.33e-13
***
## sqrt(family_income)                 2.070e-03  4.739e-04   4.368 1.34e-05
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09669 on 1485 degrees of freedom
## Multiple R-squared:  0.7129, Adjusted R-squared:  0.7085
## F-statistic: 160.3 on 23 and 1485 DF,  p-value: < 2.2e-16
```

Forward Selection with BIC

```
lmod_0 <- lm(completion_rate_200 ~ 1, data = inst_data)
lmod_forward <- step(lmod_0, scope = formula(lmod_full), direction =
"forward", k = log(nobs(lmod_full)))

## Start:  AIC=-5184.37
## completion_rate_200 ~ 1
##
##                                 Df Sum of Sq    RSS     AIC
## + family_income                  1   25.3321 23.030 -6296.6
## + sqrt(family_income)            1   24.8359 23.526 -6264.4
## + log(expenditure_per_student)   1   21.6247 26.738 -6071.4
## + faculty_salary_avg             1   18.9584 29.404 -5927.9
## + sqrt(faculty_salary_avg)       1   18.7502 29.612 -5917.3
## + cost_attendance_per_year       1   15.8308 32.531 -5775.4
## + expenditure_per_student        1   14.3942 33.968 -5710.2
## + tuition_revenue_per_student    1   13.6001 34.762 -5675.3
## + debt_median_all                1    8.2676 40.095 -5459.9
## + admission_rate                 1    5.7443 42.618 -5367.9
## + region                         9    6.5792 41.783 -5339.2
## + I(admission_rate^2)            1    4.3943 43.968 -5320.8
## + institution_type               2    2.3937 45.969 -5246.3
## + faculty_fulltime               1    1.5254 46.837 -5225.4
## <none>                                       48.362 -5184.4
##
## Step:  AIC=-6296.59
## completion_rate_200 ~ family_income
##
##                                 Df Sum of Sq    RSS     AIC
## + faculty_salary_avg             1    5.7260 17.304 -6720.6
## + sqrt(faculty_salary_avg)       1    5.2968 17.733 -6683.7
## + log(expenditure_per_student)   1    4.9501 18.080 -6654.4
## + expenditure_per_student        1    3.9399 19.090 -6572.4
## + admission_rate                 1    1.7866 21.244 -6411.1
## + I(admission_rate^2)            1    1.2095 21.821 -6370.7
## + region                         9    1.8889 21.141 -6359.9
## + tuition_revenue_per_student    1    0.8620 22.168 -6346.8
## + cost_attendance_per_year       1    0.7806 22.250 -6341.3
## + institution_type               2    0.2856 22.745 -6300.8
## + faculty_fulltime               1    0.1685 22.862 -6300.4
## <none>                                       23.030 -6296.6
## + debt_median_all                1    0.0764 22.954 -6294.3
## + sqrt(family_income)            1    0.0004 23.030 -6289.3
##
## Step:  AIC=-6720.63
## completion_rate_200 ~ family_income + faculty_salary_avg
##
##                                 Df Sum of Sq    RSS     AIC
## + log(expenditure_per_student)   1   1.16281 16.142 -6818.3
## + expenditure_per_student        1   0.57669 16.728 -6764.5
```

```
## + cost_attendance_per_year      1   0.55955 16.745 -6762.9
## + institution_type              2   0.61466 16.690 -6760.6
## + region                        9   1.07369 16.231 -6751.4
## + admission_rate                1   0.39124 16.913 -6747.8
## + I(admission_rate^2)           1   0.31489 16.989 -6741.0
## + debt_median_all               1   0.25533 17.049 -6735.7
## + sqrt(faculty_salary_avg)      1   0.21719 17.087 -6732.4
## + tuition_revenue_per_student   1   0.15419 17.150 -6726.8
## <none>                                      17.304 -6720.6
## + faculty_fulltime              1   0.05720 17.247 -6718.3
## + sqrt(family_income)           1   0.00418 17.300 -6713.7
##
## Step:  AIC=-6818.28
## completion_rate_200 ~ family_income + faculty_salary_avg +
log(expenditure_per_student)
##
##                                Df Sum of Sq    RSS     AIC
## + region                        9   0.96919 15.172 -6845.8
## + institution_type              2   0.31547 15.826 -6833.4
## + debt_median_all               1   0.22795 15.914 -6832.4
## + cost_attendance_per_year      1   0.22481 15.917 -6832.1
## + admission_rate                1   0.19557 15.946 -6829.4
## + I(admission_rate^2)           1   0.17771 15.964 -6827.7
## + sqrt(faculty_salary_avg)      1   0.10324 16.038 -6820.6
## <none>                                      16.142 -6818.3
## + faculty_fulltime              1   0.01960 16.122 -6812.8
## + tuition_revenue_per_student   1   0.00780 16.134 -6811.7
## + sqrt(family_income)           1   0.00124 16.140 -6811.1
## + expenditure_per_student       1   0.00073 16.141 -6811.0
##
## Step:  AIC=-6845.85
## completion_rate_200 ~ family_income + faculty_salary_avg +
log(expenditure_per_student) +
##     region
##
##                                Df Sum of Sq    RSS     AIC
## + debt_median_all               1   0.38928 14.783 -6877.7
## + cost_attendance_per_year      1   0.23731 14.935 -6862.3
## + institution_type              2   0.24485 14.927 -6855.8
## + sqrt(family_income)           1   0.16846 15.004 -6855.4
## + admission_rate                1   0.14795 15.024 -6853.3
## + I(admission_rate^2)           1   0.13252 15.040 -6851.8
## <none>                                      15.172 -6845.8
## + sqrt(faculty_salary_avg)      1   0.03687 15.135 -6842.2
## + faculty_fulltime              1   0.01334 15.159 -6839.9
## + expenditure_per_student       1   0.00875 15.164 -6839.4
## + tuition_revenue_per_student   1   0.00380 15.168 -6838.9
##
## Step:  AIC=-6877.75
## completion_rate_200 ~ family_income + faculty_salary_avg +
```

```
log(expenditure_per_student) +
##     region + debt_median_all
##
##                                Df Sum of Sq    RSS     AIC
## + admission_rate                1  0.237094 14.546 -6894.8
## + I(admission_rate^2)           1  0.189775 14.593 -6889.9
## + sqrt(family_income)           1  0.147690 14.635 -6885.6
## + cost_attendance_per_year      1  0.144825 14.638 -6885.3
## + sqrt(faculty_salary_avg)      1  0.107942 14.675 -6881.5
## + institution_type              2  0.176244 14.607 -6881.2
## <none>                                      14.783 -6877.7
## + faculty_fulltime              1  0.019452 14.764 -6872.4
## + tuition_revenue_per_student   1  0.001384 14.782 -6870.6
## + expenditure_per_student       1  0.000451 14.783 -6870.5
##
## Step:  AIC=-6894.83
## completion_rate_200 ~ family_income + faculty_salary_avg +
log(expenditure_per_student) +
##     region + debt_median_all + admission_rate
##
##                                Df Sum of Sq    RSS     AIC
## + sqrt(family_income)           1  0.217080 14.329 -6910.2
## <none>                                      14.546 -6894.8
## + institution_type              2  0.139394 14.406 -6894.7
## + cost_attendance_per_year      1  0.055312 14.491 -6893.3
## + sqrt(faculty_salary_avg)      1  0.043736 14.502 -6892.1
## + I(admission_rate^2)           1  0.035506 14.510 -6891.2
## + faculty_fulltime              1  0.031621 14.514 -6890.8
## + tuition_revenue_per_student   1  0.030812 14.515 -6890.7
## + expenditure_per_student       1  0.016922 14.529 -6889.3
##
## Step:  AIC=-6910.2
## completion_rate_200 ~ family_income + faculty_salary_avg +
log(expenditure_per_student) +
##     region + debt_median_all + admission_rate + sqrt(family_income)
##
##                                Df Sum of Sq    RSS     AIC
## <none>                                      14.329 -6910.2
## + cost_attendance_per_year      1  0.065976 14.263 -6909.8
## + institution_type              2  0.122081 14.207 -6908.5
## + sqrt(faculty_salary_avg)      1  0.051529 14.277 -6908.3
## + I(admission_rate^2)           1  0.043082 14.286 -6907.4
## + faculty_fulltime              1  0.027896 14.301 -6905.8
## + tuition_revenue_per_student   1  0.016090 14.313 -6904.6
## + expenditure_per_student       1  0.011947 14.317 -6904.1

(summary_forward <- summary(lmod_forward))

##
## Call:
```

```
## lm(formula = completion_rate_200 ~ family_income + faculty_salary_avg +
##     log(expenditure_per_student) + region + debt_median_all +
##     admission_rate + sqrt(family_income), data = inst_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64575 -0.05111  0.00152  0.05841  0.49957
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -5.290e-01  1.243e-01  -4.255 2.22e-05 ***
## family_income                -1.052e-06  9.080e-07  -1.159   0.2466
## faculty_salary_avg            1.929e-05  1.406e-06  13.719  < 2e-16 ***
## log(expenditure_per_student)  5.987e-02  6.971e-03   8.588  < 2e-16 ***
## regionnew_england            -1.654e-01  9.892e-02  -1.672   0.0947 .
## regionmid_east               -1.429e-01  9.880e-02  -1.446   0.1484
## regiongreat_lakes            -1.374e-01  9.881e-02  -1.390   0.1647
## regionplains                 -1.361e-01  9.881e-02  -1.377   0.1687
## regionsoutheast              -1.474e-01  9.867e-02  -1.494   0.1353
## regionsouthwest              -1.728e-01  9.900e-02  -1.746   0.0811 .
## regionrocky_mountains        -1.348e-01  9.985e-02  -1.350   0.1772
## regionfar_west               -1.052e-01  9.891e-02  -1.063   0.2879
## regionoutlying_areas          7.781e-02  1.014e-01   0.768   0.4428
## debt_median_all               5.182e-06  7.431e-07   6.973 4.64e-12 ***
## admission_rate               -8.122e-02  1.437e-02  -5.651 1.91e-08 ***
## sqrt(family_income)           2.258e-03  4.748e-04   4.756 2.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09797 on 1493 degrees of freedom
## Multiple R-squared:  0.7037, Adjusted R-squared:  0.7007
## F-statistic: 236.4 on 15 and 1493 DF,  p-value: < 2.2e-16
```

```
# Fixing the model hierarchy of the forward selection model
lmod_forward <- lm(completion_rate_200 ~ region + admission_rate +
faculty_salary_avg + family_income + sqrt(family_income) +
log(expenditure_per_student) + debt_median_all, data = inst_data)
(summary_forward <- summary(lmod_forward))
```

```
##
## Call:
## lm(formula = completion_rate_200 ~ region + admission_rate +
##     faculty_salary_avg + family_income + sqrt(family_income) +
##     log(expenditure_per_student) + debt_median_all, data = inst_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64575 -0.05111  0.00152  0.05841  0.49957
##
## Coefficients:
```

```
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -5.290e-01  1.243e-01  -4.255 2.22e-05 ***
## regionnew_england            -1.654e-01  9.892e-02  -1.672   0.0947 .
## regionmid_east               -1.429e-01  9.880e-02  -1.446   0.1484
## regiongreat_lakes            -1.374e-01  9.881e-02  -1.390   0.1647
## regionplains                 -1.361e-01  9.881e-02  -1.377   0.1687
## regionsoutheast              -1.474e-01  9.867e-02  -1.494   0.1353
## regionsouthwest              -1.728e-01  9.900e-02  -1.746   0.0811 .
## regionrocky_mountains        -1.348e-01  9.985e-02  -1.350   0.1772
## regionfar_west               -1.052e-01  9.891e-02  -1.063   0.2879
## regionoutlying_areas          7.781e-02  1.014e-01   0.768   0.4428
## admission_rate               -8.122e-02  1.437e-02  -5.651 1.91e-08 ***
## faculty_salary_avg            1.929e-05  1.406e-06  13.719  < 2e-16 ***
## family_income                -1.052e-06  9.080e-07  -1.159   0.2466
## sqrt(family_income)           2.258e-03  4.748e-04   4.756 2.17e-06 ***
## log(expenditure_per_student)  5.987e-02  6.971e-03   8.588  < 2e-16 ***
## debt_median_all               5.182e-06  7.431e-07   6.973 4.64e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09797 on 1493 degrees of freedom
## Multiple R-squared:  0.7037, Adjusted R-squared:  0.7007
## F-statistic: 236.4 on 15 and 1493 DF,  p-value: < 2.2e-16

length(coef(lmod_forward))

## [1] 16
```

Backward Selection with BIC

```
drop1(lmod_full, test = "F", k = log(nobs(lmod_full)))

## Single term deletions
##
## Model:
## completion_rate_200 ~ institution_type + region + admission_rate +
##     faculty_fulltime + faculty_salary_avg + expenditure_per_student +
##     family_income + cost_attendance_per_year + tuition_revenue_per_student
+
##     debt_median_all + I(admission_rate^2) + sqrt(faculty_salary_avg) +
##     log(expenditure_per_student) + sqrt(family_income)
##                              Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                                    13.884 -6899.3
## institution_type              2   0.02407 13.908 -6911.3  1.2874 0.2763087
## region                        9   1.22113 15.105 -6837.9 14.5125 < 2.2e-16
***
## admission_rate                1   0.11666 14.000 -6894.0 12.4778 0.0004244
***
## faculty_fulltime              1   0.01111 13.895 -6905.4  1.1880 0.2759107
## faculty_salary_avg            1   0.22383 14.107 -6882.4 23.9412 1.101e-06
***
```

```
## expenditure_per_student       1   0.11769 14.001 -6893.8 12.5881 0.0004003
***
## family_income                 1   0.00896 13.893 -6905.6  0.9583 0.3277790
## cost_attendance_per_year      1   0.08340 13.967 -6897.5  8.9204 0.0028662
**
## tuition_revenue_per_student   1   0.13985 14.024 -6891.5 14.9587 0.0001146
***
## debt_median_all               1   0.42424 14.308 -6861.2 45.3762 2.318e-11
***
## I(admission_rate^2)           1   0.06288 13.947 -6899.8  6.7260 0.0095952
**
## sqrt(faculty_salary_avg)      1   0.06893 13.953 -6899.1  7.3732 0.0066971
**
## log(expenditure_per_student)  1   0.48508 14.369 -6854.8 51.8838 9.327e-13
***
## sqrt(family_income)           1   0.17836 14.062 -6887.3 19.0775 1.342e-05
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmod_backward <- update(lmod_full, formula = . ~ . - family_income)
drop1(lmod_backward, test = "F", k = log(nobs(lmod_full)))
```

```
## Single term deletions
##
## Model:
## completion_rate_200 ~ institution_type + region + admission_rate +
##     faculty_fulltime + faculty_salary_avg + expenditure_per_student +
##     cost_attendance_per_year + tuition_revenue_per_student +
##     debt_median_all + I(admission_rate^2) + sqrt(faculty_salary_avg) +
##     log(expenditure_per_student) + sqrt(family_income)
##                              Df Sum of Sq    RSS     AIC  F value
Pr(>F)
## <none>                                    13.893 -6905.6
## institution_type              2    0.0263 13.919 -6917.4   1.4073
0.2451193
## region                        9    1.3511 15.244 -6831.4  16.0580 < 2.2e-
16 ***
## admission_rate                1    0.1156 14.008 -6900.4  12.3624
0.0004512 ***
## faculty_fulltime              1    0.0104 13.903 -6911.8   1.1134
0.2915183
## faculty_salary_avg            1    0.2224 14.115 -6889.0  23.7892 1.190e-
06 ***
## expenditure_per_student       1    0.1178 14.011 -6900.2  12.6027
0.0003972 ***
## cost_attendance_per_year      1    0.0817 13.974 -6904.1   8.7414
0.0031597 **
## tuition_revenue_per_student   1    0.1468 14.040 -6897.1  15.7042 7.757e-
05 ***
```

```
## debt_median_all                1    0.4230 14.316 -6867.7  45.2408 2.478e-
11 ***
## I(admission_rate^2)            1    0.0631 13.956 -6906.1   6.7444
0.0094969 **
## sqrt(faculty_salary_avg)       1    0.0689 13.961 -6905.5   7.3659
0.0067242 **
## log(expenditure_per_student)   1    0.4968 14.389 -6859.9  53.1434 5.018e-
13 ***
## sqrt(family_income)            1    3.1825 17.075 -6601.7 340.4127 < 2.2e-
16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmod_backward <- update(lmod_backward, formula = . ~ . - faculty_fulltime)
drop1(lmod_backward, test = "F", k = log(nobs(lmod_full)))
```

```
## Single term deletions
##
## Model:
## completion_rate_200 ~ institution_type + region + admission_rate +
##     faculty_salary_avg + expenditure_per_student +
cost_attendance_per_year +
##     tuition_revenue_per_student + debt_median_all + I(admission_rate^2) +
##     sqrt(faculty_salary_avg) + log(expenditure_per_student) +
##     sqrt(family_income)
##                               Df Sum of Sq    RSS     AIC  F value
Pr(>F)
## <none>                                     13.903 -6911.8
## institution_type               2    0.0320 13.935 -6923.0   1.7120
0.1808557
## region                         9    1.3736 15.277 -6835.5  16.3242 < 2.2e-
16 ***
## admission_rate                 1    0.1170 14.020 -6906.5  12.5142
0.0004163 ***
## faculty_salary_avg             1    0.2265 14.130 -6894.7  24.2247 9.527e-
07 ***
## expenditure_per_student        1    0.1198 14.023 -6906.2  12.8126
0.0003554 ***
## cost_attendance_per_year       1    0.0843 13.987 -6910.0   9.0166
0.0027200 **
## tuition_revenue_per_student    1    0.1523 14.055 -6902.7  16.2857 5.724e-
05 ***
## debt_median_all                1    0.4230 14.326 -6873.9  45.2371 2.482e-
11 ***
## I(admission_rate^2)            1    0.0641 13.967 -6912.2   6.8546
0.0089309 **
## sqrt(faculty_salary_avg)       1    0.0708 13.974 -6911.5   7.5696
0.0060083 **
## log(expenditure_per_student)   1    0.5113 14.414 -6864.6  54.6831 2.355e-
13 ***
```

```
## sqrt(family_income)           1    3.2765 17.180 -6599.8 350.4386 < 2.2e-
16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmod_backward <- update(lmod_backward, formula = . ~ . - institution_type)
drop1(lmod_backward, test = "F", k = log(nobs(lmod_full)))
```

```
## Single term deletions
##
## Model:
## completion_rate_200 ~ region + admission_rate + faculty_salary_avg +
##     expenditure_per_student + cost_attendance_per_year +
tuition_revenue_per_student +
##     debt_median_all + I(admission_rate^2) + sqrt(faculty_salary_avg) +
##     log(expenditure_per_student) + sqrt(family_income)
##                               Df Sum of Sq    RSS     AIC  F value
Pr(>F)
## <none>                                     13.935 -6923.0
## region                         9    1.4183 15.353 -6842.6  16.8389 < 2.2e-
16 ***
## admission_rate                 1    0.1140 14.049 -6918.0  12.1816
0.0004967 ***
## faculty_salary_avg             1    0.2496 14.185 -6903.5  26.6741 2.734e-
07 ***
## expenditure_per_student        1    0.1362 14.071 -6915.6  14.5524
0.0001419 ***
## cost_attendance_per_year       1    0.2099 14.145 -6907.7  22.4304 2.387e-
06 ***
## tuition_revenue_per_student    1    0.1984 14.133 -6909.0  21.2024 4.484e-
06 ***
## debt_median_all                1    0.4244 14.360 -6885.0  45.3450 2.352e-
11 ***
## I(admission_rate^2)            1    0.0636 13.999 -6923.4   6.7905
0.0092557 **
## sqrt(faculty_salary_avg)       1    0.0874 14.023 -6920.8   9.3433
0.0022779 **
## log(expenditure_per_student)   1    0.6480 14.583 -6861.7  69.2367 < 2.2e-
16 ***
## sqrt(family_income)            1    3.5905 17.526 -6584.3 383.6524 < 2.2e-
16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Model hierarchy for the backward selection model is already fine (all
levels of region are included)
(summary_backward <- summary(lmod_backward))
```

```
##
## Call:
## lm(formula = completion_rate_200 ~ region + admission_rate +
```

```
##     faculty_salary_avg + expenditure_per_student +
cost_attendance_per_year +
##     tuition_revenue_per_student + debt_median_all + I(admission_rate^2) +
##     sqrt(faculty_salary_avg) + log(expenditure_per_student) +
##     sqrt(family_income), data = inst_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59630 -0.05092  0.00164  0.05623  0.50547
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -4.080e-01  1.376e-01  -2.966 0.003061 **
## regionnew_england            -1.608e-01  9.842e-02  -1.634 0.102451
## regionmid_east               -1.345e-01  9.829e-02  -1.368 0.171519
## regiongreat_lakes            -1.249e-01  9.827e-02  -1.271 0.203814
## regionplains                 -1.235e-01  9.823e-02  -1.257 0.208809
## regionsoutheast              -1.376e-01  9.810e-02  -1.403 0.160787
## regionsouthwest              -1.604e-01  9.843e-02  -1.629 0.103496
## regionrocky_mountains        -1.181e-01  9.919e-02  -1.191 0.234005
## regionfar_west               -9.636e-02  9.852e-02  -0.978 0.328198
## regionoutlying_areas          7.234e-02  9.997e-02   0.724 0.469400
## admission_rate               -2.655e-01  7.607e-02  -3.490 0.000497 ***
## faculty_salary_avg            4.943e-05  9.570e-06   5.165 2.73e-07 ***
## expenditure_per_student      -2.443e-06  6.403e-07  -3.815 0.000142 ***
## cost_attendance_per_year      1.363e-06  2.877e-07   4.736 2.39e-06 ***
## tuition_revenue_per_student  -2.616e-06  5.681e-07  -4.605 4.48e-06 ***
## debt_median_all               5.317e-06  7.895e-07   6.734 2.35e-11 ***
## I(admission_rate^2)           1.480e-01  5.681e-02   2.606 0.009256 **
## sqrt(faculty_salary_avg)     -5.323e-03  1.741e-03  -3.057 0.002278 **
## log(expenditure_per_student)  8.680e-02  1.043e-02   8.321  < 2e-16 ***
## sqrt(family_income)           1.632e-03  8.332e-05  19.587  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09674 on 1489 degrees of freedom
## Multiple R-squared:  0.7119, Adjusted R-squared:  0.7082
## F-statistic: 193.6 on 19 and 1489 DF,  p-value: < 2.2e-16
```

Best Subset Selection

```
reg_best <- regsubsets(formula(lmod_full), data = inst_data, nvmax = 25)
(summary_best <- summary(reg_best))

## Subset selection object
## Call: regsubsets.formula(formula(lmod_full), data = inst_data, nvmax = 25)
## 23 Variables  (and intercept)
##                                    Forced in Forced out
## institution_typeprivate_non_profit     FALSE      FALSE
## institution_typeprivate_for_profit     FALSE      FALSE
## regionnew_england                      FALSE      FALSE
```

```
## regionmid_east                            FALSE      FALSE
## regiongreat_lakes                         FALSE      FALSE
## regionplains                              FALSE      FALSE
## regionsoutheast                           FALSE      FALSE
## regionsouthwest                           FALSE      FALSE
## regionrocky_mountains                     FALSE      FALSE
## regionfar_west                            FALSE      FALSE
## regionoutlying_areas                      FALSE      FALSE
## admission_rate                            FALSE      FALSE
## faculty_fulltime                          FALSE      FALSE
## faculty_salary_avg                        FALSE      FALSE
## expenditure_per_student                   FALSE      FALSE
## family_income                             FALSE      FALSE
## cost_attendance_per_year                  FALSE      FALSE
## tuition_revenue_per_student               FALSE      FALSE
## debt_median_all                           FALSE      FALSE
## I(admission_rate^2)                       FALSE      FALSE
## sqrt(faculty_salary_avg)                  FALSE      FALSE
## log(expenditure_per_student)              FALSE      FALSE
## sqrt(family_income)                       FALSE      FALSE
## 1 subsets of each size up to 23
## Selection Algorithm: exhaustive
##           institution_typeprivate_non_profit
institution_typeprivate_for_profit
## 1  ( 1 )  " "                                " "
## 2  ( 1 )  " "                                " "
## 3  ( 1 )  " "                                " "
## 4  ( 1 )  " "                                " "
## 5  ( 1 )  " "                                " "
## 6  ( 1 )  " "                                " "
## 7  ( 1 )  " "                                " "
## 8  ( 1 )  " "                                "*"
## 9  ( 1 )  "*"                                " "
## 10  ( 1 ) "*"                                " "
## 11  ( 1 ) "*"                                " "
## 12  ( 1 ) " "                                " "
## 13  ( 1 ) " "                                " "
## 14  ( 1 ) " "                                " "
## 15  ( 1 ) " "                                "*"
## 16  ( 1 ) " "                                "*"
## 17  ( 1 ) " "                                "*"
## 18  ( 1 ) " "                                "*"
## 19  ( 1 ) " "                                "*"
## 20  ( 1 ) " "                                "*"
## 21  ( 1 ) " "                                "*"
## 22  ( 1 ) " "                                "*"
## 23  ( 1 ) "*"                                "*"
##           regionnew_england regionmid_east regiongreat_lakes regionplains
## 1  ( 1 )  " "               " "            " "               " "
## 2  ( 1 )  " "               " "            " "               " "
```

```
## 3  ( 1 )  " "                " "             " "                 " "
## 4  ( 1 )  " "                " "             " "                 " "
## 5  ( 1 )  " "                " "             " "                 " "
## 6  ( 1 )  " "                " "             " "                 " "
## 7  ( 1 )  " "                " "             " "                 " "
## 8  ( 1 )  " "                " "             " "                 " "
## 9  ( 1 )  " "                " "             " "                 " "
## 10  ( 1 ) " "                " "             " "                 " "
## 11  ( 1 ) "*"                " "             " "                 " "
## 12  ( 1 ) "*"                " "             " "                 " "
## 13  ( 1 ) "*"                " "             " "                 " "
## 14  ( 1 ) "*"                " "             " "                 " "
## 15  ( 1 ) "*"                " "             " "                 " "
## 16  ( 1 ) "*"                " "             " "                 " "
## 17  ( 1 ) "*"                "*"             " "                 " "
## 18  ( 1 ) "*"                "*"             " "                 " "
## 19  ( 1 ) "*"                "*"             " "                 " "
## 20  ( 1 ) "*"                "*"             "*"                 "*"
## 21  ( 1 ) "*"                "*"             "*"                 "*"
## 22  ( 1 ) "*"                "*"             "*"                 "*"
## 23  ( 1 ) "*"                "*"             "*"                 "*"
##           regionsoutheast regionsouthwest regionrocky_mountains
regionfar_west
## 1  ( 1 )  " "              " "             " "                   " "
## 2  ( 1 )  " "              " "             " "                   " "
## 3  ( 1 )  " "              " "             " "                   " "
## 4  ( 1 )  " "              " "             " "                   " "
## 5  ( 1 )  " "              " "             " "                   " "
## 6  ( 1 )  " "              " "             " "                   " "
## 7  ( 1 )  " "              " "             " "                   "*"
## 8  ( 1 )  " "              " "             " "                   "*"
## 9  ( 1 )  " "              " "             " "                   "*"
## 10  ( 1 ) " "              "*"             " "                   "*"
## 11  ( 1 ) " "              "*"             " "                   "*"
## 12  ( 1 ) " "              " "             " "                   "*"
## 13  ( 1 ) " "              "*"             " "                   "*"
## 14  ( 1 ) " "              "*"             " "                   "*"
## 15  ( 1 ) " "              "*"             " "                   "*"
## 16  ( 1 ) "*"              "*"             " "                   "*"
## 17  ( 1 ) "*"              "*"             " "                   "*"
## 18  ( 1 ) "*"              "*"             " "                   "*"
## 19  ( 1 ) "*"              "*"             " "                   "*"
## 20  ( 1 ) "*"              "*"             "*"                   "*"
## 21  ( 1 ) "*"              "*"             "*"                   "*"
## 22  ( 1 ) "*"              "*"             "*"                   "*"
## 23  ( 1 ) "*"              "*"             "*"                   "*"
##           regionoutlying_areas admission_rate faculty_fulltime
## 1  ( 1 )  " "                  " "            " "
## 2  ( 1 )  " "                  " "            " "
## 3  ( 1 )  " "                  " "            " "
```

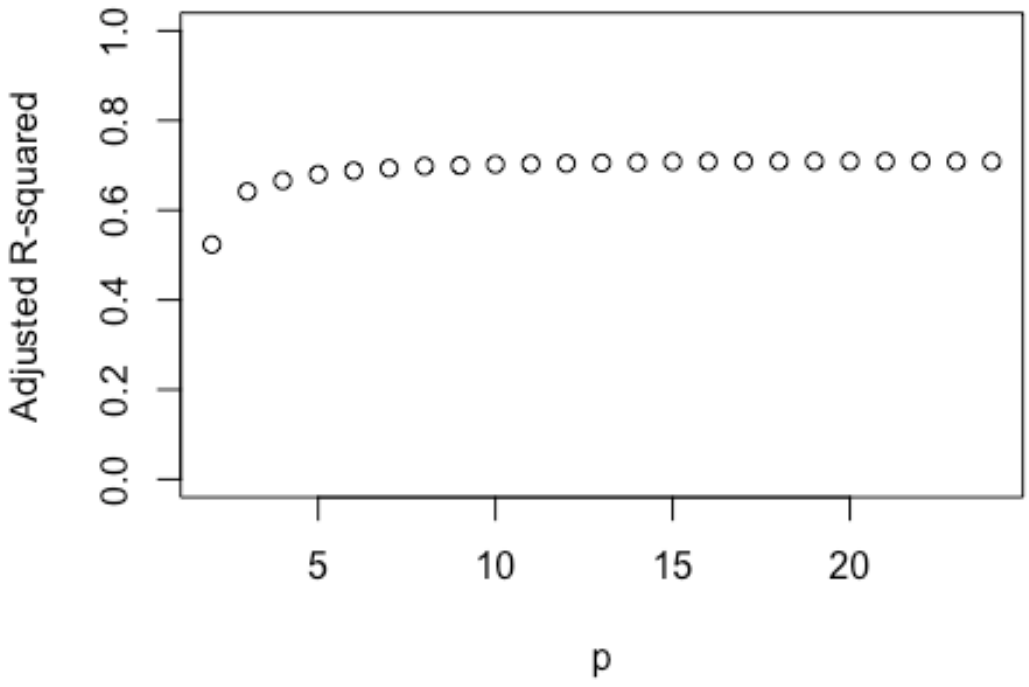
```
## 4  ( 1 )  "*"                  " "             " "
## 5  ( 1 )  "*"                  " "             " "
## 6  ( 1 )  "*"                  "*"             " "
## 7  ( 1 )  "*"                  "*"             " "
## 8  ( 1 )  "*"                  "*"             " "
## 9  ( 1 )  "*"                  "*"             " "
## 10  ( 1 ) "*"                  "*"             " "
## 11  ( 1 ) "*"                  "*"             " "
## 12  ( 1 ) "*"                  "*"             " "
## 13  ( 1 ) "*"                  "*"             " "
## 14  ( 1 ) "*"                  "*"             " "
## 15  ( 1 ) "*"                  "*"             " "
## 16  ( 1 ) "*"                  "*"             " "
## 17  ( 1 ) "*"                  "*"             " "
## 18  ( 1 ) "*"                  "*"             "*"
## 19  ( 1 ) "*"                  "*"             "*"
## 20  ( 1 ) " "                  "*"             "*"
## 21  ( 1 ) " "                  "*"             "*"
## 22  ( 1 ) "*"                  "*"             "*"
## 23  ( 1 ) "*"                  "*"             "*"
##           faculty_salary_avg expenditure_per_student family_income
## 1  ( 1 )  " "                " "                   "*"
## 2  ( 1 )  "*"                " "                   "*"
## 3  ( 1 )  "*"                " "                   "*"
## 4  ( 1 )  "*"                " "                   " "
## 5  ( 1 )  "*"                " "                   " "
## 6  ( 1 )  "*"                " "                   " "
## 7  ( 1 )  "*"                " "                   " "
## 8  ( 1 )  "*"                " "                   " "
## 9  ( 1 )  "*"                " "                   " "
## 10  ( 1 ) "*"                " "                   " "
## 11  ( 1 ) "*"                " "                   " "
## 12  ( 1 ) "*"                "*"                   " "
## 13  ( 1 ) "*"                "*"                   " "
## 14  ( 1 ) "*"                "*"                   " "
## 15  ( 1 ) "*"                "*"                   " "
## 16  ( 1 ) "*"                "*"                   " "
## 17  ( 1 ) "*"                "*"                   " "
## 18  ( 1 ) "*"                "*"                   " "
## 19  ( 1 ) "*"                "*"                   "*"
## 20  ( 1 ) "*"                "*"                   " "
## 21  ( 1 ) "*"                "*"                   "*"
## 22  ( 1 ) "*"                "*"                   "*"
## 23  ( 1 ) "*"                "*"                   "*"
##           cost_attendance_per_year tuition_revenue_per_student
debt_median_all
## 1  ( 1 )  " "                      " "                         " "
## 2  ( 1 )  " "                      " "                         " "
## 3  ( 1 )  " "                      " "                         " "
## 4  ( 1 )  " "                      " "                         " "
```

```
## 5  ( 1 )  " "                      " "                            "*"
## 6  ( 1 )  " "                      " "                            "*"
## 7  ( 1 )  " "                      " "                            "*"
## 8  ( 1 )  " "                      " "                            "*"
## 9  ( 1 )  " "                      "*"                            "*"
## 10  ( 1 ) " "                      "*"                            "*"
## 11  ( 1 ) " "                      "*"                            "*"
## 12  ( 1 ) "*"                      "*"                            "*"
## 13  ( 1 ) "*"                      "*"                            "*"
## 14  ( 1 ) "*"                      "*"                            "*"
## 15  ( 1 ) "*"                      "*"                            "*"
## 16  ( 1 ) "*"                      "*"                            "*"
## 17  ( 1 ) "*"                      "*"                            "*"
## 18  ( 1 ) "*"                      "*"                            "*"
## 19  ( 1 ) "*"                      "*"                            "*"
## 20  ( 1 ) "*"                      "*"                            "*"
## 21  ( 1 ) "*"                      "*"                            "*"
## 22  ( 1 ) "*"                      "*"                            "*"
## 23  ( 1 ) "*"                      "*"                            "*"
##           I(admission_rate^2) sqrt(faculty_salary_avg)
## 1  ( 1 )  " "                 " "
## 2  ( 1 )  " "                 " "
## 3  ( 1 )  " "                 " "
## 4  ( 1 )  " "                 " "
## 5  ( 1 )  " "                 " "
## 6  ( 1 )  " "                 " "
## 7  ( 1 )  " "                 " "
## 8  ( 1 )  " "                 " "
## 9  ( 1 )  " "                 " "
## 10  ( 1 ) " "                 " "
## 11  ( 1 ) " "                 " "
## 12  ( 1 ) " "                 "*"
## 13  ( 1 ) " "                 "*"
## 14  ( 1 ) "*"                 "*"
## 15  ( 1 ) "*"                 "*"
## 16  ( 1 ) "*"                 "*"
## 17  ( 1 ) "*"                 "*"
## 18  ( 1 ) "*"                 "*"
## 19  ( 1 ) "*"                 "*"
## 20  ( 1 ) "*"                 "*"
## 21  ( 1 ) "*"                 "*"
## 22  ( 1 ) "*"                 "*"
## 23  ( 1 ) "*"                 "*"
##           log(expenditure_per_student) sqrt(family_income)
## 1  ( 1 )  " "                          " "
## 2  ( 1 )  " "                          " "
## 3  ( 1 )  "*"                          " "
## 4  ( 1 )  "*"                          "*"
## 5  ( 1 )  "*"                          "*"
## 6  ( 1 )  "*"                          "*"
```

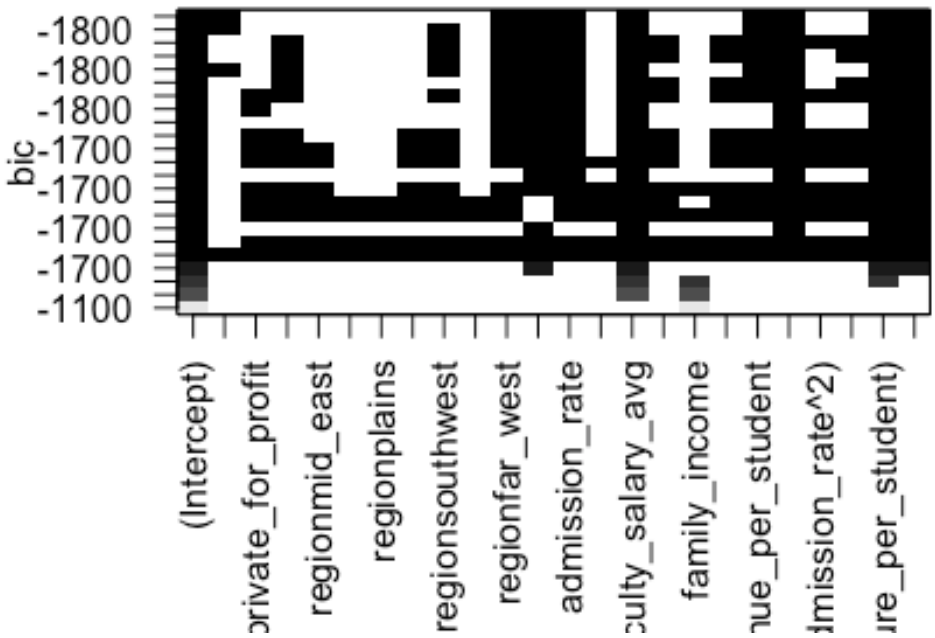
```
## 7  ( 1 )  "*"                                  "*"
## 8  ( 1 )  "*"                                  "*"
## 9  ( 1 )  "*"                                  "*"
## 10  ( 1 ) "*"                                  "*"
## 11  ( 1 ) "*"                                  "*"
## 12  ( 1 ) "*"                                  "*"
## 13  ( 1 ) "*"                                  "*"
## 14  ( 1 ) "*"                                  "*"
## 15  ( 1 ) "*"                                  "*"
## 16  ( 1 ) "*"                                  "*"
## 17  ( 1 ) "*"                                  "*"
## 18  ( 1 ) "*"                                  "*"
## 19  ( 1 ) "*"                                  "*"
## 20  ( 1 ) "*"                                  "*"
## 21  ( 1 ) "*"                                  "*"
## 22  ( 1 ) "*"                                  "*"
## 23  ( 1 ) "*"                                  "*"
```

Adj R2 by Regressors

```
plot(summary_best$adjr2 ~ seq(from = 2, to = length(summary_best$adjr2) + 1),
  xlab = "p",
  ylab = "Adjusted R-squared",
  ylim = c(0, 1))
```



BIC Scale




```
##                          (Intercept) institution_typeprivate_non_profit
##                                 TRUE                                TRUE
## institution_typeprivate_for_profit                  regionnew_england
##                                FALSE                               FALSE
##                        regionmid_east                  regiongreat_lakes
##                                FALSE                               FALSE
##                          regionplains                    regionsoutheast
##                                FALSE                               FALSE
##                       regionsouthwest               regionrocky_mountains
##                                FALSE                               FALSE
##                       regionfar_west                 regionoutlying_areas
##                                 TRUE                                TRUE
##                       admission_rate                    faculty_fulltime
##                                 TRUE                               FALSE
##                  faculty_salary_avg              expenditure_per_student
##                                 TRUE                               FALSE
##                        family_income            cost_attendance_per_year
##                                FALSE                               FALSE
##         tuition_revenue_per_student                     debt_median_all
##                                 TRUE                                TRUE
##                  I(admission_rate^2)             sqrt(faculty_salary_avg)
```

```
##                                FALSE                                FALSE
##       log(expenditure_per_student)                sqrt(family_income)
##                                TRUE                                 TRUE
```

```
# Fixing the model hierarchy of the best subset model (BIC)
lmod_best_bic <- lm(completion_rate_200 ~ institution_type + region +
admission_rate + faculty_salary_avg + tuition_revenue_per_student +
debt_median_all + log(expenditure_per_student) + sqrt(family_income), data =
inst_data)
(summary_best_bic <- summary(lmod_best_bic))
```

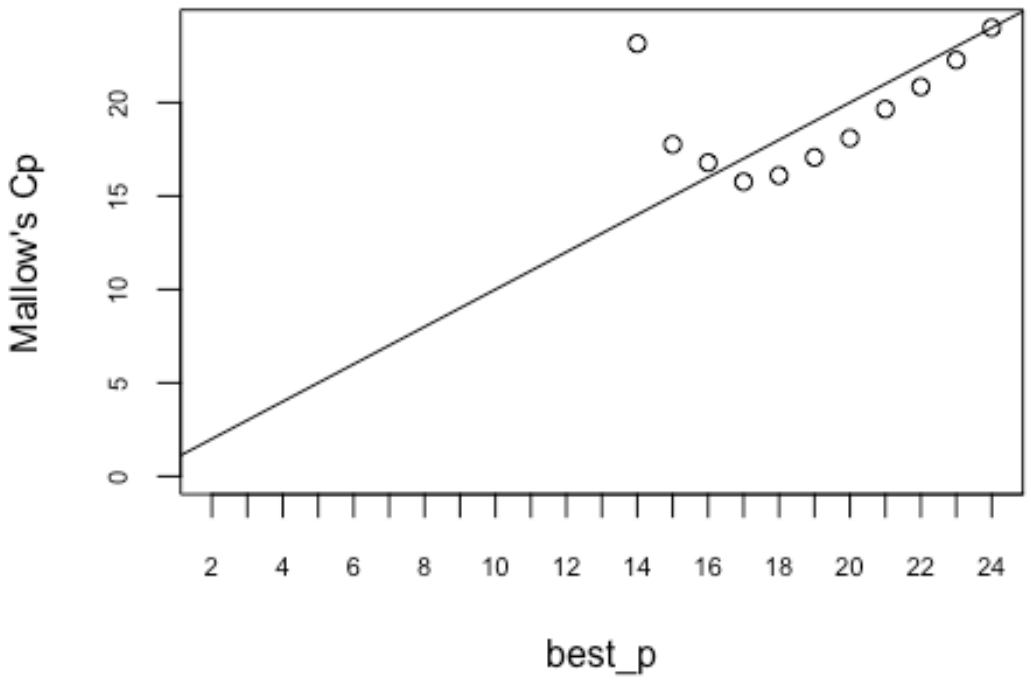
```
##
## Call:
## lm(formula = completion_rate_200 ~ institution_type + region +
##     admission_rate + faculty_salary_avg + tuition_revenue_per_student +
##     debt_median_all + log(expenditure_per_student) + sqrt(family_income),
##     data = inst_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61003 -0.05009  0.00003  0.05771  0.50545
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -4.996e-01  1.165e-01  -4.289 1.91e-05
***
## institution_typeprivate_non_profit  2.716e-02  8.032e-03   3.382 0.000738
***
## institution_typeprivate_for_profit -1.668e-02  1.881e-02  -0.887 0.375169
## regionnew_england                -1.544e-01  9.871e-02  -1.564 0.117962
## regionmid_east                   -1.328e-01  9.853e-02  -1.348 0.177804
## regiongreat_lakes                -1.282e-01  9.852e-02  -1.301 0.193461
## regionplains                     -1.279e-01  9.848e-02  -1.298 0.194339
## regionsoutheast                  -1.387e-01  9.836e-02  -1.410 0.158778
## regionsouthwest                  -1.608e-01  9.873e-02  -1.629 0.103544
## regionrocky_mountains            -1.181e-01  9.957e-02  -1.186 0.235731
## regionfar_west                   -9.348e-02  9.872e-02  -0.947 0.343808
## regionoutlying_areas              6.877e-02  1.002e-01   0.686 0.492700
## admission_rate                   -7.736e-02  1.513e-02  -5.115 3.55e-07
***
## faculty_salary_avg                2.177e-05  1.577e-06  13.803  < 2e-16
***
## tuition_revenue_per_student      -1.541e-06  5.700e-07  -2.704 0.006926
**
## debt_median_all                   5.081e-06  7.609e-07   6.678 3.40e-11
***
## log(expenditure_per_student)      6.120e-02  7.469e-03   8.194 5.37e-16
***
## sqrt(family_income)               1.716e-03  7.854e-05  21.850  < 2e-16
***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09741 on 1491 degrees of freedom
## Multiple R-squared:  0.7074, Adjusted R-squared:  0.7041
## F-statistic: 212.1 on 17 and 1491 DF,  p-value: < 2.2e-16

nrow(summary_best_bic$coefficients)

## [1] 18
```

Using Mallow's Cp




```
##                          (Intercept) institution_typeprivate_non_profit
##                                 TRUE                                FALSE
## institution_typeprivate_for_profit                   regionnew_england
##                                 TRUE                                 TRUE
##                       regionmid_east                   regiongreat_lakes
##                                FALSE                                FALSE
##                         regionplains                     regionsoutheast
##                                FALSE                                FALSE
##                       regionsouthwest                regionrocky_mountains
##                                 TRUE                                FALSE
##                        regionfar_west                 regionoutlying_areas
```

```
##                                TRUE                                TRUE
##                      admission_rate                    faculty_fulltime
##                                TRUE                               FALSE
##                  faculty_salary_avg             expenditure_per_student
##                                TRUE                                TRUE
##                       family_income            cost_attendance_per_year
##                               FALSE                                TRUE
##         tuition_revenue_per_student                      debt_median_all
##                                TRUE                                TRUE
##                 I(admission_rate^2)            sqrt(faculty_salary_avg)
##                                TRUE                                TRUE
##        log(expenditure_per_student)                 sqrt(family_income)
##                                TRUE                                TRUE
```

```
# Fixing the model hierarchy of the best subset model (Mallow's Cp)
lmod_best_mcp <- lm(completion_rate_200 ~ institution_type + region +
admission_rate + faculty_salary_avg + expenditure_per_student +
cost_attendance_per_year + tuition_revenue_per_student + debt_median_all +
I(admission_rate^2) + sqrt(faculty_salary_avg) + log(expenditure_per_student)
+ sqrt(family_income), data = inst_data)
(summary_best_mcp <- summary(lmod_best_mcp))
```

```
##
## Call:
## lm(formula = completion_rate_200 ~ institution_type + region +
##     admission_rate + faculty_salary_avg + expenditure_per_student +
##     cost_attendance_per_year + tuition_revenue_per_student +
##     debt_median_all + I(admission_rate^2) + sqrt(faculty_salary_avg) +
##     log(expenditure_per_student) + sqrt(family_income), data = inst_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59648 -0.05099  0.00109  0.05633  0.50259
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -3.898e-01  1.484e-01  -2.626 0.008736
**
## institution_typeprivate_non_profit  5.095e-03  1.140e-02   0.447 0.655036
## institution_typeprivate_for_profit -2.608e-02  1.990e-02  -1.311 0.190202
## regionnew_england                  -1.586e-01  9.860e-02  -1.608 0.107940
## regionmid_east                     -1.317e-01  9.842e-02  -1.338 0.180951
## regiongreat_lakes                  -1.233e-01  9.837e-02  -1.253 0.210254
## regionplains                       -1.222e-01  9.832e-02  -1.242 0.214273
## regionsoutheast                    -1.362e-01  9.825e-02  -1.386 0.165847
## regionsouthwest                    -1.579e-01  9.856e-02  -1.602 0.109335
## regionrocky_mountains              -1.146e-01  9.934e-02  -1.154 0.248692
## regionfar_west                     -9.383e-02  9.869e-02  -0.951 0.341843
## regionoutlying_areas                7.029e-02  1.001e-01   0.702 0.482681
## admission_rate                     -2.691e-01  7.606e-02  -3.538 0.000416
```

```
***
## faculty_salary_avg                   4.751e-05  9.654e-06   4.922 9.53e-07
***
## expenditure_per_student             -2.317e-06  6.474e-07  -3.579 0.000355
***
## cost_attendance_per_year             1.188e-06  3.958e-07   3.003 0.002720
**
## tuition_revenue_per_student         -2.419e-06  5.995e-07  -4.036 5.72e-05
***
## debt_median_all                      5.331e-06  7.927e-07   6.726 2.48e-11
***
## I(admission_rate^2)                  1.487e-01  5.679e-02   2.618 0.008931
**
## sqrt(faculty_salary_avg)            -4.897e-03  1.780e-03  -2.751 0.006008
**
## log(expenditure_per_student)         8.255e-02  1.116e-02   7.395 2.36e-13
***
## sqrt(family_income)                  1.625e-03  8.683e-05  18.720  < 2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09669 on 1487 degrees of freedom
## Multiple R-squared:  0.7125, Adjusted R-squared:  0.7085
## F-statistic: 175.5 on 21 and 1487 DF,  p-value: < 2.2e-16
```

```
length(coef(lmod_best_mcp))
```

```
## [1] 22
```

Stepwise Selection with BIC

```
lmod_step <- step(lmod_full, direction = "both", k = log(nobs(lmod_full)))
```

```
## Start:  AIC=-6899.26
## completion_rate_200 ~ institution_type + region + admission_rate +
##     faculty_fulltime + faculty_salary_avg + expenditure_per_student +
##     family_income + cost_attendance_per_year + tuition_revenue_per_student
+
##     debt_median_all + I(admission_rate^2) + sqrt(faculty_salary_avg) +
##     log(expenditure_per_student) + sqrt(family_income)
##
##                                Df Sum of Sq    RSS     AIC
## - institution_type              2   0.02407 13.908 -6911.3
## - family_income                 1   0.00896 13.893 -6905.6
## - faculty_fulltime              1   0.01111 13.895 -6905.4
## - I(admission_rate^2)           1   0.06288 13.947 -6899.8
## <none>                                      13.884 -6899.3
## - sqrt(faculty_salary_avg)      1   0.06893 13.953 -6899.1
## - cost_attendance_per_year      1   0.08340 13.967 -6897.5
## - admission_rate                1   0.11666 14.000 -6894.0
```

```
## - expenditure_per_student       1   0.11769 14.001 -6893.8
## - tuition_revenue_per_student   1   0.13985 14.024 -6891.5
## - sqrt(family_income)           1   0.17836 14.062 -6887.3
## - faculty_salary_avg            1   0.22383 14.107 -6882.4
## - debt_median_all               1   0.42424 14.308 -6861.2
## - log(expenditure_per_student)  1   0.48508 14.369 -6854.8
## - region                        9   1.22113 15.105 -6837.9
##
## Step:  AIC=-6911.29
## completion_rate_200 ~ region + admission_rate + faculty_fulltime +
##     faculty_salary_avg + expenditure_per_student + family_income +
##     cost_attendance_per_year + tuition_revenue_per_student +
##     debt_median_all + I(admission_rate^2) + sqrt(faculty_salary_avg) +
##     log(expenditure_per_student) + sqrt(family_income)
##
##                                 Df Sum of Sq    RSS     AIC
## - family_income                  1   0.01120 13.919 -6917.4
## - faculty_fulltime               1   0.01677 13.925 -6916.8
## - I(admission_rate^2)            1   0.06214 13.970 -6911.9
## <none>                                       13.908 -6911.3
## - sqrt(faculty_salary_avg)       1   0.08446 13.992 -6909.5
## - admission_rate                 1   0.11402 14.022 -6906.3
## - expenditure_per_student        1   0.12983 14.038 -6904.6
## - tuition_revenue_per_student    1   0.17066 14.078 -6900.2
## + institution_type               2   0.02407 13.884 -6899.3
## - sqrt(family_income)            1   0.18900 14.097 -6898.2
## - cost_attendance_per_year       1   0.21286 14.121 -6895.7
## - faculty_salary_avg             1   0.24458 14.152 -6892.3
## - debt_median_all                1   0.42807 14.336 -6872.9
## - log(expenditure_per_student)   1   0.58306 14.491 -6856.6
## - region                         9   1.26025 15.168 -6846.3
##
## Step:  AIC=-6917.39
## completion_rate_200 ~ region + admission_rate + faculty_fulltime +
##     faculty_salary_avg + expenditure_per_student +
cost_attendance_per_year +
##     tuition_revenue_per_student + debt_median_all + I(admission_rate^2) +
##     sqrt(faculty_salary_avg) + log(expenditure_per_student) +
##     sqrt(family_income)
##
##                                 Df Sum of Sq    RSS     AIC
## - faculty_fulltime               1    0.0161 13.935 -6923.0
## - I(admission_rate^2)            1    0.0623 13.981 -6918.0
## <none>                                       13.919 -6917.4
## - sqrt(faculty_salary_avg)       1    0.0848 14.004 -6915.5
## - admission_rate                 1    0.1127 14.032 -6912.5
## + family_income                  1    0.0112 13.908 -6911.3
## - expenditure_per_student        1    0.1306 14.050 -6910.6
## + institution_type               2    0.0263 13.893 -6905.6
## - tuition_revenue_per_student    1    0.1815 14.101 -6905.2
```

```
## - cost_attendance_per_year      1    0.2102 14.129 -6902.1
## - faculty_salary_avg            1    0.2435 14.162 -6898.5
## - debt_median_all               1    0.4264 14.345 -6879.2
## - log(expenditure_per_student)  1    0.6029 14.522 -6860.7
## - region                        9    1.3844 15.303 -6840.2
## - sqrt(family_income)           1    3.4137 17.333 -6593.7
##
## Step:  AIC=-6922.97
## completion_rate_200 ~ region + admission_rate + faculty_salary_avg +
##     expenditure_per_student + cost_attendance_per_year +
tuition_revenue_per_student +
##     debt_median_all + I(admission_rate^2) + sqrt(faculty_salary_avg) +
##     log(expenditure_per_student) + sqrt(family_income)
##
##                                Df Sum of Sq    RSS     AIC
## - I(admission_rate^2)           1    0.0636 13.999 -6923.4
## <none>                                      13.935 -6923.0
## - sqrt(faculty_salary_avg)      1    0.0874 14.023 -6920.8
## - admission_rate                1    0.1140 14.049 -6918.0
## + faculty_fulltime              1    0.0161 13.919 -6917.4
## + family_income                 1    0.0105 13.925 -6916.8
## - expenditure_per_student       1    0.1362 14.071 -6915.6
## + institution_type              2    0.0320 13.903 -6911.8
## - tuition_revenue_per_student   1    0.1984 14.133 -6909.0
## - cost_attendance_per_year      1    0.2099 14.145 -6907.7
## - faculty_salary_avg            1    0.2496 14.185 -6903.5
## - debt_median_all               1    0.4244 14.360 -6885.0
## - log(expenditure_per_student)  1    0.6480 14.583 -6861.7
## - region                        9    1.4183 15.353 -6842.6
## - sqrt(family_income)           1    3.5905 17.526 -6584.3
##
## Step:  AIC=-6923.42
## completion_rate_200 ~ region + admission_rate + faculty_salary_avg +
##     expenditure_per_student + cost_attendance_per_year +
tuition_revenue_per_student +
##     debt_median_all + sqrt(faculty_salary_avg) +
log(expenditure_per_student) +
##     sqrt(family_income)
##
##                                Df Sum of Sq    RSS     AIC
## <none>                                      13.999 -6923.4
## + I(admission_rate^2)           1    0.0636 13.935 -6923.0
## - expenditure_per_student       1    0.0931 14.092 -6920.7
## - sqrt(faculty_salary_avg)      1    0.1087 14.107 -6919.1
## + faculty_fulltime              1    0.0174 13.981 -6918.0
## + family_income                 1    0.0107 13.988 -6917.3
## + institution_type              2    0.0315 13.967 -6912.2
## - tuition_revenue_per_student   1    0.1772 14.176 -6911.8
## - admission_rate                1    0.2013 14.200 -6909.2
## - cost_attendance_per_year      1    0.2139 14.213 -6907.9
```

```
## - faculty_salary_avg            1    0.2929 14.292 -6899.5
## - debt_median_all               1    0.3834 14.382 -6890.0
## - log(expenditure_per_student)  1    0.6062 14.605 -6866.8
## - region                        9    1.3937 15.392 -6846.1
## - sqrt(family_income)           1    3.5471 17.546 -6589.9

(summary_step <- summary(lmod_step))

##
## Call:
## lm(formula = completion_rate_200 ~ region + admission_rate +
##     faculty_salary_avg + expenditure_per_student +
cost_attendance_per_year +
##     tuition_revenue_per_student + debt_median_all +
sqrt(faculty_salary_avg) +
##     log(expenditure_per_student) + sqrt(family_income), data = inst_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60999 -0.05183  0.00168  0.05730  0.51776
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -3.918e-01  1.377e-01  -2.846 0.004492 **
## regionnew_england            -1.785e-01  9.837e-02  -1.815 0.069770 .
## regionmid_east               -1.544e-01  9.818e-02  -1.573 0.115949
## regiongreat_lakes            -1.451e-01  9.816e-02  -1.478 0.139616
## regionplains                 -1.431e-01  9.814e-02  -1.458 0.145031
## regionsoutheast              -1.571e-01  9.800e-02  -1.603 0.109079
## regionsouthwest              -1.794e-01  9.835e-02  -1.825 0.068255 .
## regionrocky_mountains        -1.362e-01  9.914e-02  -1.373 0.169836
## regionfar_west               -1.171e-01  9.839e-02  -1.191 0.234014
## regionoutlying_areas          5.215e-02  9.986e-02   0.522 0.601546
## admission_rate               -7.138e-02  1.542e-02  -4.629 4.00e-06 ***
## faculty_salary_avg            5.299e-05  9.490e-06   5.584 2.79e-08 ***
## expenditure_per_student      -1.917e-06  6.089e-07  -3.149 0.001672 **
## cost_attendance_per_year      1.375e-06  2.882e-07   4.771 2.01e-06 ***
## tuition_revenue_per_student  -2.458e-06  5.660e-07  -4.344 1.50e-05 ***
## debt_median_all               4.989e-06  7.810e-07   6.388 2.24e-10 ***
## sqrt(faculty_salary_avg)     -5.889e-03  1.731e-03  -3.402 0.000688 ***
## log(expenditure_per_student)  8.323e-02  1.036e-02   8.032 1.93e-15 ***
## sqrt(family_income)           1.619e-03  8.334e-05  19.431  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09693 on 1490 degrees of freedom
## Multiple R-squared:  0.7105, Adjusted R-squared:  0.707
## F-statistic: 203.2 on 18 and 1490 DF,  p-value: < 2.2e-16
```

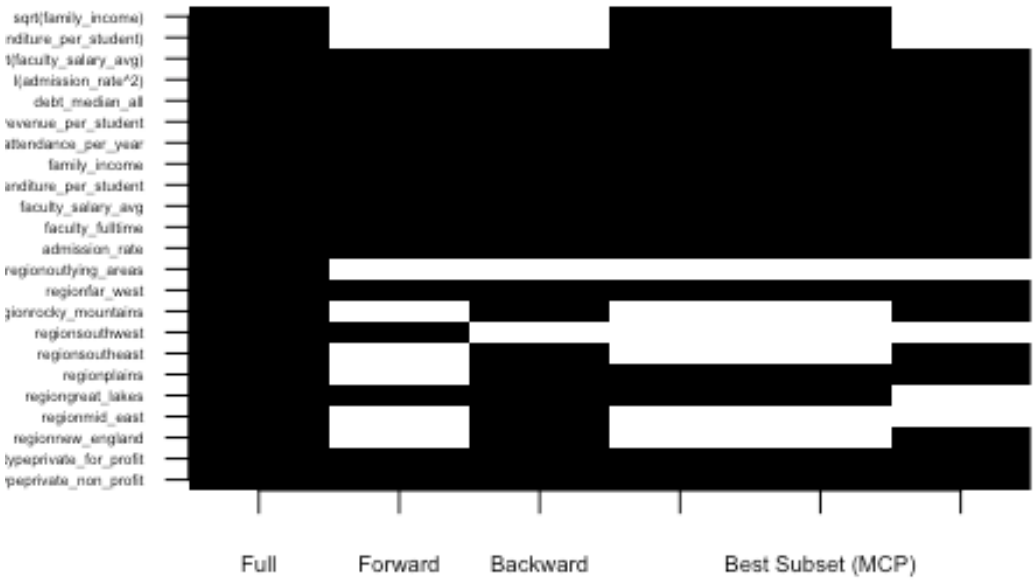
```r
# Fixing the model hierarchy of the stepwise model
lmod_step <- lm(completion_rate_200 ~ region + admission_rate +
faculty_salary_avg + expenditure_per_student + cost_attendance_per_year +
tuition_revenue_per_student + sqrt(faculty_salary_avg) +
log(expenditure_per_student) + sqrt(family_income), data = inst_data)
(summary_step <- summary(lmod_step))
```

```
##
## Call:
## lm(formula = completion_rate_200 ~ region + admission_rate +
##     faculty_salary_avg + expenditure_per_student +
cost_attendance_per_year +
##     tuition_revenue_per_student + sqrt(faculty_salary_avg) +
##     log(expenditure_per_student) + sqrt(family_income), data = inst_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67686 -0.05361  0.00203  0.05875  0.51981
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -5.084e-01  1.383e-01  -3.677 0.000245 ***
## regionnew_england            -1.424e-01  9.951e-02  -1.431 0.152575
## regionmid_east               -1.155e-01  9.929e-02  -1.163 0.245050
## regiongreat_lakes            -1.068e-01  9.927e-02  -1.075 0.282330
## regionplains                 -1.103e-01  9.930e-02  -1.111 0.266661
## regionsoutheast              -1.239e-01  9.916e-02  -1.250 0.211621
## regionsouthwest              -1.516e-01  9.955e-02  -1.523 0.128046
## regionrocky_mountains        -1.099e-01  1.004e-01  -1.095 0.273568
## regionfar_west               -8.557e-02  9.956e-02  -0.859 0.390245
## regionoutlying_areas          6.851e-02  1.011e-01   0.677 0.498318
## admission_rate               -5.695e-02  1.546e-02  -3.685 0.000237 ***
## faculty_salary_avg            4.456e-05  9.523e-06   4.679 3.14e-06 ***
## expenditure_per_student      -2.365e-06  6.129e-07  -3.859 0.000119 ***
## cost_attendance_per_year      1.689e-06  2.877e-07   5.870 5.36e-09 ***
## tuition_revenue_per_student  -2.176e-06  5.717e-07  -3.806 0.000147 ***
## sqrt(faculty_salary_avg)     -4.313e-03  1.736e-03  -2.484 0.013093 *
## log(expenditure_per_student)  8.777e-02  1.047e-02   8.380  < 2e-16 ***
## sqrt(family_income)           1.734e-03  8.246e-05  21.028  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09821 on 1491 degrees of freedom
## Multiple R-squared:  0.7026, Adjusted R-squared:  0.6992
## F-statistic: 207.2 on 17 and 1491 DF,  p-value: < 2.2e-16
```

```r
length(coef(lmod_step))
```

```
## [1] 18
```

Choosing a Model



Comparing Adjustd R2's

```
summary_full$adj.r.squared

## [1] 0.7084767

summary_forward$adj.r.squared

## [1] 0.7007426

summary_backward$adj.r.squared

## [1] 0.7081838

summary_best_bic$adj.r.squared

## [1] 0.7041033

summary_best_mcp$adj.r.squared

## [1] 0.7084626

summary_step$adj.r.squared
```

```
## [1] 0.6992279
```

Comparing number of regressors

```
nrow(summary_full$coefficients)

## [1] 24

nrow(summary_forward$coefficients)

## [1] 16

nrow(summary_backward$coefficients)

## [1] 20

nrow(summary_best_bic$coefficients)

## [1] 18

nrow(summary_best_mcp$coefficients)

## [1] 22

nrow(summary_step$coefficients)

## [1] 18
```

Cross-Validation

5-Fold Cross Validation

```
# Since all the models have a fairly similar Adjusted R^2, we use cross-
validation to pick between the forward selection model and the best subset
(BIC) model, since both models have few regressors in comparison to the other
models
set.seed(42)
cv_5fold <- trainControl(method="cv", number = 5)

forward_model <- train(formula(lmod_forward), data = inst_data, trControl =
cv_5fold, method = "lm")

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

best_bic_model <- train(formula(lmod_best_bic), data = inst_data, trControl =
cv_5fold, method = "lm")

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading
```

```
resamp <- resamples(list(forward_model, best_bic_model), modelNames =
c("forward", "best subset (bic)"))
summary(resamp, metric = c("RMSE", "MAE"))

##
## Call:
## summary.resamples(object = resamp, metric = c("RMSE", "MAE"))
##
## Models: forward, best subset (bic)
## Number of resamples: 5
##
## RMSE
##                          Min.    1st Qu.     Median       Mean    3rd Qu.
## forward             0.09537694 0.09664832 0.09808443 0.09830883 0.09949287
## best subset (bic) 0.08709261 0.09832447 0.09905632 0.09840615 0.10070448
##                          Max. NA's
## forward             0.1019416    0
## best subset (bic) 0.1068528    0
##
## MAE
##                          Min.    1st Qu.     Median       Mean    3rd Qu.
## forward             0.06769364 0.07149713 0.07375519 0.07269759 0.07513131
## best subset (bic) 0.06866389 0.07089434 0.07173381 0.07294024 0.07431835
##                          Max. NA's
## forward             0.07541068    0
## best subset (bic) 0.07909083    0
```

LOO Cross Validation

```
cv_loo <- trainControl(method="LOOCV")
forward_model <- train(formula(lmod_forward), data = inst_data, trControl =
cv_loo, method = "lm")

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

best_bic_model <- train(formula(lmod_best_bic), data = inst_data, trControl =
cv_loo, method = "lm")

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

forward_model$results

##   intercept       RMSE  Rsquared        MAE
## 1      TRUE 0.09861364 0.6965984 0.07293377

best_bic_model$results

##   intercept       RMSE  Rsquared        MAE
## 1      TRUE 0.09836109 0.6981658 0.07288658
```

Explain what you did and your conclusions.

We ran ordinary fitting (the "full" model with all the variable transformations mainly mentioned in the Data Exploration section), forward selection, backward selection, best subset regression (with BIC and Mallow's Cp), and stepwise selection (with BIC).

We then compared the R_a^2 and number of regressors of these five models, discovering that all five models had very similar R_a^2 statistics (within 1% to 2%).

Thus, we further investigated the differences between the forward selection and best subset (by BIC) models, since these two models had the fewest regressors while also looking very promising (in terms of R_a^2 and average individual-variable explanatory power).

Using both 5-fold cross validation and LOO cross validation, we saw similar statistics for both models, and thus decided to go with the forward selection model since it has two fewere regressors.

Provide your final model

```
lmod_final <- lmod_forward
summary(lmod_final)
```

```
##
## Call:
## lm(formula = completion_rate_200 ~ region + admission_rate +
##     faculty_salary_avg + family_income + sqrt(family_income) +
##     log(expenditure_per_student) + debt_median_all, data = inst_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64575 -0.05111  0.00152  0.05841  0.49957
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -5.290e-01  1.243e-01  -4.255 2.22e-05 ***
## regionnew_england            -1.654e-01  9.892e-02  -1.672   0.0947 .
## regionmid_east               -1.429e-01  9.880e-02  -1.446   0.1484
## regiongreat_lakes            -1.374e-01  9.881e-02  -1.390   0.1647
## regionplains                 -1.361e-01  9.881e-02  -1.377   0.1687
## regionsoutheast              -1.474e-01  9.867e-02  -1.494   0.1353
## regionsouthwest              -1.728e-01  9.900e-02  -1.746   0.0811 .
## regionrocky_mountains        -1.348e-01  9.985e-02  -1.350   0.1772
## regionfar_west               -1.052e-01  9.891e-02  -1.063   0.2879
## regionoutlying_areas          7.781e-02  1.014e-01   0.768   0.4428
## admission_rate               -8.122e-02  1.437e-02  -5.651 1.91e-08 ***
## faculty_salary_avg            1.929e-05  1.406e-06  13.719  < 2e-16 ***
## family_income                -1.052e-06  9.080e-07  -1.159   0.2466
## sqrt(family_income)           2.258e-03  4.748e-04   4.756 2.17e-06 ***
## log(expenditure_per_student)  5.987e-02  6.971e-03   8.588  < 2e-16 ***
## debt_median_all               5.182e-06  7.431e-07   6.973 4.64e-12 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09797 on 1493 degrees of freedom
## Multiple R-squared:  0.7037, Adjusted R-squared:  0.7007
## F-statistic: 236.4 on 15 and 1493 DF,  p-value: < 2.2e-16
```

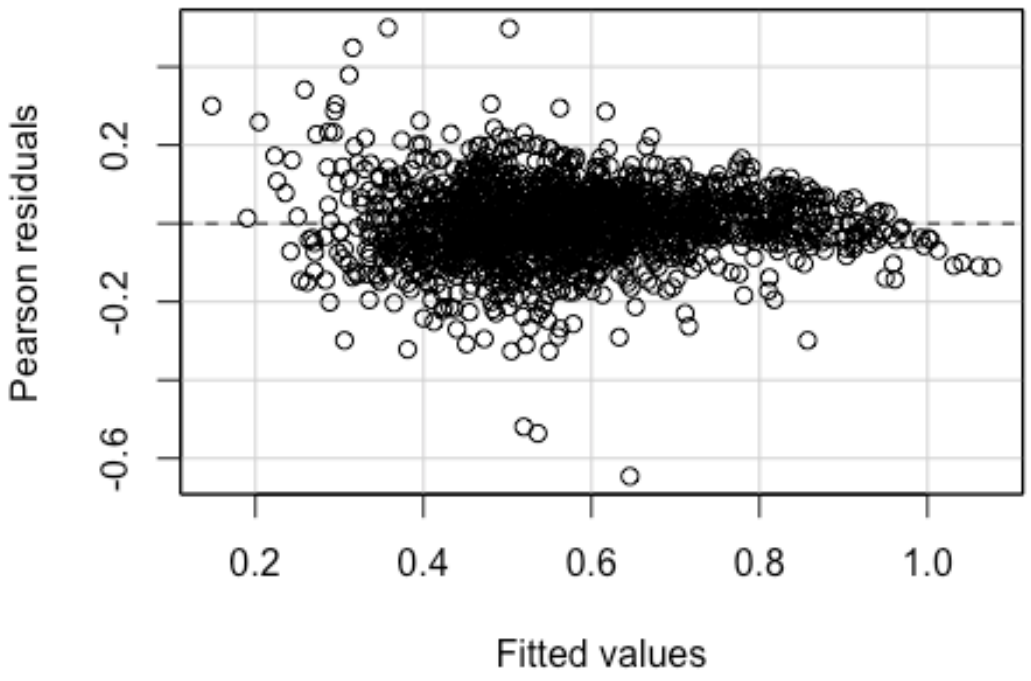
Our final model is (rounded to three significant figures)

$$
\begin{aligned}
\widehat{completion\_rate}\_200 = \; & -0.529 \\
& -0.0165\, D_{new\_england} \\
& -0.0143\, D_{mid\_east} \\
& -0.0137\, D_{great\_lakes} \\
& -0.136\, D_{plains} \\
& -0.147\, D_{southeast} \\
& -0.173\, D_{southwest} \\
& -0.135\, D_{rocky\_mountains} \\
& -0.105\, D_{far\_west} \\
& +7.78\, D_{outlying_areas} \\
& -0.0812\, admission\_rate \\
& +0.0000193\, faculty\_salary\_avg \\
& -0.00000105\, family\_income \\
& +0.00226\, \sqrt{family\_income} \\
& +0.0599\, \log(expenditure\_per\_student) \\
& +0.00000518\, debt\_median\_all\,.
\end{aligned}
$$

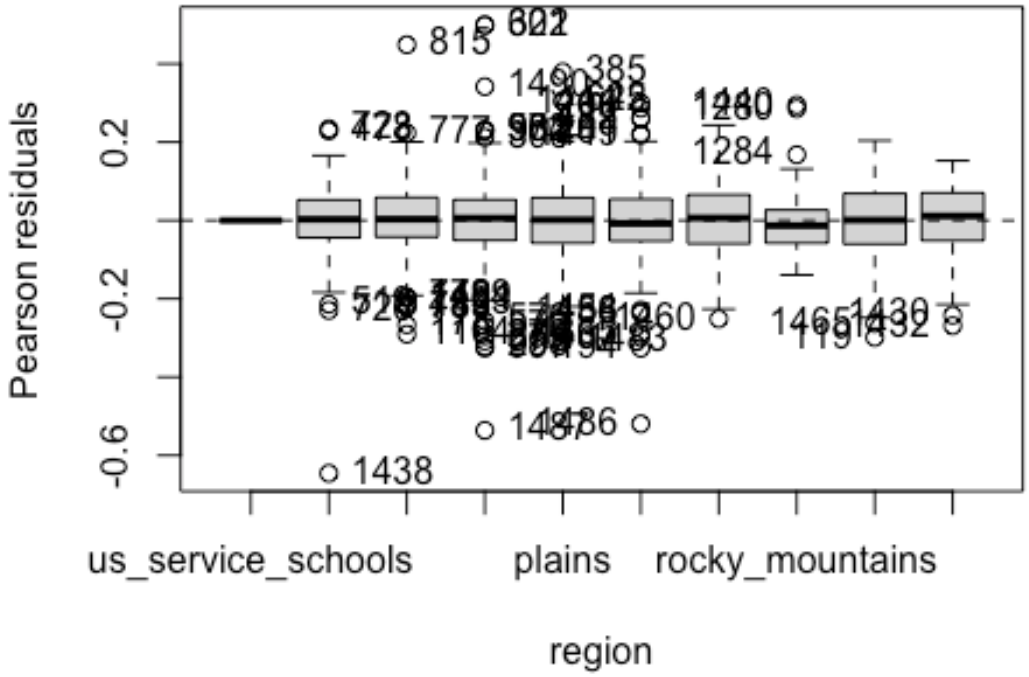
Model evaluation

Check the structure of your model

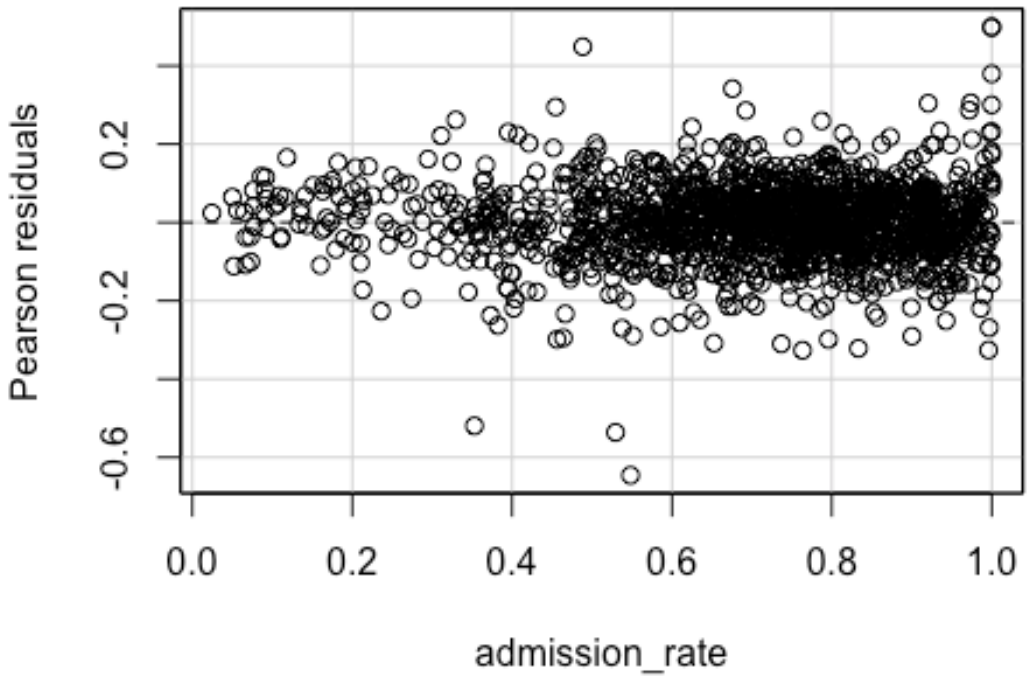
Residual Plot with Fitted Values



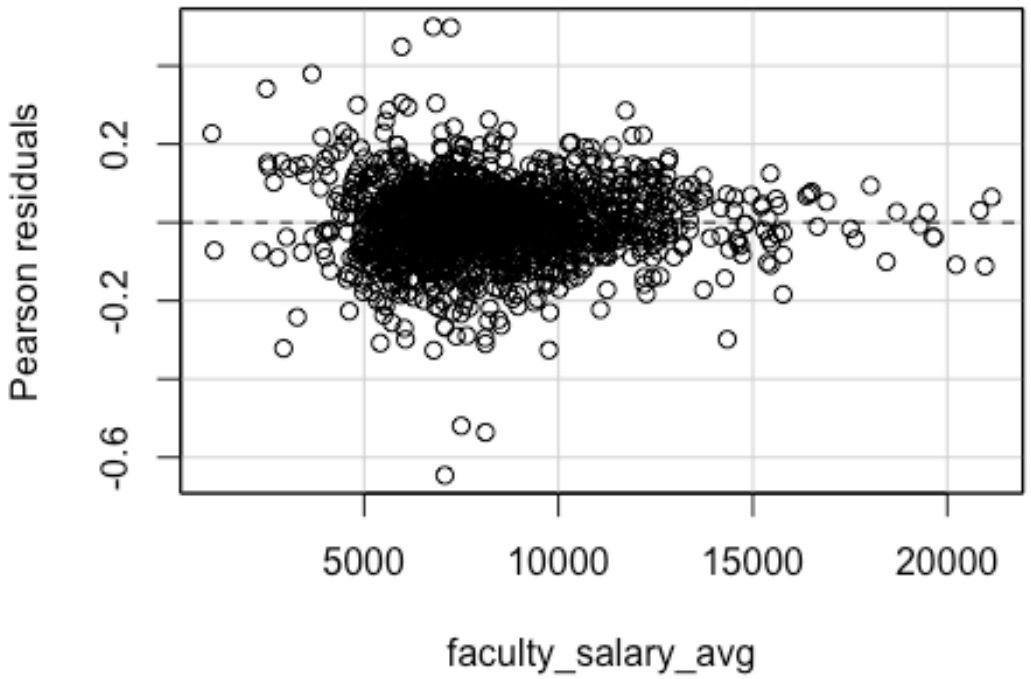
Residual Plot with `region`



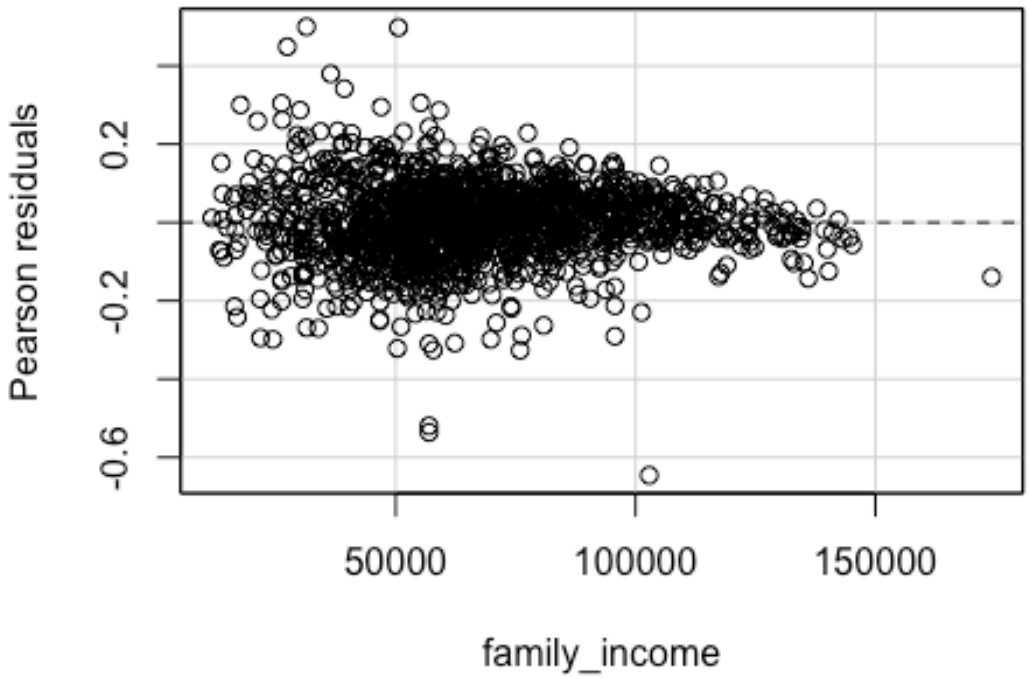
Residual Plot with `admission_rate`



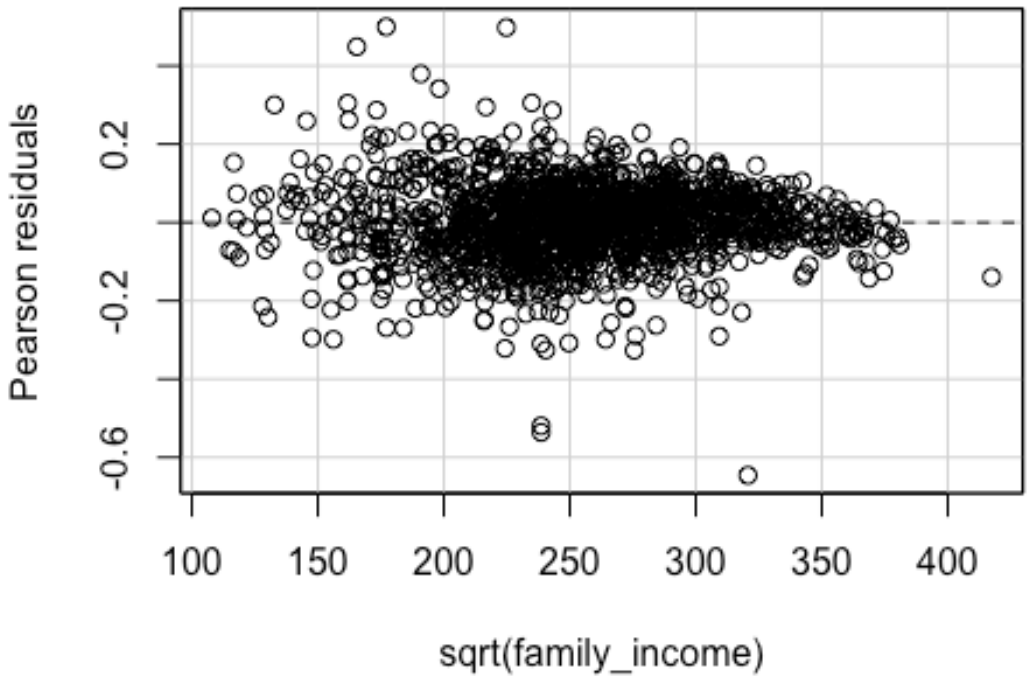
Residual Plot with `faculty_salary_avg`



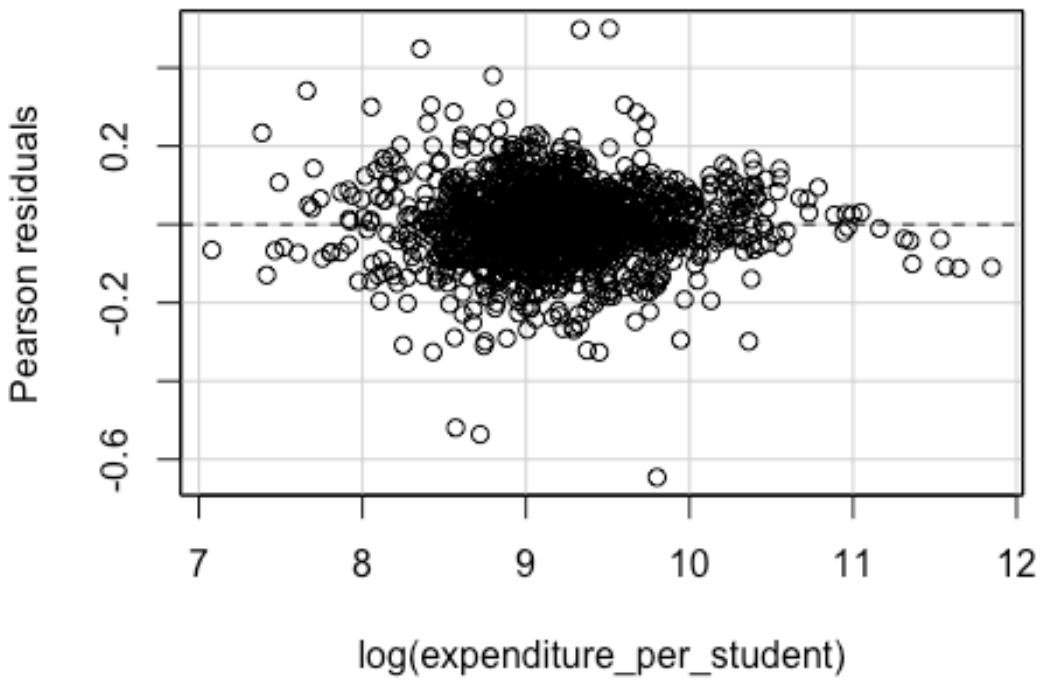
Residual Plot with `family_income`



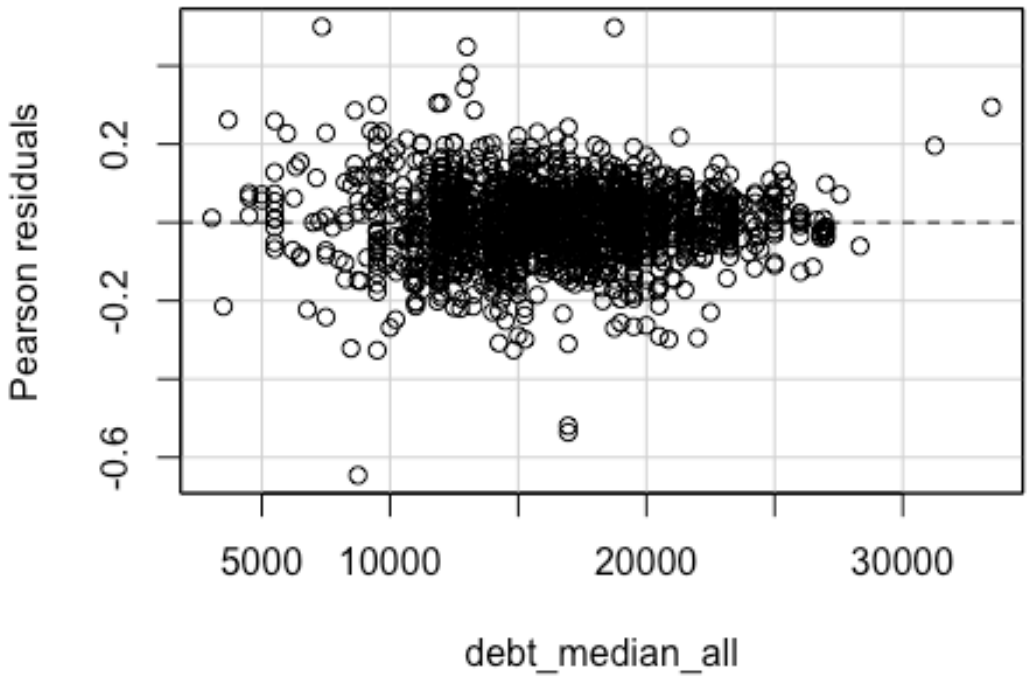
Residual Plot with `sqrt(family_income)`



Residual Plot with `log(expenditure_per_student)`



Residual Plot with `debt_median_all`



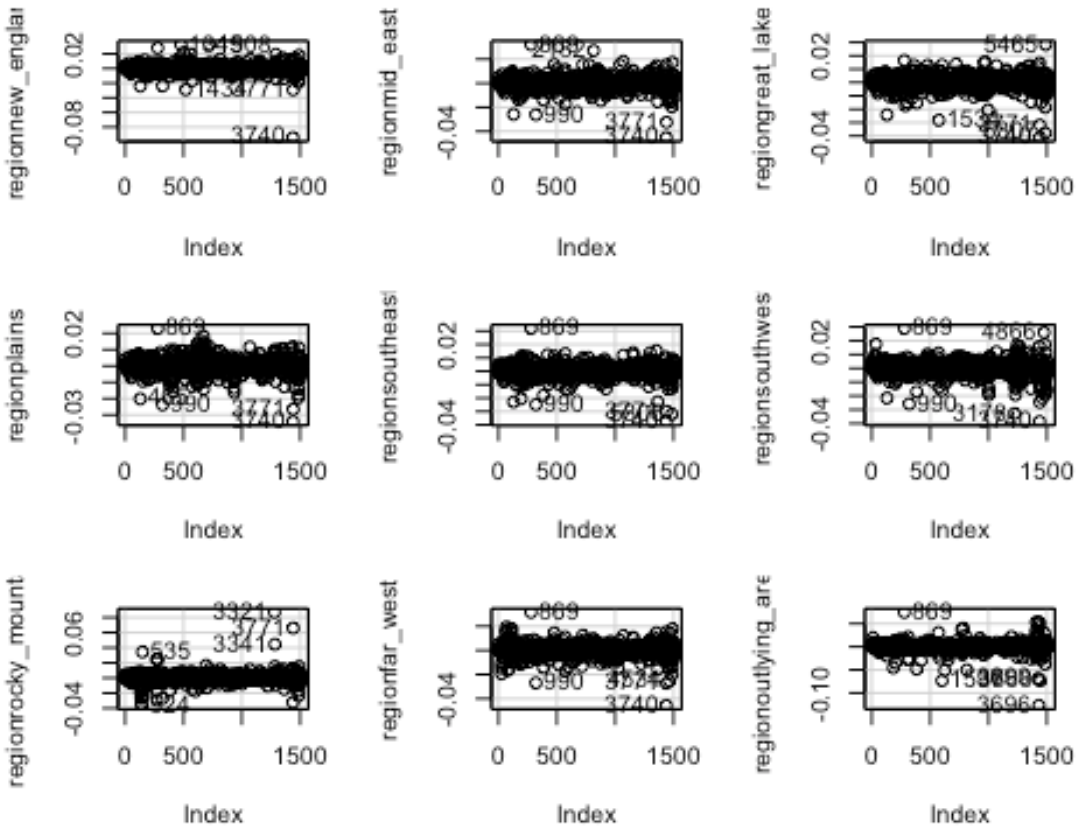
Explain what you did and your conclusions.

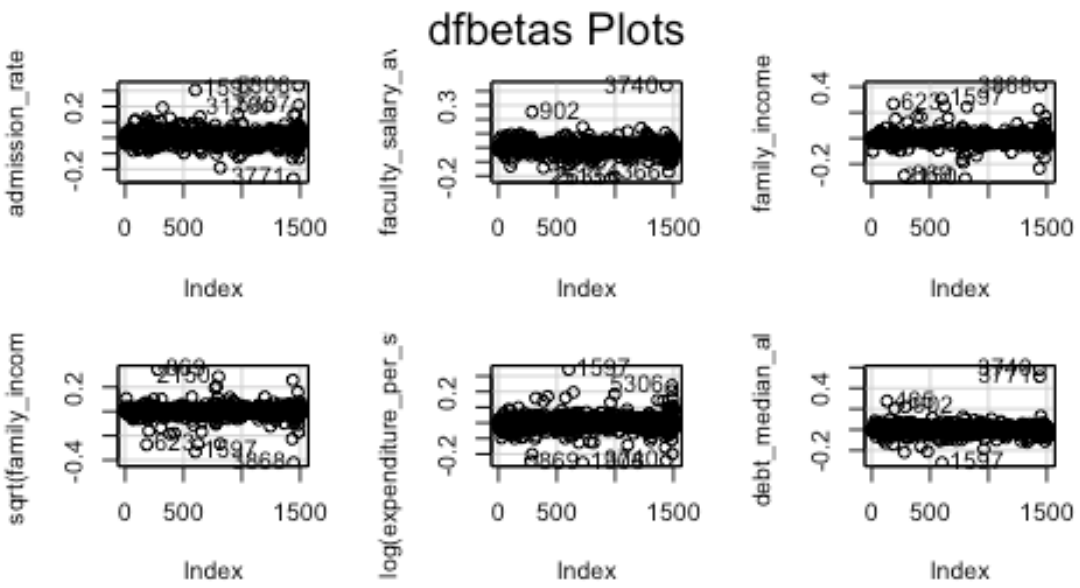
After our variable transformations to variables such as `family_income` and `expenditure_per_student`, we've been able to see that our structural assumptions are fairly satisfied. However, the right-side end of the residual plots for the fitted values and `expenditure_per_student` variable look somewhat asymmetrical and trending by some decreasing.

But overall, since we've observed the symmetry and randomness of each of our residual plots for the fitted values and for each of our model's regressors to be fairly symmetric and random, respectively, it appears that the structural assumption of our model is satisfied.

Check for influential observations

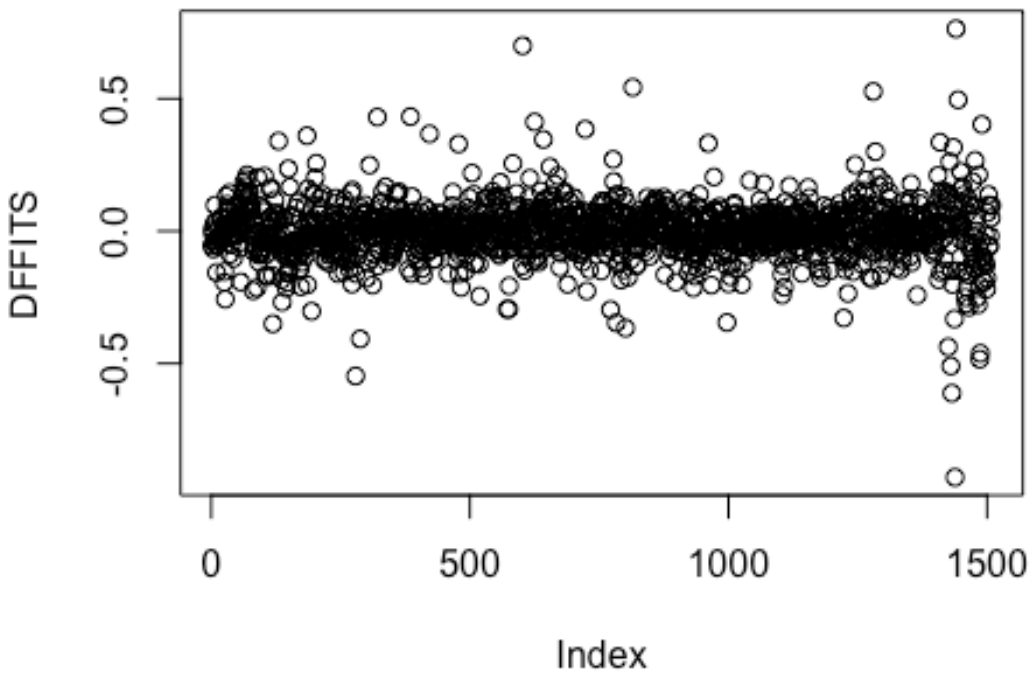
DFBETAS Plot


regionnew_englar
regionmid_east
regiongreat_lake
regionplains
regionsoutheasl
regionsouthwes
regionrocky_mount
regionfar_west
regionoutlying_ar
Index
0
500
1500




```
#inst_data[c(NUM, NUM, NUM), c("region", "admission_rate",
"faculty_salary_avg", "family_income", "expenditure_per_student",
"debt_median_all")]
```

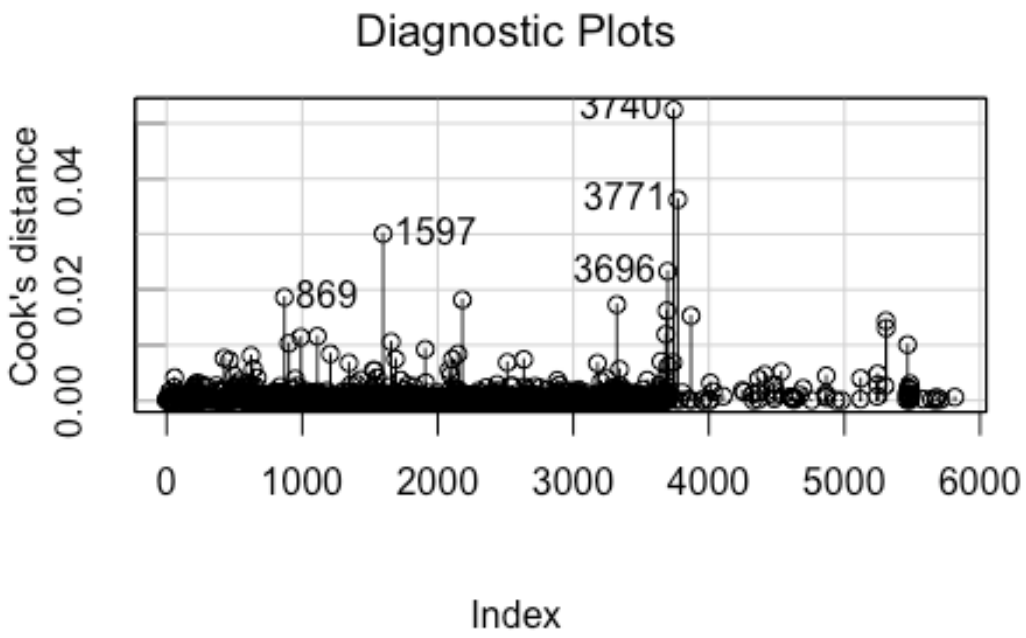
DFFITS Plot



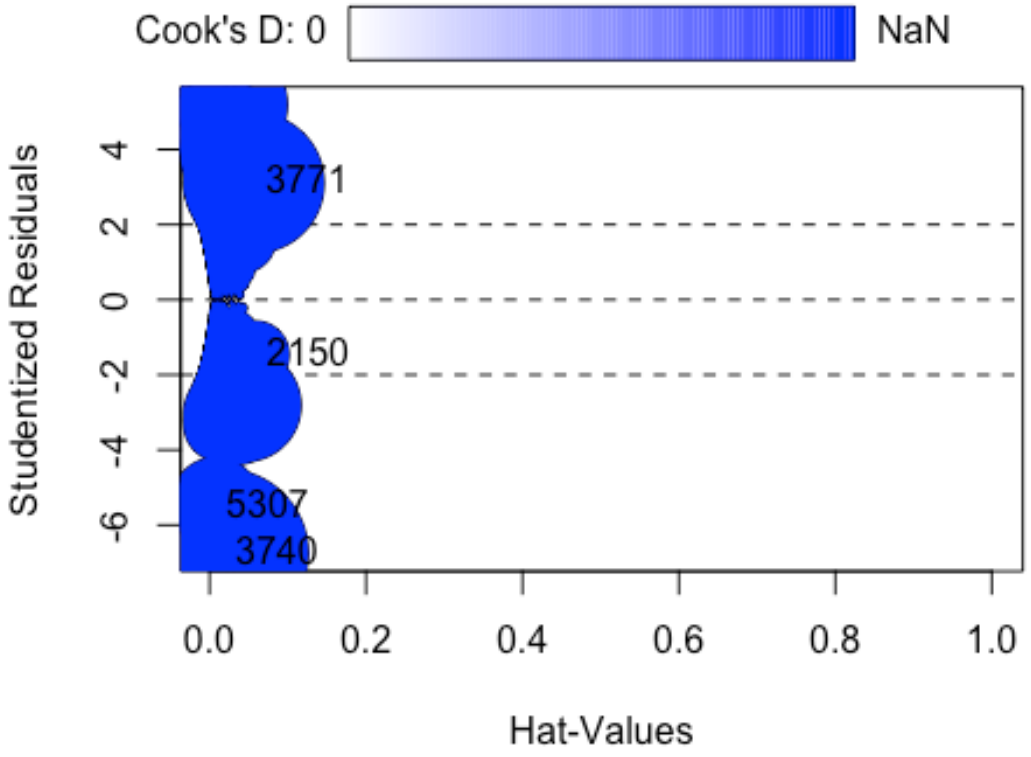

```
which(abs(DFFITS) > 0.7)

## 3740 3771
## 1438 1440
```

Influence Index Plot



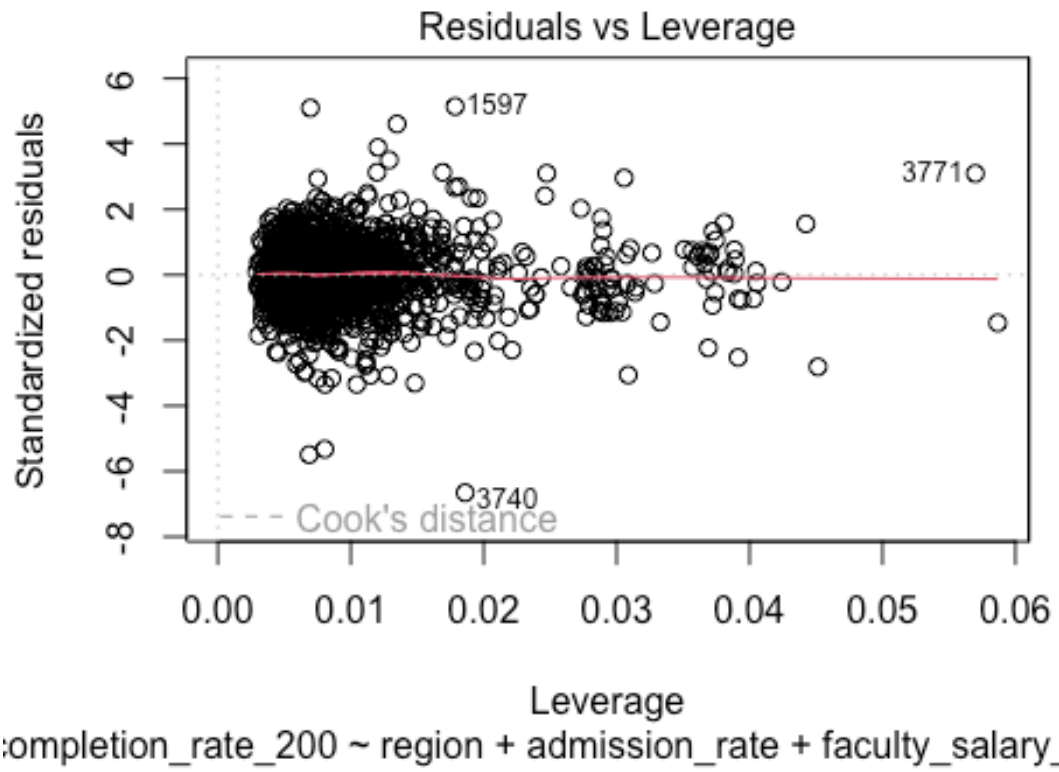
Influence Plot




```
##         StudRes         Hat        CookD
## 2150 -1.467445 0.05866374 0.008380947
## 2294        NaN 1.00000000          NaN
## 3740 -6.752434 0.01860454 0.052455773
## 3771  3.107189 0.05698223 0.036251393
## 5307 -5.545854 0.00687258 0.013042506
```

Standardized Residuals vs Leverage

```
## Warning: not plotting observations with leverage one:
##   871
```



completion_rate_200 ~ region + admission_rate + faculty_salary_avg

```
(inf_obs <- inst_original[c(1597, 3740, 869, 3771, 3696, 1438, 1440), ])
```

```
##                       institution_name   institution_type          region
## 1597      Sacred Heart Major Seminary private_non_profit     great_lakes
## 3740                  Landmark College private_non_profit     new_england
## 869          Chicago State University             public     great_lakes
## 3771             Platt College-Aurora private_for_profit rocky_mountains
## 3696 University of the Virgin Islands             public  outlying_areas
## 1438                  Emerson College private_non_profit     new_england
## 1440                 Endicott College private_non_profit     new_england
##      admission_rate completion_rate_200 faculty_fulltime
faculty_salary_avg
## 1597         1.0000              0.8572           0.3651
6779
## 3740         0.5482              0.0000           1.0000
7079
## 869          0.4642              0.1776           0.9952
8118
## 3771         0.4545              0.8572           0.3191
6133
## 3696         0.9969              0.2938           1.0000
7079
## 1438         0.4105              0.8025           0.3712
```

```
10172
## 1440         0.7040              0.7488           0.2574
10658
##      expenditure_per_student family_income cost_attendance_per_year
## 1597                   13506      31381.48                    36882
## 3740                   18063     102849.91                    77400
## 869                    20865      21800.62                    21867
## 3771                    7172      46954.40                    44393
## 3696                    8171      31398.64                    16381
## 1438                   11139     118339.44                    72119
## 1440                    8908     114013.84                    52642
##      tuition_revenue_per_student debt_median_all
## 1597                       14521            7343
## 3740                       45056            8750
## 869                         7087           22000
## 3771                       13859           33470
## 3696                        7936           10000
## 1438                       30966           21188
## 1440                       17416           21960
```

Explain what you did and your conclusions.

We created and looked at DFBETAS plots, DFFITS plots, influence index plots, influence plots ,and standardized residuals versus leverage plots.

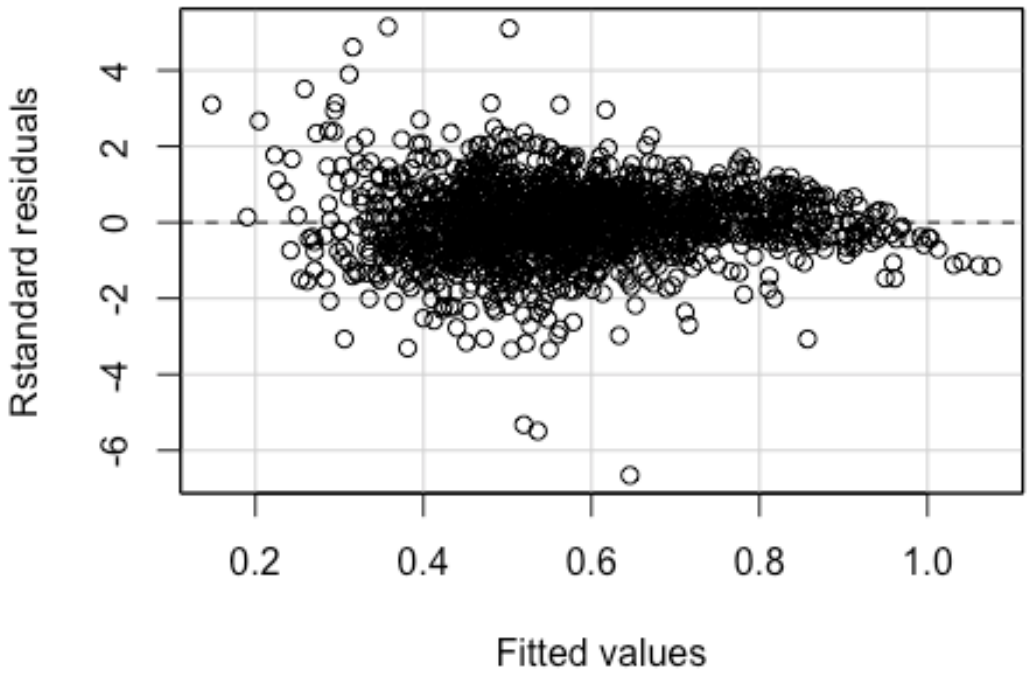
And we narrowed down 7 influential observations that were prevalent throughout most of the plots.

We see that, generally, these 7 influential observations have high admission rates, extreme completion rates, low average faculty salaries, and extreme average family incomes.

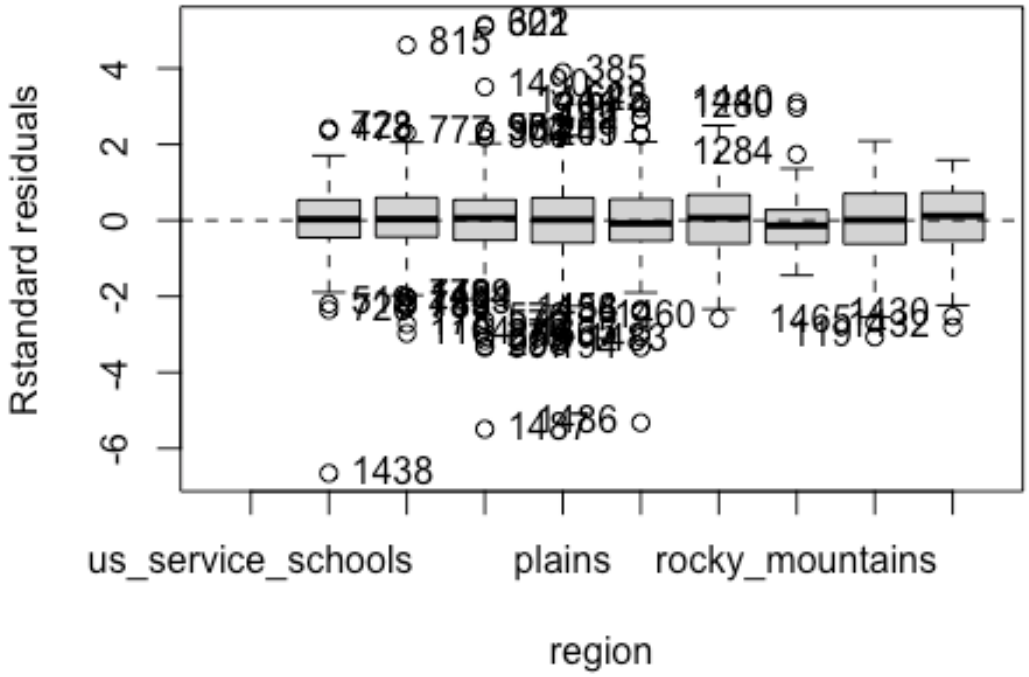
Although damaging to our overall model, we choose to keep these influential observations and similar ones to it, since excluding them has the potential to weaken the real-world explanatory power of our model for very or fairly statistically extreme institutions.

Check error assumptions

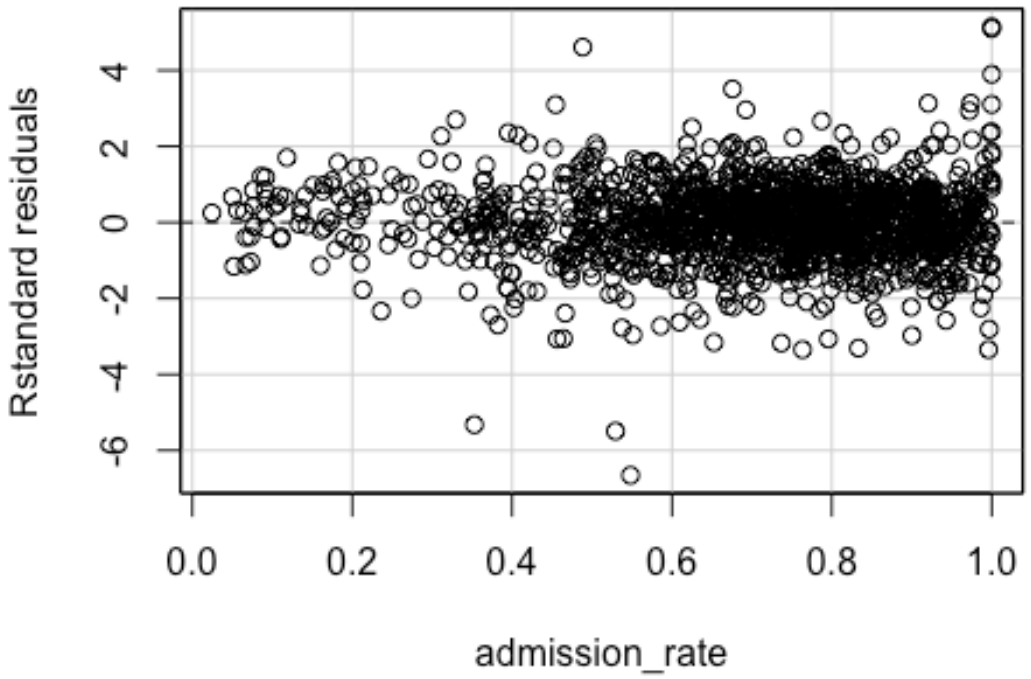
Standardized Residuals with Fitted Values



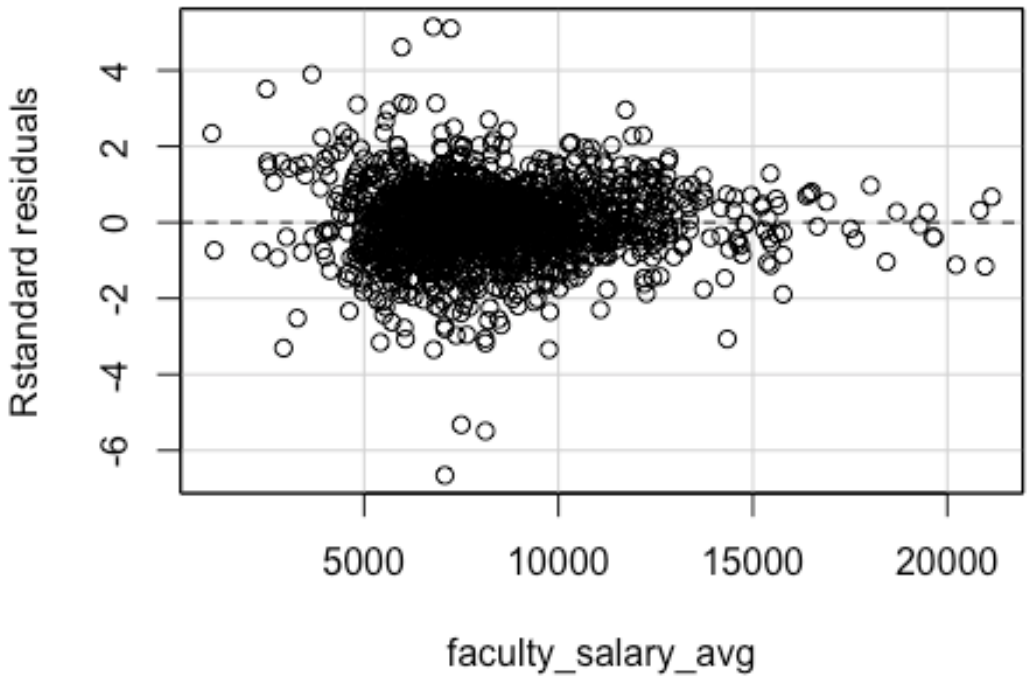
Standardized Residuals with `region`



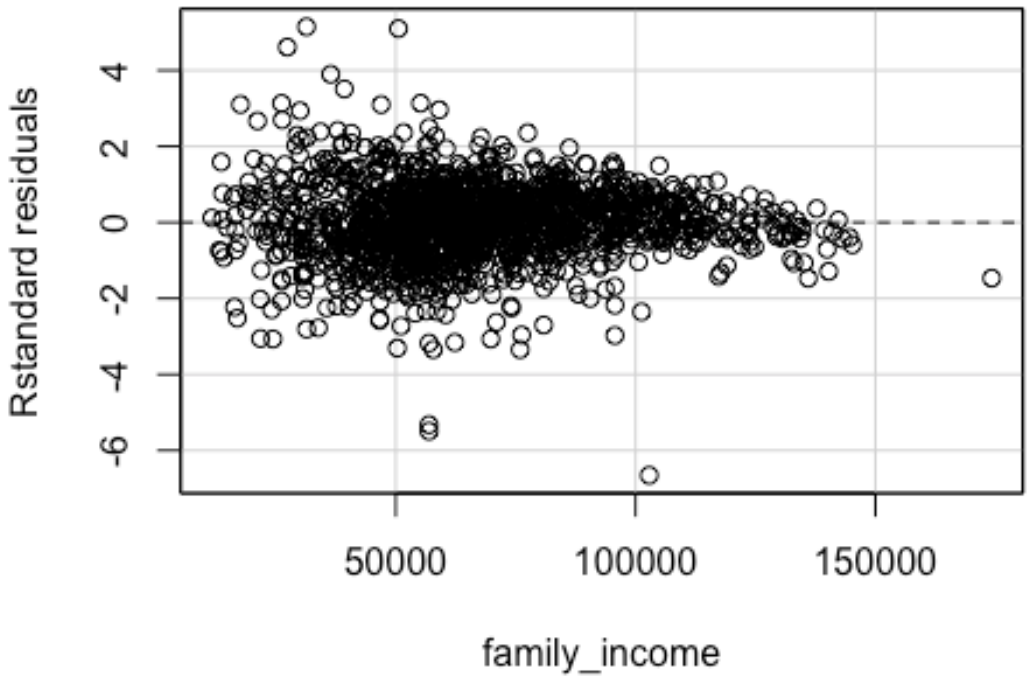
Standardized Residuals with `admission_rate`



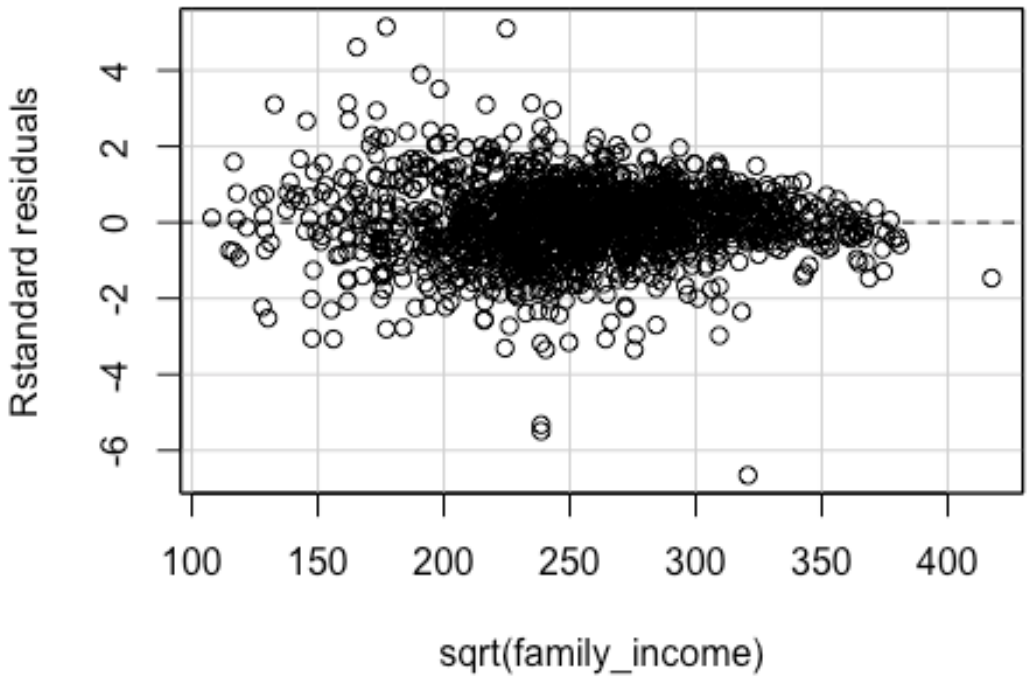
Standardized Residuals with `faculty_salary_avg`



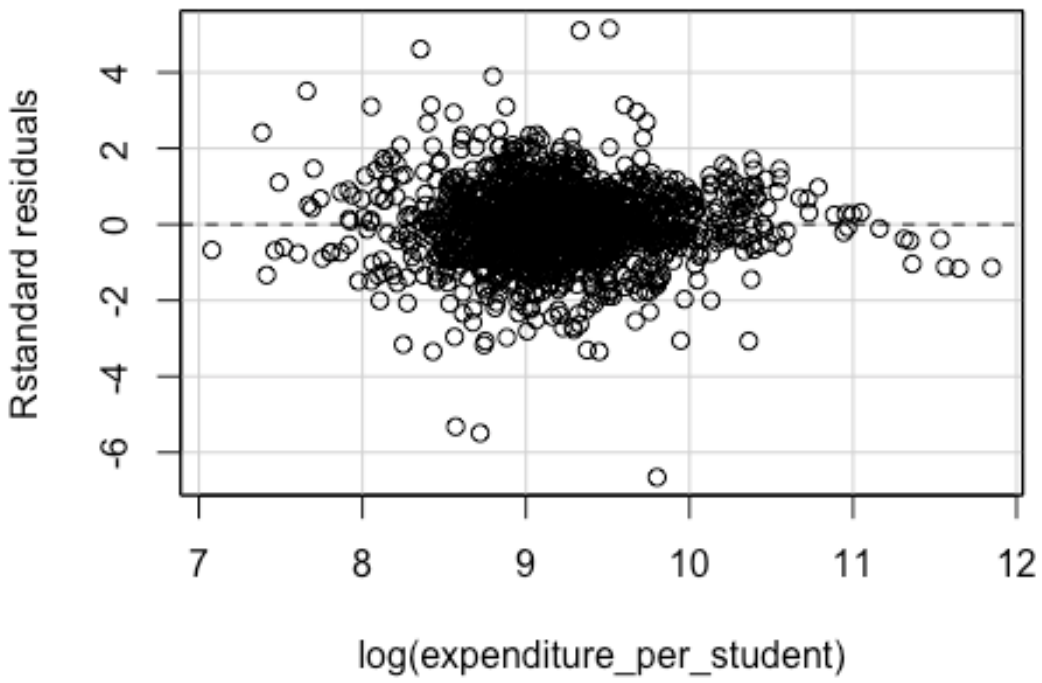
Standardized Residuals with `family_income`



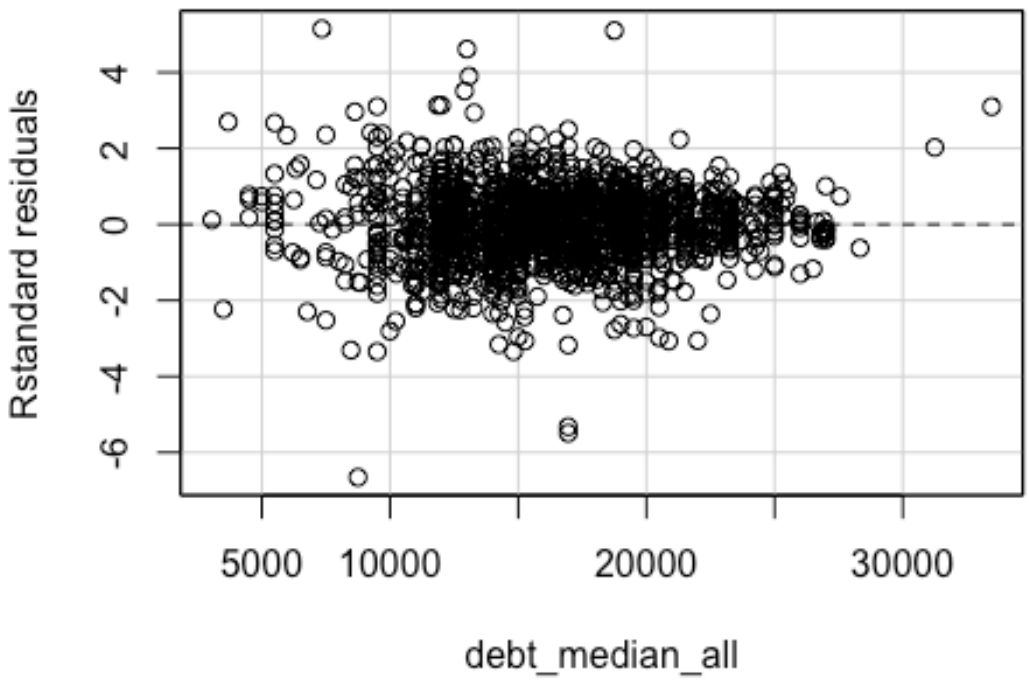
Standardized Residuals with `sqrt(family_income)`



Standardized Residuals with `log(expenditure_per_student)`

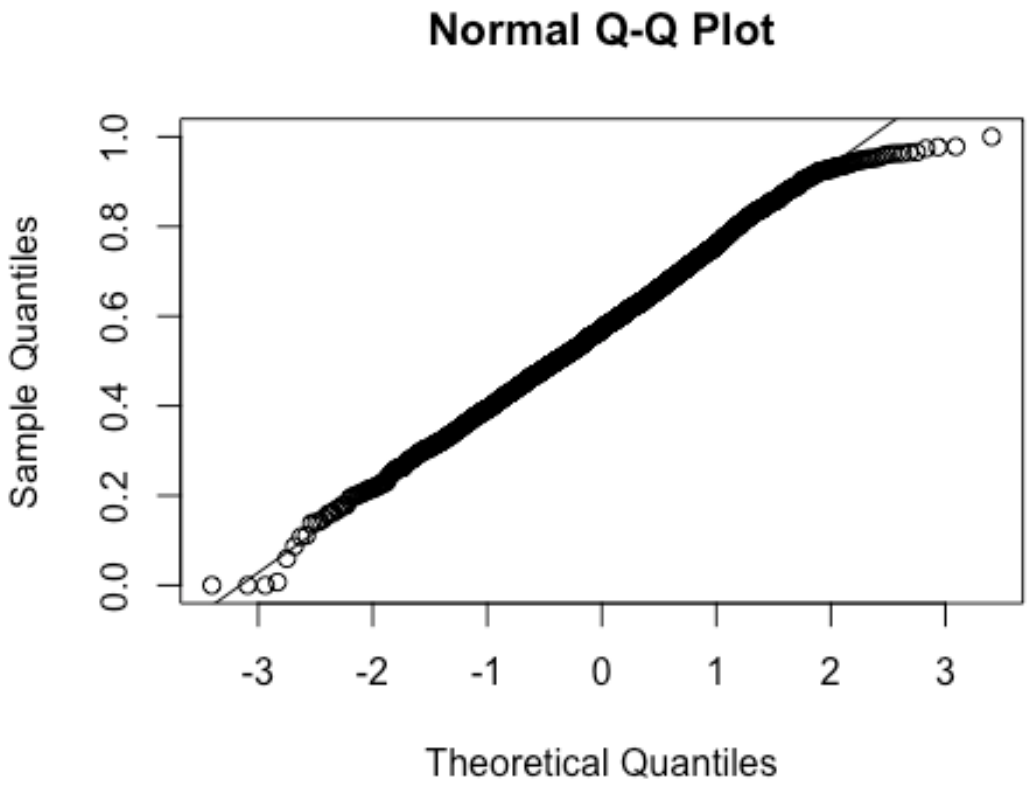


Standardized Residuals with `debt_median_all`



Checking Normality

Q-Q Plot



Shapiro-Wilk Test

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(lmod_final)
## W = 0.96492, p-value < 2.2e-16
```

Explain what you did and your conclusions.

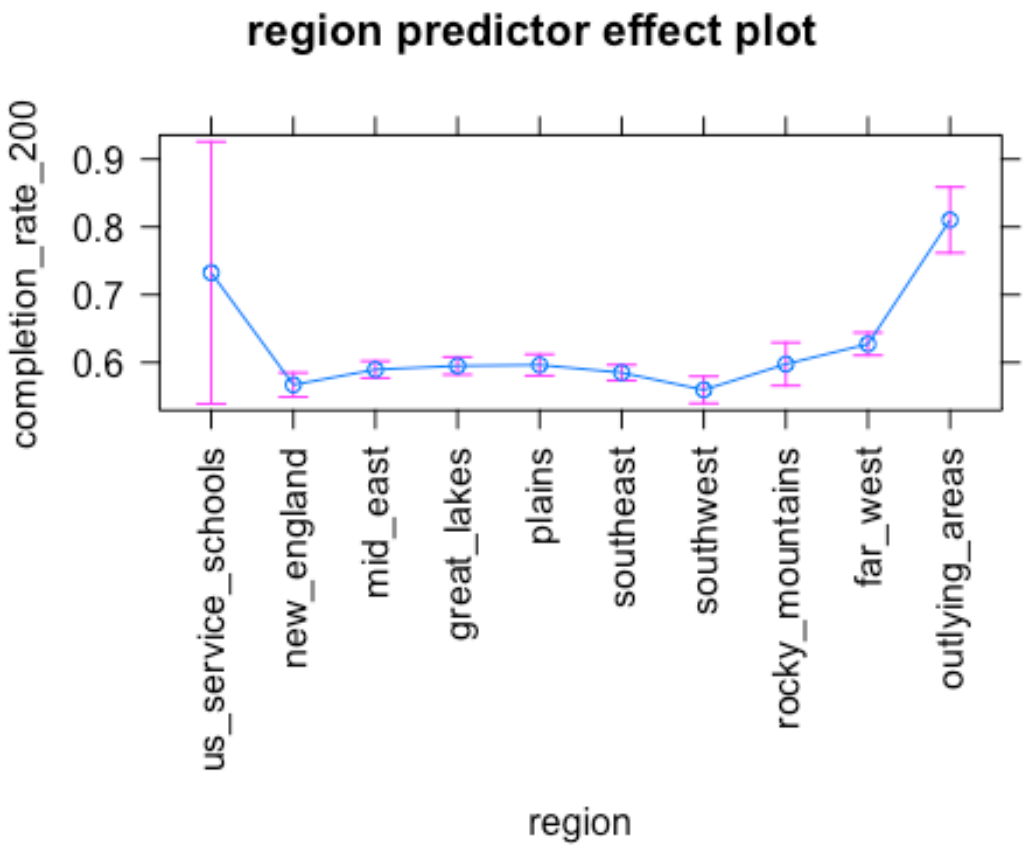
We created and analyzed standardized residual plots, for the fitted values and for regressors in our model. Unfortunately, this revealed that only the `region`, `admission_rate`, and `debt_median_all` regressors satisfy the constant-error variance assumption. The scatter plot for the fitted values revealed that the standardized residuals decrease as the fitted values increase. This same trend, interestingly, is shared by all the regressors that violate the constant-error variance assumption.

We also employed the Shapiro-Wilk normality test after observing potentially promising results with the Q-Q plot, which unfortunately revealed strong evidence that our residuals, and thus our errors, are not normally distributed.

Overall, this reveals that our model is, unfortunately, unsatisfactory in regards to the constant-variance and normality assumptions about our errors.

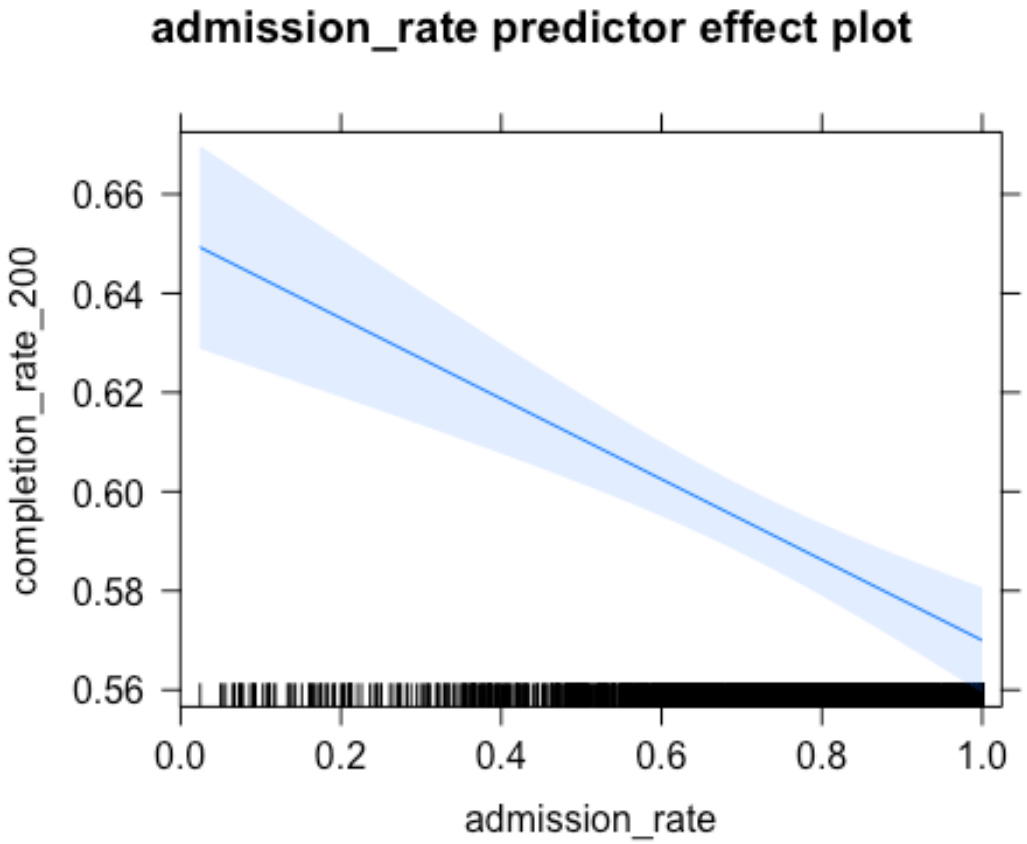
Interpretation

`region` Effect Plot



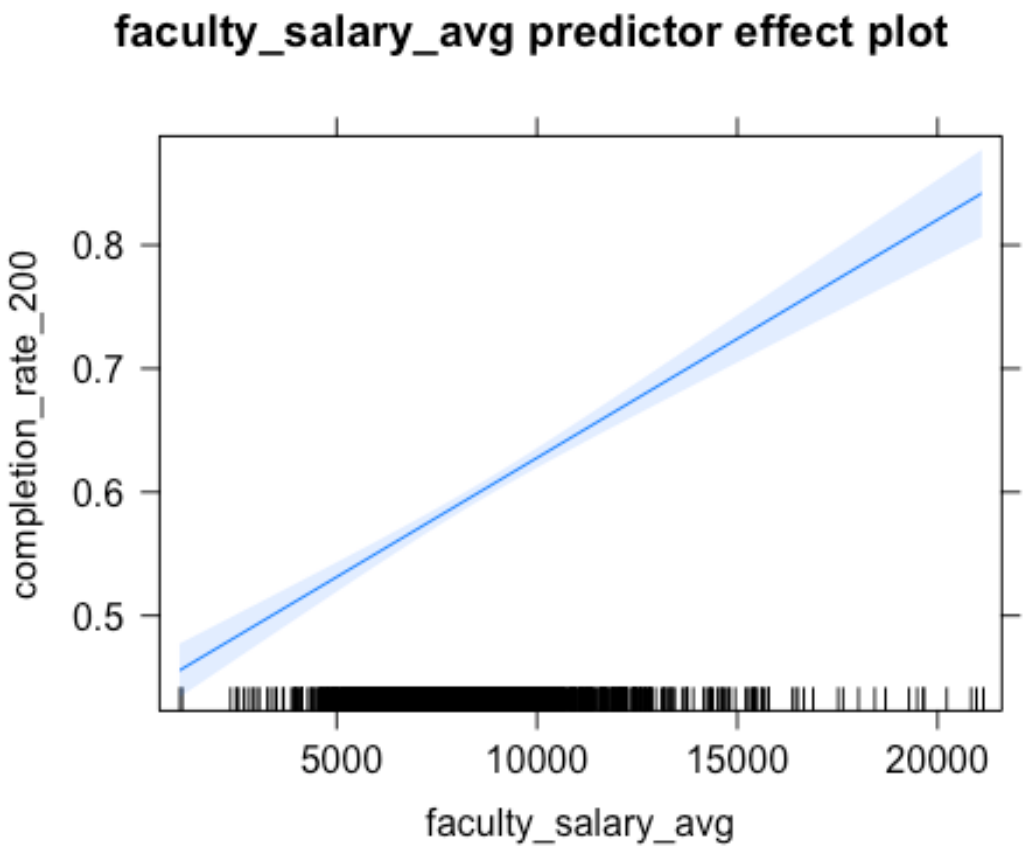
This plot demonstrates the different effects that regions have on admission rates. The `us_service_schools` region tends to have higher completion rates, but the confidence interval is so large that it cannot be considered significant, whereas the `outlying_areas` region also has higher completion rates than the other regions, but its confidence interval is reasonably small enough to consider it as significant. Each of the other regions have lower completion rates and small enough confidence intervals that they are significant.

`admission_rate` Effect Plot



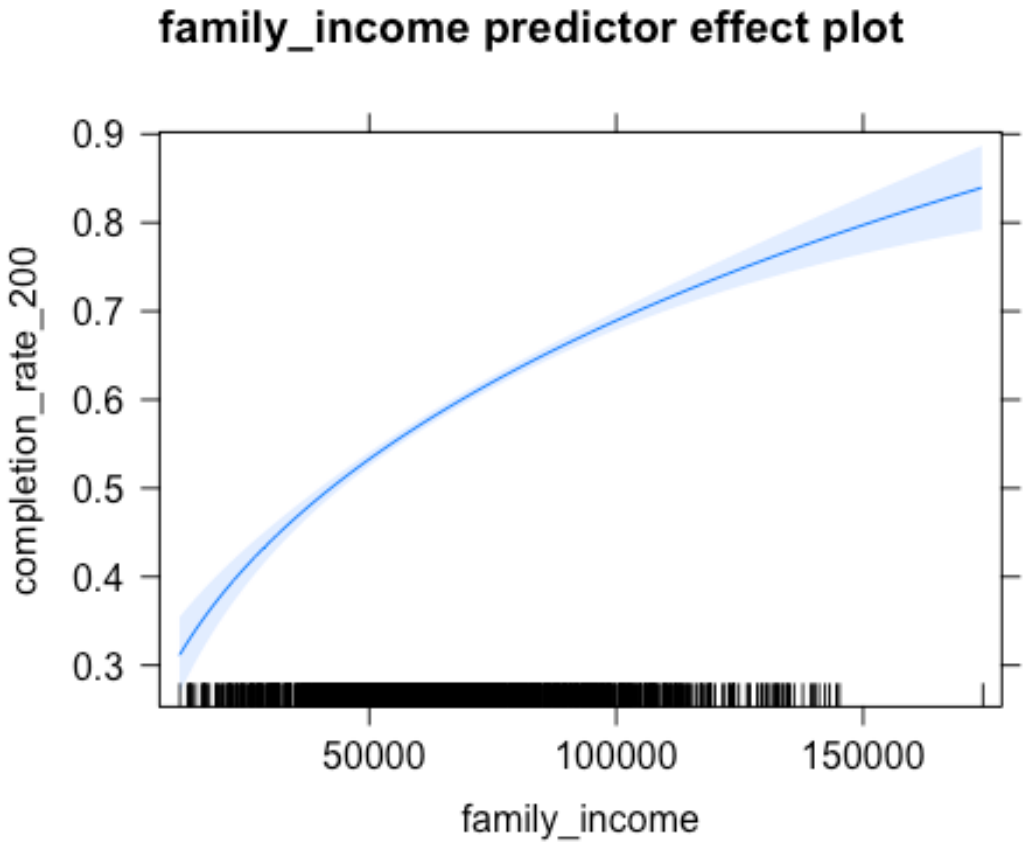
This plot shows that admission rates have a negative linear relationship with completion rates, meaning that higher admission rates have a negative impact on college completion rates. Since the associated p-value for `admission_rate` is 1.91e-08, it is statistically significant.

`faculty_salary_avg` Effect Plot



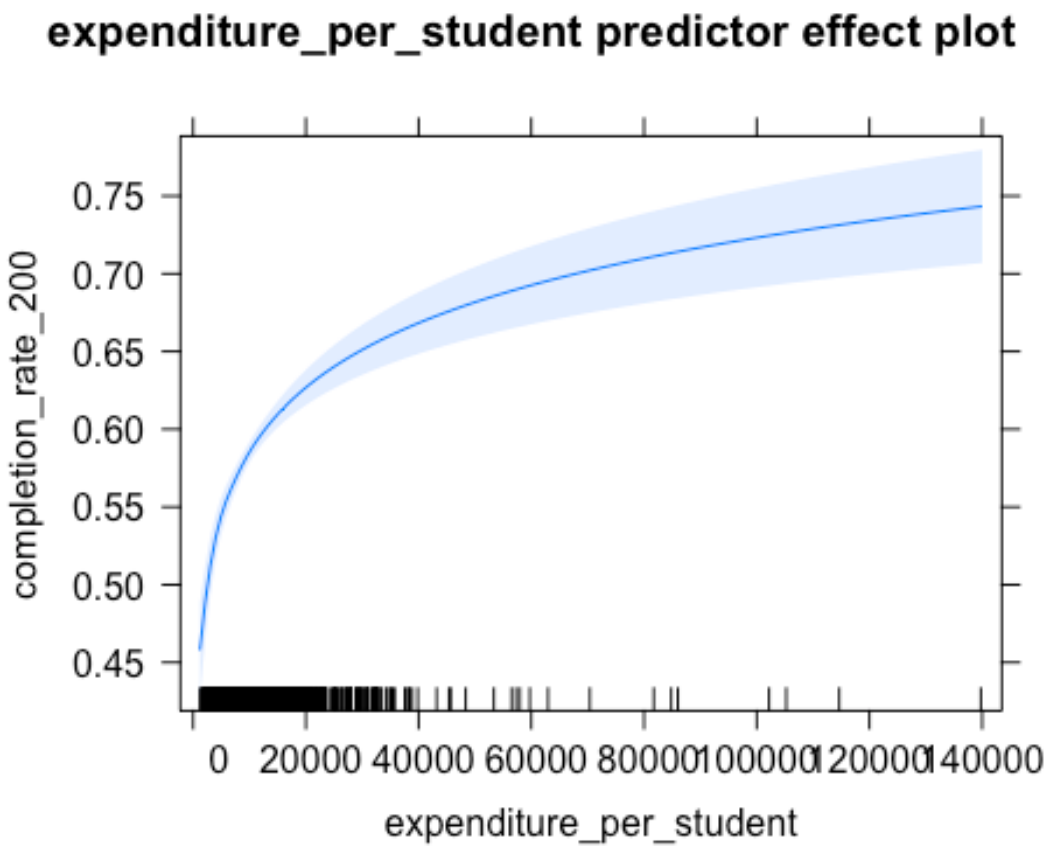
This plot shows that the average faculty salary has a positive linear relationship with completion rates, meaning that higher average faculty salaries positively impact college completion rates. Since the associated p-value for `faculty_salary_avg` is 2e-16 and its confidence interval is narrow, it is very significant.

`family_income` Effect Plot



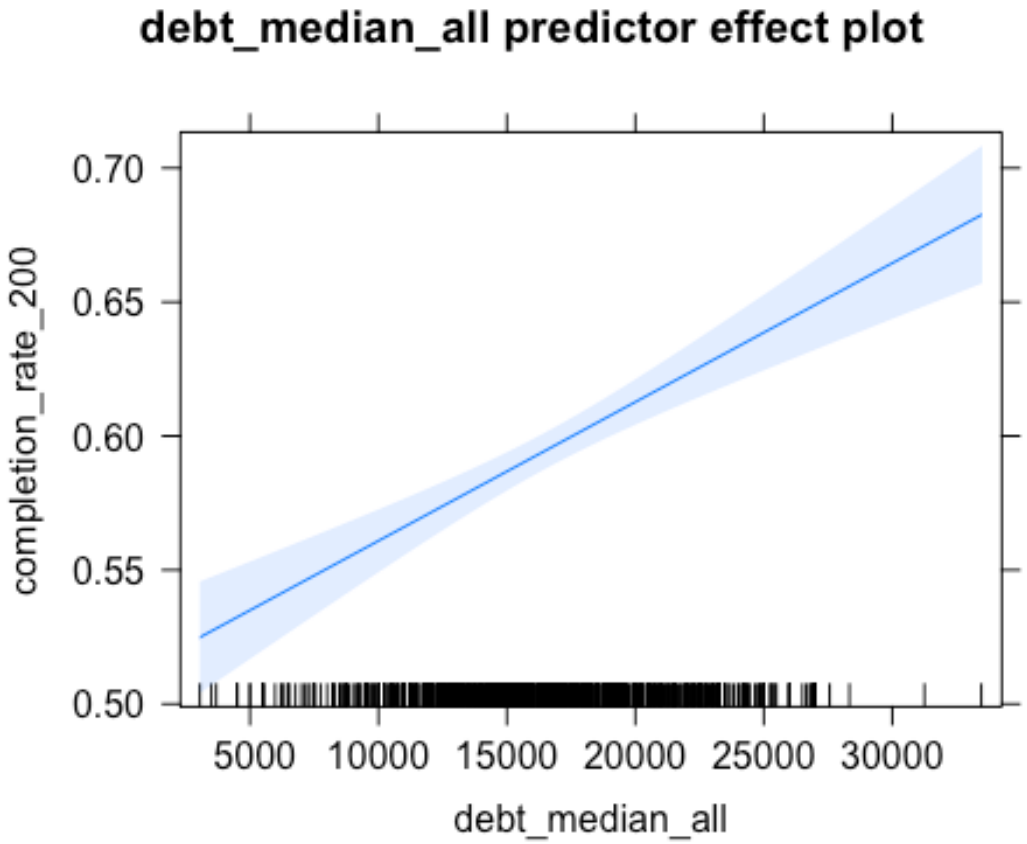
This plot shows that a student's family income has a positive, but non-linear relationship with completion rates, meaning that higher family incomes positively impact college completion rates. Since the relationship is non-linear and as determined by the different slopes, it appears that family incomes below $50,000 have a more variable impact than higher family incomes. Since the associated p-value for `family_income` is 0.2466, it is not as significant as the others.

`expenditure_per_student` Effect Plot



This plot shows that a college's expenditures per student has a positive, but non-linear relationship with completion rates, meaning that higher expenditures positively impact college completion rates. Since the relationship is non-linear and as determined by the changing slope, it appears that expenditures below $20,000 have a much more variable impact than higher expenditures. Since the associated p-value for `expenditure_per_student` is 2e-16, it is significant.

`debt_median_all` Effect Plot



This plot shows that student debt has a positive linear relationship with completion rates, meaning that higher debts seem to positively impact college completion rates. Since the associated p-value for `debt_median_all` is 4.64e-12, it is significant.

Conclusions

Summary of results

The overall relationship between the response variable and the predictors suggests that colleges tend to have higher rates of completion when they have lower admission rates, maintain higher faculty salaries, have students from higher income families, spend more money on students, and have students with higher debts.

This relationship makes sense because lower admission rates could mean that a college chooses students with better academic records, higher faculty salaries means they attract better quality professors, and spending more on students could mean that there are more resources for student success. These all logically seem like they would be associated with higher completion rates. Having students from higher income families and students with higher debts also make sense being associated with higher completion rates because

colleges that are expensive are better able to pay large faculty salaries and have higher expenses per student, and expensive colleges will most likely have either students from families with higher incomes or larger debts.

Public policy recommendations

In light of our final model, we recommend that colleges pay higher salaries to their faculty in order to attract better quality professors and that they maintain high expenses per student so that they may provide students with the resources to help them succeed and ultimately graduate.

We do not want to recommend that colleges have lower admission rates or choose students from families with higher incomes or with higher debts, but realize that these may be a secondary result from our recommendations to pay higher salaries and spend more on students.

A further exploration on whether or not lower admission rates, higher student family incomes, and higher student debts impact completion rates on their own or whether they impact completion rates through association with other factors.

Improvements

Since nearly 70% of our response values contained NA's, our data was limited and may have had some bias because our model was based on only those colleges that had all data points available. A potential improvement would be to remedy this situation by completing a more thorough study based on these factors that has available data for each institution.

Additionally, because our model did not satisfy the constant-variance and normality assumptions it would be worthwhile to consider different variables transformations and/or data sources in the future.