STA302 - VIDEO PROJECT

Admission Rates of U.S. Colleges and Universities

Understanding what factors affect admission rates significantly by means of statistical analysis

Contributors

EDA | Predictor Selection | Model Diagnostics | Interpretation

1005684127

Md Abrar Nasir

1006666644

Sumaita Imam Anika

Our Data

250 Observations of 30 Variables

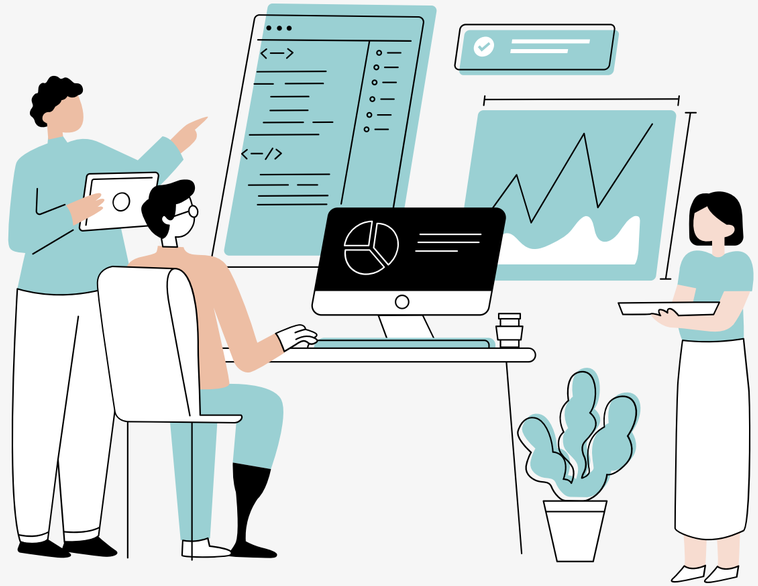
- **Institution Specific Predictors**
  10 Variables
- **Student Specific Predictors**
  15 Variables
- **School Location Identifiers**
  2 Variables

Admission Rate

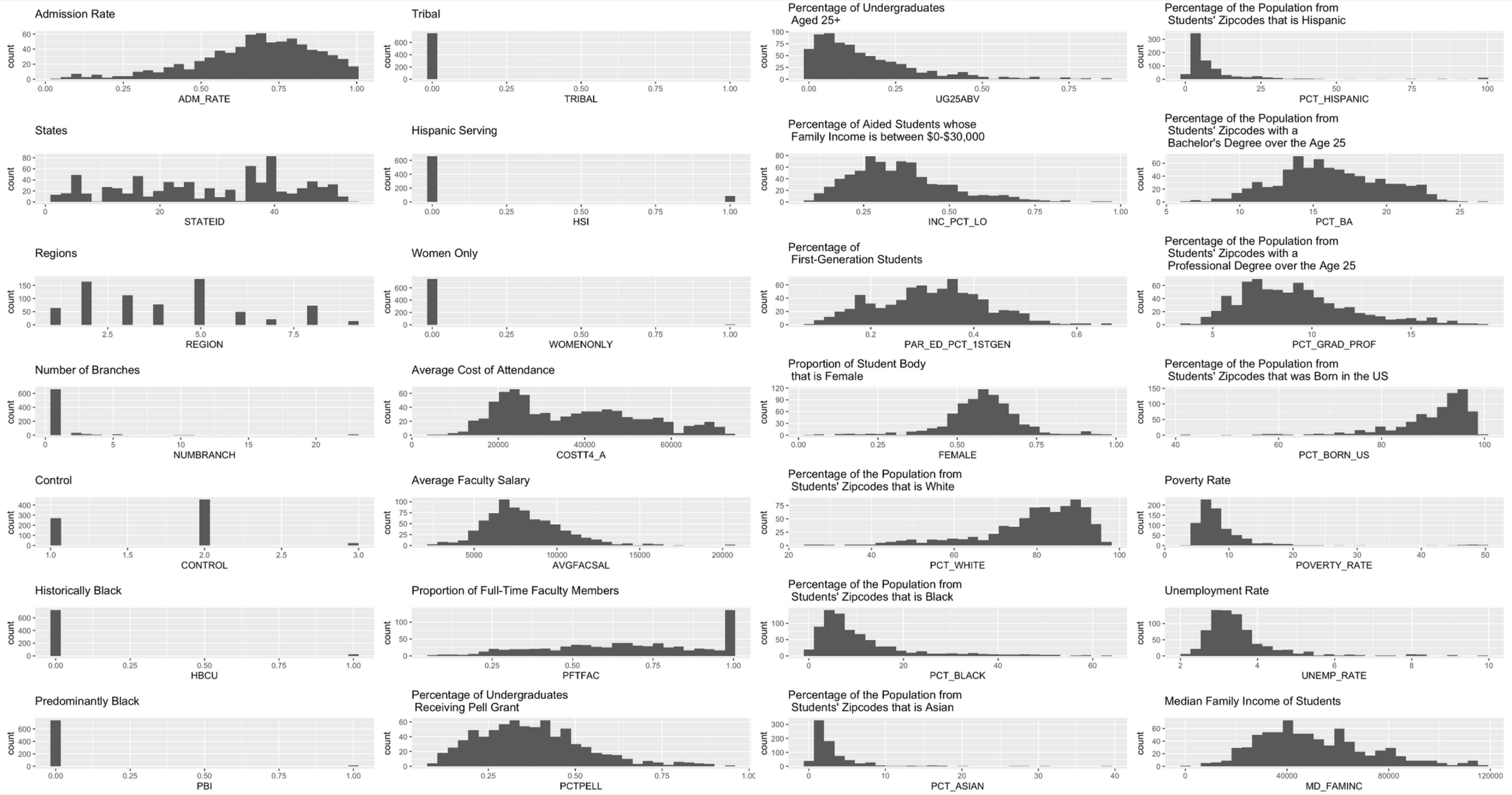
Response Variable

Data Dictionary

ADM_RATE: Admission Rate
STATEID: State Identifier
NUMBRANCH: Number of Branches
CONTROL: Control Identifier
REGION: Region Identifier
HBCU: Historically Black-Serving Institute Identifier
PBI: Predominantly Black-Serving Institute Identifier
TRIBAL: Tribal-Serving Institute Identifier
HSI: Hispanic-Serving Institute Identifier
WOMENONLY: Women-Only Institute Identifier
COSTT4_A: Average Cost of Attendance
AVGFACSAL: Average Faculty Salary
PFTFAC: Proportion of FT Faculty
PCTPELL: Percentage of UGs Receiving Pell Grant
UG25ABV: Percenatge of UGs Aged 25 and Above
INC_PCT_LO: Percentage of Aided Students whose Family Income is between $0-$30,000
PAR_ED_PCT_1STGEN: Percentage of First-Generation Students
FEMALE: Proportion of Female Student Body
MD_FAMINC: Median Family Income of Students
PCT_WHITE: Percentage of White Population in Students' Neighbourhood
PCT_BLACK: Percentage of Black Population in Students' Neighbourhood
PCT_ASIAN: Percentage of Asian Population in Students' Neighbourhood
PCT_HISPANIC: Percentage of Hispanic Population in Students' Neighbourhood
PCT_BA: Percentage of Population Aged 25+ with a Bachelor's Degree in Students' Neighbourhood
PCT_GRAD_PROF: Percentage of Population Aged 25+ with a Professional Degree in Students' Neighbourhood
PCT_BORN_US: Percentage of U.S. Born Population in Students' Neighbourhood
POVERTY_RATE: Poverty Rate in Students' Neighbourhood
UNEMP_RATE: Unemployment Rate in Students' Neighbourhood



Histogram of Variables



Observations

From the Histograms

- Histogram of the response variable is negative skewed.
- Predictor variables are also either skewed or bimodal.
- Presence of extreme observations

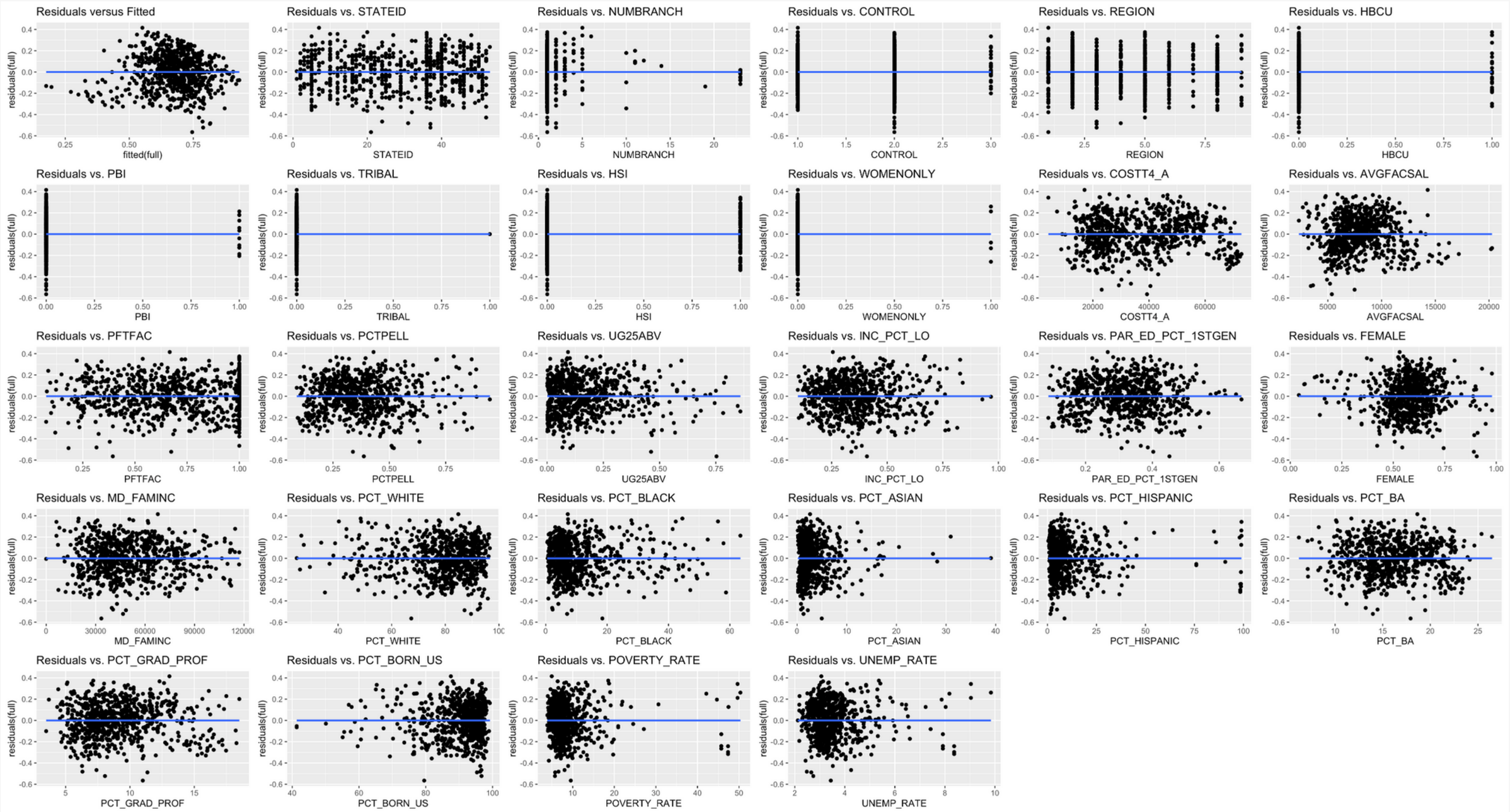
Skews in predictors and the response variable suggest possible violations of either or both linearity and normality assumptions.

Model Assumptions

Cross Examination with All Available Predictors

- Residual versus Response Plot
- Residual versus Predictor Plots
- Normal Q-Q Plot

Residual Plots



Observations

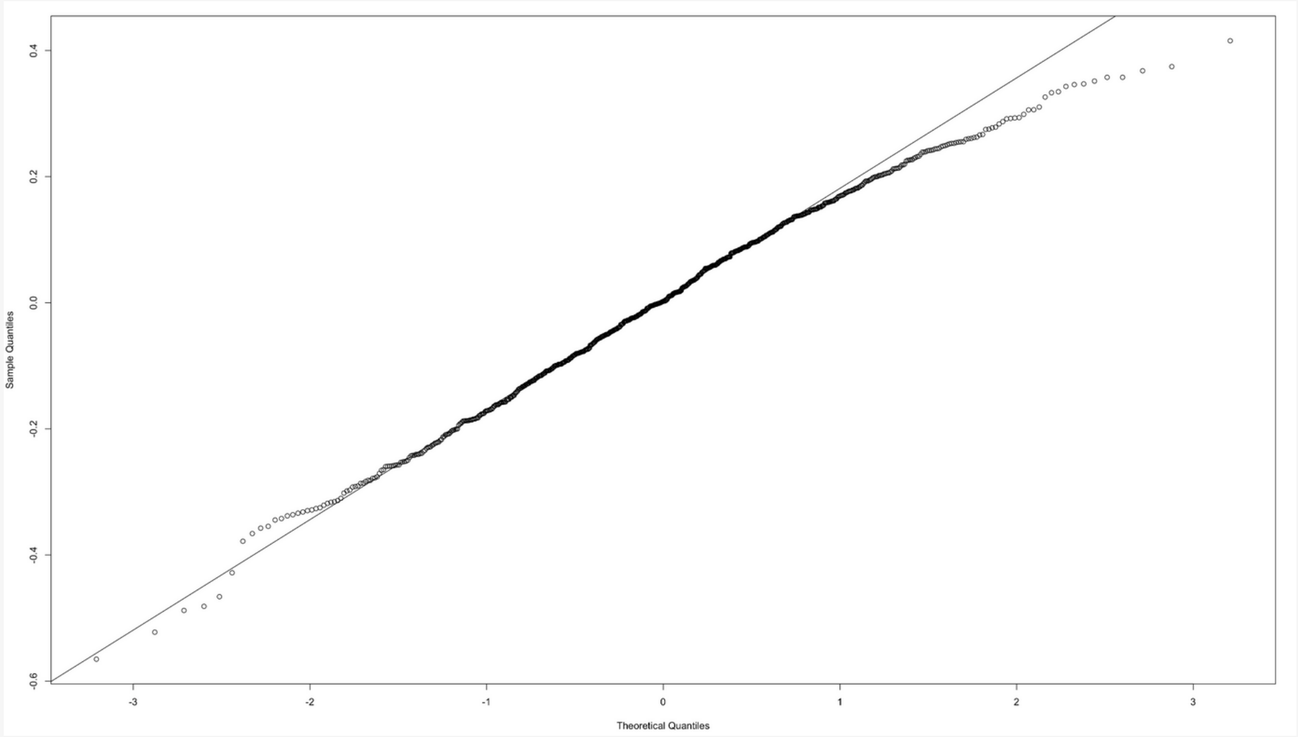
From the Residual Plots

- No fanning pattern is observed.
- No large cluster of residuals with obvious separation from the rest
- Slightly curved (systematic) pattern for the residual-fitted plot

Uncorrelated errors and constant variance assumptions are satisfied while linearity assumption is violated.

Normality Assumption

Violation Check Using Normal Q-Q Plot



Large deviations observed at the ends of the Q-Q plot

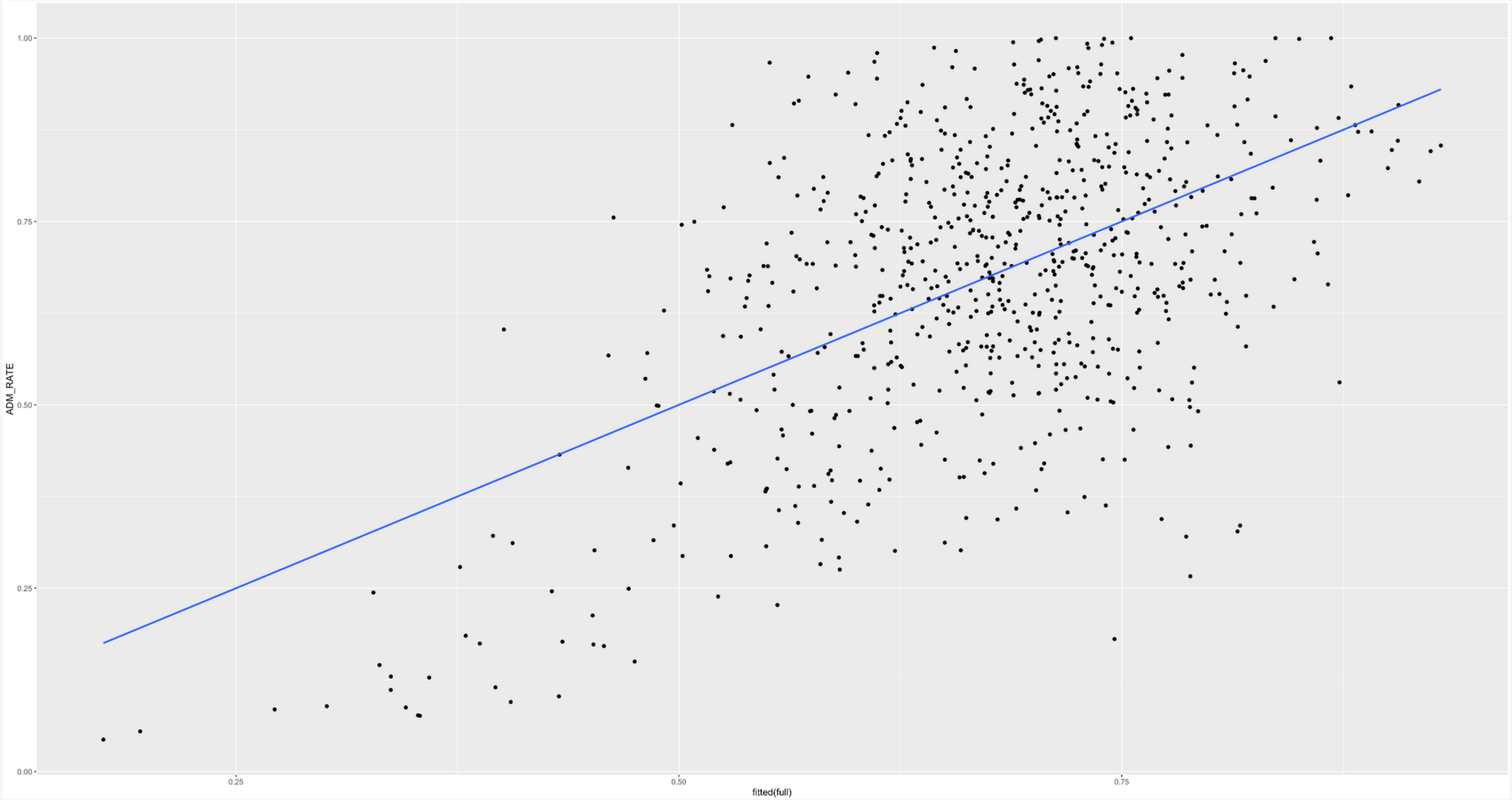
Conditions

To Complement Assumption Checks

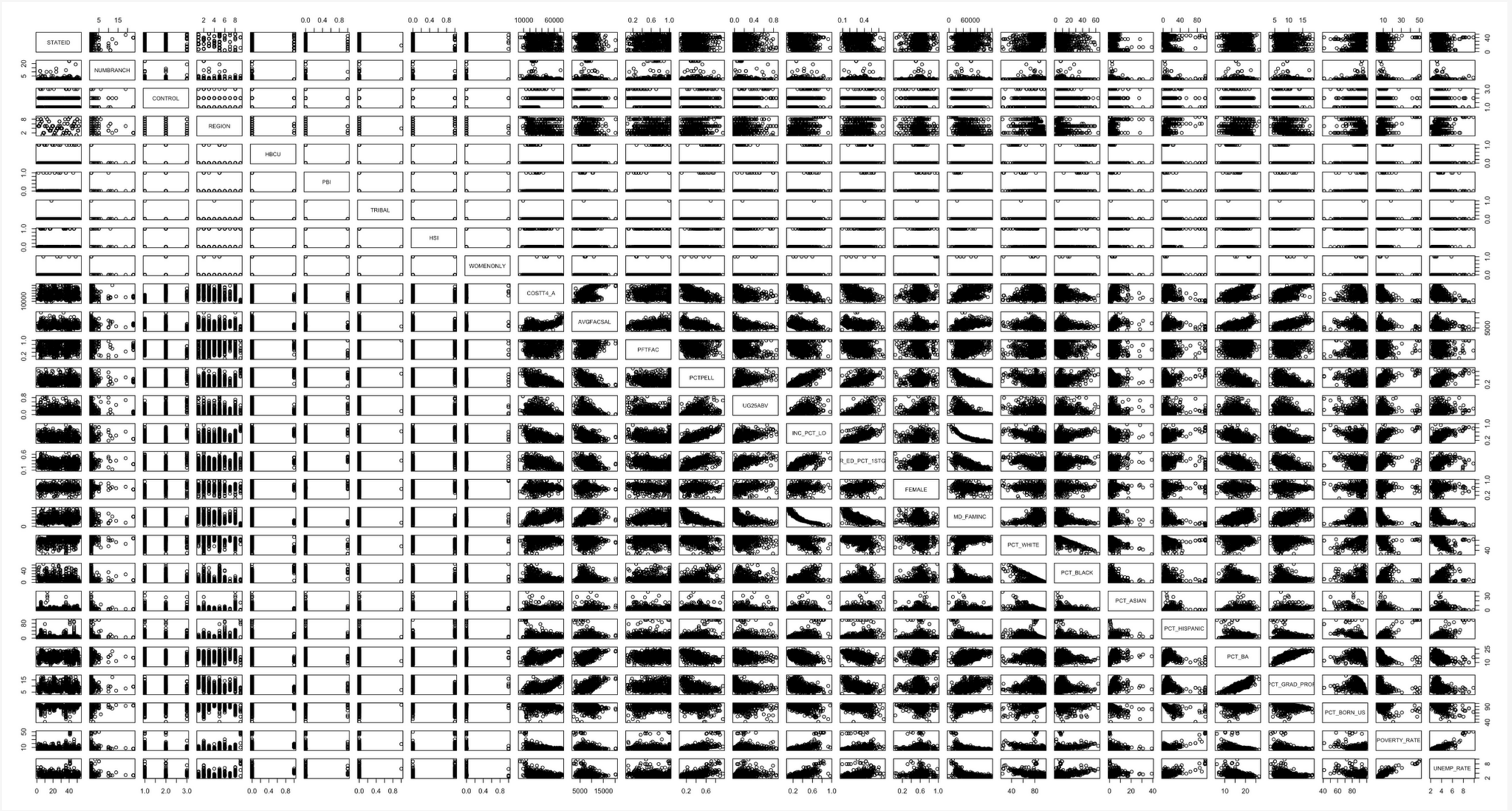
1. Conditional mean response is a single function of a linear combination of the predictors.

2. Conditional mean of each predictor is a linear function with another predictor.

Response versus Fitted Plot



Predictor Correlation Matrix



Observations

From the Correlation Matrix and Response versus Fitted Plot

- A clear non-random, non-linear pattern in the response-fitted plot
- Evidence of non-linear relationships seen in scatterplots of the correlation matrix

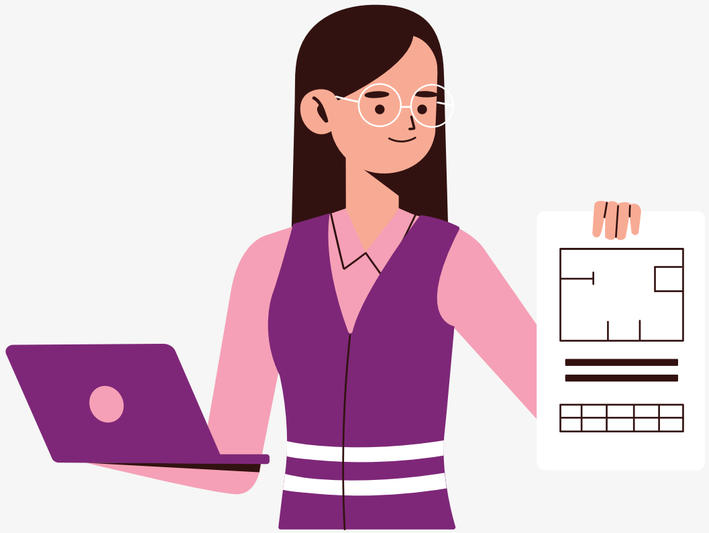
Condition 1 and Condition 2 fails.

Therefore, it is rather implied that the plots show the fitting of an incorrect model. However, it does not provide specifics of the problem.

Power Transformation

Using Box-Cox Methodology

Variables	Power
ADM_RATE	1.40
NUMBRANCH	-5.72
COSTT4_A	0.50
AVGFACSAL	0.33
PCTPELL	0.50
UG25ABV	0.33
INC_PCT_LOW	0.33
PAR_ED_PCT_1STGEN	0.76
FEMALE	2.00
MD_FAMINC	0.50
PCT_WHITE	2.93
PCT_BLACK	0.57
PCT_ASIAN	0.11
PCT_HISPANIC	0.12
PCT_GRAD_PROF	0.33
PCT_BORN_US	9.47
POVERTY_RATE	-0.33
UNEMP_RATE	-0.80



Power of variables raised to better match assumptions and correct skewness towards normality

Reduced Model

Selection of Predictors from the Transformed Data

$$\hat{Y}^{1.40} = 1.921 + 0.001x_1 - 0.066x_2 - 0.085x_3 - 0.033{x_4}^{0.33} + 0.314{x_5}^{0.76} + 0.318{x_6}^{0.33} - 0.832{x_7}^{0.33} + 0.076x_8 - 0.204{x_9}^{0.12} + \varepsilon,$$

where

- $\hat{Y}$ is admission rate (ADM_RATE)
- x_1 is state identifier (STATEID)
- x_2 is control identifier (CONTROL)
- x_3 is number of branches (NUMBRANCH)
- x_4 is average faculty salary (AVGFACSAL)
- x_5 is percentage of first-generation students (PAR_ED_PCT_1STGEN)
- x_6 is percentage of undergraduates aged 25 and above (UG25ABV)
- x_7 is percentage of aided students whose family income is between $0-$30,000 (INC_PCT_LO)
- x_8 is Hispanic-serving institute identifier (HSI)
- x_9 is percentage of Hispanic population in students' neighborhood (PCT_HISPANIC)

What does the coefficients mean?

Predictor Selection

Using individual t-statistics of predictors of interest

	t-value	p-value
STATEID	1.968	0.049465
CONTROL	-4.468	9.14E-06
NUMBRANCH	-3.803	0.000155
AVGFACSAL	-6.904	1.09E-11
PAR_ED_PCT_1STGEN	2.63	0.008722
UG25ABV	5.229	2.22E-07
INC_PCT_LOW	-5.982	3.43E-09
HIS	2.412	0.016107
PCT_HISPANIC	-2.774	0.005683

Step 1: Extract transformed predictors with significant linear relationship with transformed response variable from linear model summary table

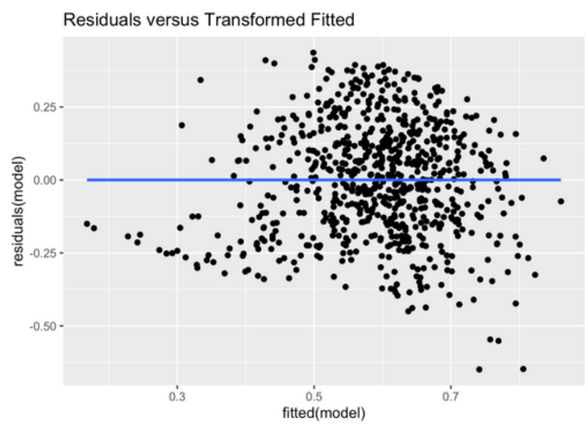
Step 2: Pursue trial-and-error to further replace and add predictors to build a model that better explains variations with all predictors have higher significance levels

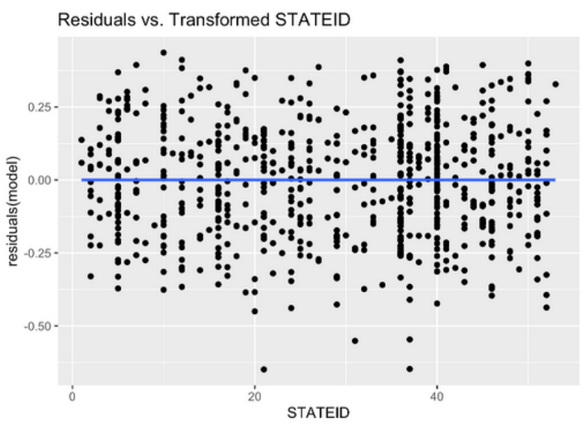
Rechecking Assumptions

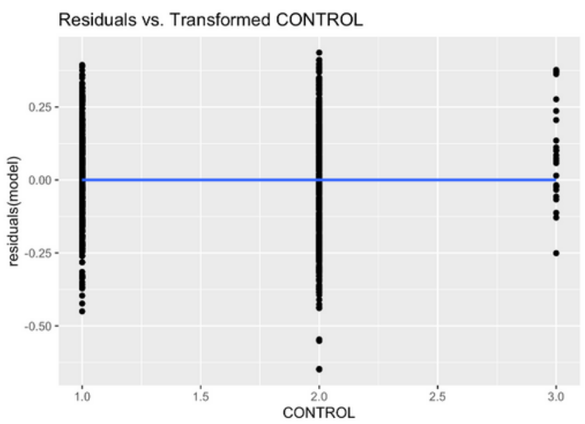
Cross Examination with Selected Predictors

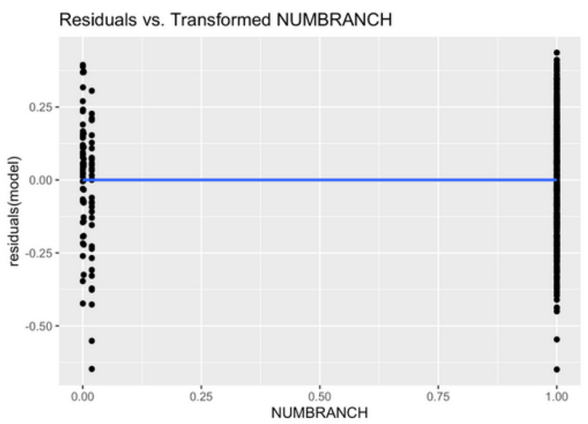
- Residual versus Response Plot
- Residual versus Predictor Plots
- Normal Q-Q Plot

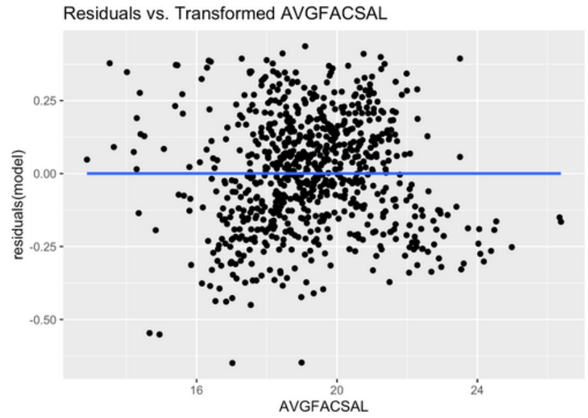
Residual Plots

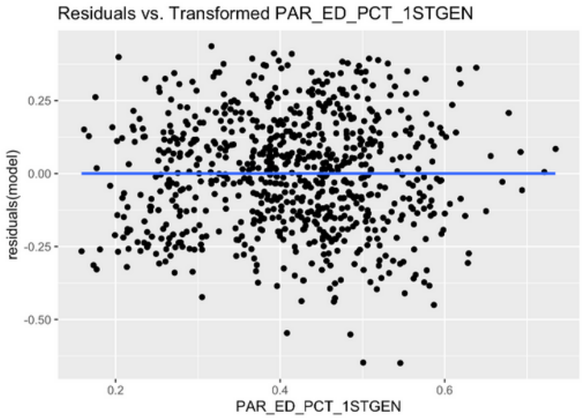


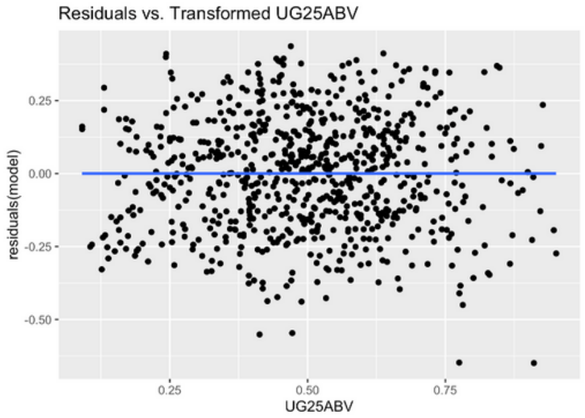


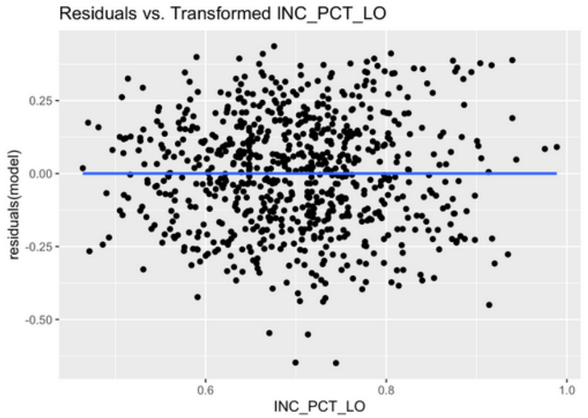


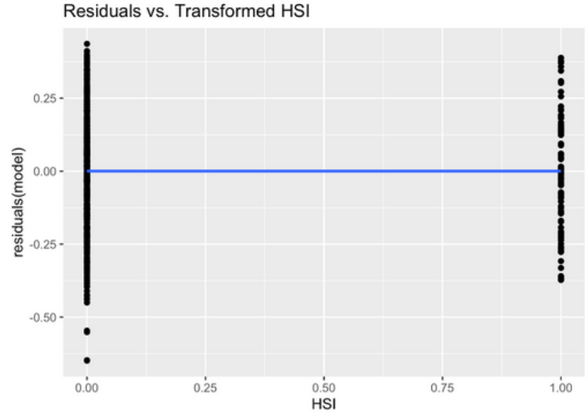


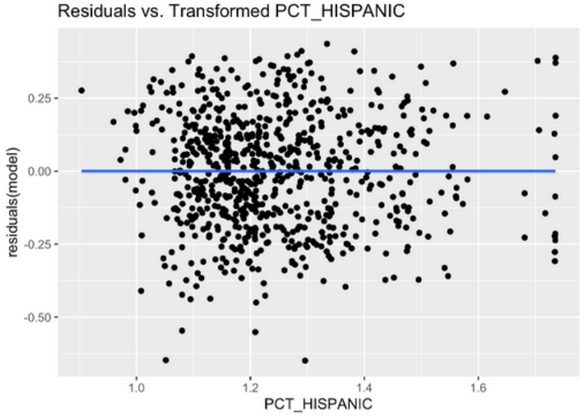




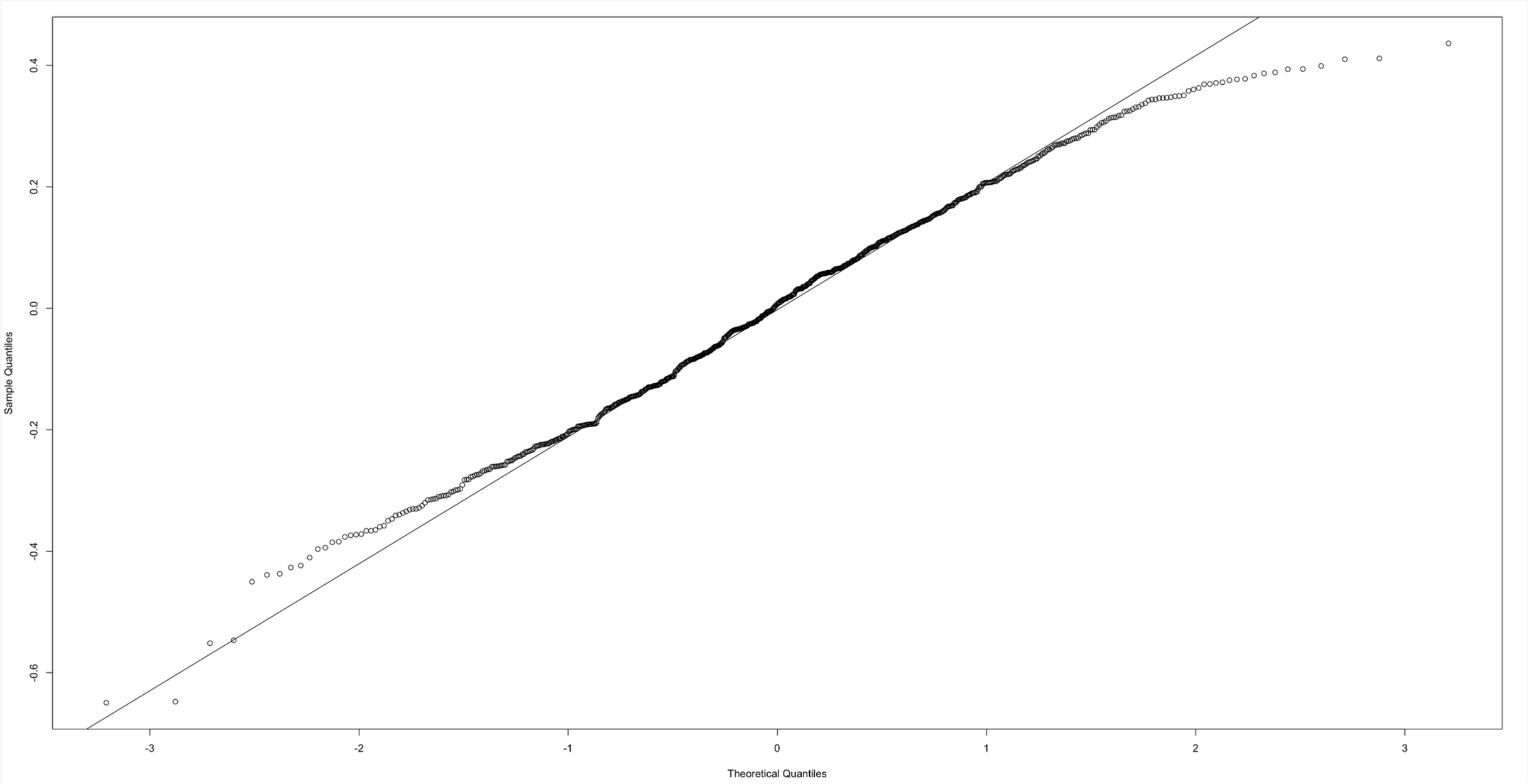








Normal Q-Q Plot



Observations

From the Transformed Residual Plots and the Normal Q-Q plot

- A clear non-random pattern in the residual-transformed fitted plot
- No evidence of non-linear relationships seen in the residual-predictors scatterplots
- Still some deviations on the ends of the Q-Q plot suggesting that normality assumption does not quite hold

Uncorrelated errors and constant variance assumptions are satisfied while linearity and normality assumptions are violated.

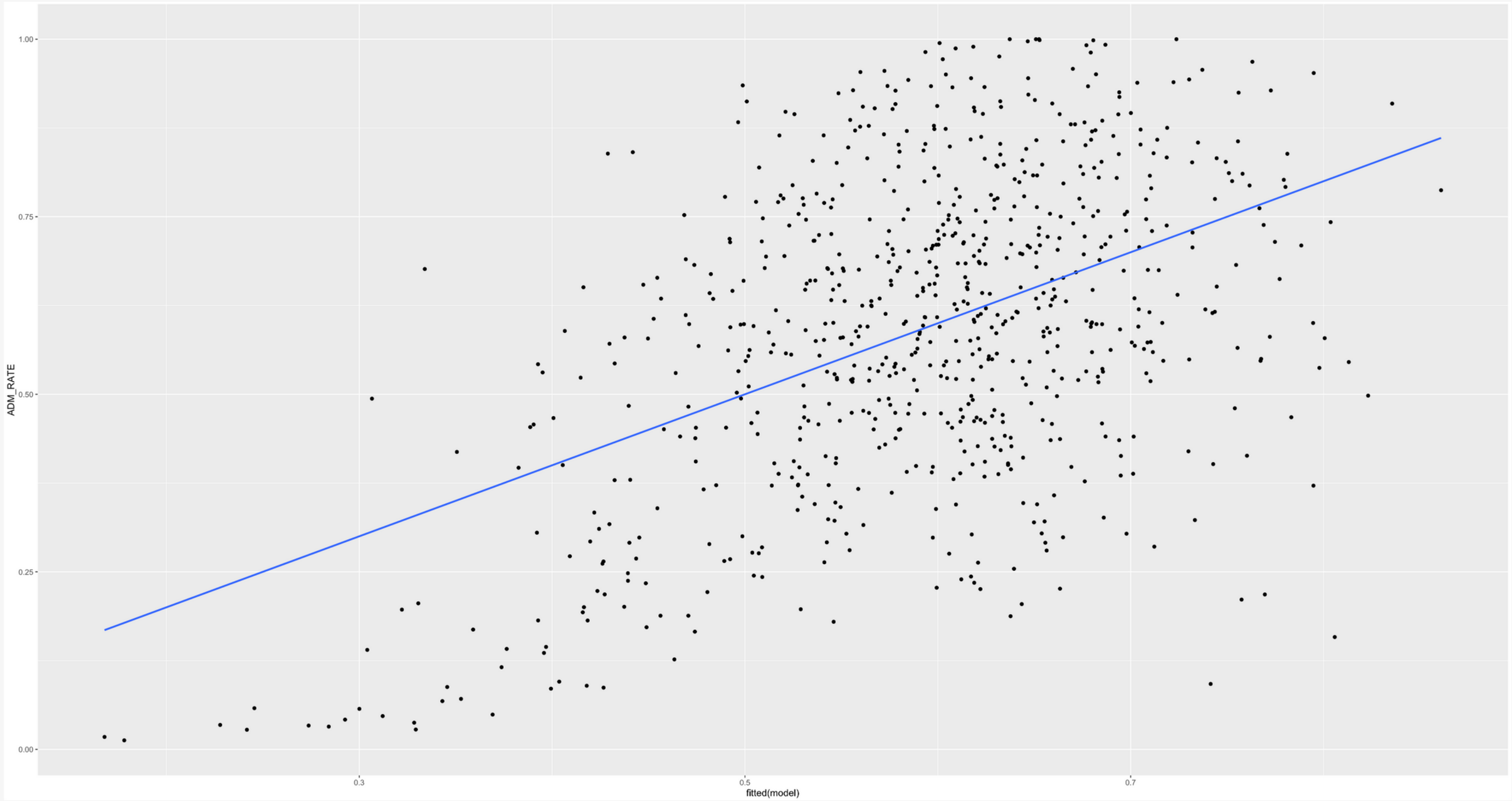
Rechecking Conditions

To Complement Assumption Checks

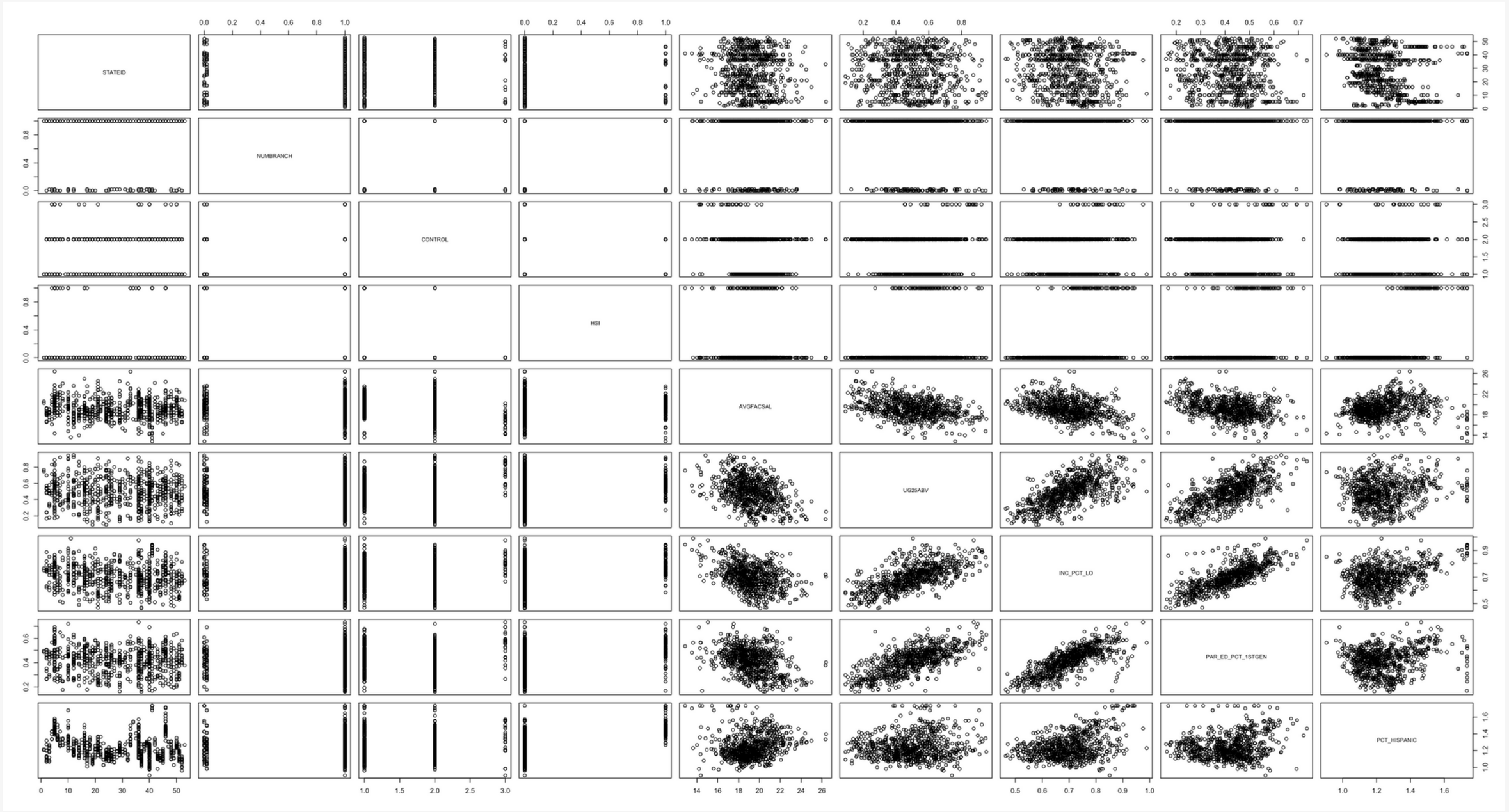
1. Conditional 1 is not satisfied: Non-random pattern in the transformed response-transformed fitted plot

2. Condition 2 is satisfied: No evidence of non-linear relation in the scatterplots

Transformed Response versus Transformed Fitted Plot



Predictor Correlation Matrix



Model Comparison

Using Adjusted R-Squared Values and Partial F-Statistics from ANOVA

	Adjusted R-Squared
Full Transformed Model	0.2171
Reduced Transformed Model	0.2097

Partial F-Statistics	0.1294

Observations

From the Adjusted R-Squared Values and Partial F-Statistics

- The excluded 18 variables only explained 0.74% of the variation as suggested the adjusted R-squared value.
- The partial F-statistics is more than the cut off value of 0.05.

We have built a model with the only predictors that significantly linearly related to the transformed response variable.

Limitations

Of the Reduced Transformed Model

- Condition 1 not satisfied
- Linearity and Normality assumptions not satisfied

The F-statistics of both the models, especially the reduced one, is less than the cut-off value of 0.05 based on the probability distribution that is not normal.

Conclusion

Choosing Best Model with Statistical Methods

- Since assumptions are violated, the relevance of p-values of the test statistics might be insignificant in the analysis of selecting the best model.

- Such the case, we are relying on partial F-statistics and adjusted R-squared to conclude that admission rates of U.S. colleges and universities are best influenced by the predictors in the reduced transformed model.