STA302 Video Project

Zewen Ma, Yuchen Zeng

Outline

Problem introduction - Zewen

EDA and initial model - Zewen

Reduced & Modified model analysis - Yuchen

Final model explanation - Yuchen

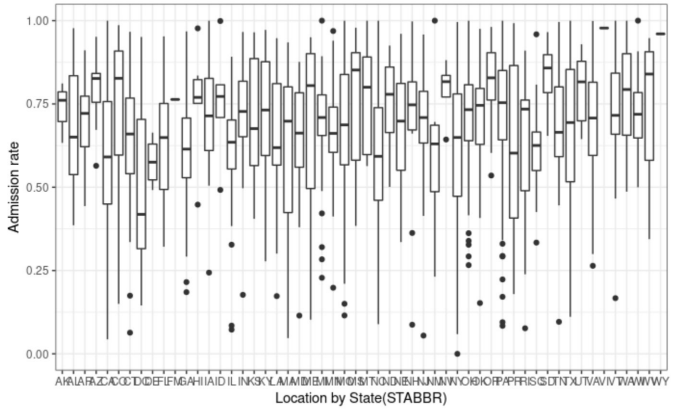
Problem

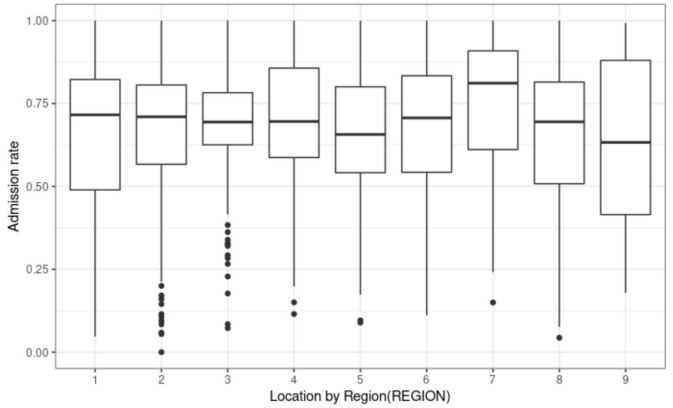
- Admission rates of colleges and universities in the United States can be drastically different from one another and this can be due to a variety of reasons.
  - Total of 1508 institutions
  - No missing variables
  - Has both categorical and numerical variables

- **GOAL:** Create models to find which of the variables/factors in the provided dataset best explains the variation in admission rates in the US.
  - What factors might have impact on the admission rates?

Predictors Selection

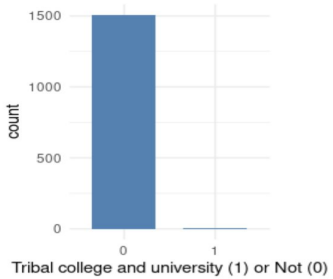
- UNITID and INSTNM are unique identifiers of observations so they can't be predictors.
- REGION and STABBR
  - Shows similar information.
  - Hard to interpret STABBR
    - Singularity in some category
    - No significant estimate
  - For both simplicity and accuracy, removed STABBR from predictors.

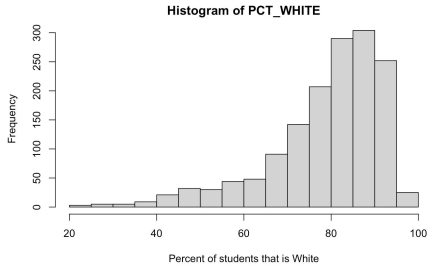


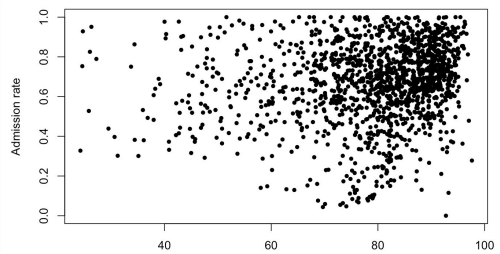


Predictors Removed

- Predictors related to race and ethinic group
  - Race and ethinic group are not a factor for the overall admission rate
    - (Sandra & Amir, 2002)
  - Categorical:
    - HBCU, PBI, TRIBAL, HSI
    - Not evenly distributed
  - Numeric:
    - PCT_WHITE, PCT_BLACK, PCT_ASIAN, PCT_HISPANIC
    - Highly skewed
    - No visible linear relationship with response
  - Predictors related to race and ethinic group are not selected

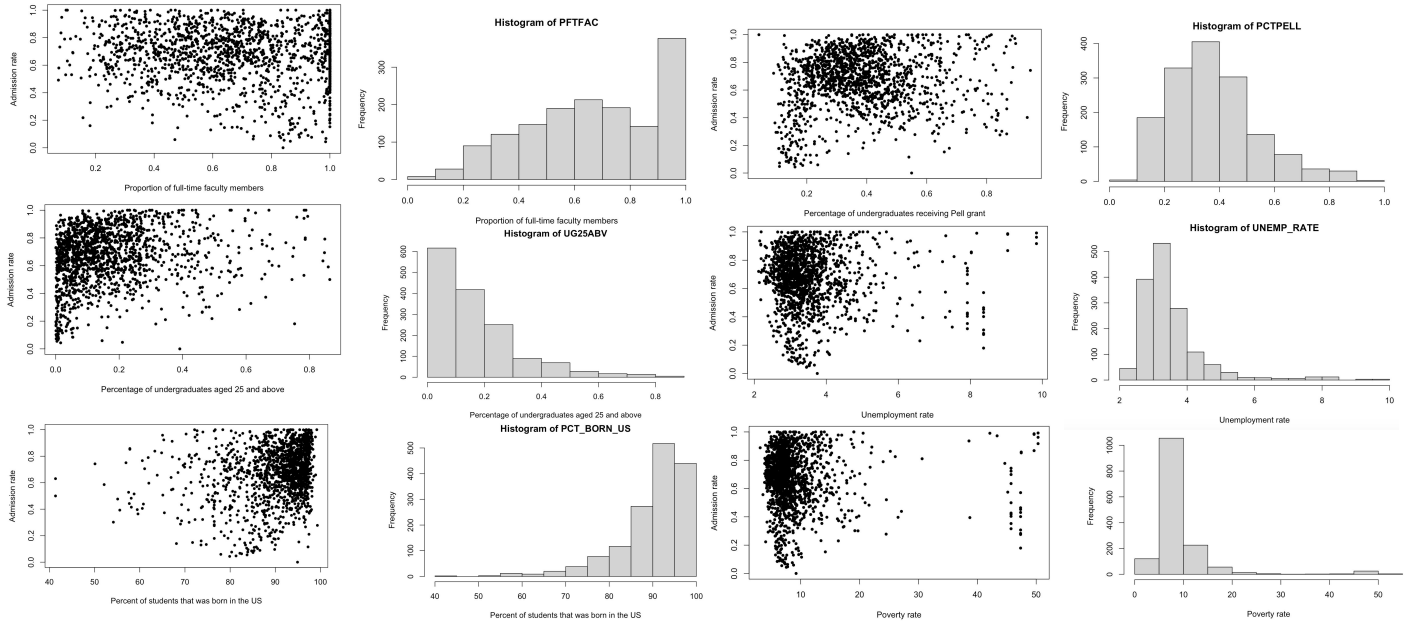






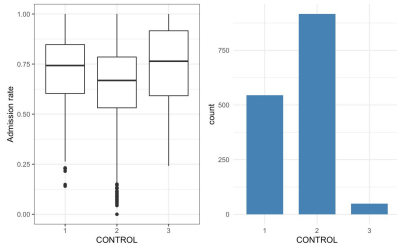
Predictors Removed

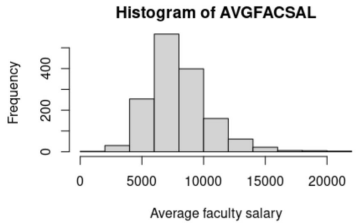
- Numeric Variables
  - Some variables are skewed and does not show linear relationship with the response
    - PFTFAC, UG25ABV, PCT_BORN_US, POVERTY_RATE, UNEMP_RATE, PCTPELL
    - Not selected as predictors

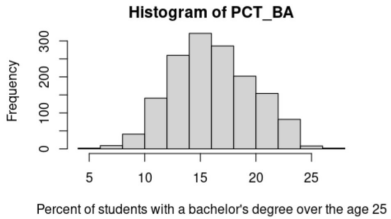


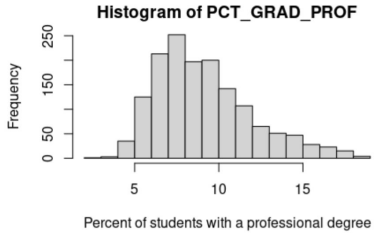
Predictors Kept

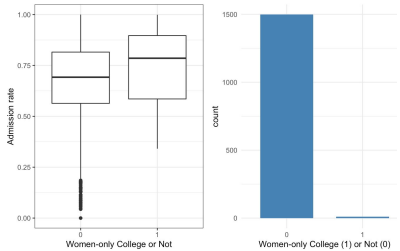
- Categorical Variables: CONTROL, WOMENONLY
  - Imbalanced distribution
- Numeric Variables: AVGFACSAL, PCT_BA. PCT_GRAD_PROF
  - The following variables are quite normal and shows a linear relationship with the response variable
    - Likely good predictors

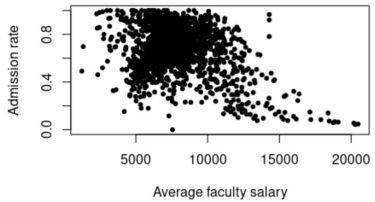


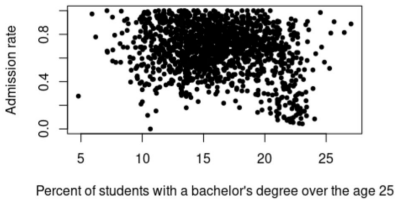


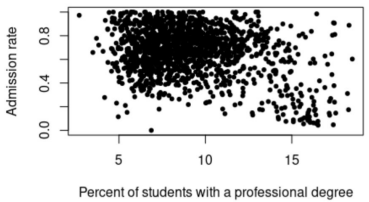






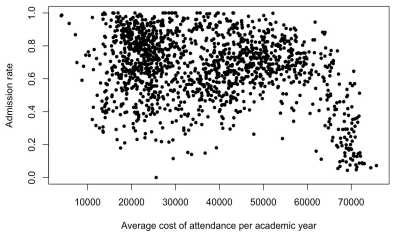


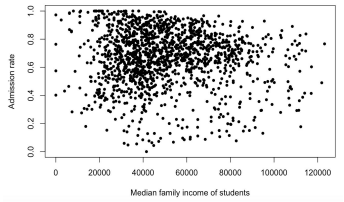


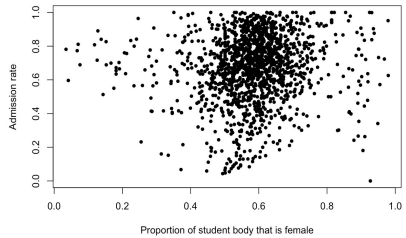


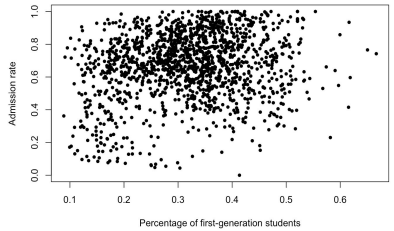
Predictors Kept

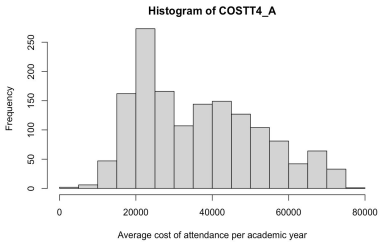
- Numeric Variables
  - COSTT4_A seems to have a linear relationship with the response
  - MD_FAMINC, FEMALE, PAR_ED_PCT_1STGEN is less skewed but no obvious relationship with the response are seen.

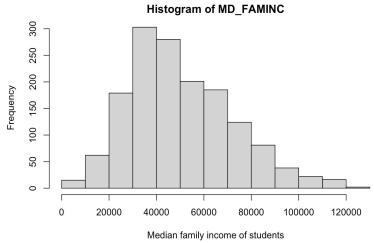


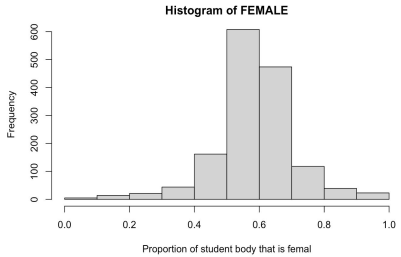


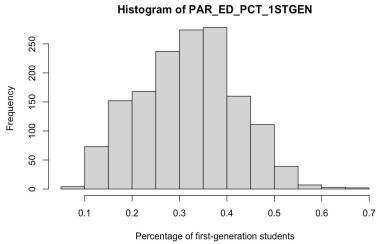








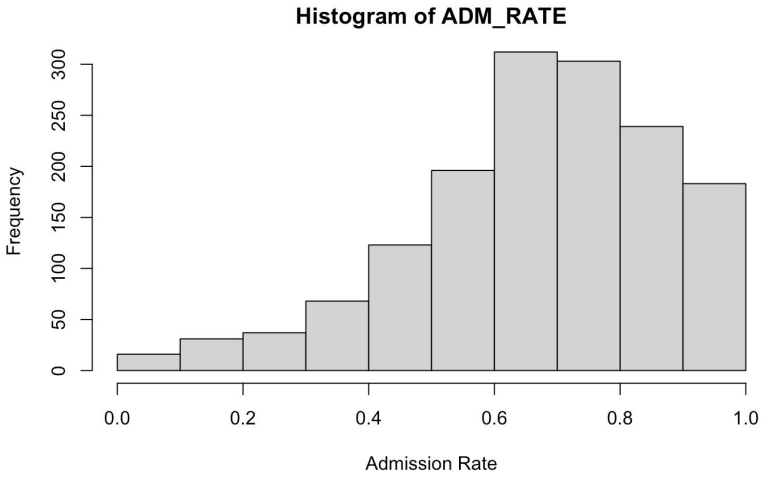


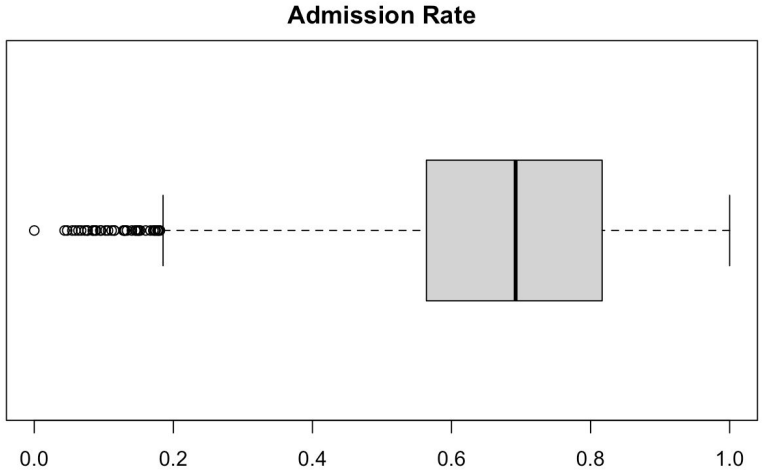


- **Initial Selection of Predictors (10):**
  REGION + CONTROL + WOMENONLY + COSTT4_A + AVGFACSAL + FEMALE + PCT_BA + MD_FAMINC + PAR_ED_PCT_1STGEN + PCT_GRAD_PROF

Response Variable

- Our response variable is admission rate, ADM_RATE
- Left Skewed
  - Transformation may be needed to satisfy assumption of normality

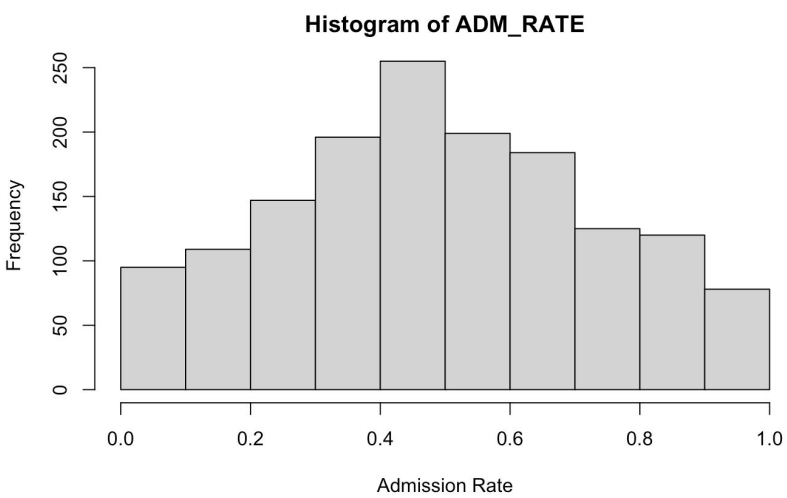


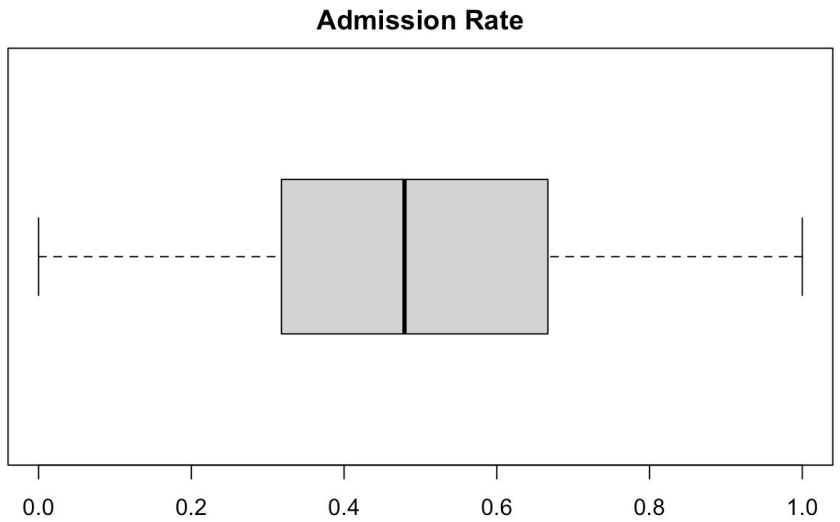


Zewen Ma

Assumption of Normality

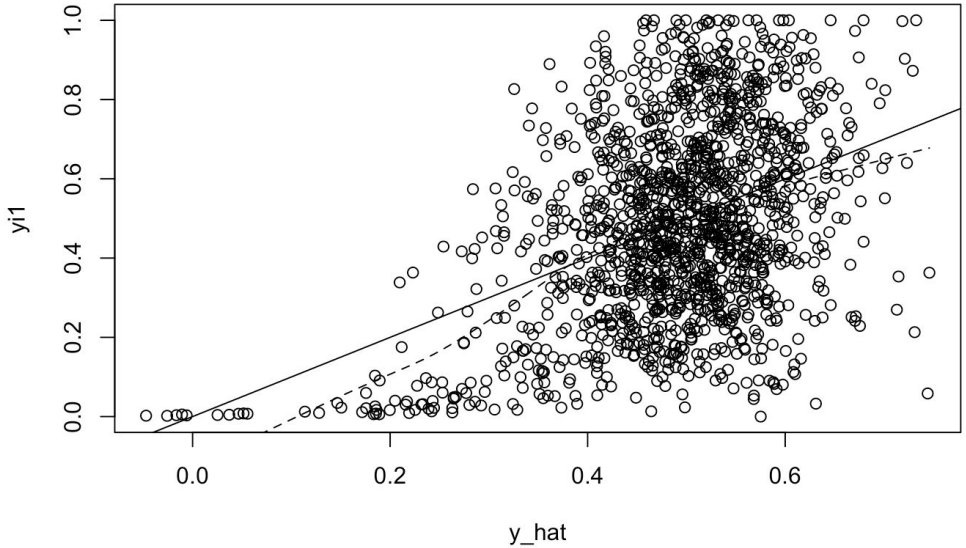
- We know distribution of admission rate is not a normal
- Transform it to squared admission rate(ADM_RATE^2)

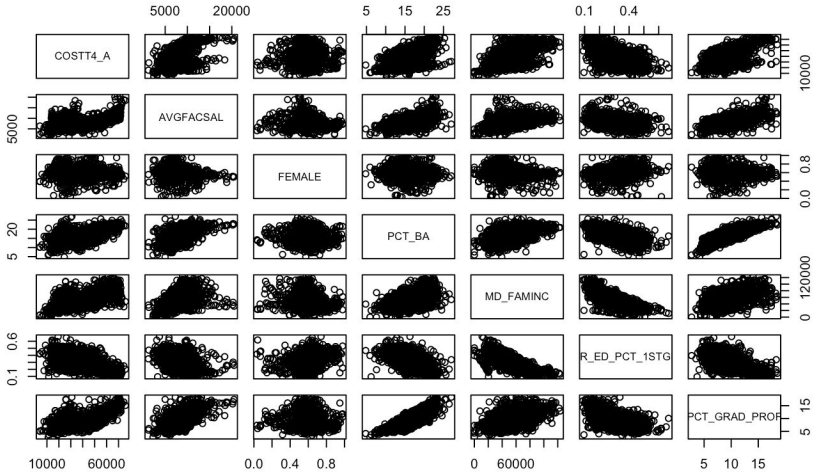




Condition 1 and 2

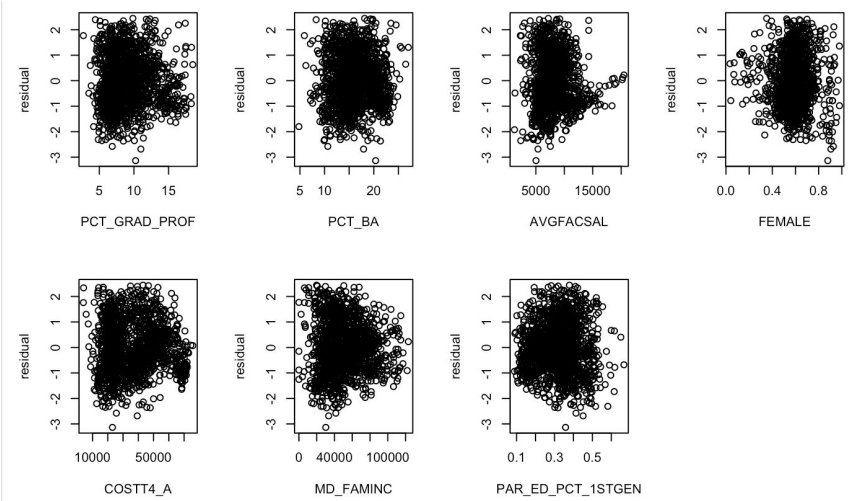
- **Initial Model:**
  - (ADM_RATE)^2 = REGION + CONTROL + WOMENONLY + COSTT4_A + AVGFACSAL + FEMALE + PCT_BA + MD_FAMINC + PAR_ED_PCT_1STGEN + PCT_GRAD_PROF

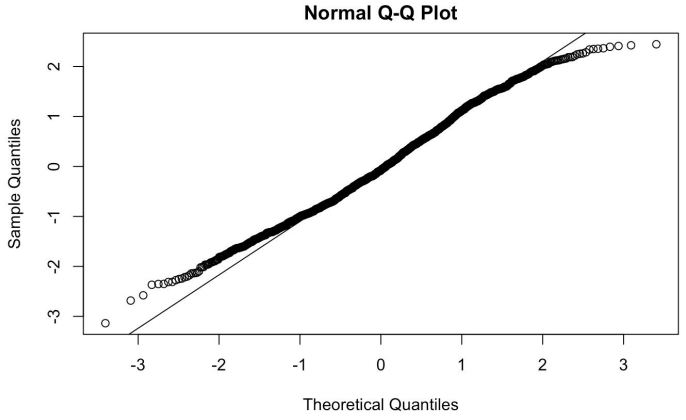


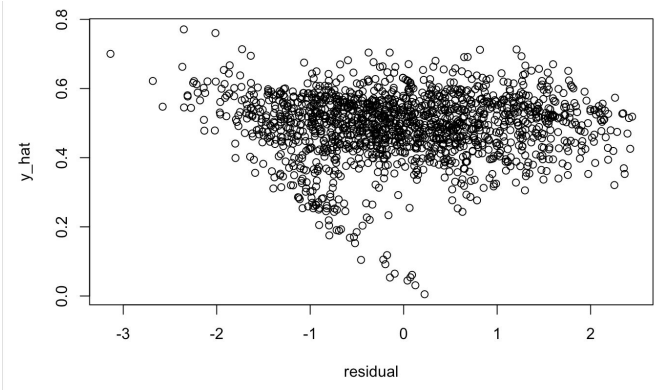


Zewen Ma

Residual Plots, QQ Plot, Residual vs. Fitted Value





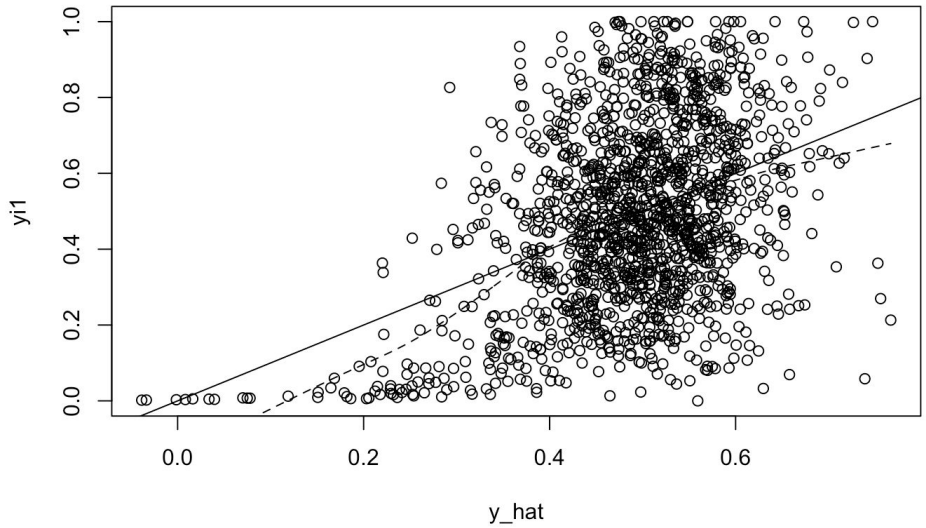


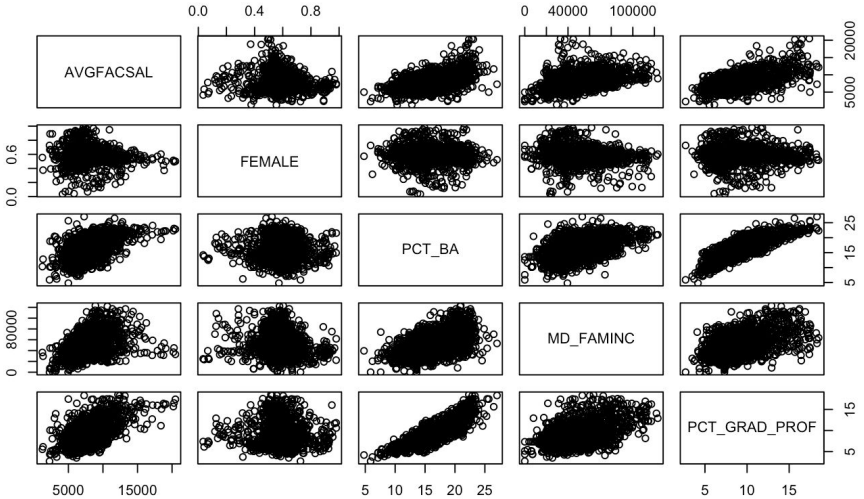
Reduce the Model

- We have checked all assumptions
- Too complicated model - We will remove predictors that are not significant, with significant level of 0.01
  - Removed WOMENONLY, COSTT4_A, PAR_ED_PCT_1STGEN
    - not significant
- **Reduced Linear Model:**
  - (ADM_RATE)^2 = REGION + CONTROL + AVGFACSAL + FEMALE + PCT_BA + MD_FAMINC + PCT_GRAD_PROF
  - Initial Adjusted R^2 = 0.1572, Reduced Adjusted R^2 = 0.1547
  - By ANOVA F Test
    - P-value: 0.05987
    - The predictors can be removed

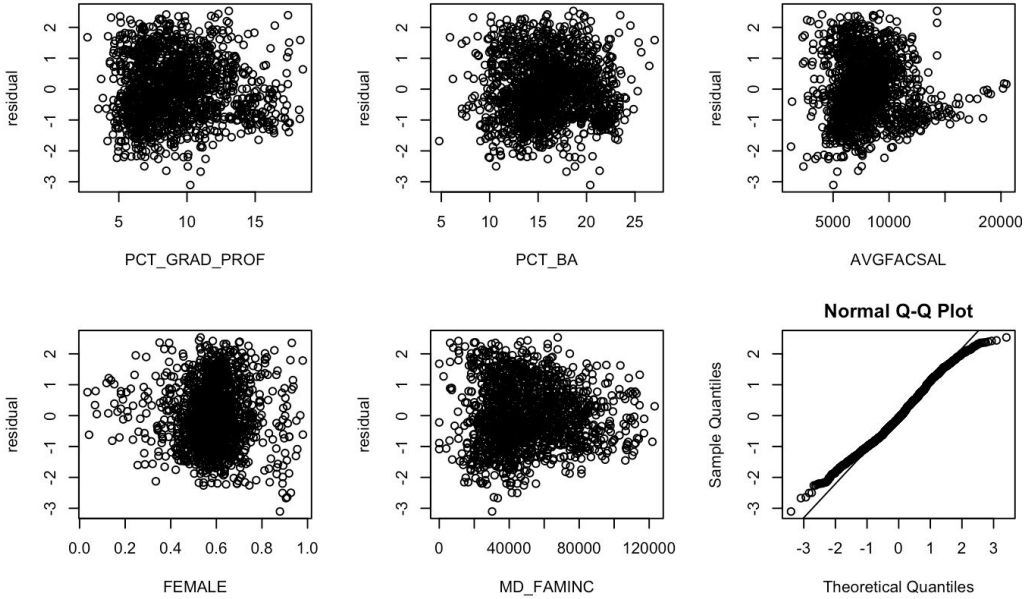
Condition 1 and 2

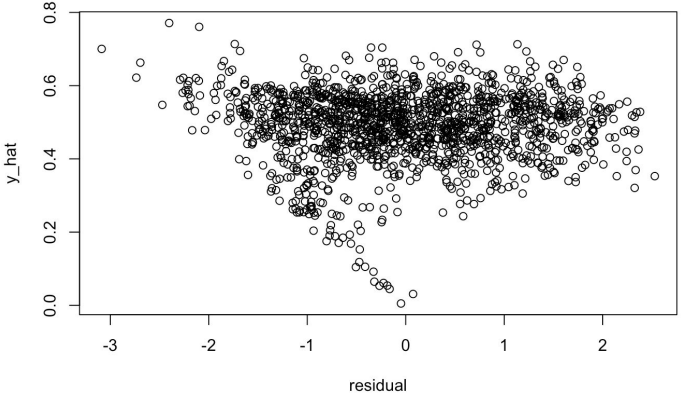
- **Reduced Model:**
  - (ADM_RATE)^2 = REGION + CONTROL + AVGFACSAL + FEMALE + PCT_BA + MD_FAMINC + PCT_GRAD_PROF





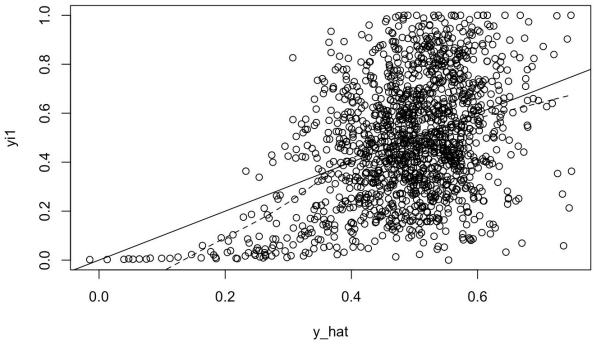
Residual Plots, QQ Plot, Residual vs. Fitted Value

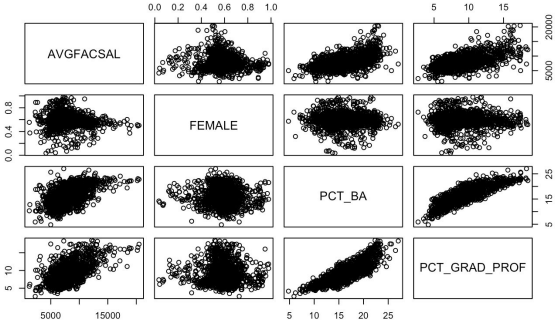




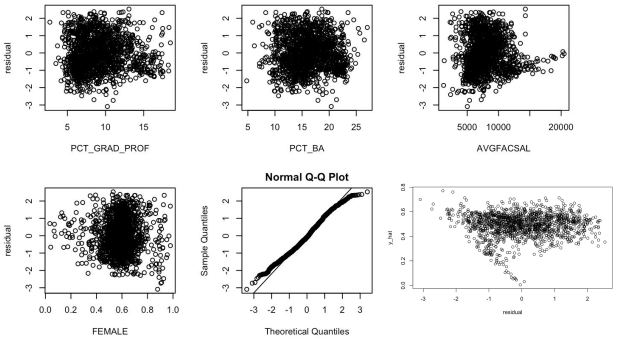
Further Reduction

- MD_FAMINC is now not significant with significant level of 0.01
  - Removed from the model
- **Reduced Linear Model:**
  - (ADM_RATE)^2 = REGION + CONTROL + AVGFACSAL + FEMALE + PCT_BA + PCT_GRAD_PROF
  - Initial Adjusted R^2 = 0.1572, Reduced Adjusted R^2 = 0.1523
  - ANOVA F Test with the naive model
    - P-value: 0.01307
    - MD_FAMINC can be removed
- Condition 1 & 2

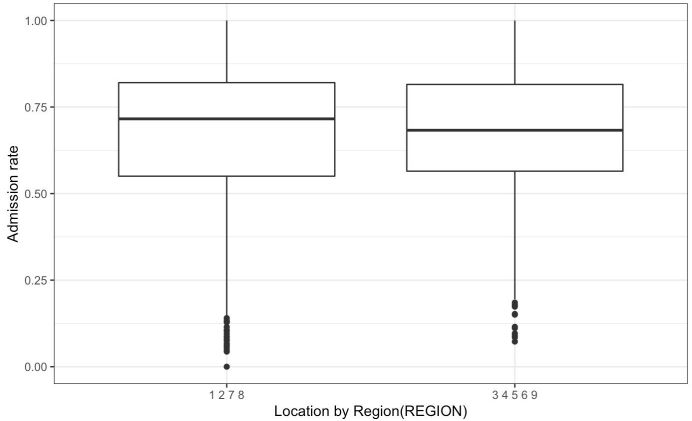




- Residual and QQ plot



Modified Linear Model



- REGION
  - REGION 1, 2, 7, 8 are not significant
  - ALso, 9 regions is hard to interpret
  - Introduce a new variable called MOD_REGION:
    - Has 3 categories: Institution in Region 3 4("3 4"), Region 1 2 7 8("1 2 7 8"), Region 5 6 9(“5 6 9”)
- **Modified Linear Model:**
  - (ADM_RATE)^2 = MOD_REGION + CONTROL + AVGFACSAL + FEMALE + PCT_BA + PCT_GRAD_PROF
  - Before R^2 = 0.1602, New R^2 = 0.1542
  - F test with p-value < 2.2e^-16

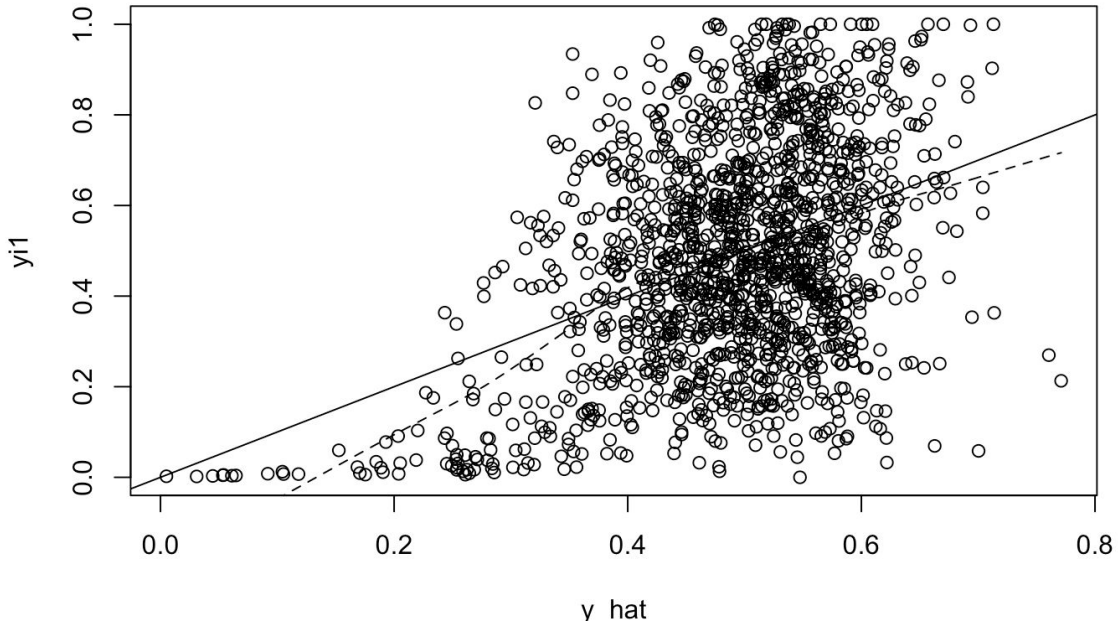
Final Linear Model

Intercept	7.399e-01	< 2e-16
Predictor	Estimate	Significance
(MOD_REGION)3 4	-5.165e-02	0.002201
(MOD_REGION)5 6 9	-1.059e-01	1.12e-11
(CONTROL)2	-1.120e-01	2.47e-16
(CONTROL)3	-8.602e-02	0.016820
AVGFACSAL	-3.175e-05	< 2e-16
FEMALE	1.751e-01	0.000304
PCT_BA	1.258e-02	0.000185
PCT_GRAD_PROF	-1.995e-02	3.16e-05

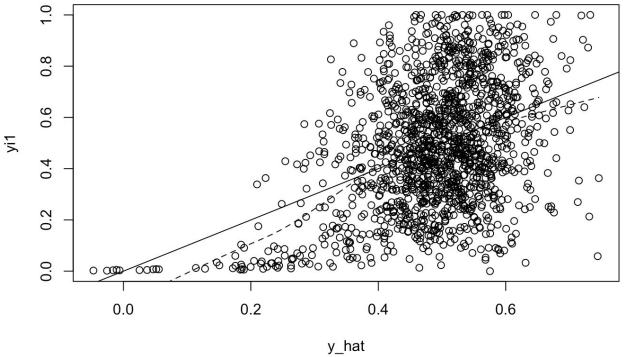
- (ADM_RATE)^2 = MOD_REGION + CONTROL + AVGFACSAL + FEMALE + PCT_BA + PCT_GRAD_PROF
- Total of 6 Predictors
- Example:

  Given all other factors held constant:
  - When average faculty salary increase by 1, the squared admission rate decrease by 2.1%

Final Model - Condition 1

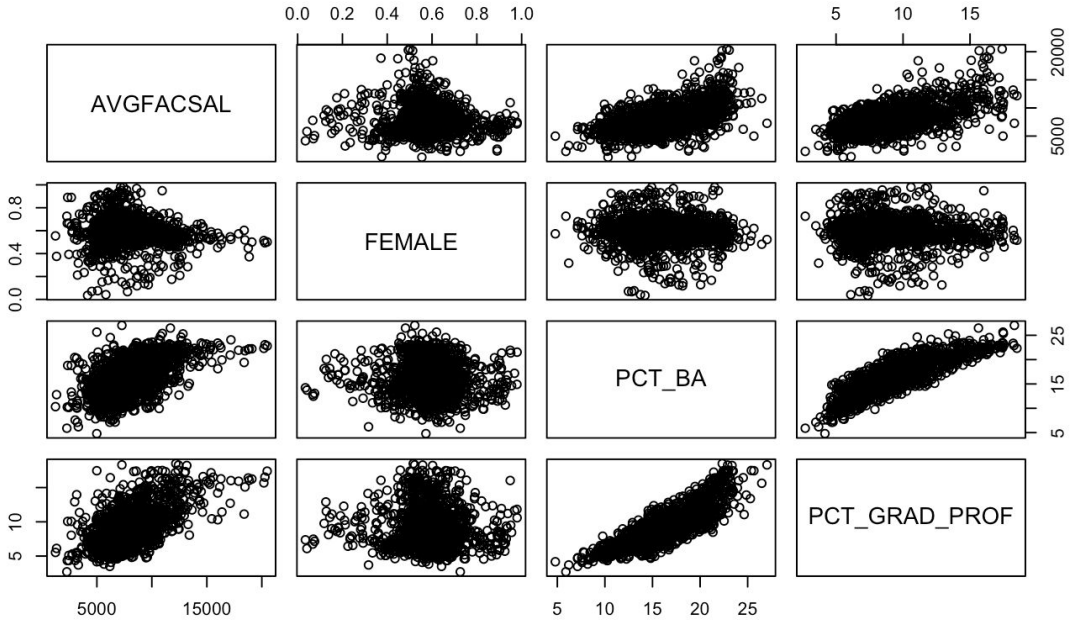


- Initial Graph



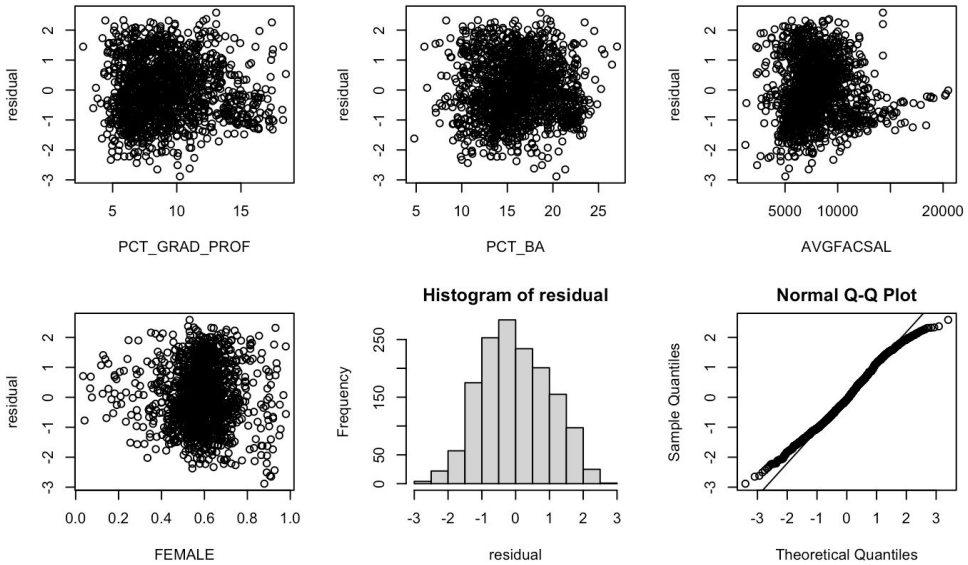
- By EDA we know numeric predictors doesn't have abnormal relationship with response
- Satisfied Assumption of Linearity

Final Model - Condition 2



Final Model - Assumption of Unrelated Error and Normality

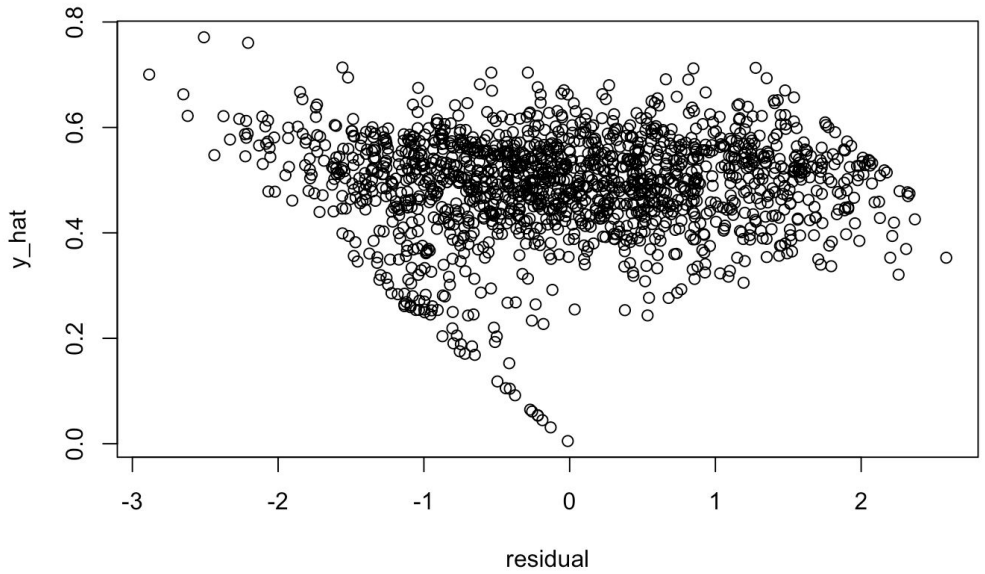
- Residual plot and QQ plot



- Unrelated Error
  - Residuals are randomly scattered
- Normality
  - Histogram of residual
    - Bell shaped, symmetric
  - Normal QQ plot
    - Mostly on the line

Final Model - Assumption of Constant Variance

- Residual vs. fitted value



- Constant Variance
  - Mostly equally scattered
  - A few bad observations

Final Interpretation

- (ADM_RATE)^2 = MOD_REGION + CONTROL + AVGFACSAL + FEMALE + PCT_BA + PCT_GRAD_PROF
- All predictor are statistically significant
- Simple - 6 predictors
- In context:
  - lower admission rate:
    - Private institution, especially private for-profit institution
    - Students live in a professionally educated neighbourhood
    - High faculty salary
    - In Great Lakes, Plains region, even lower admission rate in Southeast, Southwest, Outlying Areas
  - Higher admission rate:
    - Institution with more female students
    - Students live in a educated neighbourhood