Regression Analysis on Early Career Pay

Branden Lopez

College Tuition, Diversity, & Pay Dataset

Data from the US Department of Education

Compiled and accumulated by TuitionTracker

- Historical averages spanning 1985-2019
- Tuition & Fees
- School type, degree lengths
- Diversity group
- Net income & salary potential

Questions

- Can average early career pay be predicted by the characteristics of a school?
- What features are most important in determining average early career earnings.

Initial Observations

- Salary information is limited to the top-25 schools in each state ranked in order of median_career_salary.
- Therefore any results for this survey reflects the characteristics of 'successful' schools.

17	San Jose State University	California	63000	114700

Data Preprocessing

- Salary information is a single table. Inner joins on the school name were performed.

  - diversity_school.csv
  - historical_tuition.csv
  - salary_potential.csv
  - tuition_cost.csv
  - tuition_income.csv

- Salaries are from 2018 while the supporting tables span from the mid 80’s to 2017.
- Narrowing variables such as tuition cost to the year 2017 to reflect what drives early career salary in 2018.
- Null values are processed with simple mean imputation.
- Much more data analysis and preprocessing in ‘Datasets/Merge and Clean.ipynb’

Data Preprocessing (cont.)

- Categorical variables:
  - tuition_higher_than_national _average
    - 1 being yes; 0 being no
  - Type
    - 1 being public; 0 being private
  - Degree_length
    - 1 being two years; 0 being four years
- For a total of 14 predictors.

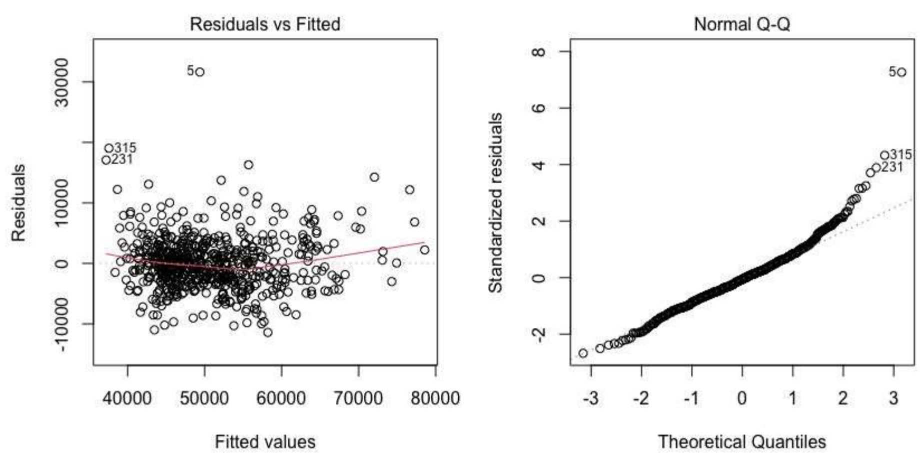
```
name                                        object
state_name                                  object
early_career_pay                           float64
make_world_better_percent                  float64
stem_percent                               float64
year                                         int64
total_price                                float64
net_cost                                   float64
tuition_higher_than_national_average         int64
type                                         int64
degree_length                                int64
room_and_board                             float64
in_state_tuition                           float64
in_state_total                             float64
out_of_state_tuition                       float64
out_of_state_total                         float64
Total Minority                             float64
total_enrollment                           float64
dtype: object
```

Verifying Assumptions

- Linear regression requires that predictors have linear trends with our response
- Constant variance.
- No autocorrelation.
- Little or no multicollinearity.

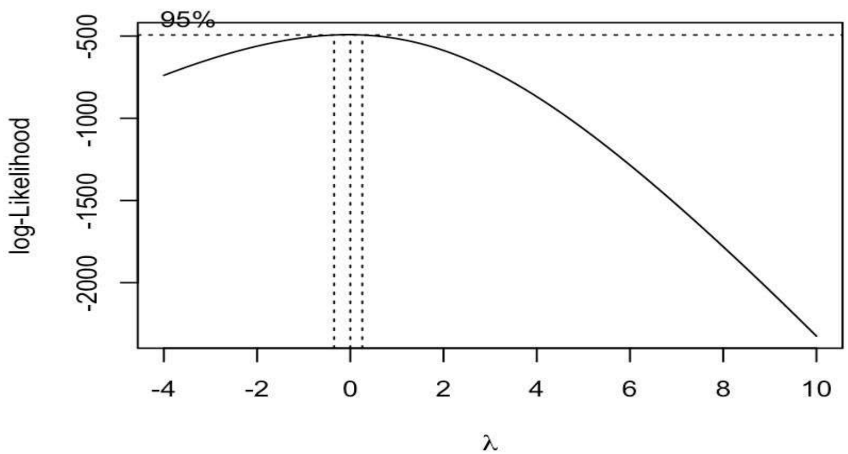
Plotting

- Residual v fitted plot indicates if variance is approximately constant.
- Normal QQ indicates if data is normally distributed.
- They also showcase outliers.



Transformations (cont.)

- Linearity plots show linear trends with the response.
- No obvious transformation from Res V fitted plot.
- Automatic method might show otherwise.



Boxcox shows lambda = 0. Meaning we transform the response of early career pay with a log transformation.

Outliers

- Verifying potential outlier with Cook's Distance.
- Outlier tend to be Health care institutions (Highlighted)

```
                                        name
Albany College of Pharmacy and Health Sciences
                             Bellevue College
                               Bellin College
             Southern New Hampshire University
             Mount Carmel College of Nursing
                            Kettering College
```

- Only obvious in 2 of the names; others are Health care verified 'by-hand'
- It might be beneficial to add a 'Health-Care' variable.
- We only omit the most significant outlier for modeling.

Multicollinearity

- Multicollinearity occurs when features are linear combinations of one another.
- In our dataset 'total_cost' of school attendance is a linear combination of 'room and board' and tuition cost.
- Linear regression requires a matrix inverse.
- A matrix inverse requires full rank, that is columns to not be linear combinations of one another.
- We eliminate granular predictors for to remove multicollinearity.

Variable Selection

- After re-verifying assumptions with transformed response. We select subset of predictors.
- For predictors > 4 summary statistics such as Mallow's CP, R-squared, Mean Squared Residual cluster around similar values.
- Forward and backward selections are used to arrive at models.
- For forward: kep variable with lowest p-value and highest F-value.
- Vice versa for Backward.

- Forward and backwards arrive at the same result!

Final Model

$$log(y) = 10.36 + 3.945 * 10^{-3} StemPercent + 5.76 * 10^{-6} TotalPrice$$
$$+ .009 Type + 8.6 * 10^{-6} RoomAndBoard - 4.805 * 10^{-6} TotalMinority$$
$$+ 5.09 * 10^{-6} TotalEnrollment$$

- The model indicates that Large-expensive institutions with a high stem percentage produce higher early career salary.

- Coefficients are not what they seem, due to log response variable.
- Consider a one percent (unit) increase in StemPercent. We need to exponentiate our coefficient.
- exp(.0003945) = 1.003953
- This says every one-unit increase in StemPercent leads to an increase of 1.003953. Or in other words, for every one-unit increase in StemPercent, Early Career Salary increases by about .4%.

- Therefore a unit increase in Total Minority decreases early career salary by 4 ten-thousandths of a percent.
- Stem percent is the most important variable in predicting early career salary,

Improvements

- Of the 935 instances only 628 where used. Inner join uses string literal matching, fuzzy matching could perform better.
- Health Care Institution Variable.

Thank you!