# Salary inference and prediction from college data

Theodoros Mamalis

## Contents

# Salary inference and prediction from college data

## Introduction and Explanation of the Data

-Data origins.

The data was obtained from Kaggle (link: https://www.kaggle.com/jessemostipak/college-tuition-diversity-and-pay/version/2?select=historical_tuition.csv). It originally came from the US Department of Education.

-Data background.

As per the Kaggle description these datasets contains diversity statistics (e.g. number of Native Americans), and tuition and fees for various United States college and universities, along with school type, degree length and state. Moreover, historical tuition averages (e.g. over a few number of years) from the National Center for Education Statistics, are also included. Moreover, it contains salary potential per university, and graduate rates, and other information. A complete description of the datasets along with the various sources from which the datasets were obrained can be found at: https://www.kaggle.com/jessemostipak/college-tuition-diversity-and-pay/version/2?select=historical_tuition.csv

-Brief introduction of the goal of this final project.

The goal of this final project is to firstly analyze the College Tuition dataset and then build inference and prediction models with the response variable being one of the dataset variables, and in specific, the sum of their estimated early- and mid-career salaries. For the first part, an Explanatory Data Analysis will be conducted which will analyze the dataset with respect to the nature of the variables included, what variables should be kept, trends between variables (e.g. the costs of college tuition by geographic area, or estimated salary trends) etc. For the second part, inference will be conducted initially, and inference conclusions will be presented. Then, prediction models will be made for that response variable and the best will be selected.

Three datasets will be used.

The first is a dataset containing enrollment information as shown below:

```
## # A tibble: 3 x 5
##   name                 total_enrollment state  category               enrollment
##   <chr>                           <dbl> <chr>  <chr>                       <dbl>
## 1 University of Phoen~           195059 Arizo~ Women                      134722
## 2 University of Phoen~           195059 Arizo~ American Indian / Ala~        876
## 3 University of Phoen~           195059 Arizo~ Asian                        1959
```

The second is a dataset containing various information including salary potential:

```
## # A tibble: 3 x 7
##    rank name  state_name early_career_pay mid_career_pay make_world_bett~
##   <dbl> <chr> <chr>                 <dbl>          <dbl>            <dbl>
## 1     1 Aubu~ Alabama               54400         104500               51
## 2     2 Univ~ Alabama               57500         103900               59
## 3     3 The ~ Alabama               52300          97400               50
## # ... with 1 more variable: stem_percent <dbl>
```

The third contains tuition cost information:

```
## # A tibble: 3 x 10
##   name  state state_code type  degree_length room_and_board in_state_tuition
##   <chr> <chr> <chr>      <chr> <chr>                  <dbl>            <dbl>
## 1 Aani~ Mont~ MT         Publ~ 2 Year                    NA             2380
## 2 Abil~ Texas TX         Priv~ 4 Year                 10350            34850
## 3 Abra~ Geor~ GA         Publ~ 2 Year                  8474             4128
## # ... with 3 more variables: in_state_total <dbl>, out_of_state_tuition <dbl>,
## #   out_of_state_total <dbl>
```

## Data wrangling

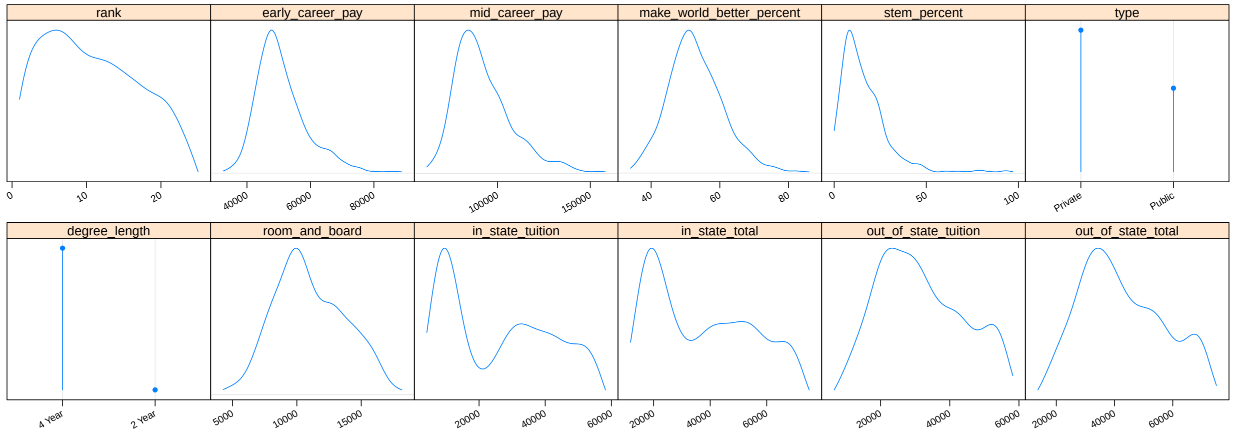
For the first dataset the column values are turned into variables to create unique names for schools to join them with the other two datasets, and also create more predictors (feature engineering). Then, the first dataset is merged with the second and the resulting dataset is merged with the third. Afterwards, some duplicate predictors, concerning the various states, are dropped. The state names, (school) type and degree length predictors are converted to factors. Moreover, each observation is characterized by the name of the university. Therefore, the name of the university acts as a unique identifier for each observation, and can be dropped. Furthermore, the early career pay and mid career pay are added, to create a column of the sum of the two, which will serve as the predicted value of this dataset. After these changes, the first few observations of the final dataset is:

```
##   rank early_career_pay mid_career_pay make_world_better_percent stem_percent
## 1   16            44400          81400                        56            3
## 2   14            46000          83600                        57           26
## 3   20            39800          71500                        61           16
##   state_code    type degree_length room_and_board in_state_tuition
## 1         CO  Public        4 Year           8782             9440
## 2         GA Private        4 Year          12330            41160
## 3         AL  Public        4 Year           5422            11068
##   in_state_total out_of_state_tuition out_of_state_total total_enrollment
## 1          18222                20456              29238             3154
## 2          53490                41160              53490              873
## 3          16490                19396              24818             5519
##   American Indian / Alaska Native Asian Black Hispanic
## 1                              37    37   191      812
## 2                               1    41   288       78
## 3                               8    20  4950       68
##   Native Hawaiian / Pacific Islander Non-Resident Foreign Total Minority
## 1                                  10                    0          1213
## 2                                   2                  104           462
## 3                                   3                  130          5105
##   Two Or More Races Unknown White Women total_pay_mid_early
## 1               126     165  1776  1728              125800
## 2                52      23   284   867              129600
## 3                56      76   208  3430              111300
```

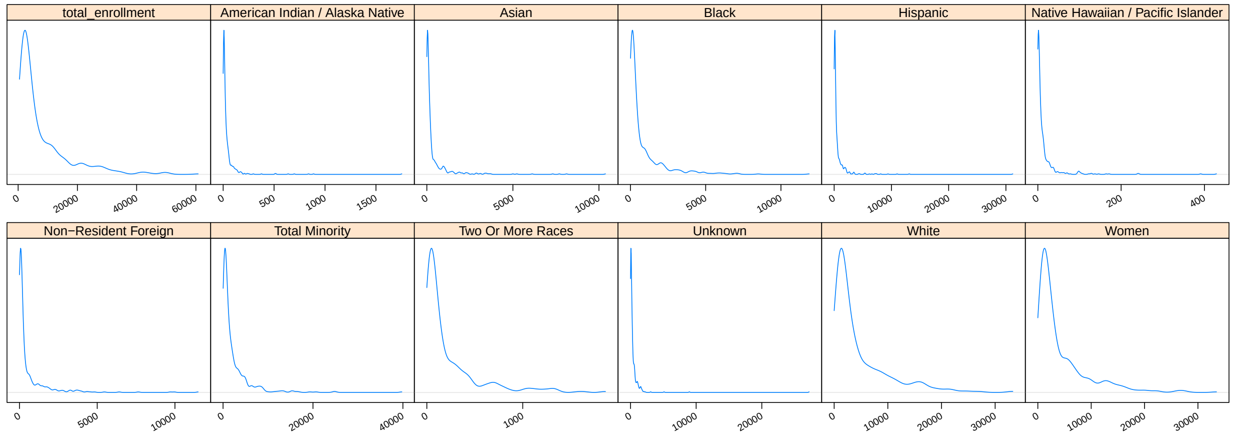
## Explanatory data analysis

The observations with missing values for the purposes of EDA are dropped. Nonetheless, the resulting data contains 598 observations/schools which is still adequately large to give clues about trends or relationships between variables in EDA.

The marginal distributions of the variables are:

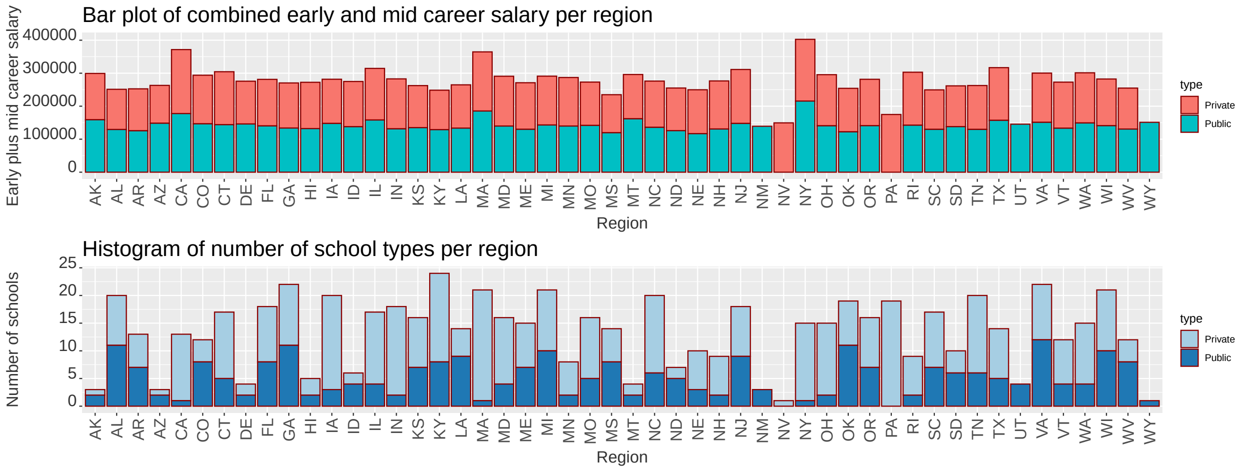


Most schools seem to have a stem percentage enrollment of approximately 15 per cent. Also, most of them are private, with a 4-year degree length.



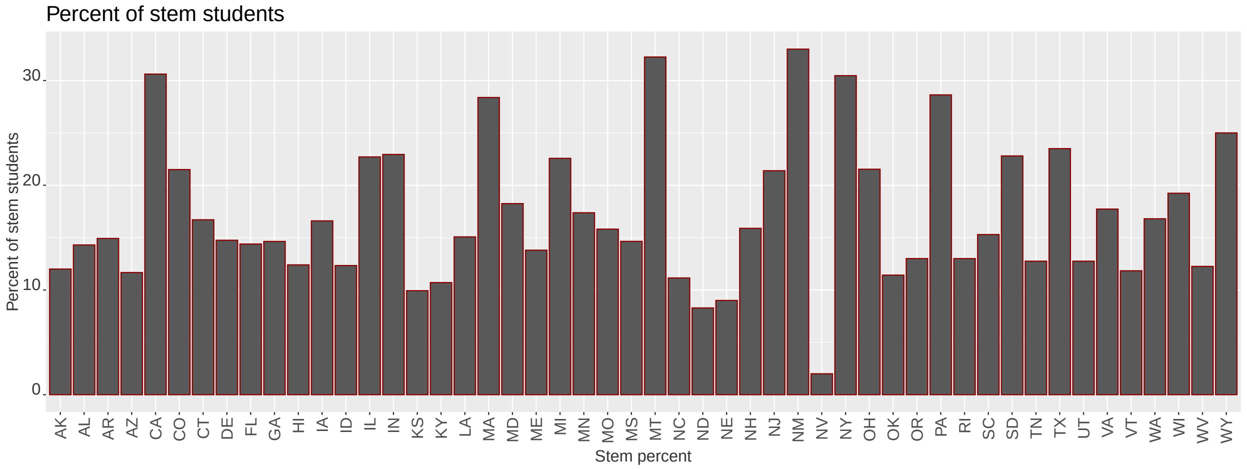
It is observed that the highest enrollments concern people identifying as White, and Two Or More Races. The former is perhaps justified since the White category is the most prevalent in the US.

Next, a plot of early plus mid career salary per region is presented, along with a plot of Plot of number of private and public universities per region:



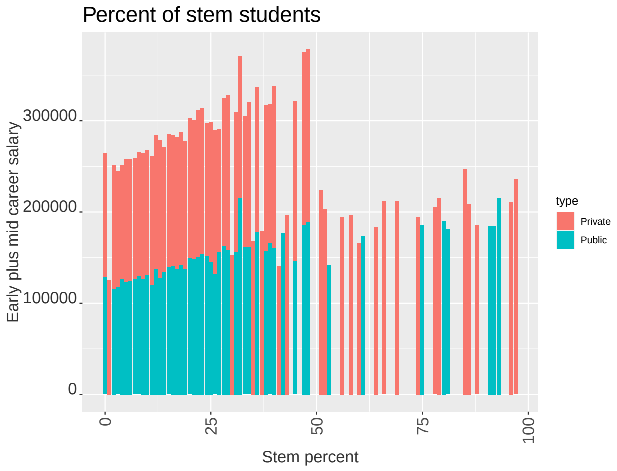
In most regions, the combined salary seems to be similar for each university types. Some regions such as MA and PA demonstrate high percentages of private schools whereas regions such as AK, AZ, and CO contain primarily public schools.

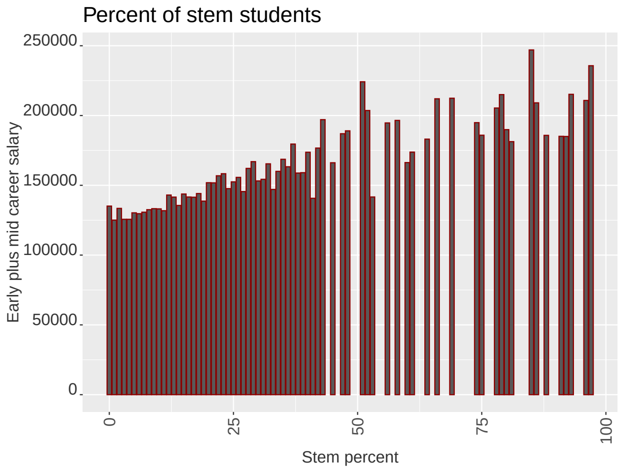
Next a plot of stem percent per region is presented:



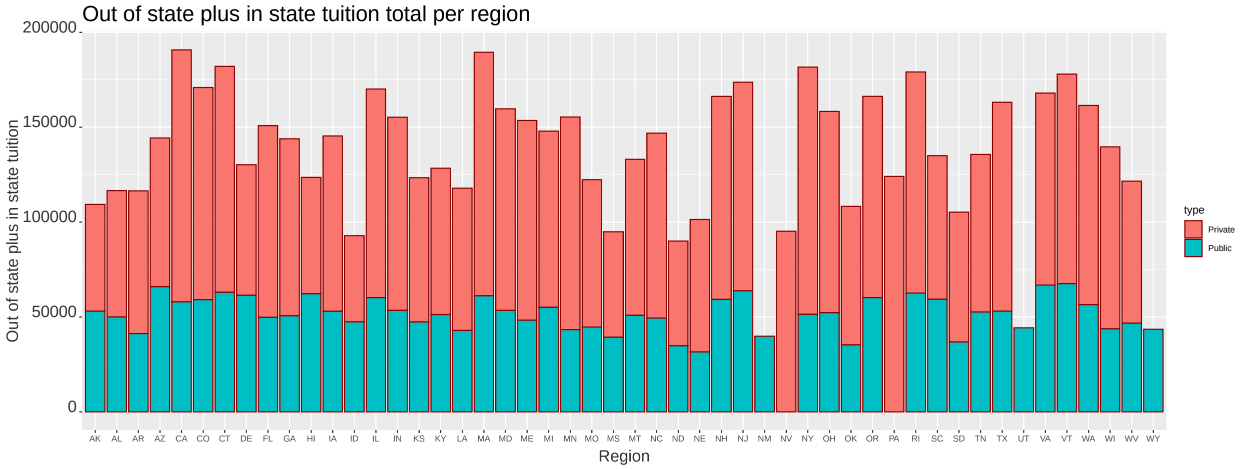
Most universities have step enrollments percentages less than 20.

Next, a plot of early plus mid career salary against stem percent per school type is presented, along with a plot of early plus mid career salary against stem percent without counting for school type, are presented:





Next, a plot of a combined in- and out-of-state tuition per region and per school type is presented:



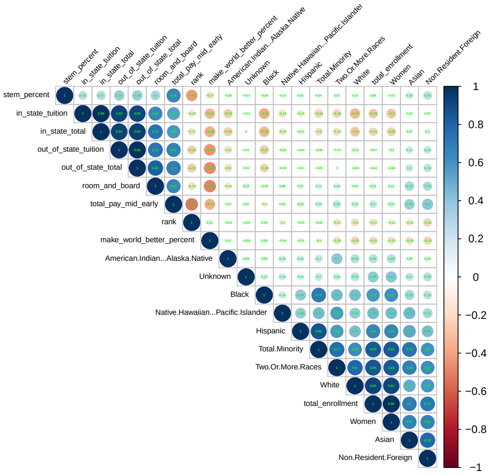
Private schools seem to charge higher tuition than public schools. Moreover, the highest public school tuition belongs to states such as AZ and VT. The highest private school tuition to states such as CA, PA and MA. Of course, these regions also contain high cost of living, so this could play a role in the tuition rates.

## Inference

In this section, inference will be made on the dataset by using ordinary linear regression. For this reason, highly correlated variables will be removed, since linear regression is susceptible to highly correlated variables (which give clues about colinearity and high Variance-Inflation-Factors(VIFs)). Moreover, the linearity and normality assumptions of the residuals will be check so that inference is reasonable. Further,the normality assumption is not necessary for good inference to be made since the number of observations is high enough. A ridge or elastic regression model could also be fit to deal with the colinearity issue during inference, however, their functions encode the factor levels as dummy variables internally, and they do not produce coefficients for every factor level in the output. These models will be used for prediction.

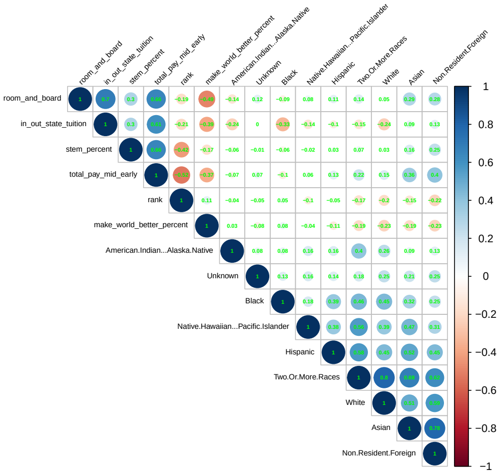
### Check for colinearity

The correlation matrix is:



There seem to be some highly correlated variables. These are `total_enrollment` with `Women`, `White` with `total_enrollment` and `Women`, `out_of_state_tuition` with `out_of_state_total`, `in_state_total` with `out_of_state_tuition` and `out_of_state_total`, `in_state_tuition` with `in_state_total` and `out_of_state_tuition` and `out_of_state_total`. The `total_enrollment` and the `Woman` variables highly correlated since the latter is dependent on the former. Therefore, the `total_enrollment` variable is removed since it is also the sum of the races variables. The `Total.Minority` variable is dependent on the minority races and therefore it is removed. Moreover, the `Women` and `White` variables seem to be correlated for some reason, therefore the `Women` variable is removed, so that `White` along with the other categories (e.g. `Black` etc) are left in the dataset.

The `in_state_total` is the sum of `room_and_board` and `in_state_tuition` and similarly for `out_of_state_total`. Therefore, the total variables are removed. Moreover, `in_state_tuition` and `out_of_state_tuition` are highly correlated. Therefore, they are added to form the `in_out_state_tuition` variable, so that their information is combined instead of just discarding information from one variable. After these changes, the correlation matrix of the resulting dataset is:



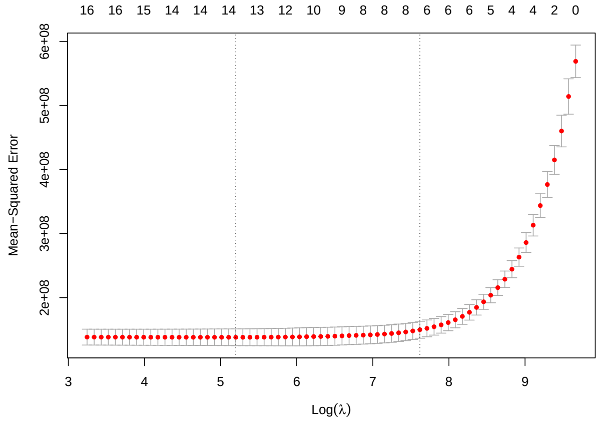
The variables seem to not be strongly correlated now.

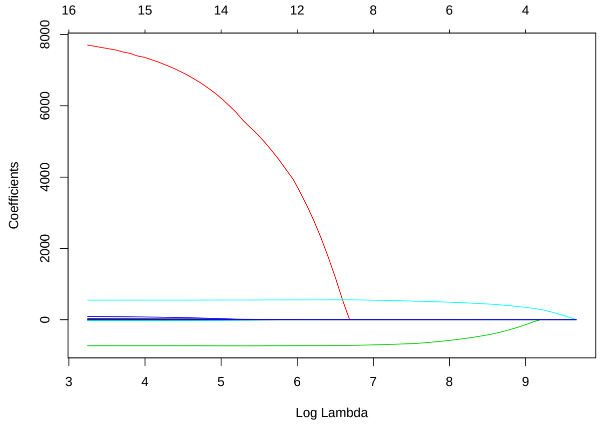
## Inference, after removing highly correlated variables

Initially a lasso model will be fit which reduces the effects of colinearity. Then a linear regression model will be fit, since it also gives coefficients for the factor levels (as opposed to a lasso fit through `glm`). Moreover, a linear regression is an unbiased estimator of the coefficients. Then the diagnostics will be checked for the linear regression model and it will be refit. The lasso model will also be refit. All of this is done for inference.

For brevity, a table of the significant, at the 0.05 level, coefficients will not be shown. Discussion on the coefficients will follow at the end of the section.

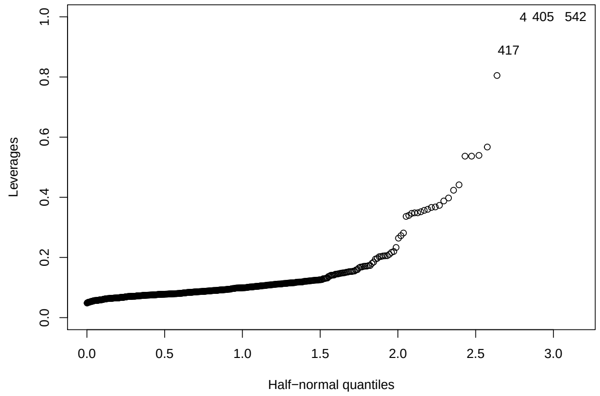
Therefore, first a lasso is fit which reduces the effects of colinearity. :

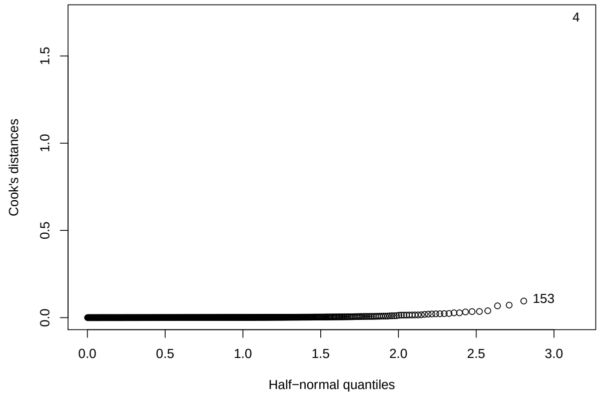




Then a linear regression is fit (all points with leverage one have been previously removed).

The number of predictors with a VIF larger than 10 is 36. The diagnostics of the linear regression model are the following:

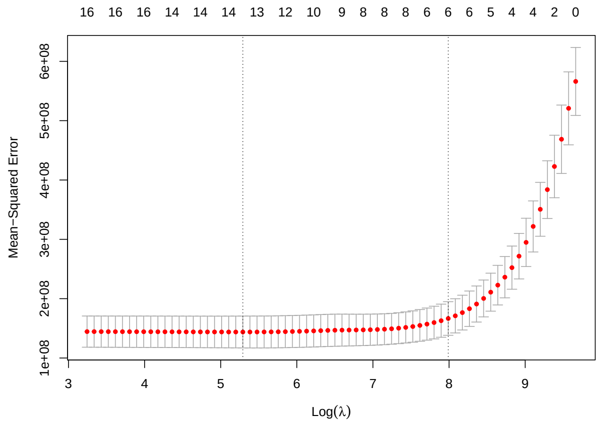


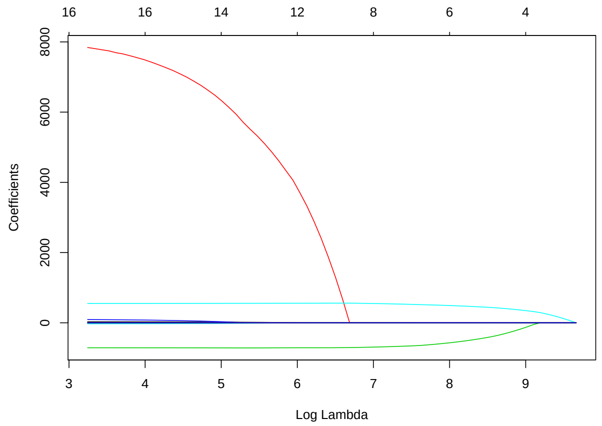


The studentized residuals are:

```
##   115     5   597   130   251
## 5.205 4.753 4.406 4.323 3.872
```

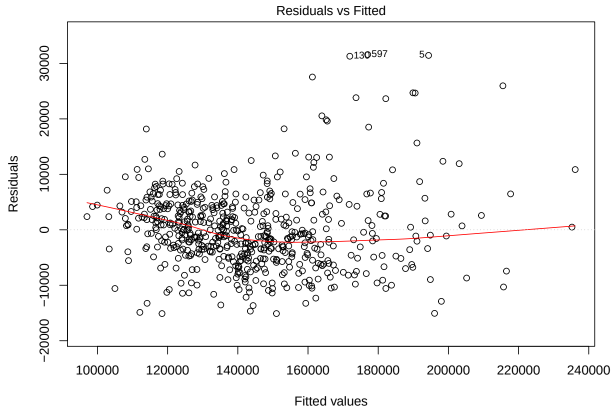
The extreme observations are 4, 405 and 452 all of which have leverage one. Moreover, observation 4 has the largest Cook's distance with the other observations being relatively close and less than one. Observation 115 has a studentized residual above five. Since these observations are few enough (four in total) they can be removed. Afterwards, the previous process is remade. That is is, initially a lasso is fit:

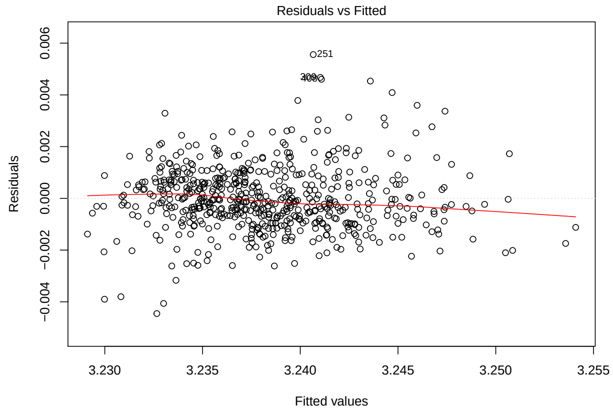




Next, a linear regression is fit.

The number of predictors with a VIF larger than 10 has now dropped considerably to 36. The linearity assumption is not verified too closely so boxcox is used, which includes $\lambda = -1/2$ (the boxcox plot is omitted for brevity). The first plot concerns the model before the boxcox transformation and the right plot the model after the boxcox tranformation:





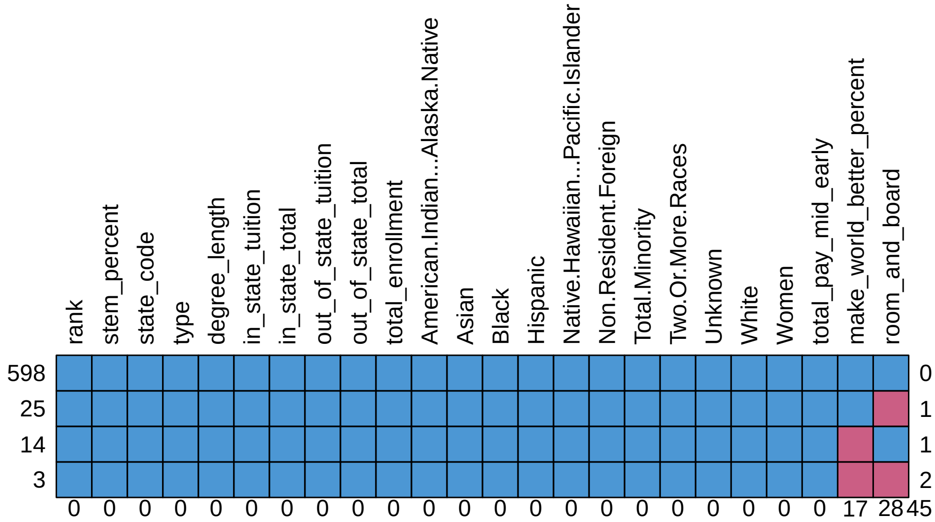
However, the coefficients of the box cox transformed model have the same sign therefore the same interpretation with the con that they are harder to interpret. Therefore, the non-boxcox model is chosen.

The linear regression model deems that stem percentage is significant to predict the salary, with the stem percentage having among the the lowest pvalues. This is to be expected from the plots which show an increase in salary as the stem percentage increases. Also rank, which gives the estimated rank w.r.t. salary of a graduate of a university (observation) in a particular region, has among the lowest pvalues, which is expected. Moreover, among the 30 regions that are deemed significant, there are most of the regions with high stem percentages in their universities, e.g. IL or MA. The type of the university is also significant. Moreover, how the percent of alumni who think they are making the world a better place is also significant. The tuition costs is significant, which is to be expected since more stem oriented schools have higher tuition costs. The White category is also significant, but probably this is because they represent most of the graduates in each region.

The lasso model chose the rank, the stem percent, room and board, Asian, Non Resident Foreign, and tuition costs. The first and second, as well as tuition, because of aforementioned reasons, were expected. Room and board was chosen probably because the higher they are, the more expensive the region and therefore the higher the salaries in order to compensate for the expensive region. Asian and Non Resident Foreign were also chosen. Regardless, the variables stem percent and rank have the highest absolute values by far compared with the others. It is noted that the nonzero coefficients for lasso were the same regardless of whether the leverage-one and highest-studentized-residual observations were included.

## Missing data

The dataset with highly correlated observations removed will be used. The missing values are:



### Filling missing values

The performance of four imputations is tested. These imputation methods are namely mean-value, stochastic regression, regression, and random-sample. Their performance is tested below:

```
## [1] 49249042
```

```
## [1] 49560768
```

```
## [1] 49476616
```

```
## [1] 49056813
```

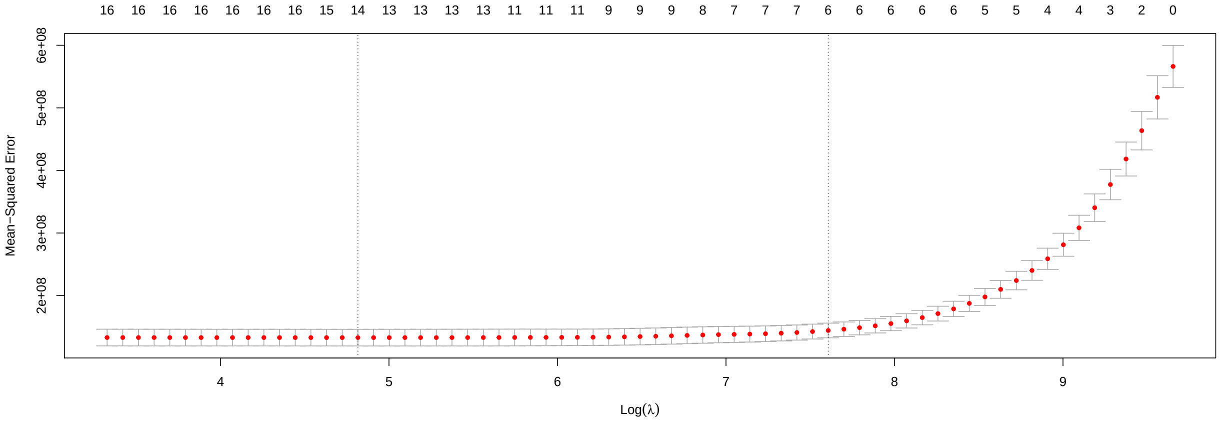
The sample method gives the better fit, and this dataset with the sample-filled values will be chosen for the rest of the experiments. The dataset is also split into training and testing data with a 80%/20% train/test split. The training data will be used for training and validation, and the testing data will be used for calculating predictions and testing performance through RMSE.

## Prediction with various models

The models under consideration will be Elastic Net, a Linear Regression model, KNN, a radial and a linear SVM, and Random Forests. The testing error measure is RMSE in all of the models.

### Linear Regressors: Elastic Net, and Linear Regression

An elastic net is fit using a 5-fold CV:



The best CV was selected by varying the $a$ parameter inside the set 0.00, 0.25, 0.50, 0.75, and 1.00. The best CV belonged to $a = 1$, i.e., a Lasso model. The minimum lambda belongs to a model with 14 parameters, and the 1 standard error lambda belongs to a model with 6 parameters.

Prediction on testing data, using `lambda.min` gives a test error of:

```
## [1] 1242
```

Prediction on testing data, using `lambda.1se` gives a test error of:

```
## [1] 727.9
```

Therefore, the model using lambda.1se will be used.

Next, a linear model is fit using both-ways stepwise AIC criterion.

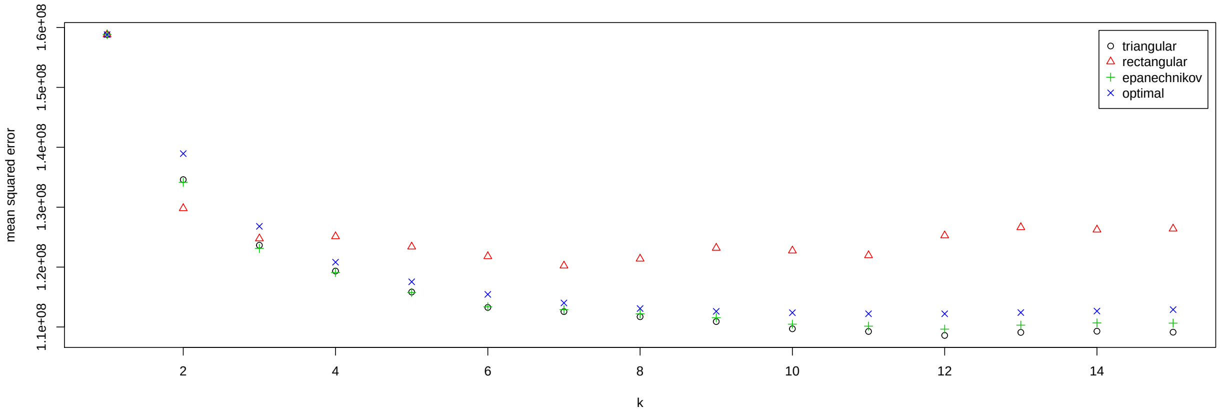
The best AIC model contains the rank, the stem percent, the room and board costs, the tuition costs, the make-world-better percent and ethnicities.

The best AIC model is crossvalidated. The test error is:

```
## [1] 1187
```

### KNN

Leave one out CV for KNN is performed, simultaneously testing for 1 to 15 neighbours and four kinds of kernels, namely triangular, rectangular, epanechnikov and optimal:



The best model is the model with the triangular kernel, and where the number of neighbours is 12. Overall, it is observed that the triangular and epanechnikov kernels yield similar performance, with the latter having slightly lower performance. The optimal kernel yields slightly lower performance than the latter too. The rectangular kernel, i.e., classical K-NN, yields the lowest performance.

Predictions on test using the best model data yields a test error of:

```
## [1] 2537
```

A knn model using 5-fold CV was investigated. However, the LOOCV model performed better, therefore the latter is chosen. The former is omitted due to space constraints.

### SVM

Next, SVM models are fit using 5-fold CV. Using radial kernel and varying cost and $\sigma$.

The mean squared error on the training data is:

```
## [1] 156853652
```

Predictions on testing data yield a testing error of:

```
## [1] 1504
```

Using linear kernel and varying cost with 5-fold CV:

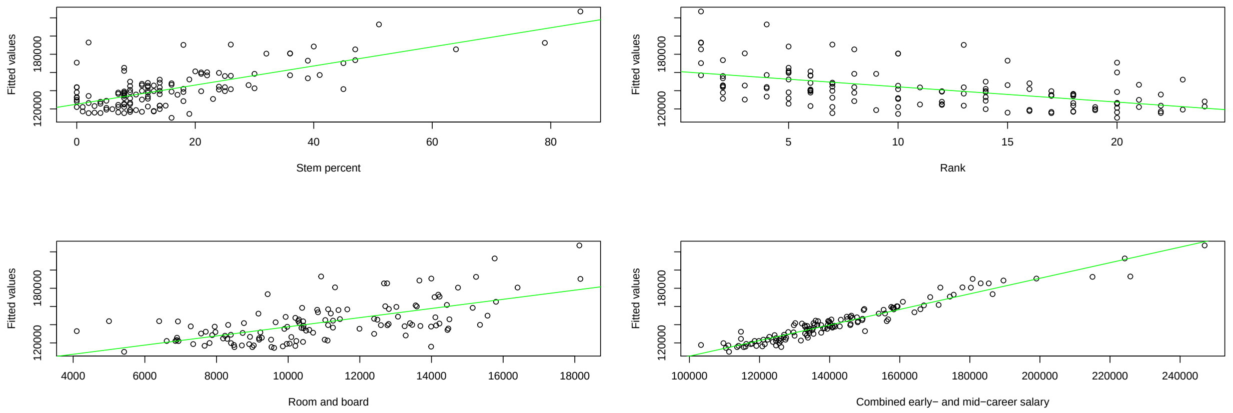
The mean squared error on the training data is:

```
## [1] 85547332
```

Predictions on test data yield a testing error of:

```
## [1] 762.4
```

Plots for SVM using the linear kernel:



The fitted value for the predicted value combined early- and mid-career salary, is fit adequately. The trend demonstrated in the rest of the plots are expected from the Inference section, and from the sign of the coefficients specifically. That is, as the stem percentage and the room-and-board costs increase the estimated salary increases. As rank increases the estimated salary decreases.

It can be observed that the significant variables that appeared during inference (i.e. linear regression and elastic net), have an adequate fit. These variables are step percent, rank and room-and-board. Therefore, the superior (as will be shown) performance of the Linear Kernel SVM , with respect to the other models, could be explained by the adequate fitting of the most important variables.

### Random Forests

A random forest model is fit below using 5-fold CV, varying `mtry`, `min.node.size` and `ntree`.

The minimum MSE is:

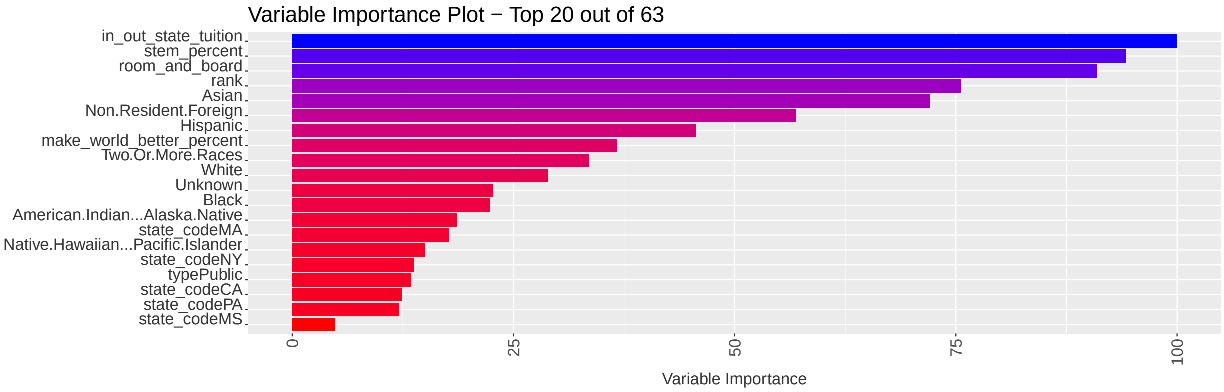
```
## [1] 200602772
```

for the model with `ntree` of 500 , `mtry` of 2 and minimum node size of 10.

Predictions on test data yield a testing error of:

```
## [1] 1064
```

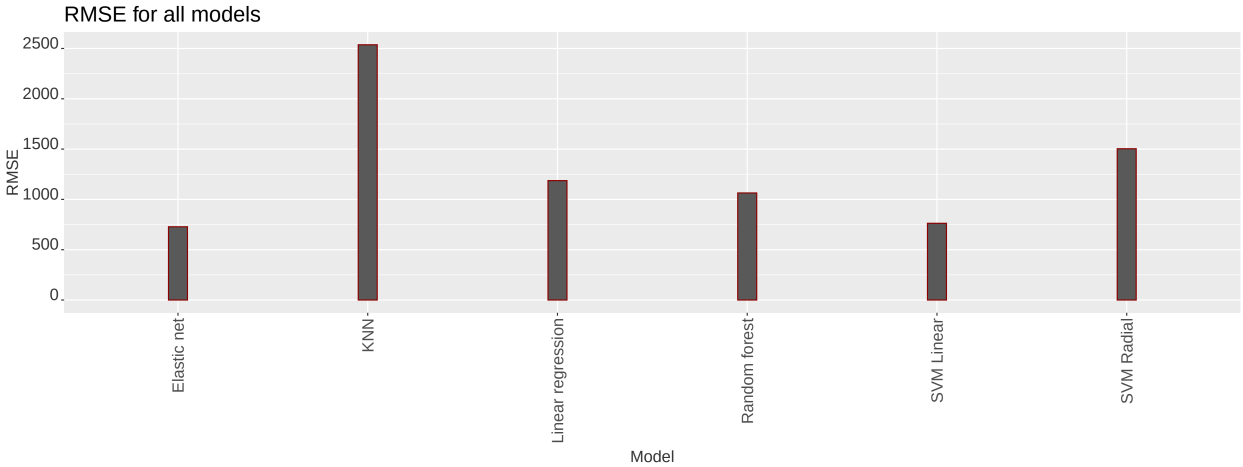
A variable importance plot is shown below:



As expected tuition costs, room and board, stem percent and rank are the most importance variables. The latter two possibly because higher living costs yield higher tuitions but also higher salaries. Stem percent means higher salary on average. Also, the way the variable rank is coded in this dataset, lower estimated-salary-rank means higher expected salary. The method for selecting the important variables was to check the mean decrease in node impurity for each variable at each level of the tree.

## Summary

The testing RMSEs for all models are:



The best performing model is the Elastic net model. The second best model is SVM Linear, whose performance is close to the best performing model. The third best model is the Random Forest Model. The fourth best model is the Linear Regression model. The fifth best performing model is the SVM Radial model. The sixth, and last, model is the triangular-kernel Nearest Neighbours model.