Data Science Practical: Case studies

Starting salary of law graduates from higher ranked schools

Group 5 *

February 4th 2022

Abstract

It is known that college graduate starting salaries differ per student, even for students who graduated in the same major. Which factors motivate this difference? We know the college a student graduated from might affect the starting salaries, but does a high ranked school actually mean a higher starting salary? And which other factors come into play? We intend to find explanatory models for the starting salary of law school graduates by using Ordinary Least Squares regression. We compare these models with another regression method: the K-th nearest neighbor regression. Uncertainties about the results are thereafter studied by means of a Monte Carlo simulation. We will use data from The Official Guide to U.S. Law Schools, 1986, Law School Admission Services, and The Gourman Report: A Ranking of Graduate and Professional Programs in American and International Universities, 1995, Washington, D.C. Furthermore, this document is written with the help of LaTeX.

Keywords: Data Science Practical, higher salaries higher ranked schools, Ordinary Least Squares, K-th nearest neighbor, Monte Carlo, Law School 1995, LaTeX.

1 Introduction

Nowadays, every student wants to know what their starting salary will be. Since, most students start looking for work after obtaining their degrees, one of the most challenging aspects is getting an agreement on the starting salary. The starting salary may certainly depend on many different factors such as the type of major pursued, overall intelligence and the quality of the schools the student graduated from. We intend to investigate if such a relationship can be found in a data set of law school graduates from the United States. We expect that graduating from a higher ranked school is positively correlated with the starting salary. We suspect this due to the fact that having graduated from a prestigious school, like Harvard, has an immense impact on a student's life in today's society. But what is perhaps even more important is the value related to the name, so the rank of the school. However, other variables may also notably affect the starting salary. The following research report reflects our statistical, machine learning and data science knowledge. The main goal of our analysis is to build an appropriate model for estimating the starting salary from American law school graduates and the school they graduated from. Our report is structured as followed: First we will disclose more information about our data, by briefly explaining the main characteristics of the data. After that, we will reveal all the methods

*Guus Bouwens (2701442), Antonio José Guerra-Librero Reja (2696947), Jaspreet Singh (2710547), Shabnam Alizada (2666615)

and techniques we use during our research and explain several key concepts. For example, the Ordinary Least Squared regression and the K-Nearest Neighbour regression. We also carry out a Monte Carlo simulation to check certain uncertainties found during our regression analysis. Furthermore, we will show our results and discuss the considerations we have made. Last but not least, in the last section we will summarize our findings and give our final conclusion.

2 Data

The source of our data is as follows: Law School Admission Services, "The Official Guide to U.S. Law Schools," LAWSCH85 datasheet, March 1, 1986 [Revised Jan. 2022]. Gourman, J., "The Gourman Report: A Ranking of Graduate and Professional Programs in American and International Universities," LAWSCH85 datasheet, March 1, 1995 [Revised Jan. 2022].

2.1 Characteristics of the data

During our research, we will make use of dataset LAWSCH85 (Law School Admission Services, 1986; Gourman, 1995). The dataset contains information regarding 156 law schools from The United States. It contains the following variables:

variable Name	explanation	type of variable
rank	law school ranking	numerical variable
salary	median starting salary	numerical / dependent variable
cost	law school cost	numerical
LSAT	median LSAT score	numerical
GPA	median college GPA	numerical
libvol	no. volumes in lib., 1000s	numerical
faculty	no. of faculty	numerical
age	age of law sch., years	numerical
clsize	size of entering class	numerical
north	=1 if law sch in north	categorical / geographical
south	=1 if law sch in south	categorical /geographical
east	=1 if law sch in east	categorical / geographical
west	=1 if law sch in west	categorical / geographical
lsalary	log(salary)	numerical / log(dependent variable)
studfac	student-faculty ratio	numerical
top10	=1 if ranked in top 10	categorical
r11_25	=1 if ranked 11-25	categorical
r26_40	=1 if ranked 26-40	categorical
r41_60	=1 if ranked 41-60	categorical
llibvol	log(libvol)	numerical
lcost	log(cost)	numerical

There are a lot of missing values for different types of variables and even though the data is produced in 1995, we assume it still holds true. As an indicator of the quality of law schools, we use the rank of the school, which is also partly divided into the different categories top10, r11_25, r26_40 and r41_60 as dummy variables. The rank of a law school is, among other things, determined by the median LSAT score and median GPA of its students. It is for example difficult to enter higher ranked schools with a low LSAT score. The LSAT (Law School Admission Test) score ranges from 120 to 180, with 180 being the highest and the GPA ranges from 0 to 4.0

with 4.0 being the highest. We use the LSAT and the GPA as indicators for the achievements of students. The amount of books in the school library is used as a indicator for knowledge among the students of a law school. Besides the basic salary, cost and libvol variables, the dataset also contains the logarithmic functions of these variables, called lsalary, lcost and llibvol.

3 Methodology

In our investigation, we calculate and compare all tests and statistics w.r.t. a 5% significance level.

3.1 Incomplete dataset

The first problem we come across are the missing values in the dataset we use to research our hypothesis. Unfortunately, deleting the data with missing values would lead to non statistically significant results, because of the small amount of input data. That is why we use different approaches to fill in the missing values in our data.
For our first approach of filling in the missing data, we use mean imputation. For this method we calculate the mean for the non-missing values and substitute this into the missing values. This would prevent the problem of reducing our data size. Unfortunately, there are some disadvantages to this method. Standard deviation and the variance of the variables, for which mean imputation is used to fill in the missing values, are biased. If we fill in the mean for values which would have been much more distant from the mean, it would result to an underestimated SD. Our second approach, is median imputation. This implies substituting the calculated median over the non-missing values, into the missing values. The disadvantages are nearly as big as using mean imputation. In case of outliers, median imputation would be a better choice. Both approaches ignore the relationship between variables, which decreases the correlation between them and so using these two creates another bias (Salgado et al., 2016).
Our third approach, for filling in the missing values, is linear interpolation. The equation for the linear interpolation function is (Chapra and Canale, 1998)

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0} (x - x_0)$$

Linear interpolation determines missing values by taking the average of the two adjacent points. If there is no earlier point, it copies the value that comes after. If there is no subsequent point, it copies the earlier value. If there are no values concurrently, it uses the earliest neighboring points. We believe that linear interpolation will protect the statistical performance of the data, or at least not degrade it as proven in *"Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set"* (Mohammed Noor et al., 2014).

3.2 Choosing between OLS and KNN (interpretability vs flexibility (simplicity))

For the regression of the model there are two options: Ordinary Least Squares (OLS) or K-nearest neighbor (KNN) regression. For both methods there are a mutual concessions and compromises:

3.2.1 KNN (dis)advantages

One of the main advantages of the KNN is that there a no assumptions about the shape of the distribution of the population, from which the data is drawn, as it is a non-parametric algorithm (Alo, 2019). So there a no parametric conditions which have to be fulfilled. Besides that, it is a very simple algorithm, easy to understand and easy to interpret Mohammed Noor et al. (2014). Besides its advantages, there are also some disadvantages. One of the problems which arise with using KNN is its high memory requirement, as all of the training data must be present in memory in order to calculate the closest K neighbors. It is also known as a slow algorithm, since we're computing the distances for all the K closest points in the testing phase. Its testing phase has a worst case run time of $\mathcal{O}$(n*k*d). Furthermore, the algorithm is very sensitive to irrelevant variables and outliers in the data.

3.2.2 OLS (dis)advantages

The OLS regression model is a linear model, for which the Ordinary Least Squares method is used to estimate the parameters of the model. The model reveals many statistics about the model and information for all variables. Unlike KNN, it is also much faster.
However, the model has a lot of limitations. The data has to meet some requirements before the OLS regression model can be used, like linearity. There has to be a linear relation between the dependent and each independent variable in the model. No multicollinearity, high correlation between two or more independent variables. Homoscedasticity, the error term of an independent variable having a constant variance across different samples in the data. If the assumptions of the linear model are violated, then the results of our hypothesis tests and confidence intervals will be inaccurate. The errors having a normal distribution is an optional condition Furthermore, OLS is sensitive to functional form, whenever the error term is not correctly specified. The OLS regression model needs a large data set to achieve reliable results and is very sensitive to outliers.

3.3 OLS Regression

After completing our dataset, we start with a simple linear regression. We add different independent variables one by one to our model and compare all possible regression models, by mostly looking at (adjusted) R-squared, the t-values and P-values per variable. The R-squared for the explainability of our model. The t-test for significance of the coefficients and F test for model significance. Besides comparing the descriptive statistics, we also check the conditions of the OLS for every combination of independent variables. For example, adjusted R-squared, VIF and condition number for detecting collinearity. Checking density histograms and also doing the Jarque-Bera test for normality. To detect outliers, we look at the plots and compare all regular variables and their log transformed plots. For linearity we plot the residuals vs the explanatory variables and also perform the Ramsey RESET test. For heteroskedasticity we check the variables their regular plots, the model its residuals and we do the Breush-Pagan test. We will explain these tests in more details.

3.4 Checking for the assumptions

Assumptions:

1. Linearity (linear in parameters)
2. Random sampling
3. No multi-collinearity (or perfect collinearity)
4. Exogeneity/Endogeneity
5. Large outliers
6. Homoskedasticity
7. Error terms should be normally distributed.

With 1 or more of these assumption violated, the results of the regression can not be trusted. To control these assumptions as much as possible, some cautions can be taken:

3.4.1 Linearity

Using a linear regression model to detect the relationship between variables, which have no linear relationship, leads to an underfitting model. That is why the linearity assumption is one of the most important conditions, which have to be present. To check for linearity, we can apply different methods.

Ramsey Reset Test
The Ramsey reset (RR) test checks whether we have a functional misspecification e.g whether the linearity is too restrictive. We do this by adding a non-linear combinations of fitted values and test whether we get more explainability of our dependent variable.

$$y_i = \beta_1 x_1 i + \beta_2 x_2 i + ... + \beta_k x_k i + \gamma \hat{y_i}^2 + \varepsilon \tag{1}$$

$$H_0 : \gamma = 0 \; H_A = \gamma \neq 0 \tag{2}$$

In case the p-value is bigger than 0.05, we fail to reject the null hypothesis and fail to detect any misspecifications (non-linearity). In case the p-value is smaller than 0.05, we reject the null hypothesis and this indicates a non-linear relationship in our data. This means the linearity assumption is violated.

3.4.2 Random sampling

The second assumption is random sampling. This assumption states that the sample used for our regression model, so our data, must be drawn at random. So every sample must have equal probability of being used. Of course, we can not test this, but using logical reasoning we understand the data we used can not be selected randomly. Our data set contains the 175 highest ranked law school, which could lead to very different results, with respect to the relationship between starting salary and the quality of the school, than a data set with randomly chosen law schools. The 175 highest ranked schools are all very close to each other in space, which is a source of non-independence. Also, one school its rank has to depend on the rank of the other schools. So our sample does not consist of random, independent draws. This does not lead to any problem, as the error term is uncorrelated with the explanatory variables (Woolridge, 2013).

3.4.3 Multicollinearity

Whenever two variables that are highly correlated are both included in a regression model, one of them will be insignificant, as they are both providing a lot of the same information. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems :

1. redundancy
2. the estimates of independent variables on the dependent variable will cause to be less precise
3. the standard error of the independent variable, which causes the collinearity, will increase. This can yield a Type II error, which means that we fail to reject the null hypothesis because it is not significant, while it does turns out to be significant.
4. Overfitting: A good model consists of independent variables which all have an unique affect on the dependent variable. When there is multicollinearity, the model is overfitted.

The Variance Inflation Factor
The Variance Inflation Factor (VIF) is how much the variance of your regression coefficient is larger than it would have been if the variable had been completely uncorrelated with all the other variables in the model. A value of 1 would mean completely uncorrelated. A rule of thumb for (perfect) collinearity is that if the VIF is bigger than 10, it is too much. A large VIF in the constant indicates that the regressors also have a large constant component. This would happen when a variable has a large mean, but only a small variance. Finding perfect collinearity happens commonly with the dummy variable trap, when one of the levels of a categorical variable is not removed and thus the dummies sum to 1 and, therefore, replicate a constant. So including the intercept in your VIF calculation is important. Standardized explanatory variables do not show higher values.

Pearson correlation matrix
To see in what magnitude certain variables have linear relationships, a correlation matrix can be used. It can give a clear view of which variables depend on each other. The sign shows which way the relationship goes.

Condition number
The condition number measures the sensitivity of a function its results to its (data) input. When two predictor variables are correlated, the coefficients of those regressors can differ greatly for small changes in the data. A bigger value for the condition number would imply some form of multicollinearity. A value below 1000 would be well conditioned. The smaller the condition number, the better the model is conditioned in this perspective (Belsley, Kuh, and Welsch, 1980).

3.4.4 Conditional mean should be zero

Another assumption of the OLS is that the conditional distribution of the residuals, given the independent variables, should have a mean of zero. Mathematically written as $E(ui/X_i) = 0$. This assures that the other factors, which are included in the error term, are unrelated to the independent variables. This assumption implies $Cov(X_i, u_i) = 0$ (not the other way around). If the assumption holds true, X_i is called exogeneous. Otherwise X_i will be endogenous. If the conditional mean is too far from zero, it means that something you did not account for has an effect you did not predict.

Best Linear Unbiased Estimators
If all of the above mentioned assumptions are true, we will have Linear Unbiased (and consistent) Estimators for the parameters ($E(\hat{\sigma}^2) = \sigma^2$). To find the Best Linear Unbiased Estimators (BLUE) $\hat{\beta}_i$ for β_i, with i being the amount of explanatory variables, we have to find the estimators with the lowest variance. This will hold true for one more assumption being correct (homoscedasticity), according to Gauss Markov Theorem's (Woolridge, 2013).

3.4.5 Homoscedasticity

When the variance of the conditional distribution of the error term given X_i is not constant, heteroskedasticity occurs. Heteroscedasticity reduces the precision of the estimates in the OLS regression. The reasoning for this is simple: the standard errors will be biased. Hence, the model is not well defined. Adding more variables can help explain the performance of the dependent variable.

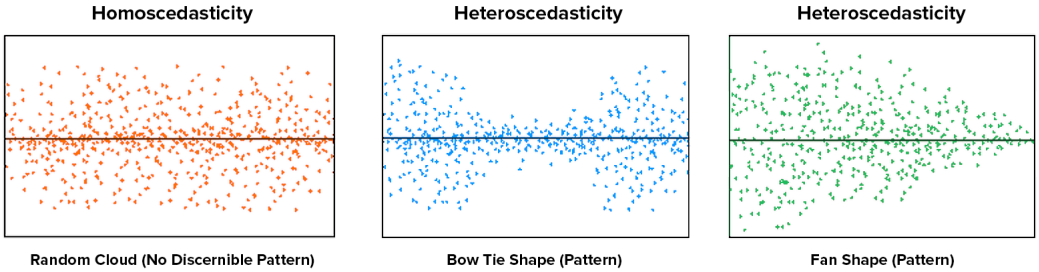


Figure 1: Homoscedasticity vs Heteroskedasticity

Check residual plots
To see whether a pattern persists for a certain variable, we can check the residuals plot of that variable, as seen in figure 1.

Breush-Pagan test
The Breush Pagan (BP) Test tests whether the variance of error term depends on a certain variable Z_1

$$H_0 : Homoscedasticity \; H_A = Heteroskedasticity \tag{3}$$

Whenever the p-value is bigger than 0.05, we can not reject the Null Hypothesis, which means $var(u/x_1, .., u_k) = \sigma^2.$, with x_i being the explanatory variables. In case the p-value is smaller than 0.05, we reject the Null Hypothesis. Then there is heteroscedasticity.

Classical Linear Model (CLM) assumptions
The classical linear model assumptions contains all of the Gauss Markov assumptions (one to five) and one plus assumption: normality. According to Woolridge (2013), the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ will have a stronger effieciency property under the CLM assumptions than under the Gauss-Markov assumptions.

3.4.6 Normality

The last assumption we check is if the error is normally distributed with mean zero and variance σ^2. If normality is present, assumption 4 (exogeneity) and assumption 5 (Homoscedasticity) will

be automatically true:

$$E(u|x_1, ..., x_k) = E(u) = 0 and Var(u|x_1, ..., x_k) = Var(u) = \sigma^2 \tag{4}$$

Jarque-Bera test
The Jarque-Bera (JB) test considers asymmetry and kurtosis, in which the null hypothesis considers a normal distribution of the residuals. A JB value of 0 and a probability of 1 would indicate that the model, including its error term, is normally distributed.

$$H_0 : Normally\ distributed\ H_A = non - normal \tag{5}$$

A p-value bigger than 0.05 indicates the null hypothesis can not be rejected, while a p-value smaller than 0.05 means the opposite.

Density histograms
A density histogram shows how frequent the variable its data is distributed over the whole column. Whenever this is compared with how a normally distributed variable behaves, we can see whether the variable throws off the normality of a model.

3.4.7 Large Outliers

Large outliers lead to incorrect OLS regression results, as OLS is very sensitive to these outliers. That is because the OLS results weigh each pair X,Y, so large outliers can affect the slope of the regression line greatly. There are different ways to deal with outliers. For example by scaling the data.

3.4.8 Significance

Testing significance is not an assumption, but still very important for selecting our model.

R^2 (adjusted)
R^2 is a statistical measure of how well the regression line approximates the data points: what fraction of the variance of the dependent variable is explained by the independent variables. The problem with R^2 is that it does not take into account the number of regressors. In theory, the model is always explained equally as good, or better, by adding more regressors. However, the estimate becomes less accurate, if the additional regressors have little influence on the dependent variable. This is why we also look at the R^2 adjusted. Adding regressors does not necessarily lead to higher R^2 adjusted. Only when the variance falls, so the addition of a regressor is actually useful, only then will the adjusted R^2 rise. Then the added regressor adds value to the model and creates a better fit (Stock and M.W., 2015).

T-test
The t-test checks whether the variable is significant to a model. Doing this for all variables in your model individually, can show if your regression is correctly explained. This test can be done by performing a hypothesis test for a given variable.

$$H_0 : 0\ vs\ H_A \neq 0 \tag{6}$$

To derive the t-test we first need to calculate the t-statistic. For a 95% double sided test: 0.05/2 for the number of residuals this model has n number of observations-number of variables,

including dependent regressor. Then the formula for the t-test is:

$$t = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \tag{7}$$

If the found value is smaller than the t-statistic, H_0 holds, otherwise the alternative is favourable: so the variable ought to be significant for this model. If the value is negative it means the effect on the model is in the opposite direction (Casella and Berger, 2002).

F test
The F value and Prob(F) statistics test the overall significance of the regression model. Specifically, they test the null hypothesis that all of the regression coefficients are equal to zero versus at least one coefficient is non zero:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \; vs \; H_A : \text{ at least one } \beta_j \text{ is nonzero} \tag{8}$$

This tests the full model against a model with no variables and with the estimate of the dependent variable being the mean of the values of the dependent variable. The F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares $\frac{MSR}{MSE}$. Its value will range from zero to an arbitrarily large number.
The value of Prob(F) is the probability that the null hypothesis for the full model is true (i.e., that all of the regression coefficients are zero). For example, if Prob(F) has a value of 0.01000 then the chance that all of the regression parameters are zero is 1 in 100. Such a low value would imply that at least some of the regression parameters are nonzero and that the regression equation does have some validity in fitting the data (i.e., the independent variables are not purely random with respect to the dependent variable) (Stock and M.W., 2015).

3.5 K-th Nearest Neighbor Algorithm

To check whether our OLS results can be improved, we use K-th nearest neighbor (KNN) regression. For KNN to work for our data set, we have to split the data set into two sets. The training and test set. The training set uses our model with the estimates. The test set uses our x-value predictions. The y values are the same as in training set. This is done such that we can find: $\hat{y}$ - y = prediction - true value. If the training set is too small, our estimates may be unreliable. because, when there are too little points to conclude your regression from, your predictions will be biased. If the training set is too large, our predictions will be imprecise since our test set is too small. A small test set renders it unable to accurately check your results from the training set. To achieve optimal results for our model, we will use a balance of 70 and 30% for the training and test set, respectively.

The first step is to calculate the distance between the new point and each training point. There are various methods for calculating this distance, of which the most commonly known methods are Euclidean, Manhattan (for continuous) and Hamming distance (for categorical).

Euclidean Distance: Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (y).

$$\text{Euclidean Distance } = |X - Y| = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{9}$$

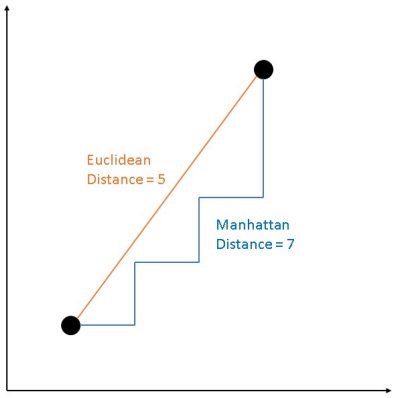


Figure 2: Euclidean vs Manhattan distance

Manhattan Distance: This is the distance between real vectors using the sum of their absolute difference.

$$\sum_{i=1}^{k} |x_i - y_i| \tag{10}$$

Hamming Distance: is used for categorical variables. If the value (x) and the value (y) are the same, the distance D will be equal to 0. Otherwise D will be equal to 1.

$$\begin{aligned} D_H &= \sum_{i=1}^{k} |x_i - y_i| \\ x = y &\Rightarrow D = 0 \\ x \neq y &\Rightarrow D = 1 \end{aligned} \tag{11}$$

Once the distance of a new observation from the points in our training set has been measured. The next step is to select the k value. This determines the number of neighbors we look at when we assign a value to any new observation. Based on the k value, the final result tends to change. We intend to find the optimal value if k by calculating the error for our train and test set. As minimal error term for any prediction model, is the goal. For a very low value of k, say 1, the model overfits on the training data, which leads to a high error rate on the validation set. On the other hand, for a high value of k, the model performs poorly on both the train and test set. The test error curve reaches a minimum at some value of k. This value of k is the optimum value of the model.

3.6 Fixing violated assumptions

Some violations can be fixed. The following are relevant to our investigation.

3.6.1 Multicollinearity

How to deal with multicollinearity:
Linearly combine the independent variables, such as standardizing two similar variables and add

them together.

3.6.2 Heteroscedasticity

To fix heteroscedasticity:
As we do not know the specification of the heteroskedasticity, we can not use weighted least squares. Hence we use heteroscedasticity-consistent standard errors to minimize the impact of inaccurately specified residuals. Heteroskedasticity-robust requires large n for consistency. In addition, there have to be few assumptions valid such as exogeneity, random sampling and no large outliers.

$$\text{Var}(\hat{\beta}) = \frac{\sum_{i=1}^{k} \left[(x_i - \bar{x})^2 \sigma_i^2 \right]}{\left(\sum_{i=1}^{k} (x_i - \bar{x})^2 \right)^2} \tag{12}$$
$$\text{where } \text{Var}(u_i \mid x_i) = \sigma_i^2$$

Another way would be to transform the dependent variable. One common transformation is to simply take the logarithmic function of the dependent variable.

3.7 Monte Carlo simulation

To check on some uncertain events regarding your regression results, a technique called Monte Carlo simulation can be used. It predicts a couple of outcomes based on a set of fixed input values. So it builds a model of potential outcomes by taking a probability distribution for any variable that is fundamentally uncertain. This process can be repeated thousands of times within the same random set of values to get a large number of likely outcomes. Finally, it shows a range of possible outcomes with the frequency of each result presented.
In our case we are mostly interested in what happens when there is a normally distributed model with non-normally distributed residuals (ϵ). Different distributions can be used for the residuals. Using distributions that are similar to your results and defining your simple model on basis of your regression results can be advantageous. So using constant values for values which are in practice hard to determine, like B_1, makes this simulation useful. As the same model runs over and over again, while maintaining the values you are not interested in, the behaviour of the value you are interested can be deciphered. The simple linear model looks like:

$$Y = B_0 + B_1 * X + \epsilon \tag{13}$$

4 Results

4.1 Descriptive statistics

After seeing the data set, we perceive plenty of missing values. Most of the variables have around 8 missing values, except age which has 45. If we delete all the rows of missing values, we will lose 42.3 percentage of our observations and that will make our investigation unreliable, as we already have little data to work with.
Furthermore we notice from (9) that not all the ranks in the range of 1-175 are included, so we are still missing in total 21 ranks. Besides the missing ranks, we have two duplicate ranks (16 and 17) in the r11_25, but they do have different values for the other variables, so schools can share a rank.

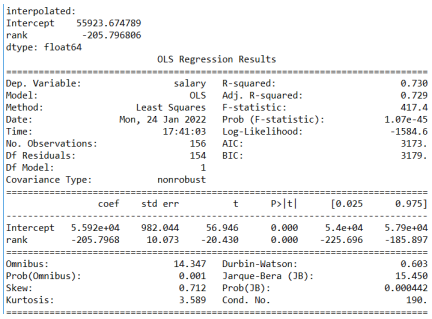
Now we will briefly analyze our dependent variable. The histogram 9a of salary shows us that

there are 3 mods i.e it's a multi-modal distribution, which can identify categorical variable of "rank" as high rank , medium rank and low rank. Further, salary is right skewed, which indicates that the right tail is heavier than the left tail. This is due to outliers on the right tail (school ranks with very high starting salaries), this is visible in the boxplot 9a. We make use of the log transformation of salary, because salary tends to be highly skewed. This ensures that there are fewer outliers and data is more normally distributed, 9b. We can also claim this based on the standard deviation. After transforming to log, it has become more precise. Besides, it also ensures that the correlation has increased between rank and salary. If we take the log transform then the correlation goes from -0.85 to -0.90 7, so more magnitude to the relationship. That's why we mainly use "lsalary" for our response variable.
Furthermore, (7) lsalary is perfectly negatively correlated with rank (higher lsalary leads to a higher rank), with the variables GPA, LSAT and libvol it's strongly positive correlated and with cost, age it is averagely positively correlated.

Finally, we will describe the variables that most relevant do this investigation. First of all, the variables LSAT, GPA, cost, grade and age approximately follow a symmetric distribution(8), The variable llibvol is moderately skewed. Other variables are highly skewed.
Furthermore, all the positive values(8) of excess kurtosis indicate that there are extreme values as outliers and the negative value indicates that there are few outliers.
Besides the transformation of salary, there are also two other variables for which we can consider taking the log transformed form. It seems wise to take log transformation for libvol and cost, since they are both highly skewed(8). You can also notice from the box plot that taking log reduces outliers12a,10a, . Furthermore, LSAT and GPA also have few outliers11a, 10b, but these variables contain a minimum and maximum of their own statistics (refer). To measure central tendency of a variable the median will be more accurate, since the mean of all the variables will be affected by these outliers; the median is more robust to outliers. Furthermore about the correlation 7, rank is negatively correlated with all the variables. For GPA, LSAT and salary even strongly. For cost, libvol and faculty they are averagely positively correlated. And last but not least, LSAT and GPA could be collinear, as they are strongly positively correlated. This can cause violation of an assumption.

4.2 Missing data interpolation

To choose between either deletion, mean, median or the linear interpolation method we compare the OLS regression results for salary with as only regressor the rank. We are mostly interested in the explainability and significance. Before interpolation the data is sorted by rank, this is because otherwise the linear interpolation method will not derive with the correct adjacent points. The best method is linear interpolation as it has the highest R^2 and the least amount of outliers.



```
interpolated:
Intercept    55923.674789
rank          -205.796806
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.730
Model:                            OLS   Adj. R-squared:                  0.729
Method:                 Least Squares   F-statistic:                     417.4
Date:                Mon, 24 Jan 2022   Prob (F-statistic):           1.07e-45
Time:                        17:41:03   Log-Likelihood:                -1584.6
No. Observations:                 156   AIC:                             3173.
Df Residuals:                     154   BIC:                             3179.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   5.592e+04    982.044     56.946      0.000     5.4e+04    5.79e+04
rank        -205.7968     10.073    -20.430      0.000    -225.696    -185.897
==============================================================================
Omnibus:                       14.347   Durbin-Watson:                   0.603
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               15.450
Skew:                           0.712   Prob(JB):                     0.000442
Kurtosis:                       3.589   Cond. No.                         190.
==============================================================================
```

Figure 3: Linear interpolation results.

The other methods their results can be found in appendix B.

4.3 OLS results

To start a regression you need to understand your data set and know what is being investigated. In our investigation we aim to predict the salary of law graduates from 156 ranked schools. All the results in this report have been computed while using the standard errors which assume that the covariance matrix of the errors is correctly specified. This way, heteroscedasticity can be found accurately. As mentioned earlier, we use the log transformation of salary as it produces fewer outliers. First we try to find our model without using any categorical variables.
Whenever we use rank as our main regressor, we can see great explainability but also a non-linear pattern in the regression and a non-normal pattern in the residuals. Trying to use transformations, like taking the log or squaring the whole column, does not help with the regression results (this can be found in appendix C).
However, using cost instead of rank does not come with these assumption violations. Hence, we can also use cost as our main regressor as it has better assumptions indications than rank. This difference in indications can be seen in appendix C.
To gain more explainability on the model while retaining the significance of coefficients, we start adding other variables. A very important assumption to keep in mind while adding variables to a model, is that of no multicollinearity, because violating this assumption becomes more likely as you add variables. So whenever a new variable is added, it is of essence to check whether some values and tests are in order. For the sake of simplicity, we will unearth some findings in advance. The variables faculty, studfac, west, south, north, east all have very poor significance to explain lsalary in any combination, and thus will not be used in the remainder of this report. We find that the variable LSAT is significant, but GPA is not. They are also jointly significant. As a t-test smaller than the t-statistic for 5% significance (=1.95) indicates that a variable is insignificant. The F-test value has to bigger than the F-statistic (2.37 for this significance level), to be significant. Moreover, the low probability indicates the probability that the null hypothesis is true, which assumes that all regression coefficients are zero. The coefficients we used are for a model that includes both LSAT and GPA (appendix D).

variable	t-test	F-test
LSAT	3.00	NA
GPA	1.75	NA
LSAT + GPA	NA	180 (Pr=0)

Table 1: Overview for t-test and F test for LSAT and GPA. NA indicating not applicable.

This brings us to including both of them into our model. Now we realise that GPA and LSAT are both measures of intelligence, and hence must be correlating in some way. Their correlation is 0.77, as can be seen in the correlation matrix (appendix A). Thus, choosing only LSAT might seem like a good solution. Even so, if you only include LSAT you will lose considerable explainability. If you standardize both LSAT and GPA and bring them together in a new column called grade, you fix this collinearity while simultaneously keeping that combined significance and explainability. This can be seen in the tables in appendix D.

In the following table the normality of each relevant variable can be seen.

variable	grade	LSAT	GPA	age	clsize	llibvol	libvol	rank	lcost
JB value	2.57	9.85	1.59	0.73	70.25	25.13	2519	9.02	33.97
P value	0.28	0.007	0.45	0.69	0	0	0	0.01	0

Table 2: Individual Jarque-Bera overview for all relevant variables.

The combination of the JB value and P value shows whether given variable is normally distributed. The lower the JB value, the better the normality and if the p-value is bigger than 0.05 this holds. Since grade is our most relevant variable, we will show that this is also visible in its density histogram. The less relevant variables their density histograms can be found in appendix E.

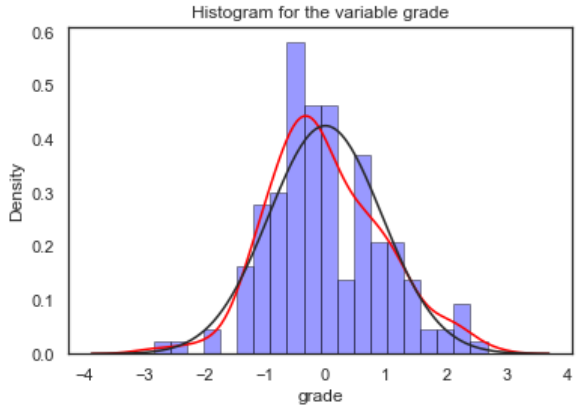


Figure 4: Red = curve of grade. Black = normal distribution.

After checking for all non-categorical variables and comparing the results and test for many combinations, the model $\widehat{lsalary} = rank + libvol + grade$ has the best explainability. The model

$\widehat{lsalary} = lcost + libvol + grade$, breaches least assumptions. Unfortunately, linearity and homoscedasticity still are far from optimal. We refer to the models with rank and lcost as model 2 and 3, respectively, for future comparisons.
Now all that remains is to check whether the categorical ranks are of any significance to predict salary. We can directly see that the original set of categorical ranks (rank 1-60) has great indications for the assumptions. The explainability is also great. This makes us wonder if differently defined categorical ranks yield even better results. We do this for the ranks 1-120 (double original), 100-176 (last 60), 50-115 (middle 60) and for all ranks 1-176 (156 observations). The double has good significance but all assumptions violated. The last ranks has poor explainability, and all assumptions, except for normality, violated. The middle ranks only has homoscedasticity. When looking at all ranks categorically, only collinearity seems to be satisfied. So defining the ranks in a different categorical order makes the regression useless. An overview of the results of all differently defined categorical ranks can be seen in appendix F.
Continuing our search for the best model, we only look at other variables to add to the original set of categorical ranks, so ranks 1-60. The variable grade has better results than GPA and LSAT individually or even combined. Including grade appears to be essential as it fixes linearity and adds considerable explainability. We find that the variables age and clsize best go together. However, the variable age had a lot of missing values in our original data set, consequently, our interpolation for that variable ought to be faulty. Therefore, we would rather not include it in our final model. Only using clsize has too poor of a significance. Now the only variable that remains is llibvol, as libvol has way worse results. Including llibvol to our model adds little explainability. It also does not satisfy the Breush-Pagan test for 1% significance, where the model with only grade added does. Additionally, the model with only grade has the best normality and collinearity conditions, even going as far as making the residuals appear to be normally distributed. Also, comparing the condition number between the two models implies that small changes in the data produce way smaller changes in the solution for the model with only grade. Hence, the best model consists of the original categorical ranks and grade. We refer to this model as model 1. A table considering all original categorical ranks models, can be found in appendix G.
To compare between the models 1, 2 and 3 the following tables are given. The models their individual variables can be found in appendix H.

Model	R-squared	R-squared adj.	Skew	Kurtosis
1	0.879	0.875	0.01	2.98
2	0.852	0.849	0.24	2.91
3	0.712	0.706	0	2.84

Table 3: Regression results for each model.

Model	JB	Cond. No	RR.2	RR.3	RR.4	BP
1	0.01 (Pr=1)	7	0.91	0.95	0.26	0.026
2	1.54 (Pr=0.46)	1720	0	0	0	0.001
3	0.17 (Pr=0.92)	1070	0.82	0	0	0

Table 4: Assumption tests and indications for each model.

We see that model 1 has the best explainability and the best assumption indications. All the Ramsey RESET test their results indicate that the model is linear. The skewness, kurtosis and JB values all indicate near perfect normality. The only assumption that is not within bounds, is that of homoscedasticity. However model 1 only considers the first 60 ranks and our data consists of 156 observations. This is why we are also interested in models 2 and 3, which do cover the whole data set. Amongst them can been deduced that the model with ranks has more explainability, however it does not satisfy a lot of OLS assumptions. The model with lcost has better assumption indications but still lacks some linearity and homoscedasticity.
To investigate further on our best model, model 1, we look into its residuals.

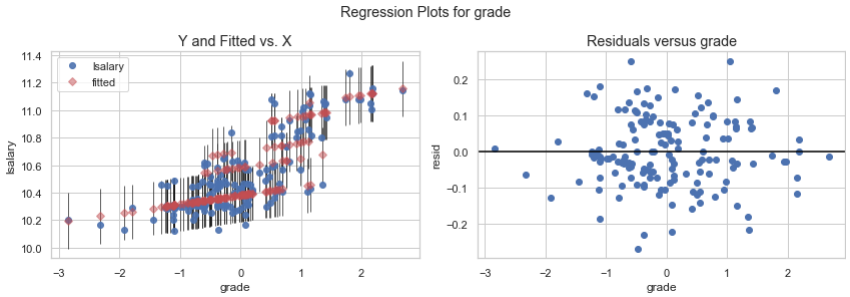


Figure 5: Regression plots for grade, for model 1.

In the Residuals versus grade plot, the following can be seen:

- The residuals are randomly scattered around the horizontal axis. Linearity seems to be correctly specified.
- The residuals are closely scattered around the horizontal axis, without any very noticeable ones. Suggesting that there are no outliers.
- The residuals form a little bit of a cluster around the middle. They also do not form a clear ring around the horizontal axis. Hence, the variances of the error terms are not equal everywhere. This implies a mild form of heteroscedasticity.

From the Y and Fitted vs. X plot can be perceived that:

- There is a positive linear relationship between lsalary and grade.
- The predicted values come close to the actual values of lsalary, this result is backed up by our high R-squared in table 3.

As it is clear that model 1 has some sort of heteroscedasticty, the use of hetero robust errors is necessary for predictions. This results in the following coefficients and (standard deviations): $\widehat{log(salary)} = 10.382\ (0.011) + 0.066\ grade(0.012) + 0.598\ top10(0.040) + 0.509\ r11\_25(0.034) + 0.318\ r26\_40(0.044) + 0.204\ r41\_60(0.021)$ for a total explainability of 87.9% of the data.

4.4 KNN results

As we mention in the methodology for KNN, no assumptions about the form of f(X) are to be made when applying the KNN regression algorithm. Our aim using the K-Nearest Neighbor regression is to compare the results when applying parametric and non parametric framework to

our choices of variables. We carry out this comparison by measuring the Mean Squared Error of our predictions with respect to the true values in both OLS and KNN. In the latest, the computed MSE varies highly with different choices of k. For this reason we display the different MSE for all the values of k up to 25 together with the yielded by the OLS regression. The optimal split between train and test data is for 70/30 in this case, yielding on average a lower MSE than when using other splits like 50/50, 60/40 or 80/20 in our three models

Model 1: lsalary $\sim$ top10 + r11_25 + r26_40 + r41_60 + grade

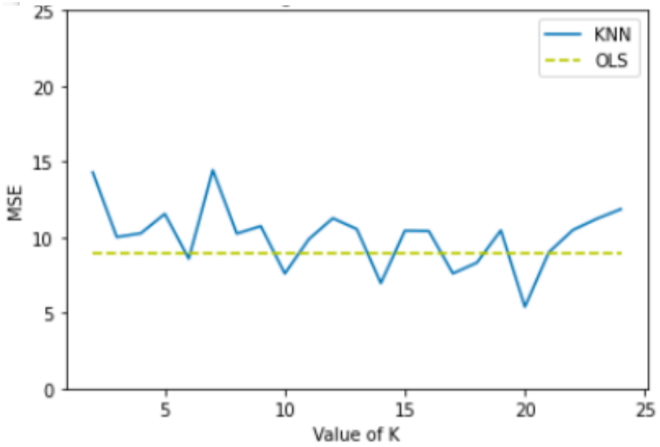


Figure 6: Comparison of OLS and KNN performance in Model 1

We find the MSE in model 1 to be 8.953 when using the Ordinary Least Squares regression. Looking at the graph we can see how the K-Nearest Neighbor approach outperforms the OLS when applied with 6, 10, 14, 17, 18 and 20 neighbors. Getting the best MSE of 5.397 when using k=20.

Model 2: lsalary $\sim$ rank + libvol + grade

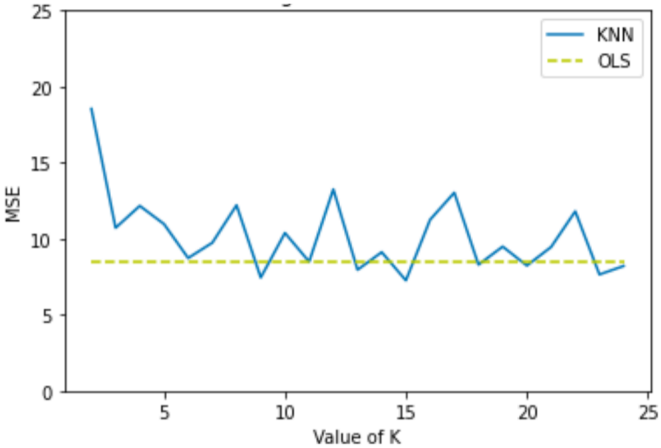


Figure 7: Comparison of OLS and KNN performance in Model 2

In our second model we can observe the same behaviour for the non-parametric approach, outperforming the OLS regression only for some of the values of k. While the parametric approach yields an MSE of 8.429, the KNN gives better predictions when applied with the values of k of 9, 13, 15, 18, 20, 23 and 24. The best performance k for this model is k=15 with an MSE of 7.260

Model 3: lsalary $\sim$ lcost+ libvol + grade

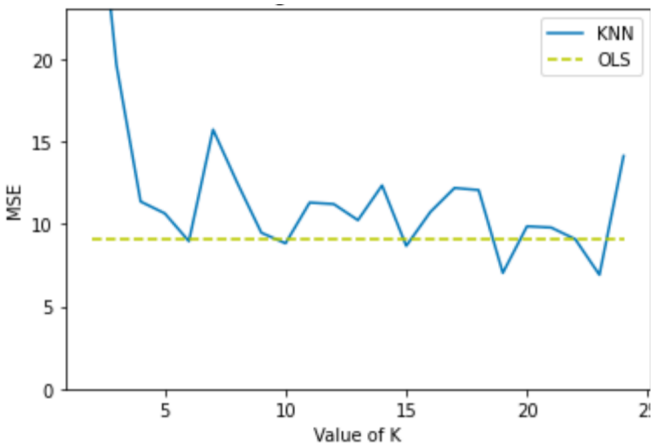


Figure 8: Comparison of OLS and KNN performance in Model 3

The last model that we consider has an MSE of 9.109 in the OLS regression. Slightly higher than in our previous models, KNN still gets better results in terms of the MSE when k equals 6, 10, 15, 19, 22 and 23. The best KNN model is the one that yields MSE of 6.911 using k=23

4.5 Monte Carlo Simulation

We want to see what happens when there is a normally distributed model with non-normal residuals. We use the following four distributions to imitate this process:

- normal:
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$
- chi-square: its independent random variables, which all have a standard normal distribution, are squared and summed. Chi square is positively skewed, it becomes more positively skewed as the degree of freedom increases.
- Laplace: is expressed in terms of absolute difference from the mean, whereas the normal distribution is expressed in terms of the squared difference from the mean. Therefore, the Laplace distribution tends to have fatter tails. It looks like the normal distribution with a very steep summit.
- lognormal: natural logarithm of the normal distribution. We are particularly interested in this distribution as we have used a couple of log transformed variables in our regression models.

We use these distributions, which are reminiscent to the normal distribution, because our considered models seem to have very good normality indications. So, using remote distributions will not be meaningful to our investigation. For our investigation we use the simple linear model that has the same values as our simple salary for rank regression. Here in the $B_0 = 55000$ and $B_1 = -200$. We researched the the difference in resulting estimated B_1, while performing each different distribution ten thousand times. We are interested what happens whenever the sample size is small, considerable or big.

ϵ distr.	n=10	n=20	n=50	n=100	n=150
normal	-207.93, 0.00	-218.88, 0.01	-207.095, 0.46	-199.91, 0.97	-202.03, 0.90
chi-square	-238.92, 0.0	-206.05, 0.30	-197.458, 0.31	-207.55, 0.59	-203.73, 0.40
Laplace	-211.07, 0.0	-223.27, 0.35	-211.414, 0.43	-205.25, 0.12	-208.29, 0.37
lognormal	-207.93, 0.0	-218.88, 0.01	-207.095, 0.46	-199.91, 0.97	-202.03, 0.9

Table 5: $\hat{B_1}$ for a normally distributed random simple regression, where the residuals' its distribution is either: normal, chi-square, Laplace or lognormal. the amount of observations is denoted by a letter n. The value to the right of -xxx, is the JB p-value.

These results indicate that as the amount of observations increase, $\hat{B_1}$ grows closer to the true value of B_1. This is backed up when you interpret the JB p-values, normality becomes better as the amount of observations increase. As a p-value bigger than 0.05 indicates normality. So it does not really matter if your normally distributed model has non-normally distributed residuals, that is, if your data set is big enough. Whenever you only have a small amount of observations, differently distributed residuals can have big influences on your regression. This varies per different distribution, as distributions closer to the normal distribution have better $\hat{B_1}$ values compared to distributions that are less reminiscent. So lognormal its values for $\hat{B_1}$ become more accurate, while needing less observations as oppossed to Laplace which has a less reminiscent form of the normal distribution. For the chi-square distribution there seems to be a threshold of normality as it becomes more positively skewed as the amount of observations increase.

5 Final Word

5.1 Conclusion

We start by filling in our incomplete data set by means of interpolation, as deleting all missing values would results into very biased modeling. Now we can actually begin the regression process and find explanatory models for the salary of law graduates in America, around 1985. We find that the log transformed form of salary, lsalary, has substantially fewer outliers and will thus yield better results. If we use rank as our main regressor we get great explainability, but poor assumption indications. Yet, whenever we use lcost as our main regressor these assumption indications become way more favourable, however it lacks some explainability. To add more explainability to the model, we include more variables. We find that the variable LSAT is individually significant, but GPA is not. They are also jointly significant. Adding them both results into collinearity, as both variables are indicators of intelligence. Standardizing both variables and bringing them together into a new variable called grade, resolves this collinearity while simultaneously maintaining that high explainability. The model $\widehat{\log(\text{ salary }) = 10.74+ -.004}$ rank

+.0002 llibvol+ .059 grade has the best explainability. The model $\widehat{\log(\text{salary})} = 9.364+ -.108$ lcost $+.0004$ llibvol+ .166 grade, breaches least assumptions. Unfortunately, linearity and homoscedasticity are still far from optimal. Now we start looking at categorical variables. We find that the original categorical ranks, 1 until 60, has great explainability and decent assumption indications. Defining the categorical ranks differently results into unacceptable regressions with awful assumption indications. The variable grade is the only one that adds good explainability, to the original categorical ranks, while improving the assumption indications. The only assumption that is not satisfied for this last model, is homoskedasticity. So, this model needs heteroscedastic robust errors to accurately predict. In conclusion, we proclaim that $\widehat{\log(\text{salary})} = 10.382+ -.597$ top10 $+.509$ r11_25+.318 r26_40 $+.204$ r41_60+ .066 grade is the best estimator to predict the salary of graduates from ranked schools in the United States, around 1985. Taking into account mean and variance to compare the different approaches to our regressions, the Mean Squared Error comes as a perfect tool to compare so. As we appreciate when computing the MSE for the parametric and non-parametric regression, the K-nearest neighbor predictions yields better predictions. We come to the conclusion that for this investigation, KNN will represent a more preferable approach to explaining the starting salaries of law school graduates. In our Monte Carlo simulation we compare a normally distributed simple linear model that has normally distributed residuals with the same model that has non-normally distributed residuals. We choose the chi-square, lognormal and Laplace distribution as non-normal distributions for the residuals. We use these distributions as the considered models in our regressions have great normality indications. So, using remote distributions will not be meaningful to our investigation. Our results indicate that as the number of observations increase, the $\hat{B}_1$ grows closer to the true value of B_1. So it does not really matter if your normally distributed model has non-normally distributed residuals, that is, if your data set is big enough. Whenever you have a small amount of observations, differently distributed residuals can have big influences on your regression. We also find that more similar distributions to the normal distribution, tend to have more accurate $\hat{B}_1$ values.

5.2 Discussion

We can not check whether the data has been drawn randomly from the population. Thus, we are not able to clarify one of the assumptions for OLS.
We are uncertain about the origin of the rank for schools. Perhaps this measure is highly correlated with some of the other variables. Or maybe rank has been incorrectly formed otherwise. We are specifically concerned about the variable rank, as our regressions are mainly formed around it. Also, our data has been interpolated by ordering the whole data set for rank. If there are any mistakes in the original form of rank, this can greatly alter our final results.
We have come up with some additional variables that might result in an increase in explainability for the starting salaries. Gender or racial differences are not considered, which have been proven, at least around the 90s (from when this data is from), to have influence on the starting salary. Also, Unemployment rate of the recent graduates for a given school could happen to be significant to our model, given that the unemployment measures the demand for workers from a specific school and this directly has an influence in how much is paid to an employee.

References

Alo, O. (2019). *Handbook of Research Methodology in Social and Behavioural Sciences*. Taraba: Department of Sociology, Federal University Wukari, pp. 59–85.

Belsley, D., K. Kuh, and R. Welsch (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley Sons.

Casella, G. and L. Berger (2002). *Statistical Inference*. Thomson Learning.

Chapra, S. and R. Canale (1998). *Numerical Methods for Engineers: With Programming and Software Applications*. Singapore: McGraw-Hill.

Mohammed Noor, N. et al. (2014). "Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set". In: *Materials Science Forum* 803.4, pp. 278–281.

Salgado, C. et al. (2016). *Secondary Analysis of Electronic Health Records*. Cambridge: Springer, Cham, pp. 143–162.

Stock, J. and W. M.W. (2015). *Introduction to Econometrics*. Pearson.

Woolridge, J. (2013). *Introductory Econometrics: A Modern Approach*. Mason, Ohio: South-Western, Cengage Learning.

	salary	cost	LSAT	GPA	Libvol	faculty	age	clsize	lsalary	llibvol	lcost	grade
mean	38688	12670	158	3.30	348	70	84	242	10.52	5.76	9.39	-0.00
std	12054	3999	4.69	0.20	188	39	40	112	0.28	0.42	0.37	0.94
min	24900	2623	140	2.73	124	17	3	70	10.12	4.82	7.87	-2.85
25%	29837	9635	155	3.20	235	45	62	164	10.30	5.46	9.17	-0.60
50%	34700	12835	158	3.30	303	58	83	225	10.45	5.71	9.46	-0.12
75%	41187	15979	161	3.40	400	86	112	284	10.63	5.99	9.68	0.59
max	78325	20518	171	3.82	1745	245	206	679	11.27	7.46	9.93	2.68

Table 6: descriptive statistics

Appendix A: descriptive statistics

	ranks	salary	cost	LSAT	GPA	Libvol	age	clsize	lsalary	top10	r11_25	r26_40	r41_60	llibvol	lcost	grade
ranks	1															
salary	-0.85	1														
cost	-0.46	0.48	1													
LSAT	-0.72	0.74	0.47	1												
GPA	-0.71	0.74	0.25	0.77	1											
Libvol	-0.61	0.69	0.34	0.57	0.57	1										
age	-0.56	0.49	0.24	0.41	0.43	0.37	1									
clsize	-0.25	0.29	0.38	0.11	0.01	0.45	0.02	1								
lsalary	-0.90	0.99	0.48	0.75	0.74	0.68	0.51	0.28	1							
top10	-0.41	0.62	0.38	0.51	0.53	0.49	0.30	0.03	0.56	1						
r11_25	-0.45	0.54	0.14	0.34	0.35	0.40	0.30	0.27	0.54	-0.09	1					
r26_40	-0.31	0.19	0.13	0.13	0.15	0.07	0.04	0.02	0.23	-0.08	-0.10	1				
r41_60	-0.24	0.04	0.12	0.09	0.03	0.01	0.07	0.05	0.10	-0.09	-0.12	-0.11	1			
llibvol	-0.69	0.72	0.38	0.61	0.58	0.93	0.36	0.47	0.72	0.44	0.43	0.12	0.08	1		
lcost	-0.44	0.42	0.97	0.44	0.19	0.30	0.21	0.36	0.42	0.30	0.14	0.14	0.11	0.34	1	
grade	-0.76	0.79	0.38	0.94	0.94	0.60	0.45	0.07	0.79	0.56	0.37	0.15	0.07	0.63	0.33	1

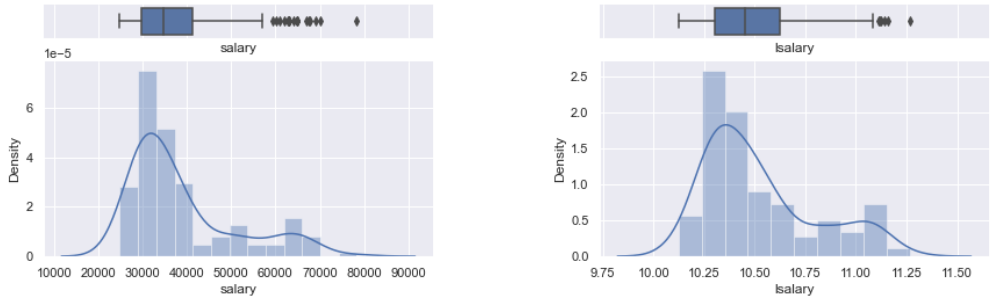
Table 7: correlation matrix

	Skewness	kurtosis(excess)
salary	1.34	0.78
cost	-0.20	-0.77
LSAT	-0.19	1.25
GPA	0.24	0.20
Libvol	3.36	19.18
faculty	1.79	1.79
age	0.15	0.21
clsize	1.32	2.11
lsalary	0.95	-0.13
studfac	1.14	3.06
llibvol	0.79	1.27
lcost	-1.03	1.11
grade	0.29	0.30

Table 8: skewness and kurtosis

	observations	duplicates	missing ranks
top10	10	0	0
r11-25	16	2	1
r26-40	13	0	2
r41-60	18	0	2
r61-175	99	0	16

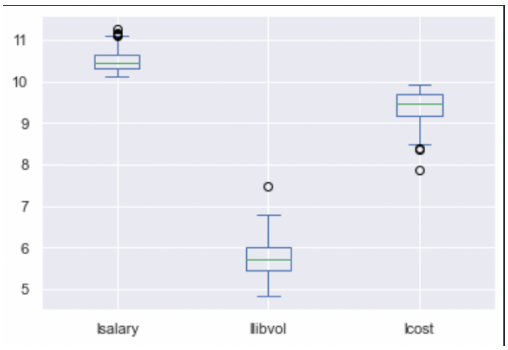
Table 9: categorical rank



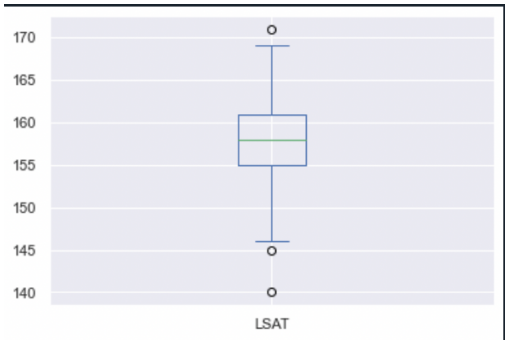
(a) Histogram and boxplot of salary

(b) Histogram and boxplot of lsalary

Figure 9: A figure with two subfigures

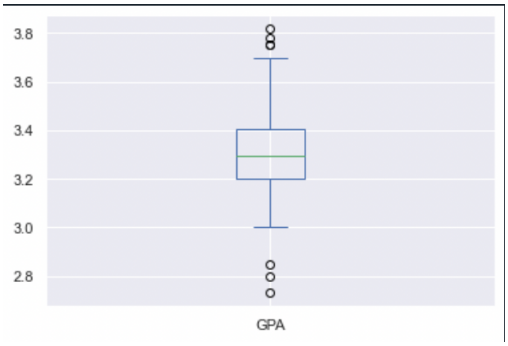


(a) boxplot of log(salary, llibvol,lcost)

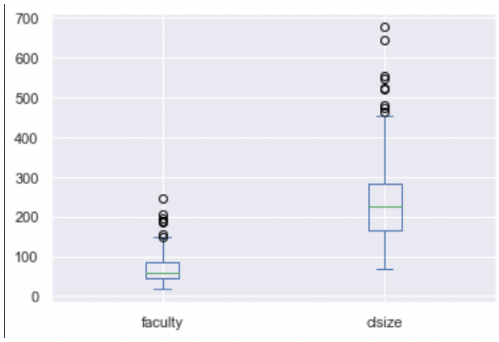


(b) boxplot of LSAT

Figure 10: A figure with two subfigures

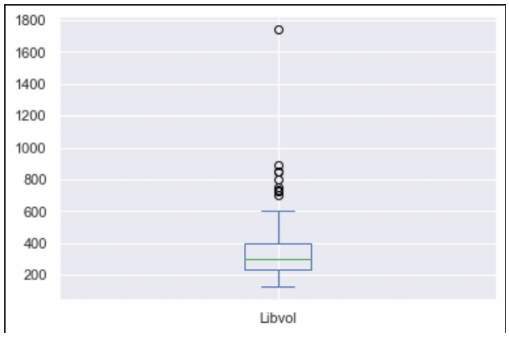


(a) boxplot of GPA



(b) boxplot of faculty and clsize

Figure 11: A figure with two subfigures



(a) boxplot of Libvol

Figure 12: A figure with two subfigures

Appendix B: interpolation methods

```
deleted
Intercept    55886.064003
rank          -208.581080
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.703
Model:                            OLS   Adj. R-squared:                  0.700
Method:                 Least Squares   F-statistic:                     208.6
Date:                Mon, 24 Jan 2022   Prob (F-statistic):           6.11e-25
Time:                        17:41:03   Log-Likelihood:                -916.32
No. Observations:                  90   AIC:                             1837.
Df Residuals:                      88   BIC:                             1842.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   5.589e+04   1306.417     42.778      0.000    5.33e+04    5.85e+04
rank        -208.5811     14.443    -14.442      0.000    -237.283    -179.879
==============================================================================
Omnibus:                        3.616   Durbin-Watson:                   0.697
Prob(Omnibus):                  0.164   Jarque-Bera (JB):                3.414
Skew:                           0.412   Prob(JB):                        0.181
Kurtosis:                       2.520   Cond. No.                         174.
------------------------------------------------------------------------------
```

Figure 13: interpolation results: deleted.

```
mean
Intercept    55480.860297
rank          -197.422616
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.693
Model:                            OLS   Adj. R-squared:                  0.691
Method:                 Least Squares   F-statistic:                     348.2
Date:                Mon, 24 Jan 2022   Prob (F-statistic):           2.29e-41
Time:                        17:41:03   Log-Likelihood:                -1592.2
No. Observations:                 156   AIC:                             3188.
Df Residuals:                     154   BIC:                             3195.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   5.548e+04   1031.500     53.787      0.000    5.34e+04    5.75e+04
rank        -197.4226     10.581    -18.659      0.000    -218.324    -176.521
==============================================================================
Omnibus:                       16.184   Durbin-Watson:                   0.727
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.892
Skew:                           0.776   Prob(JB):                     0.000130
Kurtosis:                       3.588   Cond. No.                         190.
```

Figure 14: interpolation results: mean.

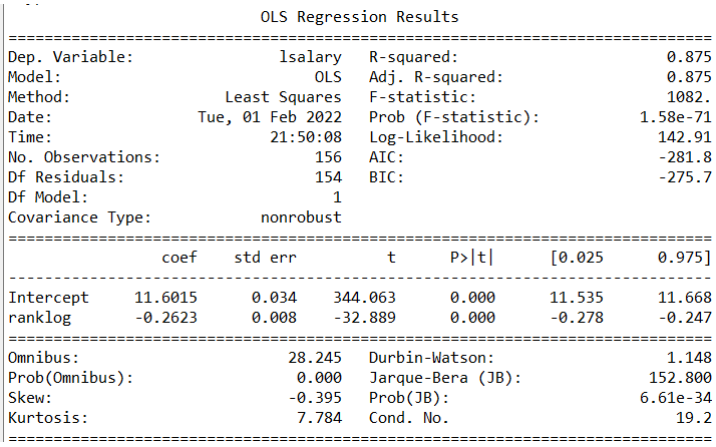
```
median
Intercept    55392.685841
rank          -198.954841
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.700
Model:                            OLS   Adj. R-squared:                  0.698
Method:                 Least Squares   F-statistic:                     359.0
Date:                Mon, 24 Jan 2022   Prob (F-statistic):           4.42e-42
Time:                        17:41:03   Log-Likelihood:                -1591.0
No. Observations:                 156   AIC:                             3186.
Df Residuals:                     154   BIC:                             3192.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   5.539e+04   1023.718     54.109      0.000    5.34e+04    5.74e+04
rank        -198.9548     10.501    -18.947      0.000    -219.699    -178.211
==============================================================================
Omnibus:                       13.841   Durbin-Watson:                   0.692
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               14.820
Skew:                           0.690   Prob(JB):                     0.000605
Kurtosis:                       3.613   Cond. No.                         190.
==============================================================================
```

Figure 15: interpolation results: median.

Appendix C: rank and cost

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                lsalary   R-squared:                       0.621
Model:                            OLS   Adj. R-squared:                  0.619
Method:                 Least Squares   F-statistic:                     252.9
Date:                Tue, 01 Feb 2022   Prob (F-statistic):           2.64e-34
Time:                        21:52:14   Log-Likelihood:                 56.256
No. Observations:                 156   AIC:                            -108.5
Df Residuals:                     154   BIC:                            -102.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      10.7559      0.020    537.620      0.000      10.716      10.795
ranksquared -2.456e-05   1.54e-06    -15.901      0.000   -2.76e-05   -2.15e-05
==============================================================================
Omnibus:                       14.392   Durbin-Watson:                   0.368
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               16.508
Skew:                           0.788   Prob(JB):                     0.000260
Kurtosis:                       2.759   Cond. No.                     1.91e+04
==============================================================================
```

Figure 16: OLS results for lsalary - ranksquared.



OLS Regression Results

Dep. Variable:	lsalary	R-squared:	0.875
Model:	OLS	Adj. R-squared:	0.875
Method:	Least Squares	F-statistic:	1082.
Date:	Tue, 01 Feb 2022	Prob (F-statistic):	1.58e-71
Time:	21:50:08	Log-Likelihood:	142.91
No. Observations:	156	AIC:	-281.8
Df Residuals:	154	BIC:	-275.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.6015	0.034	344.063	0.000	11.535	11.668
ranklog	-0.2623	0.008	-32.889	0.000	-0.278	-0.247

Omnibus:	28.245	Durbin-Watson:	1.148
Prob(Omnibus):	0.000	Jarque-Bera (JB):	152.800
Skew:	-0.395	Prob(JB):	6.61e-34
Kurtosis:	7.784	Cond. No.	19.2

Figure 17: OLS results for lsalary - ranklog.

Appendix D: LSAT, GPA and grade

Variable	Coefficient	Std error	T-value	P-value	Confidence interval	VIF
intercept	8.566	0.026	22.93	0	[7.827 , 9.304]	2289
top10	0.599	0.045	13.21	0	[0.509 , 0.688]	2.0
r11_25	0.510	0.033	15.66	0	[0.445 , 0.574]	1.6
r26_40	0.319	0.031	10.25	0	[0.258 , 0.381]	1.2
r41_60	0.203	0.026	7.71	0	[0.151 , 0.255]	1.2
LSAT	0.009	0.003	2.86	0.002	[0.003 , 0.015]	2.8
GPA	0.119	0.068	3.22	0.082	[-0.015 , 0.254]	3.0

Table 10: LSAT + GPA its variables.

Model	R-squared	R-squared adj.	Skew	Kurtosis
grade	0.879	0.875	0.01	2.98
LSAT	0.877	0.873	0	2.96
GPA	0.871	0.867	-0.06	3.10
LSAT + GPA	0.879	0.874	0.01	2.97

Table 11: Regression results for LSAT, GPA and grade models.

Model	JB	Cond. No	RR.2	RR.3	RR.4	BP
grade	0.01 (Pr=1)	7	0.91	0.95	0.26	0.026
LSAT	0.01 (Pr=0.99)	7600	0.78	0.96	0.23	0.013
GPA	0.17 (Pr=0.92)	87	0.82	0.80	0.29	0.019
LSAT + GPA	0.01 (Pr=0.99)	7610	0.99	0.99	0.22	0.043

Table 12: Assumption tests and indications for LSAT, GPA and grade models.

Appendix E: normality histograms less relevant variables

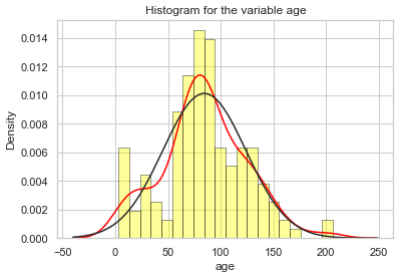


Figure 18: Red = curve of age. Black = normal distribution.

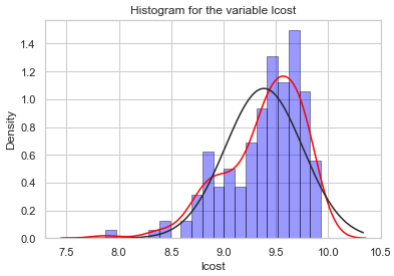


Figure 19: Red = curve of lcost. Black = normal distribution.

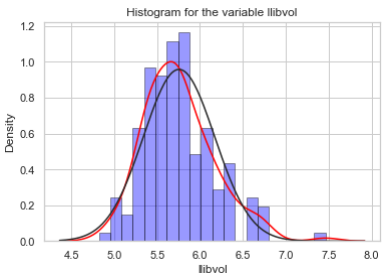


Figure 20: Red = curve of llibvol. Black = normal distribution.

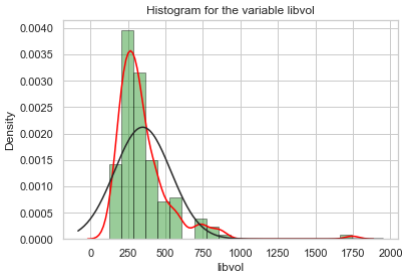


Figure 21: Red = curve of libvol. Black = normal distribution.

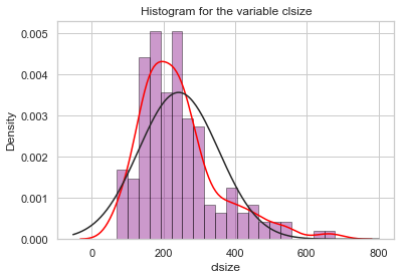


Figure 22: Red = curve of clsize. Black = normal distribution.

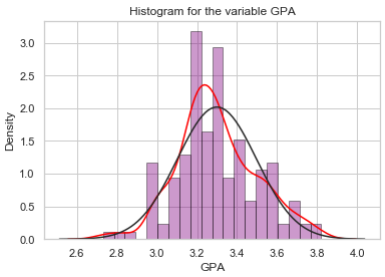


Figure 23: Red = curve of GPA. Black = normal distribution.

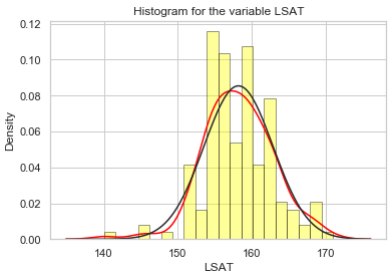


Figure 24: Red = curve of LSAT. Black = normal distribution.

Appendix F: models for differently defined categorical ranks

Model	R-squared	R-squared adj.	Skew	Kurtosis
original	0.892	0.887	-0.11	3.22
double	0.889	0.884	-0.02	4.10
last	0.751	0.739	0.12	2.93
middle	0.750	0.738	0.14	3.09
all	0.854	0.851	-0.37	4.17

Table 13: Regression results for each differently defined categorical ranks model.

Model	JB	Cond. No	RR.2	RR.3	RR.4	BP
original	0.64	2000	0.97	0.70	0.32	0.009
double	7.9 (Pr=0)	2140	0	0.01	0	0
last	0.37	1790	0	0	0	0.045
middle	0.58	1190	0	0	0	0.714
all	12.45 (Pr=0)	6	0	0	0	0

Table 14: Assumption tests and indications for each differently defined categorical ranks model.

Appendix G: variable considerations for original categorical ranks

Model with grade	R-squared	R-squared adj.	Skew	Kurtosis
without grade	0.858	0.854	-0.08	2.94
model 1	0.879	0.875	0.01	2.98
+ age + clsize	0.892	0.887	-0.11	3.22
+ llibvol	0.884	0.879	-0.06	3.12

Table 15: Regression results for each model.

Model	JB	Cond. No	RR.2	RR.3	RR.4	BP
without grade	0.21 (Pr=0.90)	5	0	0	0	0
model 1	0.01 (Pr=1)	7	0.91	0.95	0.26	0.026
+ age + clsize	0.64	2000	0.97	0.70	0.32	0.009
+ llibvol	0.17	118	0.41	0.72	0.42	0.006

Table 16: Assumption tests and indications for each model.

Variable	Coefficient	Std error	T-value	P-value	Confidence interval	VIF
intercept	10.28	0.026	389	0	[10.231 , 10.335]	12.1
top10	0.552	0.039	14.17	0	[0.476 , 0.628]	2.1
r11_25	0.451	0.037	12.25	0	[0.379 , 0.525]	1.9
r26_40	0.300	0.045	6.69	0	[0.212 , 0.388]	1.2
r41_60	0.183	0.020	9.17	0	[0.144 , 0.222]	1.2
age	0.0006	0	2.86	0.004	[0 , 0.001]	1.3
clsize	0.0003	0	3.22	0.001	[0 , 0]	1.1
grade	0.067	0.013	5.30	0	[0.043 , 0.093]	2.5

Table 17: Checking age + clsize its variables.

Variable	Coefficient	Std error	T-value	P-value	Confidence interval	VIF
intercept	10.011	0.150	66.67	0	[9.715 , 10.308]	383.6
top10	0.555	0.048	11.62	0	[0.461 , 0.649]	2.3
r11_25	0.471	0.035	13.30	0	[0.401 , 0.541]	2.0
r26_40	0.300	0.031	9.57	0	[0.238 , 0.241]	1.3
r41_60	0.190	0.026	7.17	0	[0.137 , 0.242]	1.2
llibvol	0.059	0.013	2.48	0.014	[0.013 , 0.119]	2.1
grade	0.066	0.027	4.57	0	[0.034 , 0.085]	2.5

Table 18: Checking llibvol model its variables.

Appendix H: variables for model 1, 2 and 3

Variable	Coefficient	Std error	T-value	P-value	Confidence interval	VIF
intercept	10.382	0.011	907.54	0	[10.359 , 10.405]	2.2
top10	0.597	0.045	13.20	0	[0.508 , 0.687]	2.0
r11_25	0.509	0.032	15.66	0	[0.455 , 0.573]	1.6
r26_40	0.318	0.031	10.25	0	[0.257 , 0.380]	1.2
r41_60	0.204	0.026	7.78	0	[0.152 , 0.256]	1.2
grade	0.066	0.013	5.10	0	[0.040 , 0.092]	2.4

Table 19: Model 1 its variables.

Variable	Coefficient	Std error	T-value	P-value	Confidence interval	VIF
intercept	10.74	0.036	294	0	[10.666 , 10.810]	18.2
rank	-0.004	0	-12.8	0	[-0.004 , -0.003]	2.6
libvol	0.0002	0	3.9	0	[0 , 0]	1.7
grade	0.059	0.015	4.1	0	[0.030 , 0.088]	2.6

Table 20: Model 2 its variables.

Variable	Coefficient	Std error	T-value	P-value	Confidence interval	VIF
intercept	9.364	0.322	29.1	0	[8.728 , 9.999]	725
lcost	0.108	0.035	3.1	0.002	[0.039 , 0.176]	1.1
libvol	0.0004	0	5.3	0	[0 , 0.001]	1.6
grade	0.166	0.016	10.2	0	[0.134 , 0.198]	1.6

Table 21: Model 3 its variables.