What kind of median earnings to expect from a type of major

Kelly Trinh, Hanfei Sun, and Nathan Lam

Introduction

Research question Not all college students are informed about how the choice of their major can impact their earnings after graduation. We aim to provide an overview for college students of the financial standing of graduates from different majors categories. Our research question is:

Given a major category, what factors contribute to how high my median income as a full-time, year-round worker?

We will be building an linear regression model to find an association between major category and median income.

Dataset The dataset we used is the recent-grad dataset from the fivethirtyeight package. https://www.rdocumentation.org/packages/fivethirtyeight/versions/0.6.1/topics/college_recent_grads

It contains income and other information about 16 different major categories. Below is a detailed description of each variable given in the dataset:

(this stuff wont show for me) |Header | Description | |——————|————————————————| `Rank` | Rank by median earnings `Major_code` | Major code `Major` | Major description `Major_category` | Category of major `Total` | Total number of people with major `Sample_size` | Sample size of full-time, year-round `Men` | Male graduates `Women` | Female graduates `ShareWomen` | Proportion of women `Employed` | Number employed `Full_time` | Employed 35 hours or more `Part_time` | Employed less than 35 hours `Full_time_year_round` | Employed at least 50 weeks and at least 35 hours `Unemployed` | Number unemployed `Unemployment_rate` | Unemployed / (Unemployed + Employed) `Median` | Median earnings of full-time, year-round workers `P25th` | 25th percentile of earnings `P75th` | 75th percentile of earnings `College_jobs` | Number with job requiring a college degree `Non_college_jobs` | Number with job not requiring a college degree `Low_wage_jobs` | Number in low-wage service jobs

Final model selected

The selected model we chose includes the following variables, with the two last variables being the interaction terms between dummy variable `major_category_Engineering` and variables `rank` and `sharewomen`:

$$log_{m}edian = ranksharewomen+employed+employed_{f}ulltime_{y}earround+college_{j}ob_{p}rop+major_{c}ategory_{B}iology...Life.Sci$$

We arrived at the final model by performing two round of variable selection. The first round we only consider the existing variables (no interaction terms) using a combination of forward selection, backward selection, and leave-one-out cross validation (LOOCV). The second round we considered interaction terms and used the Analysis of Variance test to determine which terms we should keep in the model. Our diagnostic plots and bootstrap distributions of the coefficients show that the data sufficiently meets the assumption linear regression involving constant variance in the errors, linearity, and normality. Therefore, we decide that an OLS model is sufficient after removing some points with high influence.

Model results and limitations

Coefficient interpretations: **(insert here)**

Here are some limitations of our model:

(model only cares about which model influences higher pay, not a comparison between majors) 1. Our final model only contain 3 out of 16 major categories. Therefore, if a student want to know about how major categories not included in the model variables are compared, the model will not provide that information. For example, if the student wants to know how major categories `Social Science` or `Health` compares, the model will not give much information.

2. The two methods of variable selection (backward/forward versus Analysis of Variance/ANOVA) each do not guarantee to select the best subset of explanatory variables. Backward/forward selection add variable step-wise, so it behaves more as a local optimizer than a global optimizer. The ANOVA test relies on the assumption that the data behaves in a normal distribution, and since it utilizes p-values, there is a possibility of wrongly rejecting the null.

3. For each major category there's only around 5-15 rows, some majors such as `Interdisciplinary` only have 1 row. The results of the model may be more reliable if there is more rows per major category.

(alternate) 3. Each major category is not equally represented proportionally, some majors such as `Interdisciplinary` only have 1 row. The results of the model may be more reliable if there is more rows per major category.

4. The model doesn't explain the following confounding variables:

- The universities the students attend: some universities have stronger connections to certain industries or are located in more advantageous locations. For example, UC Berkeley has strong connections with Silicon Valley, meaning that computer science and technology students may have better chances to land higher paying jobs. On the other hand, Arts students at Berkeley would not have the same advantages.
- Location where the students work: pay may vary highly depending on the location worked. For example, consulting jobs in New York may have higher pay compared to Nevada due to the prices of each location.
- Financial background of the students: students from higher-income families tend to be more well-equipped to land a high-paying job early in their careers.
- The industries the student go to: some students may not choose to go to an unrelated compared to their major. For example, an Arts student choose to work in an investment banking firm such as JP Morgan will have very different salaries compared to an Arts student who becomes an architect.

(alternate) -Context can heavily influence a student's trajectory. Things like financial background, location of university, and type of university can offer students more or less resources needed for better chances of higher pay and thus can influence median income.

5. Other model-building methods we should consider include: finding a better way to treat NaN values instead of omitting them, trying alternative cross-validation methods (such as k-fold cross validation), and considering non-linear models (our diagnostics show that our data show linearity but our analysis will be more rigorous if we compare our OLS model and non-linear models).

(alternate) 5. Alternative steps future studies can make: - change method on dealing with missing values; filling in missing values instead of omitting them. - Change method of cross validation - Consider nonlinear model

Conclusions on main findings

Summarize conclusions ***

Some improvements we could implement into our model: 1. We may want to manually add major categories that were not selected in our model, or we could combine some major categories together depending on their relatedness (for example, combining `Physical Sciences` and `Biology & Life Sciences`), which may allow for our final model include more major category variables. 2. We also might try an exhaustive selection method and trying out different cross validation methods. 3. Finally, we can collect more data to add more information on confounding variables or add more rows per major category.

Additional work

EDA and data cleaning

Here are some modifications we did to our data:

1. The target variable `median` is right-skewed, so we log transformed it.
2. There is only one row with missing variables, so we omitted the row.
3. When examining the correlation plots, we see that many variables are highly correlated. For example, `college_jobs`, `employed_fulltime`, `employed_parttime` are highly correlated. We transforms some of these variables into `college_job_prop` and `full_time_yearround_prop`, which gives the proportions rather than an absolute number. We believe these variables would be less correlated with other explanatory variables.
4. We standardized continuous variables to prepare for variable selection and transformed categorical variables into dummies.
5. We also performed a test-train split; the test set will be used to guard against overfitting and to build prediction intervals.

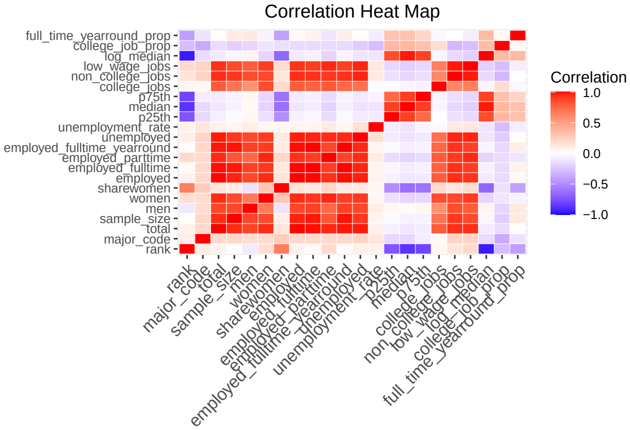


Figure 1: Figure 1: Correlation Heat Map

We also created an overview of the distribution of median incomes from different major categories.

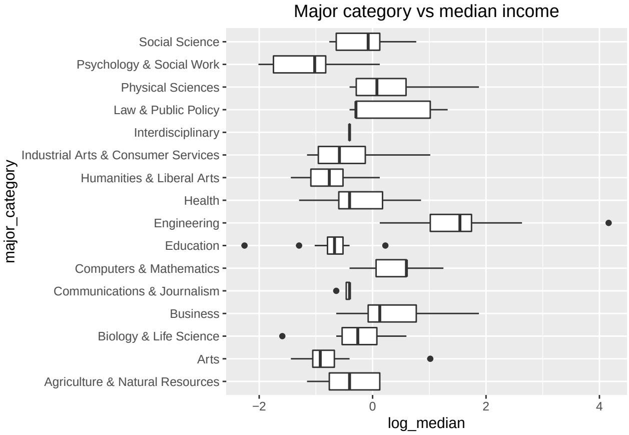


Figure 2: Figure 2

(needs improvement) Looking at Figure 1, there are a lot of variables highly correlated with each other, particularly for variables that are already similar to each other, like type of job or type of employment. High correlation can be good to help explain median income, but it also comes with the risk of collinearity. In Figure 2, we look at how log_median is spread out among various major categories. The most apparent thing is seeing how much money Engineering gets. If a category

Using the alias function on a lm object with all variables reveals which variables are collinear, this turned out to be the variables total and men. We removed men as well as the variables major_code, p25th, and p75th as they do not benefit the model.

Model selection

Methodology As mentioned above, we performed two rounds of variable selection. We started with backward/forward selection. We note that this method requires standardizing continuous variable, and this method doesn't obey the principle of marginality nor treat the variable `major_category` as a single category. Therefore, in this round of variable selection, we do not consider interaction terms. This task is left to the second round of variable selection where we used an ANOVA test that works better with categorical and interaction terms. The first round helps us select which variables to include in the model, and we will create interaction terms from the selected variables. The second round helps us determine which newly created interaction terms to include in the model in addition to the variables selected from the first round.

We used backward and forward selection to filter out the existing variables; we consider interaction terms later. We produced six models: the 3 best models selected by backward selection in terms of Mallow's Cp, BIC, and adjusted R squared; and the 3 best models selected by forward selection using the same criteria. To decide between the 6 models, we performed leave-one-out cross validation (LOOCV) and calculated the root mean squared error (RMSE). We select the model with the lowest LOOCV RMSE, which is the model selected using backward selection with adjusted R squared as a criteria (this selected model is the same model chosen from backward selection using Mallow's Cp). The selected model without interaction terms has RMSE of `0.2435908`. We ensure to check for overfitting by training the selected model using the training set, predict on the testing set, and look at the sum of error squared and the correlation between the fitted values and the actual values.

Next, we consider adding interaction terms. We used the Analysis of Variance test to determine which interaction terms to add to the model. To compare this model and the model without interaction terms, we again calculate the LOOCV RMSE for the former and compare with the RMSE of the latter. We found that the model with interaction terms perform better; specifically its LOOCV RMSE is `0.1933749`.

Variable selection without interactions: backward/forward/cross-validation

```
## Sum of Errors Squared:  6.126396
```

```
## Correlation between the predicted values and actual values:  0.8498587
```

Summary table description:

`forward.rmse` and `backward.rmse` are columns with the RMSE from prediction on the testing dataset depending on either forward selection or backward selection. Each of the row represent the criteria in which to choose the model from either forward or backward selection: `Adjusted R squared`, `BIC`, `Mallow's Cp`.

```
##             criterias forward.rmse backward.rmse
## 1 Adjusted R squared    0.2464696     0.2435908
## 2                 BIC    0.2484253     0.2484253
## 3        Mallow's Cp    0.2476821     0.2435908
```

Variable selection with interactions To examine possible interaction terms, we look at coplots, created on the training data set with standardized continuous variables. The coplots show that `sharewomen` and `employed_fulltime_yearround` seem to have interactions with `major_category`. However, these coplots are inconclusive, so we will run an Analysis of Variance test to determine which interaction terms to keep.

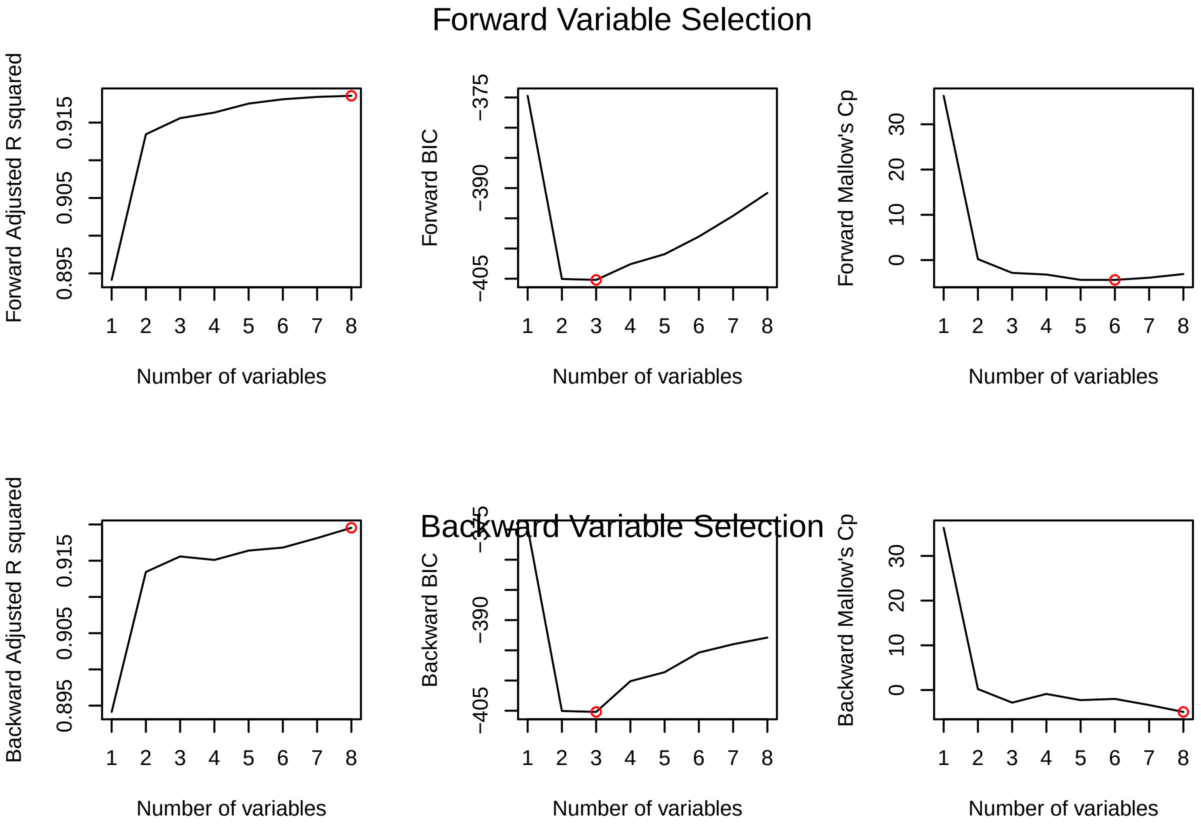
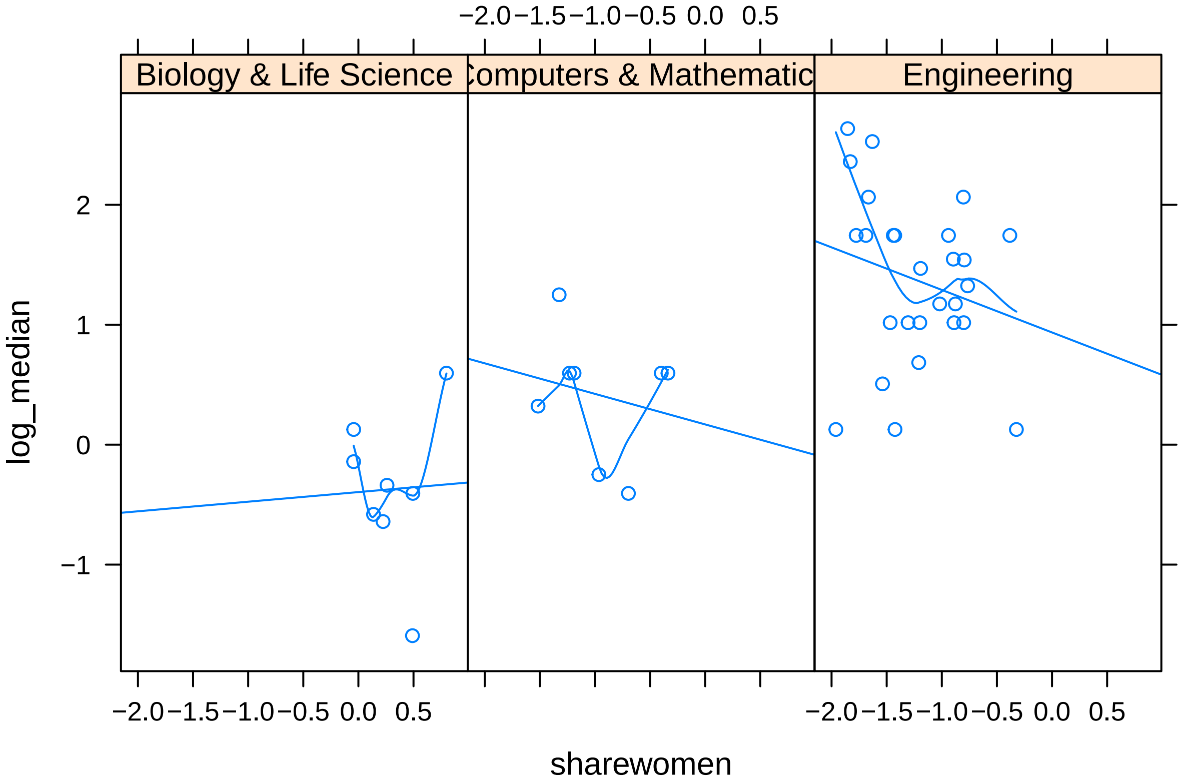
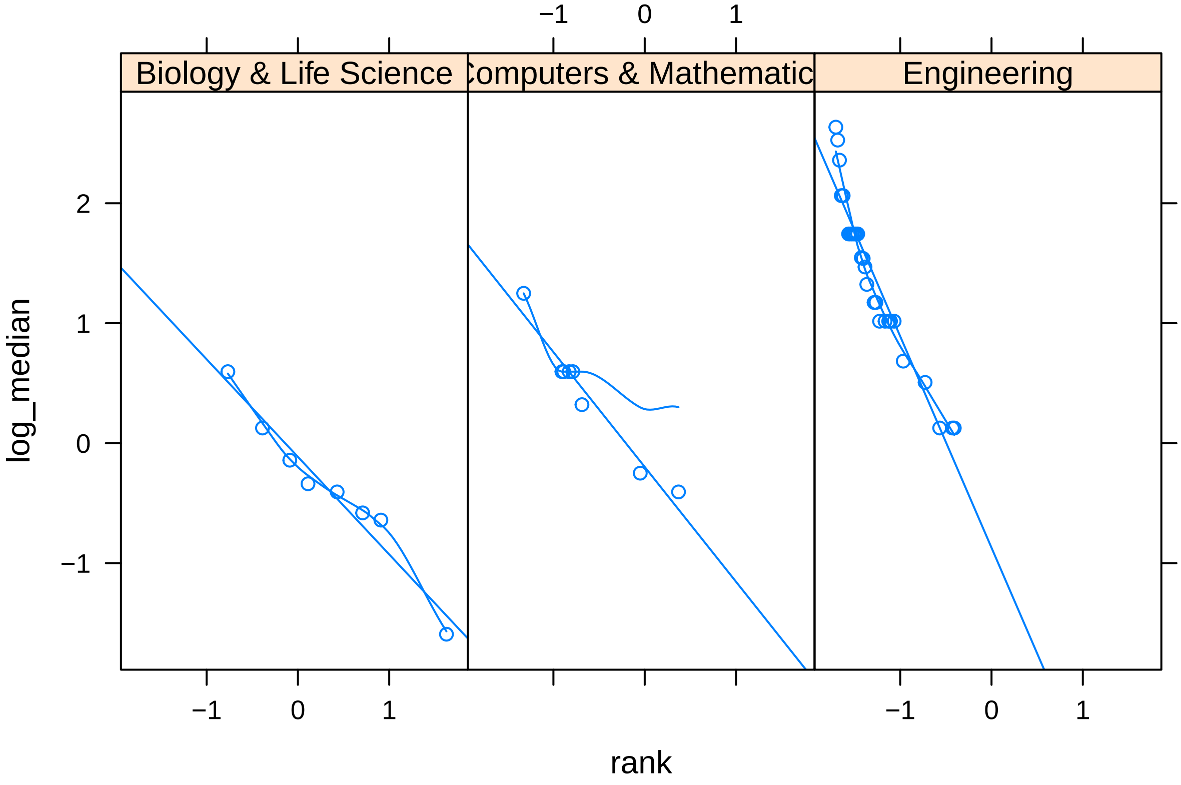
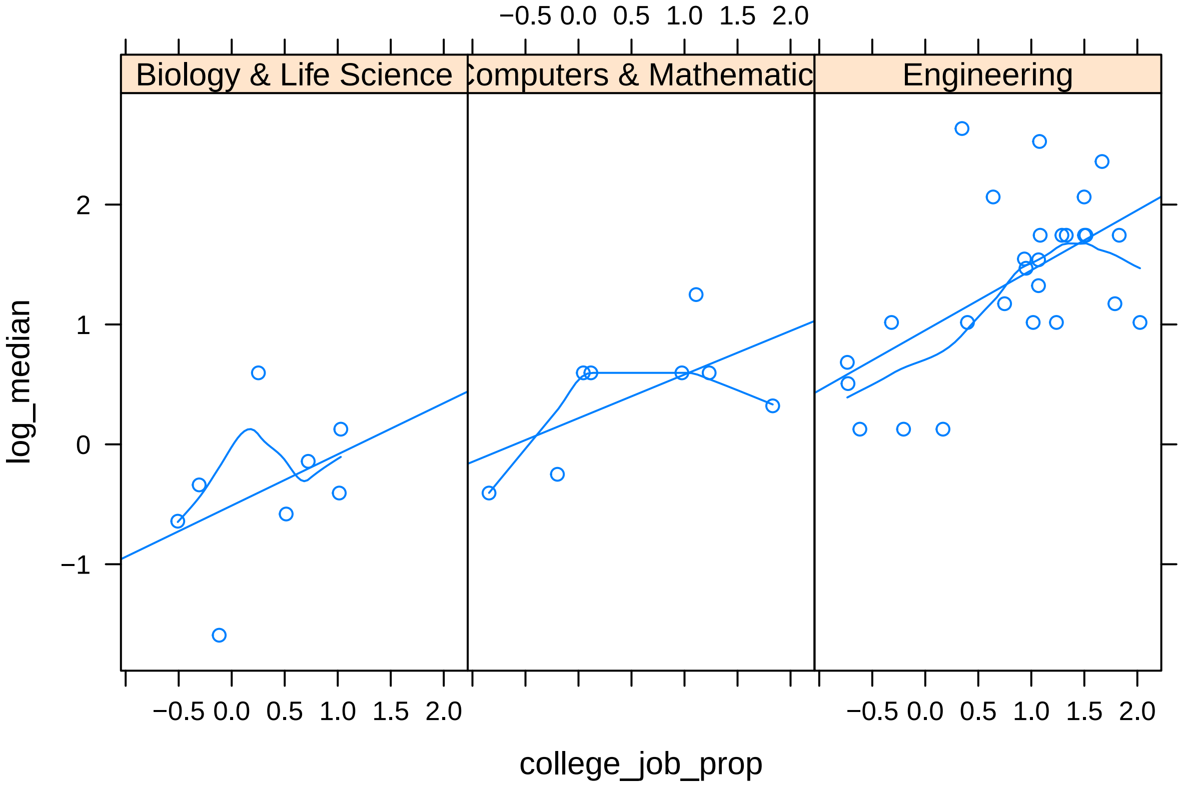
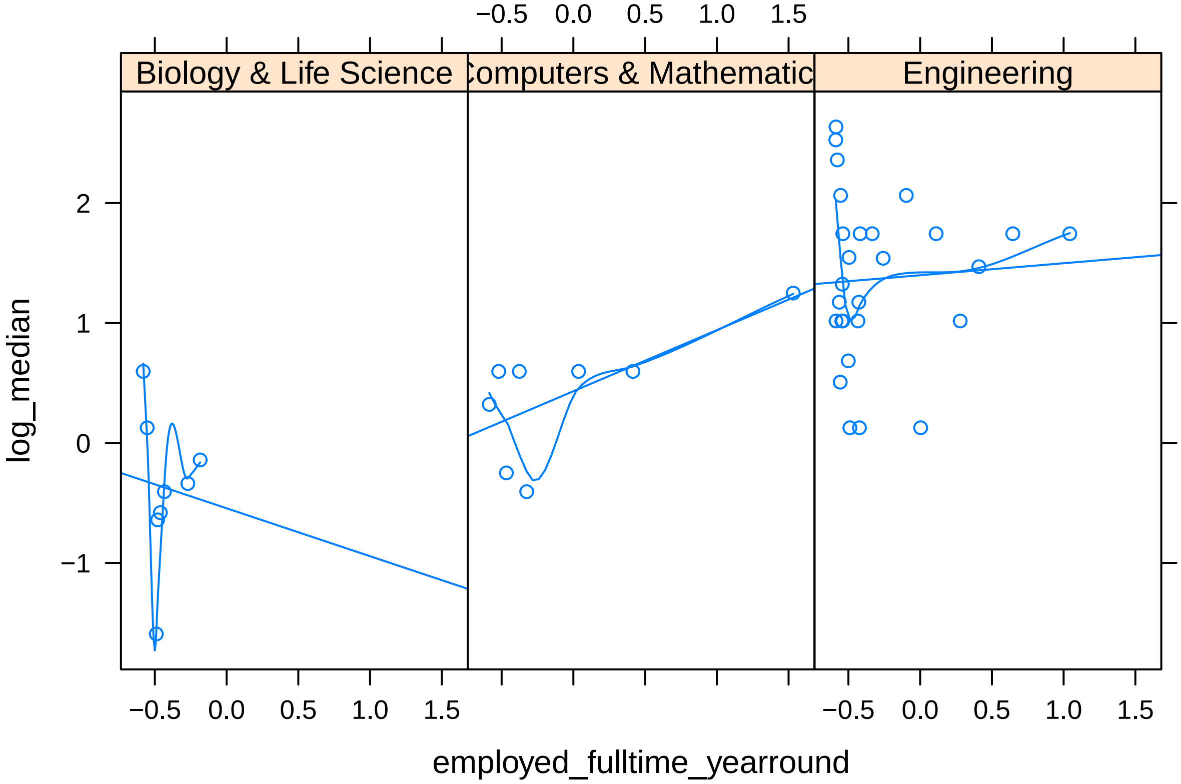


Figure 3: Figure 3: Progression of variable selection


Biology & Life Science
Computers & Mathematic
Engineering
log_median
sharewomen
−2.0 −1.5 −1.0 −0.5 0.0 0.5
−1
0
1
2


Biology & Life Science
Computers & Mathematic
Engineering
log_median
rank
−1
0
1
2





We will keep all the original terms (not interaction terms) and will use the Analysis of Variance test to determine which interaction terms to add. This test, unlike the coplots shown above, show that interaction terms `rank:major_category_Engineering` and `sharewomen:major_category_Engineering` have significant p-values. We will include those variables in our final selected model.

Results After considering the interaction terms, the selected model we chose includes the following variables: `rank`,`sharewomen`,`employed`, `employed_fulltime_yearround`, `college_job_prop`, `major_category_Biology...Life.Science`, `major_category_Computers...Mathematics`, `major_category_Engineering`, `major_category_Engineering * rank`, and `major_category_Engineering * sharewomen`.

Diagnostics

We then perform diagnostics on outliers and assumptions. From the `Residuals vs Fitted` and `Scale-Location` plots, we see that there's no pattern in the studentized residuals against the fitted values, so we conclude that the response variable has a quite linear relationship with the explanatory variables, and the errors have constant residuals. Based on the `QQ plot`, we see that most of the data lines around the theoretical line well, meaning that the response variable is somewhat normally distributed. However, according to the `Residuals vs Leverage` plot, there are some outliers with higher Cook's distance, notably. Below, we fitted the model selected above without these points. We calculate the LOOCV RMSE of this new model.

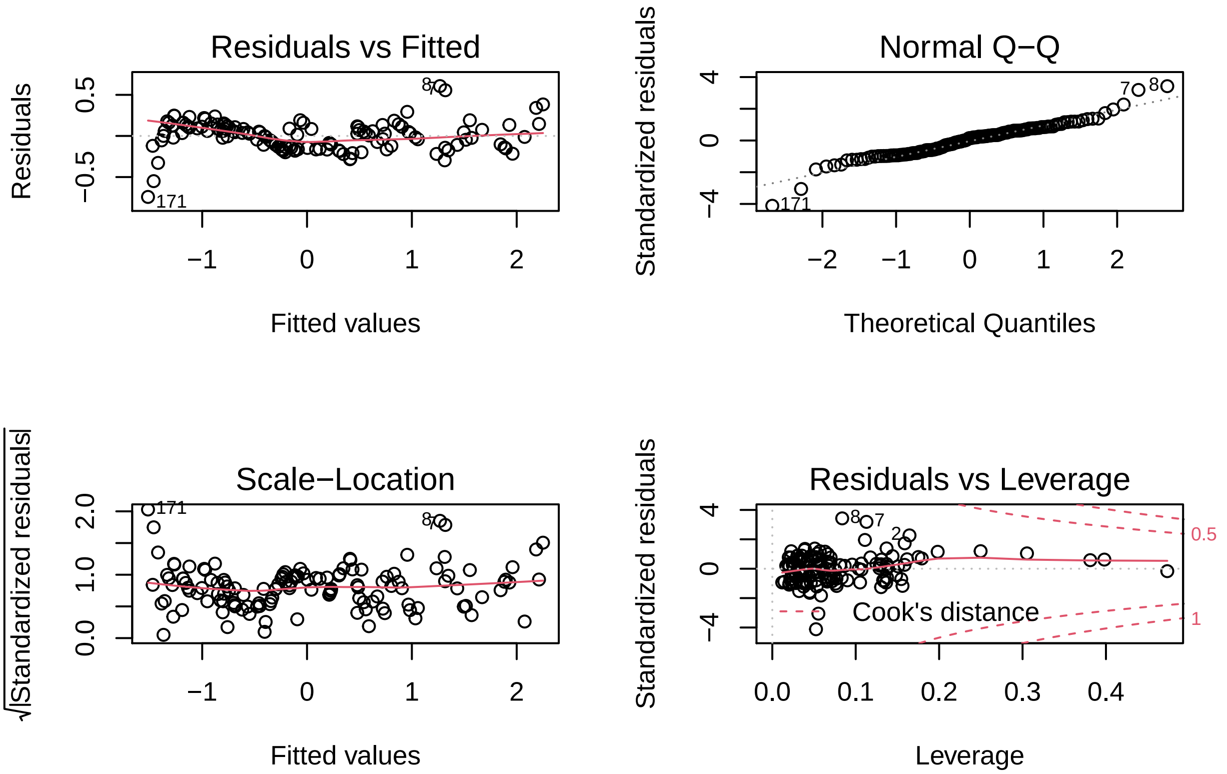


Figure 4: Figure 5: Diagnosis plots

```
##          rank major_code      total sample_size        men      women
## 171  1.706801 -0.2348409 -0.6053869  -0.5765977 -0.5922209 -0.5302517
## 7   -1.605289  1.3689298 -0.5632631  -0.4976805 -0.5220905 -0.5131642
## 8   -1.585337  0.6558129 -0.5944746  -0.5637133 -0.5674481 -0.5303489
## 2   -1.705051 -0.8790806 -0.6107644  -0.5685450 -0.5728782 -0.5518117
##      sharewomen   employed employed_fulltime employed_parttime
## 171  1.54429216 -0.6054164        -0.5977496        -0.5910490
## 7   -0.36746826 -0.5627582        -0.5435864        -0.5870371
## 8    0.04570044 -0.5900044        -0.5863175        -0.5695616
## 2   -1.85405393 -0.6074216        -0.5986093        -0.5956048
##     employed_fulltime_yearround unemployed unemployment_rate      p25th
## 171                  -0.5858389 -0.5704201         1.2184726 -1.0264879
## 7                    -0.5236024 -0.5169041         0.9084334  2.5674138
## 8                    -0.5733135 -0.5834963        -1.5764308  0.2259324
## 2                    -0.5864997 -0.5709044         1.6286645  2.7852260
##        median     p75th college_jobs non_college_jobs low_wage_jobs log_median
## 171 -1.572607 -1.976691   -0.5691383       -0.5481298    -0.5318373  -2.258713
## 7    1.907152  1.378954   -0.4999356       -0.5491344    -0.5222309   1.875002
## 8    1.907152  3.862131   -0.5371554       -0.5413485    -0.5278227   1.875002
## 2    3.038074  2.586986   -0.5662393       -0.5515204    -0.5521971   2.634460
##     college_job_prop full_time_yearround_prop major_category_Arts
## 171       -0.2415388               -0.7405737                   0
## 7          1.7220499                2.2813363                   0
## 8          0.7690312               -0.8476392                   0
## 2          0.3465613               -0.1993523                   0
##     major_category_Biology...Life.Science major_category_Business
## 171                                     0                       0
## 7                                       0                       1
## 8                                       0                       0
## 2                                       0                       0
##     major_category_Communications...Journalism
## 171                                          0
## 7                                            0
## 8                                            0
## 2                                            0
##     major_category_Computers...Mathematics major_category_Education
## 171                                      0                        1
## 7                                        0                        0
## 8                                        0                        0
## 2                                        0                        0
##     major_category_Engineering major_category_Health
## 171                          0                     0
## 7                            0                     0
## 8                            0                     0
## 2                            1                     0
##     major_category_Humanities...Liberal.Arts
## 171                                        0
## 7                                          0
## 8                                          0
## 2                                          0
##     major_category_Industrial.Arts...Consumer.Services
## 171                                                  0
## 7                                                    0
## 8                                                    0
## 2                                                    0
##     major_category_Interdisciplinary major_category_Law...Public.Policy
## 171                                0                                  0
## 7                                  0                                  0
## 8                                  0                                  0
## 2                                  0                                  0
##     major_category_Physical.Sciences major_category_Psychology...Social.Work
## 171                                0                                       0
## 7                                  0                                       0
## 8                                  1                                       0
## 2                                  0                                       0
##     major_category_Social.Science
## 171                             0
## 7                               0
## 8                               0
## 2                               0
```

```
## [1] 0.1555571
```

Prediction

```
##          fit       lwr        upr
## 165 -1.28595 -1.586913 -0.9849877
```

```
##          fit      lwr      upr actual
## 165 28141.92 26099.63 30344.02  27000
```

Reporting