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

Kelly Trinh, Hanfei Sun, and Nathan Lam

Introduction

Research question and Dataset

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 and assess its effectiveness.

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:

Variable name	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`	Number of male graduates
`Women`	Number of 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
`full_time_yearround_prop`	(Created during EDA, not present in original dataset) Proportion of jobs that are full-time year round
`college_job_prop`	(Created during EDA, not present in original dataset) Proportion of jobs that requires a college degree

Final model selected

The final model includes the following variables.(The variable selection process will be further described in "additional work".)

$$
\begin{aligned}
log(median) = {} & rank + sharewomen + employed + employed\_fulltime\_yearround + college\_job\_prop \\
& + major\_category\_Biology\ Life\&Science + major\_category\_Computers\&Mathematics \\
& + major\_category\_Engineering + major\_category\_Engineering * rank \\
& + major\_category\_Engineering * sharewomen
\end{aligned}
$$

We arrived at the final model by performing two rounds 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.

```
##
## Call:
## lm(formula = log_median ~ 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 + major_category_Engineering *
##     sharewomen, data = training_set.temp[-c(7, 8, 171), ])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.185412 -0.032547  0.005124  0.027994  0.151635
##
## Coefficients:
##                                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                1.091e+01  1.902e-02 573.773  < 2e-16
## rank                                      -4.113e-03  1.267e-04 -32.464  < 2e-16
## sharewomen                                -3.780e-02  2.972e-02  -1.272  0.20576
## employed                                   1.738e-06  6.679e-07   2.602  0.01040
## employed_fulltime_yearround               -2.708e-06  1.026e-06  -2.640  0.00936
## college_job_prop                           3.650e-02  2.501e-02   1.460  0.14695
## major_category_Biology...Life.Science     -1.622e-02  1.766e-02  -0.919  0.36014
## major_category_Computers...Mathematics    -2.413e-02  2.000e-02  -1.206  0.22999
## major_category_Engineering                 2.396e-01  3.244e-02   7.387 1.99e-11
## rank:major_category_Engineering           -4.402e-03  5.203e-04  -8.461 6.56e-14
## sharewomen:major_category_Engineering     -2.230e-01  9.436e-02  -2.363  0.01969
##
## (Intercept)                               ***
## rank                                      ***
## sharewomen
## employed                                  *
## employed_fulltime_yearround               **
## college_job_prop
## major_category_Biology...Life.Science
## major_category_Computers...Mathematics
## major_category_Engineering                ***
## rank:major_category_Engineering           ***
## sharewomen:major_category_Engineering     *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.04661 on 123 degrees of freedom
## Multiple R-squared:  0.968,  Adjusted R-squared:  0.9654
## F-statistic: 372.3 on 10 and 123 DF,  p-value: < 2.2e-16
```

Model results and limitations

Since the variables are standardized before the regression, to assess the importance of the variables, we could directly compare the magnitude of the coefficients, and we could discover that the engineering major category and its interaction with the variable "rank" are relatively more important that the other explanatory variables.

In order to make interpretations easier, we reloaded the data and fit the model using the un-standardized variables.

According to the adjusted R-squared value of our final model, roughly 95% of the variation in the log of the median income is explained by our linear model. Since the response variable is the log of the median income, the coefficients can be interpreted as the percentage that the actual median income will change given a one unit increase in the explanatory variables. For instance, the coefficient on "sharewomen", which represents the proportion of women in that major, is roughly -3.780e-02. This means that if the proportion of women in a major increases by 10 percent, the median income of the students graduating with that major is expected to decrease by a factor of eˆ(-0.00378), or by 0.37%, which indicates that there is probably a gender gap in income. For the categorical variables, the coefficient on, for example, major_category_Engineering is approximately 2.396e-01. This means that students graduating with a major in the engineering category has a median income that's higher by a factor of eˆ(0.02396), or 2.4% higher than those who are not. Interpreting the coefficient on interaction terms is slightly more difficult. The coefficient on the interaction term between "sharewomen" and the engineering major category is about -4.402e-03. This means that for students with a major in the engineering category, if the proportion of women students increase by 10%, then the median income of this group of students after graduating is expected to decrease by 0.00402%.

Here are some limitations of our model:

Firstly, our final model only contain 3 out of 16 major categories, which makes our original categorical variable "major_category" coarser. Secondly, 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. Thirdly, 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. Additionally, 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. Finally, 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).

Additional work

EDA and Data Cleaning

The histogram for the median income is right-skewed, so the log transformation would be applicable. Additionally, we omitted missing values and removed rows with missing variables.

To check for collinearity, we looked at the correlation plot for the continuous variables in the data set

(see Figure 1). Several variables are highly correlated since they are similar, such as "college_jobs", "employed_fulltime", and "employed_parttime"; therefore, we transformed two variable into "college_job_prop" and "full_time_yearround_prop". We believe these variables would be less correlated with other explanatory variables.

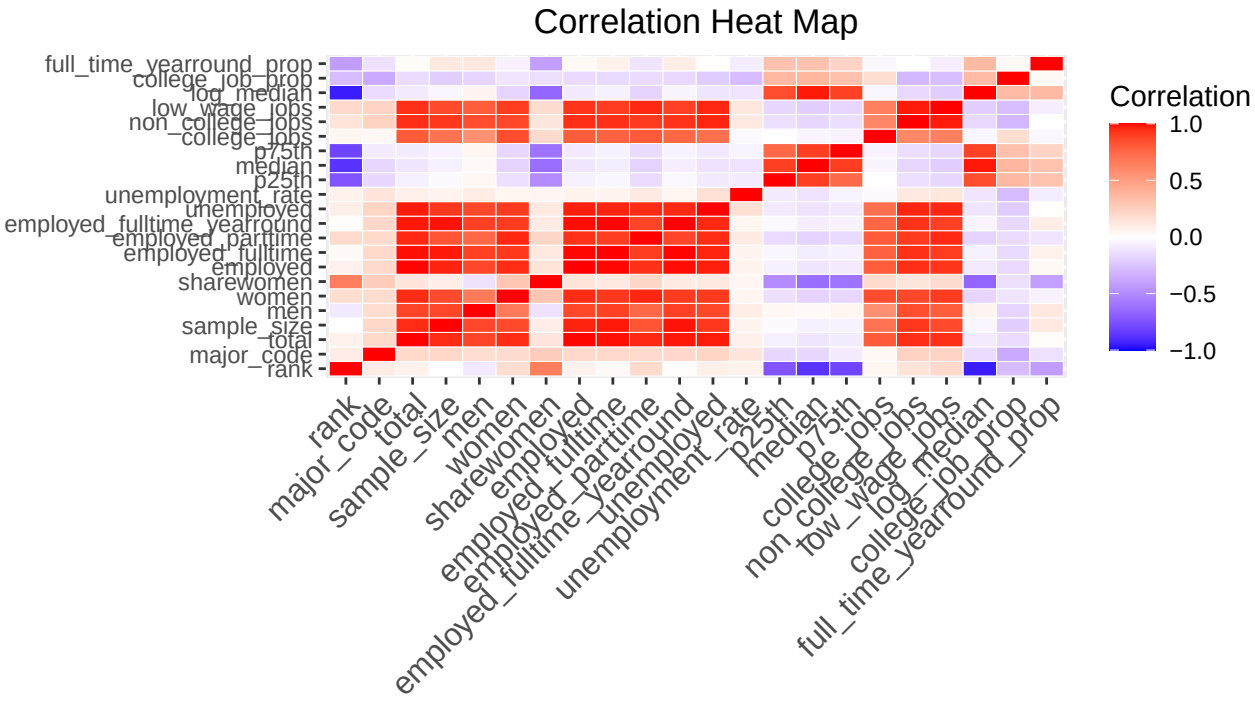


Figure 1: Correlation plot

We also standardized the continuous variables to prepare for variable selection and make it easier to assess the importance of the variables. Then, we converted the major category into dummy columns so that we could perform variable selection using the built-in function in R. To evaluate the model's performance on predicting and avoid overfitting, we conducted a train-test split which takes 80% of the cleaned data as the training set.

Before variable selection, we removed the variables, such as the 25th and the 75th percentile of income, that do not benefit the regression at all, and we used the built-in function "alias" on the linear model object, which reveals any collinearity between the cleaned explanatory variables, and it turned out the "men" and "total" are linearly dependent, so we removed the variable "men" from the cleaned design matrix.

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 `0.1943964`.

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

Figure 2 provides a summary of our variable selection process. The 3 graphs on top shows how the criterions adjusted R squared, BIC, and Mallow's Cp changed as the number of variables increase during forward selection. The 3 graphs below show the same during backward selection. Something interesting to note that BIC chooses models with fewer variables (around 3), while the other criterions prefer models of 6-7 variables.

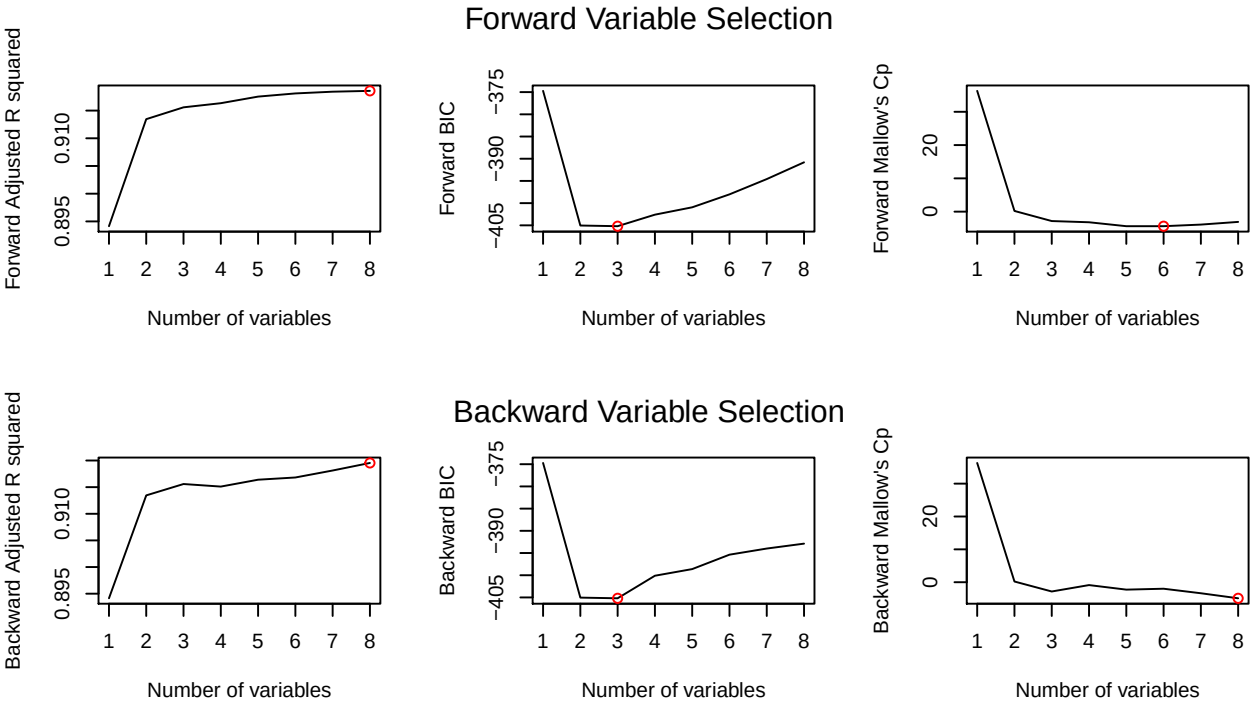


Figure 2: Variable selection summary

Below is a summary table of the results of our variable selection method:

`forward.rmse` and `backward.rmse` are columns with the LOOCV RMSE calculated from predictions on the training dataset from forward selection and 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`.

The model with the lowest LOOCV RMSE is the following model:

$$log(median) = rank + sharewomen + employed + employed\_fulltime\_yearround + college\_job\_prop + major\_category\_Biology\ Life\&Science + major\_category\_Computers\&Mathematics + major\_category\_Engineering$$

```
##            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
```

Second-stage variable selection involving possible 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, the one displayed below for example, are inconclusive because each panel either have a smaller number of points or the points are too clumped together for the fitted lines to show any reliable pattern. Therefore, we will run an Analysis of Variance test to evaluate the quality of the interaction terms under consideration to determine whether or not to keep them.

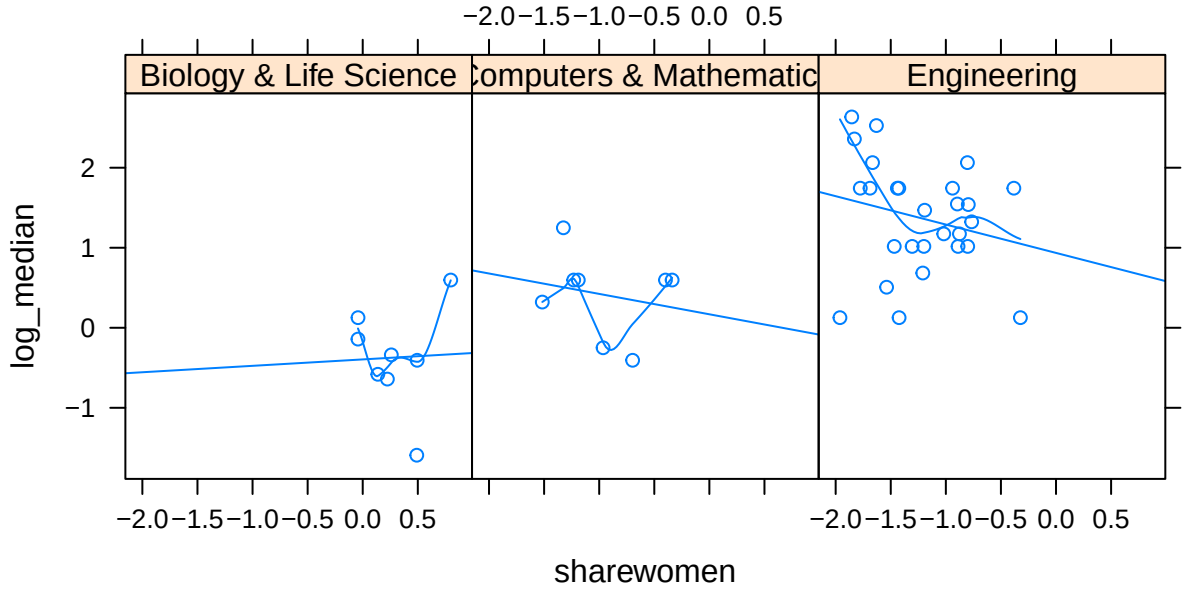


Figure 3: Coplot showing possible interactions between variables sharewomen and log(median)

We will keep all the original terms from the model selected in the first stage of model selection and will use the Analysis of Variance test to determine which interaction terms to add. This test shows 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. At this stage, our model is now:

$$log(median) = rank + sharewomen + employed + employed\_fulltime\_yearround + college\_job\_prop + major\_category\_Biology\ Life\&Science + major\_category\_Computers\&Mathematics + major\_category\_Engineering + major\_category\_Engineering * rank + major\_category\_Engineering * sharewomen$$

Diagnostics

We then perform diagnostics on outliers and checking 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.

```
## [1] 0.1943964
```

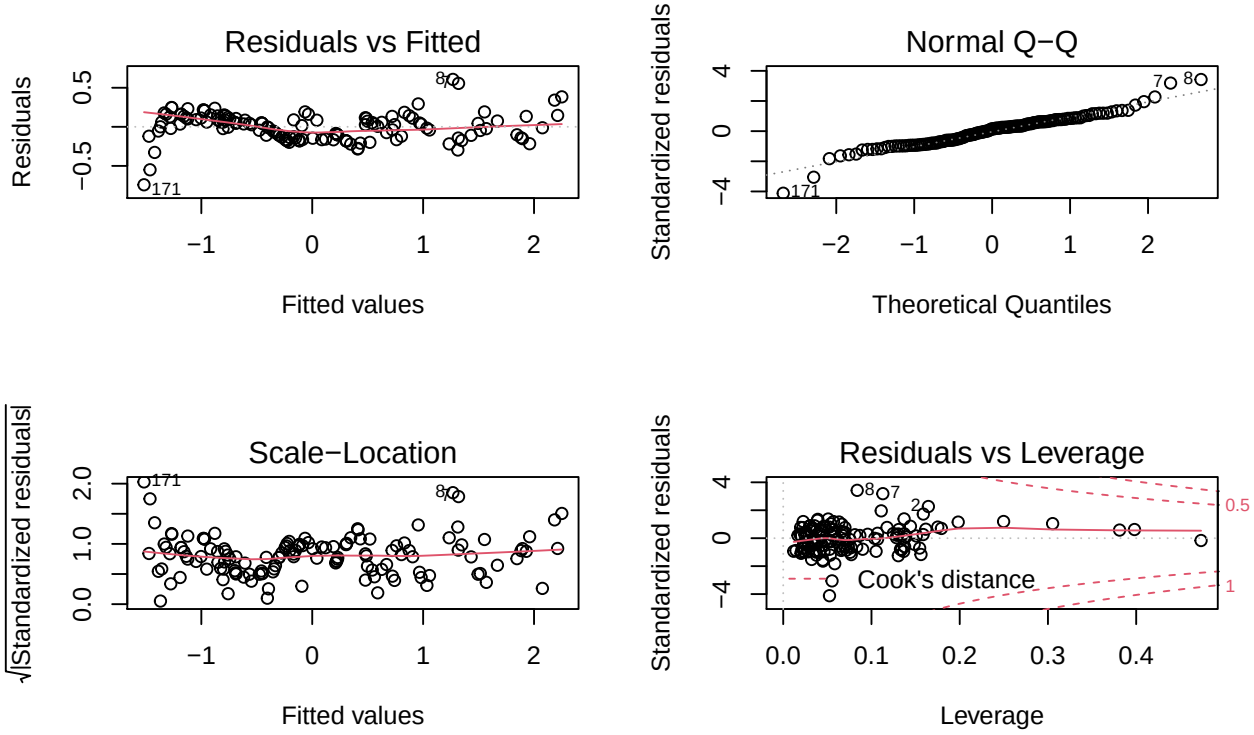


Figure 4: Diagnostics plot

Prediction

```
##          fit       lwr       upr
## 165 -1.35648 -1.734033 -0.978926
```

```
##          fit      lwr     upr actual
## 165 27649.42 25155.91 30390.1  27000
```

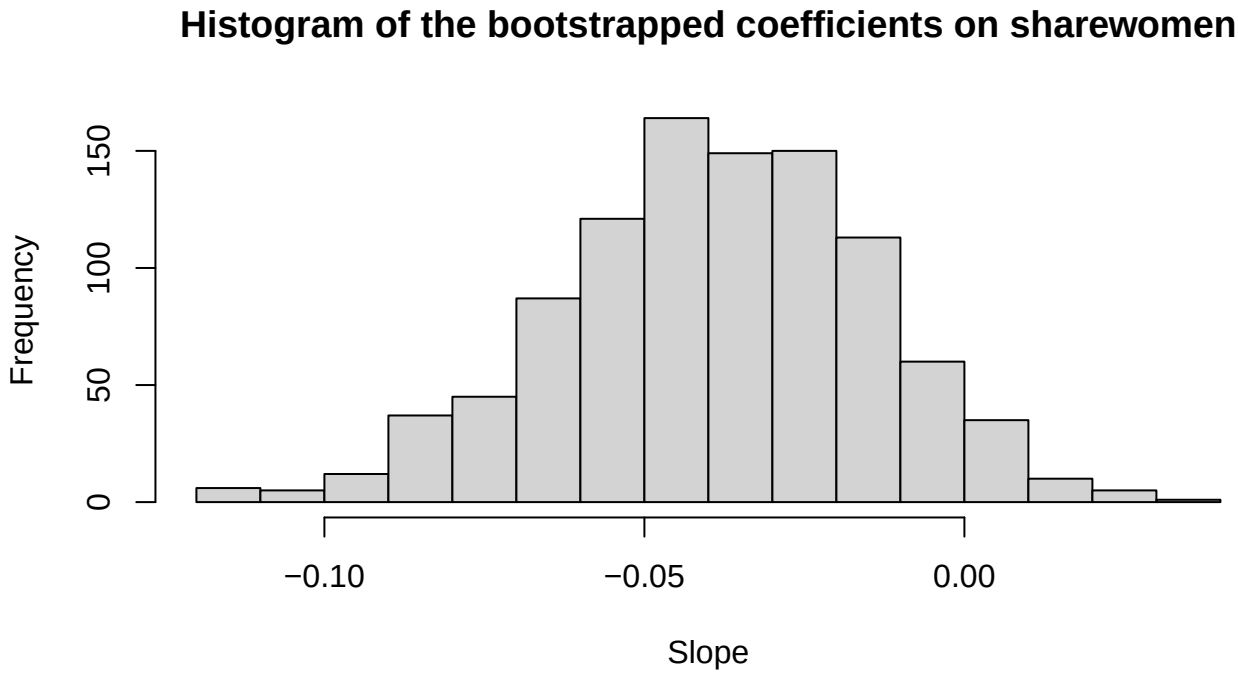
Based on the prediction intervals constructed, we could notice that the predicted value is in the prediction interval. This indicates that the uncertainty of the model is relatively low.

Bootstrap

We will use the bootstrap to assess the stability of the coefficients since different data set will result in different regression coefficients, and then, we will construct 95% confidence intervals for the coefficient on each of the explanatory variables we have selected, using the percentile method.

```
##                                   Lower        Upper
## (Intercept)                 -0.10669573  0.004226982
## rank                        -0.87904870 -0.748308546
## sharewomen                  -0.08936238  0.005818360
## employed                     0.15931940  0.640659437
## employed_fulltime_yearround -0.66072372 -0.162945484
## college_job_prop            -0.01029038  0.072265437
```

We will show, as an example, a histogram in order to see the distribution of the bootstrapped coefficients on an explanatory variable.



In general, if the distribution looks normal, we could make the conclusion that the coefficients are stable. We also noticed that among the 95% confidence intervals computed for the variables we have selected, some contain zero, indicating statistical insignificance, but we believe this could be explained by how, rather than using hypothesis testing, we followed Adjusted R-squared, Mallow's Cp, and BIC,then a series of cross validating, to select the variables.

Conclusions on main findings and possible improvements

From our model results, we see that the major categories "Biology", "Engineering", and "Computer/Math" have significant associations with logged median income. Particularly, engineering majors tend to have higher incomes, while majors with a larger proportion of women, unfortunately, seem to have lower incomes.

Some improvements we could implement into our model include: 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. We also might try an exhaustive selection method and trying out different cross validation methods. Finally, we can collect more data to add more information on confounding variables or add more rows per major category.

Appendix

```
library(fivethirtyeight) #source of data
library(corrplot) #correlation plot
library(dplyr) #data frame manipulation
library(ggplot2) #plotting
library(leaps)#variable selection: adjr2, bic, cp
library(car) # Anova
library(fastDummies)#dummy variables
library(faraway)
library(caret)
# load data
crg <- college_recent_grads
crg <- na.omit(crg)

# viewing correlations
crg["log_median"] <- log(crg$median)
crg["college_job_prop"] <- crg$college_jobs / (crg$college_jobs + crg$non_college_jobs)
crg["full_time_yearround_prop"] <- crg$employed_fulltime_yearround / crg$employed
crg <- na.omit(crg)
cont <- select_if(crg, is.numeric)
cont <- na.omit(cont)

# transform data
standardized <- data.frame(scale(cont))
major_category <- crg$major_category
crg <- cbind(standardized, major_category)

# temporary data frame to use later
temp <- crg

# transform categorical variable
crg <- dummy_cols(crg, select_columns = "major_category", remove_selected_col = TRUE,
                  remove_first_dummy = TRUE)
column_names <- make.names(names(crg),unique=TRUE)
colnames(crg) <- column_names

# test train split
set.seed(11)
#randomly take 2:8 of data for training
training_size <- sample(dim(crg)[1],nrow(crg)*0.8)
training_data <- crg[training_size,]
testing_data <- crg[-training_size,]

#Interpretation of the coefficients(un-standardized)

# recreating a table without standardized continuous variables
# this table should have the same rows as the training dataset
crg.temp <- college_recent_grads
crg.temp <- na.omit(crg.temp)

# create new variable
crg.temp["log_median"] <- log(crg.temp$median)
crg.temp["college_job_prop"] <- crg.temp$college_jobs /
```

```
   (crg.temp$college_jobs + crg.temp$non_college_jobs)
crg.temp["full_time_yearround_prop"] <- crg.temp$employed_fulltime_yearround /
   crg.temp$employed
crg.temp <- na.omit(crg.temp)

training_set.temp <- crg.temp[training_size,]

# transform categorical variable
training_set.temp  <- dummy_cols(training_set.temp , select_columns = "major_category",
                                 remove_selected_col = FALSE,remove_first_dummy = TRUE)
column_names <- make.names(names(training_set.temp ),unique=TRUE)
colnames(training_set.temp) <- column_names

final_model <- lm(log_median ~
                        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 +
                        major_category_Engineering * sharewomen,
                      data = training_set.temp[-c(7,8,171),])

summary(final_model)


#correlation between continuous variables
corrrr <- reshape2::melt(cor(cont),
                       varnames = paste0("variables", 1:2),
                       value.name = "Correlation")

ggplot(corrrr, aes(variables1, variables2, fill = Correlation)) +
  geom_tile(color="white") +
  scale_fill_gradient2(low="blue",
                       high="red",
                       mid="white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Correlation") +
  labs(x="",y="",title = "Correlation Heat Map") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 12, hjust = 1), plot.title = element_text(hjust = 0.5))


Y <- as.matrix(crg$log_median)
X <- as.matrix(subset(crg,select=-c(log_median,median, major_code, p25th, p75th)))

# Check for linear dependencies, remove "men" or "total"
alias(lm(Y~X))

#removing men because of collinearity
X <- as.matrix(subset(crg,select=-c(log_median,median, major_code, p25th, p75th,  men)))
```

```r
#########Forward Selection#########
f_model <- regsubsets(x = X, y = Y, method = "forward", nbest = 1) %>% summary()

#extracting criterion
f_adjustr2 <- f_model$adjr2
f_BIC <- f_model$bic
f_mallow_Cp <- f_model$cp

#picking best variables
f_adjr2_picked <- f_model$which[which.max(f_adjustr2),]
f_BIC_picked <- f_model$which[which.min(f_BIC),]
f_cp_picked <- f_model$which[which.min(f_mallow_Cp),]

#printing picked variables
f_adjr2_picked[f_adjr2_picked == TRUE]
f_BIC_picked[f_BIC_picked == TRUE]
f_cp_picked[f_cp_picked == TRUE]



#########Backward Selection#########
b_model <- regsubsets(x = X, y = Y, method = "backward", nbest = 1) %>% summary()

#extracting criterion
b_adjustr2 <- b_model$adjr2
b_BIC <- b_model$bic
b_mallow_Cp <- b_model$cp

#picking best variables
b_adjr2_picked <- b_model$which[which.max(b_adjustr2),]
b_BIC_picked <- b_model$which[which.min(b_BIC),]
b_cp_picked <- b_model$which[which.min(b_mallow_Cp),]

#printing picked variables
b_adjr2_picked[b_adjr2_picked == TRUE]
b_BIC_picked[b_BIC_picked == TRUE]
b_cp_picked[b_cp_picked == TRUE]

par(mfrow=c(2,3))
plot(f_adjustr2, xlab = "Number of variables",
     ylab = "Forward Adjusted R squared", type = "l")
points(which.max(f_adjustr2), max(f_adjustr2), col = "red")
plot(f_BIC, xlab = "Number of variables",ylab = "Forward BIC", type = "l")
points(which.min(f_BIC), min(f_BIC), col = "red")
plot(f_mallow_Cp, xlab = "Number of variables",ylab = "Forward Mallow's Cp", type = "l")
points(which.min(f_mallow_Cp), min(f_mallow_Cp), col = "red")
###
plot(b_adjustr2, xlab = "Number of variables",
     ylab = "Backward Adjusted R squared", type = "l")
points(which.max(b_adjustr2), max(b_adjustr2), col = "red")
plot(b_BIC, xlab = "Number of variables",ylab = "Backward BIC", type = "l")
points(which.min(b_BIC), min(b_BIC), col = "red")
plot(b_mallow_Cp, xlab = "Number of variables",ylab = "Backward Mallow's Cp", type = "l")
```

```r
points(which.min(b_mallow_Cp), min(b_mallow_Cp), col = "red")
mtext("Forward Variable Selection",side=3,line=-2,outer=TRUE)
mtext("Backward Variable Selection",side=3,line=-18,outer=TRUE)


# Backward Adjusted R-squared CV
tc <- trainControl(method = "LOOCV")
b_adjr2_mod_cv <- train(log_median ~ employed + employed_fulltime_yearround +
                          college_job_prop +
                          major_category_Biology...Life.Science+
                          major_category_Computers...Mathematics +
                          rank + sharewomen + major_category_Engineering ,
                        data = training_data, method = "lm", trControl = tc)
b_adjr2_mod_cv.rmse <- b_adjr2_mod_cv$results[,"RMSE"]

# Backward BIC CV
b_bic_mod_cv <- train(log_median ~ rank + college_job_prop +
                        major_category_Engineering,
                      data = training_data, method = "lm", trControl = tc )
b_bic_mod_cv.rmse <- b_bic_mod_cv$results[, "RMSE"]
# Backward Mallow's Cp CV
b_cp_mod_cv <- train(log_median ~ employed + employed_fulltime_yearround +
                       college_job_prop +  major_category_Biology...Life.Science
                     + major_category_Computers...Mathematics + rank +
                       major_category_Engineering + sharewomen , data = training_data ,
                     method = "lm", trControl = tc)
b_cp_mod_cv.rmse <- b_cp_mod_cv$results[, "RMSE"]
# Forward Adjusted R-squared CV
f_adjr2_model_cv <- train(log_median ~ rank + unemployment_rate +
                            college_job_prop + major_category_Arts +
                            major_category_Biology...Life.Science  +
                            major_category_Engineering  +
                            major_category_Humanities...Liberal.Arts  +
                            major_category_Industrial.Arts...Consumer.Services ,
                            data = training_data, method = "lm", trControl = tc)

f_adjr2_model_cv.rmse <- f_adjr2_model_cv$results[, "RMSE"]

# Forward BIC CV
f_BIC_model_cv <- train(log_median ~ rank + college_job_prop +
                            major_category_Engineering,
                        data = training_data, method = "lm", trControl = tc)

f_BIC_model_cv.rmse <- f_BIC_model_cv$results[, "RMSE"]

# Forward Mallow's Cp CV
f_Cp_model_cv <- train(log_median ~ rank + unemployment_rate +
                            college_job_prop + major_category_Arts +
                           major_category_Engineering +
                           major_category_Humanities...Liberal.Arts,
                       data = training_data, method = "lm", trControl = tc)

f_Cp_model_cv.rmse <- f_Cp_model_cv$results[, "RMSE"]
```

```
criterias <- c("Adjusted R squared", "BIC", "Mallow's Cp")
forward.rmse <- c(f_adjr2_model_cv.rmse, f_BIC_model_cv.rmse, f_Cp_model_cv.rmse)
backward.rmse <- c(b_adjr2_mod_cv.rmse, b_bic_mod_cv.rmse, b_cp_mod_cv.rmse)

# table of RMSE of each model
data.frame(criterias, forward.rmse, backward.rmse)
#interactions

# temp dataframe for co-plots
temp.training <- temp[temp$rank %in% training_data$rank,]
temp.coplot <- temp.training[temp.training$major_category %in%
                               c("Biology & Life Science",
                                 "Computers & Mathematics",
                                 "Engineering"), ]


# share women and median income
xyplot(log_median ~ sharewomen | major_category, data = temp.coplot ,
       panel = function(x, y) {
panel.xyplot(x, y, type = c("p", "smooth", "r"))
})

# rank and median income
xyplot(log_median ~ rank | major_category, data = temp.coplot ,
       panel = function(x, y) {
panel.xyplot(x, y, type = c("p", "smooth", "r"))
})

# college job proportion and median income
xyplot(log_median ~ college_job_prop | major_category, data = temp.coplot ,
       panel = function(x, y) {
panel.xyplot(x, y, type = c("p", "smooth", "r"))
})

# employed fulltime and median income
xyplot(log_median ~ employed_fulltime_yearround | major_category, data = temp.coplot ,
       panel = function(x, y) {
panel.xyplot(x, y, type = c("p", "smooth", "r"))
})
# model selected from backward/forward selection
model.backfor <- lm(log_median ~ employed + employed_fulltime_yearround +
                          college_job_prop +
                          major_category_Biology...Life.Science+
                          major_category_Computers...Mathematics +
                          rank + sharewomen + major_category_Engineering,
                    data = training_data)

# full model with all the interaction terms
model.interacts <- lm(log_median ~
                        rank + sharewomen +
                        employed + employed_fulltime_yearround +
                        college_job_prop +
                        major_category_Biology...Life.Science+
```

```
                        major_category_Computers...Mathematics +
                        major_category_Engineering +
                        major_category_Biology...Life.Science * rank+
                        major_category_Computers...Mathematics * rank+
                        major_category_Engineering * rank +
                        major_category_Biology...Life.Science * sharewomen+
                        major_category_Computers...Mathematics * sharewomen +
                        major_category_Engineering * sharewomen +
                        major_category_Biology...Life.Science * employed_fulltime_yearround+
                        major_category_Computers...Mathematics * employed_fulltime_yearround+
                        major_category_Engineering * employed_fulltime_yearround+
                        major_category_Biology...Life.Science * college_job_prop+
                        major_category_Computers...Mathematics *college_job_prop +
                        major_category_Engineering *college_job_prop,
                        data = training_data)


# anova test
Anova(model.interacts)

# reduced model after running Anova test
model.inter_reduced <- lm(log_median ~
                        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 +
                        major_category_Engineering * sharewomen,
                        data = training_data)

anova(model.interacts, model.inter_reduced)
# compare model with and without interactions in terms of LOOCV RMSE
interactions_cv <- train(log_median ~
                        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 +
                        major_category_Engineering * sharewomen,
                      data = training_data, method = "lm", trControl = tc)

interactions_cv.rmse <- interactions_cv$results[, "RMSE"]

interactions_cv.rmse

# outlier rows
outlier_rows <- training_data[c(7,8,171),]

# refit model with the outliers
```

```r
interactions_no_outliers <- train(log_median ~
                        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 +
                        major_category_Engineering * sharewomen,
                      data = training_data[-c(7,8,171),], method = "lm", trControl = tc)


# see CV RMSE
interactions_no_outliers.rmse <- interactions_no_outliers$results[, "RMSE"]
interactions_no_outliers.rmse
par(mfrow=c(2,2))

plot(model.inter_reduced)


# final model selected

model.final <- lm(log_median ~
                        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 +
                        major_category_Engineering * sharewomen,
                      data = training_data[-c(7,8,171),])


# new data from testing set
set.seed(11)
new_data1 <- testing_data[sample(nrow(testing_data), 1), ]


# reporting prediction intervals (standardized)
predict.standard <- predict(model.final, newdata = new_data1, interval = "predict")
predict.standard

# bring back to original scale
predict.original <- exp(predict.standard * sd(log(college_recent_grads$median)) +
                          mean(log(college_recent_grads$median)))

# get data from original data table
new_data1.rank <- new_data1[,1] * sd(college_recent_grads$rank) + mean(college_recent_grads$rank)
temp1 <- cbind(predict.original, college_recent_grads[ceiling(new_data1.rank), ]$median)
colnames(temp1) <- c(colnames(predict.standard), "actual")

# compare predicted data and actual value
```

```
temp1

# Bootstrap Alternative

r <- 1000
store <- matrix(0, nrow = r, ncol = 11)
for (i in 1:r){
   new_data <- sample_n(training_data, nrow(training_data), replace = TRUE)
   mod <- lm(log_median ~
                        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 +
                        major_category_Engineering * sharewomen,
                      data = new_data)
   store[i,] <- coef(mod)
}

cis <- c()
for (i in 1:11){
   sorted <- sort(store[,i], decreasing = FALSE)
   lower <- quantile(sorted, 0.025)
   upper <- quantile(sorted, 0.975)
   cis <- append(cis, cbind(lower, upper))
}
cis <- matrix(cis, nrow = 11, ncol = 2, byrow = TRUE)
final_model_summary <- summary(model.final)
final_model_summary <- data.frame(final_model_summary$coefficients)
variable_names <- row.names(final_model_summary)
cis <- data.frame(cis)

row.names(cis) <- variable_names
colnames(cis)  <- c("Lower", "Upper")
head(cis)
# Show an example histogram
hist(store[,3],
     main = "Histogram of the bootstrapped coefficients on sharewomen", xlab = "Slope")
```