Relationship between Gender or College with Starting Salary in the Philippines

Ethan Lee, Sean Ty, Jason Zhou, Al Xin

12/9/2020

0 Motivation

0.1 College

During the COVID pandemic, there have been questions raised about the value of colleges, in particular as certain colleges has moved to a remote setting, where a lot of benefits of colleges, such as the social networking as well as casual conversations with professors, have been curtailed by reality. This current question of the value of college has been met with a long history in economics of studying the effect of education on earnings, with some seminal articles including one in 1966 by Becker and Chiswick, as well as one written by Card in 1999 surveying the current field at the time. According to Card, this issue began trending in the late 1950s, due to its descriptive abilities with post-war growth (page 1802).

There has also been a history of working with education from different colleges, such as one written by Brewer, Eide, and Ehrenberg in 1996. Using measures of college quality, it focused on the effect of the quality of college on earnings, which is similar to what we are analyzing here with our first hypothesis. But, instead of focusing on American colleges, we are focusing on colleges in the Philippines. We also focus on the present-day as well, critical as the paper itself even noted that the 1980s cohorts had significantly more return from elite colleges than the 1972 cohort that they analyzed.

This allows us to further the current research, due to the changing time periods and as the Philippines has its own demographics and demands for skilled labor.

0.2 Gender

The motivation for understanding the impact of gender on first-time salaries stems from the current uncertainty in academia regarding whether recent efforts in the Philippines to close gender wage gaps in various industries have been truly effective; the Philippines was named the 8th best country in the world with regards to gender equality, but a gender wage gap was still found to persist in its job market. Furthermore, an Ipsos survey found that 75% of respondents in the Philippines believed that women receive the same amount of pay as men despite evidence showing otherwise. Thus, one goal of this study is to use up-to-date data that can illuminate whether a significant gender wage gap exists, especially when looking to first-time salaries. Looking at the recent time periods allows us to add onto previous research, even though we do not have the same data quality and breath they do.

1 Data source

Liyab is a startup that focuses on helping Filipinos transition to the labor market. Our data comes from a Liyab survey on the salary of respondents in their first jobs. Survey results are open to the public as a Google Sheets form and can be found at https://www.liyab.ph/liyab-first-pay-survey/. The survey is still taking responses; the data we are analyzing was downloaded on October 20th, 2020.

1.1 Predictors

The survey contains the following information from respondents:

- `year` (quantitative): the year when respondents had their first job.
- `industry` (categorical): The general industry that encompasses the respondents' occupation described by the respondent.
- `salary` (quantitative, response): The first-time monthly salary in PHP agreed upon for the respondent
- `univ` (categorical): The university attended by the respondent
- `gender` (categorical): The respondents' gender (self-reported)
- `negotiate` (categorical, binary): Indicator of whether and how the salary was negotiated

2. Hypotheses

With salary as our response variable, we have two hypotheses we would like to examine.

2.1 Gender bias

First, we want to test whether different genders earn different wages. Based on gender roles in the Philippines, we expect that males earn more than females. If there is sufficient data for non-binary individuals, we expect this group to earn less than members of the traditional gender binary. Furthermore, recent surveys have shown that residents of the Philippines believe that the gender wage gap is decreasing despite evidence showing otherwise. Thus, we hope to test these results using up-to-date data that reflects the current climate of the job market in the Philippines.

2.2 Effect of university

Second, we are interested in testing whether attending different colleges results in different first-time salaries. More specifically, we would expect that controlling for potential confounders, people from the more well-known colleges in the Philippines ("The Big Four") earn more than people from other colleges. Like our earlier hypothesis, we also focus on the present using up-to-date-data.

3. Exploratory data analysis

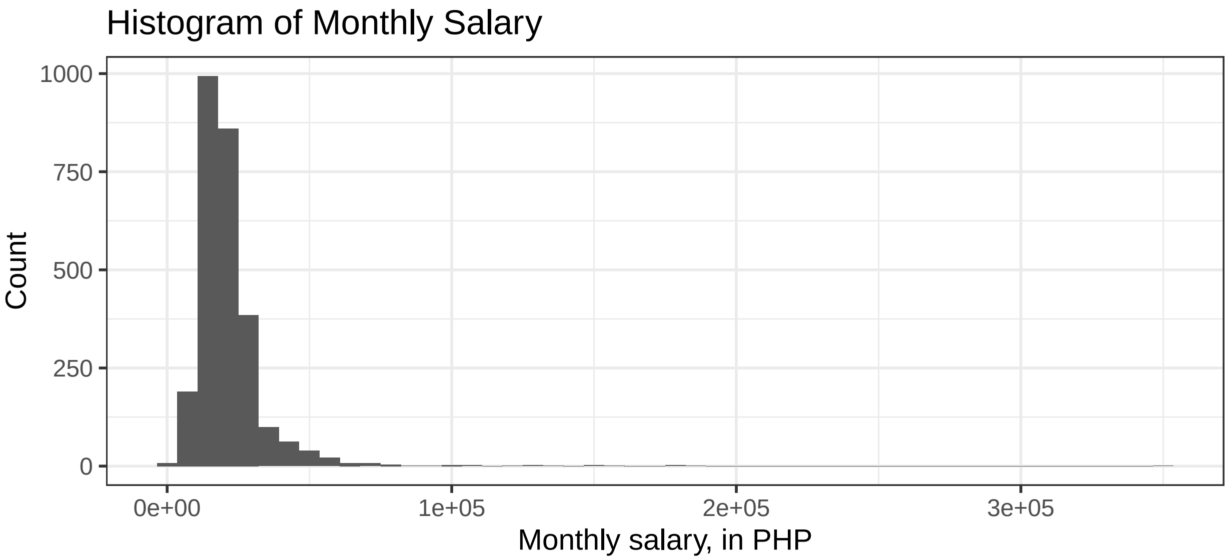
We cleaned the data manually, as the responses were non-standardized. We will address data cleaning in the appendix.

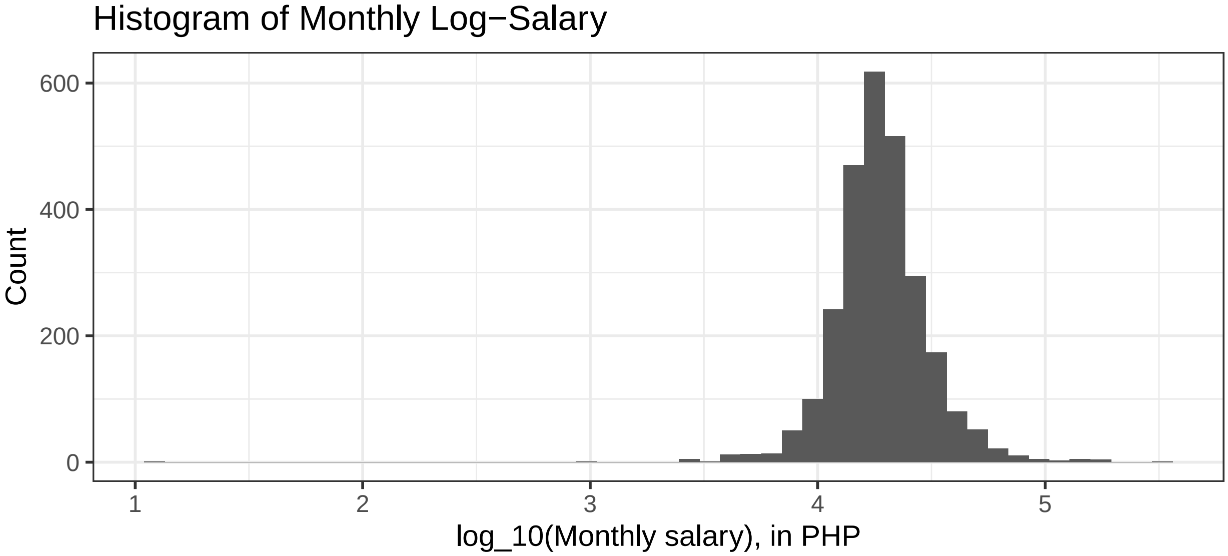
3.1 Salary

Salary is our response variable of interest. Respondents reported monthly salaries in units of PHP, the national currency of the Philippines. The histogram of salaries is right-skewed, with a measure of center, the median, at 18000 PHP, and a measure of spread, an IQR, of 9000 PHP. To address the right skew of the distribution, the base 10 log of the salary is taken, with the histogram of this transformed variable showing a more symmetric distribution, though this transformation introduces low outliers. Overall, we anticipate that using the transformed variable will be more consistent with analysis that requires an assumption of Normality compared to the untransformed variable. The center of the log-transformed salary is around 4.25 log PHP.

Observed Proportion of Genders in the Data:

Female	61.56%
Male	32.80%
Other	1.00%
N/A	4.64%





3.2 Gender

Using string pattern-matching, we were able to create categories for individuals who identified as female, male, and other. Around 61.56% of respondents identify as female, 32.80% identify as male, 1.0% responded as non-binary or some other ambiguous response, and 4.64% did not respond to this question. For the remaining analysis, we will group `NA` and non-binary respondents as `"Other"`.

Around 61.56% of respondents identify as female, 32.80% identify as male, and 4.64% responded outside of these two categories (the vast majority of this 4.64% is non-responses, or `NA`, and a small percentage responded as non-binary.) We grouped the variable gender into three categories: "M", "F", and "O", where "O" includes both `NA` responses as well as non-binary responses.

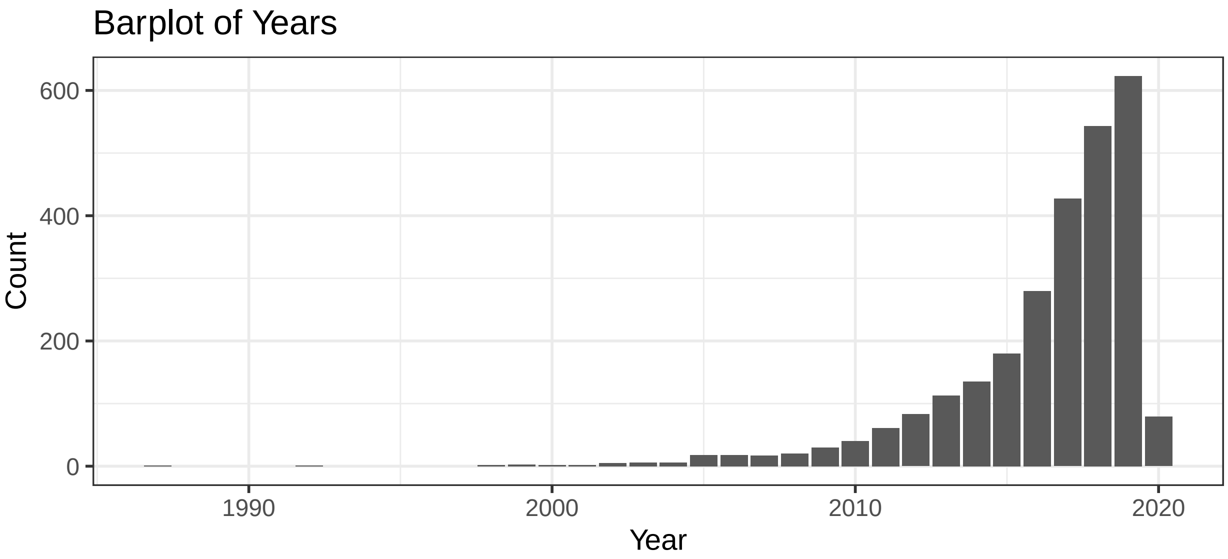
We have around double the number of females than males in our sample, which is surprising given that

Observed Proportion of Universities in the Data:

ADMU	11.16%
UPD	17.13%
UST	9.78%
DLSU	6.68%
Big Four	44.75%
NA	5.57%

the poll was mainly accessible via Twitter, and according to Statista (https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/), the vast majority of Twitter users are male (around 70% in October 2020). We are unable to explain why this is the case.

3.3 Year



Looking at the barplot of all the years, we see that it is very left-skewed; the data concentrates more towards the recent years. As it is skewed, it also suggests that we should use robust measures to estimate the center and spread. The center is at around 2017, its median, while its spread is around 4 years, its IQR. There are also a few low outliers, as seen in the barplot.

3.4 University

In the Philippines, the "Big Four" universities are regarded as prestigious and include ADMU, UPD, UST, and DLSU. To simplify the categorical variable of `univ`, we sorted for responses that referred to these universities. All other universities were grouped as `"other"`. We created an additional dummy variable `is_big4` that groups these universities together. Around 5.6% of respondents did not put an answer in this category.

Around 44.8% of respondents attended a Big Four university. This indicates we may perform reasonable analysis of the impact of attending a Big Four university on respondents' first-time salaries.

However, some sampling method may be required for this analysis, since the share of respondents who attended Big Four universities is not evenly split among the four universities; of the four colleges, the institution most attended by respondents was UPD, with 17.1% of respondents, while the institution least attended was DLSU, with 6.7% of respondents.

3.5 Industry

Industries were manually grouped into 41 unique categories. A table of their counts is large, so it was put in the appendix. Using this sorting, IT, Finance, Communications, Sales, and Retail were the five most common industries. There are 9 industries with over 100 observations and 7 industries with less than 10. Outside of investigating salary relative to industry, we may potentially use this predictor variable to understand whether any statistically significant differences in first-time salary existed between popular and less popular industries.

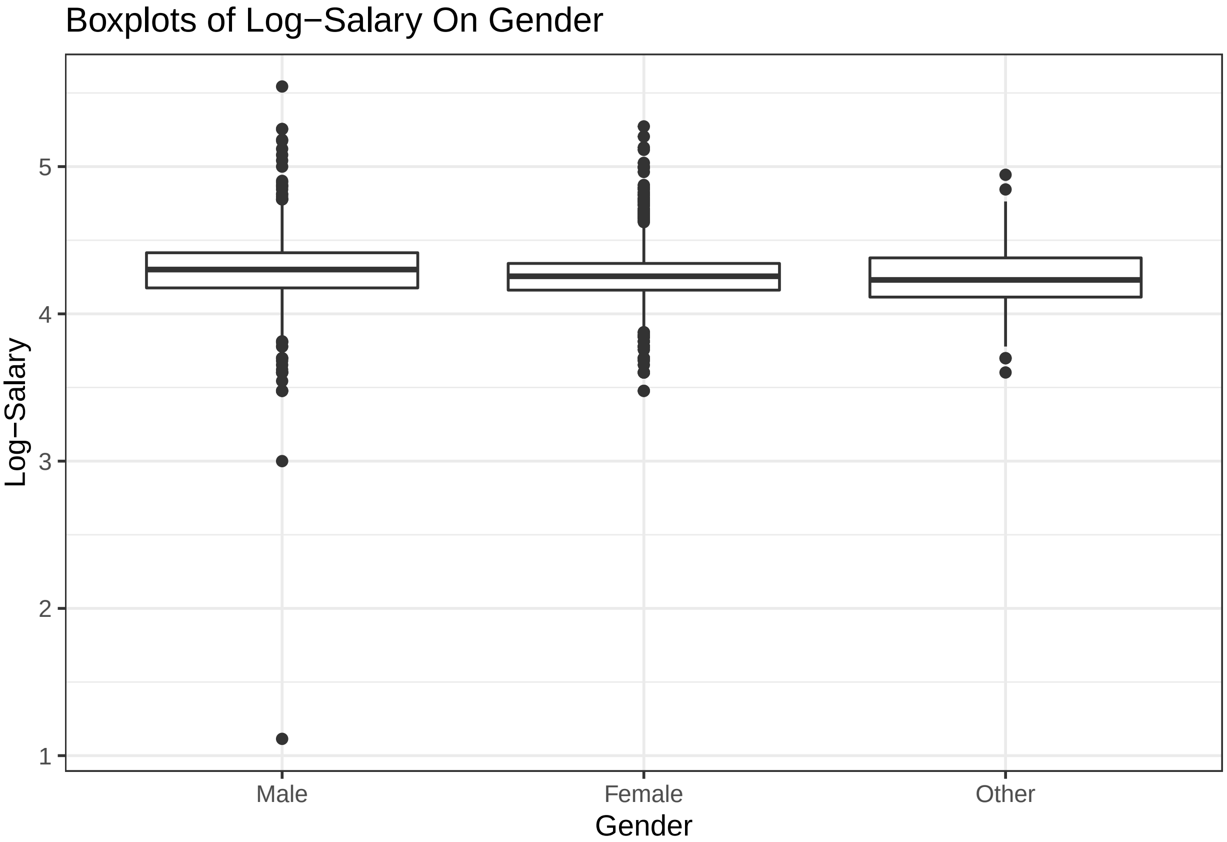
4 Exploration of relationship between salary and predictors

We would first like to investigate the variables relevant in our hypotheses. We can perform initial visualizations and tests to examine whether gender or school are associated with differences in mean salary.

4.1 log(Salary) vs. gender

A collection of initial boxplots showing the distribution of first-time log-salaries across different genders reveal that there is an outlier in the males category, something that we will discuss later when we make our baseline model.

Looking at the boxplots closer, males seem to have a slightly higher median log-salary than females or other, which suggests that our hypothesis about genders having different salaries merits further investigation.



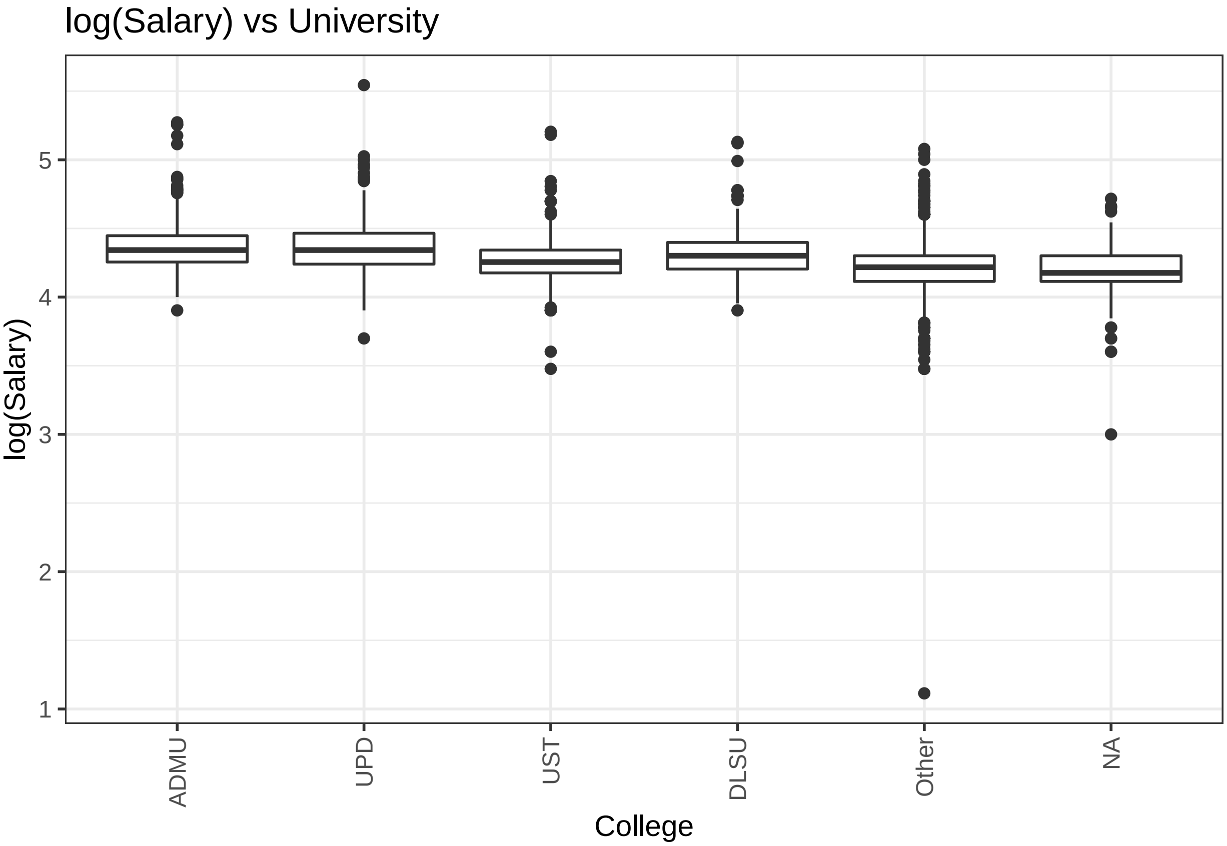
Conducting a ANOVA test between the genders with a null hypothesis that all the mean log-salaries are the same, and an alternative hypothesis that there is at least one pair of different mean log-salaries (and $\alpha = 0.05$), we get a F-statistic of 14.26 with $df = 2, 2692$ and a p-value of 6.92×10^{-7}, which is less than α. Note that the R output can be found in the appendix. This means that we reject the null hypothesis; there is statistically significant evidence that the mean log-salaries between the genders are different. While comparing means is different than comparing medians, this does support the idea that the center of the distributions of log-salaries between genders are different.

For the validity of the ANOVA test, we assume independence due to lack of reasonable dependence in responses (this may merit further research). From the boxplots, due to the relatively equal number of high and low outliers, we see that the distributions are approximately Normal, but we also needed to check constant variance. Looking at the ordered list of variances below, we see that the lowest variance is around 0.0342 and the highest variance is around 0.0613. As the latter is not more than three-fold the former, this implies that the constant variance assumption is satisfied and our conclusions from the earlier ANOVA test is valid.

	x
Female	0.0341803
Other	0.0409560
Male	0.0631371

4.2 log(Salary) vs. school

A collection of boxplots showing the distribution of first-time log-salaries across different colleges reveals that respondents who attended each of the "Big Four" universities seem to have a higher median salary than respondents who attended other colleges, with ADMU, UPD, and DLSU having higher outlier salaries as well. Respondents who attended UST in particular have a salary center and spread similar to that of respondents who attended other universities, although UST also has more high outliers than low outliers. Nevertheless, these boxplots suggest that whether the "Big Four" universities offer more prestige in terms of first-time salaries is an interesting question.



We conduct an ANOVA test between the colleges with the null hypothesis that all the mean log-salaries are the same, the alternative hypothesis that there is at least one pair of different mean log-salaries, and $\alpha = 0.05$. We get a F-statistic of 67.45 with $df = (4, 2540)$ and a p-value of 2×10^{-16}, which is less than our chosen α. We reject the null hypothesis; there is statistically significant evidence that the mean log-salaries between the colleges are different. While comparing means is different than comparing medians, this does support the idea that the center of the distributions of log-salaries between colleges are different.

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## univ           4  10.78   2.696   67.45 <2e-16 ***
## Residuals   2540 101.54   0.040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 150 observations deleted due to missingness
```

For the validity of the ANOVA test, through the voluntary sampling method, we are uncertain whether the independence assumption is violated. From the boxplots, due to the relatively equal number of high and low outliers, we see that the distributions are approximately Normal. Looking at the ordered list of variances below (where "x" denotes "variance"), we see that the lowest variance is around 0.0348 and the highest variance is around 0.0441. As the largest variance is smaller than three-fold the smallest variance, this implies that the constant variance assumption is satisfied. We can assume that we meet assumptions for the ANOVA test.

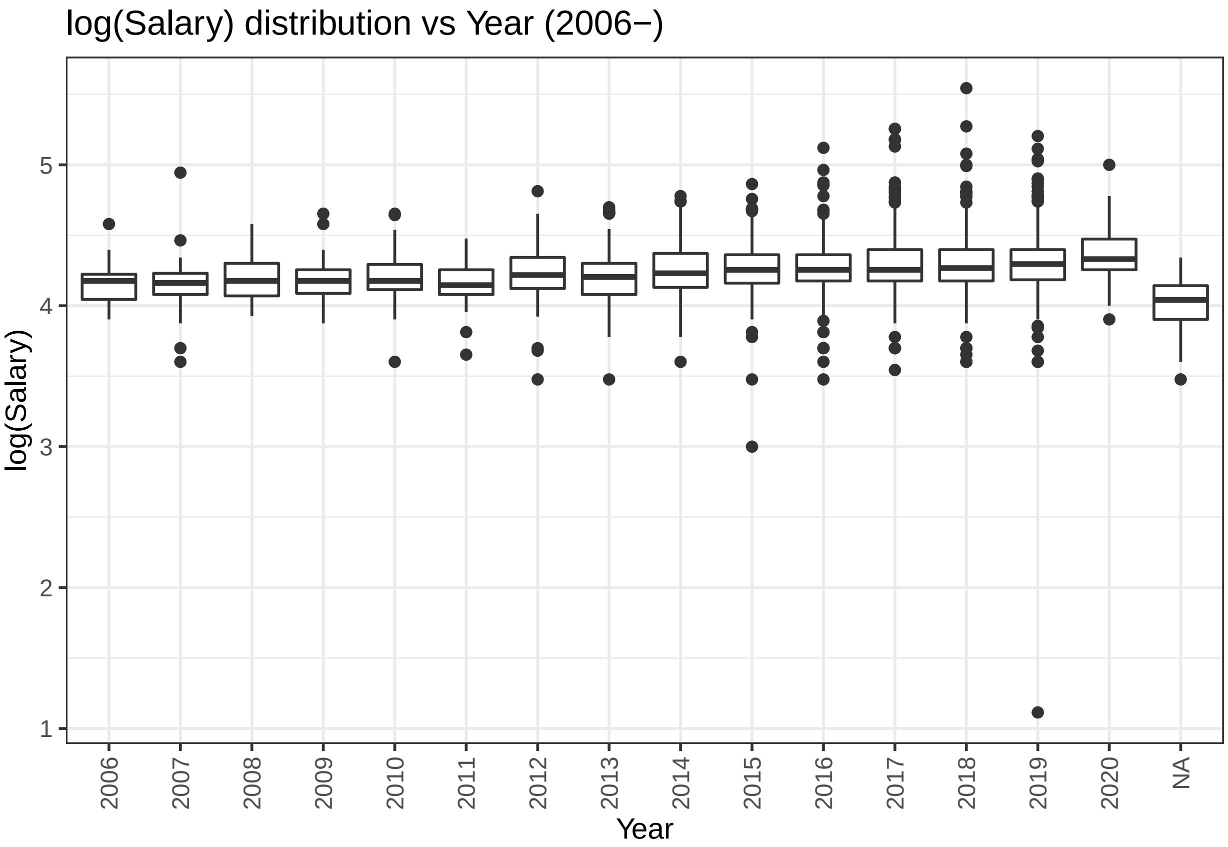
	x
UST	0.0347555
UPD	0.0376149
DLSU	0.0390912
Other	0.0413835
ADMU	0.0417097
	0.0441345

Outside of our hypotheses, we would like to visualize whether there are associations between other variables and salary. This will provide guidance as we investigate interaction effects and confounding variables.

4.3 log(Salary) vs. year

When investigating our data, we found that there are far fewer respondents from the 1980s to early 2000s. Though this data may be useful later, for exploration purposes we will consider only salaries in more recent years.

A visualization the distribution of salaries from 2006 to 2020 reveals that the median log-salary has gradually increased over time. Additionally, the spread of salaries has also seemed to increase over time, perhaps due to a diversification of available occupations (and salaries) for first-time job applicants in recent years.



We would like to investigate whether mean log-salaries are the same over time. We conducted an ANOVA test between the recent years (refer to appendix for results). Our null hypothesis is that all mean log-salaries are the same. Our alternative hypothesis is that there is at least one pair of different mean log-salaries. We consider $\alpha = 0.05$. After performing the test, we receive an F-statistic of 6.544 with $df = (14, 2634)$. Our p-value is 3.39×10^{-13}, which is significant based on our selected α. We have evidence to reject the null hypothesis and consider that there is statistically significant evidence that the mean log-salaries between the years are different. While comparing means is different than comparing medians, this does support the idea that the center of the distributions of log-salaries between years are different.

4.3.1 ANOVA assumptions

Here, through the voluntary sampling method, we are uncertain whether the independence assumption is violated. From the boxplots, due to the relatively equal number of high and low outliers, we see that the distributions are approximately Normal. Based on the ordered list of variances below, we see that the lowest variance is around 0.0234 and the highest variance is around 0.086. This violates the rule of thumb that the highest variance should be less than three-fold that of the lowest variance. This means that the constant variance assumption is likely violated.

```
##       2011       2009       2006       2008       2010       2014       2012 
## 0.02338348 0.02528083 0.02663444 0.02901554 0.03395151 0.03648352 0.03740342 
##       2016       2018       2020       2013       2017       2015       2019 
## 0.03742518 0.03891037 0.04019546 0.04240185 0.04332153 0.04504707 0.05062793 
##       2007 
## 0.08633574
```

4.3.2 Non-parametric test

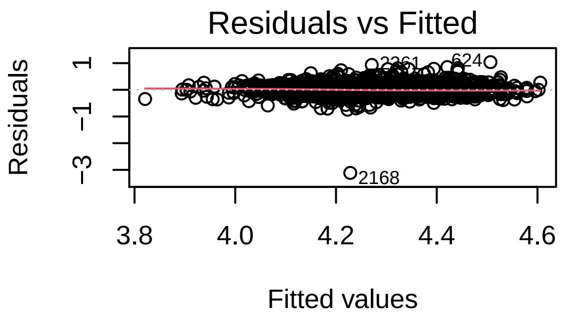
```
## 
##  Kruskal-Wallis rank sum test
## 
```

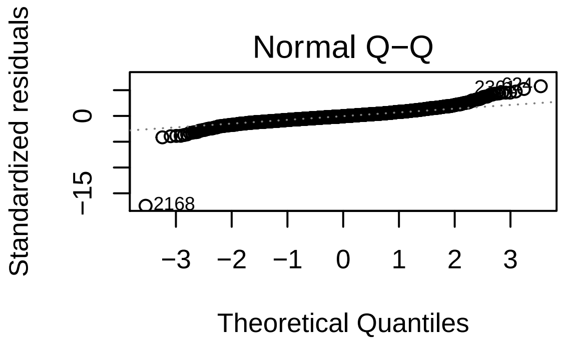
```
## data:  log_salary by yearfiltered
## Kruskal-Wallis chi-squared = 123.06, df = 14, p-value < 2.2e-16
```

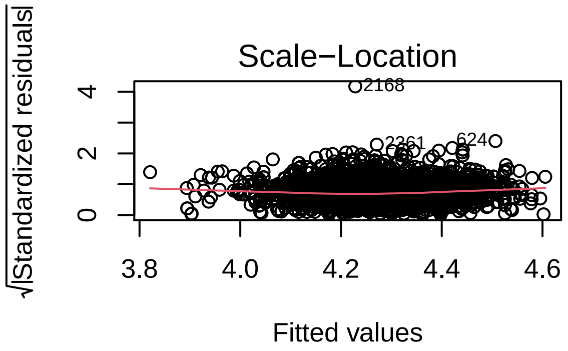
The violations in ANOVA assumptions suggest that to further prove that the centers of these distributions are different between years, we should use a non-parametric test as well. We will use the Kruskal-Wallis test, which is an extension of the Wilcox rank-sum test for multiple groups. With a null hypothesis that the median log-salaries between years are equal, and the alternative hypothesis that the median log-salaries between years are not all equal, we get a chi-squared statistic of 123.06 with $df = 14$, which also gives a p-value much less than 0.05. Therefore we reject the null hypothesis, which means that there is statistically significant evidence that the median log-salaries between years are different. This again supports the idea that the centers of the distributions of log-salaries between years are different.

5. Baseline model

Below, we consider the linear model fitting log of salary on the year, university, industry, gender, and negotiate (all available predictors in the dataset). We then proceed by checking assumptions.







5.1 Assumptions

We first check for assumptions. For independence, we note that there may be some reason to believe that this is not true. For example, this survey may have been made popular also by word-of-mouth, which makes it likely that in a family, if one person in the family fills out the survey, all working members of the family may fill out the survey. This leads to some correlation of results. Also, all the observations are correlated as well, as they would most likely have learned about the survey from Twitter.

For linearity, we see that the residuals are scattered around the line y = 0, which implies that the linearity assumption is reasonable. We do see that there is an outlier in the residuals vs fitted plot though (pointed out in the R code).

For Normality, from the Q-Q plot, we can see that the residuals do not follow the Q-Q line when we reach negative or positive residuals. Therefore, the Normality assumption is violated.

For constant variance, we see from the Residual-Fitted plot that there is approximately constant variance, as there is no fanning out or in at any point. This is supported by the Scale-Location plots, where we see that the LOESS curve fitted onto the plot is relatively a straight horizontal line, which also suggests that the constant variance assumption is reasonable.

5.1.1 Outliers

For residuals vs leverage, we see that the outlier seen in the residuals vs fitted plot appears again; the point from row 2168 seems to be an outlier. Looking closer as to why it is an outlier, we see that the self-reported salary was 13. This is nonsensical, as 13 PHP is approximately a quarter in USD, and a person would normally not be paid a quarter for their first month. As the person worked for a nonprofit, it is more likely for them to be unpaid as a volunteer or paid some baseline minimum wage instead of 13 as well.

There are multiple interpretations of 13 as well: the person could have meant 13 PHP, 13000 PHP, or 130000 PHP (the largest being within the range of the salaries seen in the dataset). That, including the fact that it could've been a typo, means that we will not try to "edit" the data. Also, as this is a baseline model to be built upon and compared to, we will also not remove this data point, as we are unsure if the input was actually a typo or not.

```
##      year  industry              role salary  univ gender negotiate
## 2168 2019 NONPROFIT TRAININGASSISTANT     13 Other   Male        No
```

5.2 Interpretations

Note that when we interpret coefficients, we are implicitly assuming that we are holding all the other variables in the baseline model constant. We note this here because we don't want to write that we are holding all the other variables constant out every time we interpret a coefficient.

5.2.1 Gender identity

The baseline (intercept) of the model assumes that the respondent is male, and the coefficient estimate for the binary variable indicating whether the respondent is female compared to being a male shows that there is a statistically significant decrease in salary when comparing a female and male respondent (p-value < 0.05).

Specifically, when a respondent's gender is reported to be female, a decrease of 0.0362 in the log of first-time salary is predicted by the model. Additionally, if the respondent's gender is not male or female (i.e. non-binary or not reported), compared to being a male, the log of the first-time salary is predicted to decrease by 0.0157 , although this decrease is not statistically significant (p-value > 0.05). Thus, these initial baseline coefficient estimates indicate that male respondents received higher pay than female respondents, potentially showing some signs of gender bias.

5.2.2 University attended

The baseline (intercept) of the model assumes that the respondent attended Ateneo de Manila University (ADMU). The coefficient estimate for the binary variable indicating whether the respondent attended UPD compared to ADMU shows that attending UPD is associated with a statistically insignificant ($p > 0.05$) decrease in log of first-time salary of 0.0132.

The coefficient estimate for the binary variable indicating whether the respondent attended UST compared to ADMU shows that attending UST is associated with a statistically significant ($p < 0.05$) decrease in log of first-time salary of 0.110.

The coefficient estimate for the binary variable indicating whether the respondent attended DLSU compared to ADMU is associated with a statistically significant ($p < 0.05$) decrease in log of first-time salary of 0.0546.

Lastly, the coefficient estimate for the binary variable indicating whether the respondent attended a college outside of the "Big Four" compared to ADMU is associated with a statistically significant ($p < 0.05$) decrease in log of first-time salary of 0.156.

The decrease in first-time salary associated with attending a college outside of the "Big Four" was the greatest in magnitude – thus, these initial baseline coefficient estimates indicate that attending a "Big Four" university is associated with higher first-time salary, potentially showing some signs of college bias. In the future, we will investigate a model grouping all Big Four colleges together as well as examine interactions.

5.3 Removing Outliers and Cutting off by Year

Looking at the Cook's distance as well, we see that the person who earned 12 PHP on their first month is extremely influential with respect to the other variables. Due to the fact that it is likely a user input error and that it is extremely influential, we decided to remove that point.

We then decided to look at the number of observations in each year. We see that for each year before 2009, there were less than 30 observations, with many having less than 5 observations for further out years, and some only having one observation. This suggests that we need to further clean the data to better answer our question of the recent assocations between first pay and gender, as well as first pay and college. Due to our focus on recent associations, combined with the slight concern of overfitting for the years that are extremely far out, we decided to truncate the data to data that is only 2009 and above.

6 Non-parametric test

We will use the transformed log version of the variable of `salary` in later analyses where Normality is an important assumption. However, we would like to also investigate the untransformed variable with non-parametric tests.

In our investigation of gender, we will examine the variable `is_male`. Narrowing our scope slightly from considering `Other` responses to gender outside of the female/male binary, this choice of variable is most relevant when considering the question of whether traditional gender roles are correlated with salary differences.

Similarly, in investigation of college, we will examine the variable `is_big4` rather than consider all Big Four colleges independently.

Our null hypothesis for gender is that the salary of male and non-male first-time employees is the same. Our alternative hypothesis is that males have higher first-time salaries.

Our null hypothesis for college graduation is that the salary of Big Four graduates and non-Big Four college graduates is identical. Our alternative is that Big Four graduates have higher first-time salaries.

From exploratory analysis, we know that based on simple averages, the mean salary for those identifying as male and those graduating from a Big Four college are higher.

Wilcox rank sum test

In both cases, we will perform a two-sided Wilcox rank sum test.

Gender

The test performed on `salary` against `is_male` is significant, with $p = 1.64 \times 10^{-14}$. This provides evidence to reject the null. The alternative is that males tend to have higher salaries.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  salary by is_male
## W = 599490, p-value = 1.641e-14
## alternative hypothesis: true location shift is not equal to 0
```

College

The test performed on `salary` against `is_big4` is significant, with $p < 2.2 \times 10^{-16}$. This provides evidence to reject the null. The alternative is that Big Four graduates have higher salaries.

```
## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  salary by is_big4
## W = 487435, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Bootstrapping

The models have been generated pseudorandomly with respect to the random seed 139. Results have been saved.

Gender

A bootstrap performed with 10 000 replications yields a basic 95% confidence interval of $(2525, 5343)$ PHP (where salary from male employees is greater). The confidence interval lies entirely above zero, providing evidence that the salary for males is higher.

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
## 
## CALL : 
## boot.ci(boot.out = gender_boot, type = c("norm", "basic"))
## 
## Intervals : 
## Level      Normal              Basic         
## 95%   (2559, 5436 )   (2525, 5343 )  
## Calculations and Intervals on Original Scale
```

College

A bootstrap performed with 10 000 replications yields a basic 95% confidence interval of $(5192, 7732)$ PHP (where salary from Big Four graduates is greater). The confidence interval lies above zero, providing evidence that the salary for males is higher.

The CI for difference in Big Four/non-Big Four colleges is disjoint and greater than the estimate for gender. This provides tentative evidence that the increase in salary for Big Four graduates over non-Big Four graduates is greater than that of males over non-males.

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
## 
## CALL : 
## boot.ci(boot.out = big4_boot, type = c("norm", "basic"))
## 
## Intervals : 
## Level      Normal              Basic         
## 95%   (5269, 7792 )   (5192, 7732 )  
## Calculations and Intervals on Original Scale
```

Permutation

When performing the test, a two-sided test is used even though the alternate hypothesis is one-sided. This helps increase the accuracy (reduce frequency of Type I errors) in the test as the lower bound for significant t-value in one-sided tests is more generous that two-sided tests.

Gender

A permutation test using ‘perm::permTS“ indicates that with 999 MC replications, we have evidence suggesting a significant difference in the salaries of male vs non-male workers ($p = 0.002$.) The sample estimate of difference is 3552.842 PHP in favor of males.

```
##
##  Exact Permutation Test Estimated by Monte Carlo
##
## data:  salary by is_male
## p-value = 0.002
## alternative hypothesis: true mean is_male=0 - mean is_male=1 is not equal to 0
## sample estimates:
## mean is_male=0 - mean is_male=1
##                       -3552.842
##
## p-value estimated from 999 Monte Carlo replications
## 99 percent confidence interval on p-value:
##  0.00000000 0.01057916
```

College

A permutation test indicates that with 999 MC replications, we have evidence suggesting a significant difference in the salaries of Big Four vs non-Big Four graduates ($p = 0.002$). The sample estimate of difference is 6505.069 PHP in favor of Big Four graduates.

Again, we have a trend where the difference between Big Four and non-Big Four salaries appears to be wider than that of males and non-males.

```
##
##  Exact Permutation Test Estimated by Monte Carlo
##
## data:  salary by is_big4
## p-value = 0.002
## alternative hypothesis: true mean is_big4=0 - mean is_big4=1 is not equal to 0
## sample estimates:
## mean is_big4=0 - mean is_big4=1
##                       -6505.069
##
## p-value estimated from 999 Monte Carlo replications
## 99 percent confidence interval on p-value:
##  0.00000000 0.01057916
```

Comments

These non-parametric tests support the inference that some difference in salaries exists between gender groups or graduates of different colleges. However, we cannot make claims of causality. We have additional evidence supporting a correlation between gender identity and starting income as well as evidence supporting a correlation between college and starting income.

It is possible that disparities in the data are the result of confounding variables. For example, if we find that `INDUSTRY` (or specific classes in `INDUSTRY`) can be predictive of salary, it raises the question of whether income differences is due to different demographic representation in different industries.

7 Sequential Variable Selection; New Baseline; Interaction; Mixed Effects

Before proceeding to gain further insights into the data through various statistical tests and modeling, the dataset was split at an 80-20 ratio to create train and test datasets for future model comparisons.

We will also take industries with less than 10 entries and group them under `"OTHER"` to simplify the model building.

7.1 Fitting linear models: All variables and interactions

First we consider what level of polynomials year can be, as year is the only numeric predictor we have. Doing sequential selection for only year, we see that the higher polynomials of year were dropped out when we use AIC. This suggests that it is okay for us to consider only year as a linear effect.

Doing a Chi-squared test to check for correlation between university and gender, we get a Chi-squared statistic of 10.451 with $df = 8$ and a p-value of 0.2348. This means that we do not reject the null hypothesis that there is no dependence between the university a person goes to and their gender.

Fitting with all of the variables for a second baseline model, we see that university is a very important predictor in general. We also see that the coefficient estimates for college check out with what we expect. Compared to ADMU, we see that the other colleges are associated with lower income. From highest to lowest, we see that UPD, DLSU, UST, and then Other do worse.

```
##
##  Pearson's Chi-squared test
##
## data:  table(liyab_nona$univ, liyab_nona$gender)
## X-squared = 10.451, df = 8, p-value = 0.2348
```

	df	AIC
lm_poly	45	-1399.595
lm_all	42	-1401.436

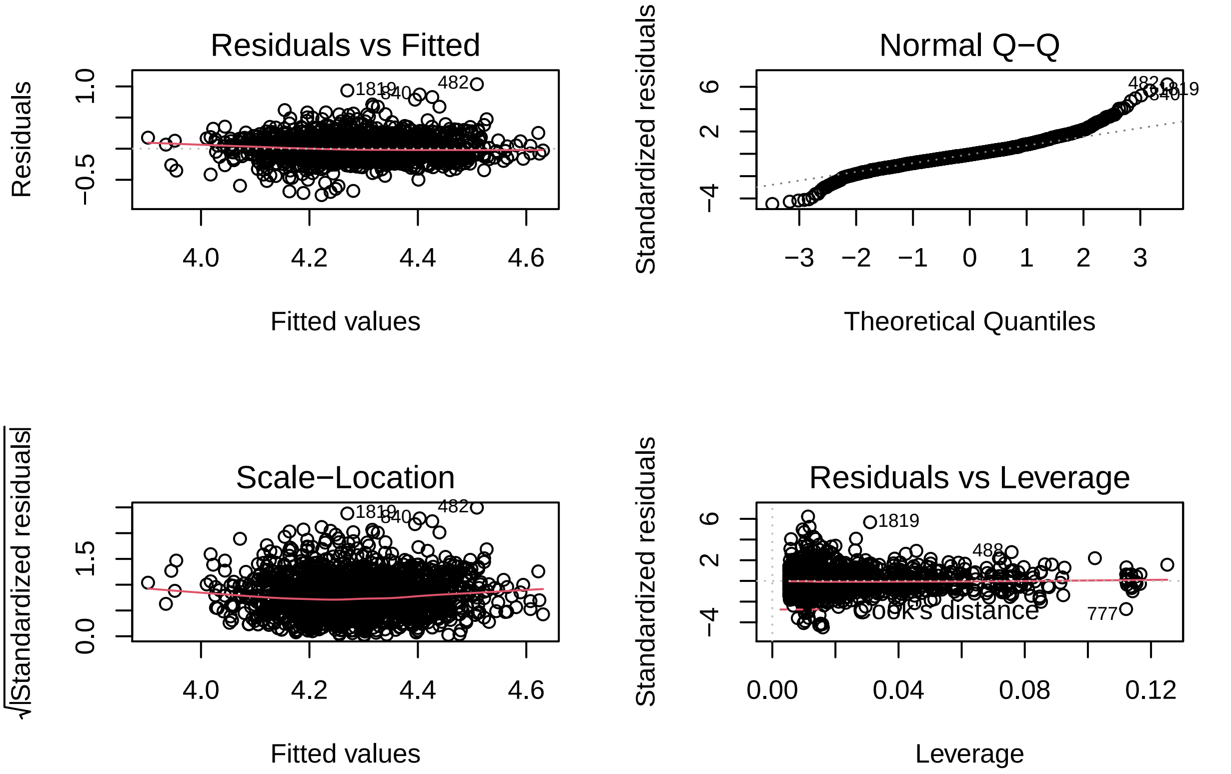
	df	BIC
lm_poly	45	-1148.532
lm_all	42	-1167.110

7.1.1 Interpretations for lm_all

According to `lm_all`, although there is risk of overfitting, gender remains an important predictor; there is statistically significant evidence that females earn less than males. Female are associated with earning around $1 - e^{-0.042882} = 4.2\%$ less than males.

7.1.2 Assumption Checking for lm_all

We now check assumptions for the whole model without interaction effects.



The first plot suggests that the residuals have about mean zero, except possibly for small values of `log_salary`, suggesting that the linearity assumption is satisfied for the most part. Homoskedasticity seems to be roughly fine as well from the residual plot. The second plot, on the other hand, suggests that our distribution of residuals is heavy-tailed, so Normality could be an issue although this can be remedied by the Central Limit Theorem. The Scale-Location plot suggests that as we estimate extremely low or high values, our residuals increase in magnitude, although the fitted line looks roughly horizontal, so it shouldn't be a big problem. The residuals vs leverage plot suggests that some observations still have more leverage than others.

The plots for polynomial regression are very similar, and discussion will be omitted (the plots can be seen in the appendix).

7.2 Sequential Selection

We now consider interaction effects and try to control for observable confounders through sequential variable selection.

Using AIC as the criterion, we see that the final model from sequential selection uses industry, university, year, gender, as well as the interaction between year and gender as predictors. Although we cannot interpret the p-values due to multiple testing issues, we see that at the very least university as well as gender remain in the model, which are positive signs of at least some usefulness of gender and university as predictors for log-salary.

As for industries, sequential and regular OLS models seem to agree that: architecture, education, labor, have lower wages than average while data, energy, it, legal, retail have higher wages than average.

```
## Sequential model formula:

## log_salary ~ industry + univ + year + gender + year:gender
```

7.3 Mixed Effects Modelling

We proceeded by creating two mixed effects models, one of which accounted for differences in average salary that might exist between different job industries, and the other of which accounted for different salaries that might exist in different years. Note that the formulas for both mixed effects models came from our earlier sequential modelling.

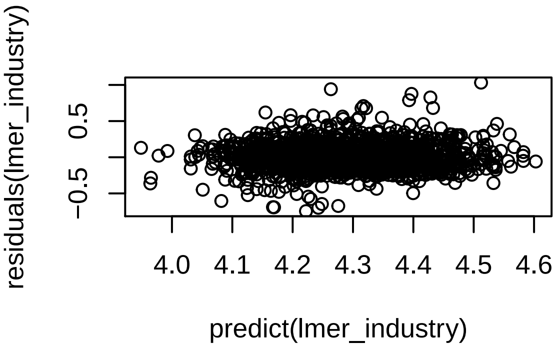
The first model used a random intercept clustered by industry and used the university, year, gender, and interaction effect between year and gender as fixed effects; these effects were used to account for the information discovered from the previously implemented sequential step model. The fixed effect coefficient estimates of this model showed that attending a university outside of the “Big Four” was associated with a statistically significant decrease in first-time salary compared to attending ADMU ($\beta = -.1424$, $t = -11.292$). Furthermore, among UPD, UST, and DLSU, attending UST was associated with the largest decrease in first-time salary compared to attending ADMU ($\beta = -0.1024$, $t = -6.157$). Additionally, the fixed effect estimates associated with gender and the interaction between year and being female were not found to be statistically significant.

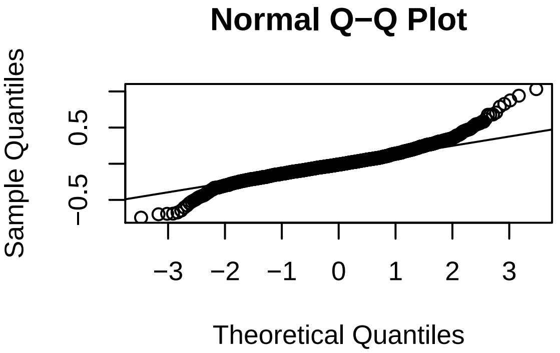
The second model used a random intercept clustered by year as well as a random slope assigned to gender. The fixed effect coefficient estimates of this model show similar results to the previous model, with attending a university outside the “Big Four” associated with a statistically significant decrease in first-time salary compared to attending ADMU ($\beta = -.1424$, $t = -11.255$). In this model, however, being female was associated with a statistically significant decrease in first-time salary compared to male employees ($\beta = -0.03134$, $t = -3.041$).

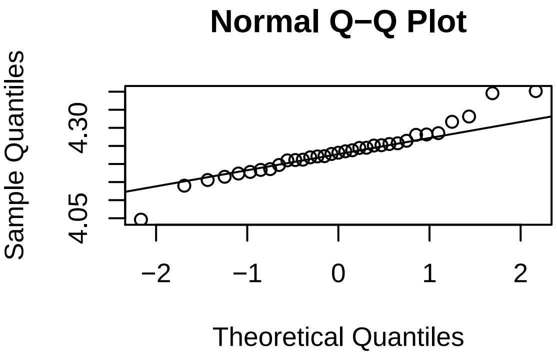
7.3.1 Assumption Checking

7.3.1.1 Industry Model

Checking the residual plot, we again see that linearity and constant variance are relatively satisfied by the first plot below. Unfortunately, Normality of residuals is still an issue as can be seen by the QQplot below, but the random intercepts do seem Normally distributed based on the random intercept QQplot.

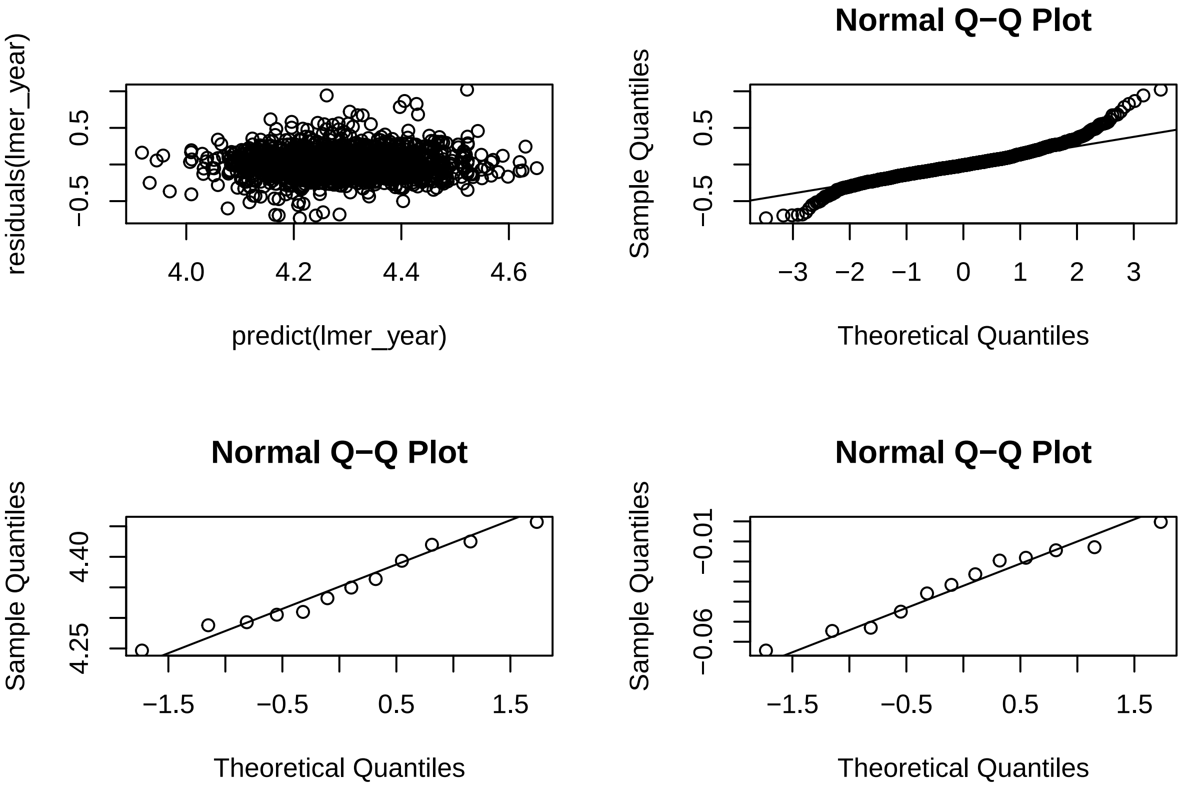






7.3.1.2 Year Model

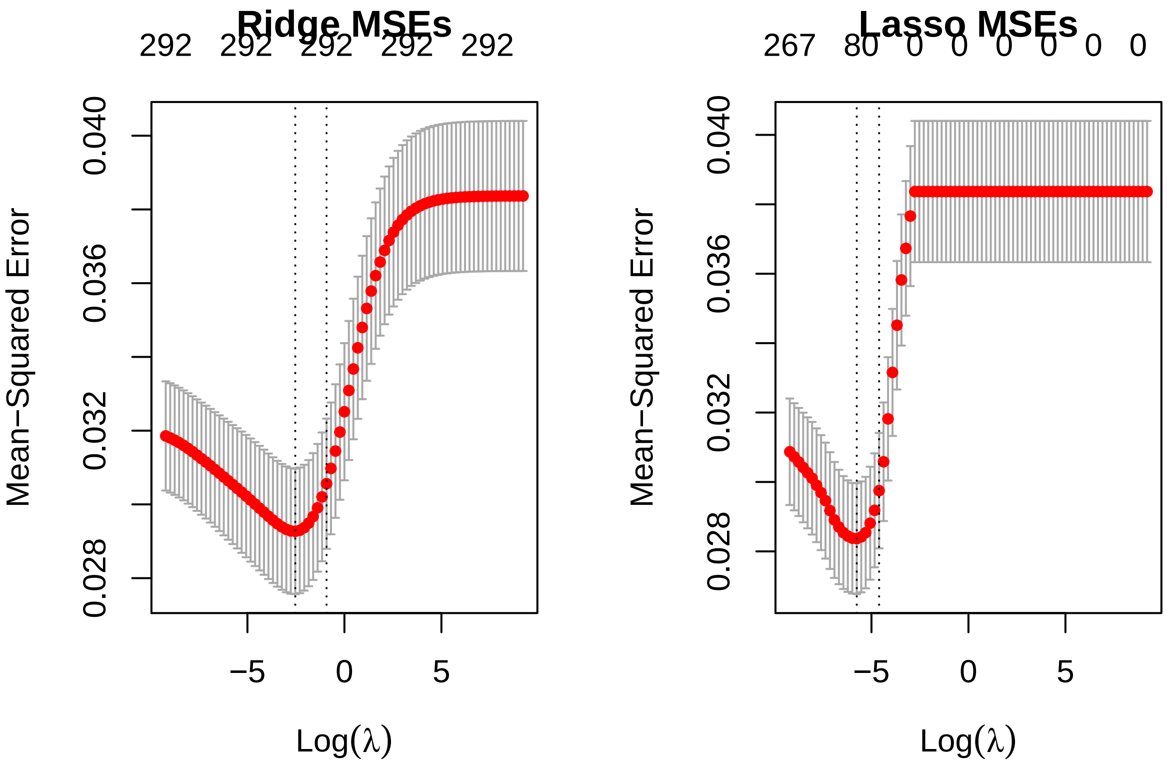
Checking the residual plots for the mixed effects model using year as a clustering variable, we again see that linearity and constant variance are relatively satisfied. Unfortunately, Normality of residuals is still an issue based on the first QQplot below. The random intercepts and random slopes do seem Normally distributed based on their corresponding QQplots.



8 More Flexible Models

8.1 Ridge + Lasso Models

We can also look to cross-validated models to further understand the impact of gender and university in strong predictive models. Although we can suspect that a LASSO model would be preferred in this case as there are most likely only a few predictors in the data with a strong effect on the first-time salary. However, we can compare a ridge and LASSO approach by comparing their mean cross-validated errors as below.

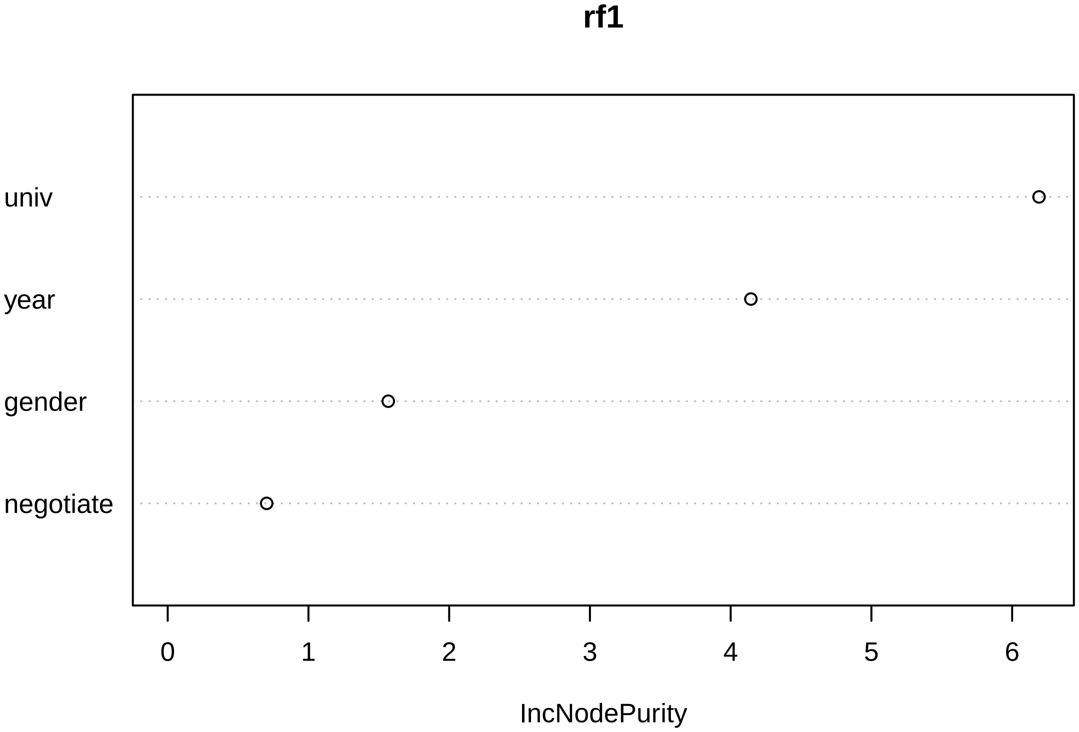


As shown, the LASSO approach produced a lower mean cross-validated error at 0.028, further showing why it is chosen over the Ridge approach.

Using cross-validation to select the best lambda value for a final LASSO model, the optimal lambda value was found to be 0.003162278. The predictor variables indicating whether an employees attended DLSU was not deemed to be significant by the final LASSO model, and the predictor indicating whether an employee identified with a gender other than male or female was also not found to be significant. Furthermore, negative coefficient estimates were associated with each of the predictors indicating whether an employee attended a university outside of the "Big Four" or attended UST or was female; however, numerous interaction effects were included in the final LASSO model, indicating that confounding variables most likely existed with regards to which industry an employee worked in, when the salary was decided, and whether it was negotiated.

8.2 RF Model

Due to the large amount of interaction effects included in the LASSO model, a random forest model was fitted as well in order to account for them.



First, the data was split into train and test datasets at a 40-60 split. Then, the values tested for maximum nodes used in the random forest were 2, 5, 10, 50, and 100, while the values used for the number of predictors tried at each split were 1, 2, 3, and 4, and 50 trees were used for the model. The optimal corresponding values were found to be 50 and 4 respectively. Based on the RMSE values calculated for the train and test sets used to generate the random forest model, there doesn't seem to be any overfitting, with the RMSE's very close to each other in value. However, the model only explains 8.46 percent of variance, a very low value compared to previous models. Nevertheless, we can still use the model to identify the most important variables, which were found to be the university attended and the year; gender was the third-most important predictor, although around 4 times less than the university predictor.

9 Model comparison

With all predictive models now generated, we can compare their performances in terms of R^2 values, AIC, BIC, and RMSE with relation to the pre-split test dataset where applicable.

	name	rsquared	rmse_test	rmse_train	aic	bic
1	lm_all	0.28	0.18	0.17	-1401.44	-1167.11
2	lm_interact	0.39	0.19	0.15	-1236.24	348.24
3	lm_poly	0.29	0.18	0.17	-1399.59	-1148.53
4	lm_seq	0.29	0.18	0.17	-1405.72	-1165.82
5	lmer_industry	0.15	0.18	0.17	-1285.84	-1218.89
6	lmer_year	0.25	0.18	0.17	-1139.41	-877.19
Rsquared	model_lasso	0.20	0.18	0.16		
11	model_rf		0.19	0.18		

Using the table above, we can see that using the test RMSE value (which is applicable for all of the generated models), the mixed effects model using industry to cluster the data with a random intercept performed the best with the lowest RMSE at 0.1783645. Additionally, the train data RMSE was found to be 0.1655381, a very close value to the test data RMSE, thus showing that the strong performance of the model was not due to overfitting.

10 External Validity

We can be almost sure that this is not representative of the general Philippine population. Citing general results from the EDA, we saw that a large portion of the respondents reported to be female. Note that a large number of respondents also came from the Big 4, even though the Philippines as an enrollment population of over 3.4 million. This means that this population is especially educated as well. In general, these observations combined by the fact that this is a survey sample suggests that we have little external validity. In particular, these results can be generalized to those that would respond to the Liyab survey, although we are unable to accurately pinpoint the particular characteristics of this population to exactly say which part of the general Philippine population this actually is. Nevertheless, having some room for generality is better than not having any. Note that this lack of generality could have also affected our results: it is more likely that those who are educated have greater absolute wages, and so those with more education may also have higher differences between wages, although this is unclear directly from the data. Similar things occur with gender as well, where distributionally, the preferences of each gender may still be different even after controlling for industry, with one possible reason because of social expectations mixed with culture.

11 Conclusions

Because the models generated in this study were predictive, any causal claims cannot be made regarding the hypotheses we outlined. Furthermore, we can look more deeply into the coefficient estimates of our best-performing predictive model in order to gain insights on whether gender and university were found to be significant, as well as what changes in first-time salary they were associated with.

Based on the calculated RMSE values, the best-performing model was the mixed effects model using a random intercept clustered by industry and using gender, university, year and the interaction effect between gender and year as fixed effects. Firstly, this model suggests that the relationship between university attended and first-time salary is significant. Additionally, based on this model, attending a university outside of the "Big Four" was associated with a decrease in first-time salary when looking to the fixed effect estimates, matching the hypothesis statement that attending a "Big Four" university is associated with a higher first-time salary. Regarding the second hypothesis on gender's role in first-time salary, the mixed effects model indicated that female employees earned higher first-time salaries than male employees based on fixed effect estimates, although this was not found to be statistically significant ($t = 0.9668$); thus, this result may have been due to confounding variables caused by multicollinearity. The correlation matrix of the model shows that the predictor indicating whether an employee is female and the variable for year had a correlation of 0.741, a high value.

University and gender are important predictors in other models, in particular, in our Lasso and random forest models. A small discrepancy arises when we consider the random forest model, whose variable importance plot appears to suggest that gender is not as important of a predictor compared to year, although this likely is due to the fact that we have not adjusted for inflation. The random forest model however does still have university as the most important predictor, which suggests that the university one attends is very important compared to the other predictors.

In general, across models, we also see that going to a "Big Four" university is correlated with higher wages as well. Unfortunately, we cannot see whether the sign of most coefficients for university are statistically significant, either due to multiple testing issues (sequential models), or due to the fact that we did not have any coefficient because it was nonparametric. Nevertheless, the consistent positive association suggests that this requires further research. Our current analyses show that indeed going to a more well-known college, like the "Big Four", is associated with higher earnings. Moreover, across the "Big Four" universities, the university coefficients in the linear regression models suggest that ADMU and UPD graduates are roughly equal with the highest earnings, followed by DLSU, followed by UST. This result makes sense as it checks out with the public perception of the colleges in the Philippines.

We also see that models have different results for our hypothesis on gender, albeit even though most do not have statistically significant results. While our linear models, including those from sequential selection suggest that being female is associated with lower first-time salaries, as mentioned before, our best model, the mixed effects model using industry as a cluster indicated that female employees earned higher first-time salaries than male employees based on fixed effect estimates. This suggests that indeed industry is an extremely important variable to control for, and due to the conflicting results more research, perhaps with better data as well, is necessary. We also see in some linear models that those who identify as "Other" in gender are also associated with lower earnings as well.

In general, we conclude that our results, even though most are not statistically significant, are in line with the hypothesis that a more well-known college is associated with higher wages, but our results with gender are mixed.

12 Discussion

In general, we had multiple problems with our data, much of it stemming from the fact that it is a survey that was popularized on Twitter, where we lose the power to control for our data collection methods. This caused us to have weak external validity as well as relatively weaker conclusions.

Note that much of this data cannot be obtained through the Philippine Census, where the main use is to obtain population counts for distributing Congressional seats, nor through the national identification system as it does not track income (Source: https://psa.gov.ph/philsys). Nevertheless, with more funding and time, it would still be better to conduct surveys on-site or through locals, where we can survey a more representative sample of the Philippine population to allow for not only better external validity but also internal validity (with internal validity problems mostly noted in the EDA section). This would also allow us to expand on our dataset as well, controlling for possible variables such as family background, a very common control variable for education as suggested in Card (page 1841). To control for unseen variables for education, it is possible to also consider twin studies and conduct a regression on their differences. For the wage differentials between gender, it would be interesting to get more data to control for variables such as whether it is a part-time or a full-time job.

Nevertheless, although the data is subject to a lot of data defects, we are able to get results that are interesting as they expand the questions of education and gender affecting wages to a more underexplored country, the Philippines.

References

Becker, Gary S., and Barry R. Chiswick. “Education and the Distribution of Earnings.” The American Economic Review, vol. 56, no. 1/2, 1966, pp. 358–369. JSTOR, www.jstor.org/stable/1821299. Accessed 10 Dec. 2020.

Brewer, Dominic J., et al. “Does It Pay to Attend an Elite Private College? Cross-Cohort Evidence on the Effects of College Type on Earnings.” The Journal of Human Resources, vol. 34, no. 1, 1999, pp. 104–123. JSTOR, www.jstor.org/stable/146304. Accessed 10 Dec. 2020.

Card, David, 1999. “The causal effect of education on earnings,” Handbook of Labor Economics, in: O. Ashenfelter & D. Card (ed.),Handbook of Labor Economics, edition 1, volume 3, chapter 30, pages 1801-1863, Elsevier.

“Gender Pay Gap Still Exists in the Philippines—Study.” Investing in Women, 15 Oct. 2019, investinginwomen.asia/posts/video-gender-pay-gap-philippines/.

Sanchez, Martha Jean. “Share of Respondents in the Philippines Who Thought That Women Would Be Paid the Same as Men for the Same Work as of January 2020.” Statista, 28 Jan. 2020, www.statista.com/statistics/1091112/philippines-views-on-gender-pay-gap/.

Appendix

Data cleaning can be found in the file `0-liyab_cleaning.Rmd`.

3 Exploratory Data Analysis

3.1 Salary

Histograms of the untransformed and transformed version of the monthly salary:

```r
ggplot(liyab, aes(x = salary)) +
  geom_histogram(bins = 50) +
  labs(title = "Histogram of Monthly Salary", x = "Monthly salary, in PHP", y = "Count")

liyab <- liyab %>%
  mutate(log_salary = log(salary, 10))

ggplot(liyab, aes(x = log_salary)) +
  geom_histogram(bins = 50) +
  labs(title = "Histogram of Monthly Salary", x = "log_10(Monthly salary), in PHP", y = "Count")
```

3.2 Gender

Data cleaning code below; we used “Male”, “Female”, and “Other” as the categories for gender as there were a variety of responses to the gender identification question on the survey and we wanted to simplify the data.

```r
# factoring gender
liyab_gender <- liyab$is_male * 1 + liyab$is_female * 2
liyab_gender <- as.factor(liyab_gender)
liyab_gender <- factor(liyab_gender, levels = c(1:2,0))
liyab$genderfactor <- as.factor(liyab_gender)
levels(liyab$genderfactor) <- c("Male", "Female", "Other")
```

Table of proportions showing the distribution of responses across gender categories. It is interesting to note that there were almost double the proportion of female respondents to male respondents.

3.3 Year

Histogram of years; this plot shows that a high proportion of responses came after 2009, which is why we focused on years after then in this study.

```r
ggplot(data = liyab) +
  geom_bar(aes(x = year)) +
  labs(y = "Count", x = "Year", title = "Barplot of Years")
```

3.4 University

Table of proportions of respondents’ universities. Respondents who attended “Big Four” universities made up close to half of respondents (44.77%).

```r
cat(
  "Proportion ADMU:", mean(liyab$is_admu, na.rm = T),
  "\nProportion UPD:", mean(liyab$is_upd, na.rm = T),
  "\nProportion UST:", mean(liyab$is_ust, na.rm = T),
  "\nProportion DLSU:", mean(liyab$is_dlsu, na.rm = T),
  "\nProportion Big Four:", mean(liyab$is_big4, na.rm = T),
  "\nProportion NA:", mean(is.na(liyab$is_big4))
)
```

Data Cleaning in order to assign categories for responses to the university question on the survey:

```
liyab_colleges <- liyab$is_admu * 1 + liyab$is_upd * 2 + liyab$is_ust * 3 + liyab$is_dlsu * 4
liyab_colleges <- as.factor(liyab_colleges)
liyab_colleges <- factor(liyab_colleges, levels = c(1:4, 0))

liyab$univ <- as.factor(liyab_colleges)
levels(liyab$univ) <- c("ADMU", "UPD", "UST", "DLSU", "Other")
```

3.5 Industry

A table showing the number of responses whose jobs fell under the different industries we assigned to the dataset. Later on in the study, industries with less than 10 respondents were included as "OTHER", so that splitting the train and test data would not exclude certain industries.

```
sort(table(liyab$industry), decreasing = T)
```

```
##
##             IT        FINANCE COMMUNICATIONS          SALES         RETAIL
##            318            293            249            200            160
##      EDUCATION     GOVERNMENT          MEDIA     HEALTHCARE    ENGINEERING
##            155            151            146            109             77
##          LABOR  MANUFACTURING           FOOD     REALESTATE      RELATIONS
##             74             72             61             58             52
##      NONPROFIT         ENERGY   ARCHITECTURE        TRADING     CONSULTING
##             50             36             34             34             33
##          OTHER       BUSINESS        ACADEME         DESIGN       RESEARCH
##             33             32             28             25             25
##    RECRUITMENT    HOSPITALITY      INSURANCE          LEGAL TRANSPORTATION
##             23             22             19             17             17
## PHARMACEUTICAL  ENTERTAINMENT         TRAVEL           DATA            ART
##             16             11             11             10              9
##    ENVIRONMENT        SCIENCE    AGRICULTURE         BEAUTY        FASHION
##              7              7              6              6              6
## ANIMALINDUSTRY
##              2
```

4 Exploration of relationships between salary and predictors

4.1 log(Salary) vs. gender

Plots showing the distribution of log of salaries based on each gender:

```
ggplot(liyab, aes(x = gender, y = log_salary)) +
  geom_boxplot() +
  labs(title = "Boxplots of Log-Salary On Gender", x = "Gender", y = "Log-Salary")
```

ANOVA test showing whether the gender of an employee has a statistically significant impact on the mean log salary:

```
summary(aov(log_salary ~ gender, data = liyab))
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## gender          2   1.46  0.7289   18.09 1.58e-08 ***
## Residuals   2691 108.44  0.0403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assumption Checking for the ANOVA test (described in paper):

```
gender_var <- sapply(
  levels(liyab$gender),
  function(x) {
    liyab %>%
      filter(gender == x) %>%
      .$log_salary %>%
      var()
  }
)
```

```
gender_var[order(gender_var)]
```

```
##     Female      Other       Male
## 0.03418030 0.04095598 0.06313715
```

-

4.2 log(Salary) vs. school

Plot of log salary per university:

```
ggplot(data=liyab, aes(univ, log_salary)) +
  geom_boxplot() +
  labs(
    y = "log(Salary)",
    x = "College",
    title = "log(Salary) vs University") +
  theme(
    axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

ANOVA test showing whether the university of an employee has a statistically significant impact on the mean log salary:

```
summary(aov(log_salary ~ univ, data = liyab))
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## univ             4  10.46  2.6147   72.26 <2e-16 ***
## Residuals     2539  91.87  0.0362
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 150 observations deleted due to missingness
```

Assumptions on the previous ANOVA test:

```
univ_var <- sapply(
  levels(liyab$univ),
  function(x) {
    liyab %>%
      filter(univ == x) %>%
      .$log_salary %>%
      var()
  }
)
univ_var <- c(
  univ_var,
  var(filter(liyab, is.na(liyab$univ))$log_salary))
```

```
univ_var[order(univ_var)]
```

```
##      Other        UST        UPD       DLSU       ADMU
## 0.03452915 0.03475549 0.03761487 0.03909116 0.04170972 0.04413446
```

4.3 log(Salary) vs. year

More Data Cleaning to filter out years with almost no responses:

```
liyab$yearfactor <- as.factor(liyab$year)

# filter out the later years because of lack of data
# 2005 is the year where we first have over 10 observations
#table(liyab$year) > 10

liyab$yearfiltered <- liyab$yearfactor
levels(liyab$yearfiltered)[levels(liyab$yearfiltered) %in% c(1987, 1992, 1998:2005)] <- NA
```

Plot showing the log salary for each year:

```
ggplot(data=liyab, aes(yearfiltered, log_salary)) +
  geom_boxplot() +
  labs(
    y = "log(Salary)",
    x = "Year",
    title = "log(Salary) distribution vs Year (2006-)") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

ANOVA test showing whether the year when a salary was secured has an impact on the log salary.

```
summary(aov(log_salary ~ yearfiltered, data = liyab))
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## yearfiltered   14   4.04 0.28825   7.509 1.09e-15 ***
## Residuals    2633 101.07 0.03839                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 46 observations deleted due to missingness
```

4.3.1 ANOVA assumptions

```
year_var <- sapply(
  levels(liyab$yearfiltered),
  function(x) {
    liyab %>%
      filter(year == x) %>%
      .$log_salary %>%
      var()
  }
)
year_var[order(year_var)]
```

```
##       2011       2009       2006       2008       2010       2019       2014 
## 0.02338348 0.02528083 0.02663444 0.02901554 0.03395151 0.03435500 0.03648352 
##       2012       2016       2018       2020       2013       2017       2015 
## 0.03740342 0.03742518 0.03891037 0.04019546 0.04240185 0.04332153 0.04504707 
##       2007 
## 0.08633574
```

4.3.2 Non-parametric test

Kruskal Test:

```r
kruskal.test(log_salary ~ yearfiltered, data = liyab)
```

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  log_salary by yearfiltered
## Kruskal-Wallis chi-squared = 123.73, df = 14, p-value < 2.2e-16
```

5. Baseline model

More Data Cleaning in order to change year and negotiate to numbers and factors respectively.

```r
liyab$year <- as.numeric(liyab$year)
liyab$negotiate <- as.factor(liyab$negotiate)
```

Summary Output of the baseline model:

```r
summary(baseline_lm <- lm(log_salary ~ year + univ + industry + gender + negotiate, data = liyab))
```

```
## 
## Call:
## lm(formula = log_salary ~ year + univ + industry + gender + negotiate, 
##     data = liyab)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.74991 -0.09625 -0.01112  0.08082  1.03749 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           -3.060e+01  2.078e+00 -14.725  < 2e-16 ***
## year                   1.736e-02  1.032e-03  16.829  < 2e-16 ***
## univUPD               -1.330e-02  1.327e-02  -1.002 0.316253    
## univUST               -1.093e-01  1.503e-02  -7.270 4.78e-13 ***
## univDLSU              -5.397e-02  1.667e-02  -3.238 0.001222 ** 
## univOther             -1.540e-01  1.131e-02 -13.621  < 2e-16 ***
## industryAGRICULTURE   -4.670e-02  7.691e-02  -0.607 0.543736    
## industryANIMALINDUSTRY -8.298e-02  1.249e-01  -0.664 0.506643    
## industryARCHITECTURE  -2.276e-01  4.462e-02  -5.100 3.66e-07 ***
## industryART           -3.030e-02  6.870e-02  -0.441 0.659238    
## industryBEAUTY         2.150e-02  7.697e-02   0.279 0.779988    
## industryBUSINESS      -3.098e-02  4.564e-02  -0.679 0.497356    
## industryCOMMUNICATIONS  5.460e-03  3.521e-02   0.155 0.876788    
## industryCONSULTING     2.902e-02  4.460e-02   0.651 0.515348    
## industryDATA           1.275e-01  6.565e-02   1.943 0.052137 .  
## industryDESIGN        -9.762e-02  4.818e-02  -2.026 0.042866 *  
## industryEDUCATION     -8.671e-02  3.629e-02  -2.389 0.016948 *  
## industryENERGY         1.623e-01  4.395e-02   3.694 0.000226 ***
## industryENGINEERING    4.425e-04  3.924e-02   0.011 0.991002    
## industryENTERTAINMENT -1.010e-01  6.118e-02  -1.652 0.098727 .  
## industryENVIRONMENT   -5.586e-02  7.230e-02  -0.773 0.439824    
## industryFASHION       -7.937e-02  7.686e-02  -1.033 0.301835    
## industryFINANCE        2.976e-02  3.495e-02   0.851 0.394602    
## industryFOOD           -6.122e-02  4.037e-02  -1.517 0.129473    
## industryGOVERNMENT      3.101e-02  3.621e-02   0.856 0.391873    
## industryHEALTHCARE     -1.635e-02  3.776e-02  -0.433 0.665021    
## industryHOSPITALITY    -9.658e-02  5.126e-02  -1.884 0.059651 .  
## industryINSURANCE       3.212e-02  5.122e-02   0.627 0.530714    
## industryIT              9.163e-02  3.484e-02   2.630 0.008594 ** 
## industryLABOR          -1.145e-01  3.926e-02  -2.915 0.003585 ** 
## industryLEGAL           1.923e-01  5.408e-02   3.557 0.000383 ***
## industryMANUFACTURING  -2.315e-03  3.904e-02  -0.059 0.952718    
## industryMEDIA          -7.643e-02  3.630e-02  -2.105 0.035359 *  
## industryNONPROFIT       3.705e-03  4.145e-02   0.089 0.928771    
## industryOTHER           1.479e-02  4.465e-02   0.331 0.740528    
## industryPHARMACEUTICAL  4.596e-02  5.394e-02   0.852 0.394346    
## industryREALESTATE     -2.711e-02  4.106e-02  -0.660 0.509141    
## industryRECRUITMENT    -1.051e-02  4.990e-02  -0.211 0.833158    
## industryRELATIONS      -3.420e-02  4.156e-02  -0.823 0.410615    
## industryRESEARCH        2.707e-02  4.760e-02   0.569 0.569593    
## industryRETAIL          7.791e-02  3.613e-02   2.156 0.031157 *  
## industrySALES          -3.247e-02  3.555e-02  -0.913 0.361205    
## industrySCIENCE         5.405e-03  7.691e-02   0.070 0.943984    
## industryTRADING         4.880e-02  4.445e-02   1.098 0.272451    
## industryTRANSPORTATION  1.913e-02  5.316e-02   0.360 0.718938    
## industryTRAVEL         -6.273e-02  6.121e-02  -1.025 0.305531    
## genderFemale           -4.138e-02  7.402e-03  -5.590 2.52e-08 ***
## genderOther            -1.977e-02  2.020e-02  -0.979 0.327867    
## negotiateYes           -3.880e-03  9.870e-03  -0.393 0.694239    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1694 on 2485 degrees of freedom
##   (160 observations deleted due to missingness)
## Multiple R-squared:  0.3019, Adjusted R-squared:  0.2884 
## F-statistic: 22.39 on 48 and 2485 DF,  p-value: < 2.2e-16
```

Assumption checking for the baseline model:

```
par(mfrow = c(2, 2))
plot(baseline_lm, which = c(1:3))
```

Outlier identified in the data:

```
#outlier
liyab[2168, 1:7]
```

```
##      year  industry                              role salary  univ gender negotiate
## 2169 2020 NONPROFIT PARTNERSHIPSASSOCIATETRAINEE  10000 Other Female        No
```

5.3 Removing Outliers and Cutting off by Year

Cook's distance plot:

```
plot(baseline_lm, which = c(4))
```

Removing the Outlier:

```
# cook's distance calc
liyab <- liyab[-2168,]
```

Table of years:

```
# justification for cutting off by year
table(liyab$year)
```

```
## 
## 1987 1992 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011
##    1    1    2    3    2    2    5    6    6   18   18   17   20   30   40   61
## 2012 2013 2014 2015 2016 2017 2018 2019 2020
##   83  113  135  180  280  427  543  622   79
```

Removing years under the cutoff:

```
cutoff = 2009
liyabf <- liyab[liyab$year >= cutoff,]
liyabf$yearfactor <- factor(liyabf$year)
```

6. Non-parametric tests

Splitting the data:

```
set.seed(139)

# store full dataset in liyab_full
liyab_full <- liyabf
nrow(liyab_full)

# get rid of industries with less than 10 observations
for(i in 1:nrow(liyab_full)){
  liyab_industries = table(liyab_full$industry)
  if(liyab_industries[names(liyab_industries) == liyab_full$industry[i]] < 10){
    liyab_full$industry[i] = "OTHER"
  }
}

# create train and test sets
test.id = sample(seq(1,nrow(liyab_full), 1), 540)
liyab_test = liyab_full[test.id,]
liyab_train = liyab_full[-test.id,]
```

7 Sequential Variable Selection; New Baseline; Interaction; Mixed Effects

Note that for section 7, a lot of the commented code was still originally run, but for the sake of not spamming the appendix with R output some of the stuff is commented out for sanity's sake.

More data cleaning:

```
liyab_new <- liyab_train

# create a copy without univ, industry, role, gender
# univ, industry, gender oh cos already have indicators
# role, too many to use. will instead use when we "narrow down"
# also remove salary since we will use log

liyab_new$role <- NULL
liyab_new$salary <- NULL
liyab_new$genderfactor <- NULL
liyab_new$yearfactor <- NULL
liyab_new$yearfiltered <- NULL

# drop the indicators
liyab_new[, 6:52] <- NULL
```

```r
# transform year by subtracting 2009 so it will be 0 to 11
liyab_new$year <- liyab_new$year - min(liyab_new$year)

# change negotiate to 1 if yes, 0 if no
require(dplyr)
liyab_new <- liyab_new %>%
      mutate(negotiate = ifelse(negotiate == "No",0,1))

# making necessary changes to test set
liyab_test <- liyab_test %>%
      mutate(negotiate = ifelse(negotiate == "No",0,1))
liyab_test$year <- liyab_test$year - min(liyab_test$year)
liyab_test <- liyab_test[complete.cases(liyab_test), ]
```

7.1 Fitting Linear Models: All base terms and interaction terms

R output for summary and for the Chi-Squared test, as well as code for everything else.

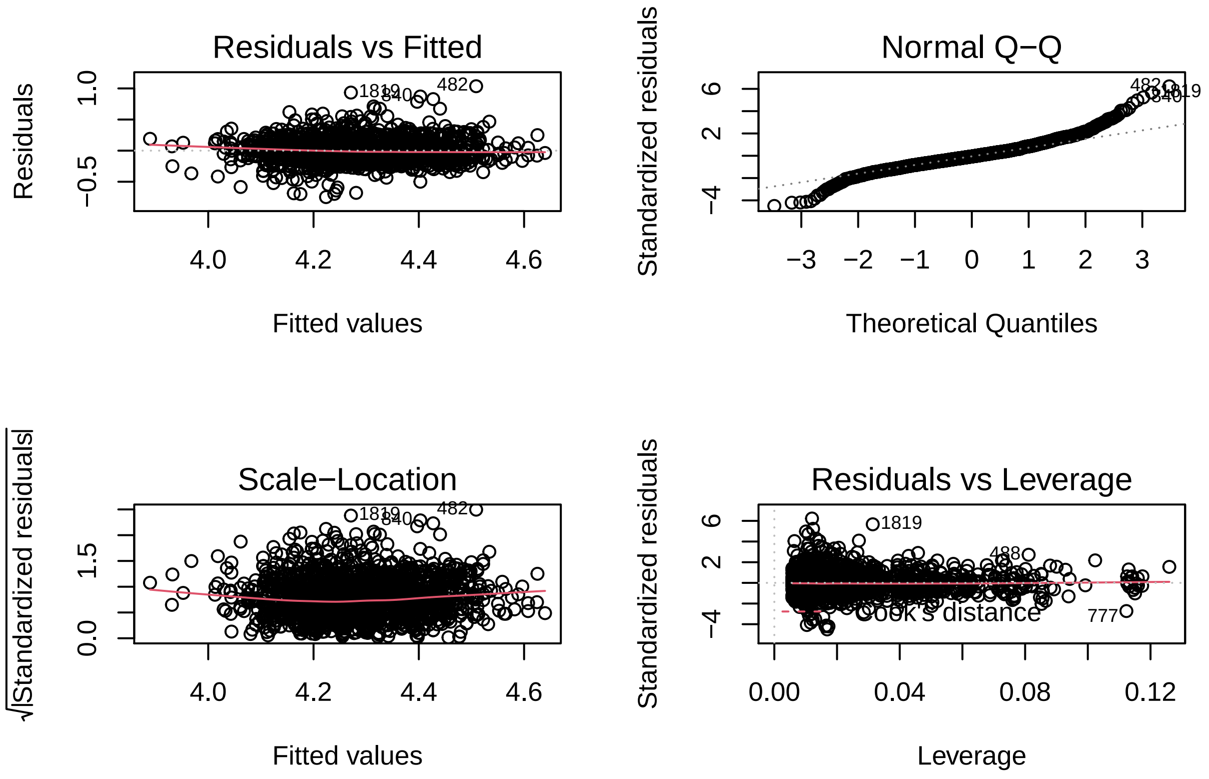
```
## 
##  Pearson's Chi-squared test
## 
## data:  table(liyab_nona$univ, liyab_nona$gender)
## X-squared = 10.451, df = 8, p-value = 0.2348
```

```
## 
## Call:
## lm(formula = log_salary ~ poly(year, degree = 4, raw = TRUE) + 
##     (. - year), data = liyab_nona)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.74660 -0.09510 -0.01069  0.07872  1.03516 
## 
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                          4.312e+00  5.295e-02  81.436  < 2e-16 ***
## poly(year, degree = 4, raw = TRUE)1 -4.448e-02  3.272e-02  -1.359 0.174255    
## poly(year, degree = 4, raw = TRUE)2  1.753e-02  1.102e-02   1.591 0.111678    
## poly(year, degree = 4, raw = TRUE)3 -1.941e-03  1.406e-03  -1.381 0.167490    
## poly(year, degree = 4, raw = TRUE)4  7.394e-05  6.023e-05   1.228 0.219765    
## industryARCHITECTURE                -2.154e-01  5.196e-02  -4.145 3.54e-05 ***
## industryBUSINESS                    -1.702e-02  5.480e-02  -0.311 0.756178    
## industryCOMMUNICATIONS               1.352e-02  4.284e-02   0.316 0.752401    
## industryCONSULTING                   2.339e-02  5.312e-02   0.440 0.659670    
## industryDESIGN                      -5.876e-02  5.685e-02  -1.034 0.301481    
## industryEDUCATION                   -6.261e-02  4.394e-02  -1.425 0.154349    
## industryENERGY                       2.116e-01  5.191e-02   4.076 4.76e-05 ***
## industryENGINEERING                  7.142e-03  4.622e-02   0.155 0.877223    
## industryENTERTAINMENT               -7.329e-02  6.916e-02  -1.060 0.289443    
## industryFINANCE                      3.999e-02  4.228e-02   0.946 0.344271    
## industryFOOD                        -7.956e-02  4.830e-02  -1.647 0.099682 .  
## industryGOVERNMENT                   6.405e-02  4.355e-02   1.471 0.141590    
## industryHEALTHCARE                  -5.196e-03  4.549e-02  -0.114 0.909085    
## industryHOSPITALITY                 -5.701e-02  6.049e-02  -0.942 0.346128    
## industryINSURANCE                    3.719e-02  6.046e-02   0.615 0.538578    
## industryIT                           1.093e-01  4.231e-02   2.584 0.009838 ** 
## industryLABOR                       -9.208e-02  4.683e-02  -1.966 0.049406 *  
## industryLEGAL                        2.381e-01  6.336e-02   3.757 0.000177 ***
## industryMANUFACTURING               -7.247e-04  4.653e-02  -0.016 0.987575    
## industryMEDIA                       -5.629e-02  4.378e-02  -1.286 0.198673    
## industryNONPROFIT                    2.766e-02  4.901e-02   0.564 0.572567    
## industryOTHER                        1.948e-02  4.507e-02   0.432 0.665686    
## industryPHARMACEUTICAL               7.337e-02  6.182e-02   1.187 0.235461    
## industryREALESTATE                  -8.920e-03  4.866e-02  -0.183 0.854569    
## industryRECRUITMENT                  8.770e-03  5.544e-02   0.158 0.874317    
## industryRELATIONS                   -1.342e-02  4.968e-02  -0.270 0.787106    
## industryRESEARCH                     3.335e-02  5.469e-02   0.610 0.542036    
## industryRETAIL                       9.672e-02  4.357e-02   2.220 0.026547 *  
## industrySALES                       -1.490e-02  4.290e-02  -0.347 0.728318    
## industryTRADING                      6.556e-02  5.291e-02   1.239 0.215477    
## industryTRANSPORTATION               4.602e-02  5.957e-02   0.773 0.439901    
## industryTRAVEL                      -4.880e-02  6.918e-02  -0.705 0.480619    
## univUPD                             -5.935e-03  1.490e-02  -0.398 0.690461    
## univUST                             -9.990e-02  1.671e-02  -5.980 2.66e-09 ***
## univDLSU                            -4.148e-02  1.906e-02  -2.177 0.029620 *  
## univOther                           -1.426e-01  1.265e-02 -11.276  < 2e-16 ***
## genderFemale                        -4.253e-02  8.292e-03  -5.129 3.20e-07 ***
## genderOther                         -1.796e-02  2.274e-02  -0.790 0.429887    
## negotiate                            3.537e-03  1.102e-02   0.321 0.748238    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1673 on 1913 degrees of freedom
## Multiple R-squared:  0.2858, Adjusted R-squared:  0.2697 
## F-statistic:  17.8 on 43 and 1913 DF,  p-value: < 2.2e-16
```

```
## 
## Call:
## lm(formula = log_salary ~ ., data = liyab_nona)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.74537 -0.09616 -0.01117  0.07964  1.03522 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             4.254011   0.044069  96.530  < 2e-16 ***
## year                    0.016492   0.001569  10.508  < 2e-16 ***
## industryARCHITECTURE   -0.215836   0.051914  -4.158 3.36e-05 ***
## industryBUSINESS       -0.015203   0.054739  -0.278 0.781244    
## industryCOMMUNICATIONS  0.014138   0.042799   0.330 0.741178    
## industryCONSULTING      0.023977   0.053102   0.452 0.651660    
## industryDESIGN         -0.059522   0.056857  -1.047 0.295294    
## industryEDUCATION      -0.062669   0.043938  -1.426 0.153946    
## industryENERGY          0.211047   0.051911   4.066 4.99e-05 ***
## industryENGINEERING     0.007319   0.046162   0.159 0.874046    
## industryENTERTAINMENT  -0.072657   0.069180  -1.050 0.293733    
## industryFINANCE         0.040602   0.042275   0.960 0.336961    
## industryFOOD           -0.078841   0.048309  -1.632 0.102840    
## industryGOVERNMENT      0.063797   0.043550   1.465 0.143109    
## industryHEALTHCARE     -0.004507   0.045496  -0.099 0.921102    
## industryHOSPITALITY    -0.057072   0.060501  -0.943 0.345639    
## industryINSURANCE       0.038059   0.060463   0.629 0.529115    
## industryIT              0.109103   0.042302   2.579 0.009979 ** 
## industryLABOR          -0.091836   0.046834  -1.961 0.050037 .  
## industryLEGAL           0.236197   0.063263   3.734 0.000194 ***
## industryMANUFACTURING  -0.003047   0.046499  -0.066 0.947767    
## industryMEDIA          -0.054288   0.043741  -1.241 0.214710    
## industryNONPROFIT       0.026872   0.048958   0.549 0.583150    
## industryOTHER           0.019900   0.045031   0.442 0.658599    
## industryPHARMACEUTICAL  0.071954   0.061761   1.165 0.244148    
## industryREALESTATE     -0.008977   0.048647  -0.185 0.853609    
## industryRECRUITMENT     0.008226   0.055381   0.149 0.881932    
## industryRELATIONS      -0.013626   0.049651  -0.274 0.783786    
## industryRESEARCH        0.033756   0.054696   0.617 0.537200    
## industryRETAIL          0.097282   0.043543   2.234 0.025588 *  
## industrySALES          -0.014259   0.042875  -0.333 0.739495    
## industryTRADING         0.065654   0.052891   1.241 0.214643    
## industryTRANSPORTATION  0.046609   0.059532   0.783 0.433768    
## industryTRAVEL         -0.048930   0.069193  -0.707 0.479562    
## univUPD                -0.005836   0.014893  -0.392 0.695213    
## univUST                -0.100290   0.016689  -6.009 2.22e-09 ***
## univDLSU                  -0.040775   0.019050  -2.140 0.032445 *
## univOther                 -0.142455   0.012646 -11.265  < 2e-16 ***
## genderFemale              -0.042882   0.008282  -5.178 2.48e-07 ***
## genderOther               -0.019104   0.022728  -0.841 0.400708
## negotiate                  0.003149   0.011018   0.286 0.775065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1673 on 1916 degrees of freedom
## Multiple R-squared:  0.2842, Adjusted R-squared:  0.2693
## F-statistic: 19.02 on 40 and 1916 DF,  p-value: < 2.2e-16
```

Assumption-checking graphs for the polynomial regression. These graphs look similar to those of the linear regression with all predictors included.

```
par(mfrow = c(2, 2))
plot(lm_poly)
```



7.2 Sequential Selection

Fitting, analyzing, and saving data for model comparison:

```
# now interactions between variables
lm_0 <- lm(log_salary~1, data = liyab_nona)
lm_interact <- lm(log_salary~.^2, data = liyab_nona)

require(glmnet)
lm_seq <- step(lm_0, scope = list(lower = lm_0, upper = lm_interact),
               direction = "both", trace = 0)

summary(lm_seq)
```

```
##
## Call:
## lm(formula = log_salary ~ industry + univ + year + gender + year:gender,
##     data = liyab_nona)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.74310 -0.09669 -0.00997  0.08000  1.03112 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             4.224444   0.046275  91.290  < 2e-16 ***
## industryARCHITECTURE   -0.215506   0.051856  -4.156 3.38e-05 ***
## industryBUSINESS       -0.020949   0.054724  -0.383 0.701897    
## industryCOMMUNICATIONS  0.011806   0.042752   0.276 0.782462    
## industryCONSULTING      0.021271   0.053039   0.401 0.688431    
## industryDESIGN         -0.056964   0.056796  -1.003 0.316007    
## industryEDUCATION      -0.065276   0.043927  -1.486 0.137440    
## industryENERGY          0.207608   0.051860   4.003 6.49e-05 ***
## industryENGINEERING     0.004995   0.046085   0.108 0.913699    
## industryENTERTAINMENT  -0.079758   0.069161  -1.153 0.248967    
## industryFINANCE         0.038594   0.042228   0.914 0.360859    
## industryFOOD           -0.080335   0.048215  -1.666 0.095839 .  
## industryGOVERNMENT      0.061542   0.043519   1.414 0.157481    
## industryHEALTHCARE     -0.006764   0.045455  -0.149 0.881722    
## industryHOSPITALITY    -0.057966   0.060416  -0.959 0.337453    
## industryINSURANCE       0.036462   0.060383   0.604 0.546024    
## industryIT              0.107523   0.042278   2.543 0.011061 *  
## industryLABOR          -0.095252   0.046779  -2.036 0.041864 *  
## industryLEGAL           0.237790   0.063220   3.761 0.000174 ***
## industryMANUFACTURING  -0.005404   0.046464  -0.116 0.907426    
## industryMEDIA          -0.055399   0.043670  -1.269 0.204747    
## industryNONPROFIT       0.026958   0.048879   0.552 0.581341    
## industryOTHER           0.018664   0.044972   0.415 0.678169    
## industryPHARMACEUTICAL  0.072682   0.061697   1.178 0.238925    
## industryREALESTATE     -0.010092   0.048520  -0.208 0.835255    
## industryRECRUITMENT     0.008795   0.055295   0.159 0.873645    
## industryRELATIONS      -0.017592   0.049597  -0.355 0.722848    
## industryRESEARCH        0.030204   0.054641   0.553 0.580485    
## industryRETAIL          0.094195   0.043500   2.165 0.030480 *  
## industrySALES          -0.016915   0.042811  -0.395 0.692799    
## industryTRADING         0.064664   0.052716   1.227 0.220112    
## industryTRANSPORTATION  0.038978   0.059535   0.655 0.512739    
## industryTRAVEL         -0.049621   0.069029  -0.719 0.472323    
## univUPD                -0.004723   0.014880  -0.317 0.750977    
## univUST                -0.100321   0.016673  -6.017 2.12e-09 ***
## univDLSU               -0.040267   0.019037  -2.115 0.034541 *  
## univOther              -0.141893   0.012640 -11.225  < 2e-16 ***
## year                    0.020634   0.002539   8.127 7.84e-16 ***
## genderFemale            0.002228   0.026011   0.086 0.931759    
## genderOther             0.127197   0.074898   1.698 0.089618 .  
## year:genderFemale      -0.005913   0.003229  -1.831 0.067267 .  
## year:genderOther       -0.019131   0.009343  -2.048 0.040726 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1671 on 1915 degrees of freedom
## Multiple R-squared:  0.2865, Adjusted R-squared:  0.2713 
## F-statistic: 18.76 on 41 and 1915 DF,  p-value: < 2.2e-16
```

```
cat("Sequential model formula:")
```

```
## Sequential model formula:
```

```
formula(lm_seq)
```

```
## log_salary ~ industry + univ + year + gender + year:gender
```

```
temp <- BIC(lm_poly, lm_all, lm_interact, lm_seq)
#temp

#AIC(lm_interact, lm_seq)

#output is too large to include in appendix
#summary(lm_interact)
```

```r
#RMSE(liyab_test$log_salary, predict(lm_seq, new=liyab_test))
#RMSE(liyab_test$log_salary, predict(lm_interact, new=liyab_test))

# this allows for model comparison later
lm_seq_df <- data.frame(name = "lm_seq", rsquared = 0.2866, rmse = 0.180255, aic = -1402.411, bic = -1162.9485)
```

7.3 Mixed Effects Modelling

R output for the summary of the mixed effects models:

```r
# Try out mixed-effects model
# Industry as the cluster variable
# Year as cluster variable
lmer_industry <- lmer(log_salary ~ gender + univ + year + year:gender +
                  (1|industry),
                data = liyab_nona)
lmer_year <- lmer(log_salary~ gender + industry + univ + negotiate +
                 (1 + gender|year), data = liyab_nona)
summary(lmer_industry)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: log_salary ~ gender + univ + year + year:gender + (1 | industry)
##    Data: liyab_nona
##
## REML criterion at convergence: -1309.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.4574 -0.5719 -0.0656  0.4651  6.1739
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  industry (Intercept) 0.005535 0.0744
##  Residual             0.027934 0.1671
## Number of obs: 1957, groups:  industry, 33
##
## Fixed effects:
##                       Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)          4.234e+00  2.715e-02  3.706e+02 155.923  < 2e-16 ***
## genderFemale         1.081e-03  2.600e-02  1.919e+03   0.042   0.9668
## genderOther          1.262e-01  7.480e-02  1.924e+03   1.687   0.0918 .
## univUPD             -3.743e-03  1.483e-02  1.936e+03  -0.252   0.8007
## univUST             -1.024e-01  1.664e-02  1.929e+03  -6.157 9.01e-10 ***
## univDLSU            -3.952e-02  1.901e-02  1.925e+03  -2.079   0.0378 *
## univOther           -1.424e-01  1.261e-02  1.930e+03 -11.292  < 2e-16 ***
## year                 2.051e-02  2.536e-03  1.923e+03   8.086 1.08e-15 ***
## genderFemale:year   -5.882e-03  3.227e-03  1.920e+03  -1.823   0.0685 .
## genderOther:year    -1.902e-02  9.328e-03  1.926e+03  -2.039   0.0416 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) gndrFm gndrOt unvUPD unvUST unDLSU unvOth year   gndrF:
## genderFemal -0.602
## genderOther -0.204  0.215
## univUPD     -0.360  0.021  0.017
## univUST     -0.300  0.014 -0.016  0.525
## univDLSU    -0.272  0.033 -0.019  0.458  0.416
## univOther   -0.410  0.042 -0.017  0.700  0.629  0.548
## year        -0.733  0.741  0.259  0.042  0.017  0.032  0.043
## gendrFml:yr  0.568 -0.949 -0.201 -0.018 -0.024 -0.032 -0.047 -0.780
## gndrOthr:yr  0.196 -0.202 -0.953 -0.028  0.017  0.011  0.008 -0.273  0.210
```

```r
summary(lmer_year)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: log_salary ~ gender + industry + univ + negotiate + (1 + gender |
##     year)
##    Data: liyab_nona
##
## REML criterion at convergence: -1233.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3902 -0.5728 -0.0599  0.4664  6.1118
##
## Random effects:
##  Groups   Name         Variance  Std.Dev. Corr
##  year     (Intercept)  0.0046594 0.06826
##           genderFemale 0.0004324 0.02079  -1.00
##           genderOther  0.0035572 0.05964  -1.00  1.00
##  Residual              0.0279523 0.16719
## Number of obs: 1957, groups:  year, 12
##
## Fixed effects:
##                          Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)             4.349e+00  4.707e-02  2.169e+02  92.396  < 2e-16 ***
## genderFemale           -3.134e-02  1.030e-02  1.166e+01  -3.041 0.010553 *
## genderOther             1.303e-02  2.879e-02  1.106e+01   0.452 0.659702
## industryARCHITECTURE   -2.186e-01  5.201e-02  1.908e+03  -4.203 2.75e-05 ***
## industryBUSINESS       -2.257e-02  5.480e-02  1.907e+03  -0.412 0.680501
## industryCOMMUNICATIONS  8.177e-03  4.283e-02  1.907e+03   0.191 0.848618
## industryCONSULTING      2.045e-02  5.315e-02  1.907e+03   0.385 0.700472
## industryDESIGN         -5.902e-02  5.686e-02  1.907e+03  -1.038 0.299441
## industryEDUCATION      -6.766e-02  4.396e-02  1.907e+03  -1.539 0.124003
## industryENERGY          2.061e-01  5.189e-02  1.906e+03   3.973 7.37e-05 ***
## industryENGINEERING     3.508e-03  4.623e-02  1.908e+03   0.076 0.939520
## industryENTERTAINMENT  -8.179e-02  6.918e-02  1.906e+03  -1.182 0.237250
## industryFINANCE         3.477e-02  4.230e-02  1.907e+03   0.822 0.411235
## industryFOOD           -8.362e-02  4.833e-02  1.906e+03  -1.730 0.083741 .
## industryGOVERNMENT      6.066e-02  4.356e-02  1.907e+03   1.392 0.163934
## industryHEALTHCARE     -1.006e-02  4.554e-02  1.907e+03  -0.221 0.825091
## industryHOSPITALITY    -6.237e-02  6.059e-02  1.907e+03  -1.029 0.303385
## industryINSURANCE       3.468e-02  6.046e-02  1.906e+03   0.574 0.566265
## industryIT              1.038e-01  4.234e-02  1.907e+03   2.452 0.014293 *
## industryLABOR          -9.719e-02  4.683e-02  1.906e+03  -2.075 0.038078 *
## industryLEGAL           2.375e-01  6.339e-02  1.908e+03   3.746 0.000185 ***
## industryMANUFACTURING  -7.476e-03  4.660e-02  1.908e+03  -0.160 0.872558
## industryMEDIA          -6.057e-02  4.380e-02  1.908e+03  -1.383 0.166791
## industryNONPROFIT       2.501e-02  4.904e-02  1.907e+03   0.510 0.610051
## industryOTHER           1.473e-02  4.510e-02  1.907e+03   0.327 0.744025
## industryPHARMACEUTICAL  6.946e-02  6.185e-02  1.908e+03   1.123 0.261581
## industryREALESTATE     -1.361e-02  4.868e-02  1.906e+03  -0.280 0.779890
## industryRECRUITMENT     5.149e-03  5.545e-02  1.907e+03   0.093 0.926023
## industryRELATIONS      -2.072e-02  4.971e-02  1.907e+03  -0.417 0.676942
## industryRESEARCH        2.795e-02  5.472e-02  1.906e+03   0.511 0.609575
## industryRETAIL          9.100e-02  4.361e-02  1.907e+03   2.087 0.037055 *
## industrySALES          -2.100e-02  4.290e-02  1.907e+03  -0.490 0.624523
## industryTRADING         5.970e-02  5.299e-02  1.907e+03   1.127 0.260086
## industryTRANSPORTATION  3.651e-02  5.962e-02  1.906e+03   0.612 0.540340
## industryTRAVEL         -5.089e-02  6.921e-02  1.906e+03  -0.735 0.462300
## univUPD                -5.782e-03  1.492e-02  1.908e+03  -0.387 0.698454
## univUST                -1.001e-01  1.674e-02  1.906e+03  -5.981 2.64e-09 ***
## univDLSU               -4.267e-02  1.912e-02  1.908e+03  -2.232 0.025745 *
## univOther              -1.424e-01  1.265e-02  1.906e+03 -11.255  < 2e-16 ***
## negotiate               4.424e-03  1.101e-02  1.908e+03   0.402 0.687945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
#r2(lmer_industry)
#r2(lmer_year)
```

```
#AIC(lmer_industry, lmer_year)
#BIC(lmer_industry, lmer_year)
#RMSE(liyab_test$log_salary, predict(lmer_industry, newdata =liyab_test))
#RMSE(liyab_test$log_salary, predict(lmer_year, newdata =liyab_test))

summary(liyab_nona)
```

```
##       year                 industry      univ         gender
##  Min.   : 0.000   FINANCE       :220   ADMU : 222   Male  : 667
##  1st Qu.: 6.000   IT            :218   UPD  : 337   Female:1230
##  Median : 8.000   COMMUNICATIONS:167   UST  : 197   Other :   60
##  Mean   : 7.668   SALES         :156   DLSU : 123
##  3rd Qu.:10.000   RETAIL        :116   Other:1078
##  Max.   :11.000   GOVERNMENT    :114
##                   (Other)       :966
##    negotiate        log_salary
##  Min.   :0.0000   Min.   :3.477
##  1st Qu.:0.0000   1st Qu.:4.176
##  Median :0.0000   Median :4.255
##  Mean   :0.1451   Mean   :4.280
##  3rd Qu.:0.0000   3rd Qu.:4.398
##  Max.   :1.0000   Max.   :5.544
##
```

```
lmer_industry_df <- data.frame(name = "lmer_industry", rsquared = 0.295, rmse = 0.1795364, aic = -1281.677, bic = -1214.8502)

# plot(log_salary~year, data = liyab_nona)
# abline(summary(lmer_industry)$coef[1:2,1], col = 1, lty = 1, lwd = 3)
# abline(summary(lmer_year)$coef[1:2,1], col = 1, lty = 1, lwd = 3)
```

7.31 Assumption Checking

7.311 Industry Model

Plots to check assumptions for industry mixed effets model.

```
# for industry
par(mfrow = c(2, 2))
# checking linearity and constant variance
plot(residuals(lmer_industry)~predict(lmer_industry))

# checking normality of residuals
qqnorm(residuals(lmer_industry))
qqline(residuals(lmer_industry))
qqnorm(coef(lmer_industry)$industry[['(Intercept)']])
qqline(coef(lmer_industry)$industry[['(Intercept)']])
```

Checking the residual plot, we again see that linearity and constant variance are relatively satisfied. Unfortunately, Normality of residuals is still an issue, but at least the random intercepts do seem Normally distributed as well.

7.312 Year Model

Plots:

```
# for year
par(mfrow = c(2, 2))
# checking linearity and constant variance
plot(residuals(lmer_year)~predict(lmer_year))

# checking normality of residuals
```

```r
qqnorm(residuals(lmer_year))
qqline(residuals(lmer_year))
qqnorm(coef(lmer_year)$year[['(Intercept)']])
qqline(coef(lmer_year)$year[['(Intercept)']])
```

8 More Flexible Models

8.1 Ridge + Lasso Models

Ridge + LASSO model outputs:

```r
set.seed(139)
XInteract = model.matrix(lm_interact)[,-1]

# Ridge model
modelRidge = cv.glmnet(XInteract, liyab_nona$log_salary, alpha = 0, lambda=10^seq(-4,4,0.1))
modelRidge$lambda.min
```

```
## [1] 0.07943282
```

```r
modelRidge.min = glmnet(XInteract, liyab_nona$log_salary, alpha = 0, lambda=modelRidge$lambda.min)

set.seed(139)
XInteract = model.matrix(lm_interact)[,-1]
# LASSO model
modelLasso = cv.glmnet(XInteract, liyab_nona$log_salary, alpha = 1, lambda=10^seq(-4,4,0.1))
modelLasso$lambda.min
```

```
## [1] 0.003162278
```

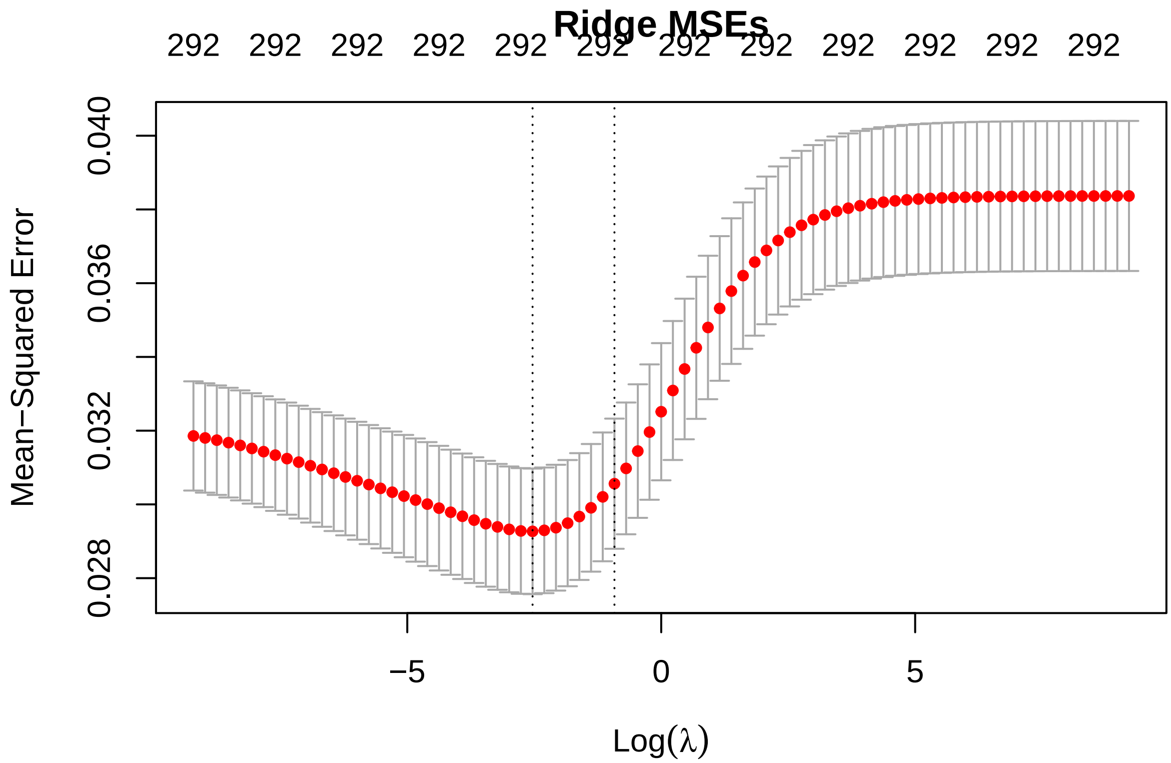
```r
modelLasso.min = glmnet(XInteract, liyab_nona$log_salary, alpha = 1, lambda=modelLasso$lambda.min)

modelLasso$lambda.min
```

```
## [1] 0.003162278
```

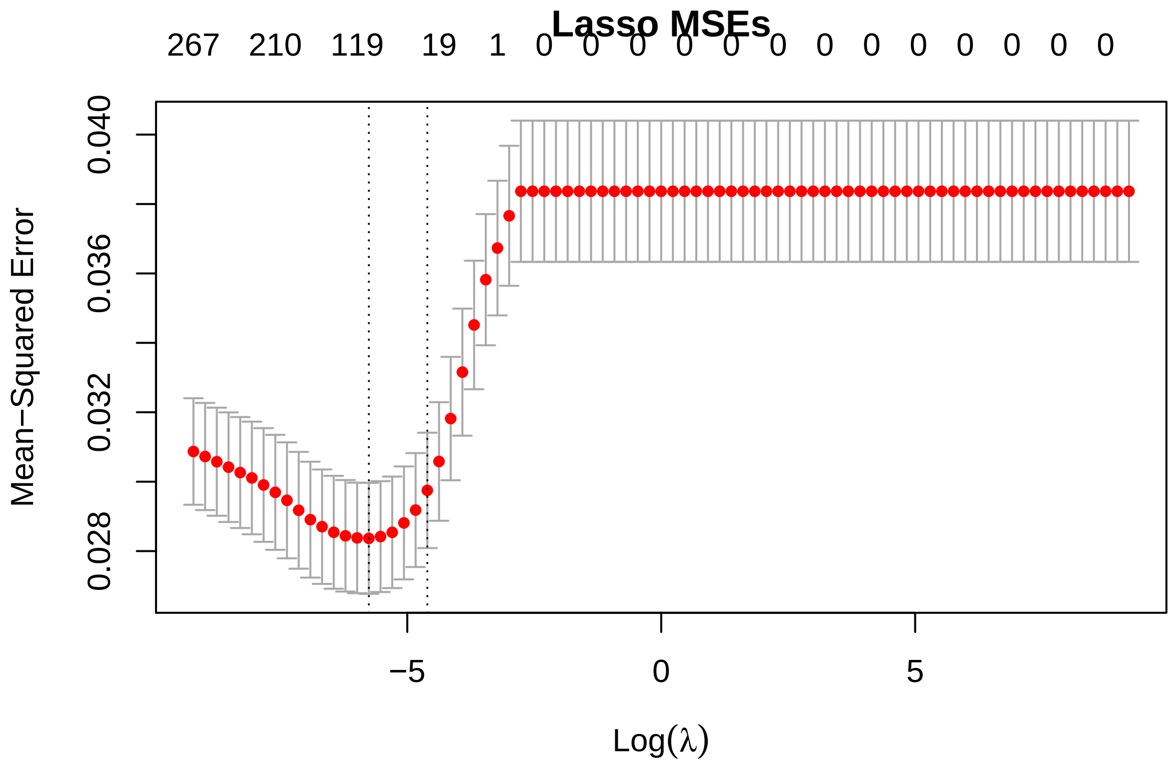
```r
data.frame(ridge=min(modelRidge$cvm),lasso=min(modelLasso$cvm))
```

```
##        ridge      lasso
## 1 0.02926904 0.02836649
```

```r
plot(modelRidge, main = "Ridge MSEs")
```



```r
plot(modelLasso, main = "Lasso MSEs")
```


Lasso MSEs
Mean−Squared Error
Log(λ)