Wages And Gender

Efran Himel, Madelyn Nelson, Tiffany Quang, Isabelle Hauge

Background

- Topic: Assess the impact of sex on wages after accounting for other variables and to produce a regression model for predicting wages.

- Data from the 1985 Current Population Survey

- Examines the impact of sex on wage while taking into consideration possible confounding variables.

Our Data

- 245 out of the 534 observations were females
- The average wage for the observed females was $7.88 per hour

- 289 out of the 534 observations were males
- The average wage for the observed males was $9.99 per hour

- This is a difference of more than $2 per hour!

Variables

We couldn't say that the difference in wages between males and females was strictly based on their **sex**, so the data includes other variables that may affect wages

- **Education** - number of years of education
- **South** - qualitative variable indicating whether or not the person lives in the South (1 = person lives in the South, 0 = person lives elsewhere)
- **Experience** - number of years of work experience
- **Union** - whether or not the person is a union member (1 = union member, 0 = not a union member)
- **Age** - in years
- **Race** - variable indicating person's race (1 - other, 2 - Hispanic, 3 - White)
- **Occupation** - occupational category (1 - management, 2 - Sales, 3 - Clerical, 4 - Service, 5 - Professional, 6 - Other)
- **Sector** - sector of work (0 - other, 1 - manufacturing, 2 - construction)
- **Marr** - marital status (0 - unmarried, 1 - married)

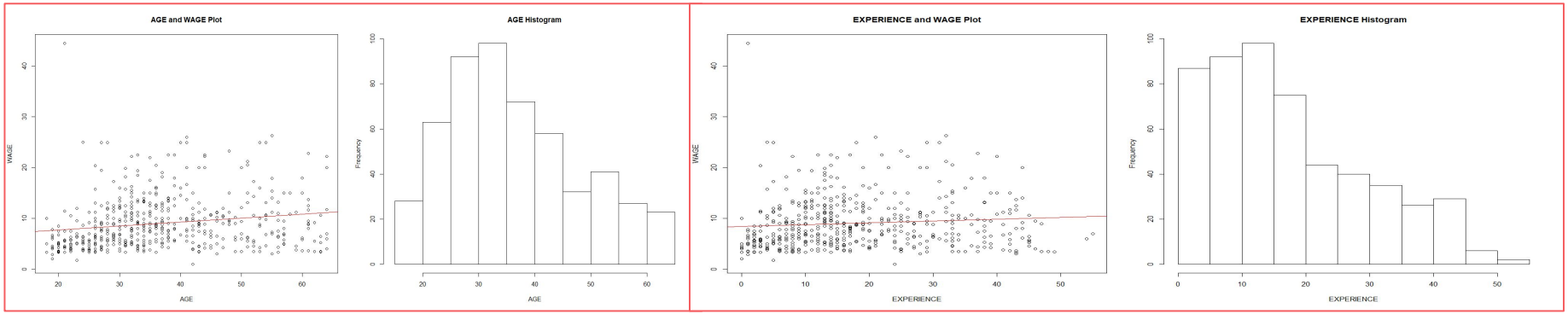
Data Overview

Here's a snapshot of the table of data we worked with.

WAGE	SEX	EDUCATION	SOUTH	EXPERIENCE	UNION	AGE	RACE	OCCUPATION	SECTOR	MARR
5.1	1	8	0	21	0	35	2	6	1	1
4.95	1	9	0	42	0	57	3	6	1	1
6.67	0	12	0	1	0	19	3	6	1	0
4	0	12	0	4	0	22	3	6	0	0
7.5	0	12	0	17	0	35	3	6	0	1
13.07	0	13	0	9	1	28	3	6	0	0
4.45	0	10	1	27	0	43	3	6	0	0
19.47	0	12	0	9	0	27	3	6	0	0
13.28	0	16	0	11	0	33	3	6	1	1

Exploring Data

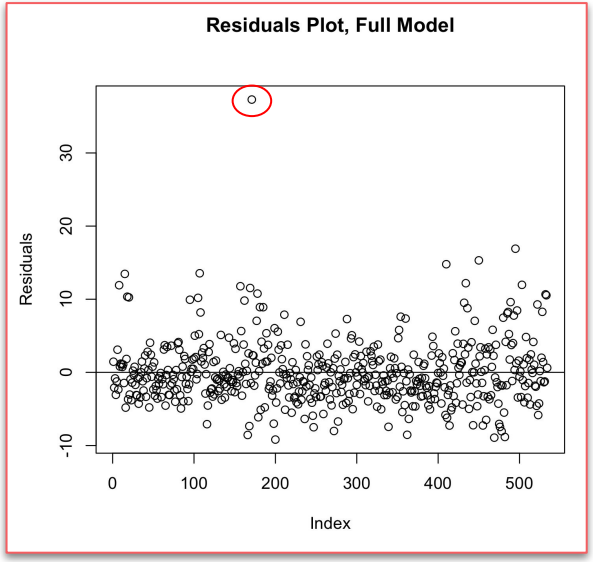
- Plotted all X variables against our Y variable (wage) to gain a general idea of the distributions of the X's against Y.
- Regression and histograms were also utilized for initial analysis of outliers and linear relationships with Y.
- Afterwards we hypothesized the X values South, Sector and Marriage may be the least impactful variables.



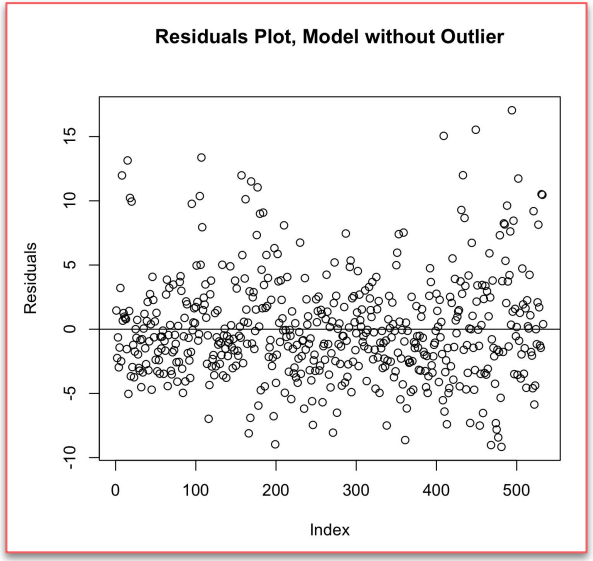
Developing Models

- Produced residual plots for collinearity and Q-Q plots for examining the normality of quantitative data (Age, Experience, Education, Wage)
- We also used BIC and AIC algorithms to hone in on the variable combinations with the highest predictive capacity and cross validation to test robustness
- Findings:
  - Some of our variables did not seem to be normal
  - Sex was present in all the highest scoring BIC AIC models which indicated to us it may have a strong importance in understanding wage
  - Other prominent variables in all our models highest scoring models: Education, Sex, Union, Age/Experience
  - All BIC AIC models performed poorly in cross validation which could indicate there are more key variables at play in determining wage or our data is too small

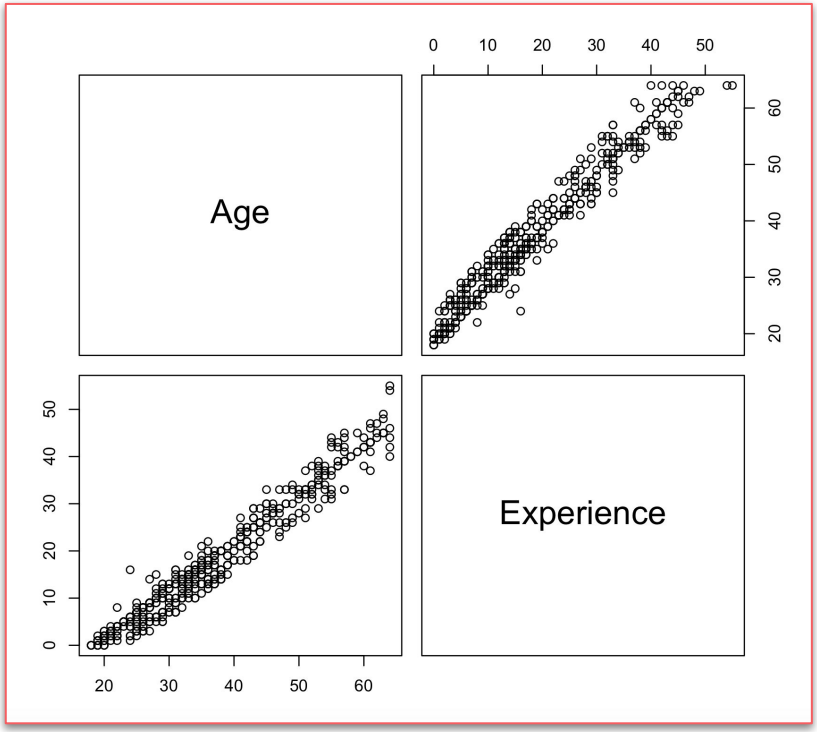
Removing the Outlier



During our preliminary data exploration, we discovered that there was one outlier that was significantly different from the rest of the data, so we decided that we should consider removing that data point for the purpose of our study.



Collinearity



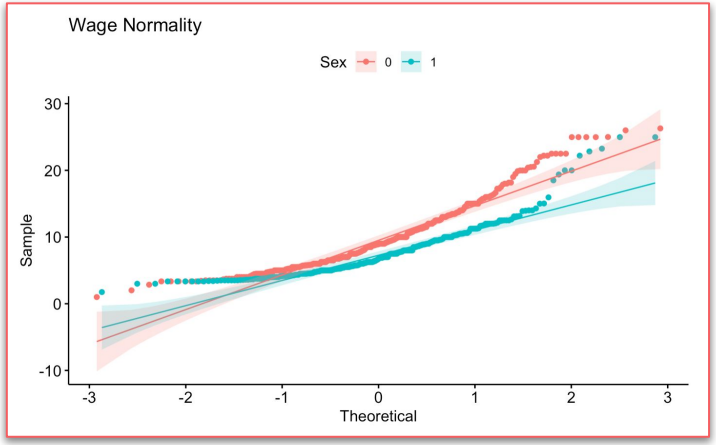
Using the pairs() function in R, we found that Age and Experience had a high correlation of 0.977885.

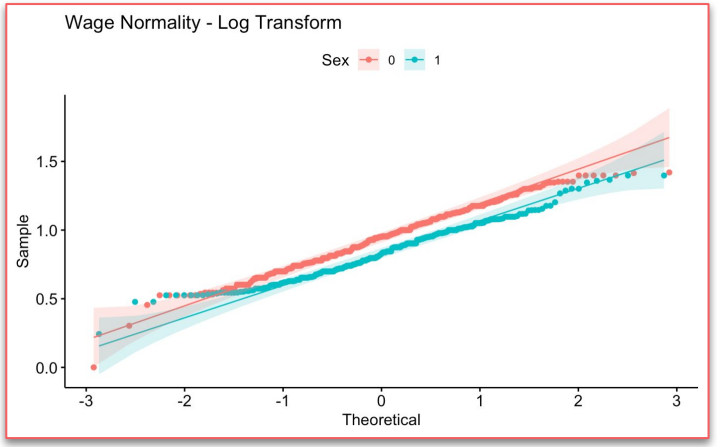
Due to this fact, we decided that our final model would only include one of these variables.

We had to do more testing to decide which one of these variables would remain in the model.

Log Transformation of Wage

When we looked at the Q-Q plots of the quantitative variables, we found that our wage variable showed a right skew on the normality plot.





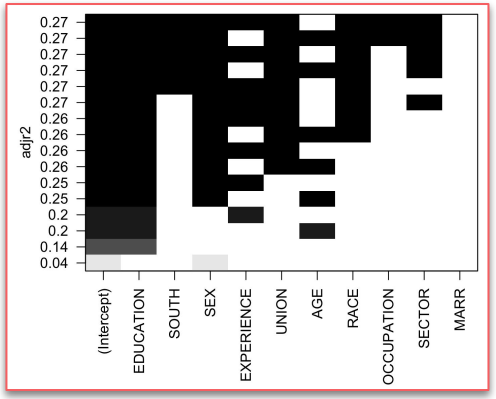
We tried a log transformation of this variable, and the new Q-Q plot showed much less of a skew, so we decided that this was something to consider when testing.

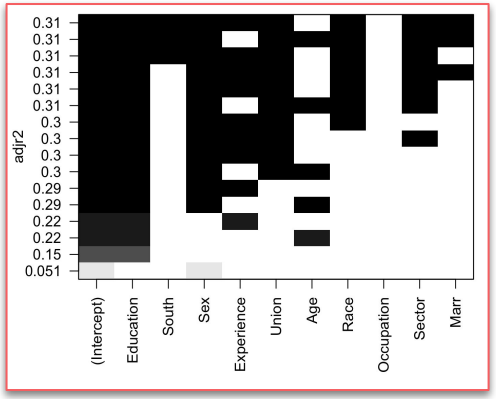
Adjusted R-Squared Testing

Outlier

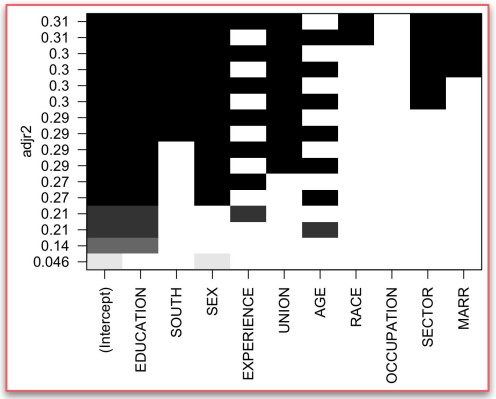
No Outlier

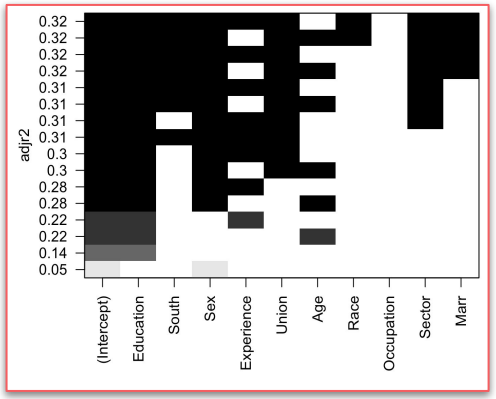
No Log Transform



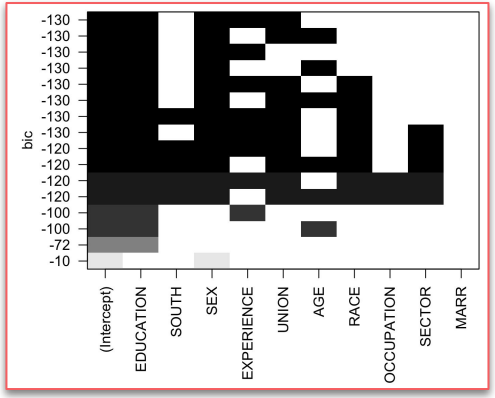
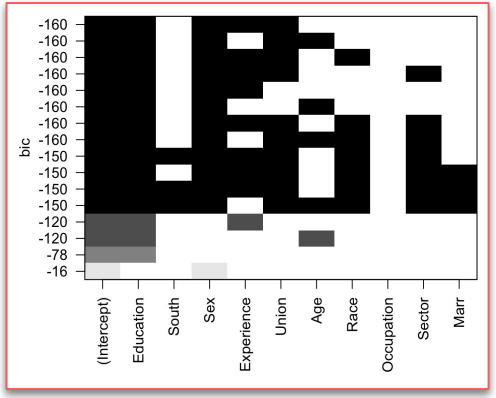
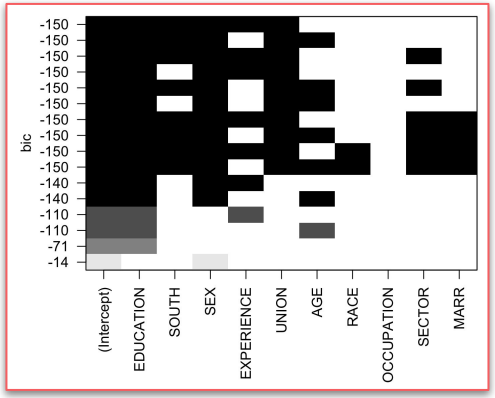
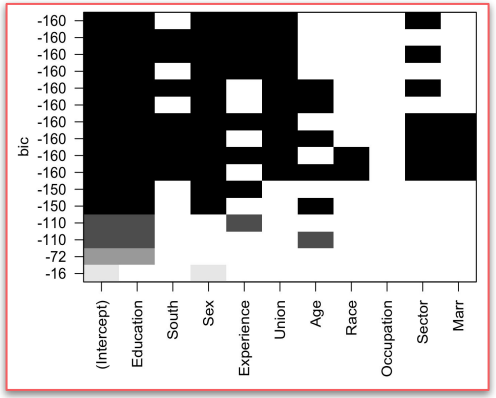


Log Transform





BIC Testing

	Outlier	No Outlier
No Log Transform		
Log Transform		

Choosing Model Based on VIF

- We want to find which multicollinear variable (experience or age) to include in our regression model.
    - log(Wage) ~ **Age** + Education + Sex + Union

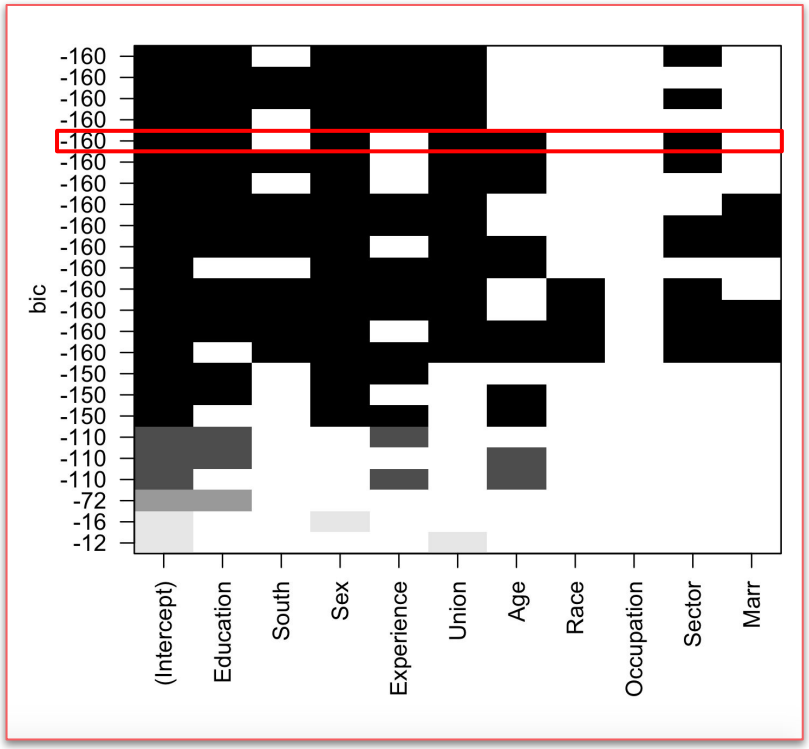
```
      AGE EDUCATION       SEX     UNION
 1.048054  1.023019  1.036125  1.043629
```

    - log(Wage) ~ Education + Sex + Union + **Experience**

```
EDUCATION       SEX     UNION EXPERIENCE
 1.143552  1.036290  1.043612   1.170085
```

- The model including **age** has VIFs values closer to 1, which is what we want.

Improving Model Based on BIC



- The BIC values suggested that the variable Sector could be added to our model after we performed a log transformation and removed an outliers.

- We were not sure if we should include an additional variable as the model could become overfit.

ANOVA Test

- We performed an ANOVA test to determine whether or not to include Sector to our original model:
  - Null Hypothesis: The model without Sector is a better performing model.
  - alpha = 0.05

```
Analysis of Variance Table

Model 1: log(WAGE) ~ AGE + EDUCATION + SEX + UNION
Model 2: log(WAGE) ~ AGE + EDUCATION + SEX + UNION + SECTOR
  Res.Df     RSS Df Sum of Sq     F  Pr(>F)
1    528 101.094
2    527  99.882  1    1.2115 6.392 0.01175 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- With a p-value = 0.01175, we reject the null hypothesis and are 95% confident that adding Sector to our model significantly improves it.

Analysis of Summary Output

```
Call:
lm(formula = log(WAGE) ~ AGE + EDUCATION + SEX + UNION + SECTOR,
    data = new_cps85wages)

Residuals:
     Min       1Q   Median       3Q      Max
-2.06638 -0.30082  0.01688  0.31312  1.19755

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.500965   0.124580   4.021 6.64e-05 ***
AGE          0.011950   0.001652   7.235 1.65e-12 ***
EDUCATION    0.088626   0.007415  11.952  < 2e-16 ***
SEX1        -0.221107   0.039068  -5.660 2.50e-08 ***
UNION        0.203689   0.050220   4.056 5.75e-05 ***
SECTOR       0.091908   0.036353   2.528   0.0118 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4353 on 527 degrees of freedom
Multiple R-squared:  0.3132,    Adjusted R-squared:  0.3067
F-statistic: 48.06 on 5 and 527 DF,  p-value: < 2.2e-16
```

Regression Model:

$$\log(\hat{Y}) = 0.500965 + 0.011950(\textbf{age}) + 0.088626(\textbf{education}) - 0.221107(\textbf{sex}) + 0.203689(\textbf{union}) + 0.091908(\textbf{sector})$$

Based off of our study's population, we received a **negative** β value for our variable Sex. We can interpret this as sex possibly having a negative impact for women, as the qualitative placeholder values was 0 for male and 1 for female.

Conclusion

- In our data from 1985, we found sex is an important predictor variable in the model we chose to predict wage.
- In our model, sex has a negative coefficient which indicates that being female could have a negative impact on wage.
- Using log transformations of our data allowed us to produce our models by making our data follow a normal distribution
- Learning Outcomes:
    - Deeper appreciation for slowly analyzing data and building a the best fit model by understanding features of the data set
    - Algorithmic methods are useful to honing in models of potential interest
    - Although our findings are interesting, we can't make any sweeping general statements given the scope and limitations of our data and tests