STAT230 Project Final Report: Are employers as blind to identifiers like age, disability, and sex as they claim to be? Personal factors affecting wages or salary income of individuals

Daniel Jang, Cassandra Jin, Dhyey Mavani

December 11, 2022

Abstract

Our project was an investigation of whether employers are as blind to identifiers like age, disability, and sex as they claim to be when considering personal factors affecting wages or salary income of individuals. More specifically, we examined the presence and nature of the relationships between the wages or salary income of an individual in the past 12 months and the factors that lie in the person's background and social status, including age, marriage status, location, race, sex, and more. After observing the correlations between personal wages and many potential explanatory variables, we proceeded to conduct analysis and build models for cube root wages versus six variables: hours worked per week, years married, income-to-poverty ratio, self-employment income, ability to speak English, and class of worker. The kitchen sink model, bestsubsets method, and nested F-test yielded a model predicting personal wages with hours worked per week, income-to-poverty ratio, self-employment income, and ability to speak English. This model mostly satisfied the conditions for multiple linear regression, but we additionally conducted bootstrapping to obtain bootstrap confidence intervals for the coefficients of the explanatory variables, all of which proved to be significant.

Background and Meaning

In most cases, the personal salary income of a person majorly dictates their living standards and ability to provide for others. Before one may enter a job and serve a community, they are scrutinized by hiring committees, and although everyone has supposedly equal opportunity, we show through this investigation that this is rarely the case. Instead of being aware of an applicant's background and social status, like their age, sex, and ability/disability to perform, we should be specially cognizant of not allowing such factors to influence their held position and corresponding income if they fulfill their role well enough otherwise. The findings of this project could be applied to policy considerations that aim to raise low household incomes and may provide insight on the impacts that social circumstances have on household income.

Methods

Data

The American Community Survey (ACS) is a randomly-selected census survey designed by the national government to determine how best to distribute state and federal funds across the United States. The ACS serves as a sample population of the United States, and we are utilizing the 2019 ACS's housing dataset within this analysis, specifically personal wages and salary income as our response variable. Our findings may then be applied to the US population in 2019.

Variables

Response Variable:

Personal wages or salary income in the past 12 months (WAGP): Univariate analysis of this quantitative variable shows extreme right skew, so we performed multiple transformations in order to best satisfy the necessary conditions ultimately settled on the cube root transformation. Thus, we renamed WAGP to Cuberoot_wages. The original variable is measured in dollars, and since the distribution is so skewed, we observe the median = \$30000 and IQR = \$37450 as measures of distribution of WAGP. After the cube root transformation, however, our new response variable is considerably symmetric with a mean of 30.409 cube root dollars and a standard deviation of 12.268 cube root dollars, which transform to mean = 28119.42 and SD = \$1846.38.

Explanatory Variables:

Usual hours worked per week past 12 months (WKHP): We renamed this quantitative variable to Work_hours_per_week, and since the distribution mostly demonstrates having a normal, symmetric distribution, we observe that the mean is 35.636 and standard deviation is 11.413 measured in hours.

Year last married (MARHYP): We mutated this variable by subtracting all of the values from 2019 in order to obtain the number of years married, calling the new variable Years_married. The original distribution of MARHYP has median = 1994 and IQR = 28, measured in years, while the median and IQR of Years_married are 18 and 18.5 years, respectively.

Income-to-poverty ratio (POVPIP): This quantitative variable is measured as percentage and demonstrate a very strong left skew, so we use median = 449, IQR = 239. We also renamed it to Income_to_poverty_ratio.

Self-employment income past 12 months (SEMP): The density distribution of this quantitative variable drops dramatically as the income increases, so we observe median = \$0, IQR = \$0. However, it ranges from \$0 all the way up to \$117000 with a mean of \$3182.8. We renamed the variable to Self_employment_income.

Ability to speak English (ENG): This categorical variable captures a person's English-speaking ability on a scale from none to 4. The count distribution is as follows: b/N/A (less than 5 years old/speaks only English) (88.52%), 1/very well (7.54%), 2/well (2.42%), 3/not well (1.20%), 4/not at all (0.32%). We renamed this variable to English_ability.

Class of worker (COW) (renamed to Class_of_worker): This categorical variable records the nature of a worker's occupation. It originally contains 10 categories: b/N/A (less than 16 years old/NILF who last worked more than 5 years ago or never worked); 1/Employee of a private for-profit company or business, or of an individual, for wages, salary, or commissions; 2/Employee of a private not-for-profit, tax-exempt, or charitable organization; 3/Local government employee (city, county, etc.); 4/State government employee; 5/Federal government employee; 6/self-employed in own not incorporated business, professional practice, or farm; 7/self-employed in own incorporated business, professional practice or farm; 8/Working without pay in family business or farm; 9/unemployed and last worked 5 years ago or earlier or never worked. For more significant models and findings, we merged the categories into 3 more comprehensive groups: private employee (72.7273%), public employee (18.1818%), and self-employed or unemployed (9.0909%), and we renamed the variable to Class_of_worker.

Statistical Methods

We first carried out univariate data analysis by observing the summary statistics (favstats - descriptive statistics for quantitative variables and a frequency table for qualitative variables). Upon analysis of our response variable, WAGP, we decided to apply a cube root transformation for the most satfisfactory data to work with. For each quantitative variable, we were interested in the data's shape, spread, and center, and by using ggplot, we determined which information about the center and spread to report (depending on symmetry of the distribution). For the qualitative variables, we outputted histograms capturing the frequencies of the categories and reported these frequencies. Before doing so, we made sure to compile certain categories so as to obtain significant groups for analysis. Our next step was to find the most significant predictors of WAGP from a large pool of potential variables. We used ggpairs() to check the correlations between WAGP and every predictor variable, and although correlation is not a comprehensive statistic of the strength of an interaction between two variables, it gave us a starting point to understanding the predictors relative to our response variable and helped to at least eliminate very weak ones.

After we narrowed down our predictor variables, we used simple linear regression to understand and analyse the relationship between WAGP and each predictor variable, creating scatterplots for the quantitative ones and boxplots for the qualitative and checking the necessary conditions (equal variance - fitted vs. residuals plot, normality - qq plot, linearity - scatterplot, independence, randomization, and errors centered around 0). We then entered model building with multiple variables, first by constructing a kitchen sink model and then using the bestsubsets method. After further conducting nested F-tests and looking at VIFs, we finally obtained a cohesive MLR model for predicting cube root WAGP that mostly satisfied the conditions for MLR.

Results

Univariate Analysis

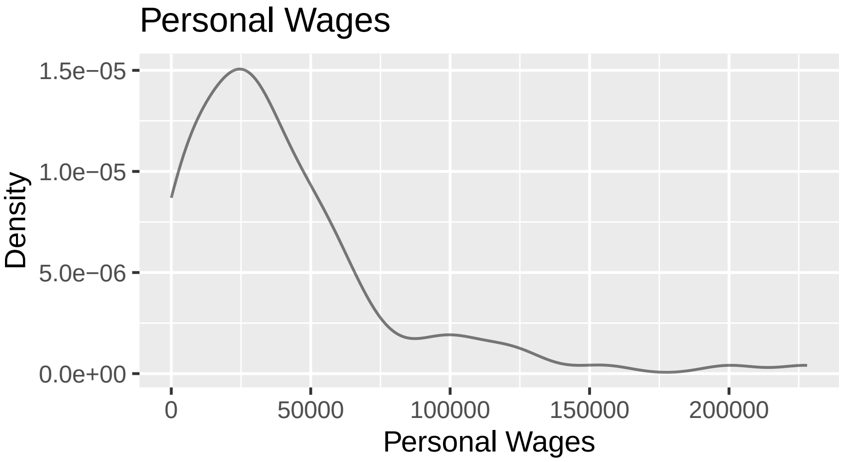

Figure 1: Personal wages or salary income in the past 12 months, measured in dollars; right-skewed distribution, median = $30000, IQR = $37450.

Bivariate Analysis

By observing the ggpairs plots between all of our quantitative predictor variables and our response variable, we decided to use quantitative variables income-to-poverty ratio, hours worked per week, years married, and self-employment income. The highest correlation values in the plots also suggested that we use the qualitative variables ability to speak English and class of worker.

To pick the best model, we first attempted simple linear regression models with each one of the four quantitative variables we chose, making sure to check against the conditions of SLR as well. Through this process, we applied combinations of re-expressions (on the explanatory or response variable) to create the best possible model that also fit the conditions relatively well. Next, we produced MLR models to see if additional predictors would be better in predicting cube root personal wages. When we found an MLR model with a considerable amount of variability accounted for, we used multiple methods compare different predictors, keeping in mind parsimony through bestsubsets selection to ensure that we had the simplest and most effective model, and we finally checked for multicollinearity.

The four single-predictor models that we investigated were:

mod1 <- lm(I(Cuberoot_wages) ~ POVPIP, data = acs_vars_pums)

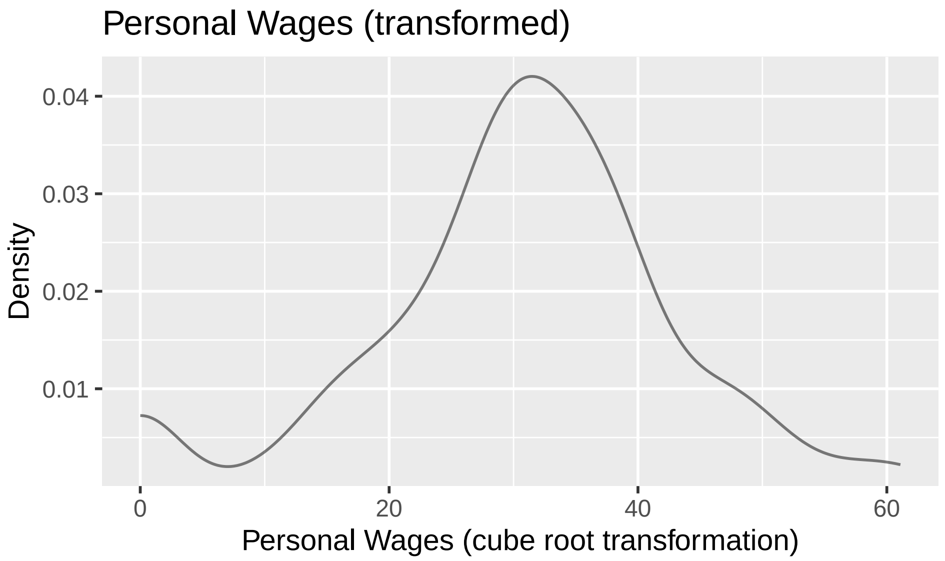

Figure 2: Personal wages with cube root transformation applied (renamed to Cuberoot_wages), quantitative response variable measured in cube root dollars; mean = 30.409, sd = 12.268.

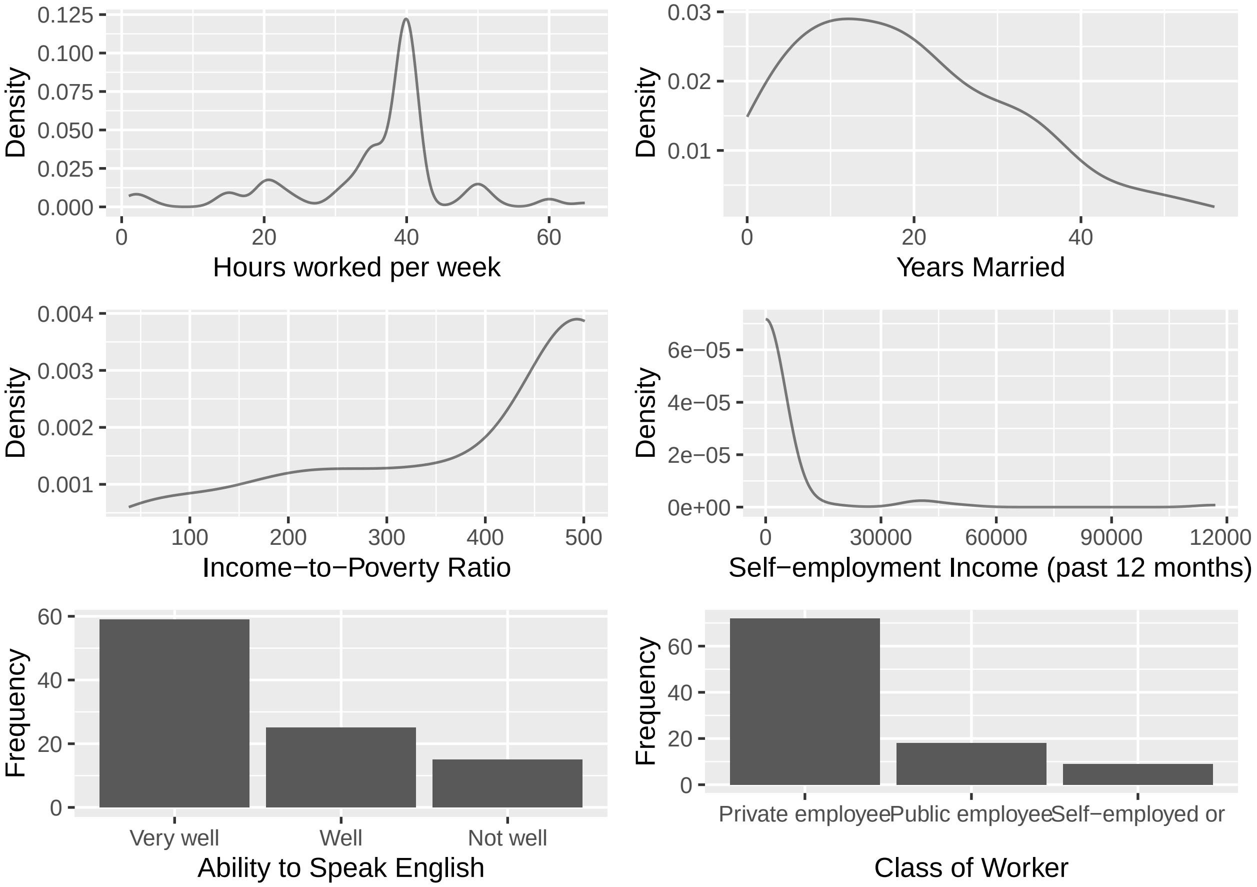

Figure 3: EDAs of potential explanatory variables WKHP, MARHYP, POVPIP, SEMP, ENG, COW. All distribution statistics are listed in Variables section above.

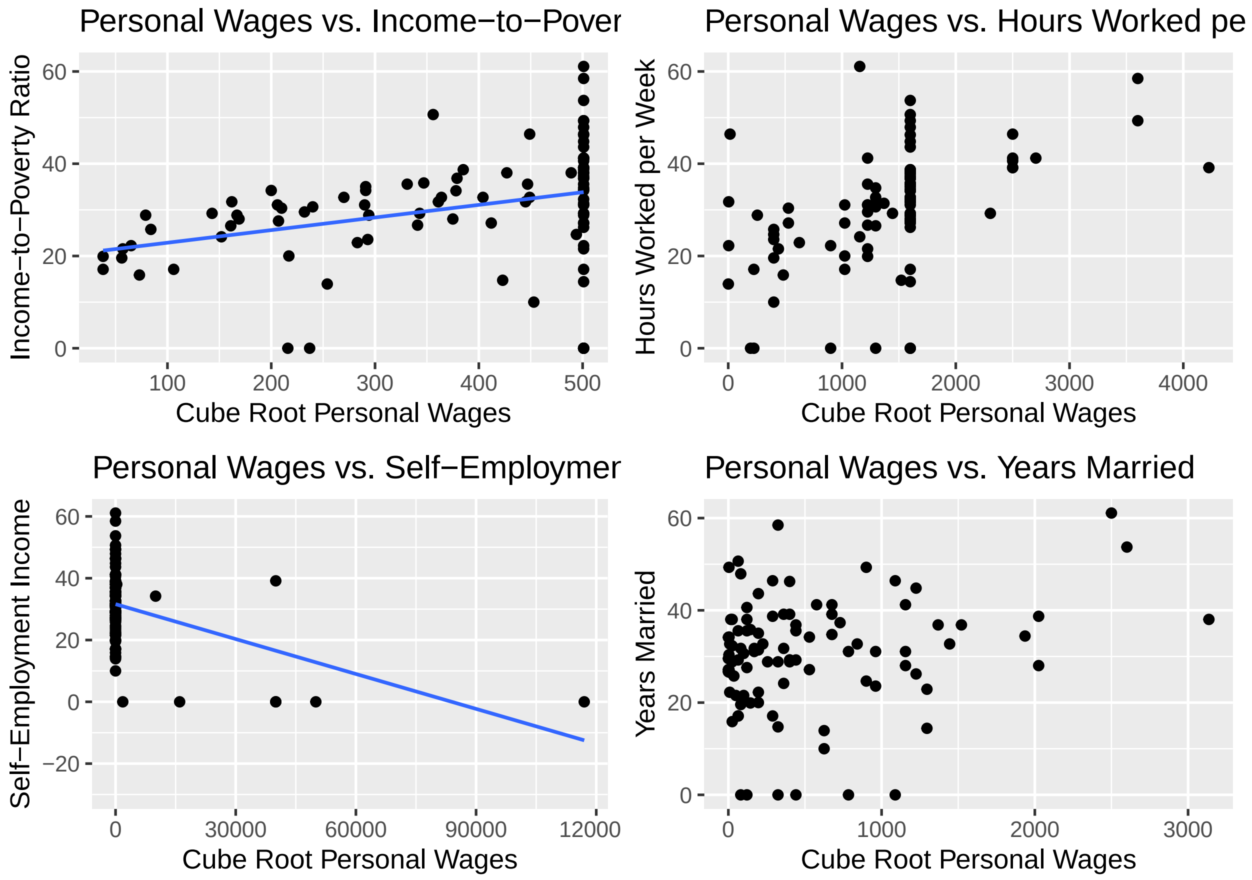

Figure 4: The bivariate distributions of cube root personal wages and each of our potential predictor variables in the form of scatterplots.

mod2 <- lm(I(Cuberoot_wages) ~ I(WKHPˆ2), data = acs_vars_pums)

mod3 <- lm(I(Cuberoot_wages) ~ SEMP, data = acs_vars_pums)

mod4 <- lm(I(Cuberoot_wages) ~ I(Years_marriedˆ2), data = acs_vars_pums)

Based on the scatterplots, residuals vs. fitted plots, and qq plots of these SLR models, none of these models did well in predicting cube root personal wages, despite our addition of variable re-expressions. mod2 demonstrated a high R^2 value compared to the others, but the conditions for this SLR model did not hold at all. Conversely, mod4 met the conditions well but had the lowest R^2 value, so none of these four models ended up being useful.

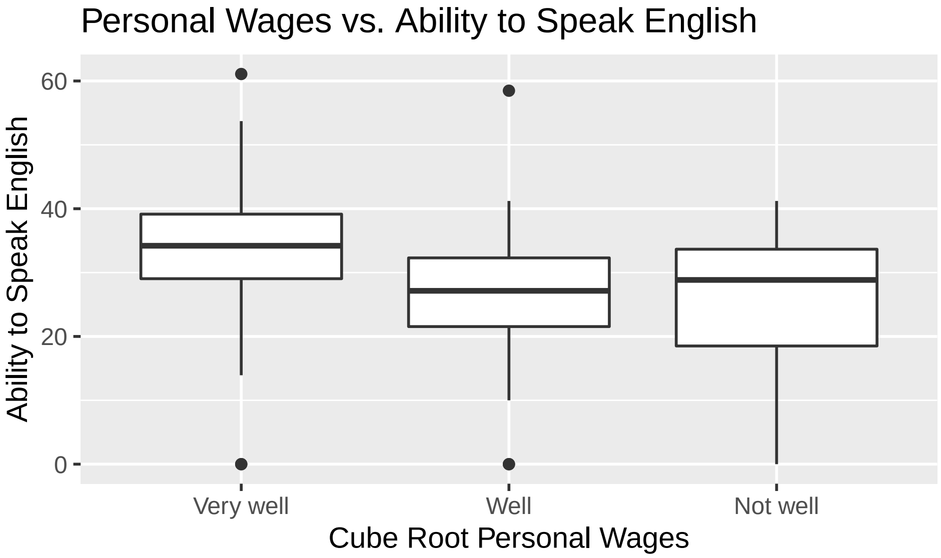

Figure 5: Besides the 'Not well' group having greater spread, there did not appear to be significant difference between cube root personal wages of the 'Well' and 'Not well' categories. However, the distribution of the cube root personal wages of the 'Very well' group was slightly more removed from the other two.

Multiple Linear Regression

```
##              rsq   adjr2      cp     rss POVPIP I.WKHP.2. SEMP
## 1  ( 1 ) 0.24521 0.23743 53.3873 11132.5                *     
## 2  ( 1 ) 0.43826 0.42655 17.4359  8285.3                *    *
## 3  ( 1 ) 0.50815 0.49262  5.6943  7254.3      *         *    *
## 4  ( 1 ) 0.52186 0.50151  5.0000  7052.2      *         *    *
##          I.Years_married.2.
## 1  ( 1 )                   
## 2  ( 1 )                   
## 3  ( 1 )                   
## 4  ( 1 )                  *
```

Table A: Since our simple linear regression models did not capture significant predictability of our response variable, we next used the bestsubsets method to identify the best model with 1, 2, 3, and 4 quantitative predictors. Based on this output and our own judgement, the best model we could produce with our quantitative predictors was with income-to-poverty ratio, hours worked per week, and self-employment income.

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.36e+01   2.59e+00    5.24  9.6e-07 ***
## POVPIP       2.32e-02   6.31e-03    3.67  0.00039 ***
```

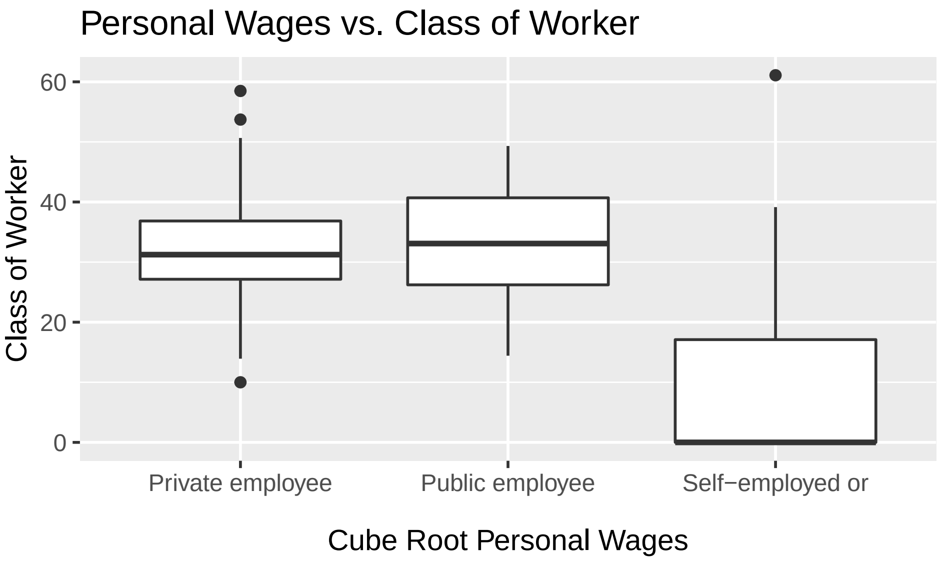

Figure 6: Public employees had greater variability in cube root personal wages than private employees did, while self-employed workers had an extremely right skewed distribution of cube root personal wages that was also very low in magnitude.

```
## I(WKHP^2)     6.75e-03   1.29e-03    5.24  9.6e-07 ***
## SEMP         -4.14e-04   6.22e-05   -6.65  1.8e-09 ***
##
## Residual standard error: 8.74 on 95 degrees of freedom
## Multiple R-squared:  0.508,  Adjusted R-squared:  0.493
## F-statistic: 32.7 on 3 and 95 DF,  p-value: 1.3e-14
```

Table B: The model containing only our three chosen quantitative predictor variables had an R^2 value of 49.3%, a residual squared error of 8.74 cube root dollars ($667.63), and an F-test p-value of 1.3e-14.

```
## `geom_smooth()` using formula 'y ~ x'
```

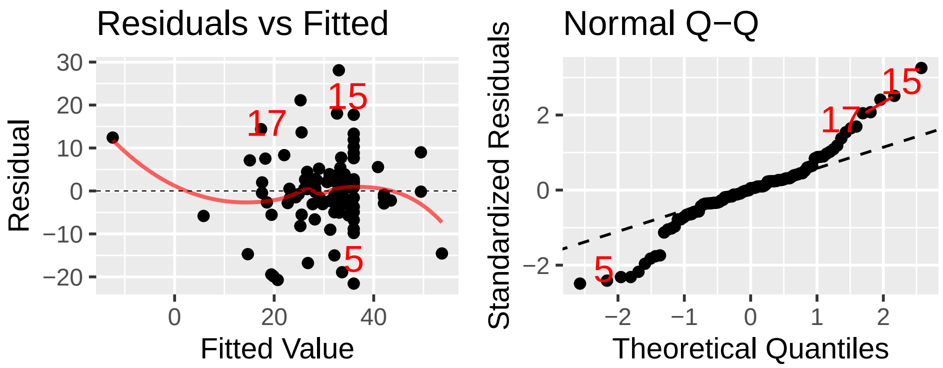

Figure 7: The residuals vs fitted plot of our mondel containing only quantitative explanatory variables demonstrated some heteroscedasticity, and both ends of the qqplot stray rather far from the normal line.

```
##
## (Intercept)
## POVPIP
## I(WKHP^2)
## SEMP
## I(Years_married^2)
## as.factor(ENG)2
## as.factor(ENG)3
## as.factor(Class_of_worker)Public employee
## as.factor(Class_of_worker)Self-employed or\n
##
## (Intercept)
## POVPIP
## I(WKHP^2)
## SEMP
## I(Years_married^2)
## as.factor(ENG)2
## as.factor(ENG)3
## as.factor(Class_of_worker)Public employee
## as.factor(Class_of_worker)Self-employed or\n
##
## (Intercept)
## POVPIP
## I(WKHP^2)
## SEMP
## I(Years_married^2)
## as.factor(ENG)2
## as.factor(ENG)3
## as.factor(Class_of_worker)Public employee
## as.factor(Class_of_worker)Self-employed or\n
##
## (Intercept)
## POVPIP
## I(WKHP^2)
## SEMP
## I(Years_married^2)
## as.factor(ENG)2
## as.factor(ENG)3
## as.factor(Class_of_worker)Public employee
## as.factor(Class_of_worker)Self-employed or\n
##
## (Intercept)
## POVPIP
## I(WKHP^2)
## SEMP
## I(Years_married^2)
## as.factor(ENG)2
## as.factor(ENG)3
## as.factor(Class_of_worker)Public employee
## as.factor(Class_of_worker)Self-employed or\n
##
## Residual standard error: 7.8 on 90 degrees of freedom
## Multiple R-squared:  0.629,  Adjusted R-squared:  0.596
## F-statistic:   19 on 8 and 90 DF,  p-value: <2e-16
```

Table C: The kitchen-sink model with all six predictors income-to-poverty ratio, hours worked per week, self-employment income, years married, ability to speak English, and class of worker had an R^2 value of 59.6%, a residual squared error of 7.8 cube root dollars ($474.55), and an F-test p-value <2e-16.

```
## `geom_smooth()` using formula 'y ~ x'
```

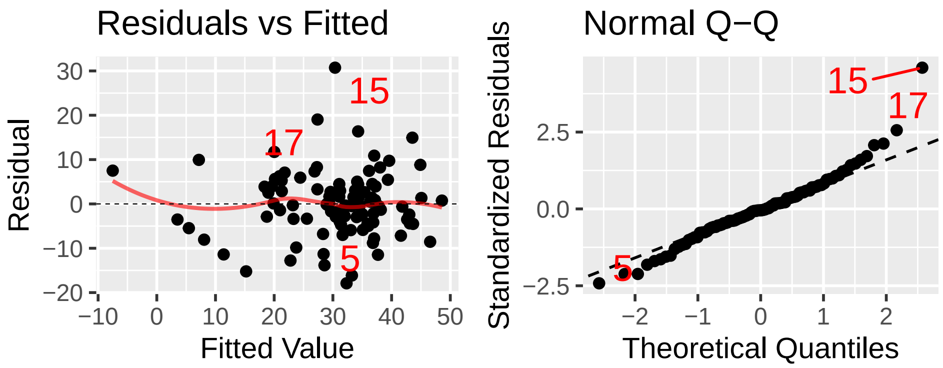

Figure 8: The kitchen-sink model better satisfies MLR conditions, as the points in the residuals vs fitted plot are much more evenly spread and they also nearly entirely follow the normal line in the qqplot.

```
##              rsq   adjr2      cp     rss POVPIP I.WKHP.2. SEMP
## 1  ( 1 ) 0.24521 0.23743 87.8863 11132.5                *     
## 2  ( 1 ) 0.43826 0.42655 43.1115  8285.3                *    *
## 3  ( 1 ) 0.50815 0.49262 28.1751  7254.3      *         *    *
## 4  ( 1 ) 0.56121 0.54253 17.3203  6471.8      *         *    *
## 5  ( 1 ) 0.60020 0.57871  9.8716  5896.7      *         *    *
## 6  ( 1 ) 0.62150 0.59682  6.7104  5582.5      *         *    *
## 7  ( 1 ) 0.62711 0.59842  7.3523  5499.9      *         *    *
## 8  ( 1 ) 0.62856 0.59555  9.0000  5478.4      *         *    *
##          I.Years_married.2. as.factor.ENG.2 as.factor.ENG.3
## 1  ( 1 )
## 2  ( 1 )
## 3  ( 1 )
## 4  ( 1 )                                  *
## 5  ( 1 )                                  *
## 6  ( 1 )                  *               *
## 7  ( 1 )                  *               *               *
## 8  ( 1 )                  *               *               *
##          as.factor.Class_of_worker.Public.employee
## 1  ( 1 )
## 2  ( 1 )
## 3  ( 1 )
## 4  ( 1 )
## 5  ( 1 )
## 6  ( 1 )
## 7  ( 1 )
## 8  ( 1 )                                         *
##          as.factor.Class_of_worker.Self.employed.or................................................
## 1  ( 1 )
## 2  ( 1 )
## 3  ( 1 )
## 4  ( 1 )
## 5  ( 1 )
## 6  ( 1 )
## 7  ( 1 )
## 8  ( 1 )
```

Table D: Observing R^2 values and Mallow's Cp, we decided that the best model while accounting for parsimony was that with four predictors: income-to-poverty ratio, hours worked per week, self-employment income, and ability to speak English.

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.60e+01   2.64e+00    6.05  3.1e-08 ***
## POVPIP            2.35e-02   6.17e-03    3.81  0.00025 ***
## I(WKHP^2)         6.48e-03   1.23e-03    5.27  8.6e-07 ***
## SEMP             -4.03e-04   6.15e-05   -6.56  3.0e-09 ***
## as.factor(ENG)2 -7.01e+00   1.98e+00   -3.53  0.00065 ***
## as.factor(ENG)3 -2.71e+00   2.54e+00   -1.07  0.28747
##
## Residual standard error: 8.29 on 93 degrees of freedom
## Multiple R-squared:  0.567,  Adjusted R-squared:  0.543
## F-statistic: 24.3 on 5 and 93 DF,  p-value: 1.42e-15
```

Table E: The final model with explanatory variables income-to-poverty ratio, hours worked per week, self-employment income, and ability to speak English had an R^2 value of 54.3%, a residual squared error of 8.29 cube root dollars ($569.72), and an F-test p-value of 1.42e-15.

```
## `geom_smooth()` using formula 'y ~ x'
```

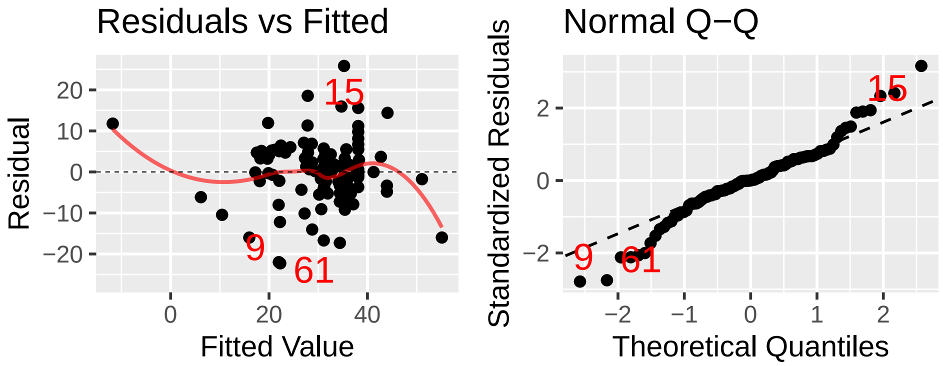

Figure 9: The residuals vs fitted plot of our final model demonstrated random spread and homoscedacisity. The normal qq plot was slightly concerning, since there were many points on the right end that deviated far from the normal distribution for errors. However, considering that this was real world data regarding income, we accounted for the fact that it was bound to be skewed. It would be wise to be cautious of this model, especially on the higher end of personal income, but we proceeded with caution.

```
## Analysis of Variance Table
##
## Model 1: I(Cuberoot_wages) ~ POVPIP + I(WKHP^2) + SEMP
## Model 2: I(Cuberoot_wages) ~ POVPIP + I(WKHP^2) + SEMP + as.factor(ENG)
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     95 7254
## 2     93 6393  2       861 6.26 0.0028 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table F: Model 2, the one containing the additional variable of ability to speak English, had a significant p-value of 0.0028, indicating that the predictor ENG captures a significant of predictability for cube root personal wages.

```
##                   GVIF Df GVIF^(1/(2*Df))
## POVPIP          1.2282  1          1.1082
## I(WKHP^2)       1.1369  1          1.0662
## SEMP             1.1180  1          1.0574
## as.factor(ENG) 1.1322  2          1.0315
```

Table G: To ensure that there was no multicollinearity present among the predictors, we checked the VIFs of each predictor in mod7 and found that none were over 5. Thus, there was no issue of multicollinearity.

Interpretations

Kitchen Sink Model

Once we had drawn out six explanatory variables from our univariate and bivariate analyses (income-to-poverty ratio, hours worked per week, self-employment income, years married, ability to speak English, and class of worker), we threw all of them into a kitchen sink model while maintaining the re-expressions, since they were necessary to force our data to be linear. We attempted one without any re-expression, but all of the coefficient terms were less significant as a result, and the following step of applying the bestsubsets method also yielded lower R^2 values. The kitchen-sink model with re-expression gave us an R^2 value of 54.2%, residual standard error of 8.3 cube root dollars ($571.79), and an F-test p-value of 5.09e-15. Although the coefficients of all six variables had p-values less than 0.05 and thus proved to be significant, some, like those for years married and class of worker, strayed closer to the cutoff than others did. So, we decided to try a more concise method of finding significant predictor variables.

Automated Variable Selection (Best Subsets Method)

After viewing the fit of the kitchen sink model, we used the bestsubsets method to find which of the six variables contribute the most to predicting personal wages. We performed two rounds of the automated variable selection method: the first was in order to find the most significant combination of quantitative predictor variables, and the second was to find the best model composed of both quantitative and qualitative predictors. From the first round, we found that the best model (good fit and satisfying conditions of MLR) we could produce with exclusively quantitative predictors consisted of income-to-poverty ratio, hours worked per week, and self-employment income, captured in mod5 and as shown in Table B and Figure 7.

We then used bestsubsets again on the kitchen-sink model, i.e. all predictor variables, to see if any variables did not contribute significantly to the model. From this, we found that the model containing the four variables income-to-poverty ratio, hours worked per week, self-employment income, and ability to speak English was our most desired in terms of fit. In comparison to using the kitchen-sink model, using the bestsubsets method allowed us to have both more control over the selection of variables and more freedom to customize the combination of high R^2, low Mallow's Cp, and optimal parsimony.

This model with the four predictors income-to-poverty ratio, hours worked per week, self-employment income, and ability to speak English had an R^2 value of 51.3%, a residual standard error of 8.56 cube root dollars ($627.22), and the F-test p-value was 7.84e-15. Furthermore, all of the variables demonstrated considerable significance with p-values well below 0.05, and the model mostly satisfied the conditions of MLR (Figure 9). Thus, we decided to use this as our final model for predicting personal wages.

ANOVA for Qualitative Predictor

As a last concrete step, we conducted an ANOVA F-test to determine whether or not our qualitative predictor, ability to speak English, truly contributed to predicting wages. In the case of not having significant predictability, we would be able to remove the variable for a simpler model. However, as seen in Table F, we found that Model 2 (model including ENG) demonstrated a significant p-value of 0.029, meaning that we should include the explanatory variable. We also took the conditions for multiple linear regression into consideration, and our conclusion from observing the conditions relatively agreed with our decision to keep four explanatory variables.

Bootstrapping

Due to our conditions being slightly concerning, we did bootstrap confidence intervals on all of the predictor variable coefficients to ensure that they were still significant at 95% confidence (if significant, would not include 0). We found that all the bootstrapping distributions were fairly symmetric (normal), so we obtained 95% confidence intervals and found all terms to be significant. We did not include them in this file, as we did not want to have multiples of 10,000 randomizations slow our knitting process significantly. However, the 95% confidence intervals are as listed below and we see that none include 0.

POVPIP: (0.0095403, 0.0324650)

WKHPˆ2: (0.0037758, 0.0100915)

SEMP: (-0.00090182, -0.00025880)

ENG: (-4.88787, -0.84019)

Conclusion

Through our investigation, we have concluded that we may predict personal wages or salary income in the past 12 months with some accuracy (although it would be wise to exercise caution when using this model for prediction due to the fit of conditions). Our final explanatory variables were income-to-poverty ratio, usual hours worked per week past 12 months, self employment income in the last 12 months, and ability to speak English. We passed this model through a rigorous set of comparisons with models re-expressed variables and also many containing other numbers and combinations of variables. As a final measure of determining the variables' combined strength of predictability, we checked their multicollinearity and found no significant amount. Our model managed to account for 51.3% of total variability in personal wages. Although the condition plots were not perfect, the fact that none of the 95% bootstrap confidence intervals from the four predictors contained 0 gave us some assurance that all of the predictors' coefficients were indeed significant. The regression equation that we built from our output in Table E indicates that, as income-to-poverty ratio, hours worked per week, and ability to speak English increase (most influential), predicted cube root personal wages does as well, and as self-employment income increases, predicted cube root personal wages decreases slightly.