DS 803: Fundamentals of Statistical Analysis
Final Project: Analyzing Education and Income
Shawn Bedard, Claire Ling, Hannah Wirth

Introduction

This project investigates the following questions: What is the relationship between an individual's income and their educational attainment? Do individuals with higher educational attainment have higher levels of income (measured in wages/salary)? What is the average level of income based on the average level of education? Is the average level of income estimated in this analysis comparable to the actual average income of New Hampshire individuals?

The American Community Survey (ACS) was selected as a dataset to investigate the defined research questions. The ACS is an annual survey sent to approximately 3.5 million American households to collect social, economic, housing, and demographic information.

Specifically, the 2013 ACS 1-year estimates from the New Hampshire population dataset was chosen for analysis. This dataset was sourced from the Census Bureau's Public Use Microdata Sample (PUMS). The dataset focusing on individual persons (rather than households) where one row of data represents one individual's response was used. Using the New Hampshire population data allowed for specific analysis of the New Hampshire population.

The investigation into the defined research questions was carried out in R, including all data cleaning and statistical analyses.

Data Preparation

The 2013 ACS 1-year estimates dataset originally contained 283 different columns. To simplify analysis and calculations, a subset of this dataset was taken to include only three variables: age (AGEP), wages or salary income in the past 12 months (WAGP), and educational attainment (SCHL). While the analysis focused only on income and educational attainment, the age variable was necessary to filter out certain age groups. The subsetted data was further filtered to only contain responses of individuals who were between the age of 15 and 65, the working population age. Individuals younger than 15 and older than 65 may be considered outliers as they are either too young to possess a significant amount of education or to earn income (i.e. a preschooler cannot earn income) or too old to be earning a working income (i.e. a 90-year-old man or woman is more likely to not be working). After filtering, the resulting dataset contained 11,475 observations.

Sampling

To conduct statistical analysis on the data, a random sample of 50% of the data was taken. The random sample resulted in 5,738 observations.

```
> head(sample)
  AGEP   WAGP SCHL
1   18      0   14
2   44  23000   19
3   35 100000   21
4   52  85000   21
5   50      0   16
6   52  65000   16
```

Figure 1: First six rows of sample dataset

Variables

Variable	*Description*
AGEP	Age
WAGP	Wages or salary income past 12 months
SCHL	Educational attainment

Variable distributions

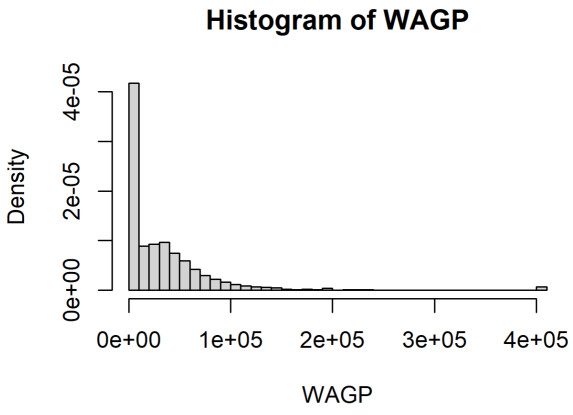

Figure 2: Density plot of WAGP

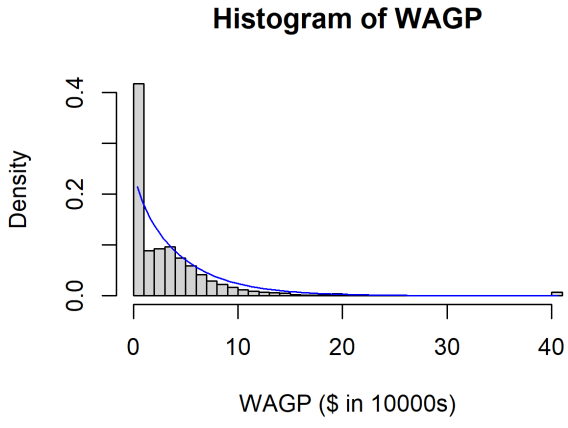

Figure 3: WAGP, gamma distribution

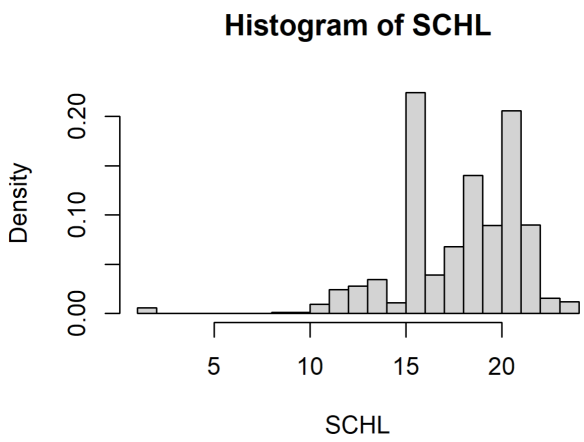

Figure 4: Density plot of SCHL

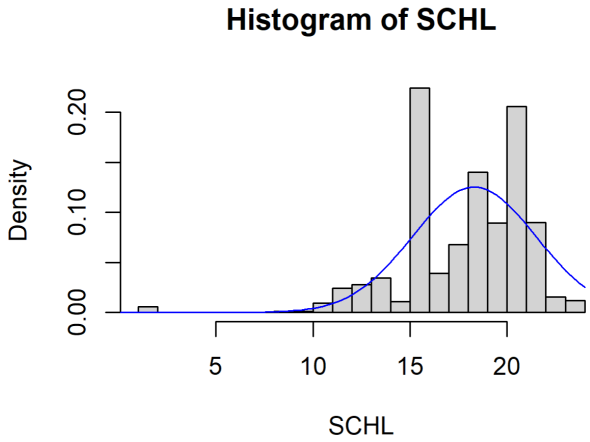

Figure 5: SCHL, normal distribution

Income is roughly gamma distributed, described with a shape of 0.93 and rate of 0.205, while educational attainment is roughly normally distributed, described with a mean of 18.3 and standard deviation of 3.18.

Summary Statistics

```
> summary(sample)
      AGEP            WAGP              SCHL
 Min.   :15.00   Min.   :     0   Min.   : 1.0
 1st Qu.:29.00   1st Qu.:     0   1st Qu.:16.0
 Median :45.00   Median : 20000   Median :19.0
 Mean   :42.26   Mean   : 33763   Mean   :18.3
 3rd Qu.:55.00   3rd Qu.: 49850   3rd Qu.:21.0
 Max.   :65.00   Max.   :406000   Max.   :24.0
```

Figure 6: Summary statistics of AGEP, WAGP, and SCHL

The AGEP variable has a minimum age of 15 and a maximum age of 65, which was defined during the data preparation stage. The median age of the sample is 45 years and the average age is approximately 42 years.

The WAGP variable has a minimum income of $0 and a maximum income of $406,000. The median income is $20,000. The average income of the sample is $33,763.

The SCHL variable has a minimum educational attainment of 1, corresponding to “No schooling completed”, and a maximum educational attainment of 24, corresponding to “Doctorate degree”. The median educational attainment is 19.0, corresponding to “1 or more years of college credit, no degree”, and the average educational attainment is 18.3, approximately corresponding to “Some college, but less than 1 year”.

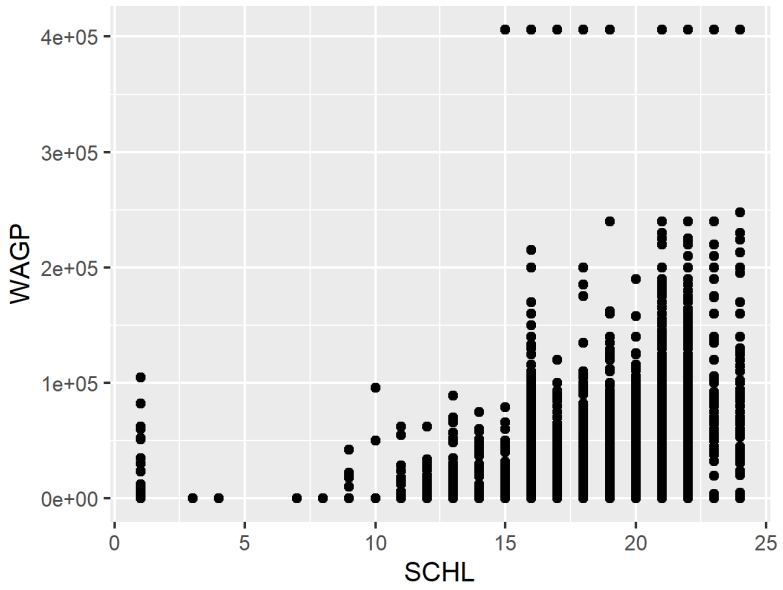

Figure 7: Scatterplot of WAGP vs. SCHL

```
cor(sample$SCHL, sample$WAGP)
[1] 0.3179663
```

A simple scatterplot of WAGP versus SCHL appears to show that as educational attainment increases, income also increases. The correlation calculation of WAGP and SCHL is positive, confirming the scatterplot's inference, and also indicates a low correlation between the two variables.

Sample means

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

```
mean(sample$WAGP)
mean(sample$SCHL)
> mean(sample$WAGP)
[1] 33763.01
> mean(sample$SCHL)
[1] 18.30185
```

The sample mean of income, $\bar{x}_{WAGP}$, is equal to $33,763, and the sample mean of educational attainment, $\bar{x}_{SCHL}$, is equal to 18.3.

Confidence intervals – sample mean

```
x1 = sample$WAGP
sigma1 = sd(sample$WAGP)
WAGP.CI=function(obs, sigma, alpha){
  n=length(obs)
  ME1=qgamma(alpha/2, 0.93, 0.205)*sigma/sqrt(n)
  ME2=qgamma(1-alpha/2, 0.93, 0.205)*sigma/sqrt(n)
  c(mean(obs)-ME1, mean(obs)+ME2)
}
WAGP.CI(x1, sigma1, 0.05)
```

```
[1] 33704.63 44901.95
```

The confidence interval of the sample mean of income, $C_{\bar{x},\alpha\,WAGP}$, at a 95% confidence level ($\alpha = 0.05$) is equal to $(33704.63,\ 44901.95)$. This can be interpreted as being 95% confident that the population parameter, mean income, is between $33,704.63 and $44,901.95. The actual sample mean is equal to $33,763.

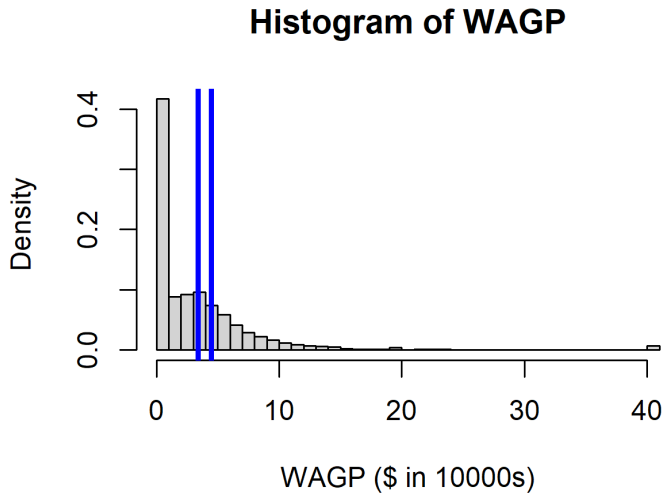

Figure 8: WAGP confidence interval

Figure 7 shows a plot of the sample data with the confidence interval plotted in blue.

```
x2 = sample$SCHL
sigma2 = sd(sample$SCHL)
SCHL.CI=function(obs, sigma, alpha){
  n=length(obs)
  ME=qnorm(1-alpha/2)*sigma/sqrt(n)
  c(mean(obs)-ME, mean(obs)+ME)
}
SCHL.CI(x2, sigma2, 0.05)
```

```
[1] 18.21945 18.38424
```

The confidence interval of the sample mean of educational attainment, $C_{\bar{x},\alpha\,SCHL}$, at a 95% confidence level ($\alpha = 0.05$) is equal to $(18.21945,\ 18.38424)$. This can be interpreted as being 95% confident that the population parameter, mean educational attainment, is between 18.21 and 18.38. These values can be rounded to 18, corresponding to an educational attainment of "Some college, but less than 1 year".

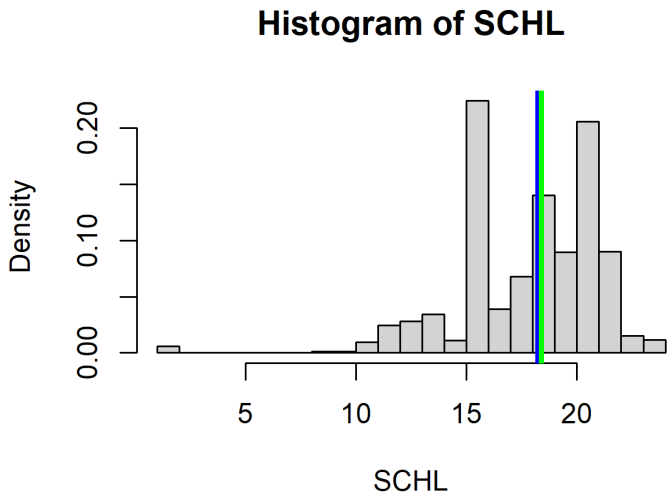

Figure 9: SCHL confidence interval

Figure 8 shows a plot of the sample data with the lower bound of the confidence interval in blue and the upper bound in green.

Estimating Population Parameters

Method of Moments

The Method of Moments estimator for the gamma distribution was used to determine the mean of WAGP.

$$E(X_i) = \mu$$

$$E(X) = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\mu}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{x}$$

```
> mean(sample$WAGP)
[1] 33763.01
mean(sample$WAGP)
```

The Method of Moments estimator for the normal distribution was used to determine the mean of SCHL.

$$E(X_i) = \mu$$

$$E(X) = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\mu}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{x}$$

```
mean(sample$SCHL)
> mean(sample$SCHL)
[1] 18.30185
```

The normal Method of Moments estimator of the mean can be described as the sample mean. The Method of Moments estimation for SCHL resulted in a value of 18.30185.

Maximum Likelihood Estimator

The Maximum Likelihood Estimator (MLE) for the gamma distribution was used to determine the mean of WAGP using a shape of 0.93 and rate of 0.205 (determined in the variable distributions).

$$\hat{\mu}_{MLE} = \frac{\bar{x}}{a}$$

```
gamma_mle = mean(standard_wage)/gamma_dist$estimate[1]*10000
gamma_mle
```

```
> gamma_mle
   shape
36249.8
```

The gamma MLE distribution can be estimated as the sample mean over the shape parameter, where $\bar{x}$ = sample mean and a = shape parameter. The MLE estimation for WAGP resulted in a value of $36,249.80.

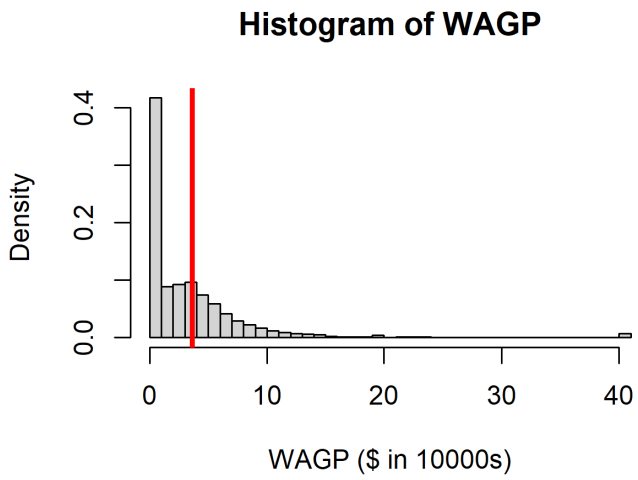

Figure 10: MLE of WAGP

Figure 9 shows a plot of the sample data and the MLE of WAGP plotted in red.

The MLE was also used for the normal distribution to estimate the mean of SCHL.

$$\hat{\mu}_{MLE} = \bar{x}$$

```
mean(sample$SCHL)
```

```
> mean(sample$SCHL)
[1] 18.30185
```

The normal distribution MLE can be estimated as the mean of the data. The MLE estimation for SCHL resulted in a value of 18.3. This is roughly equivalent to “Some college, but less than 1 year”.

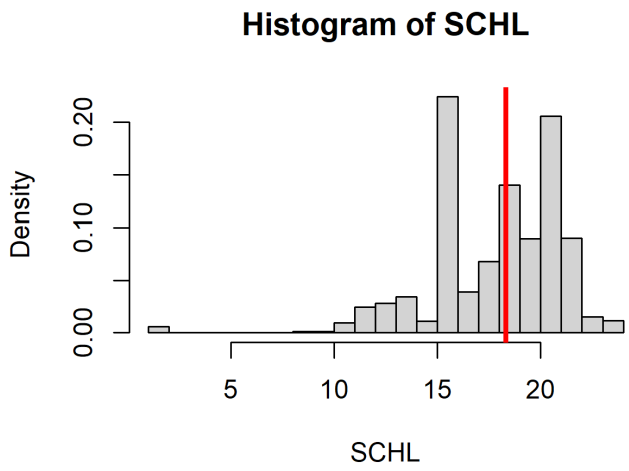

Figure 11: MLE of SCHL

Figure 10 shows a plot of the sample data and the MLE of SCHL plotted in red.

Inferential Analysis

Hypothesis Testing

Hypothesis testing was used to determine whether the average income of individuals in high school or with the highest educational attainment of a high school degree is greater than $25,000.

H_0: μ_{WAGP} is equal to 25,000

H_a: μ_{WAGP} is greater than 25,000

```
sample_hs <- subset(sample, SCHL >= 12 & SCHL <= 17)
```

A new subset of the sample data was taken to only include individuals with a high school educational attainment.

```
test.stat_hs = (mean(sample_hs$WAGP)-25000)/(sd(sample_hs$WAGP)/sqrt(5738))
pval = 1 - pt(test.stat_hs, 5737)
list(test.stat_hs = test.stat_hs, pval = pval)
```

```
$test.stat_hs
[1] -14.8541

$pval
[1] 1
```

Given that the p-value is equal to 1, which is greater than a 0.05 level of significance, this can be interpreted as failing to reject the null. It cannot be said that the average income of individuals with a high school educational attainment will be greater than $25,000.

A second hypothesis test was calculated to determine whether the average income of individuals in college or with the highest educational attainment of a college or other higher-ed degree is greater than $50,000.

H_0: μ_{WAGP} is equal to 50,000

H_a: μ_{WAGP} is greater than 50,000

```
sample_col <- subset(sample, SCHL >= 18 & SCHL <= 24)
```

A subset of the sample data was taken to only include individuals with a college or higher-ed (masters degree, doctorate degree) educational attainment.

```
test.stat_col = (mean(sample_col$WAGP)-50000)/(sd(sample_col$WAGP)/sqrt(5738))
pval_col = 1 - pt(test.stat_col, 5737)
list(test.stat_col=test.stat_col, pval_col=pval_col)
```

```
$test.stat_col
[1] -9.448912

$pval_col
[1] 1
```

Given that the p-value is equal to 1, which is greater than a 0.05 level of significance, this can be interpreted as failing to reject the null. It cannot be said that the average income of individuals with a college level educational attainment will be greater than $50,000.

Linear Regression

Least Squares

To estimate the simple linear regression equation of WAGP and SCHL, the following assumptions can be made:

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

$$V(Y_i|X_i) = \sigma^2$$

$$Cov(Y_i, Y_j) = 0, \forall i \neq j$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$WAGP_i = \beta_0 + \beta_1 SCHL_i + \varepsilon_i$$

Using the least squares method, β_0 and β_1 were estimated for the regression equation, $WAGP_i = \beta_0 + \beta_1 SCHL_i + \varepsilon_i$. A simple linear regression was estimated in R using the linear model (lm) function to determine the beta and coefficient estimates. The as.factor() function was used to register SCHL, a categorical variable, as a factor. This allows the regression to calculate an intercept at every level of educational attainment.

```
sample_reg <- subset(sample, SCHL >= 12)
```

```
> head(sample_reg)
  AGEP   WAGP SCHL
1   18      0   14
2   44  23000   19
3   35 100000   21
4   52  85000   21
5   50      0   16
6   52  65000   16
```

To conduct the regression analysis, a subset of the data was taken to exclude any data points that included individuals with an educational attainment less than high school. This was done to remove results that include individuals that have education levels that are unable to make income.

```
model = lm(sample_reg$WAGP ~ as.factor(sample_reg$SCHL))
model
```

```
> model = lm(sample_reg$WAGP ~ as.factor(sample_reg$SCHL))
> model

Call:
lm(formula = sample_reg$WAGP ~ as.factor(sample_reg$SCHL))

Coefficients:
                    (Intercept)  as.factor(sample_reg$SCHL)13
                           2497                          3154
as.factor(sample_reg$SCHL)14  as.factor(sample_reg$SCHL)15
                           3808                         18262
as.factor(sample_reg$SCHL)16  as.factor(sample_reg$SCHL)17
                          21618                         16826
as.factor(sample_reg$SCHL)18  as.factor(sample_reg$SCHL)19
                          22329                         22007
as.factor(sample_reg$SCHL)20  as.factor(sample_reg$SCHL)21
                          33595                         48645
as.factor(sample_reg$SCHL)22  as.factor(sample_reg$SCHL)23
                          58043                         92788
as.factor(sample_reg$SCHL)24
                          79815
```

The resulting regression equation can be described as:

$$WAGP = 2497 + 3154\,SCHL_{13} + 3808\,SCHL_{14} + 18262\,SCHL_{15} + 21618\,SCHL_{16}$$
$$16826\,SCHL_{17} + 22329\,SCHL_{18} + 22007\,SCHL_{19} + 33595\,SCHL_{20} + 48645\,SCHL_{21}$$
$$58043\,SCHL_{22} + 92788\,SCHL_{23} + 79815\,SCHL_{24}$$

Holding all else equal, the regression equation estimates that income will be equal to $2,497. Suppose that a person has a educational attainment of 21 (associated with “Bachelor’s degree”), holding all else equal, the regression equation estimates that income will be equal to $2497 + 48645$, or an income of $51,142.

```
summary(model)
> summary(model)

Call:
lm(formula = sample_reg$WAGP ~ as.factor(sample_reg$SCHL))

Residuals:
   Min     1Q Median     3Q    Max
-95285 -24115  -6234  12864 386677
```

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                         2497       3853   0.648 0.516951
as.factor(sample_reg$SCHL)13        3154       5268   0.599 0.549451
as.factor(sample_reg$SCHL)14        3808       5034   0.756 0.449436
as.factor(sample_reg$SCHL)15       18262       6917   2.640 0.008308 **
as.factor(sample_reg$SCHL)16       21618       4057   5.328 1.03e-07 ***
as.factor(sample_reg$SCHL)17       16826       4912   3.426 0.000618 ***
as.factor(sample_reg$SCHL)18       22329       4493   4.969 6.92e-07 ***
as.factor(sample_reg$SCHL)19       22007       4175   5.271 1.41e-07 ***
as.factor(sample_reg$SCHL)20       33595       4347   7.728 1.29e-14 ***
as.factor(sample_reg$SCHL)21       48645       4075  11.937  < 2e-16 ***
as.factor(sample_reg$SCHL)22       58043       4344  13.363  < 2e-16 ***
as.factor(sample_reg$SCHL)23       92788       6202  14.961  < 2e-16 ***
as.factor(sample_reg$SCHL)24       79815       6773  11.785  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45590 on 5619 degrees of freedom
Multiple R-squared:  0.1406,    Adjusted R-squared:  0.1388
F-statistic:  76.6 on 12 and 5619 DF,  p-value: < 2.2e-16
```

The summary statistics of the model show that the intercepts of $SCHL_{16}$, $SCHL_{17}$, $SCHL_{18}$, $SCHL_{19}$, $SCHL_{20}$, $SCHL_{21}$, $SCHL_{22}$, $SCHL_{23}$, and $SCHL_{24}$ are statistically significant. The R-squared value of 0.1406 indicates that the regression only accounts for about 14% of the variance.

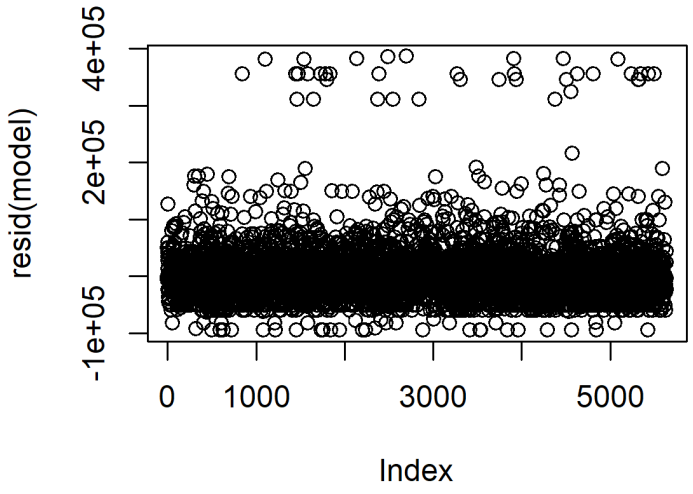

Figure 12: Plot of residuals from model

```
durbinWatsonTest(model)
```

```
> durbinWatsonTest(model)
 lag Autocorrelation D-W Statistic p-value
   1    0.0007751057      1.998445   0.914
 Alternative hypothesis: rho != 0
```

The Durbin-Watson test on the regression model resulted in a p-value of 0.914, which supports failing to reject the null hypothesis. There is evidence to support that the residuals are not autocorrelated.

```
bptest(model)
```

```
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 165.62, df = 12, p-value < 2.2e-16
```

The studentized Breusch-Pagan test on the regression model resulted in a p-value less than 2.2e-16, which supports rejecting the null hypothesis. There is evidence to support that there is some non-constant variance in the model.

```
confint(model)
```

```
> confint(model)
                                    2.5 %    97.5 %
(Intercept)                     -5056.496  10050.92
as.factor(sample_reg$SCHL)13 -7174.501  13482.18
as.factor(sample_reg$SCHL)14 -6061.256  13677.33
as.factor(sample_reg$SCHL)15  4702.350  31821.00
as.factor(sample_reg$SCHL)16 13663.860  29571.77
as.factor(sample_reg$SCHL)17  7196.960  26455.22
as.factor(sample_reg$SCHL)18 13520.584  31138.05
as.factor(sample_reg$SCHL)19 13821.537  30191.52
as.factor(sample_reg$SCHL)20 25072.173  42116.83
as.factor(sample_reg$SCHL)21 40656.260  56634.05
as.factor(sample_reg$SCHL)22 49527.974  66558.49
as.factor(sample_reg$SCHL)23 80629.333 104946.69
as.factor(sample_reg$SCHL)24 66537.776  93092.27
```

A 95% confidence interval was calculated for the intercept and slope values.

Additional Regression Analysis

After analyzing the results of the initial regression equation, ($WAGP_i = \beta_0 + \beta_1 SCHL_i + \varepsilon_i$), further analysis was done with additional variables from the original dataset to determine whether other variables may improve the R-squared value as well as improve the income estimates.

$$WAGP_i = \beta_0 + \beta_1 SCHL_i + \beta_2 AGEP_i + \varepsilon_i$$

The first additional regression equation includes the variable of age, AGEP. AGEP is measured in years and ranges from 15 to 65 years (originally defined in the data preparation stage). Similar to the original regression dataset, the data only includes individuals with an educational attainment of high school or above.

```
model2 = lm(sample_reg$WAGP ~ as.factor(sample_reg$SCHL) + sample_reg$AGEP)
model2
```

```
> model2 = lm(sample_reg$WAGP ~ as.factor(sample_reg$SCHL) + sample_reg$AGEP)
> model2

Call:
lm(formula = sample_reg$WAGP ~ as.factor(sample_reg$SCHL) + sample_reg$AGEP)

Coefficients:
                  (Intercept)  as.factor(sample_reg$SCHL)13
                      -6068.7                        2175.5
as.factor(sample_reg$SCHL)14  as.factor(sample_reg$SCHL)15
                       2549.9                       11617.5
as.factor(sample_reg$SCHL)16  as.factor(sample_reg$SCHL)17
                      13398.8                        8325.1
as.factor(sample_reg$SCHL)18  as.factor(sample_reg$SCHL)19
                      15526.2                       15487.8
as.factor(sample_reg$SCHL)20  as.factor(sample_reg$SCHL)21
                      24360.5                       39957.4
as.factor(sample_reg$SCHL)22  as.factor(sample_reg$SCHL)23
                      47966.6                       81840.6
as.factor(sample_reg$SCHL)24               sample_reg$AGEP
                      69996.7                         382.5
```

The resulting second regression equation can be described as:

$$WAGP = -6068.7 + 2175.5\,SCHL_{13} + 2549.9\,SCHL_{14} + 11617.5\,SCHL_{15} + 13398.8\,SCHL_{16} + 8325.1\,SCHL_{17} + 15526.2\,SCHL_{18} + 15487.8\,SCHL_{19} + 24360\,SCHL_{20} + 39957.4\,SCHL_{21} + 47966.6\,SCHL_{22} + 81840.6\,SCHL_{23} + 69996.7\,SCHL_{24} + 382.5\,AGEP$$

Holding all else equal, the second regression equation estimates that income will be equal to -$6,068.70. Suppose that a person has a educational attainment of 21 (associated with

"Bachelor's degree"), holding all else equal, the regression equation estimates that income will be equal to $-6068.7 + 39957.4$, or an income of $33,888.70. Suppose the same person also has an age of 22 - their income would be equal to $-6068.7 + 39957.4 + 382.5(22)$, or an income of $42,303.70.

```
summary(model2)
```

```
> summary(model2)

Call:
lm(formula = sample_reg$WAGP ~ as.factor(sample_reg$SCHL) + sample_reg$AGEP)

Residuals:
    Min      1Q  Median      3Q     Max
-100636  -22530   -5120   12329  392801


Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -6068.7     3951.3  -1.536 0.124625
as.factor(sample_reg$SCHL)13      2175.5     5234.8   0.416 0.677724
as.factor(sample_reg$SCHL)14      2549.9     5003.1   0.510 0.610302
as.factor(sample_reg$SCHL)15     11617.5     6912.8   1.681 0.092900 .
as.factor(sample_reg$SCHL)16     13398.8     4138.9   3.237 0.001214 **
as.factor(sample_reg$SCHL)17      8325.1     4975.4   1.673 0.094337 .
as.factor(sample_reg$SCHL)18     15526.2     4531.0   3.427 0.000615 ***
as.factor(sample_reg$SCHL)19     15487.8     4214.1   3.675 0.000240 ***
as.factor(sample_reg$SCHL)20     24360.5     4446.0   5.479 4.46e-08 ***
as.factor(sample_reg$SCHL)21     39957.4     4168.6   9.585  < 2e-16 ***
as.factor(sample_reg$SCHL)22     47966.6     4466.5  10.739  < 2e-16 ***
as.factor(sample_reg$SCHL)23     81840.6     6287.3  13.017  < 2e-16 ***
as.factor(sample_reg$SCHL)24     69996.7     6821.1  10.262  < 2e-16 ***
sample_reg$AGEP                    382.5       43.8   8.733  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45290 on 5618 degrees of freedom
Multiple R-squared:  0.1521,    Adjusted R-squared:  0.1501
F-statistic: 77.52 on 13 and 5618 DF,  p-value: < 2.2e-16
```

The summary statistics of the model show that the intercepts of $SCHL_{18}$, $SCHL_{19}$, $SCHL_{20}$, $SCHL_{21}$, $SCHL_{22}$, $SCHL_{23}$, $SCHL_{24}$, and $AGEP$ are statistically significant. The R-squared value of 0.1521 indicates that the regression only accounts for about 15% of the variance. This is a slight improvement over the original regression equation.

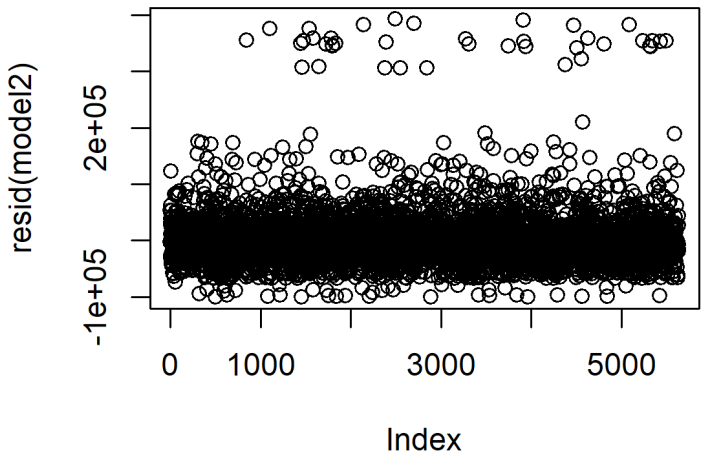

Figure 13: Plot of residuals from model2

```
durbinWatsonTest(model2)
```

```
> durbinWatsonTest(model2)
 lag Autocorrelation D-W Statistic p-value
   1     0.003629066      1.992736    0.76
 Alternative hypothesis: rho != 0
```

The Durbin-Watson test on the regression model resulted in a p-value of 0.76, which supports failing to reject the null hypothesis. There is evidence to support that the residuals are not autocorrelated.

```
bptest(model2)
```

```
> bptest(model2)

        studentized Breusch-Pagan test

data:  model2
BP = 182.35, df = 13, p-value < 2.2e-16
```

The studentized Breusch-Pagan test on the regression model resulted in a p-value less than 2.2e-16, which supports rejecting the null hypothesis. There is evidence to support that there is some non-constant variance in the model.

```
vif(model)
```

```
> vif(model)
                                GVIF Df GVIF^(1/(2*Df))
as.factor(sample_reg$SCHL) 1.206311 12        1.007846
sample_reg$AGEP            1.206311  1        1.098322
```

The addition of the age variable allowed for testing of multicollinearity. The VIF values are very low, pointing to evidence of no multicollinearity.

```
confint(model2)
```

```
> confint(model2)
                                     2.5 %     97.5 %
(Intercept)                     -13814.7846  1677.3607
as.factor(sample_reg$SCHL)13     -8086.6639 12437.7448
as.factor(sample_reg$SCHL)14     -7258.0027 12357.8305
as.factor(sample_reg$SCHL)15     -1934.2306 25169.2370
as.factor(sample_reg$SCHL)16      5285.0583 21512.5670
as.factor(sample_reg$SCHL)17     -1428.7141 18078.8269
as.factor(sample_reg$SCHL)18      6643.6336 24408.7927
as.factor(sample_reg$SCHL)19      7226.4159 23749.1058
as.factor(sample_reg$SCHL)20     15644.6273 33076.3858
as.factor(sample_reg$SCHL)21     31785.3905 48129.4489
as.factor(sample_reg$SCHL)22     39210.5785 56722.5869
as.factor(sample_reg$SCHL)23     69515.0769 94166.1086
as.factor(sample_reg$SCHL)24     56624.6170 83368.7868
sample_reg$AGEP                    296.6609   468.3979
```

A 95% confidence interval was calculated for the intercept and slope values.

Conclusions

The following answers to the initial research questions are as follows:

What is the relationship between an individual's income and their educational attainment?

> There appears to be a relationship between income and educational attainment. The linear regression model indicates that as educational attainment increases, income also increases.

Do individuals with higher educational attainment have higher levels of income (measured in wages/salary)?

Individuals with higher educational attainment do have higher levels of income. If an individual has an educational attainment of 18 (“Some college, but less than 1 year”), the first regression equation estimates their income to be equal to $24,826. In comparison, if an individual has an educational attainment of 24 (“Doctorate degree”), the regression estimates their income to be equal to $82,312.

What is the average level of income based on the average level of education?

The average level of education was determined to be 18 (rounded from 18.3), which is equivalent to “Some college, but less than 1 year” - this corresponds to an average income of $24,826.

Is the average level of income estimated in this analysis comparable to the actual average income of New Hampshire individuals?

The average level of income estimated in this analysis was calculated in the summary statistics to be approximately $33,763. Some external research through the US Census Bureau found that the average individual income in New Hampshire is $37,025. These two numbers are somewhat comparable.

In addition to answering the research questions, it was found that the WAGP variable best fit a gamma distribution and the SCHL variable best fit a normal distribution. This information was useful in calculating estimators and confidence intervals.

The hypothesis testing revealed that there is evidence to suggest that the income of individuals with a high school level educational attainment will not be greater than $20,000. The US Bureau of Labor Statistics (BLS) estimates the average salary for a high school diploma holder is $38,792. Similarly, hypothesis testing also revealed that there is evidence to suggest that the income of individuals with a college or higher-ed educational attainment will not be greater than $50,000. The BLS estimation of average salary for this group ranges from $64,896 to $97,916. The discrepancy in these numbers may be attributed to unique factors affecting the New Hampshire work population. More investigation into such factors would be beneficial to this analysis.

The results of the first regression equation that only considered educational attainment resulted in estimated incomes that were lower than the average income by education level. The results of the second regression equation resulted in estimates of income that were further from the average incomes by education level. For instance, Forbes estimates the median earnings of a 25 to 34 year old to be $59,600. The first regression equation, only considering educational attainment, estimated income for bachelor’s degree holders to be $51,142. The second regression equation, considering educational attainment and age, estimated income for bachelor’s degree holders to be $42,303.70.

To further investigate the relationship between educational attainment and income, more analysis with other variables (household data, demographics, etc.) may be beneficial. Although age was already considered, adding other variables to the regression equation may result in more accurate income estimates. Additionally, conducting the same analysis with population data from other states rather than specifically focusing on New Hampshire may yield interesting results. For instance, an analysis comparing New Hampshire population data and southern states population data may show a stronger or weaker relationship between educational attainment and income.

Furthermore, more work regarding the issue of non-constant variance found in the regression model will be needed. This could be done by transforming the data using logs or square roots.

Issues with Results and Analysis

One of the major issues we found with our analysis is the differences between our estimated income averages and the actual national and New Hampshire averages. This may be due to the estimation techniques we used compared to the source for the actual averages. It's also possible we used a different dataset or a different subset of data. An overall issue with using a survey is that the data is not verified or cleaned, which can lead to incorrect results.

Another issue we had was that we were not able to drill down into the specifics of income breakdown by major or sector. For instance, it would have been insightful to determine whether individuals in the data science field make more than individuals in the health field.

Overall, our analysis was just the first step in discovering the relationship between education and income. Deeper analysis is necessary to uncover more significant factors that explain the relationship between education and income, including the field the education is in, number of years worked in the field, race, state, and country.