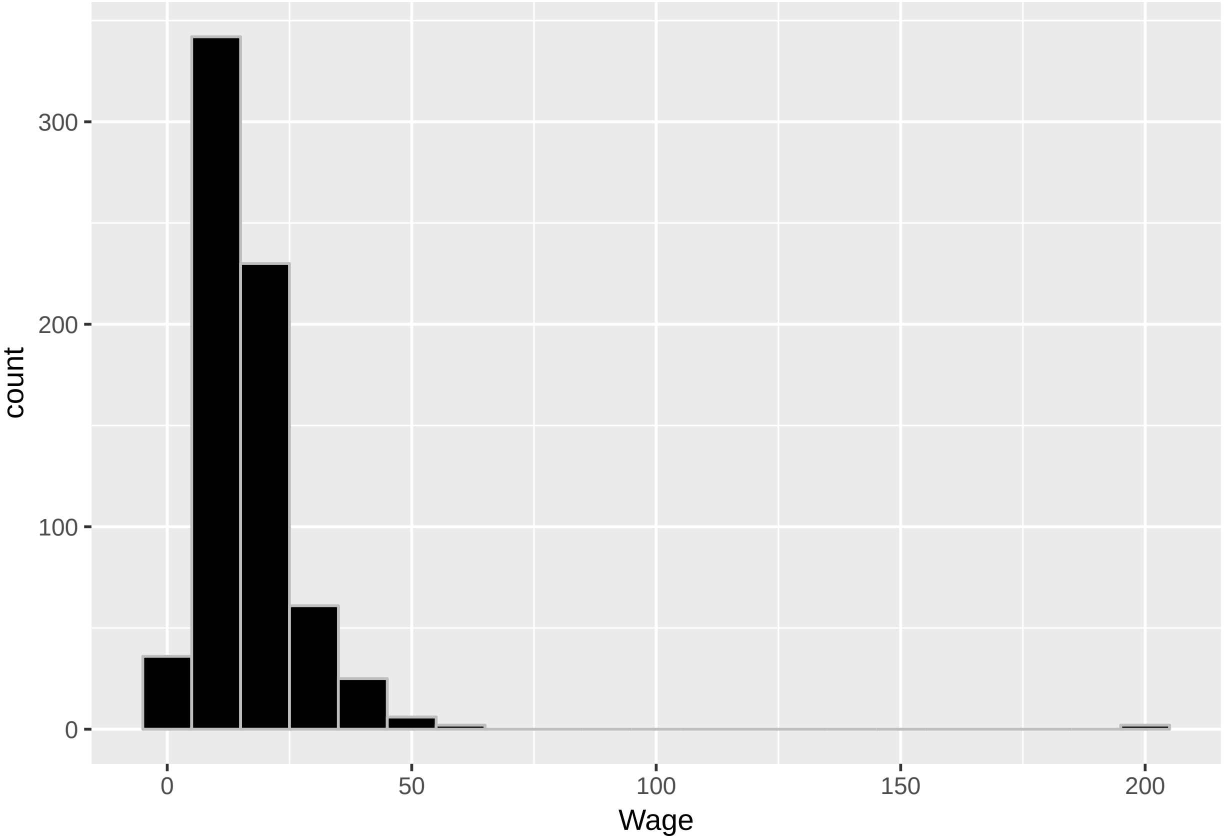
PSet 6

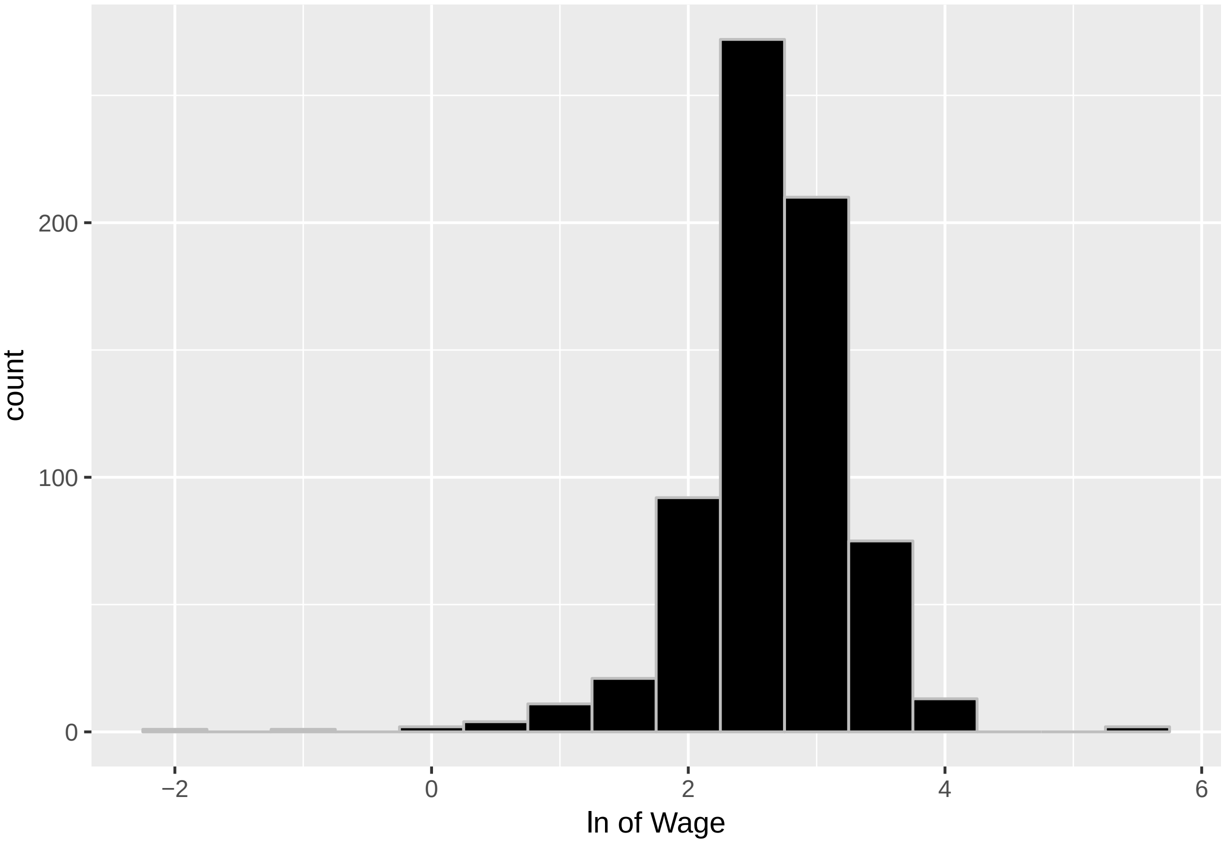
Anya Conti

April 6, 2017

1.



The data for wage seems to be skewed to the right, yet for our purposed, we need to assume normal distribution with no skew.



The data seems to look much more like a normal distribution now, without a very apparent skew.

2.a. The summary of the linear model is shown. For discussion of statistical significance, assume $\alpha = 0.05$ and two-tailed test. Something to keep in mind is that a lot of these variables might actually be highly correlated, especially the ones with both their own effect and the interaction effect like female and experience, and female experience, or where two different functions of the variable are in the model like experience and experience squared. As such, dropping variables in general would not be suggested.

The intercept refers to the case of a white male with less than high school education, no experience, no union coverage, no pension, and no health insurance paid by employer. The other variables show how someone with different characteristics compares to this baseline group. As expected, almost every education category seems to increment wage so that the intercept for that category (and wage in general, holding all other variables constant) is higher than a lower level of education. (i.e. AS is greater than HS, MS is greater than AS, etc). The exceptions to this are that Scoll (some college) is actually slightly below HS (high school), though, they are so close anyways that there does not seem to be much of a difference. Also, pd (professional degree) is negative, implying that people with a professional degree actually earn 0.238 less than people with less than a high school level of education, holding all other variables constant. This might be because of the types of jobs which people with a professional degree would hold. Also, it is not statistically significantly different from 0, with a p-value of 0.669. Of the other education categories, only AS, MS, and BS are statistically significantly different from 0, with p-values less than 0.01. The p-values of HS is 0.0878, which would be statistically significantly different from 0 if we used a one-tail test, or $\alpha = 0.10$ instead. The p-value of Scoll is 0.131998, which would be statistically significantly different from 0 if we used a one-tail test AND $\alpha = 0.10$ instead. PhD has has a p-value of 0.0953, which would be statistically significantly different from 0 if we used a one-tail test, or $\alpha = 0.10$ instead. At low levels of experience, the addition of more experience is positive (coefficient of exp is positive), but at extremely high levels would eventually be negative because the coefficient for the sqare term is negative. These are both statistically different from 0 with p-values less than 0.01. Unexpectedly, being female actually has a postive impact on the intercept of wage given no experience. This is not statistically significant from 0 however, with a p-value of 0.845. What is expected is that being female means that additional years of experience actually have a lower effect on wage compared to with being male. This term is statistically different from 0 with a p-value of 0.0302. The effects of experience, and being female are both explored more later on. As expected, having a pension, some health insurance paid by employer (hipaids), and all health insurance paid by employer (hipaida) all have a positive impact on wage, with all health insurance paid by employer having a higher impact than some health insurance paid by the employer. With p-values all less than 0.01, they are all statistically significantly less than 0. As expected, being non-white has a negative effect on wage though this is not statistically significantly different from 0 with a p-value of 0.181.

```
##
## Call:
## lm(formula = lnWage ~ hs + scoll + as + bs + ms + pd + phd +
##     exp + expsq + female + expfe + unioncov + pension + hipaids +
##     hipaida + nonwhite, data = Wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0105 -0.2428  0.0398  0.3022  2.7261
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7254656  0.0965539  17.871  < 2e-16 ***
## hs           0.1270996  0.0743360   1.710 0.087755 .
## scoll        0.1257655  0.0833953   1.508 0.131998
## as           0.3902390  0.0908152   4.297 1.98e-05 ***
## bs           0.4535787  0.0890711   5.092 4.57e-07 ***
## ms           0.4929275  0.1453076   3.392 0.000733 ***
## pd          -0.2383180  0.5564480  -0.428 0.668578
## phd          0.9294526  0.5564810   1.670 0.095329 .
## exp          0.0455367  0.0066150   6.884 1.32e-11 ***
## expsq       -0.0006932  0.0001371  -5.056 5.50e-07 ***
## female       0.0172433  0.0883595   0.195 0.845334
## expfe       -0.0079710  0.0036694  -2.172 0.030177 *
## unioncov    -0.1194222  0.5530033  -0.216 0.829089
## pension      0.2022340  0.0461696   4.380 1.37e-05 ***
## hipaids      0.1597701  0.0464277   3.441 0.000614 ***
## hipaida      0.2713099  0.0792204   3.425 0.000652 ***
## nonwhite    -0.0956531  0.0714494  -1.339 0.181094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5503 on 687 degrees of freedom
## Multiple R-squared:  0.2633, Adjusted R-squared:  0.2461
## F-statistic: 15.34 on 16 and 687 DF,  p-value: < 2.2e-16
```

2.b. Some of the binary variables for education do appear as if they can be combined. For example, the coeffiecients δ_1 and δ_2 for high school (HS) and some college (Scoll) are 0.1270996 and 0.1257655 respectively. Since these are so similar, we can use an F-test with the hypotheses: $H_0 : \delta_1 = \delta_2$ and $H_A : \delta_1 \neq \delta_2$

```
## Linear hypothesis test
##
## Hypothesis:
## hs - scoll = 0
##
## Model 1: restricted model
## Model 2: lnWage ~ hs + scoll + as + bs + ms + pd + phd + exp + expsq +
##     female + expfe + unioncov + pension + hipaids + hipaida +
##     nonwhite
##
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    688 208.03
## 2    687 208.03  1 0.0001568 5e-04 0.9819
```

Since the p-value of 0.9819 is far greater than $\alpha = 0.05$, we fail to reject the null hypothesis that the two coefficients are equal, and thus the two groups could be combined because there is no statistically significant difference between their effects on wage.

Also, the coefficient δ_4 for BS of 0.4535787 is close to both the coefficient δ_3 for AS of 0.390239 and coefficient δ_5 for MS of 0.4929275. First we can use an F-test to see if AS and BS are similar enough to combine with the hypotheses: $H_0 : \delta_3 = \delta_4$ and $H_A : \delta_3 \neq \delta_4$

```
## Linear hypothesis test
##
## Hypothesis:
## as - bs = 0
##
## Model 1: restricted model
## Model 2: lnWage ~ hs + scoll + as + bs + ms + pd + phd + exp + expsq +
##     female + expfe + unioncov + pension + hipaids + hipaida +
##     nonwhite
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    688 208.21
## 2    687 208.03  1   0.18233 0.6021  0.438
```

Since the p-value of 0.438 is far greater than $\alpha = 0.05$, we fail to reject the null hypothesis that the two coefficients are equal, and thus the two groups could be combined because there is no statistically significant difference between their effects on wage.

Then we can use an F-test to see if BS and MS are similar enough to combine with the hypotheses: $H_0 : \delta_4 = \delta_5$ and $H_A : \delta_4 \neq \delta_5$

```
## Linear hypothesis test
##
## Hypothesis:
## bs - ms = 0
##
## Model 1: restricted model
## Model 2: lnWage ~ hs + scoll + as + bs + ms + pd + phd + exp + expsq +
##     female + expfe + unioncov + pension + hipaids + hipaida +
##     nonwhite
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    688 208.06
## 2    687 208.03  1  0.024164 0.0798 0.7777
```

Since the p-value of 0.7777 is far greater than $\alpha = 0.05$, we fail to reject the null hypothesis that the two coefficients are equal, and thus the two groups could be combined because there is no statistically significant difference between their effects on wage.

Since we failed to reject both of those last two hypotheses that $\delta_3 = \delta_4$ and $\delta_4 = \delta_5$, we can then use an F-test to see if AS, BS and MS are all similar enough to combine into one group with the hypotheses: $H_0 : \delta_3 = \delta_4 = \delta_5$ and H_A : at least two of them aren't equal.

```
## Linear hypothesis test
##
## Hypothesis:
## as - bs = 0
## bs - ms = 0
##
## Model 1: restricted model
## Model 2: lnWage ~ hs + scoll + as + bs + ms + pd + phd + exp + expsq +
##     female + expfe + unioncov + pension + hipaids + hipaida +
##     nonwhite
##
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    689 208.29
## 2    687 208.03  2   0.26287 0.434 0.6481
```

Since the p-value of 0.6481 is far greater than $\alpha = 0.05$, we fail to reject the null hypothesis that the three coefficients are equal, and thus the three groups could be combined because there is no statistically significant difference between their effects on wage.

2.c. The percent change in wage for males for every additional year of experience, holding all other variables in the model constant, is given by the following expression: $(\beta_2 + 2\beta_3 * exp) * 100$ $(0.455367 - 0.001386374 * exp) * 100$ This means that for a male with no experience, the percent change in wage for 1 additional year of experience would be 45.5367 % , holding all other variables in the model constant.
For a male with 5 years of experience, the percent change in wage for 1 additional year of experience would be 44.843513 % , holding all other variables in the model constant.
For a male with 10 years of experience, the percent change in wage for 1 additional year of experience would be 44.150326 % , holding all other variables in the model constant.
For a male with 20 years of experience, the percent change in wage for 1 additional year of experience would be 42.763952 % , holding all other variables in the model constant.

The percent change in wage for females for every additional year of experience, holding all other variables in the model constant, is given by the following expression: $[(\beta_2 + \gamma_f) + 2\beta_3 * exp] * 100$ $(0.03756572 - 0.001386374 * exp) * 100$ This means that for a female with no experience, the percent change in wage for 1 additional year of experience would be 3.756572 % , holding all other variables in the model constant.
For a female with 5 years of experience, the percent change in wage for 1 additional year of experience would be 3.063385 % , holding all other variables in the model constant.
For a female with 10 years of experience, the percent change in wage for 1 additional year of experience would be 2.370198 % , holding all other variables in the model constant.
For a female with 20 years of experience, the percent change in wage for 1 additional year of experience would be 0.983824 % , holding all other variables in the model constant.

2.d. H_0: $\delta_f = \gamma_f = 0$ and H_A: at least one does not equal 0 OR H_0: $R\beta = 0$ and H_A: $R\beta \neq 0$ where R is a matrix consisting of 2 rows and 17 columns, the first row of all 0s except the 11th entry being 1 and the second row of all 0s except the 12th entry being 1; β is a vector of all the coefficients; and r is a vector with 2 entries of 0

```
## Linear hypothesis test
##
## Hypothesis:
## female = 0
## expfe = 0
##
## Model 1: restricted model
## Model 2: lnWage ~ hs + scoll + as + bs + ms + pd + phd + exp + expsq +
##     female + expfe + unioncov + pension + hipaids + hipaida +
##     nonwhite
```

```
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1    689 213.32                                  
## 2    687 208.03  2    5.2894 8.7338 0.0001796 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test has a p-value of 0.0001796 which is statistically significant at an alpha level of $\alpha = 0.05$ so we reject the null hypothesis in favor of the alternative that the female variable has some effect on wage. This is indeed consistent with the conclusions from part a where we decided that these variables (in this case fe and expfe) might not appear statistically significant on their own because of multicollinearity

2.e.

There is a 13.2397062% increase in wage when someone has a high school degree compared to someone with less than high school education, holding all other variables in the model constant.

There is a 56.7703609% increase in wage when someone has a bachelor's degree compared to someone with less than high school education, holding all other variables in the model constant.

There is a 61.990959% increase in wage when someone has a masters degree compared to someone with less than high school education, holding all other variables in the model constant.

2.f. The percent change of wage for being female compared to being male, holding all other variables in the model constant, is given by the following expression: $[e^{0.0172433-0.0079710*exp} - 1] * 100$

This means that for a woman with no experience, her wage is 1.7392824 % higher than a male with no experience, holding all other variables within the model constant.

For a woman with 5 years of experience, her wage is 2.2357972 % lower than a male with 5 years of experience, holding all other variables within the model constant.

For a woman with 10 years of experience, her wage is 6.0555654 % lower than a male with 10 years of experience, holding all other variables within the model constant.

For a woman with 20 years of experience, her wage is 13.2532039 % lower than a male with 20 years of experience, holding all other variables within the model constant.