Economics 104 – Problem Set 2 – 2018 Spring

NAMES: Ben Goodman, Ronaldo, Dillon Heffernan
RECITATION #: 204, 203
TA'S NAME: Peng Shao, Ruizhi Ma

Instructions:

- Maximum 15 pages (not including the appendix) with 1.5 line spacing.
- Minimum Font Size: MS Word (12) and Latex (10) (For typed submissions)
- Append your code.
- Use tables to format your results.
- Clearly mark all graphs and tables.

1. **Inflation Adjustment and Outliers**

   To adjust for inflation across time, all wage data has been normalized to 1999 dollars using CPI ratios on the PUMS website.

   Outliers were removed from all three sets of CPS data based on three criteria to eliminate entries which are impossible due to data recording and/or misreporting errors. Namely, observations were dropped if their hourly income was reported below the minimum hourly wage for tipped-professions of \$2.13, or if their recorded education or experience were strictly less than 0. Hourly wages below the tipped-profession and sub-zero experience and education are impossible for legally employed workers. Removing these outliers cuts the size of the 1995, 2004 and 2012 CPS data set by 3, 15, and 6 observations, respectively.

2. **Model Construction**

   a) ***Parameter Selection***

      i. **Guiding Principles**

         A. **Economic Intuition.** Common sense dictates that education, experience, gender, race and union status will have a significant impact on wages. That does not mean that education is a better predictor of wages than $education^2$, for example. But all serious models should include each of the above variables, in some form.

         B. **Simpler is better.** *Ceteris paribus*, a model with fewer parameters is better than a model with more, due to overfitting concerns. Furthermore, because of the size of the datasets we should be careful about dimensionality issues. For example, the 1995 data set has only approximately 200 nonwhite people and could be vulnerable to sampling error within further subsets of this subgroup. As such, "basic" dummy variables are preferred to their cross products with subsets consisting of low sample sizes.

         C. **Model Selection Tools (SIC):** We guide our model selection, particularly when determining the inclusion of Taylor Series Expansions terms using the SIC statistic. Cases where our previous goals came into conflict with the SIC prompted us to double check our code or data, usually resolving the issue. When in doubt, we chose to follow the criteria set out by Krass and Wasserman (1995)[1], which indicates a SIC difference of less than 2 was negligible, with the strength of evidence against the second model increasing the higher the difference goes.

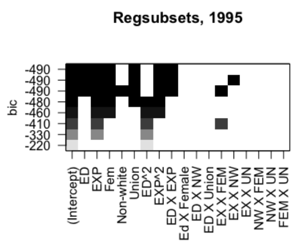
      ii. **Tools** As stated above our model selection tool was the Schwartz Information Criterion, also known as the Bayesian Information Criterion, though we also looked at the t-stats

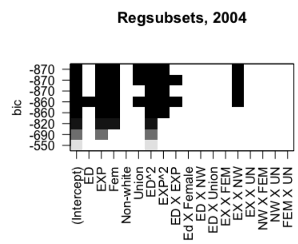
[1] Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association, 90(431), 928-934.

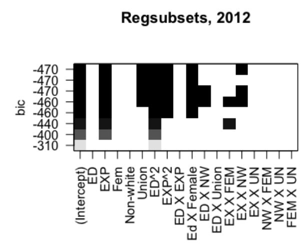
of individual coefficients. For each data set we began with what we referred to as a "kitchen sink" model that included a first order Taylor Expansion of Education and Experience, all of the dummy variables and their cross products with each other as well as education and experience. We fed those models into the regsubsets function from the leaps package, which systematically tests the BIC of all permutations of model specifications and returns the one with the lowest BIC. We then compared the model regsubsets suggested, the "kitchen sink" model, a naive regression on just an intercept, and a model with tweaks chosen based on economic intuition.

iii. **Results**

A. Regsubset plots







B. "Best" model included covariates according to regsubsets, tweaks marked in **bold**

95: Experience, education, ex^2, ed * ex, nonwhite, female, union

04: Experience, ex^2, ed^2, female, union, nonwhite *** experience**

12: Experience, ex^2, ed^2, union, female *** ed**, non-white x ex

C. BIC table

Year	Regsubsets	Kitchen Sink	Intercept	Tweaked Regsubsets
1995	1714.64	1769.15	2213.71	-
2004	2392.54	2459.29	3272.3	2393.72
2012	1617.635	1674.23	2093.54	1619.1

iv. **Economic Intuition and Tweaks**

A. 1995: No tweaks were made to the model suggested by regsubset. As expected, the linear terms for Education, Experience, Non-White, Female and Union are included. The quadratic Experience^2 suggests some non-linearity. The interaction term between education and experience is interesting - we will interpret this in a later section.

B. 2004: The Taylor expansion most efficiently capturing the DGP is now experience + ex^2 + ed^2. The main change to this model is the cross product of experience and race was found to be more significant than just race alone. However, when we tested exchanging race for race*ex the more complicated term only resulted in a small (-

1.18) improvement in BIC score. So sticking with our "simpler is better" principle, we chose to stick with Non-White. Thus, the change in the Taylor expansion was the only difference in this years model.

C. 2012: Regsubsets said that the same Taylor Expansion terms as the 2004 model were still the best. Female was removed in favor of a negative female * education term. This result suggests the gender wage gap is worse the more education a woman has. As this is counter intuitive, and results in a more complicated model, we chose to keep the original female term, resulting in a minimal increase in the BIC (+1.46).

   Experience * nonwhite was more interesting. Unlike the 2004 model swapping it for the basic nonwhite resulted in a significant (+5.5) degradation in the BIC. Including it at this point in the evolution of the model makes some intuitive sense if we consider two things. First, one would expect that racial discrimination in pay has decreased in 2012 from its levels in 1995 and 2004. Second, a workers current wage can be viewed as the result of a number of wage decisions over their career. Given these two facts one could see that a nonwhite worker that started their career in 1982 would have 95% of their wage decisions made at pre-2010 levels of discrimination, whereas someone who started in 2010 would have none of their wage decisions made at that level of discrimination. Thus, we get a negative experience * nonwhite term, workers in the labor force longer have accumulated more discrimination in their pay. It is worth mentioning that there is some empirical evidence to back this up, with Couch and Daly (2002) finding that little black/white wage convergence occurred before 1990, but during the 90s convergence was fastest for workers with less than 10 years experience[2]. A caveat, the study mentioned above was also conducted using CPS data, albeit over a much longer timer period (1970-1999.)

b) ***Bivariate Statistics*** Our selected models are further justified via a series of bivariate statistics. Without controlling for other factors, we see in bivariate statistics that for all three data sets, gender, race and union status have visible relationships with ln(wage), and that there are non-linear patterns present in the relationships between education and experience and ln(wage). Below are the figures which comprise the Bivariate analysis. Comments are contained in figure captions.

[2]Couch, K., & Daly, M. C. (2002). BlackWhite Wage Inequality in the 1990s: a Decade of Progress.Economic Inquiry,40(1), 31-41.

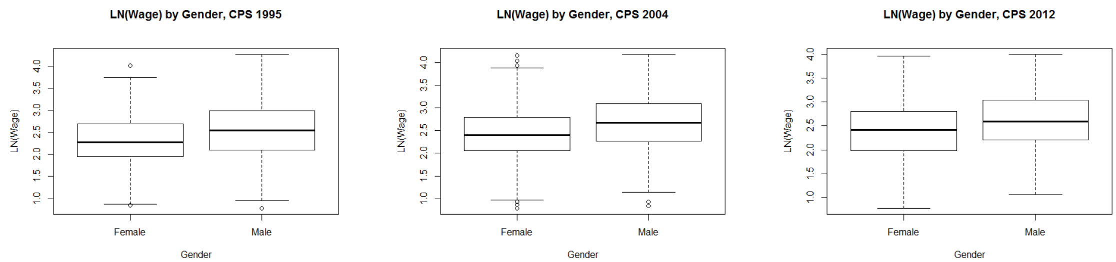


Figure 1: Ln(Wage) by Gender. Across all three data sets, the wage distribution for males is higher than the female distribution at all quartile levels.

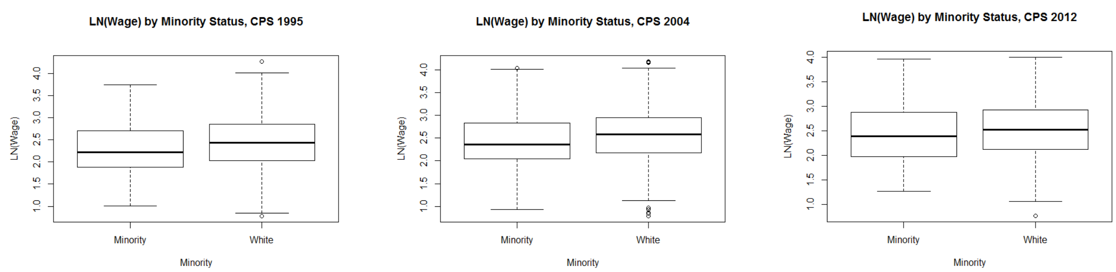


Figure 2: Ln(Wage) by Race. Across all three data sets, the mean white ln(wage) is higher than the mean nonwhite ln(wage).

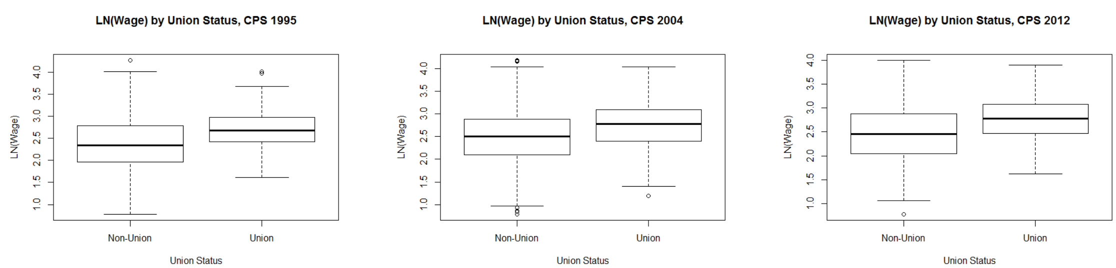


Figure 3: Ln(Wage) by Union Status. Across all three data sets, the minimum, 1st quartile, mean, and third quartile union member's ln(wage) is higher than the non-union member's wage at those same levels. Intuitively, we would expect to see higher averages for the union members, but smaller rightwards tails (ie: Union jobs receive a pay premium, but that premium is not enough to have unionized jobs such as teaching or factory work surpass well compensated professions in the upper quarter of the income distribution like doctors and lawyers.)

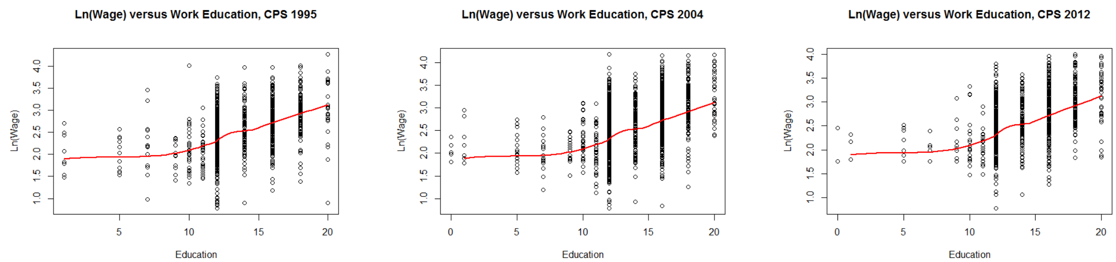


Figure 4: Ln(Wage) versus years of education. Across all three data sets we observe a non-linear relationship between the variables. The lines in the graphs are loess fitted smoothers.

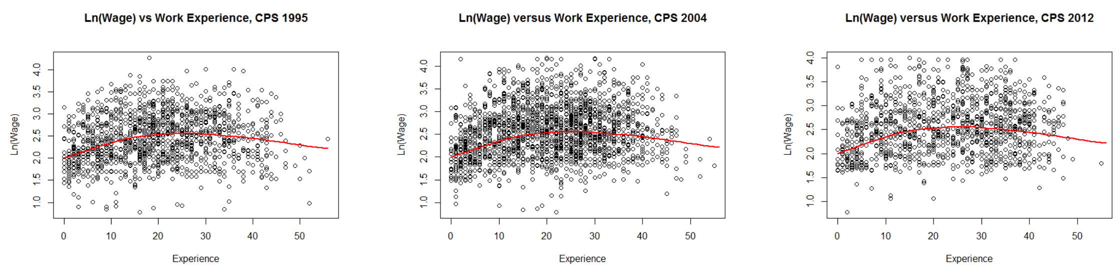


Figure 5: Ln(Wage) versus years of experience. Across all three data sets we observe a non-linear relationship between the variables. The lines in the graphs are loess fitted smoothers.

c) ***Response Normality***

We now turn to investigate whether the response variable, ln(Wage) is normally distributed for each years data set.

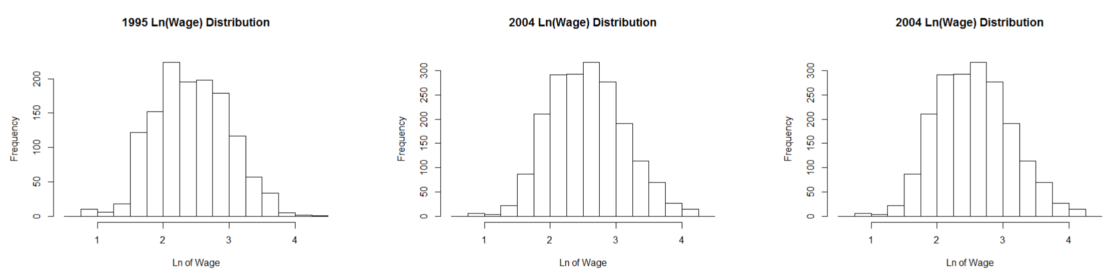


Figure 6: Histograms of response variable distribution. Distributions appear to be fairly normally distributed.

We now formally test for normality of residuals with the Jarque Bera test.

H_0: The residuals are normally distributed.

H_1: The residuals are not normally distributed. $\alpha = 0.01$.

Year	Skewness	Kurtosis	Jarque-Bera P-Value
1995	0.1090427	2.767	$= 0.0635$
2004	0.243	2.83	$< 2.2 * 10^{-16}$
2012	0.311	2.595	$< 2.2 * 10^{-16}$

We see that based on Jarque-Bera, we have very strong evidence to reject the null hypothesis for the 2004 and 2012 data sets, but do not have sufficient evidence to do so for the 1995 data set. Although the 2004 and 2012 data sets test for non-normality, their histograms both look fairly normal, and with a high number of observations we feel comfortable moving on with our analysis assuming normality.

d) ***Fitted Models***

Table 1: CPS 1995 Fitted Model

	Dependent variable:
	Ln(Wage) in 1999 Dollars
Education	0.123^{***} (SE = 0.009)
Experience	0.067^{***} (SE = 0.007)
Female	-0.236^{***} (SE = 0.025)
Non-White	-0.093^{***} (SE = 0.035)
Union	0.201^{***} (SE = 0.034)
Experience^2	-0.001^{***} (SE = 0.0001)
Education $*$ Experience	-0.002^{***} (SE = 0.0004)
Constant	0.464^{***} (SE = 0.131)
Observations	1,320
Adjusted R^2	0.337
Residual Std. Error	0.453 (df = 1312)
F Statistic	96.746^{***} (df = 7; 1312)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: 2004 CPS Fitted Model

	Ln(Wage) in 1999 Dollars
Experience	0.032*** (SE = 0.003)
Female	−0.239*** (SE = 0.020)
Non-White	−0.095*** (SE = 0.025)
Union	0.109*** (SE = 0.031)
Education^2	0.004*** (SE = 0.0001)
Experience^2	−0.0005*** (SE = 0.0001)
Constant	1.596*** (SE = 0.040)
Observations	1,925
Adjusted R^2	0.379
Residual Std. Error	0.444 (df = 1918)
F Statistic	196.927*** (df = 6; 1918)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: 2012 CPS Fitted Model

	Ln(Wage) in 1999 Dollars
Experience	0.032*** (SE = 0.004)
Female	−0.178*** (SE = 0.026)
Union	0.147*** (SE = 0.040)
Experience∗ Non-White	−0.004*** (SE = 0.001)
Education^2	0.004*** (SE = 0.0002)
Experience^2	−0.0005*** (SE = 0.0001)
Constant	1.501*** (SE = 0.053)
Observations	1,223
Adjusted R^2	0.342
Residual Std. Error	0.460 (df = 1216)
F Statistic	106.654*** (df = 6; 1216)
Note:	*p<0.1; **p<0.05; ***p<0.01

e) ***Heteroskedasticity***

I. *White's Test*

We performed a regression analysis of the squared residuals on a second order Taylor series expansion of all covariates for each year's fitted model.

H_0: The residuals for each year's model are homoskedastic.

H_1: The residuals for each year's model are heteroskedastic.

We will reject at the 0.01 α level.

Running White's test in R, for each data set we get a p value that is less than $2 * 10^{-7}$, and so for each data set we have very strong evidence to reject the null hypothesis and adopt the H_1. If the residuals are not homoskedastic, then by definition they are heteroskedastic. We thus deal with heteroskedasticity below.

II. *White's Robust Standard Errors*

Given that our models were found to be heteroskedastic, we had to adjust the standard errors of our co-efficients by using White's Robust Standard Errors. Since the standard errors are used in the calculation of the t-stats, we also had to adjust the t-stats of our co-efficients. The adjusted t-stats are shown in tables 4,5 and 6 below.

Table 4: Robust T-Stats CPS 1995 Model

(Intercept)	Education	Experience	Female	Nonwhite	Union	Experience^2	Education$*$Experience
3.578	12.559	9.694	-9.332	-2.741	6.415	-8.320	-4.659

The absolute value of all our t-stats are greater than 2.58 which is the critical value of a standard normal distribution for a 99 percent hypothesis test. This means that for each co-efficient, with 99 percent confidence, we can reject the null hypothesis that it is equal to zero.

Table 5: Robust T-Stats CPS 2004 Model

(Intercept)	Experience	Female	Nonwhite	Union	Education^2	Experience^2
42.606	11.016	-11.716	-4.007	3.931	25.859	-7.285

Similar to those of 1995, the absolute value of all our t-stats are greater than 2.58. Hence, for each co-efficient, with 99 percent confidence, we can reject the null hypothesis that it is equal to zero.

Table 6: Robust T-Stats CPS 2012 Model

(Intercept)	Experience	Female	Union	Experience∗Nonwhite	Education^2	Experience^2
31.243	8.600	-6.766	3.974	-2.764	20.272	-5.715

Similar to those of 1995 and 2004, the absolute value of all our t-stats are greater than 2.58. Hence, for each co-efficient, with 99 percent confidence, we can reject the null hypothesis that it is equal to zero.

III. *Estimating Standard Errors for Prediction*

We use our regression calculated as part of White's test to estimate standard errors for prediction for a given observation. These estimates are used to re-scale the residuals to analyze residual non-normality and to calculate density forecasts in latter sections.

f) ***Non-Normality of Residuals After Adjusting for Heteroskedasticity***

To inspect normality of residuals, we re-scale each residual for each observation t by the following formula: $\frac{e_t}{\sqrt{\hat{e_t^2}}}$, where $\sqrt{\hat{e_t^2}}$ is the square root of the estimation for observation t provided via the regression made via White's test. The distribution of each year's rescaled residuals are as follows:

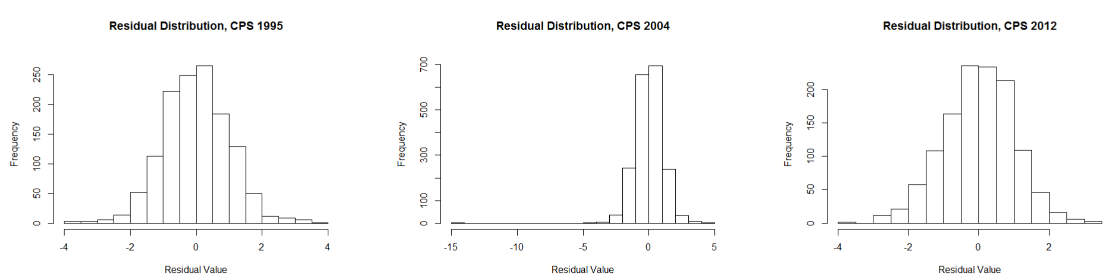


Figure 7: Histograms of residuals for each year's models. There is one massive outlier in the CPS 2004 data set. All three years appear to have fairly thick tails.

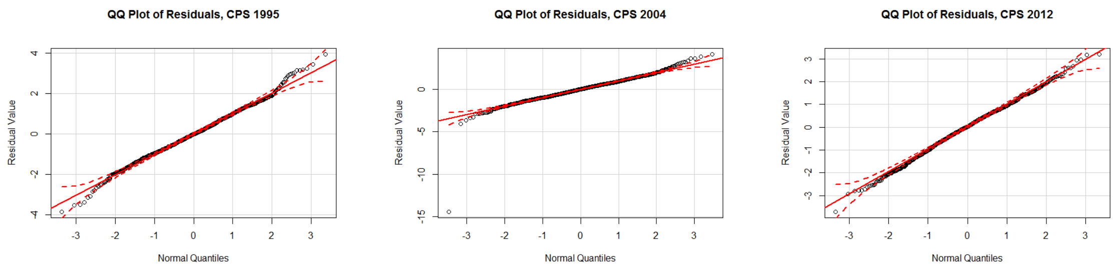


Figure 8: QQ plots of residuals for each year's models. The QQ plots for 1995 and 2004 show much more lack-of-normality in their tales than the QQ plot for 2012.

We now formally test for normality of residuals with the Jarque Bera test.

H_0: The residuals are normally distributed.

H_1: The residuals are not normally distributed. $\alpha = 0.01$.

Year	Skewness	Kurtosis	Jarque-Bera P-Value
1995	0.0526	6.604	$< 2.2 * 10^{-16}$
2004	-1.277	24.40	$< 2.2 * 10^{-16}$
2012	-0.131	6.122	< 0.0965

We see that based on Jarque-Bera, we have very strong evidence to reject the null hypothesis for the 1995 and 2004 data sets, but do not have sufficient evidence to do so for the 2012 data set. Although the 1995 and 2004 data sets exhibit non-normality, dealing with it in addition to the heteroskedasticity that we have already dealt with is outside the scope of this class.

3. **CPS 1995 Model Fitting, Interpretation and Analysis**

Table 7: Exponentiated Coefficients for 1995 Fitted Model

(Intercept)	Education	Experience	Female	Non-White	Union	Experience ^{2}	Education $*$ Experience
1.590	1.131	1.069	0.790	0.911	1.222	0.999	0.998

Interpretation of Exponentiated Intercept Co-Efficients:

Our 'base case' intercept has a value of 1.590. This means that a male who is white, a union member, has no education and has no experience is expected to earn an hourly wage of $1.59. However, this intercept would change based on whether a person is Female, Non-White or a Union member. Our 0.79 Female coefficient means that, *ceteris paribus*, a female worker is expected to have an hourly wage that is on average 21 percent lower than that of a male worker. Our Non-White coefficient of 0.911 means that, *ceteris paribus*, a non-white worker is expected to have an hourly wage that is on average 8.9 percent lower than that of a white worker. Our 1.222 Union coefficient means that, holding all other traits constant, a union-member is expected to have an hourly wage that is 22.2 percent higher than that of a non-union member.

Interpretation of Exponentiated Slope Co-Efficients:
Our 1.131 education coefficient means that our model predicts the marginal effect of an extra year of education to be a 13 percent increase in a person's hourly wage. Our 1.069 coefficient for experience means that for a person with zero years of experience, our model predicts the marginal effect of an extra year of experience to be a 6.9 percent increase in a person's hourly wage. However, we must pay attention to the fact that our exponentiated coefficient for Experience ^{2} is less than 1. This means that for each additional year of experience, the marginal effect of an extra year of experience falls. Finally, our co-efficient of 0.998 for Education$*$Experience indicates that as a person gathers more experience, the marginal of effect of gaining an extra year of education falls by 0.2 percent. Likewise, for each additional year of education that a person has, the marginal effect on hourly wage of gaining an extra year of experience would fall by 0.2 percent. This can be explained by the fact that a percentage increase in a high wage is more costly to an employer than a percentage increase in a low wage. Therefore, given that more educated and experienced people are expected to have higher wages, it would become more costly to the employer to increase their wage by an additional percent.

4. **CPS 2004 Model Fitting, Interpretation and Analysis**

Table 8: Exponentiated Coefficients for 2004 Fitted Model

(Intercept)	Experience	Female	Non-White	Union	Education^2	Experience^2
4.932	1.033	0.788	0.909	1.115	1.004	0.9995

The coefficients of the 2004 model are broadly similar to the 1995 model. With the intercept we once again have the expected wage of a non-union white male with no experience or education, this time at \$4.93. The female, non-white and union and dummy variables act as scalars on this intercept. For example a white female non-union worker is estimated to earn on average .788 times what her male co-worker–holding constant all other traits–earns. Interestingly, this is very close to the often cited statistic that female workers make 80 cents on the dollar. Both the race penalty and the union bonus are weaker than they were in 1995. This broadly matches what we would expect, as it is documented that unions weakened and racial equity increased between 1995 and 2004. The experience and education terms are Taylor series approximations of the non-linear function that determines their impact on wages. Since different terms are used it is difficult to do a direct comparison.

5. **CPS 2012 Model Fitting, Interpretation and Analysis**

Table 9: Exponentiated Coefficients for 2012 Fitted Model

(Intercept)	Experience	Female	Union	Experience $*$ Non-White	Education^2	Experience^2
4.484	1.033	0.837	1.158	0.996	1.004	0.9995

Interpretation of Intercept Co-Efficients - our 'base case' intercept has a value of 4.484 This means that a male who is white, a union member, has no education and has no experience is expected to earn an hourly wage of $1.59. However, this intercept would change based on whether a person is Female, Non-White or a member of a Union. Our 0.837 Female coefficient means that, ceteris paribus, a female worker is expected to have an hourly wage that is 16.3 percent lower than that of a male worker. Our 1.158 Union coefficient suggests that a union-member is expected to have an hourly wage that is 15.8 percent higher than that of a non-union member.

Interpretation of Slope Co-Efficients - Our 1.004 Education^2 coefficient means that for a person with zero years of education, the marginal effect of gaining an extra year of education would be a 0.4 percent increase in their hourly wage. However, the quadratic nature of the term means that this marginal effect changes depending on the level of education that one has. Furthermore, because our exponentiated co-efficient is greater than 1, the marginal effect of an extra year of education increases as education increases. Our 1.033 coefficient for experience means that for a person with zero years of experience, our model predicts the marginal effect of an extra year of experience to be a 3.3 percent increase in a person's hourly wage. However, we must pay attention to the fact that our exponentiated coefficient for Experience^2 is less than 1. This means that for each additional year of experience, the marginal effect of an extra year of experience falls.

We must also pay attention to the co-efficient of 0.996 for our Experience$*$Non-White. This means that the marginal effect of an additional year of education is 0.4 percent lower for a non-white person compared to that of a white person.

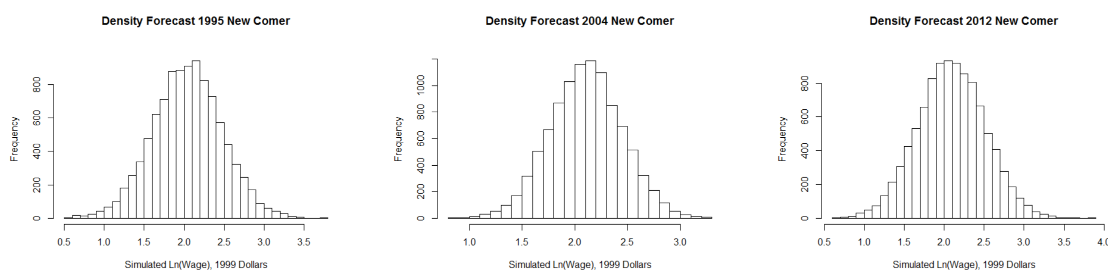
6. **Prediction for ln(Wage) of a White, Female, Unioninized worker with 3 years of Experience and 12 Years of Education.**

   (a) 1995: The point prediction of ln(Hourly Wage) for such a new observation is 2.036, which when exponentiated is \$7.66. A 95% confidence interval encapsulates ln(Wages) between 1.185 and 2.887, or \$3.27 and \$17.94.

   (b) 2004: The point prediction of ln(Hourly Wage) for such a new observation is 2.108, which when exponentiated is \$8.23. A 95% confidence interval encapsulates ln(Wages) between 1.442 and 2.773, or \$4.22 and \$16.01.

   (c) 2012: The point prediction of ln(Hourly Wage) for such a new observation is 2.091, which when exponentiated is \$8.10. A 95% confidence interval encapsulates ln(Wages) between 1.265 and 2.918, or \$3.54 and \$18.50.

   (d) Density Predictions



7. **Conclusion and Across Year Discussion** The exponentiated Female co-efficient is less than one for all three years. Once again, this is quite understandable given the history of gender-based wage discrimination. It is important to note, however, that the 1995 and 2004 co-efficients are quite similar whereas the 2012 co-efficient is higher. This suggests that between 2004 and 2012, advances have been made in reducing the level of gender based wage discrimination in the workplace.

   The exponentiated Union co-efficient is greater than 1 for all three years. This is quite understandable given that union workers, through collective bargaining, have more power to negotiate for better average wages. For all three years, the exponentiated co-efficient for experience is greater than 1, which suggests a positive relationship between experience and wage. Of course, this is quite understandable because employers tend to place a higher value on more experienced workers. For all three years, the exponentiated coefficient for Experience^2 is less than one which means that the marginal effect of experience decreases as one's experience increases. This observed dynamic within the experience terms directly ties into the economic phenomenon of employees trading off more experience for falling behind the technological frontier as they age. Initially, Experience accumulates faster than one falls behind technological advancement. At some point this tradeoff is equal, and average worker productivity at that age is maximized. Then technological lag overtakes the benefits of experience, and worker productivity begins to fall. This economic story is nicely represented in the models coefficients. $Education^2$ is included in both our 2004 and 2012 model. Both co-efficients are greater one which suggest a positive relationship between education and wages. Of course, this relationship is trivial to explain, because, ceteris paribus, employers tend to value workers who have more education, ie. have higher levels of human capital. It is also important to note the positive coefficient of this quadratic term, which suggests that as a person gains more education, the marginal effect of an extra year of education increases. A possible explanation for this is that highly educated individuals have a great opportunity to access jobs that are on the extreme positive end of the wage spectrum. Unlike the 2004 and 2012 models, the 1995 model has a linear term for education. It's coefficient is greater than 1 which suggests a positive relationship between education and wage. As in the previous paragraph, the explanation for this relationship is trivial. The 1995 model includes an interaction term between Experience and Education. Its meaning was explained previously in the interpretation of the 1995 model.

As discussed above, we find that across all years white workers, holding all other factors equal, on average are compensated more than non-white workers. One difference in the models is that the term is simply Non-White in the 1995 and 2004 models, whereas it is Non-White*Experience in the 2012 model. As discussed in previous sections, this is a subtle difference. The difference is that race-based wage discrimination is modeled as being universal in 1995 and 2004, whereas in 2012, the racial wage bias is modeled more as an effect due to compounded wage discrimination in previous years. This difference in modeling is encouraging for racial wage-discrimination in the US.

As all the data is standardized to 1999 dollars we can make direct comparisons between wage predictions across years. We see a clear increase in the point prediction from 1995 to 2004, as well as a small decrease from 2004 to 2012. We could interpret this as being broadly in line with economic trends. The boom years of the 90s leading to higher wages only to stagnate or fall during the Great Recession. This trend is particularly relevant for a union worker with only a high school education, a demographic that has been hit particularly hard during the last financial downturn. It is also worth noting that the range of the 1995 and 2012 interval predictions are roughly similar, but the 2004 range is approximately 25% lower. We were unable to determine a satisfying economic interpretation for why the range would contract only to expand again. It may very well be statistical noise as the subsets of white union women is quite small across the 95, 04 and 12 data sets, with 62, 78 and 44 observations observed, respectively.