Problem Set 2

Ajjit Narayanan

I. Introduction

The purpose of this paper is to build a multiple regression model to explain and forecast wages for people based on years of education, experience, and demographic characteristics. We use CPS wage data from three points in time, 1995, 2004, and 2012. To begin, we focus in on the 1995 data to help motivate our model, and then replicate the step for the other years.

II. Data Cleaning

To start our exploratory data analysis, we look at a short summary of all our independent and dependent variables from the 1995 dataset.

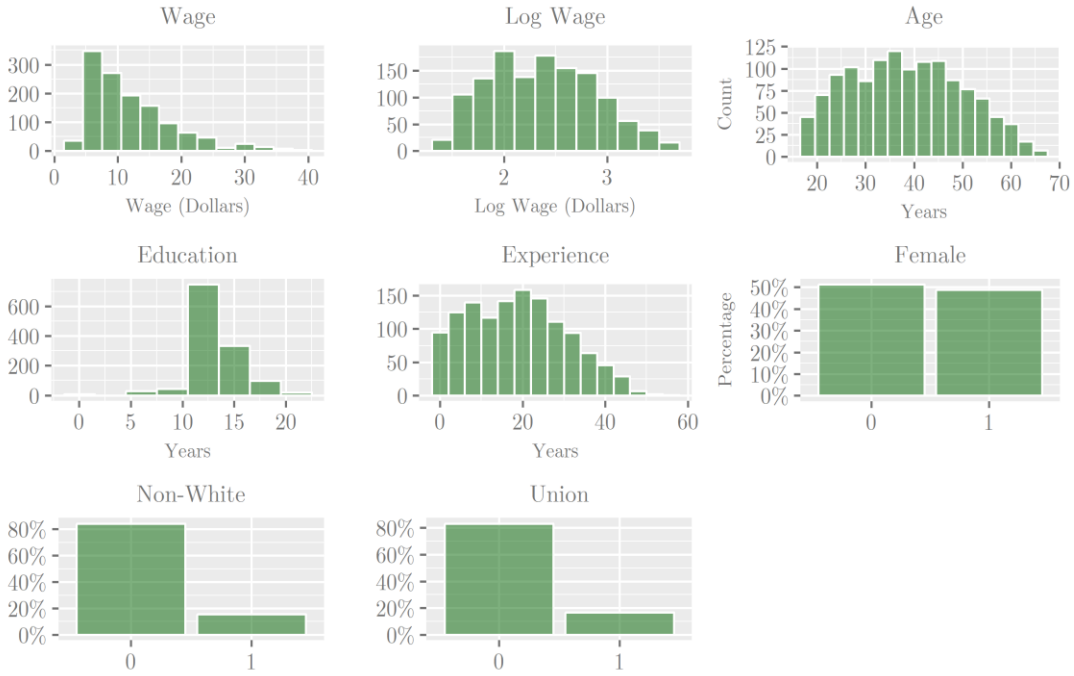
FEMALE	NONWHITE	UNION	EDUC	AGE	EXPER	WAGE	LNWAGE
0:655	0:1078	0:1064	Min. : 1.00	Min. :18.0	Min. : 0.00	Min. : 4.25	Min. :1.447
1:623	1: 200	1: 214	1st Qu.:12.00	1st Qu.:29.0	1st Qu.: 9.00	1st Qu.: 7.00	1st Qu.:1.946
NA	NA	NA	Median :12.00	Median :38.0	Median :19.00	Median :10.29	Median :2.331
NA	NA	NA	Mean :13.13	Mean :38.4	Mean :19.27	Mean :12.21	Mean :2.368
NA	NA	NA	3rd Qu.:16.00	3rd Qu.:47.0	3rd Qu.:27.00	3rd Qu.:15.38	3rd Qu.:2.733
NA	NA	NA	Max. :20.00	Max. :65.0	Max. :56.00	Max. :38.46	Max. :3.650

Some curious things to note are that there seems to be one person with negative years of experience and there are 46 people who make under the 1995 minimum wage of $4.25. We decide to throw out the one person with negative years of experience because this is probably a data entry error and so the other variables for that person could be tainted. As for the 36 people who seem to have wages under the minimum wage, these are either data entry errors or these people are actually being paid under the table below the minimum wage. If it were the latter case, it is likely that this population faces different wage dynamics than the rest of the population and could skew our results, so we throw out these observations as well.

Finally, we also throw out 8 outliers with very high wages, which we define as above $40 an hour. There are a couple of reasons for this:

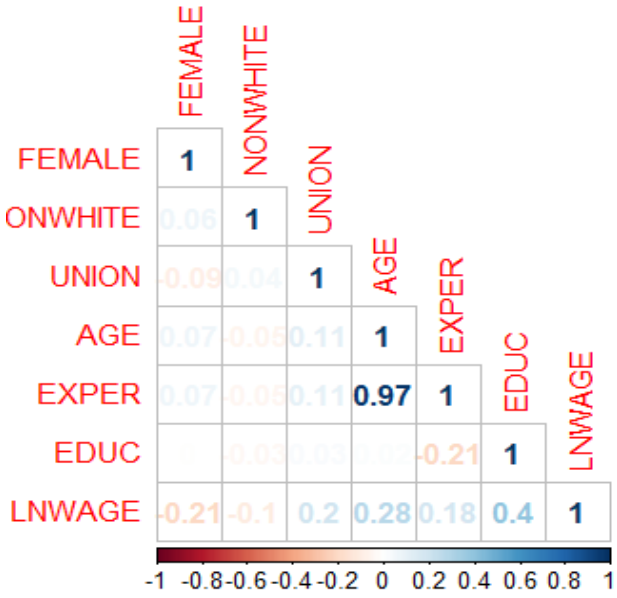
1. The wage dynamics for people with very high incomes are probably different from the general population, so their regression coefficients are probably different
2. These outliers make up a very small percentage of our data and throwing them out will likely improve model fit for the middle-waged workers.

After the data cleaning, we now have 1278 observations in our data to use for modeling. We now look at the distribution of all our dependent and independent variables.



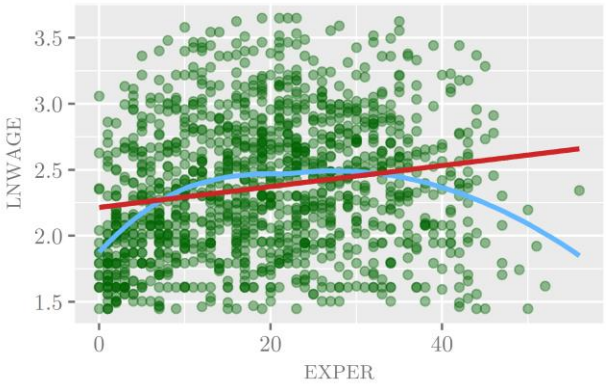
We will use Log Wage as our dependent variable as it pulls in the long right tail of wage and conforms better to a normal distribution. Age and experience predictably have a very similar distribution and probably have very high collinearity, so we should think about dropping one of them. It also seems as if their distributions are left censored. The other continuous variable, Education, is not normally distributed as there is a large spike at 12, which is equivalent to a high school education. However, no transformations will make this better, so we leave it

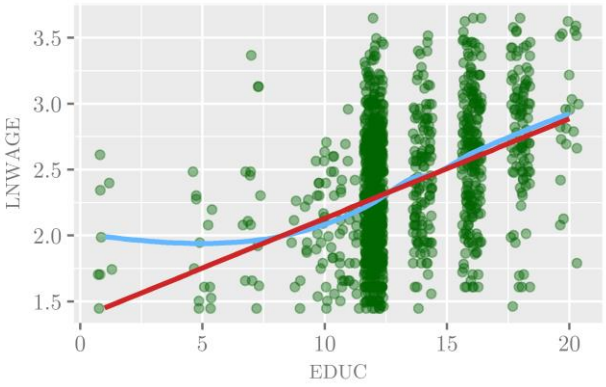
untransformed. To further assess collinearity between our variables, we look at a correlation matrix between all predictors and dependent variables below.



As we expected, the only variables to have troublingly high or low correlations are age and experience, meaning we shouldn't have both in our model. We choose to keep experience as we believe that is more predictive of salaries than the raw age of individuals. We also see later that the regression results are identical regardless of which variable we choose to keep. Otherwise collinearity is not a problem.

In order to assess the assumption of linearity between the dependent and independent variables, we plot the scatterplots between our continuous independent variables against our dependent variable. We also overlay the OLS line between the 2 variables in red and the LOESS line in blue. The LOESS line is the locally weighted scatterplot smoothing line that helps us identify non-linear relationships.





So, there is a mostly linear relationship and perhaps a quadratic relationship between our continuous predictors and the dependent variable. In particular, the Experience variable seems to have a slightly negative quadratic relationship.

III. Model Selection

Now we move on to model selection. We run multiple regressions including and excluding various regressors and nonlinear terms, the results of which are shown below. Specifically, we report the R-squared, adjusted R-squared, the SIC, and the 5-fold RMSE as a form of cross validation for our models. The 5-fold RMSE is calculated by dividing our data into 5 folds, using 4 of them for parameter estimation and using those parameters to estimate the root mean squared error for the excluded fifth fold.

1995 Regression Models

R^2	adj.R^2	sigma	f.statistic	p.value	df	SIC	RMSE	predictors
0.2749596	0.2726814	0.43730	120.6911	0	5	1550.559	0.4393	Exper, Educ, ExperxEduc, Exper^2
0.2768267	0.273984	0.43690	97.38288	0	6	1554.417	0.4392	Exper, Educ, ExperxEduc, Exper^2, Educ^2
0.3045249	0.3017911	0.42846	111.3931	0	6	1504.506	0.4293	Female, Nonwhite, Union, Exper, Educ
0.3461307	0.3425267	0.41577	96.0406	0	8	1439.976	0.4174	Female, Nonwhite, Union, Exper, Educ, Exper^2, Exper*Educ
0.3479251	0.3422594	0.41585	61.40861	0	12	1465.076	0.4177	Female, Nonwhite, Union, Exper, Educ, Exper^2, Educ^2, ExperxEduc, NonwhitexUnion, FemalexUnion, FemalexNonwhite
0.3109421	0.3060513	0.42715	63.57709	0	10	1521.272	0.4284	Female, Nonwhite, Union, Exper, Educ, ExperxEduc, NonwhitexUnion, FemalexUnion, FemalexNonwhite

The above table indicates that when examining model metrics that adjust for degrees of freedom, such as the adjusted R^2 and the SIC, the model that comes out far on top is the highlighted one with the following predictors: Experience, Education, Experience^2, Experience*Education, and binary variables for Female, Union, and Nonwhite. A further robustness check is that this model has the lowest 5-fold RMSE, suggesting that it is also the best for out of sample prediction.

One last idea to improve model fit is to split up the data into females and males and run separate regressions on them. This would imply that different labor market dynamics are affecting men and women, which is entirely possible. To provide some intuition for this, the below histograms show log wage and experience broken down by gender.

Furthermore, the median years of experience is 20 for women and 18 for men. At the same time, men also have higher median wages than women. In short, women make lower wages but have higher levels of education, so the betas for the two groups might be different. Here are the results of the gender separated models.

Female

term	estimate	std.error	statistic	p.value	SIC	adj.R2	RMSE
(Intercept)	0.2899634	0.1848798	1.568389	0.1173038	671.97	0.2736	0.416
NONWHITE	-0.0530330	0.0421676	-1.257672	0.2089868			
UNION	0.1584424	0.0479348	3.305374	0.0010036			
EXPER	0.0512331	0.0097571	5.250833	0.0000002			
EDUC	0.1256478	0.0128880	9.749230	0.0000000			
EXPER2	-0.0004729	0.0001158	-4.084744	0.0000500			
EXPER:EDUC	-0.0018078	0.0005297	-3.413013	0.0006846			

Males

term	estimate	std.error	statistic	p.value	SIC	adj.R2	RMSE
(Intercept)	0.3636738	0.1663688	2.185949	0.0291764	784.02	0.3577	0.431
NONWHITE	-0.1435955	0.0489087	-2.935990	0.0034427			
UNION	0.1691728	0.0424503	3.985202	0.0000751			
EXPER	0.0788434	0.0089107	8.848164	0.0000000			
EDUC	0.1197343	0.0116270	10.297914	0.0000000			
EXPER2	-0.0008803	0.0001174	-7.498667	0.0000000			
EXPER:EDUC	-0.0021786	0.0004713	-4.622123	0.0000046			

Under these models, the intercept is no longer significant for men and women. For women, the nonwhite binary variable is also no longer significant at the 5 percent significance level. The SIC is no longer comparable to past models as we are working with a subset of the data now. However, we can compare the adjusted R^2 and the cross-validation RMSE scores. The RMSE scores are marginally larger for men and women. The adjusted R^2 is slightly larger for men, going up from 0.3425267 to 0.357664. However, the adjusted R^2 drops for women to 0.2736124. Given that in the end, we are trying to come up with point, interval and density prediction for a woman, this does not bode well. This lower adjusted R^2 and RMSE combined with the insignificant coefficients causes us to discard these split models and go back to our initial model. In order to further assess our chosen model, we take a look at the model coefficients and t-statistics.

1995 Coefficients

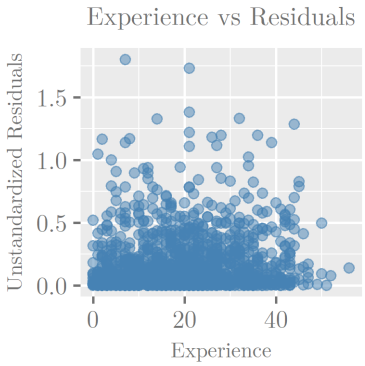
term	estimate	std.error	t.statistic	p.value	SIC	adj.R2	RMSE
(Intercept)	0.4301458	0.1237500	3.475927	0.0005263	1439.98	0.3425	0.417
FEMALE	-0.2153243	0.0235059	-9.160431	0.0000000			
NONWHITE	-0.1004658	0.0322150	-3.118605	0.0018580			
UNION	0.1752847	0.0316835	5.532368	0.0000000			
EXPER	0.0666757	0.0065793	10.134090	0.0000000			
EDUC	0.1223391	0.0086313	14.173853	0.0000000			
EXPER2	-0.0007006	0.0000826	-8.483641	0.0000000			
EXPER:EDUC	-0.0020467	0.0003520	-5.813723	0.0000000			

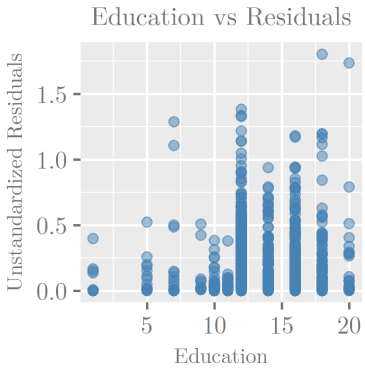
So, all the coefficients and the intercept have highly significant t-statistics at the 5% confidence level and are thus statistically significant. This also means our model doesn't have any extraneous variables and is parsimonious.

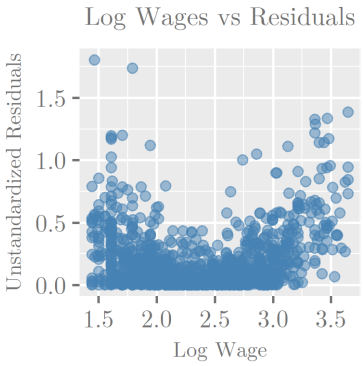
IV. Model Assumption Checks

We have already evaluated and corrected for non-linearity by using squared and interaction terms in our initial model. Now we evaluate whether the assumption of homoscedasticity has

been met in our model. First, we look at the graphs of some variables against the raw squared residuals as well as the results of a BGP test.



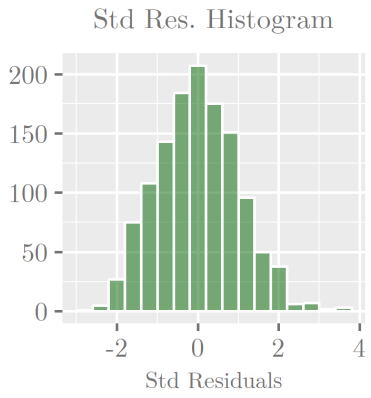


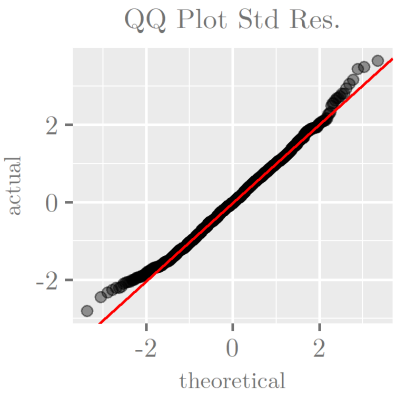


bgp_statistic	p.value	parameter
54.83334	0	7

So, there is clear heteroscedasticity in our residuals as graphically the variance of residuals are much higher at more extreme values of log wage and the values of the independent variables. The p-value of 0 from the BGP test confirms this.

The next thing to do would be to assess the normality of our residuals. However, the heteroscedasiticity means the disturbances are likely drawn from a different distribution at each observation. So, we use the estimated squared residuals/variances from the BGP regression to standardize our residuals. We put in each of our data points into the BGP regression to obtain a sample variance which we then take the square root of to obtain the sample standard deviations. We then divide the residuals by the standard deviation of the residual to obtain the standardized residual. We can then examine whether or not the standardized residuals are normal. Below is a QQ-plot, a residual histogram and a Jarque-Bera-Test to assess the normality of these standardized residuals.



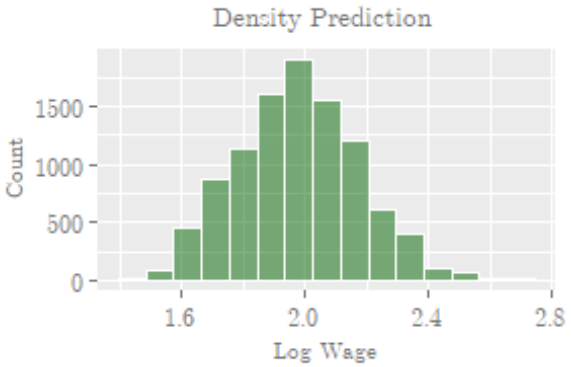


JB.test.stat	p.value	df	method
9.771	0.008	2	Jarque Bera Test

So while the histograms looks normal and symmetric, the QQ-plot shows us that the tails are fatter than the normal distribution on the left and thinner than the normal distribution on the right. And finally, the JB test has a p-value of essentially 0. Thus we reject the null of homoscedasticity.

V. Prediction

The heteroscedasticity doesn't affect the point estimate of the Log Wage for the new person. However, because of the non-normality and heteroscedasticity of the standardized residuals, we will use a simulation algorithm to obtain the interval and density estimates for the new person. We take 10,000 draws from the standardized residual distribution, multiply it by the standard deviation of the new person and then add the point prediction of that new person to each draw to obtain a density. So here are the point, interval (95%) and density estimates for the new person



Point estimate	X2.5.	X97.5.
1.978217	1.627544	2.359008

VI. Repeating for 2004 and 2012

Now we move on to the 2004 and 2012 data. For those years we employ the same data cleaning methods. In the 2012 data we remove 2 observations with negative experience, 15 observation with wages higher than 60 dollars an hour, and 48 observations that were making lower than the 2004 minimum wage of 5.15. For the 2012 data we remove 2 people with negative experience, 14 people with wages higher than 63 dollars an hour. and 26 people who make less than the minimum wage of 7.25 an hour. To clarify, the reason we use 60 dollars and 63 dollars respectively as the outlier cutoff for wages is becuase this is approximately the inflation adjusted value of 40 dollars in 1975, which was our original cutoff value. This data pruning leaves us with 1878 observations in 2012 and 1185 observations in 2012.

The plots of the distributions of the variables and the pairwise scatterplots indicate approximately the same relationship as in the 1995 data, so we won't show them again and instead proceed directly to model selection and evaluation. Below are the summary statistics of a few different regression models on the 2004 and 2012 data with the lowest SIC model highlighted

2004 Regression Models

r.squared	adj.r.squared	sigma	f.statistic	p.value	df	SIC	RMSE	predictors
0.3201046	0.3186526	0.4357373	220.4589	0	5	2249.586	0.435788	Exper, Educ, ExperxEduc, Exper^2
0.3897162	0.387104	0.41327	149.1887	0	9	2076.886	0.4140421	Exper, Educ, ExperxEduc, Exper^2, Educ^2, Female, NonWhite, Union
0.3590593	0.3573474	0.4231834	209.7414	0	6	2146.318	0.4237155	Female, Nonwhite, Union, Exper, Educ
0.3850516	0.3827497	0.4147354	167.2722	0	8	2083.647	0.4154007	Female, Nonwhite, Union, Exper, Educ, Exper^2, Exper*Educ
0.3904649	0.3868717	0.4133482	108.6682	0	12	2097.194	0.414421	Female, Nonwhite, Union, Exper, Educ, Exper^2, Educ^2 ExperxEduc, NonwhitexUnion, FemalexUnion
0.3642995	0.3612367	0.4219009	118.9434	0	10	2161.053	0.4226661	Female, Nonwhite, Union, Exper, Educ, ExperxEduc, NonwhitexUnion, FemalexUnion, FemalexNonwhite

2012 Regression Models

r.squared	adj.r.squared	sigma	f.statistic	p.value	df	SIC	RMSE	predictors
0.304778	0.3024214	0.442169	129.3249	0	5	1466.269	0.4428276	Exper, Educ, ExperxEduc, Exper^2
0.3465913	0.3421463	0.4293943	77.97405	0	9	1421.075	0.4311569	Exper, Educ, ExperxEduc, Exper^2, Educ^2, Female, NonWhite, Union
0.3192108	0.3163236	0.4377407	110.5627	0	6	1448.487	0.4381888	Female, Nonwhite, Union, Exper, Educ
0.3431401	0.3392335	0.4303439	87.83693	0	8	1420.24	0.4310615	Female, Nonwhite, Union, Exper, Educ, Exper^2, Exper*Educ
0.3486945	0.3425868	0.4292506	57.09075	0	12	1438.487	0.4313401	Female, Nonwhite, Union, Exper, Educ, Exper^2, Educ^2 ExperxEduc, NonwhitexUnion, FemalexUnion
0.3249225	0.3197516	0.4366419	62.83786	0	10	1466.813	0.4381655	Female, Nonwhite, Union, Exper, Educ, ExperxEduc, NonwhitexUnion, , FemalexUnion, FemalexNonwhite

So for the 2004 data, the best model based on the SIC, adjusted R^2 and RMSE is almost the same model as in 1995, but with an added $EDUC^2$ term. And for the 2012 data, the best model, as based on SIC and RMSE, is the exact same specification as the model for 1995. There are some other models with lower adjusted R^2 values, but we prefer the SIC as a selection criteria given its oracle property and the robustness check provided by the slighlty lower RMSE.

Looking into the coefficients and t-statistics of the of the best models for 2004 and 2012 respectivaly, we can see below that all of the beta coefficients have associated p-values of less than 0.05 and are statistically significant.

2004 Coefficients

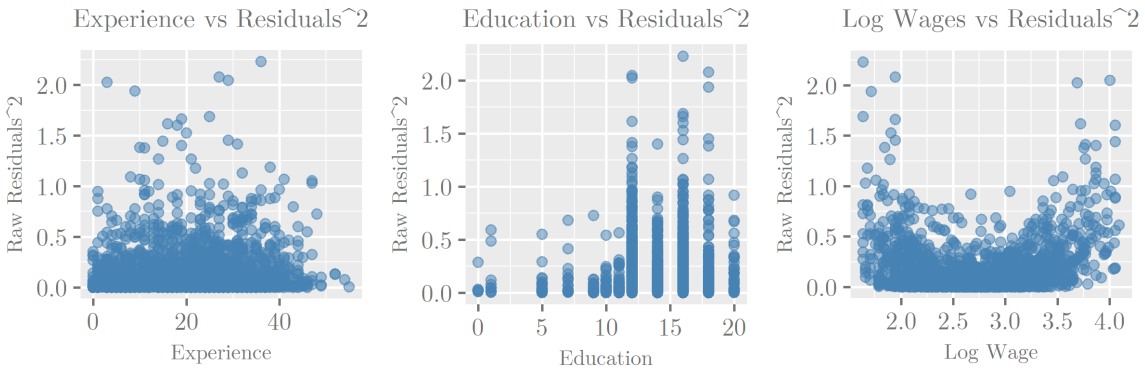
term	estimate	std.error	t.statistic	p.value	SIC	adj.R2	RMSE
(Intercept)	1.2832139	0.1744233	7.356895	0.0000000	2076.89	0.3871	0.414
EXPER	0.0470905	0.0056096	8.394555	0.0000000			
EDUC	0.0529986	0.0208909	2.536926	0.0112638			
EXPER2	-0.0005554	0.0000695	-7.993137	0.0000000			
EDUC2	0.0024276	0.0006423	3.779611	0.0001620			
FEMALE1	-0.2311514	0.0192753	-11.992133	0.0000000			
NONWHITE1	-0.1014759	0.0238483	-4.255057	0.0000219			
UNION1	0.1023419	0.0290422	3.523903	0.0004355			
EXPER:EDUC	-0.0010344	0.0003146	-3.287699	0.0010288			

2012 Coefficients

term	estimate	std.error	t.statistic	p.value	SIC	adj.R2	RMSE
(Intercept)	0.8375842	0.1410422	5.938536	0.0000000	1420.24	0.3392	0.431
FEMALE1	-0.1718186	0.0252227	-6.812061	0.0000000			
NONWHITE1	-0.0682748	0.0334632	-2.040298	0.0415435			
UNION1	0.1416845	0.0380668	3.721999	0.0002070			
EXPER	0.0523299	0.0068420	7.648344	0.0000000			
EXPER2	-0.0005209	0.0000869	-5.993025	0.0000000			
EDUC	0.1243376	0.0095029	13.084156	0.0000000			
EXPER:EDUC	-0.0014061	0.0003569	-3.940146	0.0000862			

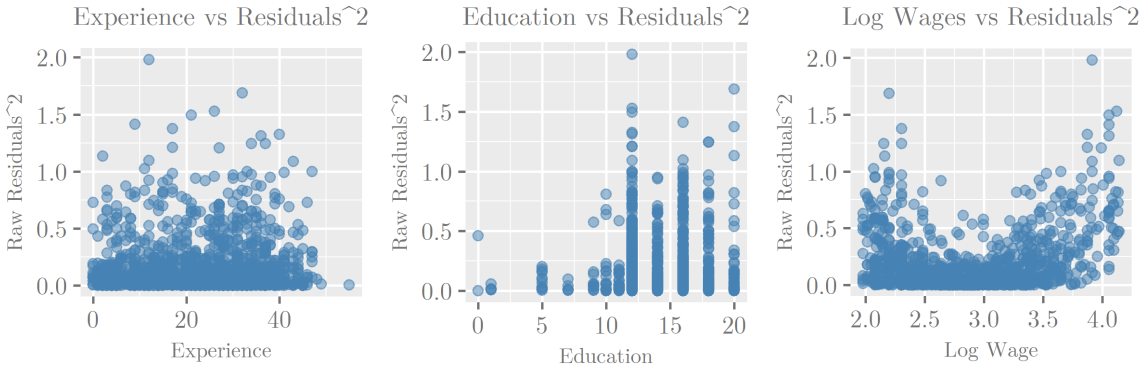
Next we evaluate the heteroscedasticity of the residuals in the model. The below graphs of the raw residuals squared show the same pattern of heterocedasticity as in the 1995 data where the variance is higher at extreme values of log wage. Furthermore, the BGP test returns p-values that are essentially 0 in both years, leading us to reject the null of homoscedasticity

2004 Heteroscedasticity tests



bgp_statistic	p.value	parameter
75.65389	0	8

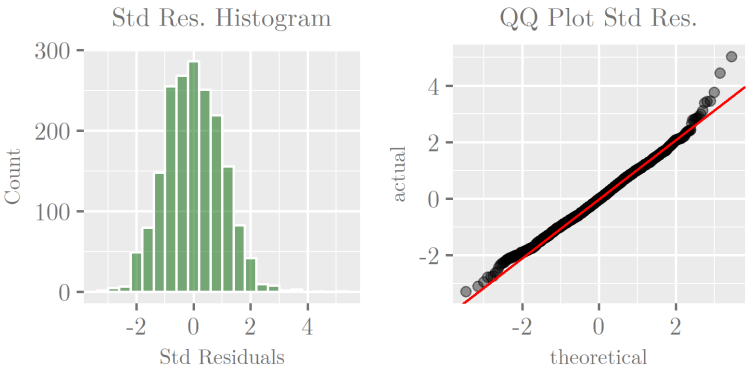
2012 Heteroscedasticity tests



bgp_statistic	p.value	parameter
37.09987	1.1e-05	8

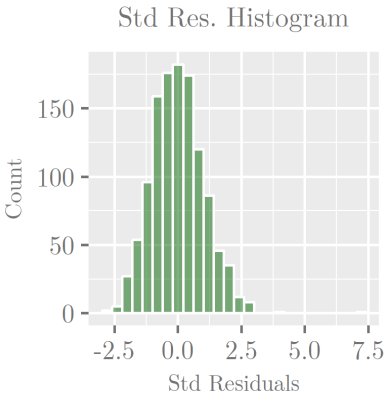
Because of the heteroscedasticity in both models, we compute the standardized residuals and analyze that distribution to test for normality of the residuals in the model. As a side note, we use the White Regression instead of the BGP regression to obtain the estimates of the variance of the residuals because using the BGP regression gave us back negative values which we can't take the square root of. Below is a qqplot, residual histogram and JB test for normality on the standardized residuals for 2004 and 2012

2004 tests for normality



JB.test.stat	p.value	df	method
27.845	0	2	Jarque Bera Test

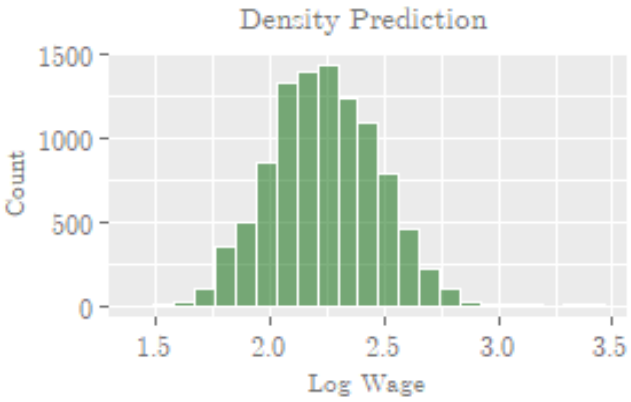
2012 Tests for normality



JB.test.stat	p.value	df	method
187.455	0	2	Jarque Bera Test

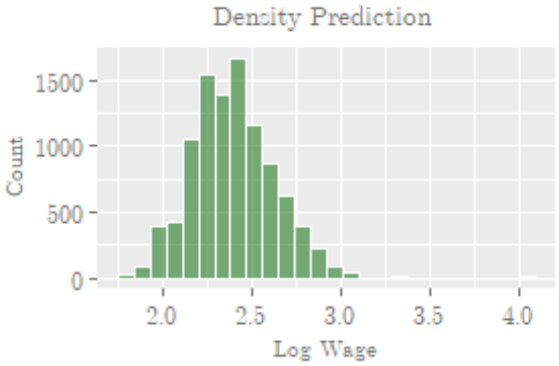
So as the graphs show, the residuals seem to be a little right skewed, and have a fat right tail and somewhat of a thin left tail compared to the normal. The JB test confirms the non-normality of these histograms as they return p-values of essentially 0. Thus we again have to use a simulation algorithm from the standardized residuals to obtain interval and density predictions. Below are the point, interval, and density predictions for the new person using the simulation algorithm for the 95% Confidence Interval and the density for 2004 and 2012.

Predictions on 2004 Data



Point_estimate	X2.5.	X97.5.
2.238996	1.787369	2.704106

Prediction on 2012 Data



Point_estimate	X2.5.	X97.5.
2.401182	1.973099	2.874365

VII. Conclusion

In conclusion, we now we have point, interval, and density estimates for the new person in all 3 time periods. It is worth mentioning that these models do suffer from heteroscedasticity and while we have made some effort to correct for this when building interval and density prediction, more effort should be placed into finding better predictors to help explain the heteroscedasticity and make it a better model.

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning=FALSE)

library(psych)
library(knitr)
library(extrafont)
loadfonts(quiet=T)
library(gridExtra)
library(ggplot2)
library(scales)
library(Blossom)
library(quantreg)
library(lmtest)
library(corrplot)
library(reshape2)
library(PerformanceAnalytics)
library(broom)
library(tidyr)
library(data.table)
library(DAAG)
library(MASS)
library(plotly)
library(tseries)
library(dplyr)

wage_95 =tbl_df(read.csv("output95_update.csv",colClasses=c(rep("factor",3),rep("numeric",5))))
wage_04 =tbl_df( read.csv("output04_update.csv",colClasses=c(rep("factor",3),rep("numeric",5))))
wage_12 = tbl_df(read.csv("output12_update.csv",colClasses=c(rep("factor",3),rep("numeric",5))))

# kable(summary(wage_95))
# summary(wage_04)
# summary(wage_12)

data_95 = wage_95 %>%
  filter(EXPER>= 0,
         WAGE >= 4.25,
         WAGE <= 40)# %>%
  # mutate(DIFF = (AGE-EXPER)) %>%
  # filter(DIFF>=14)
kable(summary(data_95))


darkgray_theme=theme(text=element_text(size=8,  family="LM Roman 10", color = "gray45"), axis.text= elemen
t_text(size=9,  family="LM Roman 10", color = "gray45"), plot.title = element_text(hjust = 0.5), axis.tick
s = element_line(color = "gray45"))

wage_hist = ggplot(data = data_95, aes(WAGE)) +
    geom_histogram(col = "white",
                   fill = "darkgreen",
                   binwidth = 3,
                   alpha=.5) +
                   labs(title="Wage") +
                   labs(x="Wage (Dollars)", y="")+
                   darkgray_theme
lnwage_hist = ggplot(data = data_95, aes(LNWAGE)) +
    geom_histogram(col = "white",
                   fill = "darkgreen",
                   binwidth = .2,
                   alpha=.5) +
                   labs(title="Log Wage") +
                   labs(x="Log Wage (Dollars)", y="")+
                   darkgray_theme
# grid.arrange(wage_hist, lnwage_hist, ncol =2)


# layout(matrix(c(1,2,3,4), 2, 2, byrow = TRUE))
age_hist = ggplot(data = data_95, aes(AGE)) +
    geom_histogram(col = "white",
```

```
                  fill = "darkgreen",
                  binwidth = 3,
                  alpha=.5) +
                  labs(title="Age") +
                  labs(x="Years", y="Count")+
                  darkgray_theme

educ_hist = ggplot(data = data_95, aes(EDUC)) +
    geom_histogram(col = "white",
                  fill = "darkgreen",
                  binwidth = 3,
                  alpha=.5) +
                  labs(title="Education") +
                  labs(x="Years", y=" ")+
                  darkgray_theme

exper_hist = ggplot(data = data_95, aes(EXPER)) +
    geom_histogram(col = "white",
                  fill = "darkgreen",
                  binwidth = 4,
                  alpha=.5) +
                  labs(title="Experience") +
                  labs(x="Years", y=" ")+
                  darkgray_theme

female_hist = ggplot(data = data_95, aes(data_95$FEMALE)) +
    geom_bar(aes(y = (..count..)/sum(..count..)),
                  col = "white",
                  fill = "darkgreen",
                  alpha=.5) +
                  labs(title="Female") +
                  labs(x="", y="Percentage")+
                  scale_y_continuous(labels=percent) +
                  darkgray_theme

nonwhite_hist = ggplot(data = data_95, aes(as.character(data_95$NONWHITE))) +
    geom_bar(aes(y = (..count..)/sum(..count..)),
                  col = "white",
                  fill = "darkgreen",
                  alpha=.5) +
                  labs(title="Non-White") +
                  labs(x="", y="")+
                  scale_y_continuous(labels=percent) +
                  darkgray_theme

union_hist = ggplot(data = data_95, aes(as.character(data_95$UNION))) +
    geom_bar(aes(y = (..count..)/sum(..count..)),
                  col = "white",
                  fill = "darkgreen",
                  alpha=.5) +
                  labs(title="Union") +
                  labs(x="", y="")+
                  scale_y_continuous(labels=percent) +
                  darkgray_theme

grid.arrange(wage_hist, lnwage_hist, age_hist, educ_hist, exper_hist, female_hist, nonwhite_hist, union_hi
st, ncol = 3)

data_95 = data_95 %>%
  dplyr::select(FEMALE, NONWHITE, UNION, AGE, EXPER, EDUC, LNWAGE)

data_95[] <- lapply(data_95, function(x) as.numeric(as.character(x)))

corrplot(cor(data_95), method ="number", type = "lower")


# ggplot(data = data_95, aes(x = EXPER, y = LNWAGE)) +
#     geom_point(aes(color = as.factor(data_95$NONWHITE)) )
```

```r
exper_scatter = ggplot(data = data_95, aes(x = EXPER, y =LNWAGE)) +
    geom_point(color = "darkgreen", alpha =0.4) +
    geom_smooth(method = "loess", se = F, color = "steelblue1")+
    geom_smooth(method = "lm", se = F, color = "firebrick3")+
    darkgray_theme

educ_scatter = ggplot(data = data_95, aes(x = EDUC, y =LNWAGE)) +
    geom_jitter(color = "darkgreen", alpha =0.4) +
    geom_smooth(method = "loess", se = F, color = "steelblue1")+
    geom_smooth(method = "lm", se = F, color = "firebrick3")+
    darkgray_theme

age_scatter = ggplot(data = data_95, aes(x = AGE, y =LNWAGE)) +
    geom_point(color = "darkgreen", alpha =0.4) +
    geom_smooth(method = "loess", se = F, color = "steelblue1")+
    geom_smooth(method = "lm", se = F, color = "firebrick3")+
    darkgray_theme

# ggplot()+
#   geom_point(data = data_95, aes(x = FEMALE, y = LNWAGE))

grid.arrange(exper_scatter, educ_scatter, ncol = 2)

### Dropped the negative experience person
data_95 = data_95 %>%
  filter(EXPER >= 0)%>%
  mutate(EXPER2 = EXPER^2,
         EDUC2  = EDUC^2,
         AGE2   = AGE^2  )


attach(data_95, warn.conflicts = F)
age_ols = lm(LNWAGE ~ FEMALE + NONWHITE + UNION + AGE + EDUC, data = data_95)
age_ols = lm(LNWAGE ~ FEMALE + NONWHITE + UNION + AGE + EDUC, data = data_95)
exper_ols = lm(LNWAGE ~ FEMALE + NONWHITE + UNION + EXPER + EDUC, data = data_95)
exper2_ols = lm(LNWAGE ~FEMALE + NONWHITE + UNION + EXPER + EDUC + EXPER2+
                  EXPER*EDUC, data = data_95)
exper_all2_ols = lm(LNWAGE ~FEMALE + NONWHITE+UNION+EXPER+EDUC+EXPER2+EDUC2
                    +EXPER*EDUC+NONWHITE*UNION+FEMALE*UNION+FEMALE*NONWHITE, data = data_95)
exper_int_binary = lm(LNWAGE ~FEMALE + NONWHITE+UNION+EXPER+EDUC+EXPER*EDUC+
                      NONWHITE*UNION+FEMALE*UNION+FEMALE*NONWHITE, data = data_95)
exper2_educ_ols = lm(LNWAGE ~EXPER+EDUC+EXPER*EDUC+EXPER2, data = data_95)
exper2_educ2_ols = lm(LNWAGE ~EXPER+EDUC+EXPER*EDUC+EXPER2+EDUC2, data = data_95)


# summary(exper_ols)
# summary(exper2_ols)
# summary(exper_all2_ols)

# cv_age_ols = CVlm(data =data_95, form.lm = age_ols, m = 5, printit = F, plotit = F )
# tidy_age   =tidy(summary(age_ols))
# glance_age = glance(summary(age_ols))%>%
#   mutate(SIC = BIC(age_ols),
#          RMSE = sqrt(attr(cv_age_ols, "ms")),
#          predictors = " Female, Nonwhite, Union, Age, Educ")%>%
#   setnames("statistic","f.statistic")

cv_exper_ols = CVlm(data =data_95, form.lm = exper_ols, m = 5,
                    printit = F, plotit = F )
tidy_exper   =tidy(summary(exper_ols))
glance_exper = glance(summary(exper_ols))%>%
  mutate(SIC = BIC(exper_ols),
         RMSE = sqrt(attr(cv_exper_ols, "ms")),
         predictors = " Female, Nonwhite, Union, Exper, Educ")%>%
  setnames("statistic","f.statistic")

cv_exper2_ols = CVlm(data =data_95, form.lm = exper2_ols, m = 5,
                     printit = F, plotit = F )
tidy_exper2  =  tidy(summary(exper2_ols))
```

```r
glance_exper2 = glance(summary(exper2_ols))%>%
  mutate(SIC = BIC(exper2_ols),
         RMSE = sqrt(attr(cv_exper2_ols, "ms")),
         predictors = " Female, Nonwhite, Union, Exper, Educ, Exper^2, Exper*Educ")%>%
  setnames("statistic","f.statistic")

cv_exper_all2_ols = CVlm(data =data_95, form.lm = exper_all2_ols, m = 5,
                         printit = F, plotit = F )
tidy_exper_all2 = tidy(summary(exper_all2_ols))
glance_exper_all2 = glance(summary(exper_all2_ols))%>%
  mutate(SIC = BIC(exper_all2_ols),
         RMSE = sqrt(attr(cv_exper_all2_ols, "ms")),
         predictors = c("Female, Nonwhite, Union, Exper, Educ, Exper^2, Educ^2 ExperxEduc, NonwhitexUnion,
FemalexUnion, FemalexNonwhite")) %>%
  setnames("statistic","f.statistic")

cv_exper_int_binary = CVlm(data =data_95, form.lm = exper_int_binary, m = 5,
                           printit = F, plotit = F )
tidy_exper_int_binary = tidy(summary(exper_int_binary))
glance_exper_int_binary = glance(summary(exper_int_binary))%>%
  mutate(SIC = BIC(exper_int_binary),
         RMSE = sqrt(attr(cv_exper_int_binary, "ms")),
         predictors = c("Female, Nonwhite, Union, Exper, Educ, ExperxEduc, NonwhitexUnion, FemalexUnion, F
emalexNonwhite")) %>%
  setnames("statistic","f.statistic")

cv_exper2_educ_ols = CVlm(data =data_95, form.lm = exper2_educ_ols, m = 5,
                          printit = F, plotit = F )
tidy_exper2_educ_ols = tidy(summary(exper2_educ_ols))
glance_exper2_educ_ols = glance(summary(exper2_educ_ols))%>%
  mutate(SIC = BIC(exper2_educ_ols),
         RMSE = sqrt(attr(cv_exper2_educ_ols, "ms")),
         predictors = c(" Exper, Educ, ExperxEduc, Exper^2")) %>%
  setnames("statistic","f.statistic")

cv_exper2_educ2_ols = CVlm(data =data_95, form.lm = exper2_educ2_ols, m = 5,
                           printit = F, plotit = F )
tidy_exper2_educ2_ols = tidy(summary(exper2_educ2_ols))
glance_exper2_educ2_ols = glance(summary(exper2_educ2_ols))%>%
  mutate(SIC = BIC(exper2_educ2_ols),
         RMSE = sqrt(attr(cv_exper2_educ2_ols, "ms")),
         predictors = c(" Exper, Educ, ExperxEduc, Exper^2, Educ^2")) %>%
  setnames("statistic","f.statistic")


kable(list(glance_exper2_educ_ols, glance_exper2_educ2_ols, glance_exper,
           glance_exper2, glance_exper_all2, glance_exper_int_binary),
      caption = "Regression Models")




female_hist_lnwage = ggplot(data_95, aes(x = LNWAGE, fill = as.factor(FEMALE)))+
  geom_histogram(position= "identity",
                 bins = 20,
                 alpha = 0.4,
                 col = "white")+
  labs(title = "Log Wage broken down by Sex",
       x = "Ln Wage",
       y = "Count") +
  scale_fill_discrete(name="Sex",
                      breaks=c("0", "1"),
                      labels=c("Male", "Female"))+
   darkgray_theme


female_hist_exper = ggplot(data_95, aes(x = EXPER, fill = as.factor(FEMALE)))+
  geom_histogram(position= "identity",
```

```
                 bins = 20,
                 alpha = 0.4,
                 col = "white")+
  labs(title = "Experience broken down by Sex",
       x = "Experience",
       y = "Count") +
  scale_fill_discrete(name="Sex",
                      breaks=c("0", "1"),
                      labels=c("Male", "Female"))+
   darkgray_theme


data_95_female = data_95 %>%
  filter(FEMALE == 1, EXPER>=0)

data_95_male = data_95 %>%
  filter(FEMALE == 0, EXPER>=0)

grid.arrange(female_hist_lnwage, female_hist_exper, ncol =2)

# kable(summary(data_95_female), caption = "Females")
# kable(summary(data_95_male), caption = "Males")


female_95_lm = lm(LNWAGE ~FEMALE + NONWHITE + UNION + EXPER + EDUC + EXPER2+
                    EDUC2 + EXPER*EDUC, data = data_95_female )

male_95_lm = lm(LNWAGE ~ FEMALE + NONWHITE + UNION + EXPER + EDUC + EXPER2+
                  EDUC2 + EXPER*EDUC, data = data_95_male )


female_95_lm = lm(LNWAGE ~FEMALE + NONWHITE + UNION + EXPER + EDUC + EXPER2
                  + EXPER*EDUC, data = data_95_female)


cv_female_95 = CVlm(data =data_95_female, form.lm = female_95_lm, m = 3,
                    printit = F, plotit = F )
kable(tidy(summary(female_95_lm))%>%
        mutate(SIC = c(round(BIC(female_95_lm),2),"","","","","",""),
               adj.R2 = c(round(summary(female_95_lm)$adj.r.squared, 4), "",
                          "","","","",""),
               RMSE = c(round(sqrt(attr(cv_female_95, "ms")),3),"","","","",
                        "","" )),
      caption = "Female")




male_95_lm = lm(LNWAGE ~FEMALE + NONWHITE + UNION + EXPER + EDUC + EXPER2+
                  + EXPER*EDUC, data = data_95_male)
cv_male_95 = CVlm(data =data_95_male, form.lm = male_95_lm, m = 3,
                    printit = F, plotit = F )
kable(tidy(summary(male_95_lm))%>%
        mutate(SIC = c(round(BIC(male_95_lm),2),"","","","","",""),
               adj.R2 = c(round(summary(male_95_lm)$adj.r.squared, 4), "",
                          "","","","",""),
               RMSE = c(round(sqrt(attr(cv_male_95, "ms")),3),"","","","",
                        "","" )),
      caption = "Male")



 kable(tidy(summary(exper2_ols)) %>% setnames("statistic","t.statistic")%>%
         mutate(SIC = c(round(BIC(exper2_ols),2),"","","","","","",""),
               adj.R2 = c(round(summary(exper2_ols)$adj.r.squared, 4), "","",
                          "","","","",""),
               RMSE = c(round(sqrt(attr(cv_exper2_ols, "ms")),3),"","","","","",
                        "","" )))
```

```r
#plots to test heteroscedasticity
data_95$resid = exper2_ols$residuals

lnwage_vsresid = ggplot(data = data_95, aes(LNWAGE, y =resid^2)) +
  geom_point(col = "steelblue",
             alpha=.5) +
             labs(title="Log Wages vs Residuals") +
             labs(x="Log Wage", y="Unstandardized Residuals")+
             darkgray_theme
exper_vsresid = ggplot(data = data_95, aes(EXPER, y =resid^2)) +
  geom_point(col = "steelblue",
             alpha=.5) +
             labs(title="Experience vs Residuals") +
             labs(x="Experience", y="Unstandardized Residuals")+
             darkgray_theme

educ_vsresid = ggplot(data = data_95, aes(EDUC, y =resid^2)) +
  geom_point(col = "steelblue",
             alpha=.5) +
             labs(title="Education vs Residuals") +
             labs(x="Education", y="Unstandardized Residuals")+
             darkgray_theme

grid.arrange(exper_vsresid, educ_vsresid, lnwage_vsresid, ncol = 3)


#formal tests for heteroscedasticity

bgp_ols = lm(exper2_ols$residuals^2 ~ EDUC+EXPER+EXPER2+NONWHITE+FEMALE+UNION+EXPER*EDUC,
             data = data_95)
bgp_stat = nrow(data_95)*(summary(bgp_ols)$r.squared)
#### bgpstat = 36, p-value is 0, we reject the null of homoscedasticity

# equivalently could use this one liner from lmtest pkg
kable(tidy(bptest(exper2_ols))%>% setnames("statistic","bgp_statistic")%>%
        select(bgp_statistic, p.value, parameter))

# kable(tidy(summary(bgp_ols)))

 whitelm<-lm(exper2_ols$residuals^2 ~ EDUC+EXPER+EXPER2+EDUC^2+NONWHITE+FEMALE+UNION+
               EXPER*EDUC+NONWHITE*FEMALE+NONWHITE*UNION+UNION*FEMALE, data = data_95)

sd_resid = sqrt(predict(bgp_ols, data_95))
std_resid = exper2_ols$residuals / sd_resid
bgp_stat = nrow(data_95)*(summary(bgp_ols)$r.squared)

data_95$std_resid = std_resid

std_res_hist = ggplot()+
  geom_histogram(aes(data_95$std_resid),
                 col = "white",
                 fill = "darkgreen",
                 binwidth = .4,
                 alpha=.5) +
                 labs(title="Std Res. Histogram") +
                 labs(x="Std Residuals", y="")+
                 darkgray_theme


#Code for creating qqline
y     <- quantile(std_resid, c(0.25, 0.75)) # Find the 1st and 3rd quartiles
x     <- qnorm( c(0.25, 0.75))         # Find the matching normal values on the x-axis
slope <- diff(y) / diff(x)             # Compute the line slope
int   <- y[1] - slope * x[1]           # Compute the line intercept

std_res_qqplot = ggplot()+
  geom_qq(aes(sample = std_resid),
          alpha = 0.4) +
  geom_abline(intercept=int,
```

```
              slope=slope,
              col = "red")+
              ylab("actual")+
  ggtitle("QQ Plot Std Res.")+
  darkgray_theme

### stat = 36, p-value is 0, we reject the null of homoscedasticity

grid.arrange(std_res_hist, std_res_qqplot, ncol = 2)

# Print out table of JB test
kable(tbl_df(t(unlist(jarque.bera.test(data_95$std_resid))))%>%
        setnames("statistic.X-squared","JB.test.stat") %>%
        setnames("p.value.X-squared", "p.value") %>%
        setnames("parameter.df", "df") %>%
        mutate(JB.test.stat = (round(as.numeric(JB.test.stat), 3)),
               p.value      = (round(as.numeric(p.value),3))) %>%
        select(JB.test.stat, p.value, df, method))


# Dont need plot of std residuals vs xs and ys bc its the same as above
# lnwage_vs_stdresid = ggplot(data = data_95, aes(LNWAGE, y =std_resid^2)) +
#   geom_point(col = "steelblue",
#              alpha=.5) +
#              labs(title="Log Wages vs Residuals") +
#              labs(x="Log Wage", y="Unstandardized Residuals")+
#              darkgray_theme
# exper_vs_stdresid = ggplot(data = data_95, aes(EXPER, y =std_resid^2)) +
#   geom_point(col = "steelblue",
#              alpha=.5) +
#              labs(title="Experience vs Residuals") +
#              labs(x="Experience", y="Unstandardized Residuals")+
#              darkgray_theme
#
# educ_vs_stdresid = ggplot(data = data_95, aes(EDUC, y =std_resid^2)) +
#   geom_point(col = "steelblue",
#              alpha=.5) +
#              labs(title="Education vs Residuals") +
#              labs(x="Education", y="Unstandardized Residuals")+
#              darkgray_theme
# grid.arrange(exper_vs_stdresid, educ_vs_stdresid, lnwage_vs_stdresid, ncol = 3)


# Simulation algorithm for density prediction

# draw_std_res =  sample(data_95$std_resid, 10000, replace = TRUE)
# draw_fitted_val=sample(exper2_ols$fitted.values, 10000, replace = TRUE)
# full_draw = draw_std_res +draw_fitted_val
new_arrival = data.frame(UNION =1, FEMALE = 1, NONWHITE = 0, EDUC = 12,
                         EXPER = 3, EXPER2 = 9, EDUC2 = 12^2, AGE = 0,
                         AGE2 = 0, xxx = 10, yyy = 10)

sample_std_res = sample(data_95$std_resid, 10000, replace = TRUE)
new_arrival_sd = sqrt(predict(bgp_ols,new_arrival))

# new_arrival_sd=sqrt(predict(whitelm,new_arrival))
new_arrival_mean = predict(exper2_ols, new_arrival)
new_arrival_dist = sample_std_res*new_arrival_sd + new_arrival_mean


new_arrival_sd = sqrt(predict(bgp_ols,new_arrival))
dist_new_arrival = sample_std_res*new_arrival_sd + new_arrival_mean


new_arrival_distribution_hist = ggplot()+
  geom_histogram(aes(dist_new_arrival),
                 col = "white",
                 fill = "darkgreen",
                 binwidth = .09,
```

```
                alpha=.5) +
                labs(title="Density Prediction ") +
                labs(x="Log Wage", y="Count")+
                darkgray_theme

new_person_predictions = cbind(Point_estimate = new_arrival_mean,
                               data.frame(as.list(quantile(new_arrival_dist, c(0.025,0.975)))))


new_arrival_distribution_hist
kable(new_person_predictions)


#Wage_04 has 1941 rows, 2 of which have negative experoence, 15 of which have WAGE > 60, and 48 rows less
than 5.15 (min wage)
data_04 = wage_04 %>%
  filter(EXPER>= 0,
         WAGE >= 5.15,
         WAGE <= 60)# %>%
  # mutate(DIFF = (AGE-EXPER)) %>%
  # filter(DIFF>=14)
#kable(summary(data_04))

#wage_12 has 1229 rows, 2 of which have negative experience, 16 of whom have wages higher than 63, and 24
rows less thatn 7.35

data_12 = wage_12 %>%
  filter(EXPER>= 0,
         WAGE >= 7.25,
         WAGE <= 63)

### Dropped the negative experience person
data_04 = data_04 %>%
  filter(EXPER >= 0)%>%
  mutate(EXPER2 = EXPER^2,
         EDUC2  = EDUC^2,
         AGE2   = AGE^2  )


# attach(data_95, warn.conflicts = F)
age_ols_04 = lm(LNWAGE ~ FEMALE + NONWHITE + UNION + AGE + EDUC, data = data_04)
exper_ols_04 = lm(LNWAGE ~ FEMALE + NONWHITE + UNION + EXPER + EDUC, data = data_04)
exper2_ols_04 = lm(LNWAGE ~FEMALE + NONWHITE + UNION + EXPER+EXPER2+EDUC+
                   EXPER*EDUC, data = data_04)
exper_all2_ols_04 = lm(LNWAGE ~FEMALE + NONWHITE+UNION+EXPER+EDUC+EXPER2+EDUC2
                     +EXPER*EDUC+NONWHITE*UNION+FEMALE*UNION+FEMALE*NONWHITE, data = data_04)
exper_int_binary_04 = lm(LNWAGE ~FEMALE + NONWHITE+UNION+EXPER+EDUC+EXPER*EDUC+
                       NONWHITE*UNION+FEMALE*UNION+FEMALE*NONWHITE, data = data_04)
exper2_educ_ols_04 = lm(LNWAGE ~EXPER+EDUC+EXPER*EDUC+EXPER2, data = data_04)
exper2_educ2_ols_04 = lm(LNWAGE ~EXPER+EDUC+EXPER*EDUC+EXPER2+EDUC2+FEMALE+
                         NONWHITE+UNION, data = data_04)


# summary(exper_ols)
# summary(exper2_ols)
# summary(exper_all2_ols)

# cv_age_ols = CVlm(data =data_95, form.lm = age_ols, m = 5, printit = F, plotit = F )
# tidy_age   =tidy(summary(age_ols))
# glance_age = glance(summary(age_ols))%>%
#   mutate(SIC = BIC(age_ols),
#          RMSE = sqrt(attr(cv_age_ols, "ms")),
#          predictors = " Female, Nonwhite, Union, Age, Educ")%>%
#   setnames("statistic","f.statistic")

cv_exper_ols_04 = CVlm(data =data_04, form.lm = exper_ols_04, m = 5,
                     printit = F, plotit = F )
tidy_exper_04   =tidy(summary(exper_ols_04))
glance_exper_04 = glance(summary(exper_ols_04))%>%
```

```
  mutate(SIC = BIC(exper_ols_04),
         RMSE = sqrt(attr(cv_exper_ols_04, "ms")),
         predictors = " Female, Nonwhite, Union, Exper, Educ")%>%
  setnames("statistic","f.statistic")

cv_exper2_ols_04 = CVlm(data =data_04, form.lm = exper2_ols_04, m = 5,
                        printit = F, plotit = F )
tidy_exper2_04  =  tidy(summary(exper2_ols_04))
glance_exper2_04 = glance(summary(exper2_ols_04))%>%
  mutate(SIC = BIC(exper2_ols_04),
         RMSE = sqrt(attr(cv_exper2_ols_04, "ms")),
         predictors = " Female, Nonwhite, Union, Exper, Educ, Exper^2, Exper*Educ")%>%
  setnames("statistic","f.statistic")

cv_exper_all2_ols_04 = CVlm(data =data_04, form.lm = exper_all2_ols_04, m = 5,
                            printit = F, plotit = F )
tidy_exper_all2_04 = tidy(summary(exper_all2_ols_04))
glance_exper_all2_04 = glance(summary(exper_all2_ols_04))%>%
  mutate(SIC = BIC(exper_all2_ols_04),
         RMSE = sqrt(attr(cv_exper_all2_ols_04, "ms")),
         predictors = c("Female, Nonwhite, Union, Exper, Educ, Exper^2, Educ^2 ExperxEduc,
                        NonwhitexUnion, FemalexUnion, FemalexNonwhite")) %>%
  setnames("statistic","f.statistic")

cv_exper_int_binary_04 = CVlm(data =data_04, form.lm = exper_int_binary_04, m = 5,
                              printit = F, plotit = F )
tidy_exper_int_binary_04 = tidy(summary(exper_int_binary_04))
glance_exper_int_binary_04 = glance(summary(exper_int_binary_04))%>%
  mutate(SIC = BIC(exper_int_binary_04),
         RMSE = sqrt(attr(cv_exper_int_binary_04, "ms")),
         predictors = c("Female, Nonwhite, Union, Exper, Educ, ExperxEduc, NonwhitexUnion,
                        FemalexUnion, FemalexNonwhite")) %>%
  setnames("statistic","f.statistic")

cv_exper2_educ_ols_04 = CVlm(data =data_04, form.lm = exper2_educ_ols_04, m = 5,
                             printit = F, plotit = F )
tidy_exper2_educ_ols_04 = tidy(summary(exper2_educ_ols_04))
glance_exper2_educ_ols_04 = glance(summary(exper2_educ_ols_04))%>%
  mutate(SIC = BIC(exper2_educ_ols_04),
         RMSE = sqrt(attr(cv_exper2_educ_ols_04, "ms")),
         predictors = c(" Exper, Educ, ExperxEduc, Exper^2")) %>%
  setnames("statistic","f.statistic")

cv_exper2_educ2_ols_04 = CVlm(data =data_04, form.lm = exper2_educ2_ols_04, m = 5,
                              printit = F, plotit = F )
tidy_exper2_educ2_ols_04 = tidy(summary(exper2_educ2_ols_04))
glance_exper2_educ2_ols_04 = glance(summary(exper2_educ2_ols_04))%>%
  mutate(SIC = BIC(exper2_educ2_ols_04),
         RMSE = sqrt(attr(cv_exper2_educ2_ols_04, "ms")),
         predictors = c(" Exper, Educ, ExperxEduc, Exper^2, Educ^2, Female, NonWhite, Union")) %>%
  setnames("statistic","f.statistic")


kable(list(glance_exper2_educ_ols_04, glance_exper2_educ2_ols_04, glance_exper_04,
           glance_exper2_04, glance_exper_all2_04,
           glance_exper_int_binary_04),
      caption = "Regression Models")


### Dropped the negative experience person
data_12 = data_12 %>%
  filter(EXPER >= 0)%>%
  mutate(EXPER2 = EXPER^2,
         EDUC2  = EDUC^2,
         AGE2   = AGE^2  )


# attach(data_95, warn.conflicts = F)
age_ols_12 = lm(LNWAGE ~ FEMALE + NONWHITE + UNION + AGE + EDUC, data = data_12)
```

```r
exper_ols_12 = lm(LNWAGE ~ FEMALE + NONWHITE + UNION + EXPER + EDUC, data = data_12)
exper2_ols_12 = lm(LNWAGE ~FEMALE + NONWHITE + UNION + EXPER+EXPER2+EDUC+
                   EXPER*EDUC, data = data_12)
exper_all2_ols_12 = lm(LNWAGE ~FEMALE + NONWHITE+UNION+EXPER+EDUC+EXPER2+EDUC2
                    +EXPER*EDUC+NONWHITE*UNION+FEMALE*UNION+FEMALE*NONWHITE, data = data_12)
exper_int_binary_12 = lm(LNWAGE ~FEMALE + NONWHITE+UNION+EXPER+EDUC+EXPER*EDUC+
                     NONWHITE*UNION+FEMALE*UNION+FEMALE*NONWHITE, data = data_12)
exper2_educ_ols_12 = lm(LNWAGE ~EXPER+EDUC+EXPER*EDUC+EXPER2, data = data_12)
exper2_educ2_ols_12 = lm(LNWAGE ~EXPER+EDUC+EXPER*EDUC+EXPER2+EDUC2+FEMALE+
                       NONWHITE+UNION, data = data_12)
ols_final_12 = lm(LNWAGE ~ EXPER+ EXPER2+ EDUC2+ FEMALE+ UNION+ EDUC*EXPER, data = data_12)

# summary(exper_ols)
# summary(exper2_ols)
# summary(exper_all2_ols)

# cv_age_ols = CVlm(data =data_95, form.lm = age_ols, m = 5, printit = F, plotit = F )
# tidy_age   =tidy(summary(age_ols))
# glance_age = glance(summary(age_ols))%>%
#   mutate(SIC = BIC(age_ols),
#          RMSE = sqrt(attr(cv_age_ols, "ms")),
#          predictors = " Female, Nonwhite, Union, Age, Educ")%>%
#   setnames("statistic","f.statistic")

cv_exper_ols_12 = CVlm(data =data_12, form.lm = exper_ols_12, m = 5,
                     printit = F, plotit = F )
tidy_exper_12   =tidy(summary(exper_ols_12))
glance_exper_12 = glance(summary(exper_ols_12))%>%
  mutate(SIC = BIC(exper_ols_12),
         RMSE = sqrt(attr(cv_exper_ols_12, "ms")),
         predictors = " Female, Nonwhite, Union, Exper, Educ")%>%
  setnames("statistic","f.statistic")

cv_exper2_ols_12 = CVlm(data =data_12, form.lm = exper2_ols_12, m = 5,
                      printit = F, plotit = F )
tidy_exper2_12  =  tidy(summary(exper2_ols_12))
glance_exper2_12 = glance(summary(exper2_ols_12))%>%
  mutate(SIC = BIC(exper2_ols_12),
         RMSE = sqrt(attr(cv_exper2_ols_12, "ms")),
         predictors = " Female, Nonwhite, Union, Exper, Educ, Exper^2, Exper*Educ")%>%
  setnames("statistic","f.statistic")

cv_exper_all2_ols_12 = CVlm(data =data_12, form.lm = exper_all2_ols_12, m = 5,
                        printit = F, plotit = F )
tidy_exper_all2_12 = tidy(summary(exper_all2_ols_12))
glance_exper_all2_12 = glance(summary(exper_all2_ols_12))%>%
  mutate(SIC = BIC(exper_all2_ols_12),
         RMSE = sqrt(attr(cv_exper_all2_ols_12, "ms")),
         predictors = c("Female, Nonwhite, Union, Exper, Educ, Exper^2, Educ^2 ExperxEduc,
                        NonwhitexUnion, FemalexUnion, FemalexNonwhite")) %>%
  setnames("statistic","f.statistic")

cv_exper_int_binary_12 = CVlm(data =data_12, form.lm = exper_int_binary_12, m = 5,
                          printit = F, plotit = F )
tidy_exper_int_binary_12 = tidy(summary(exper_int_binary_12))
glance_exper_int_binary_12 = glance(summary(exper_int_binary_12))%>%
  mutate(SIC = BIC(exper_int_binary_12),
         RMSE = sqrt(attr(cv_exper_int_binary_12, "ms")),
         predictors = c("Female, Nonwhite, Union, Exper, Educ, ExperxEduc, NonwhitexUnion,
                        FemalexUnion, FemalexNonwhite")) %>%
  setnames("statistic","f.statistic")

cv_exper2_educ_ols_12 = CVlm(data =data_12, form.lm = exper2_educ_ols_12, m = 5,
                         printit = F, plotit = F )
tidy_exper2_educ_ols_12 = tidy(summary(exper2_educ_ols_12))
glance_exper2_educ_ols_12 = glance(summary(exper2_educ_ols_12))%>%
  mutate(SIC = BIC(exper2_educ_ols_12),
         RMSE = sqrt(attr(cv_exper2_educ_ols_12, "ms")),
         predictors = c(" Exper, Educ, ExperxEduc, Exper^2")) %>%
```

```
  setnames("statistic","f.statistic")

cv_exper2_educ2_ols_12 = CVlm(data =data_12, form.lm = exper2_educ2_ols_12, m = 5,
                          printit = F, plotit = F )
tidy_exper2_educ2_ols_12 = tidy(summary(exper2_educ2_ols_12))
glance_exper2_educ2_ols_12 = glance(summary(exper2_educ2_ols_12))%>%
  mutate(SIC = BIC(exper2_educ2_ols_12),
         RMSE = sqrt(attr(cv_exper2_educ2_ols_12, "ms")),
         predictors = c(" Exper, Educ, ExperxEduc, Exper^2, Educ^2, Female, NonWhite, Union")) %>%
  setnames("statistic","f.statistic")


kable(list(glance_exper2_educ_ols_12, glance_exper2_educ2_ols_12, glance_exper_12,
           glance_exper2_12, glance_exper_all2_12,
           glance_exper_int_binary_12),
      caption = " 2012 Regression Models")



### FINAL MODEL FOR 2004 = exper2_educ2_ols_04
 kable(tidy(summary(exper2_educ2_ols_04)) %>% setnames("statistic","t.statistic")%>%
         mutate(SIC = c(round(BIC(exper2_educ2_ols_04),2),"","","","","","","","","),
                adj.R2 = c(round(summary(exper2_educ2_ols_04)$adj.r.squared, 4), "","",
                           "","","","","","","),
                RMSE = c(round(sqrt(attr(cv_exper2_educ2_ols_04, "ms")),3),"","","","","",
                         "","","" )))
 ### FINAL MODEL FOR 2012 = exper2_ols_12
kable(tidy(summary(exper2_ols_12)) %>% setnames("statistic","t.statistic")%>%
         mutate(SIC = c(round(BIC(exper2_ols_12),2),"","","","","","","","","),
                adj.R2 = c(round(summary(exper2_ols_12)$adj.r.squared, 4), "","",
                           "","","","","","","),
                RMSE = c(round(sqrt(attr(cv_exper2_ols_12, "ms")),3),"","","","","",
                         "","" )))
#plots to test heteroscedasticity
data_04$resid = exper2_educ2_ols_04$residuals
data_04$fitted = exper2_educ2_ols_04$fitted.values

lnwage_vsresid_04 = ggplot(data = data_04, aes(LNWAGE, y =resid^2)) +
  geom_point(col = "steelblue",
             alpha=.5) +
             labs(title="Log Wages vs Residuals^2") +
             labs(x="Log Wage", y="Raw Residuals^2")+
             darkgray_theme
exper_vsresid_04 = ggplot(data = data_04, aes(EXPER, y =resid^2)) +
  geom_point(col = "steelblue",
             alpha=.5) +
             labs(title="Experience vs Residuals^2") +
             labs(x="Experience", y="Raw Residuals^2")+
             darkgray_theme

educ_vsresid_04 = ggplot(data = data_04, aes(EDUC, y =resid^2)) +
  geom_point(col = "steelblue",
             alpha=.5) +
             labs(title="Education vs Residuals^2") +
             labs(x="Education", y="Raw Residuals^2")+
             darkgray_theme

grid.arrange(exper_vsresid_04, educ_vsresid_04, lnwage_vsresid_04, ncol = 3)


#formal tests for heteroscedasticity

bgp_ols_04 = lm(exper2_educ2_ols_04$residuals^2 ~ EDUC+EXPER+EXPER2+EDUC2+NONWHITE+FEMALE+UNION+EXPER*EDUC
,
              data = data_04)
bgp_stat_04 = nrow(data_04)*(summary(bgp_ols_04)$r.squared)
#### bgpstat = 36, p-value is 0, we reject the null of homoscedasticity

# equivalently could use this one liner from lmtest pkg
```

```r
kable(tidy(bptest(exper2_educ2_ols_04))%>% setnames("statistic","bgp_statistic")%>%
        select(bgp_statistic, p.value, parameter))

# kable(tidy(summary(bgp_ols)))


data_04$sq_resid = exper2_educ2_ols_04$residuals^2

whitelm_04<-lm(exper2_educ2_ols_04$residuals^2  ~ EDUC+EXPER+EXPER2+EDUC2+NONWHITE+FEMALE+UNION+
               EXPER*EDUC+NONWHITE*FEMALE+NONWHITE*UNION+UNION*FEMALE, data = data_04)
#plots to test heteroscedasticity
data_12$resid = exper2_educ2_ols_12$residuals
data_12$fitted = exper2_educ2_ols_12$fitted.values

lnwage_vsresid_12 = ggplot(data = data_12, aes(LNWAGE, y =resid^2)) +
  geom_point(col = "steelblue",
             alpha=.5) +
             labs(title="Log Wages vs Residuals^2") +
             labs(x="Log Wage", y="Raw Residuals^2")+
             darkgray_theme
exper_vsresid_12 = ggplot(data = data_12, aes(EXPER, y =resid^2)) +
  geom_point(col = "steelblue",
             alpha=.5) +
             labs(title="Experience vs Residuals^2") +
             labs(x="Experience", y="Raw Residuals^2")+
             darkgray_theme

educ_vsresid_12 = ggplot(data = data_12, aes(EDUC, y =resid^2)) +
  geom_point(col = "steelblue",
             alpha=.5) +
             labs(title="Education vs Residuals^2") +
             labs(x="Education", y="Raw Residuals^2")+
             darkgray_theme

grid.arrange(exper_vsresid_12, educ_vsresid_12, lnwage_vsresid_12, ncol = 3)


#formal tests for heteroscedasticity

bgp_ols_12 = lm(exper2_educ2_ols_12$residuals^2 ~ EDUC+EXPER+EXPER2+EDUC2+NONWHITE+FEMALE+UNION+EXPER*EDUC
,
              data = data_12)
bgp_stat_12 = nrow(data_12)*(summary(bgp_ols_12)$r.squared)
#### bgpstat = 36, p-value is 0, we reject the null of homoscedasticity

# equivalently could use this one liner from lmtest pkg
kable(tidy(bptest(exper2_educ2_ols_12))%>% setnames("statistic","bgp_statistic")%>%
        select(bgp_statistic, p.value, parameter))

# kable(tidy(summary(bgp_ols)))


data_12$sq_resid = exper2_educ2_ols_12$residuals^2

whitelm_12<-lm(exper2_educ2_ols_12$residuals^2  ~ EDUC+EXPER+EXPER2+EDUC2+NONWHITE+FEMALE+UNION+
               EXPER*EDUC+NONWHITE*FEMALE+NONWHITE*UNION+UNION*FEMALE, data = data_12)

sd_resid_04 = sqrt(predict(whitelm_04, data_04))
std_resid_04= exper2_educ2_ols_04$residuals / sd_resid_04
bgp_stat_04 = nrow(data_04)*(summary(bgp_ols_04)$r.squared)

data_04$sd_resid_04 = sd_resid_04
data_04$std_resid_04 = std_resid_04


std_res_hist_04 = ggplot()+
  geom_histogram(aes(data_04$std_resid_04),
                 col = "white",
                 fill = "darkgreen",
```

```
                binwidth = .4,
                alpha=.5) +
                labs(title="Std Res. Histogram") +
                labs(x="Std Residuals", y="Count")+
                darkgray_theme


#Code for creating qqline
y1     <- quantile(std_resid_04, c(0.25, 0.75)) # Find the 1st and 3rd quartiles
x1    <- qnorm(c(0.25, 0.75))             # Find the matching normal values on the x-axis
slope1 <- diff(y1) / diff(x1)             # Compute the line slope
int1   <- y1[1] - slope * x1[1]           # Compute the line intercept

std_res_qqplot_04 = ggplot()+
  geom_qq(aes(sample = std_resid_04),
          alpha = 0.4) +
  geom_abline(intercept=int1,
              slope=slope1,
              col = "red")+
              ylab("actual")+
  ggtitle("QQ Plot Std Res.")+
  darkgray_theme

### stat = 36, p-value is 0, we reject the null of homoscedasticity

grid.arrange(std_res_hist_04, std_res_qqplot_04, ncol = 2)

# Print out table of JB test
kable(tbl_df(t(unlist(jarque.bera.test((data_04$std_resid_04)))))%>%
        setnames("statistic.X-squared","JB.test.stat") %>%
        setnames("p.value.X-squared", "p.value") %>%
        setnames("parameter.df", "df") %>%
        mutate(JB.test.stat = (round(as.numeric(JB.test.stat), 3)),
               p.value      = (round(as.numeric(p.value),3))) %>%
        select(JB.test.stat, p.value, df, method))


sd_resid_12 = sqrt(predict(whitelm_12, data_12))
std_resid_12= exper2_educ2_ols_12$residuals / sd_resid_12
bgp_stat_12 = nrow(data_12)*(summary(bgp_ols_12)$r.squared)

data_12$sd_resid_12 = sd_resid_12
data_12$std_resid_12 = std_resid_12


std_res_hist_12 = ggplot()+
  geom_histogram(aes(data_12$std_resid_12),
                col = "white",
                fill = "darkgreen",
                binwidth = .4,
                alpha=.5) +
                labs(title="Std Res. Histogram") +
                labs(x="Std Residuals", y="Count")+
                darkgray_theme


#Code for creating qqline
y1     <- quantile(std_resid_12, c(0.25, 0.75)) # Find the 1st and 3rd quartiles
x1    <- qnorm(c(0.25, 0.75))             # Find the matching normal values on the x-axis
slope1 <- diff(y1) / diff(x1)             # Compute the line slope
int1   <- y1[1] - slope * x1[1]           # Compute the line intercept

std_res_qqplot_12 = ggplot()+
  geom_qq(aes(sample = std_resid_12),
          alpha = 0.4) +
  geom_abline(intercept=int1,
              slope=slope1,
              col = "red")+
              ylab("actual")+
```

```
  ggtitle("QQ Plot Std Res.")+
  darkgray_theme

### stat = 36, p-value is 0, we reject the null of homoscedasticity

grid.arrange(std_res_hist_12, std_res_qqplot_12, ncol = 2)

# Print out table of JB test
kable(tbl_df(t(unlist(jarque.bera.test((data_12$std_resid_12)))))%>%
        setnames("statistic.X-squared","JB.test.stat") %>%
        setnames("p.value.X-squared", "p.value") %>%
        setnames("parameter.df", "df") %>%
        mutate(JB.test.stat = (round(as.numeric(JB.test.stat), 3)),
               p.value      = (round(as.numeric(p.value),3))) %>%
        select(JB.test.stat, p.value, df, method))

# Simulation algorithm for density prediction

new_arrival_04 = data.frame(UNION =as.factor(1), FEMALE = as.factor(1), NONWHITE = as.factor(0),
                         EDUC = 12, EXPER = 3, EXPER2 = 9, EDUC2 = 12^2, AGE = 0,
                         AGE2 = 0, xxx = 10, yyy = 10)

sample_std_res_04 = sample((data_04$std_resid_04), 10000, replace = TRUE)
new_arrival_sd_04 = sqrt(predict(whitelm_04,new_arrival_04))

# new_arrival_sd=sqrt(predict(whitelm,new_arrival))
new_arrival_mean_04 = predict(exper2_educ2_ols_04, new_arrival_04)
new_arrival_dist_04 = sample_std_res_04*new_arrival_sd_04 + new_arrival_mean_04




new_arrival_distribution_hist_04 = ggplot()+
  geom_histogram(aes(new_arrival_dist_04),
                 col = "white",
                 fill = "darkgreen",
                 binwidth = .09,
                 alpha=.5) +
                 labs(title="Density Prediction ") +
                 labs(x="Log Wage", y="Count")+
                 darkgray_theme

new_person_predictions_04 = cbind(Point_estimate = new_arrival_mean_04,
                               data.frame(as.list(quantile(new_arrival_dist_04, c(0.025,0.975)))))

new_arrival_distribution_hist_04
kable(new_person_predictions_04)

# Simulation algorithm for density prediction

new_arrival_12 = data.frame(UNION =as.factor(1), FEMALE = as.factor(1), NONWHITE = as.factor(0),
                         EDUC = 12, EXPER = 3, EXPER2 = 9, EDUC2 = 12^2, AGE = 0,
                         AGE2 = 0, xxx = 10, yyy = 10)

sample_std_res_12 = sample((data_12$std_resid_12), 10000, replace = TRUE)
new_arrival_sd_12 = sqrt(predict(whitelm_12,new_arrival_12))

# new_arrival_sd=sqrt(predict(whitelm,new_arrival))
new_arrival_mean_12 = predict(exper2_ols_12, new_arrival_12)
new_arrival_dist_12 = sample_std_res_12*new_arrival_sd_12 + new_arrival_mean_12




new_arrival_distribution_hist_12 = ggplot()+
  geom_histogram(aes(new_arrival_dist_12),
                 col = "white",
                 fill = "darkgreen",
                 binwidth = .09,
```

```
                alpha=.5) +
                labs(title="Density Prediction ") +
                labs(x="Log Wage", y="Count")+
                darkgray_theme

new_person_predictions_12 = cbind(Point_estimate = new_arrival_mean_12,
                                  data.frame(as.list(quantile(new_arrival_dist_12, c(0.025,0.975)))))

new_arrival_distribution_hist_12
kable(new_person_predictions_12)
```