

Project1 - Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials

Ieva Jankevičiūtė Ilja Jurčenko

2023-11-28

R markdown project

This project analyses data from Wooldridge Source: M. Blackburn and D. Neumark (1992), “Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials,” Quarterly Journal of Economics 107, 1421-1436. It loads lazily data from **wooldridge package**.

In this exploration, we aim to unpack the hypothesis suggesting that married people have higher wages. As we navigate the intricacies of a linear regression model, our focus extends beyond scrutinizing the “married” variable alone. We delve into not only the coefficient and significance of marital status but also other pertinent correlations with wage.

A data.frame with 935 observations on 17 variables: Removing missing values, total 663 observations

wage: monthly earnings.

hours: average weekly hours.

IQ: IQ score.

KWW: knowledge of world work score.

educ: years of education.

exper: years of work experience.

tenure: years with current employer.

age: age in years.

married: =1 if married.

black: =1 if black.

south: =1 if live in south.

urban: =1 if live in SMSA (Standard Metropolitan Statistical Area).

sibs: number of siblings.

brthord: birth order.

meduc: mother's education.

feduc: father's education.

lwage: natural log of wage.

```
library(wooldridge)
library(Hmisc)

data('wage2')
# Removing missing values, total 663 observations
wage2=na.omit(wage2)
#describe(wage2)
head(wage2)

##    wage hours   IQ KWW educ exper tenure age married black south urban sibs
```

```

## 1 769 40 93 35 12 11 2 31 1 0 0 1 1
## 3 825 40 108 46 14 11 9 33 1 0 0 1 1
## 4 650 40 96 32 12 13 7 32 1 0 0 1 4
## 5 562 40 74 27 11 14 5 34 1 0 0 1 10
## 7 600 40 91 24 10 13 0 30 0 0 0 1 1
## 9 1154 45 111 37 15 13 1 36 1 0 0 0 2
##   brthord meduc feduc    lwage
## 1      2     8     8 6.645091
## 3      2    14    14 6.715384
## 4      3    12    12 6.476973
## 5      6     6    11 6.331502
## 7      2     8     8 6.396930
## 9      3    14     5 7.050990

```

Data visualisation

Visualizing data to provide some insight on the research. Figure 1, 2, 3 are simple scatter plots between IQ and Wage colored by binary class. It shows some tendency that people with higher IQ tend to have higher wages.

Figure 1 - shows people with higher IQ tend to have higher wages and are married.

Figure 2 - shows people with lower IQ and lower wages tend to be African.

Figure 3 - shows people with higher IQ and higher wages live in cities.

Figure 4 - provides first information on dependencies between variables and its distributions. The lines are regression lines provided via LM parameter. The interesting dependencies are distinguished by the slope of the line. The dots are blurred for better readability. Ellipses in the scatterplot matrix represent confidence ellipses for the scatterplots. These ellipses provide a visual representation of the bivariate relationships between pairs of variables and show the spread and orientation of the data.

It is hard to read correlation between the data via big scatterplot that is why Figure 5 provides Pearson correlation information.

Figure 5 - provides information on correlation. Crossed variables are not significant. Color saturation reflects the impact. Blue have positive correlation, while red negative.

From the figures the lwage is most impacted by IQ, KWW, Educ, Age, Married, Meduc, Feduc. The interesting thing is that exper has very low correlation and not significant in Figure 5.

Figure 1: Scatter Plot of IQ and Wage Colored By Married Status

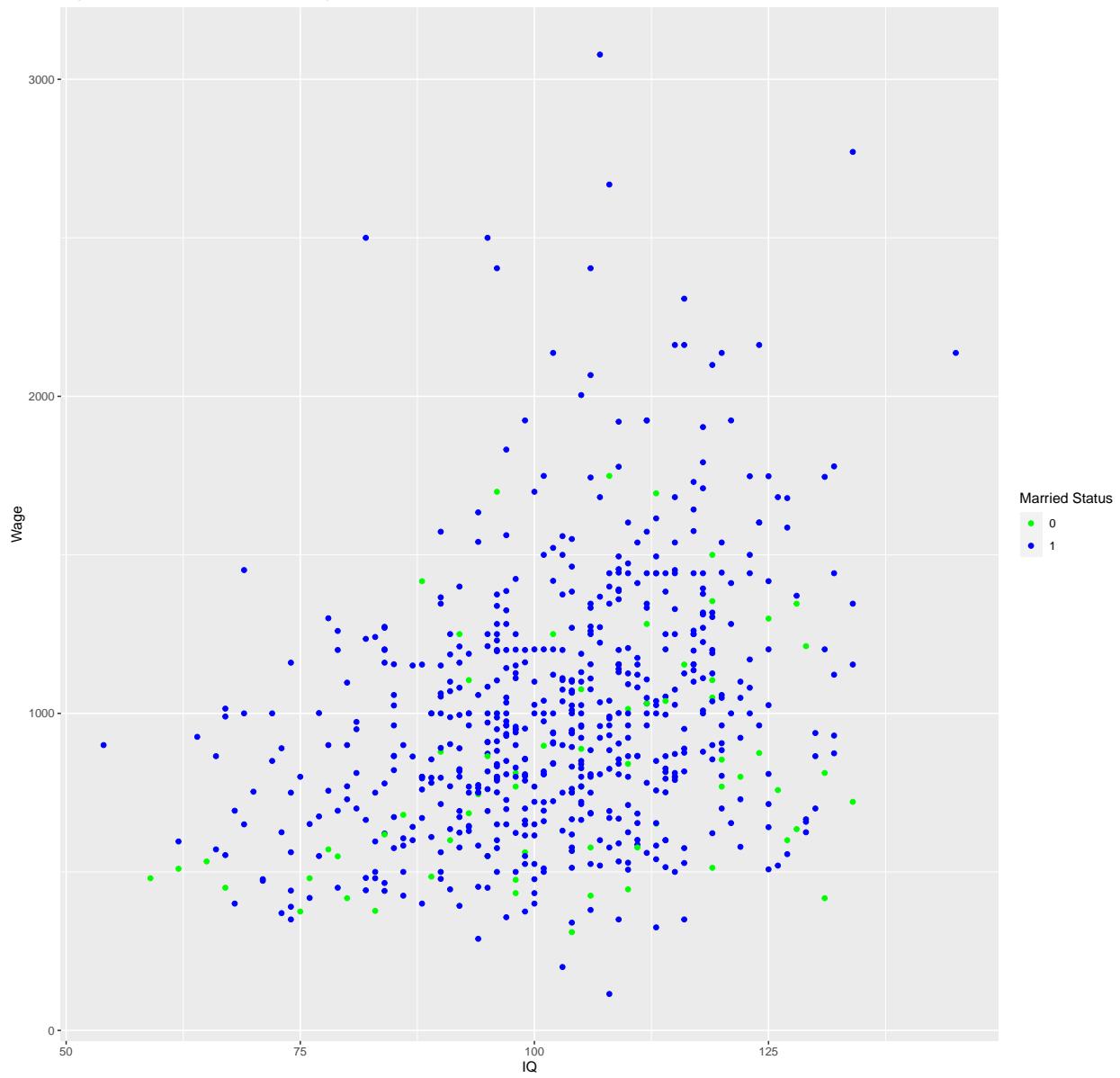


Figure 1: Dependency between IQ and Wage, distinguished by married status

Figure 2: Scatter Plot of IQ and Wage Colored By Race

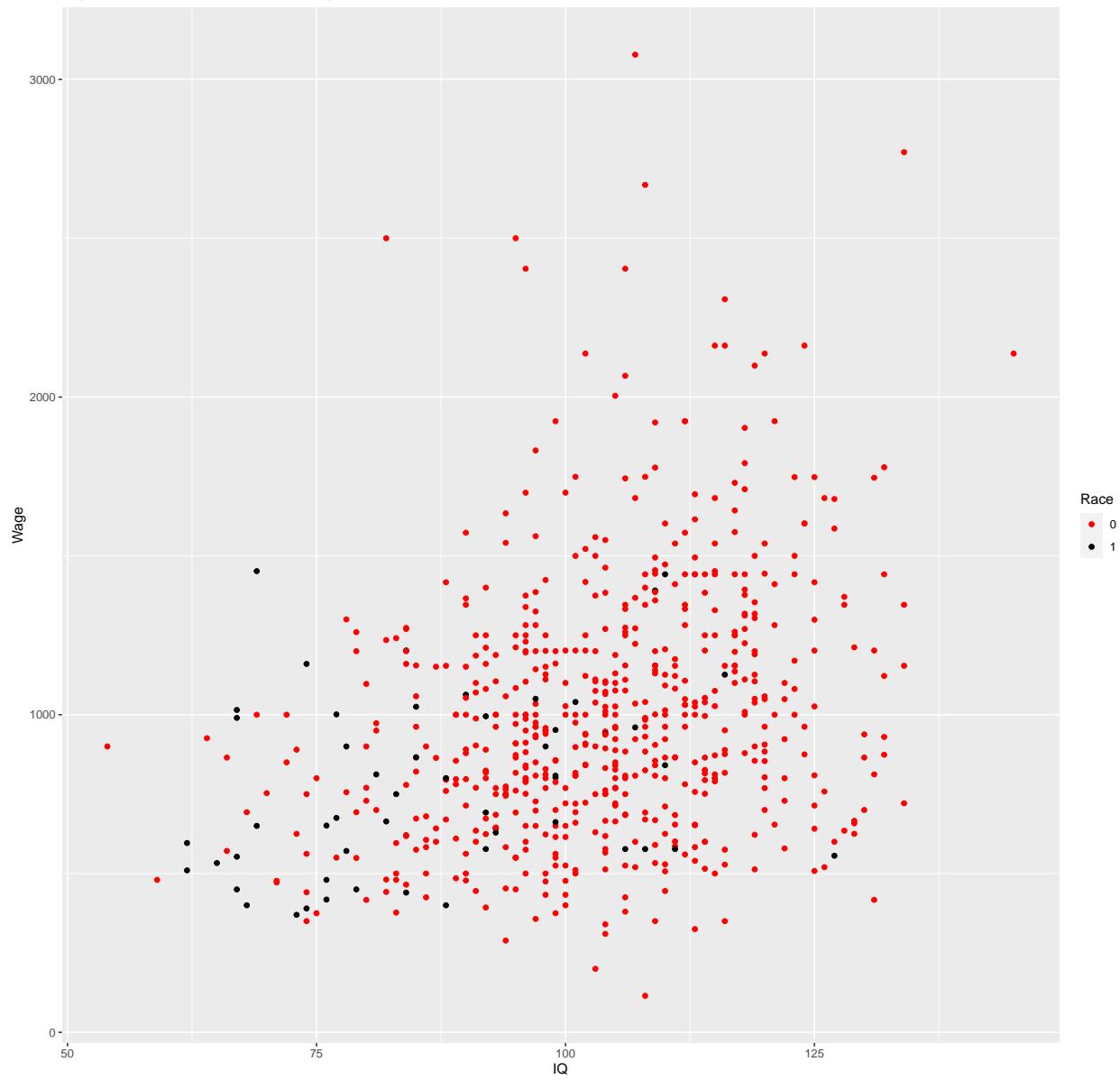


Figure 2: Dependency between IQ and Wage, distinguished by race

Figure 3: Scatter Plot of IQ and Wage Colored By Urban

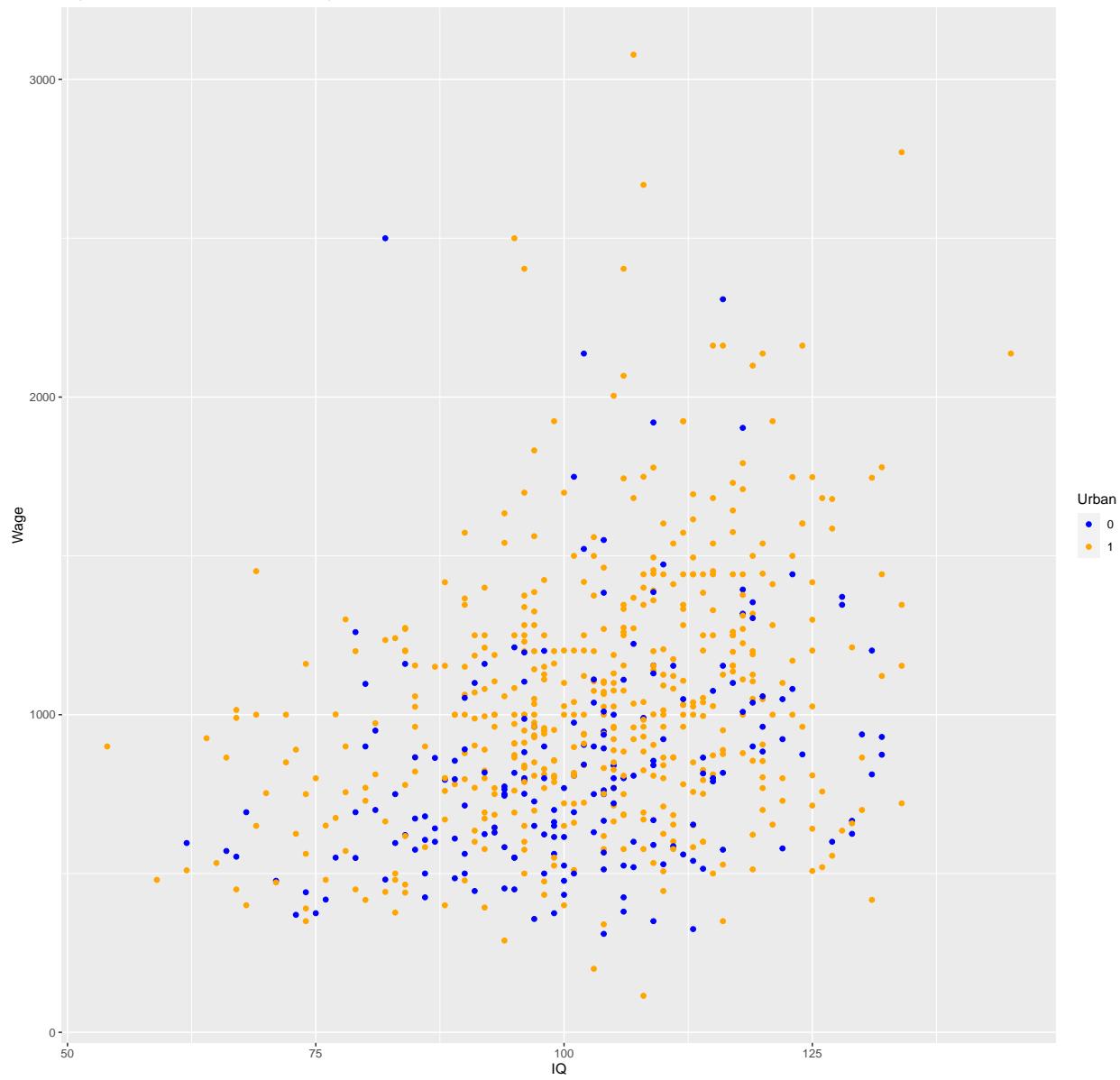
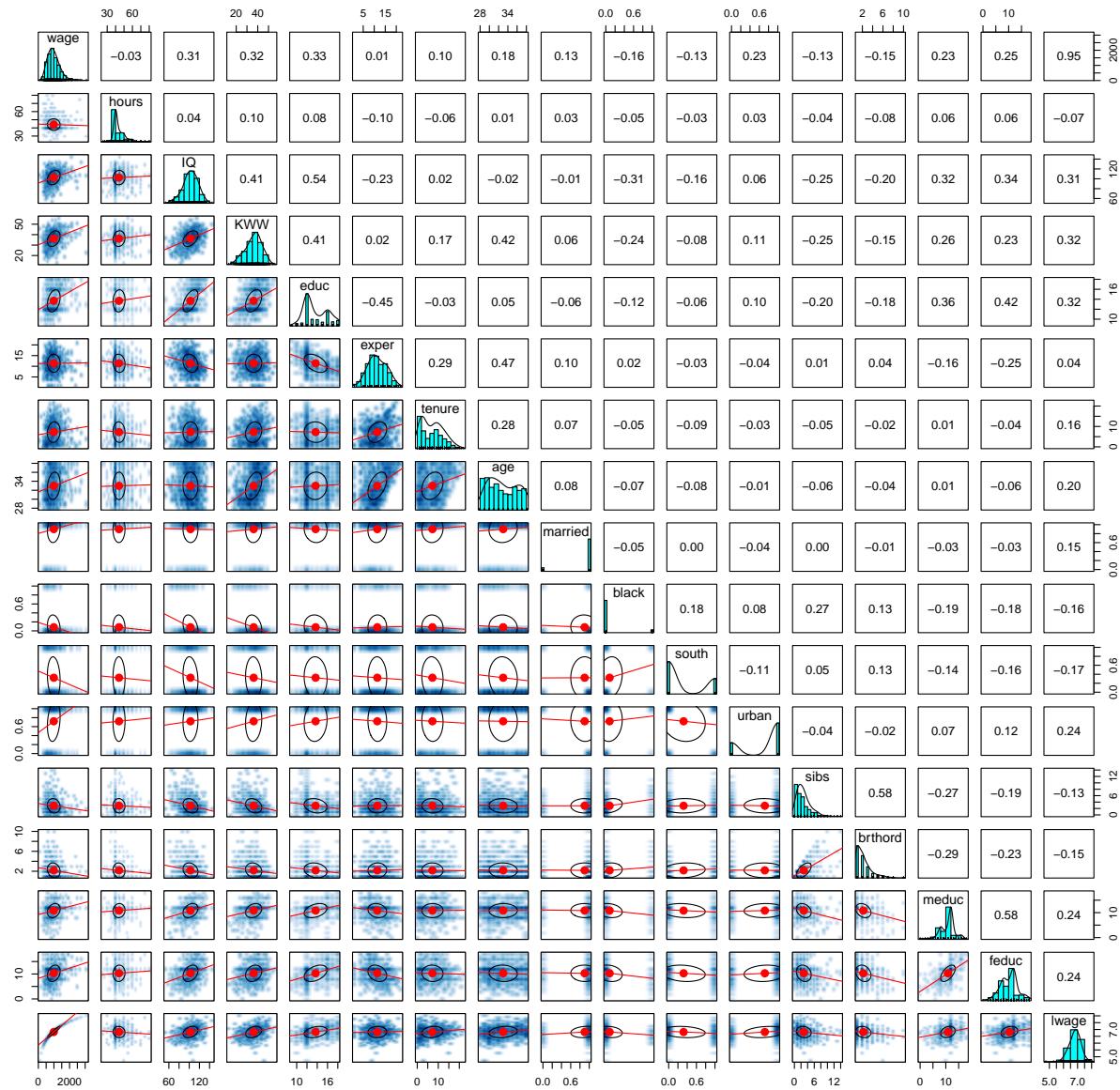


Figure 3: Dependency between IQ and Wage, distinguished by location

Figure 4: Scatter Plot Matrix for Wage2 Dataset and correlations



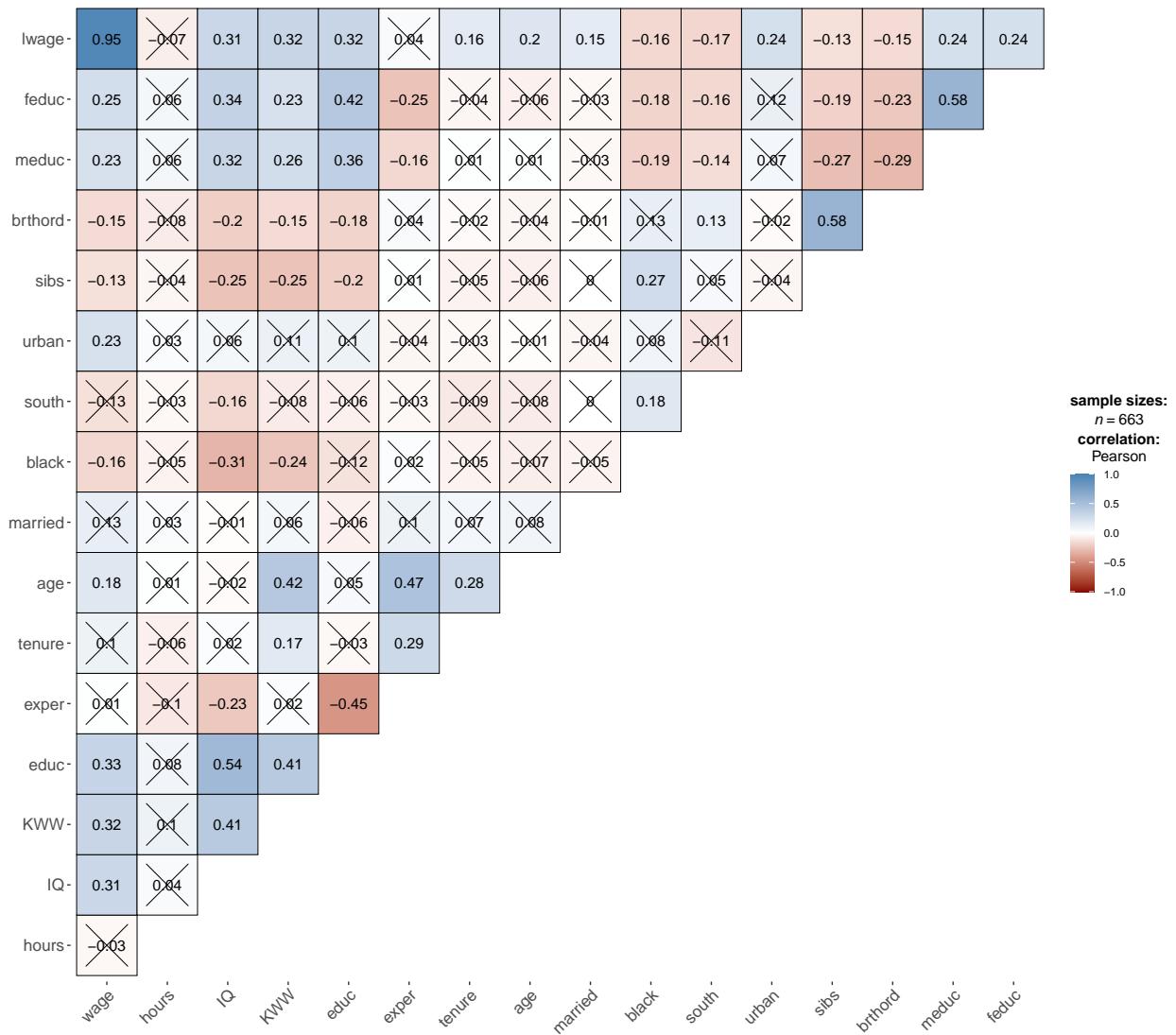


Figure 5. correlation matrix between variables. The most interesting is first row, where wage correlations are calculated between other variables
X = non-significant at $p < 0.05$ (Adjustment: Holm)

Normality

Checking the normality of the data. This block of code checks for normality of the data. Wage is skewed to left as expected, but the log of wage is more towards normal distribution

```
library('ggpubr')
par(mfrow=c(2,2))
shapiro.test(wage2$wage)
```

```
##
## Shapiro-Wilk normality test
##
```

```

## data: wage2$wage
## W = 0.93802, p-value = 5.575e-16

shapiro.test(wage2$lwage)

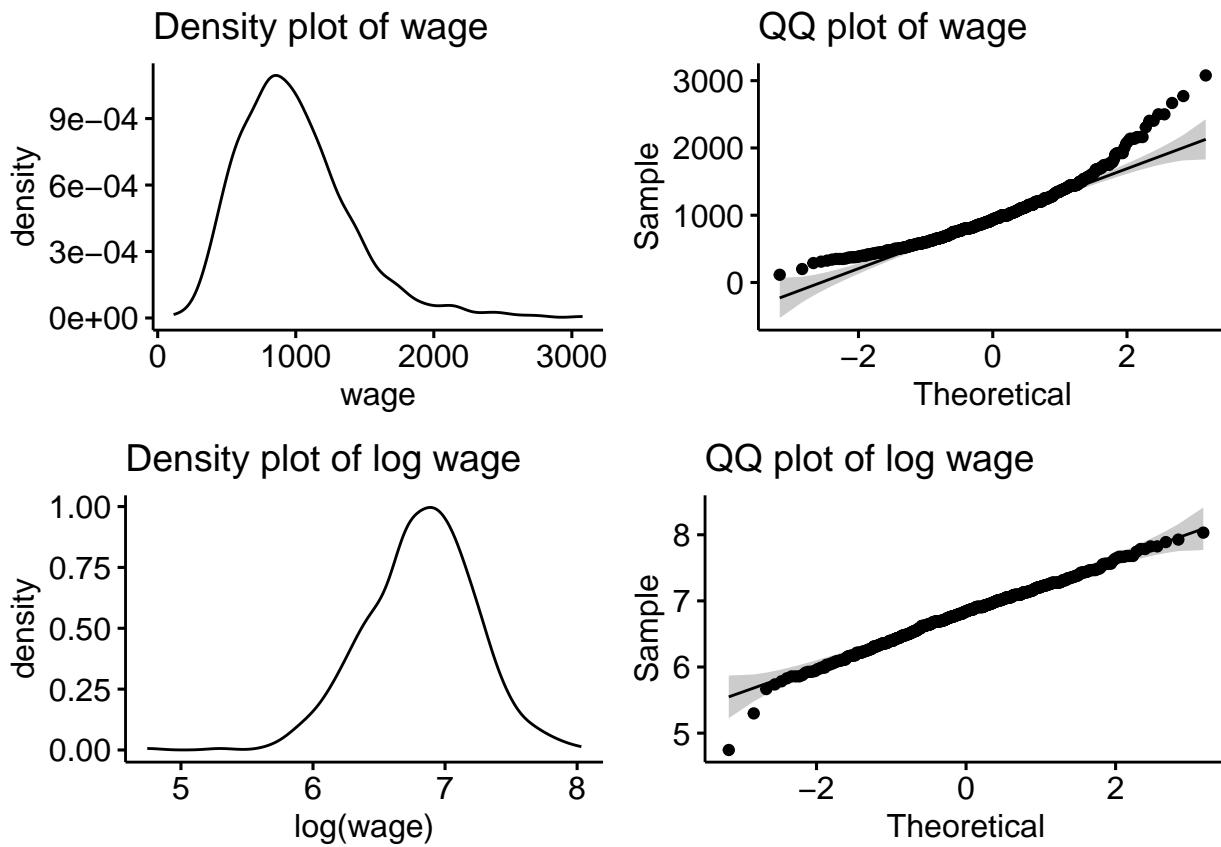
##
## Shapiro-Wilk normality test
##
## data: wage2$lwage
## W = 0.99131, p-value = 0.0006163

plot0 <- ggdensity(wage2$wage, main="Density plot of wage", xlab="wage")
plot1 <- ggqqplot(wage2$wage, title="QQ plot of wage")

plot2 <- ggdensity(wage2$lwage, main="Density plot of log wage", xlab="log(wage)")
plot3 <- ggqqplot(wage2$lwage, title="QQ plot of log wage")

ggarrange(plot0, plot1, plot2, plot3, ncol=2, nrow=2)

```



Linear regression

Creating a model with lwage as a dependent variable. Removing actual wage from the model and creating a linear regression model. Additionally, adding a stepwise selection process for more simple model with only necessary variables. The direction of stepwise is both forward and backward.

Model0:

Variables Included: Intercept, hours, IQ, KWW, educ, exper, tenure, age, married, black, south, urban, sibs, brthord, meduc, feduc.

Significant Variables (at 5% level): Intercept, hours, IQ, educ, exper, tenure, married, urban.

Adjusted R-squared: 0.2761, indicating that approximately 27.61% of the variability in the response variable is explained by the model.

Residual Standard Error: 0.3507, representing the typical difference between the observed and predicted values.

Model1: with stepwise

Variables Included: Intercept, hours, IQ, KWW, educ, exper, tenure, age, married, black, south, urban, meduc.

Variables Removed: sibs, brthord, feduc, sibs.

Significant Variables (at 5% level): Intercept, hours, IQ, educ, exper, tenure, married, urban, meduc.

Adjusted R-squared: 0.275, indicating that approximately 27.5% of the variability in the response variable is explained by the model.

Residual Standard Error: 0.351, similar to Model0.

For predictive model 27% R^2 is low, but for descriptive is is sufficient.

Model1 removes sibs, brthord, feduc, sibs.

```
# Remove the wage from the model
wage2 <- subset(wage2, select = -wage)

# Standard model
model0 <- lm(wage2$lwage~., data = wage2)

library(MASS)
# step select smaller model
fit <- lm(wage2$lwage~., data = wage2)
model1 <- stepAIC(fit, direction='both', trace=0)

summary(model0)

##
## Call:
## lm(formula = wage2$lwage ~ ., data = wage2)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -1.96887 -0.19460  0.00923  0.22401  1.34185
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.156439  0.225286 22.888 < 2e-16 ***
## hours      -0.006548  0.001934 -3.385 0.000754 ***
## IQ          0.003186  0.001223  2.604 0.009425 **
## KWW         0.003735  0.002390  1.562 0.118662
## educ        0.041267  0.008942  4.615 4.74e-06 ***
## exper       0.010749  0.004435  2.424 0.015629 *
## tenure      0.007102  0.002894  2.454 0.014401 *
```

```

## age          0.009107  0.005977  1.524 0.128058
## married     0.200760  0.045998  4.365 1.48e-05 ***
## black        -0.105141  0.055667 -1.889 0.059373 .
## south        -0.049076  0.030753 -1.596 0.111019
## urban         0.195658  0.031240  6.263 6.88e-10 ***
## sibs          0.009619  0.007876  1.221 0.222423
## brthord      -0.018465  0.011569 -1.596 0.110975
## meduc         0.009633  0.006167  1.562 0.118753
## feduc         0.005590  0.005398  1.036 0.300804
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3507 on 647 degrees of freedom
## Multiple R-squared:  0.2925, Adjusted R-squared:  0.2761
## F-statistic: 17.84 on 15 and 647 DF,  p-value: < 2.2e-16

summary(model1)

```

```

##
## Call:
## lm(formula = wage2$lwage ~ hours + IQ + KWW + educ + exper +
##       tenure + age + married + black + south + urban + meduc, data = wage2)
##
## Residuals:
##      Min      1Q      Median      3Q      Max
## -2.01294 -0.20059  0.01195  0.22524  1.32195
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.134460  0.217379 23.620 < 2e-16 ***
## hours      -0.006371  0.001932 -3.298 0.00103 **
## IQ          0.003273  0.001221  2.681 0.00753 **
## KWW         0.003546  0.002377  1.492 0.13612
## educ        0.042825  0.008799  4.867 1.42e-06 ***
## exper       0.010333  0.004424  2.336 0.01981 *
## tenure      0.006943  0.002895  2.398 0.01676 *
## age          0.009068  0.005958  1.522 0.12849
## married     0.203743  0.046006  4.429 1.11e-05 ***
## black        -0.096556  0.054384 -1.775 0.07629 .
## south        -0.057885  0.030457 -1.901 0.05781 .
## urban        0.196127  0.031141  6.298 5.55e-10 ***
## meduc        0.013480  0.005348  2.521 0.01195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.351 on 650 degrees of freedom
## Multiple R-squared:  0.2882, Adjusted R-squared:  0.275
## F-statistic: 21.93 on 12 and 650 DF,  p-value: < 2.2e-16

```

Collinearity

If higher than 5 indicates collinearity. Here we do not have this issue, but if we did, we would remove one of the variables with high VIF from the model.

```

library(car)
vif(model0)

##    hours      IQ      KWW      educ      exper      tenure      age      married
## 1.032256 1.737404 1.743251 2.142799 1.919560 1.152359 1.803937 1.022335
##    black      south      urban      sibs      brthord      meduc      feduc
## 1.249702 1.114358 1.061829 1.676644 1.594288 1.631398 1.695625

vif(model1)

##    hours      IQ      KWW      educ      exper      tenure      age      married
## 1.027821 1.727252 1.720755 2.071801 1.907037 1.151330 1.789792 1.021123
##    black      south      urban      meduc
## 1.190886 1.091371 1.053447 1.224933

```

Heteroscedasticity

Model 1 and 2 have a heteroscedasticity, the null hypothesis about Homoscedasticity is rejected. That is why recalculating robust standard errors with coefest and heteroscedasticity-robust covariance matrix estimator (HC3).

Model0:

Breusch-Pagan Test:

Test Statistic (BP): 38.914 Degrees of Freedom (df): 15 p-value: 0.00066 (indicating rejection of the null hypothesis of homoscedasticity) T-Test of Coefficients:

The table displays the estimated coefficients for each predictor variable along with their standard errors, t-values, and p-values. Variables like Intercept, hours, IQ, educ, exper, tenure, married, black, and urban have significant coefficients.

Model1:

Breusch-Pagan Test:

Test Statistic (BP): 37.256 Degrees of Freedom (df): 12 p-value: 0.000203 (indicating rejection of the null hypothesis of homoscedasticity) T-Test of Coefficients:

Similar to Model0, this table provides coefficients with standard errors, t-values, and p-values. Variables like Intercept, hours, IQ, educ, exper, tenure, married, black, south, and urban have significant coefficients.

```

bptest(model0)

##
##  studentized Breusch-Pagan test
##
## data: model0
## BP = 38.914, df = 15, p-value = 0.00066

```

```
model0.fixed <- coeftest(model0, vcov. = vcovHC(model0, type = 'HC3'))
model0.fixed
```

```
##  
## t test of coefficients:  
##  
##           Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 5.1564394 0.2211557 23.3159 < 2.2e-16 ***  
## hours      -0.0065480 0.0023551 -2.7803 0.005589 **  
## IQ          0.0031857 0.0011042  2.8851 0.004044 **  
## KWW         0.0037348 0.0027302  1.3679 0.171809  
## educ        0.0412666 0.0094044  4.3880 1.336e-05 ***  
## exper       0.0107493 0.0047579  2.2592 0.024201 *  
## tenure      0.0071016 0.0030538  2.3255 0.020353 *  
## age          0.0091075 0.0062267  1.4626 0.144050  
## married     0.2007601 0.0479763  4.1846 3.252e-05 ***  
## black        -0.1051412 0.0520714 -2.0192 0.043881 *  
## south        -0.0490755 0.0337257 -1.4551 0.146116  
## urban        0.1956579 0.0311713  6.2769 6.331e-10 ***  
## sibs          0.0096190 0.0074648  1.2886 0.198005  
## brthord      -0.0184648 0.0122953 -1.5018 0.133642  
## meduc        0.0096331 0.0059212  1.6269 0.104250  
## feduc        0.0055897 0.0055074  1.0150 0.310504  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
bptest(model1)
```

```
##  
## studentized Breusch-Pagan test  
##  
## data: model1  
## BP = 37.256, df = 12, p-value = 0.000203
```

```
model1.fixed <- model1.fixed <- coeftest(model1, vcov. = vcovHC(model1, type = 'HC3'))
model1.fixed
```

```
##  
## t test of coefficients:  
##  
##           Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 5.1344598 0.2194234 23.3998 < 2.2e-16 ***  
## hours      -0.0063708 0.0023217 -2.7440 0.006237 **  
## IQ          0.0032725 0.0011007  2.9730 0.003058 **  
## KWW         0.0035465 0.0027280  1.3001 0.194043  
## educ        0.0428251 0.0094180  4.5472 6.487e-06 ***  
## exper       0.0103329 0.0047388  2.1805 0.029578 *  
## tenure      0.0069429 0.0030515  2.2752 0.023217 *  
## age          0.0090684 0.0062250  1.4568 0.145663  
## married     0.2037429 0.0480420  4.2409 2.549e-05 ***  
## black        -0.0965560 0.0504821 -1.9127 0.056229 .  
## south        -0.0578847 0.0329506 -1.7567 0.079438 .
```

```

## urban      0.1961270  0.0311978  6.2866 5.953e-10 ***
## meduc      0.0134802  0.0049654  2.7148  0.006807 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Normality of residuals

For the first set of results:

All four tests (Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling) indicate a rejection of the null hypothesis of normality. The p-values are very close to zero.

For the second set of results:

Again, all tests suggest a rejection of the null hypothesis of normality. The p-values are low, particularly for Shapiro-Wilk and Anderson-Darling.

```

library(olsrr)

ols_test_normality(model0)

## -----
##          Test       Statistic     pvalue
## -----
## Shapiro-Wilk      0.9804    0.0000
## Kolmogorov-Smirnov 0.0383    0.2838
## Cramer-von Mises   110.23    0.0000
## Anderson-Darling    1.5062    7e-04
## -----


ols_test_normality(model1)

## -----
##          Test       Statistic     pvalue
## -----
## Shapiro-Wilk      0.9804    0.0000
## Kolmogorov-Smirnov 0.0358    0.3624
## Cramer-von Mises   109.7335   0.0000
## Anderson-Darling    1.3792    0.0014
## -----

```

Conclusion

The coefficient for “married” appeared statistically significant in both models suggesting a positive association between marital status and wages. Additionally with visual analysis and correlation calculations it has also appeared among descriptive variables.

This implies that, according to the model, being married is not merely incidental but significantly associated with higher earnings. Beyond the exploration of marital status, the top variables are education, hours, urban in the models.

Additionally, from the analysis the top variables which describe higher wages are IQ, KWW (knowledge of work), Education. The negative correlation with wages are race, south and number of siblings. From the logical point of explanation families with more siblings tend to have lower education levels, living in the south provides less opportunities and race due to economic disparities.

In conclusion, the analysis lends support to the hypothesis that married do, indeed, earn higher wages. The statistics show a strong connection, but it doesn't prove that one thing causes the other. Other factors we didn't look at could also play a role.

Appendix

```
# check the data
head(wage2)
```

```
##   hours   IQ KWW educ exper tenure age married black south urban sibs brthord
## 1    40   93  35   12   11     2   31      1     0     0     1     1     2
## 3    40  108  46   14   11     9   33      1     0     0     1     1     2
## 4    40   96  32   12   13     7   32      1     0     0     1     4     3
## 5    40   74  27   11   14     5   34      1     0     0     1    10     6
## 7    40   91  24   10   13     0   30      0     0     0     1     1     2
## 9    45  111  37   15   13     1   36      1     0     0     0     2     3
##   meduc feduc   lwage
## 1     8     8 6.645091
## 3    14    14 6.715384
## 4    12    12 6.476973
## 5     6    11 6.331502
## 7     8     8 6.396930
## 9    14     5 7.050990
```

```
# summary of data
summary(wage2)
```

```
##       hours          IQ          KWW          educ          exper
##  Min. :25.00  Min. : 54.0  Min. :13.00  Min. : 9.00  Min. : 1.0
##  1st Qu.:40.00 1st Qu.: 94.0 1st Qu.:32.00 1st Qu.:12.00 1st Qu.: 8.0
##  Median :40.00 Median :104.0 Median :37.00 Median :13.00 Median :11.0
##  Mean   :44.06 Mean   :102.5 Mean   :36.19 Mean   :13.68 Mean   :11.4
##  3rd Qu.:48.00 3rd Qu.:113.0 3rd Qu.:41.00 3rd Qu.:16.00 3rd Qu.:15.0
##  Max.   :80.00 Max.   :145.0 Max.   :56.00 Max.   :18.00 Max.   :22.0
##       tenure         age         married        black
##  Min.   : 0.000  Min.   :28.00  Min.   :0.0000  Min.   :0.00000
##  1st Qu.: 3.000 1st Qu.:30.00 1st Qu.:1.0000 1st Qu.:0.00000
##  Median : 7.000 Median :33.00  Median :1.0000  Median :0.00000
##  Mean   : 7.217 Mean   :32.98  Mean   :0.9005  Mean   :0.08145
##  3rd Qu.:11.000 3rd Qu.:36.00 3rd Qu.:1.0000 3rd Qu.:0.00000
##  Max.   :22.000 Max.   :38.00  Max.   :1.0000  Max.   :1.00000
##       south         urban         sibs        brthord
##  Min.   :0.0000  Min.   :0.0000  Min.   : 0.000  Min.   : 1.000
##  1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.: 1.000 1st Qu.: 1.000
##  Median :0.0000 Median :1.0000 Median : 2.000  Median : 2.000
##  Mean   :0.3228 Mean   :0.7195 Mean   : 2.846  Mean   : 2.178
##  3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.: 4.000 3rd Qu.: 3.000
##  Max.   :1.0000 Max.   :1.0000  Max.   :14.000 Max.   :10.000
##       meduc        feduc        lwage
##  Min.   : 0.00  Min.   : 0.00  Min.   :4.745
##  1st Qu.: 9.00 1st Qu.: 8.00 1st Qu.:6.550
##  Median :12.00 Median :11.00 Median :6.843
##  Mean   :10.83 Mean   :10.27 Mean   :6.814
##  3rd Qu.:12.00 3rd Qu.:12.00 3rd Qu.:7.090
##  Max.   :18.00 Max.   :18.00  Max.   :8.032
```

```
# description of data
library(psych)
describe(wage2)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
## hours	1	663	44.06	7.16	40.00	42.89	0.00	25.00	80.00	55.00	1.60
## IQ	2	663	102.48	14.69	104.00	103.00	13.34	54.00	145.00	91.00	-0.32
## KWW	3	663	36.19	7.53	37.00	36.46	7.41	13.00	56.00	43.00	-0.32
## educ	4	663	13.68	2.23	13.00	13.54	1.48	9.00	18.00	9.00	0.45
## exper	5	663	11.40	4.26	11.00	11.38	4.45	1.00	22.00	21.00	0.03
## tenure	6	663	7.22	5.06	7.00	6.87	5.93	0.00	22.00	22.00	0.43
## age	7	663	32.98	3.06	33.00	32.92	4.45	28.00	38.00	10.00	0.16
## married	8	663	0.90	0.30	1.00	1.00	0.00	0.00	1.00	1.00	-2.67
## black	9	663	0.08	0.27	0.00	0.00	0.00	0.00	1.00	1.00	3.05
## south	10	663	0.32	0.47	0.00	0.28	0.00	0.00	1.00	1.00	0.76
## urban	11	663	0.72	0.45	1.00	0.77	0.00	0.00	1.00	1.00	-0.97
## sibs	12	663	2.85	2.24	2.00	2.55	1.48	0.00	14.00	14.00	1.57
## brthord	13	663	2.18	1.49	2.00	1.89	1.48	1.00	10.00	9.00	1.79
## meduc	14	663	10.83	2.82	12.00	10.91	1.48	0.00	18.00	18.00	-0.59
## feduc	15	663	10.27	3.29	11.00	10.24	2.97	0.00	18.00	18.00	-0.01
## lwage	16	663	6.81	0.41	6.84	6.82	0.38	4.74	8.03	3.29	-0.32
##			kurtosis	se							
## hours			3.69	0.28							
## IQ			0.02	0.57							
## KWW			-0.29	0.29							
## educ			-0.99	0.09							
## exper			-0.47	0.17							
## tenure			-0.75	0.20							
## age			-1.24	0.12							
## married			5.13	0.01							
## black			7.34	0.01							
## south			-1.43	0.02							
## urban			-1.05	0.02							
## sibs			3.55	0.09							
## brthord			3.85	0.06							
## meduc			1.25	0.11							
## feduc			-0.03	0.13							
## lwage			0.83	0.02							