Stat 333 Project

Wages and Hours

Qiuying(Autumn) Li

12/15/2015

I. Introduction

The data are from a national sample of 6000 households with a male head earning less than $15,000 annually in 1966. The data were classified into 39 demographic groups for analysis. The study was undertaken in the context of proposals for a guaranteed annual wage (negative income tax). At issue was the response of labor supply (average hours) to increasing hourly wages. The study was undertaken to estimate this response from available data.

Variable	Description
HRS	Average hours worked during the year
RACE	Percent of white respondents
WAGE	Average hourly wage ($)
ERSP	Average yearly earnings of spouse ($)
SCHOOL	Average highest grade of school completed
ERNO	Average yearly earnings of other family members
NEIN	Average yearly non-earned income
ASSET	Average family asset holdings (Bank account, etc.)
AGE	Average age of respondent
DEP	Average number of dependents

Table 1. List of the initial variables

II. Background Information

During the 1960's the United States experienced its longest uninterrupted period of economic expansion in history. In the 1960's housing and computer industry overpowered automobiles, chemicals, and electrically powered consumer durables, which were the leading sectors in the 1950s. As the development of the economics, the rate of the wage increase as well.

At the same time, United States also experienced African-American Civil Rights Movement, Hispanic and Chicano Movement and labor right movements. This counterrevolution brought revolutions among the states and deeply divided the dynamics of politics and the society in the 50 states. As the changes of social, political and economic systems became more intensive than ever, wage, working hours and the race issue are interesting topics in 1960's[1].

III. Scientific Underpinnings

Among the all the variables of the collected data, the relationship between the amount of the working hour and the rate of the wage is interesting to me . I wonder is there any causality relationship between them, in other words, would the higher rate of the wage cause the the increase of the working hour.

I collected this data by using google search, and I found this data at statistic library of Carnegie Mellon University.

The data are from a national sample of 6000 households with a male head earning less than $15,000 annually in 1966. The U.S. Department of Labor selected 6000 random sample from the list of all the male head earning less than $15,000 annually in 1966. Collecting data from a random sample of people might have required traveling to home all over the country, which would have been time consuming and expensive, so U.S. Department of Labor decided to collect data with telephone, which was cheaper and convenient. This is an observational study, in which the researcher dose not actively control the value of any variables but simply observe the values as they naturally exist.

The purpose of finding causality relationship between rate of wage and working hours, is to provide guidance for legislation department, employers and employees guidance to balance the rate of wage and amount of the woking hour. Legislation department is able to enact the legal working time and the minimum working wage, employers are capable of offering the attractive wage level to job applicants , and employees could modify their working hours in order to make enough money and have enough leisure time.

IV **Analytical Methods and Interpretation** :

A. Simple linear regression of HRS and WAGE

A.1 Plotting scatterplot of HRS versus WAGE

The two scatterplot of the wage and working hours show the if there is linearity relation between the two variables.However, correlation can be heavily influenced by outliers. By plotting Figure.1, there are some outliers far away from the cloud of the data set, then I removed these outliers and get Figure 2, even though there is a little departure from the ideal pattern, the scatterplot for WAGE and HRS is good, there is no sign of curvature and most of the points are fairly distributed on the either side of the line with constant among of the variability.

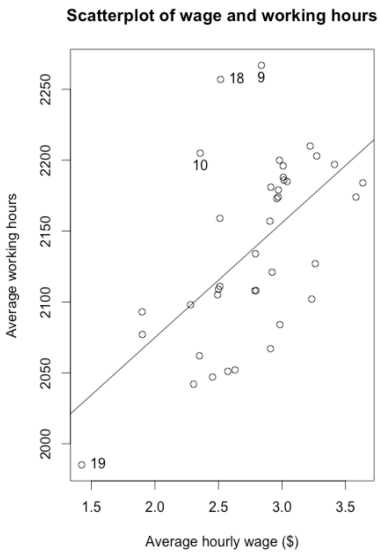


Figure.1

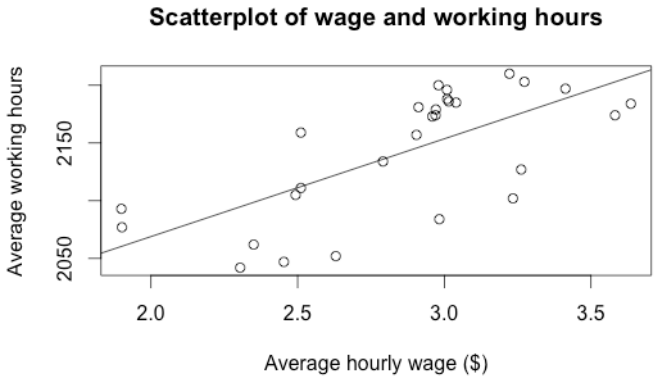


Figure. 2

A. 2 Fitting simple linear regression in R

The response variable HRS can be predicted by a linear function of a regressor variable WAGE. I can estimate $\hat{HRS}$, the intercept $\hat{\beta}_0$, and the slop $\hat{\beta}_1$: $\hat{HRS} = \hat{\beta}_0 + \hat{\beta}_1 * WAGE$

lm(formula = HRS ~ WAGE)				
	stimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**1913.01**	52.89	36.167	< 2e-16 ***
WAGE	**80.94**	18.83	4.298	**0.00012** ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 52.99 on 37 degrees of freedom				
Multiple R-squared: **0.333,** Adjusted R-squared: 0.3149				
F-statistic: 18.47 on 1 and 37 DF, p-value: 0.0001203				
$\hat{HRS}$=1913.01+80.94*WAGE.				

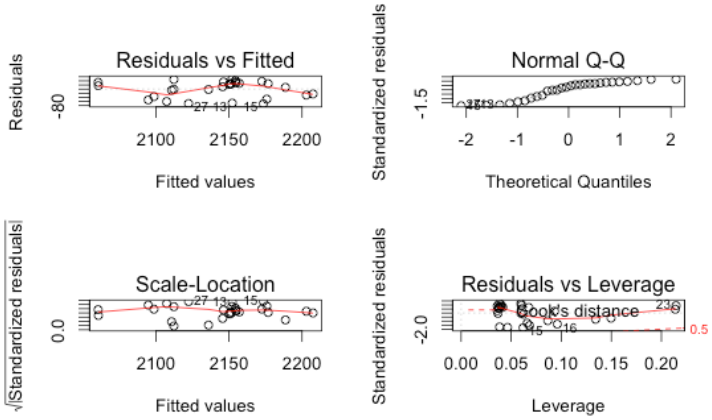
A. 3 Interpreting the results

From that output we see have the estimates of coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, their standard errors, the value of the coefficient of determination $R^2 = 0.3333$, which means the 33.33% variability in the HRS of this sample is explained by WAGE. The variance estimate 52.99. The hypothesis test that E(HRS | WAGE) = β_0 versus the alternative that E(HRS | WAGE) = $\beta_0 + \beta_1$, HRS is summarized by the F-statistic which in this case is given by 18.47and has a p-value less than 2e-16 or strong evidence against the hypothesis that E(HRS | WAGE) = β_0 in favor of the other hypothesis. The intercept $\beta_0 = 1913.01$ and the slope is β_1=80.94 to produce the least squares prediction equation : $\widehat{HRS}$=1913.01+80.94*WAGE.

The slope $\hat{\beta}_1 = 80.94$ indicates that HRS is predicted to go up by about 80.94 for one dollar increase in WAGE.

The intercept $\beta_0 = 1913.01$ indicates that the HRS will be 1913.01 if the WAGE is $0.

A. 4 Residual Analysis



a.Linearity

Checking plot of residuals versus predicted values, there are a point locating far away from the 0 line, but in my opinion, it is not necessary to be considered as potential outliers, and most of the points are symmetrically distributed around horizontal line in the residual plot;

b. Homoscedasticity

look at a plot of residuals versus predicted values, even though there is a little departure from the ideal pattern, there is a constant variance around the horizontal line in general ;

c. Normality

By looking at QQ-plot, the distribution is normal because the points on such a plot should fall close to the diagonal reference line.

d.Residual versus leverage : There is no points with high cook distance (>0.5); we note that one point has relatively high leverage. I have removed it, see Appendix 1.

B. Accounting for Confounding Variables

Even though there is a positive correlation between wage and working hour, it dose not mean that there is a causal relation between these two variable. This data set is from a random observational study, it is hard to avoid confounding variables in observational studies. For this reason, observational studies can almost never be used to establish causality. Multiple linear regression provides a powerful way to account for confounding variables by including them as additional explanatory variables in the model.

B.1 Potential Confounding Variable - RACE

The United States experienced African-American Civil Rights Movement in 1954– 1968, noted legislative achievements during this phase of the Civil Rights Movement were passage of the Civil Rights Act of 1964, that banned discrimination based on "race, color, religion, or national origin" in employment practices and public accommodations[2]. The data were collected in 1966, so it may be interesting to think if Civil Rights Act of 1964 had true effect between race problems and employment in 1966, in other words, I wonder could RACE(Percent of white respondents) be a potential confounding variable which has association with both HRS and WAGE.

B1.1 Checking association between HRS&RACE, and WAGE$RACE

P-value of lm(HRS~RACE)
0.000484
Small p-value shows a significant association between HRS and RACE

> cor(HRS, RACE)	> cor(WAGE, RACE)
-0.5325138	-0.3866377
Higher RACE has negative correlation with HRS	Higher RACE has negative correlation with WAGE

In a word, Since RACE have a significant association withHRS, and at the same time, RACE has association with both HRS and WAGE. To test whether there is a causality relation between HRS and WAGE, even after accounting for variable RACE. By including both RACE and WAGE as explanatory variable in a multiple regression model.

B.1.2 Fitting multiple linear regression model in R

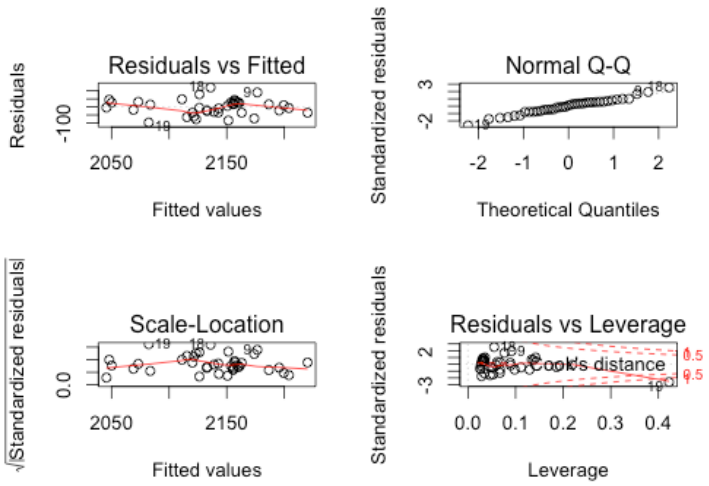
Contrast of Two models

Model 2: lm(formula = HRS ~ WAGE + RACE)					**Model 1: lm(formula = HRS ~ WAGE)**				
	stimate	Std. Error	t value	Pr(>|t|)		stimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2015.8010	61.9359	32.547	< 2e-16	(Intercept)	1913.01	52.89	36.167	**< 2e-16**
WAGE	61.2078	18.8752	3.243	**0.00255**	**WAGE**	80.94	18.83	4.298	**0.00012**
RACE	-1.3385	0.4951	-2.703	**0.01041**					
Residual standard error: 48.98 on 36 degrees of freedom					Residual standard error: 52.99 on 37 degrees of				
Multiple R-squared: 0.4455, Adjusted R-squared: 0.4147					Multiple R-squared: 0.333; Adjusted R-squared: 0.3149				
F-statistic: 14.46 on 2 and 36 DF, p-value: 2.454e-05					F-statistic: 18.47 on 1 and 37 DF, p-value: 0.0001203				
$\widehat{HRS} = 2015.80+61.2*WAGE-1.33*RACE$					$\widehat{HRS} = 1913.01+80.94*WAGE$				

B.1.3 Interpreting results of Model 2 and contrast of two models

From the table of Model 2, the p-values for the coefficients of WAGE and RACE are both very small, so there are strong evidence that both terms are important in model 2. R square is 0.4455, which 44.55% variability in the HRS of this sample is explained by WAGE and RACE.

In model 2, the $\beta_0 = 2015.80$ indicates that HRSA will be 2015.80 if both WAGE is $0 and RACE is 0. The coefficient of WAGE is 61.2078. If holding RACE fixed, and if the WAGE increase by $1, we expect the the amount of the HRS to increase by 61.2078; The coefficient of RACE is -1.3385. If let the WAGE stayed exactly the same, if the RACE increase by 1 percent, we expect the amount of the HRS to decrease 1.3385.



Based on the contraction of two models, after accounting for whether or not include RACE, both R-square(from 0.33 to 0.45) and Adjusted R(0.31 to 0.41) increase; Residual standard error decrease (decrease from 52.99 to 48.98), and both WAGE and RACE are significant, which might indicate that RACE could a additional predictor. However, this result also indicates it could be other potential confounding variables in the data. As a result, it is essential to continue finding other potential confounding variables in the rest of the variables.

B. 1. 4 Residual Analysis

a.Linearity

Checking plot of residuals versus predicted values, there are some points (#18,#9 and #19) are a little far away from the horizontal line, but most of the points are symmetrically distributed around horizontal line in the residual plot;

b. Homoscedasticity

Looking at a plot of residuals versus predicted values, there is a constant variance around the horizontal line;

c. Normality

By looking at Q-Q plot, the distribution is normal because the points on such a plot should fall close to the diagonal reference line.

d. Residual versus Leverage

There is two points with high cook distance (>0.5); I noted that 2 points have relatively high leverage. I have removed these points, see Appendix 2

B.2 Potential Confounding Variable - ASSET

ASSET (Average family asset holdings (Bank account, etc.)) could be another potential confounding variable for HRS and rate of WAGE. First of all, bank account is an obvious indicator of the wealth, the more bank deposits a person has, the richer the person is. In other words, rich people might have higher wage than poor people. However, some people might not have very high wage, they can improve wealth by increase working hours, the longer time they work, they more money they could have in the bank. In a word, ASSET might have correlation with rate of wages, it also might have correlation with working hours.

P-value of the Linear regression model of HRS&ASSET
lm(HRS~ASSET)
p-value <4.25e-07
HRS and ASSET are significantly associated

B.2.1 Finding association between HRS&ASSET, and WAGE&ASSET

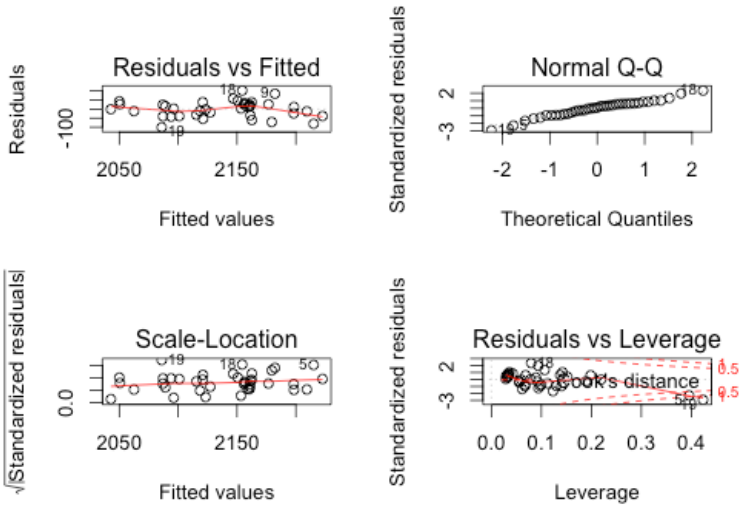
>cor(HRS, ASSET)	>cor(WAGE, ASSET)
>0.7095824	>0.777595
Positive correlation between HRS and ASSET	Positive correlation between WAGE and ASSET

Since ASSET have correlation with both HRS and WAGE, in addition, HRS and ASSET are significantly associated, I account for ASSET by including it in the model as an additional explanatory variable, with the relevant output below:

B.2.2 Fitting the multiple regression model and make contrast of two models

Contrast of Two models								
Model 3: lm(formula = HRS ~ WAGE + RACE+ASSET)					**Model 2: lm(formula = HRS ~ WAGE+RACE)**			
	stimate	Std. Error	t value	Pr(>|t|)		stimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.059e+03	5.985e+01	34.406	<2e-16	(Intercept)	2015.8010	61.9359	32.547	< 2e-16
WAGE	1.195e+01	2.577e+01	0.464	**0.6457**	**WAGE**	80.94	18.83	4.298	**0.00012**
RACE	-7.712e-01	5.085e-01	-1.517	**0.1383**	**RACE**	-1.3385	0.4951	-2.703	**0.01041**
ASSET	1.161e-02	4.455e-03	2.606	**0.0134 ***					
Residual standard error: 45.46 on 35 degrees of freedom					Residual standard error: 48.98 on 36 df				
Multiple R-squared: 0.5356, Adjusted R-squared: 0.4958					F-statistic: 18.47 on 1 and 37 DF, p-value: 0.0001203				
F-statistic: 13.46 on 3 and 35 DF, p-value: 5.337e-06					F-statistic: 14.46 on 2 and 36 DF, p-value: 2.454e-05				

B.2.3 Interpreting the Model 3 and results of contrast



After accounting for ASSET, WAGE (p-value = 0.6457) and RACE(p-value=0.1383) are no longer significant predictors of HRS in model 3. In other words, based solely on this dataset, ASSET is the real confounding variable, and we do not have significant evidence to support that there is a causality relation between HRS and WAGE.

B. 2.4 Residual Analysis :

a.Linearity

Checking plot of residuals versus predicted values, most of the points are symmetrically distributed around horizontal line in the residual plot;

b. Homoscedasticity

Looking at a plot of residuals versus predicted values, there is a constant variance around the horizontal line;

c. Normality

By looking at QQ-plot, the distribution is normal because the points on such a plot should fall close to the diagonal reference line.

d. Residual versus Leverage

There is three points with high cook distance (>0.5); I note that three points have relatively high leverage. I have removed these points.See Appendix 3.

V. Conclusion and Further Studies

I was tempting to find the causal relation between HRS and WAGE based on this data set. However, according to the analysis, there is no causal relation between HRS and WAGE. First of all, I find the significant association between HRS and WAGE, and then I accounted the confounding variables by using multiple linear regression. In this process, I find WAGE and

RACE can be predictors to explain the 44.5%variability in HRS. However, when I add ASSET in the model, both WAGE and RACE are not significant anymore, but ASSET become significant. Thus, ASSET is the real confounding variable, there is no causal relation between HRS and WAGE in this dataset, which indicates that people who have higher wage would not cause longer working hours.

Based solely on this data, ASSET is the the real confounding variable, and it plays the major role to predict the the variability of the HRS, which indicates that people have more ASSET tend to work longer, and rich people will become richer. This phenomenon can be related to Matthew effect in sociology, the Matthew effect is the phenomenon where "the rich get richer and the poor get poorer". In both its original and typical usage it is meant metaphorically to refer to issues of fame or status but it may also be used literally to refer to cumulative advantage of economic capital [3].

It may be tempting to make to make causal conclusion once we have accounted all the confounding variables, multiple linear regression can only allows us to account for confounding variables in the data, there could be other confounding variable outside of the data set.

VI. Limitations of the analysis:

a.The data were classified into 39 demographic groups for analysis, some data might be deleted in the process of the classification. I deleted several missing values; In addition, it might be a subset of a much larger data set, and running the same model on the larger data set might have the different results.

b.In order to save time and money, US labor department collected data by telephone survey. The telephone survey in 1966 reached only people wealthy enough to own a telephone, causing the sample to be wealthier than the population, which might cause the sample bias. What is more, some people own the phones but they are not willing to answer the questions,and some people might provide false information.

c. In the residual analysis, all three models satisfy the assumption of linearity, homoscedasticity and normality. However, since the data is not time series, it is hard to check independence assumption by checking statistical analysis.

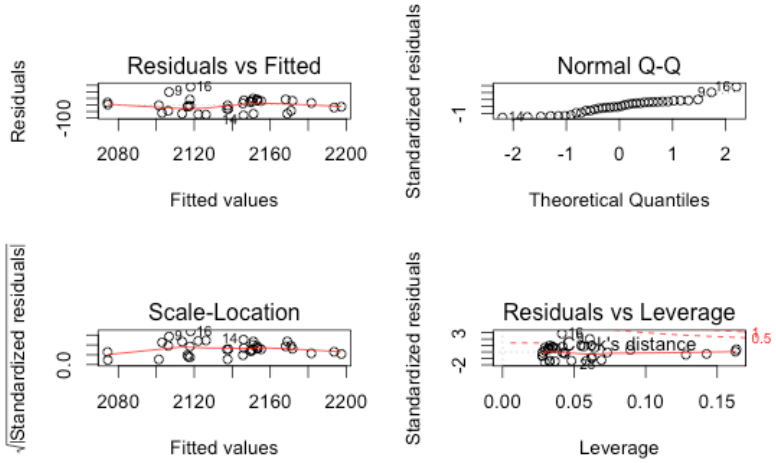
Works Cited

1.”The 1960s." History.com. *A&E Television Networks*, n.d. Web. 24 Apr. 2014.

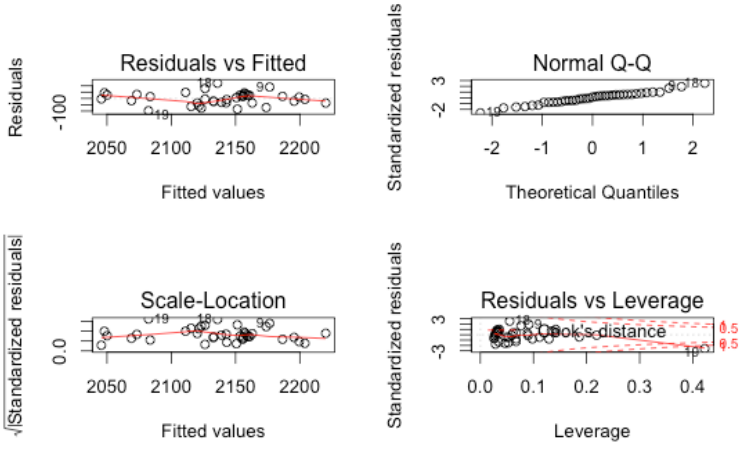
2.African-American Civil Rights Movement (1954–68), retrieved from http://en.wikipedia.org/wiki/African-American_Civil_Rights_Movement_%281954%E2%80%9368%29

3. Matthew effect, retrieved from http://en.wikipedia.org/wiki/Matthew_effect

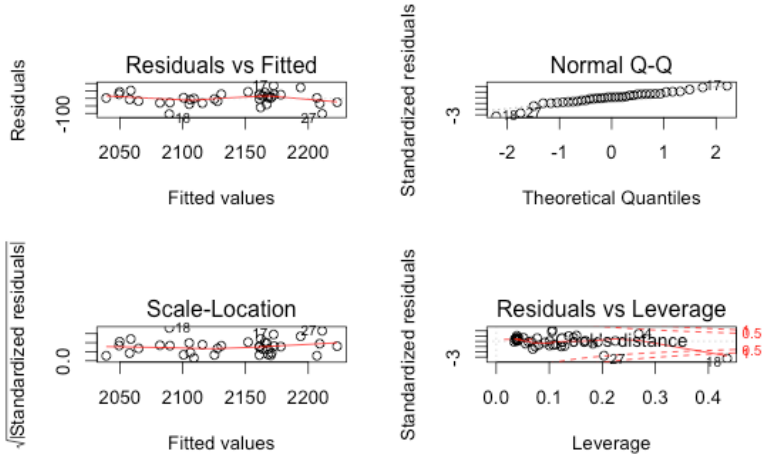
Appendix 1



Appendix 2



Appendix 3


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage