Factors Affecting Wage: A General Linear Model Analysis

By Yimi Zhao

1 Objective

In this report, we wish to determine **whether wages are related to the characteristics stated in this dataset**. We try to figure out the answer to this problem through various general linear model data analysis strategies, including multiple regression, variable transformation, collinearity, categorical variables, ANOVA, tests of significance, model building strategies and interpretation.

2 About the Dataset

The Current Population Survey (CPS) is used to supplement census information between census years. For over 60 years, the Census Bureau has conducted the CPS for the Bureau of Labor Statistics as a source of data on employment and unemployment. Each month the survey contacts about 50,000 households and collects basic data on the characteristics of households and their labor force status: employment, job search, occupation, weeks worked, hours worked last week, etc. In most months, the survey asks supplemental questions on a variety of other subjects: income, poverty, education, migration, etc.

The dataset involved in this report contains 534 observations on 11 variables sampled from the Current Population Survey of 1985 with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership. The dataset format is:

- **wage:** Wage (in dollars per hour).
- **education:** Number of years of education.
- **experience:** Number of years of potential work experience (age - education – 6).
- **age:** Age in years.
- **ethnicity:** Factor with levels "cauc", "hispanic", "other".
- **region:** Factor. Does the individual live in the South?
- **gender:** Factor indicating gender.
- **occupation:** Factor with levels "worker" (tradesperson or assembly line worker), "technical" (technical or professional worker), "services" (service worker), "office" (office and cleri- cal worker), "sales" (sales worker), "management" (management and administration).
- **sector:** Factor with levels"manufacturing"(manufacturing or mining),"construction","other".
- **union:** Factor. Does the individual work on a union job?
- **married:** Factor. Is the individual married?

Load the dataset from the source:

```
#load the dataset
>library(AER)
>data(CPS1985)
```

3 Data Analysis

3.1 Data Summary

The summary of the data shows close median and mean values of wage, education, experience and age. And both wage and experience vary a lot, which indicates a sign of data skewness.

> summary(CPS1985)

```
      wage          education       experience         age           ethnicity     occupation
 Min.   : 1.000   Min.   : 2.00   Min.   : 0.00   Min.   :18.00   cauc    :440   worker    :156
 1st Qu.: 5.250   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:28.00   hispanic: 27   technical :105
 Median : 7.780   Median :12.00   Median :15.00   Median :35.00   other   : 67   services  : 83
 Mean   : 9.024   Mean   :13.02   Mean   :17.82   Mean   :36.83                  office    : 97
 3rd Qu.:11.250   3rd Qu.:15.00   3rd Qu.:26.00   3rd Qu.:44.00                  sales     : 38
 Max.   :44.500   Max.   :18.00   Max.   :55.00   Max.   :64.00                  management: 55
   region       gender             sector        union     married
 south:156   male  :289   manufacturing: 99   no :438   no :184
 other:378   female:245   construction : 24   yes: 96   yes:350
                          other        :411
```

Figure 1 wage vs each variable

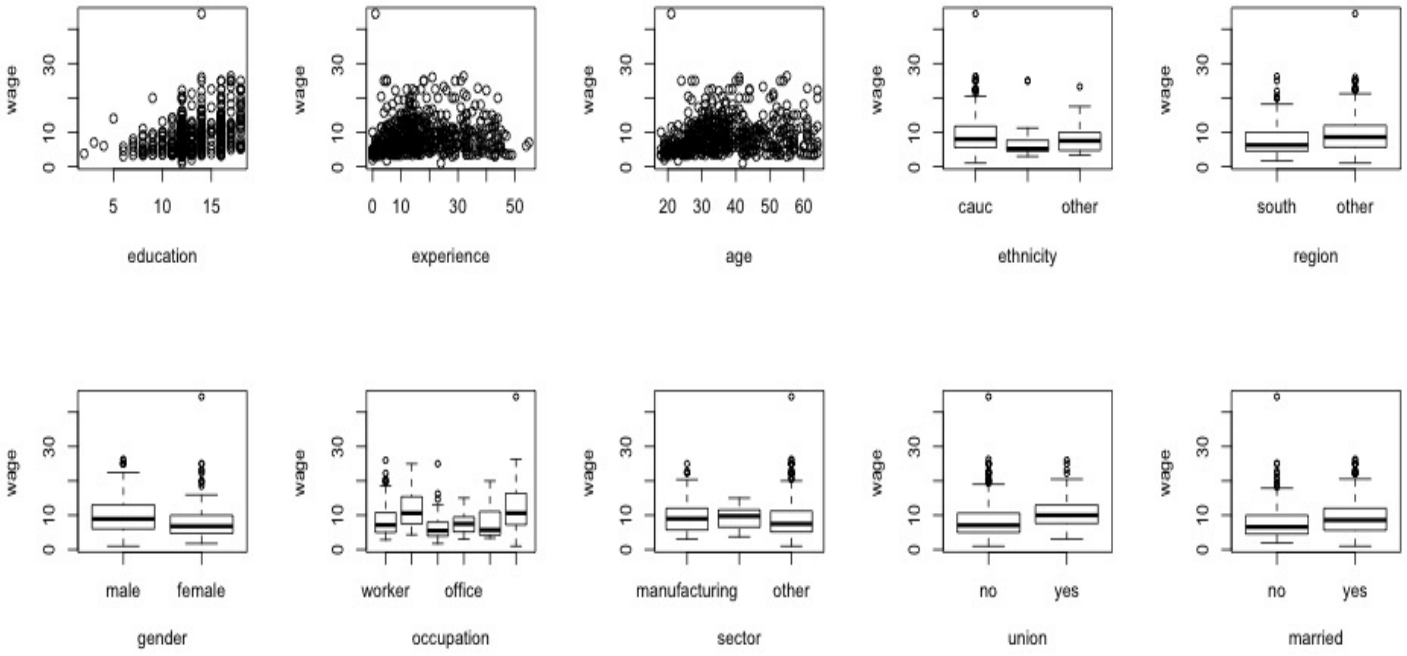
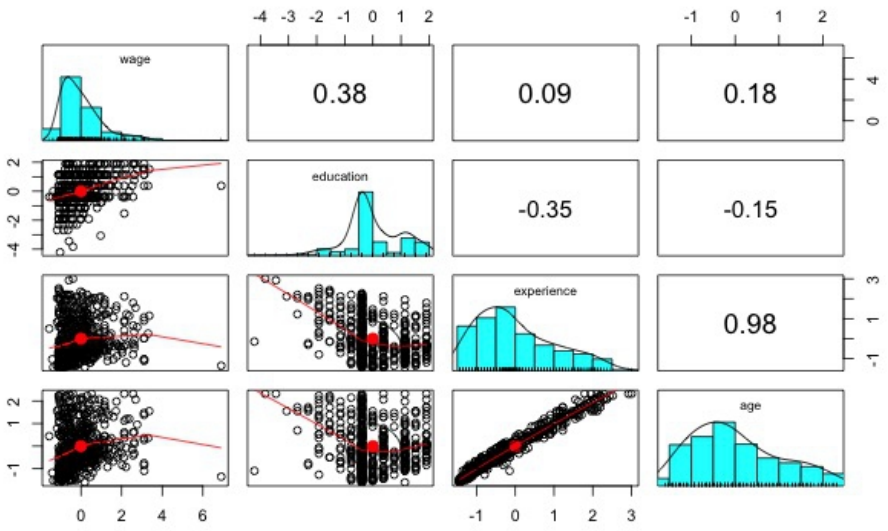


Figure 1 displays the relationship between wage and each of the other 10 variables. Clearly, we can tell from the plots:

- wage rises as education or experience increase;
- wage varies a lot among different occupations;
- wage varies slightly based on gender, union, married or not and region;
- outlier exists.

Figure 2 correlation plot among four numeric variables



Furthermore, Figure 2 shows age and work experience were almost perfectly correlated (corr=0.98). This also means collinearity, therefore, we are going to use only experience to represent the effect of both age and experience in the following model fitting.

3.2 Model Fitting and Regression Diagnostics

First, we fit a full model, using wage as the response and all other variables as predictors. However, the summary is beyond our expectation. Only gender, occupation and union are significant at 5% level, while neither education nor experience is significant. As common sense, education and experience are definitely related to wage rate. Therefore, we have to check our model and improve it.

```
> full0<-lm(wage~.,data=CPS1985)
Coefficients:          Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.30356    6.61313   0.046  0.96341
education               0.81279    1.08688   0.748  0.45491
experience              0.24485    1.08178   0.226  0.82103
age                    -0.15803    1.08092  -0.146  0.88382
ethnicityhispanic      -0.60651    0.86963  -0.697  0.48584
ethnicityother         -0.83787    0.57450  -1.458  0.14532
regionother             0.56274    0.41982   1.340  0.18070
genderfemale           -1.94252    0.41943  -4.631 4.60e-06 ***
occupationtechnical     1.95695    0.73319   2.669  0.00784 **
occupationservices     -0.68485    0.67900  -1.009  0.31363
occupationoffice        0.02232    0.68588   0.033  0.97405
occupationsales        -0.77332    0.85332  -0.906  0.36523
occupationmanagement    3.29052    0.80050   4.111 4.59e-05 ***
sectorconstruction     -0.56349    0.99154  -0.568  0.57008
sectorother            -1.04089    0.54923  -1.895  0.05863 .
unionyes                1.60169    0.51272   3.124  0.00188 **
marriedyes              0.30050    0.41120   0.731  0.46523
---
Residual standard error: 4.282 on 517 degrees of freedom
Multiple R-squared: 0.3265,      Adjusted R-squared: 0.3056
F-statistic: 15.66 on 16 and 517 DF,  p-value: < 2.2e-16
```

From the previous data summary section, we found that age and experience are highly correlated. Thus, we remove “age” and only keep “experience” to eliminate the effect of collinearity. In this new model, education and experience become significant as we expected, and the R-squared and adjusted R-squared almost keep the same.

```
# remove age(eliminate collinearity)
> full<-lm(wage~.-age, CPS1985)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.64046    1.42667  -0.449  0.65368
education             0.65458    0.10061   6.506 1.82e-10 ***
experience            0.08671    0.01726   5.025 6.94e-07 ***
ethnicityhispanic    -0.60811    0.86874  -0.700  0.48425
ethnicityother       -0.83819    0.57395  -1.460  0.14479
regionother           0.56404    0.41933   1.345  0.17919
genderfemale         -1.93933    0.41846  -4.634 4.54e-06 ***
occupationtechnical   1.96211    0.73164   2.682  0.00756 **
occupationservices   -0.68658    0.67825  -1.012  0.31188
occupationoffice      0.02001    0.68505   0.029  0.97671
occupationsales      -0.77399    0.85251  -0.908  0.36435
occupationmanagement  3.28997    0.79973   4.114 4.53e-05 ***
sectorconstruction   -0.56175    0.99053  -0.567  0.57088
sectorother          -1.03964    0.54865  -1.895  0.05866 .
unionyes              1.60095    0.51221   3.126  0.00187 **
marriedyes            0.29743    0.41027   0.725  0.46881
---
Residual standard error: 4.278 on 518 degrees of freedom
Multiple R-squared:  0.3265, Adjusted R-squared: 0.307
F-statistic: 16.74 on 15 and 518 DF,  p-value: < 2.2e-16
```

Next, we are going to do the Regression Diagnostics by checking the assumptions. According to the diagnostic plots, we found non-constant variance of residuals(errors), non-normality of residuals (errors) and outliers.

```
# outlier 1
     wage education experience age ethnicity region gender occupation  sector union  married
170  44.5    14         1       21      cauc   other female  management  other    no      no
```

Figure 3 Diagnostic plots of full model

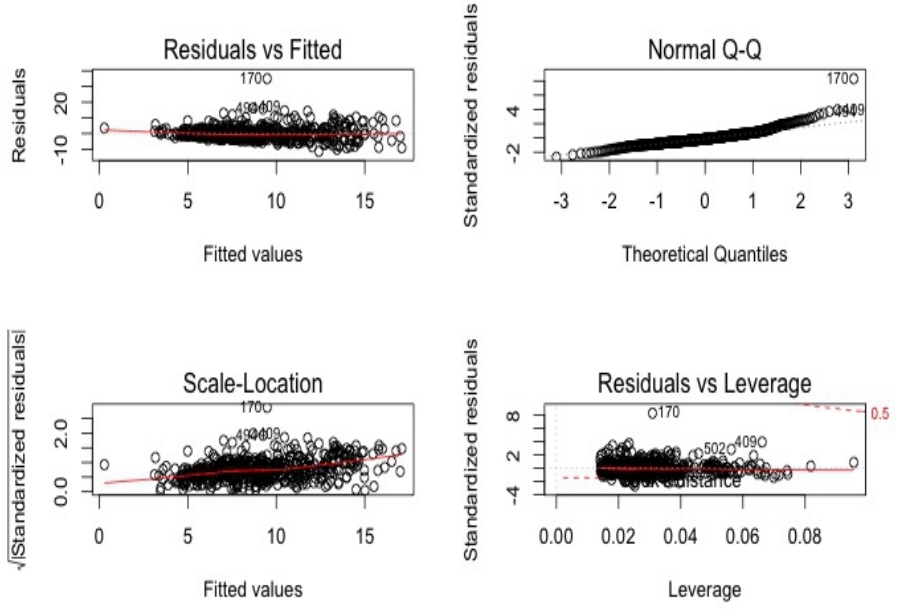


Figure 4 Density plot of wage

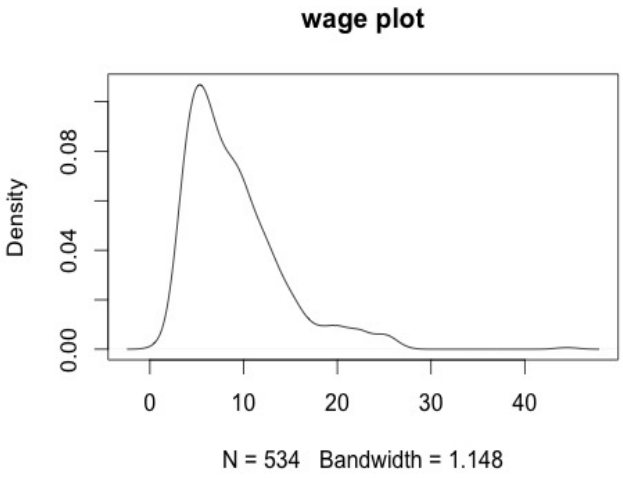
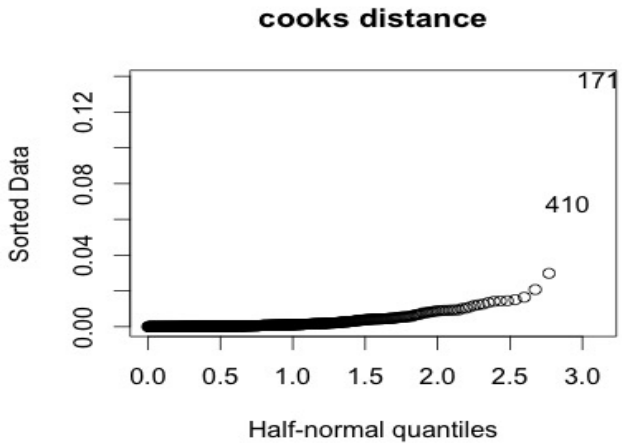


Figure 5 Influential(Cook's distance)



The outlier is a 21 years old female, has 14 years education, only one year experience in a management position, but gets the highest wage(44.5 dollars per hour) in our all 534 observations. The Cook's distance also indicates she is an influential observation to the model, so we are going to remove it in the new model fitting. Furthermore, our response variable "wage" seems quite rightly skewed, to stabilize the variance we log-transformed wage.

```
# refit model
> md1<-lm(log(wage)~.-age, CPS1985,subset=(rownames(CPS1985)!="170"))
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.004797   0.140395   7.157 2.84e-12 ***
education            0.068276   0.009899   6.897 1.55e-11 ***
experience           0.009824   0.001700   5.779 1.30e-08 ***
ethnicityhispanic   -0.106583   0.085429  -1.248  0.21273
ethnicityother      -0.077105   0.056433  -1.366  0.17243
regionother          0.089516   0.041238   2.171  0.03041 *
genderfemale        -0.228456   0.041228  -5.541 4.79e-08 ***
occupationtechnical  0.212875   0.071935   2.959  0.00323 **
occupationservices  -0.114280   0.066692  -1.714  0.08721 .
occupationoffice     0.059902   0.067362   0.889  0.37428
occupationsales     -0.100298   0.083817  -1.197  0.23200
occupationmanagement 0.228480   0.079086   2.889  0.00403 **
sectorconstruction  -0.028043   0.097398  -0.288  0.77352
sectorother         -0.117777   0.053946  -2.183  0.02947 *
unionyes             0.208387   0.050365   4.138 4.10e-05 ***
marriedyes           0.071045   0.040389   1.759  0.07916 .
---
Residual standard error: 0.4206 on 517 degrees of freedom
Multiple R-squared:  0.371, Adjusted R-squared: 0.3528
F-statistic: 20.33 on 15 and 517 DF,  p-value: < 2.2e-16
```

In this new model, **sector and region became significant, and the Adjusted R-squared increased by 15%(from 0.307 to 0.353)**.

However, a new regression diagnostic still shows non-constant variance and influential outlier. The new outlier is a 42 years old male, has 12 years education and 24 years experience in a management

position, but only gets 1 dollar per hour as the lowest wage in all our 534 observations. Then, we remove the new outlier and refit our model.

```
#outlier 2
    wage education experience age ethnicity region gender occupation sector union married
199  1     12         24       42   cauc     other  male  management  other   no     yes

# refit a model
> data<- CPS1985[-c(171,200),]
> md2<-lm(log(wage)~.-age,data)
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.037428   0.136297   7.612 1.30e-13 ***
education            0.065002   0.009618   6.758 3.79e-11 ***
experience           0.009778   0.001649   5.931 5.53e-09 ***
ethnicityhispanic   -0.116293   0.082881  -1.403  0.16118
ethnicityother      -0.081565   0.054744  -1.490  0.13686
regionother          0.095996   0.040015   2.399  0.01679 *
genderfemale        -0.238632   0.040029  -5.961 4.64e-09 ***
occupationtechnical  0.226133   0.069813   3.239  0.00128 **
occupationservices  -0.111413   0.064692  -1.722  0.08563 .
occupationoffice     0.068555   0.065357   1.049  0.29470
occupationsales     -0.095734   0.081305  -1.177  0.23955
occupationmanagement 0.281184   0.077250   3.640  0.00030 ***
sectorconstruction  -0.029553   0.094474  -0.313  0.75455
sectorother         -0.113732   0.052331  -2.173  0.03021 *
unionyes             0.205092   0.048856   4.198 3.17e-05 ***
marriedyes           0.076621   0.039188   1.955  0.05110 .
---
Residual standard error: 0.408 on 516 degrees of freedom
Multiple R-squared: 0.3916,      Adjusted R-squared: 0.374
F-statistic: 22.15 on 15 and 516 DF,  p-value: < 2.2e-16
```

In this model, **the significant variables are the same, some of their coefficients changed(married is almost significant now), and the Adjusted R-squared keeps increasing by another 6%(from 0.353 to 0.374)**.

3.3 Variable Selection

Here, we apply both Backward Elimination and AIC Criterion to select the appropriate variables for our model. Both ways yield the same result: to remove “ethnicity” and keep “married” based on previous md2 model.

```
 # AIC Criterion-Based Selection
> step(md2)
Step:  AIC=-938.21
log(wage) ~ education + experience + region + gender + occupation + sector + union + married

             Df Sum of Sq    RSS     AIC
<none>                    86.528 -938.21
- sector      2    0.7841 87.312 -937.41
- married     1    0.6947 87.223 -935.95
- region      1    1.1701 87.698 -933.06
- union       1    2.7114 89.239 -923.79
- occupation  5    6.5451 93.073 -909.41
- gender      1    5.7645 92.292 -905.90
- experience  1    6.0451 92.573 -904.28
- education   1    8.4260 94.954 -890.77
```

```
# backward elimination
> summary(update(md2,.~.-ethnicity))

# final model
> md.final<-lm(log(wage)~education + experience + region + gender + occupation + sector +
union + married, data=data)
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.983949   0.133078   7.394 5.78e-13 ***
education            0.067324   0.009479   7.102 4.07e-12 ***
experience           0.009908   0.001647   6.016 3.39e-09 ***
regionother          0.105307   0.039789   2.647 0.008377 **
genderfemale        -0.235255   0.040047  -5.874 7.59e-09 ***
occupationtechnical  0.217745   0.069598   3.129 0.001855 **
occupationservices  -0.124534   0.064382  -1.934 0.053622 .
occupationoffice     0.057715   0.065220   0.885 0.376607
occupationsales     -0.097829   0.081436  -1.201 0.230184
occupationmanagement 0.269148   0.077080   3.492 0.000521 ***
sectorconstruction  -0.019350   0.094480  -0.205 0.837806
sectorother         -0.110320   0.052340  -2.108 0.035534 *
unionyes             0.196122   0.048679   4.029 6.45e-05 ***
marriedyes           0.079978   0.039217   2.039 0.041919 *
---
Residual standard error: 0.4087 on 518 degrees of freedom
Multiple R-squared:  0.3872, Adjusted R-squared:  0.3718
F-statistic: 25.17 on 13 and 518 DF,  p-value: < 2.2e-16
```

Although some level of factor variables are not significant in the summary, such as "sectorconstruction" of section, "occupationoffice" of occupation, because they have at least one level is significant, we need to keep the variable in our model. Finally, the final model we reached includes variables of education, experience, region, occupation, sector, union and married. It turns out that "ethnicity" will not statistically affect a person's hourly wage rate.

Plus, we found the coefficient of experience is too small compared with other coefficients. To make the model coefficients more readable, we can rescale the formula from "experience" to "experience /10".

```
# rescale experience coefficient
summary(lm(formula = log(wage) ~ education + I(experience/10) + region + gender +
+ occupation + sector + union + married, data = data))
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.983949   0.133078   7.394 5.78e-13 ***
education            0.067324   0.009479   7.102 4.07e-12 ***
I(experience/10)     0.099080   0.016470   6.016 3.39e-09 ***
regionother          0.105307   0.039789   2.647 0.008377 **
genderfemale        -0.235255   0.040047  -5.874 7.59e-09 ***
occupationtechnical  0.217745   0.069598   3.129 0.001855 **
occupationservices  -0.124534   0.064382  -1.934 0.053622 .
occupationoffice     0.057715   0.065220   0.885 0.376607
occupationsales     -0.097829   0.081436  -1.201 0.230184
occupationmanagement 0.269148   0.077080   3.492 0.000521 ***
sectorconstruction  -0.019350   0.094480  -0.205 0.837806
sectorother         -0.110320   0.052340  -2.108 0.035534 *
unionyes             0.196122   0.048679   4.029 6.45e-05 ***
marriedyes           0.079978   0.039217   2.039 0.041919 *
---
Residual standard error: 0.4087 on 518 degrees of freedom
Multiple R-squared:  0.3872,     Adjusted R-squared:  0.3718
F-statistic: 25.17 on 13 and 518 DF,  p-value: < 2.2e-16
```

3.3 Diagnostics and Interpretation

The diagnostics on md.final shows the regression assumptions fit. Figure 6 below indicates constant variance of errors while Figure 7 testifies the normality. Also, we didn't find any outliers or influential observations. Therefore, we continue to study the model for parsimony.

Figure6 Residuals vs Fitted

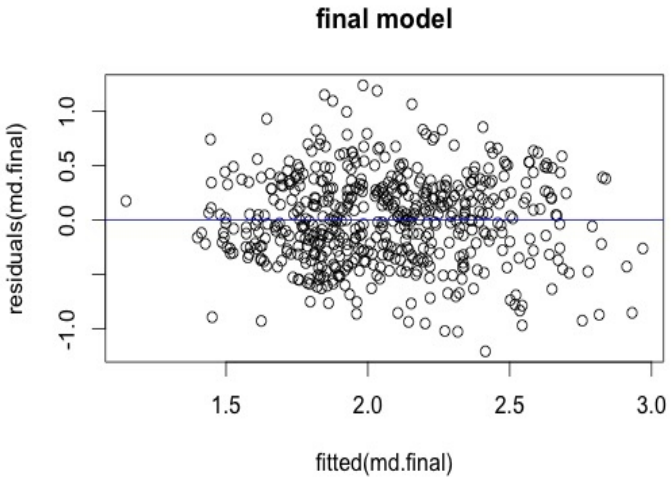
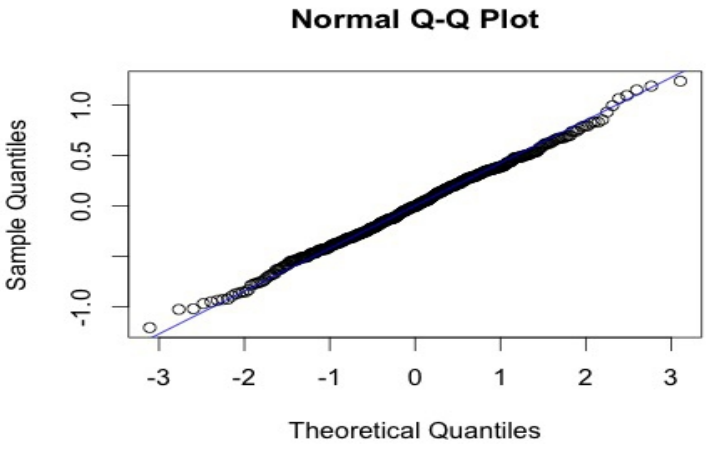


Figure7 qqplot of md.final

#




```
# study the model for parsimony
#analysis of variance
> anova(md.final)
Analysis of Variance Table
Response: log(wage)
             Df Sum Sq Mean Sq  F value    Pr(>F)
education     1 20.913 20.9125 125.1930 < 2.2e-16 ***
experience    1 10.772 10.7725  64.4894 6.574e-15 ***
region        1  1.616  1.6161   9.6746  0.001971 **
gender        1  9.891  9.8908  59.2112 7.206e-14 ***
occupation    5  7.108  1.4215   8.5098 9.342e-08 ***
sector        2  0.739  0.3696   2.2128  0.110422
union         1  2.930  2.9304  17.5431 3.301e-05 ***
married       1  0.695  0.6947   4.1591  0.041919 *
Residuals   518 86.528  0.1670

# drop one variable test
> drop1(md.final,test="F")
Single term deletions
Model: log(wage) ~ education+experience+region+gender+occupation+sector+union+married
           Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                  86.528 -938.21
education   1    8.4260 94.954 -890.77 50.4423 4.073e-12 ***
experience  1    6.0451 92.573 -904.28 36.1891 3.387e-09 ***
region      1    1.1701 87.698 -933.06  7.0047  0.008377 **
gender      1    5.7645 92.292 -905.90 34.5093 7.594e-09 ***
occupation  5    6.5451 93.073 -909.41  7.8364 3.989e-07 ***
sector      2    0.7841 87.312 -937.41  2.3471  0.096658 .
union       1    2.7114 89.239 -923.79 16.2316 6.445e-05 ***
married     1    0.6947 87.223 -935.95  4.1591  0.041919 *
```

We apply both sequential analysis of variance (ANOVA) and 1 variable deletion to test each of the 2 variables against the full model. Both results show that “sector” doesn't significantly contribute. Therefore, we drop it and refit a model.

```
# refit a final model and test the model
> md.final2<-lm(log(wage)~ education + experience + region + gender + occupation + union +
married, data)
> anova(md.final2)
            Df Sum Sq Mean Sq  F value    Pr(>F)
education    1 20.913 20.9125 124.5477 < 2.2e-16 ***
experience   1 10.772 10.7725  64.1570 7.587e-15 ***
region       1  1.616  1.6161   9.6248  0.002024 **
gender       1  9.891  9.8908  58.9060 8.235e-14 ***
occupation   5  7.108  1.4215   8.4660 1.024e-07 ***
union        1  2.855  2.8554  17.0059 4.337e-05 ***
married      1  0.725  0.7249   4.3172  0.038219 *
Residuals  520 87.312  0.1679

> drop1(md.final2,test="F")
           Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                  87.312 -937.41
education   1    8.5679 95.880 -889.61 51.0272 3.091e-12 ***
experience  1    6.5882 93.900 -900.71 39.2369 7.875e-10 ***
region      1    1.3013 88.613 -931.54  7.7500  0.005567 **
gender      1    5.6288 92.941 -906.17 33.5232 1.219e-08 ***
occupation  5    6.8889 94.201 -907.01  8.2055 1.796e-07 ***
union       1    2.6349 89.947 -923.59 15.6923 8.493e-05 ***
married     1    0.7249 88.037 -935.01  4.3172  0.038219 *

> summary(md.final2)
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.916231   0.127896   7.164 2.70e-12 ***
education            0.067826   0.009495   7.143 3.09e-12 ***
experience           0.010277   0.001641   6.264 7.87e-10 ***
regionother          0.110625   0.039738   2.784  0.00557 **
genderfemale        -0.230674   0.039841  -5.790 1.22e-08 ***
occupationtechnical  0.168914   0.065996   2.559  0.01077 *
occupationservices  -0.184545   0.058158  -3.173  0.00160 **
occupationoffice     0.004085   0.060490   0.068  0.94619
occupationsales     -0.149021   0.078006  -1.910  0.05663 .
occupationmanagement 0.218080   0.073461   2.969  0.00313 **
unionyes             0.193145   0.048757   3.961 8.49e-05 ***
marriedyes           0.081667   0.039305   2.078  0.03822 *
---
Residual standard error: 0.4098 on 520 degrees of freedom
Multiple R-squared:  0.3816, Adjusted R-squared:  0.3685
F-statistic: 29.17 on 11 and 520 DF,  p-value: < 2.2e-16
```

This new final model drops “sector” variable, the Adjusted R-squared declined a little(from 0.3718 to 0.3685, but we do have a simpler model. Furthermore, both ANOVA and drop one variable tests testify that all the variables in our new model are significantly contribute. Thus, we are going to use this model for the interpretation.

Because our response variable "wage" is log-transformed in the model fitting, to correctly interpret our model, we need to anti-log transform(exponential-transform) the coefficients back.

```
# original coefficients
> round(coef(md.final2),3)
         (Intercept)            education           experience
               0.916                0.068                0.010
         regionother         genderfemale  occupationtechnical
               0.111               -0.231                0.169
  occupationservices     occupationoffice      occupationsales
              -0.185                0.004               -0.149
occupationmanagement            unionyes           marriedyes
               0.218                0.193                0.082
#anti-log transform the coefficients
> round(exp(coef(md.final2)),3)
         (Intercept)            education           experience
               2.500                1.070                1.010
         regionother         genderfemale  occupationtechnical
               1.117                0.794                1.184
  occupationservices     occupationoffice      occupationsales
               0.831                1.004                0.862
occupationmanagement            unionyes           marriedyes
               1.244                1.213                1.085
```

The interpretation we come up with is as follows. **When other predictors are held constant**:

- If education is increased by one year, we expect the hourly wage to increase by 7%;
- Similarly, if a person's experience is increased by one year, we would expect the hourly wage to increase by 1%;
- For people live in places other than South, their the average hourly wage is expected to be 1.117 times of people live in South:

$$\ln\left(\frac{wage\ of\ other}{wage\ of\ south}\right)=0.111 \quad => \quad \frac{wage\ of\ other}{wage\ of\ south} = e^{0.111} = 1.117$$

- For the female, their average hourly wage is 79.4% of the male:

$$\ln\left(\frac{wage\ of\ female}{wage\ of\ male}\right)=-0.231 \quad => \quad \frac{wage\ of\ female}{wage\ of\ male} = e^{-0.231} = 0.794$$

- Similarly, the average hourly wage for married people is expected to be 1.085 times ($e^{0.082}$ =1.085) of unmarried people; people in union would expect their average hourly pay be 1.213 times ($e^{0.193}$ =1.213) of their peers not in a union;
- Compared with the hourly wage of "workers(tradesperson or assembly line worker)", the technical or professional workers attain 1.184 times of wage($e^{0.169}$ =1.184), the service workers get 83.1% ($e^{-0.185}$ =0.831), the office and clerical workers almost have the same wage rate ($e^{0.004}$ =1.004), the sales workers, on the contrary, could only get 86.2% ($e^{-0.149}$ =0.862), while the management and administration staff hold the highest hourly wage, which reaches 1.244 times ($e^{0.218}$ = 1.244), on average.