ANALYTICAL METHODS OF BUSINESS

(QMB 6304 Regression Project)

FINAL PROJECT

ON

DETERMINING WAGES OF INDIVIDUALS

Submitted by

Karna Haritha	**U46427632**
Sri Venkata Likhitha Duggi	**U37870238**
Rahul Gomedhikam	**U10270893**

USF UNIVERSITY OF SOUTH FLORIDA

SOURCE OF DATA

This dataset pertains to the difference in wages paid for individuals considering the factors like education, experience, female or not, race and ethnicity .

The data is sourced from GitHub from the mentioned link:
https://github.com/bharathirajatut/sample-excel-dataset/blob/master/cps2.xls

The dataset consists of 2995 observations and 10 features out of which 5 are attributes of interest.

This dataset consists of following columns: wage, education, experience, female, race, white, black, Hispanic, Asian, others.

- **Wage**: A numeric vector in which the wages of the employees are conveyed.
  The variable wage is going to be our **dependent variable(Y)**
- **Education**: Years of education. This is a continuous variable; education is going to be our first **independent variable (X1).**
- **Experience**: Number of years of work experience. This is a continuous variable; experience is going to be our **second independent variable (X2).**
- **Female**: If the employee is female or not (Female=1, if not=0).This variable is a binary variable and will be **third independent variable(X3). There are 2 levels which are 0 and 1**.
- **Race** : Hispanic/ white/ black/ Asian/ others

X1	Education
X2	Experience
X3	Female
Y	Wage

UNIVERSITY OF SOUTH FLORIDA

Complete listing of 100 sample observations

	wage	education	experience	female	race	White	Black	Hispanic	Asian	Other
503	38.46	16	31	0	White	1	0	0	0	0
2035	19.23	16	14	1	White	1	0	0	0	0
2967	20.31	18	29	1	Other	0	0	0	0	1
470	7.21	12	9	0	White	1	0	0	0	0
1990	13.39	18	3	1	White	1	0	0	0	0
1540	7.69	12	4	0	Hispanic	0	0	1	0	0
823	12.02	14	28	1	White	1	0	0	0	0
2886	25.48	16	16	0	White	1	0	0	0	0
1122	21.15	16	12	0	White	1	0	0	0	0
2951	17.05	20	32	0	White	1	0	0	0	0
183	33.33	16	1	1	White	1	0	0	0	0
2347	19.23	18	25	0	Asian	0	0	0	1	0
1528	23.08	13	16	0	White	1	0	0	0	0
2514	10.42	12	55	1	White	1	0	0	0	0
2956	17.09	16	46	1	White	1	0	0	0	0
1331	12.62	12	42	1	White	1	0	0	0	0
456	19.23	12	26	0	Hispanic	0	0	1	0	0
1817	11.54	16	17	0	Asian	0	0	0	1	0
2372	25.00	14	29	0	White	1	0	0	0	0
1092	37.50	12	18	0	Black	0	1	0	0	0
510	6.25	12	36	1	White	1	0	0	0	0
948	5.98	12	15	1	Hispanic	0	0	1	0	0
288	16.48	11	10	0	Other	0	0	0	0	1
347	13.90	16	3	0	White	1	0	0	0	0
1191	34.79	16	12	1	White	1	0	0	0	0
1808	9.70	13	3	1	White	1	0	0	0	0
971	25.00	13	40	0	White	1	0	0	0	0
2676	12.50	13	13	1	Other	0	0	0	0	1
2991	24.04	12	23	0	Other	0	0	0	0	1
605	9.51	13	25	1	Black	0	1	0	0	0
1325	35.04	16	30	0	White	1	0	0	0	0
1182	25.87	12	10	0	Black	0	1	0	0	0
733	20.71	13	32	0	White	1	0	0	0	0
1631	11.92	12	13	0	White	1	0	0	0	0
1889	19.23	13	28	0	Black	0	1	0	0	0
223	19.47	18	6	0	White	1	0	0	0	0
2780	16.48	13	13	1	Black	0	1	0	0	0
2299	8.37	12	1	1	White	1	0	0	0	0
1567	6.92	16	1	1	Hispanic	0	0	1	0	0
1718	14.42	14	4	1	Black	0	1	0	0	0
1449	13.74	14	4	1	Hispanic	0	0	1	0	0
1513	6.67	16	23	1	White	1	0	0	0	0
502	27.87	12	48	1	White	1	0	0	0	0
1195	28.85	14	12	0	White	1	0	0	0	0
519	10.42	12	32	1	White	1	0	0	0	0
449	7.42	12	19	0	White	1	0	0	0	0
2441	19.05	13	56	0	White	1	0	0	0	0
998	11.00	12	16	1	Hispanic	0	0	1	0	0
660	30.77	12	7	1	Hispanic	0	0	1	0	0
1934	8.65	14	4	1	White	1	0	0	0	0
2411	14.42	12	41	1	White	1	0	0	0	0
2894	9.62	4	57	0	Hispanic	0	0	1	0	0
1624	14.42	16	6	1	Other	0	0	0	0	1
1411	16.11	12	18	0	Asian	0	0	0	1	0
1902	21.63	16	17	0	White	1	0	0	0	0
1444	29.33	18	24	1	Asian	0	0	0	1	0
2419	25.00	16	24	1	White	1	0	0	0	0
2931	16.07	13	23	1	White	1	0	0	0	0
2478	14.42	10	21	0	White	1	0	0	0	0
1012	54.09	16	14	1	Hispanic	0	0	1	0	0
2095	50.00	18	15	0	White	1	0	0	0	0
1463	28.00	12	41	1	White	1	0	0	0	0
708	26.67	13	25	1	White	1	0	0	0	0

UNIVERSITY OF SOUTH FLORIDA

2060	14.42	6	18	0	Hispanic	0	0	1	0	0
947	20.16	14	11	0	White	1	0	0	0	0
1145	38.70	14	5	1	White	1	0	0	0	0
16	25.00	13	36	0	White	1	0	0	0	0
1976	20.67	18	2	1	White	1	0	0	0	0
1430	19.87	13	24	1	White	1	0	0	0	0
978	13.89	16	11	1	Asian	0	0	0	1	0
949	5.98	12	17	0	Hispanic	0	0	1	0	0
2691	32.69	11	41	0	White	1	0	0	0	0
556	22.98	12	38	0	White	1	0	0	0	0
2204	6.84	14	47	1	White	1	0	0	0	0
2974	15.77	16	29	1	Asian	0	0	0	1	0
757	15.38	16	38	0	Hispanic	0	0	1	0	0
2329	20.34	18	9	0	White	1	0	0	0	0
1578	13.59	14	27	1	White	1	0	0	0	0
1209	20.67	13	15	0	White	1	0	0	0	0
1868	15.38	10	9	1	White	1	0	0	0	0
2469	16.15	12	45	1	White	1	0	0	0	0
2714	8.01	9	18	1	Hispanic	0	0	1	0	0
2538	46.15	13	61	0	White	1	0	0	0	0
871	13.94	12	12	0	Asian	0	0	0	1	0
1816	6.15	18	43	1	Asian	0	0	0	1	0
1420	16.83	14	35	0	White	1	0	0	0	0
1161	18.46	13	36	1	White	1	0	0	0	0
2298	12.50	13	17	1	White	1	0	0	0	0
1387	44.07	14	16	1	Hispanic	0	0	1	0	0
2615	52.00	16	14	1	White	1	0	0	0	0
1867	25.00	13	31	0	White	1	0	0	0	0
1315	11.54	16	0	1	White	1	0	0	0	0
2586	23.08	16	12	1	White	1	0	0	0	0
1142	28.85	14	37	1	White	1	0	0	0	0
1061	38.46	14	10	0	Hispanic	0	0	1	0	0
2270	6.41	12	42	1	Hispanic	0	0	1	0	0
2779	8.14	12	12	0	Hispanic	0	0	1	0	0
1682	16.83	12	25	1	White	1	0	0	0	0
328	20.67	16	8	1	White	1	0	0	0	0
1608	7.21	8	18	1	Hispanic	0	0	1	0	0

REGRESSION ANALYSIS

1) REGRESSION MODEL for WAGE, EDUCATION

Equation for the model:

$Y = 4.4193 + 1.1017 * X1$

where $\beta 0$(intercept)=4.4193 and $\beta 1$(slope)=1.1017

IV's and DV's

Here, the **independent variable is education** and **dependent variable is wage**.

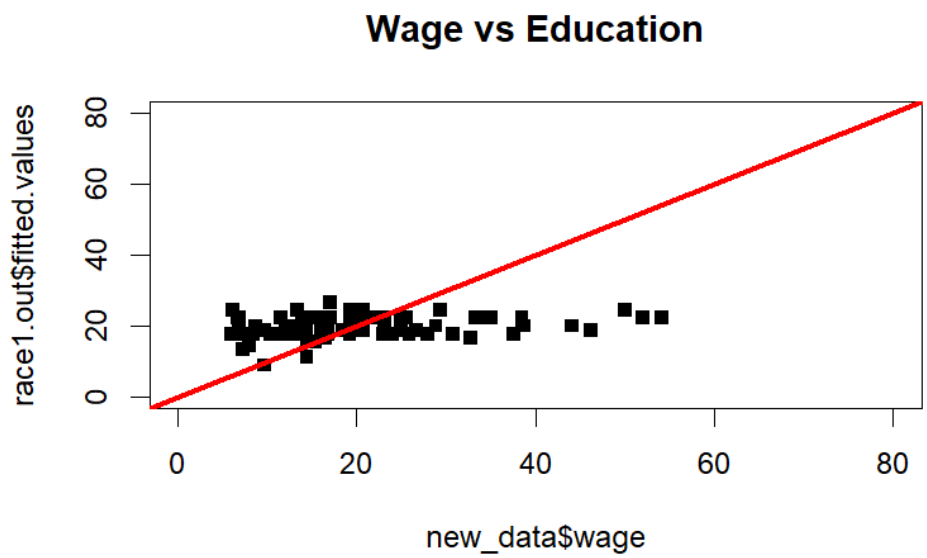
CODE:

```
race1.out=lm(wage~education,data=new_data)
summary(race1.out)
plot(new_data$wage,race1.out$fitted.values,
     pch=15,
     xlim=c(0,80),ylim=c(0,80),
     main="Wage vs Education")
abline(0,1,lwd=3,col="red")
```

Output:

```
Call:
lm(formula = wage ~ education, data = new_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.100  -7.220  -1.896   5.202  32.043 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   4.4193     5.6498   0.782  0.43598   
education     1.1017     0.4044   2.724  0.00763 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.43 on 98 degrees of freedom
Multiple R-squared:  0.0704,	Adjusted R-squared:  0.06091 
F-statistic: 7.422 on 1 and 98 DF,  p-value: 0.007632
```

Graph:



Significant term: education

2) REGRESSION MODEL for WAGE, EXPERIENCE

Equation for the model:

Y= 19.11597+ 0.01967X2

where β0(intercept)= 19.11597and β1(slope)= 0.01967

IV's and DV's

Here, the **independent variable is experience** and **dependent variable is wage**.

CODE:

```
race2.out=lm(wage~experience,data=new_data)
summary(race2.out)
plot(new_data$wage,race2.out$fitted.values,
     pch=15,
     xlim=c(0,80),ylim=c(0,80),
     main="Wage vs Experience")
abline(0,1,lwd=3,col="red")
```

Output:

```
Call:
lm(formula = wage ~ experience, data = new_data)

Residuals:
    Min      1Q  Median      3Q     Max
-13.812  -7.483  -2.805   5.201  34.699

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.11597    1.96255   9.740 4.43e-16 ***
experience   0.01967    0.07487   0.263    0.793
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.82 on 98 degrees of freedom
Multiple R-squared:  0.0007036, Adjusted R-squared:  -0.009493
F-statistic: 0.06901 on 1 and 98 DF,  p-value: 0.7933
```

Graph:

Significant term: Intercept

USF
UNIVERSITY OF
SOUTH FLORIDA

3) REGRESSION MODEL for WAGE, FEMALE

Equation for the model:

$$Y = 21.278 - 3.149 * X3$$

where β0(intercept)= 21.278 and β1(slope)= -3.149

IV's and DV's

Here, the **independent variable** is binary level **Female** variable and **dependent variable** is **wage**.

CODE:

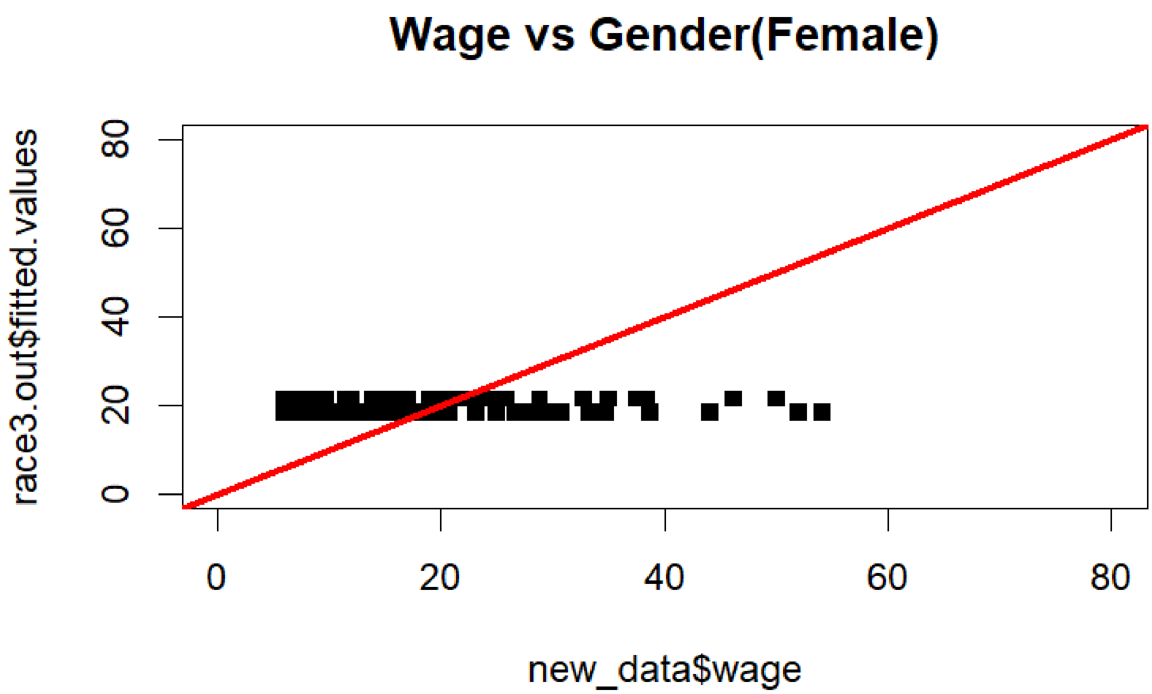
```
race3.out=lm(wage~female,data=new_data)
summary(race3.out)
plot(new_data$wage,race3.out$fitted.values,
     pch=15,
     xlim=c(0,80),ylim=c(0,80),
     main="Wage vs Gender(Female)")
abline(0,1,lwd=3,col="red")
```

Output:

```
Call:
lm(formula = wage ~ female, data = new_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.298  -7.181  -2.048   3.722  35.961 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   21.278      1.596  13.333   <2e-16 ***
female        -3.149      2.152  -1.463    0.147    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.71 on 98 degrees of freedom
Multiple R-squared:  0.02139,	Adjusted R-squared:  0.0114 
F-statistic: 2.142 on 1 and 98 DF,  p-value: 0.1465
```

Graph:

UNIVERSITY OF SOUTH FLORIDA



Significant term: Intercept

Multiple Regression

4) Wage vs (Education, Experience)

Equation for the model:

Y= 1.7594 + 1.1876 *X1 + 0.0677 * X2

where β0(intercept)= 1.7594 and β1(slope)= 1.1876, β2(slope)= 0.0677

IV's and DV's

Here, the independent variables are Education, and experience whereas dependent variable is wage.

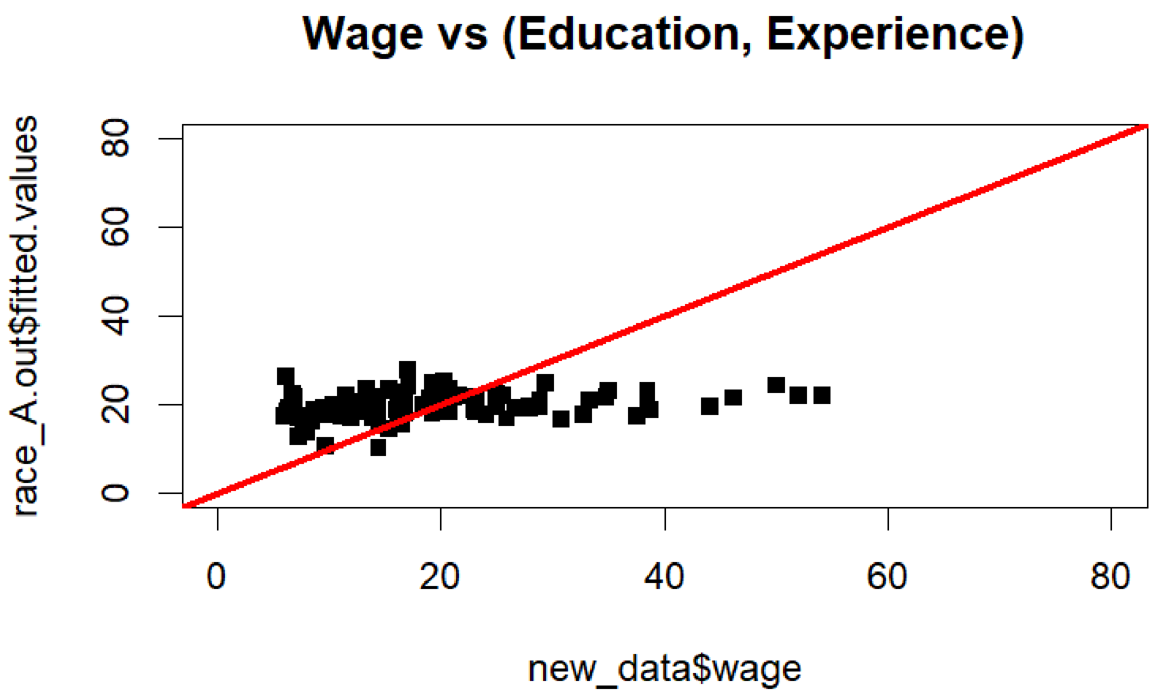
CODE:

```
race_A.out=lm(wage ~ education+experience,data=new_data)
summary(race_A.out)
plot(new_data$wage,race_A.out$fitted.values,
     pch=15,
     xlim=c(0,80),ylim=c(0,80),
     main="Wage vs (Education, Experience)")
abline(0,1,lwd=3,col="red")
```

Output:

```
Call:
lm(formula = wage ~ education + experience, data = new_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.898  -7.202  -2.210   4.687  32.381 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   1.7594     6.3619   0.277  0.78272   
education     1.1876     0.4156   2.858  0.00522 **
experience    0.0677     0.0742   0.912  0.36382   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.44 on 97 degrees of freedom
Multiple R-squared:  0.07831,   Adjusted R-squared:  0.0593 
F-statistic: 4.121 on 2 and 97 DF,  p-value: 0.01916
```

Graph:



Significant term: Education

5) Wage vs (Experience, Female)

Equation for the model:

Y= 5.4905 + 1.1743 *X1 – 3.7590 * X3

where β0(intercept)= 5.4905 , and β1(slope)= 1.1743 , β2(slope)= – 3.7590

IV's and DV's

Here, the independent variables are Female, and experience whereas dependent variable is wage.

CODE:

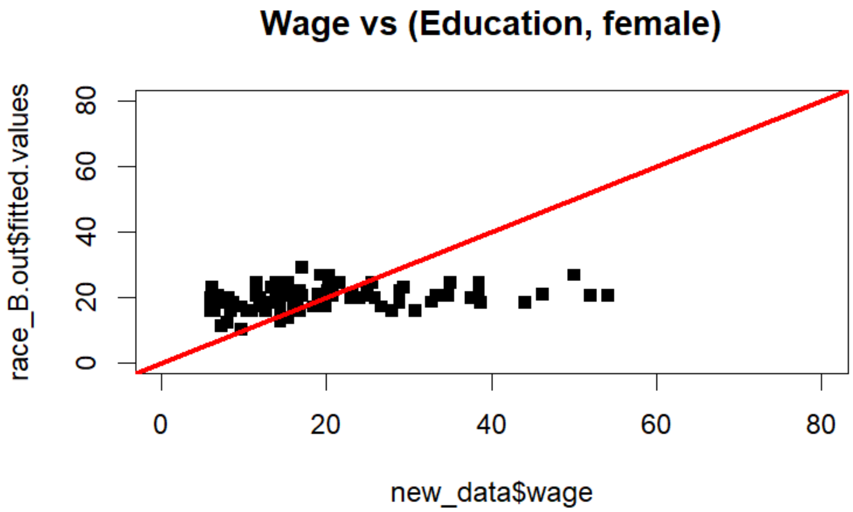
UNIVERSITY OF SOUTH FLORIDA

```
race_B.out=lm(wage ~ education+female,data=new_data)
summary(race_B.out)
plot(new_data$wage,race_B.out$fitted.values,
     pch=15,
     xlim=c(0,80),ylim=c(0,80),
     main="Wage vs (Education, female)")
abline(0,1,lwd=3,col="red")
```

Output:

```
Call:
lm(formula = wage ~ education + female, data = new_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.719  -6.762  -2.063   4.244  33.570 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   5.4905     5.6175   0.977  0.33080   
education     1.1743     0.4019   2.922  0.00433 **
female       -3.7590     2.0842  -1.804  0.07440 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.32 on 97 degrees of freedom
Multiple R-squared:  0.1006,	Adjusted R-squared:  0.08202 
F-statistic: 5.423 on 2 and 97 DF,  p-value: 0.005856
```

Graph:



Significant term : **Education**

6) Wage vs (Education, Female)

Equation for the model:

Y= 20.98202 + 0.01295 *X2 – 3.12584 * X3

where β0(intercept)= 20.98202 and β1(slope)= 0.01295, β2(slope)= – 3.12584

IV's and DV's

Here, the **independent variables are Education, and Female** whereas dependent variable is wage.

CODE:

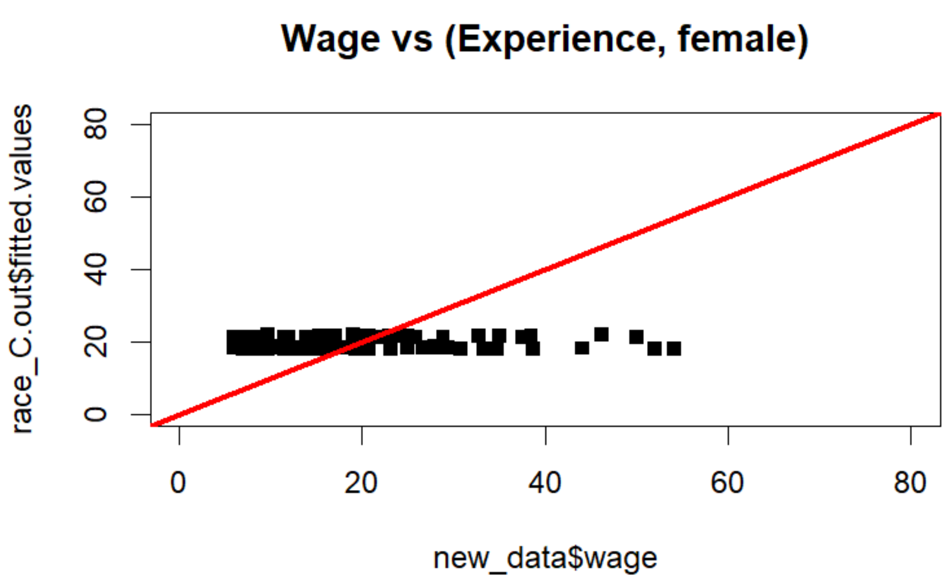
```
race_C.out=lm(wage ~ experience+female,data=new_data)
summary(race_C.out)
plot(new_data$wage,race_C.out$fitted.values,
     pch=15,
     xlim=c(0,80),ylim=c(0,80),
     main="Wage vs (Experience, female)")
abline(0,1,lwd=3,col="red")
```

UNIVERSITY OF SOUTH FLORIDA

Output:

```
Call:
lm(formula = wage ~ experience + female, data = new_data)

Residuals:
    Min       1Q   Median       3Q      Max
-15.222   -7.078  -2.102    3.623   36.052

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.98202    2.34159   8.961 2.35e-14 ***
experience   0.01295    0.07461   0.174    0.863
female1     -3.12584    2.16695  -1.443    0.152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.76 on 97 degrees of freedom
Multiple R-squared:  0.02169,    Adjusted R-squared:  0.001519
F-statistic: 1.075 on 2 and 97 DF,  p-value: 0.3452
```

Graph:



Significant term: Intercept

7) Multiple regression model between Wage Vs Female, Education and Experience.

Equation for the model:

Y= 3.01862 + 1.25204 *X1 + 0.06239 *X2 – 3.6863 * X3

where β0(intercept)= 3.01862 and β1(slope)= 1.25204, β2(slope)= 0.06239, β3(slope)= –3.6863

IV's and DV's

Here, the **independent variables are Education, Experience and Female** whereas **dependent variable is wage.**

CODE:

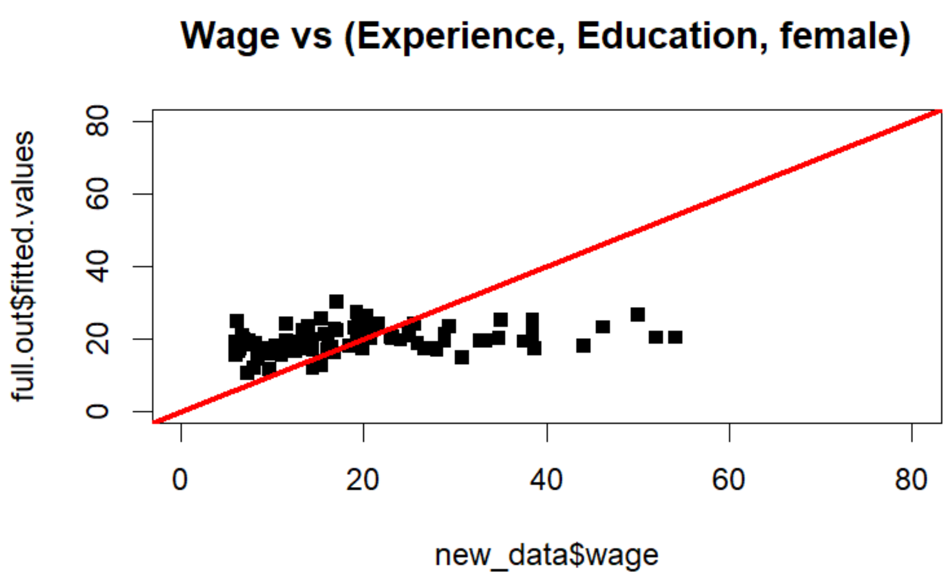
```
full.out=lm(wage ~ education+experience+female,data=new_data)
summary(full.out)
plot(new_data$wage,full.out$fitted.values,
     pch=15,
     xlim=c(0,80),ylim=c(0,80),
     main="Wage vs (Experience, Education, female)")
abline(0,1,lwd=3,col="red")
```

Output:

UNIVERSITY OF SOUTH FLORIDA

```
Call:
lm(formula = wage ~ education + experience + female, data = new_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.402  -6.236  -2.456   3.272  33.852 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  3.01862    6.33405   0.477  0.63475   
education    1.25204    0.41273   3.034  0.00311 **
experience   0.06239    0.07347   0.849  0.39788   
female1     -3.68630    2.08894  -1.765  0.08080 . 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.33 on 96 degrees of freedom
Multiple R-squared:  0.1073,	Adjusted R-squared:  0.07937 
F-statistic: 3.845 on 3 and 96 DF,  p-value: 0.01201
```

Graph:



Significant Term: Education

8) Multiple regression model using an interaction term

Equation for the model:

Y= -4.73504 + 1.66103 X1 + 0.32617X2 – 0.01939 X1X2

where β0(intercept)= -4.73504 and β1(slope)= 1.66103, β2(slope)= 0.32617, β3(slope)= – 0.01939

IV's and DV's

Here, the **independent variables are Education, Experience** and **dependent variable is wage**. The interaction term is **education * experience**.

CODE:

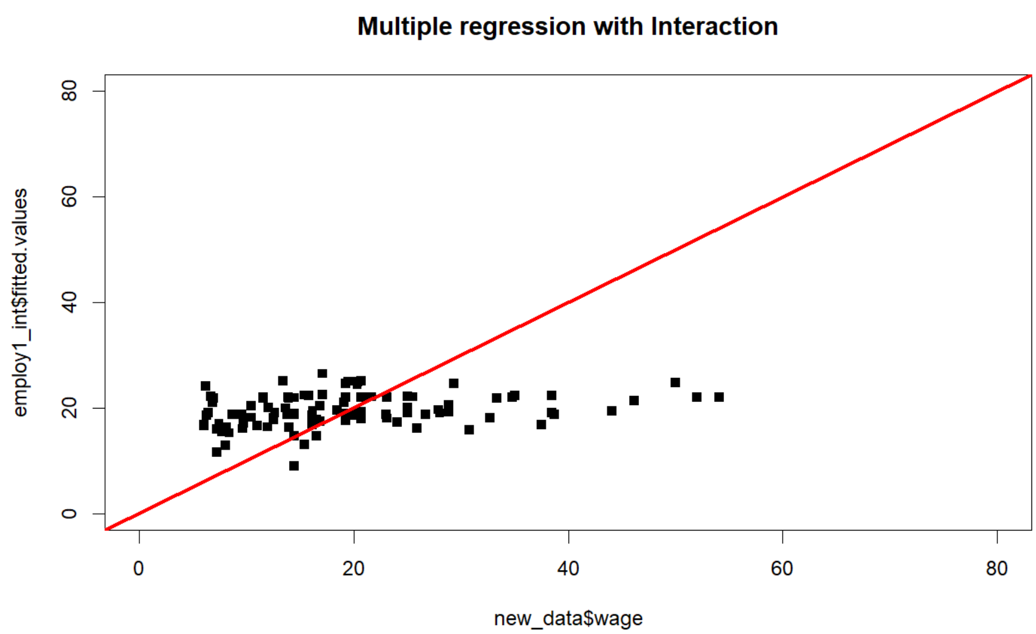
```
employ1_int= lm(wage ~
education+experience+I(education*experience),data=new_data)
summary(employ1_int)
plot(new_data$wage,employ1_int$fitted.values,
     pch=15,
     xlim=c(0,80),ylim=c(0,80),
     main="Multiple regression with Interaction between
Education and Experience")
abline(0,1,lwd=3,col="red")
```

UNIVERSITY OF SOUTH FLORIDA

Output:

```
Call:
lm(formula = wage ~ education + experience + I(education * experience),
    data = new_data)

Residuals:
    Min      1Q  Median      3Q     Max
-18.032  -7.146  -2.436   5.071  32.025

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -4.73504   10.35302  -0.457   0.6484
education                  1.66103    0.72596   2.288   0.0243 *
experience                 0.32617    0.33309   0.979   0.3299
I(education * experience) -0.01939    0.02436  -0.796   0.4280
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.46 on 96 degrees of freedom
Multiple R-squared:  0.08435,   Adjusted R-squared:  0.05574
F-statistic: 2.948 on 3 and 96 DF,  p-value: 0.03668
```

Graph:



Significant Term: Education

9) Regression models using squared terms of education-

Equation for the model:

Y= -7.42072 + 2.62600 X1 + 0.07067X2 – 0.05448 X1^2

where β0(intercept)= -7.42072 and β1(slope)= 2.62600, β2(slope)= 0.07067, β3(slope)= – – 0.05448

IV's and DV's

Here, the **independent variables are Education, Experience, Education ^ 2** whereas **dependent variable is wage.**

CODE:

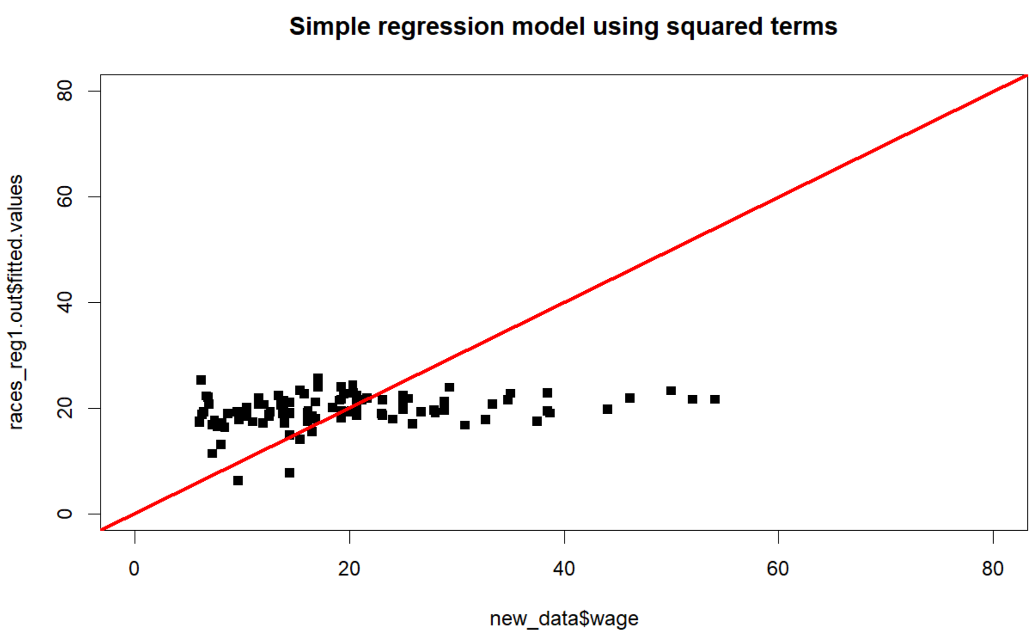
```
races_reg1.out=                    lm(wage                    ~
education+experience+I(education^2),data=new_data)
summary(races_reg1.out)
plot(new_data$wage,races_reg1.out$fitted.values,
     pch=15,
     xlim=c(0,80),ylim=c(0,80),
     main="Simple regression model using squared terms")
abline(0,1,lwd=3,col="red")
```

Output:

UNIVERSITY OF SOUTH FLORIDA

```
Call:
lm(formula = wage ~ education + experience + I(education^2),
    data = new_data)

Residuals:
    Min      1Q  Median      3Q     Max
-19.083  -7.120  -2.177   4.734  32.453

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.42072   15.91125  -0.466    0.642
education       2.62600    2.32147   1.131    0.261
experience      0.07067    0.07458   0.948    0.346
I(education^2) -0.05448    0.08651  -0.630    0.530

Residual standard error: 10.48 on 96 degrees of freedom
Multiple R-squared:  0.0821,    Adjusted R-squared:  0.05342
F-statistic: 2.862 on 3 and 96 DF,  p-value: 0.04081
```

Graph:



10) Regression models using squared terms of experience-

Equation for the model:

Y= 1.207580 + 1.18142X1 + 0.140042X2 - 0.001375 X2^2

where β0(intercept)= 1.207580 and β1(slope)= 1.18142, β2(slope)= 0.140042 , β3(slope)= -0.001375

IV's and DV's

Here, the independent variables are Education, Experience, Experience ^ 2, whereas dependent variable is wage.

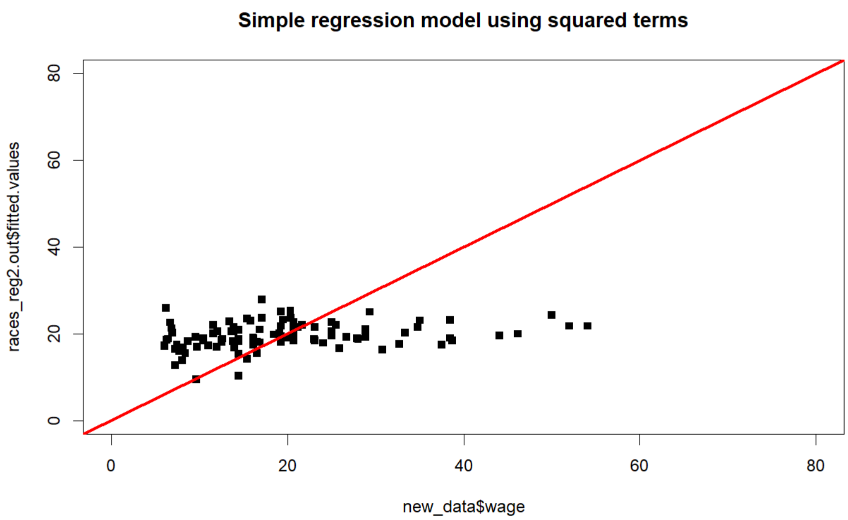
CODE:

```
races_reg2.out= lm(wage ~
education+experience+I(experience^2),data=new_data)
summary(races_reg2.out)
plot(new_data$wage,races_reg2.out$fitted.values,
     pch=15,
     xlim=c(0,80),ylim=c(0,80),
     main="Simple regression model using squared terms")
abline(0,1,lwd=3,col="red")
```

Output:

```
Call:
lm(formula = wage ~ education + experience + I(experience^2),
    data = new_data)

Residuals:
    Min      1Q  Median      3Q     Max
-19.802  -7.176  -1.876   4.417  32.289

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.207580   6.635365   0.182  0.85597
education        1.181412   0.417997   2.826  0.00573 **
experience       0.140042   0.245129   0.571  0.56913
I(experience^2) -0.001375   0.004440  -0.310  0.75740
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.49 on 96 degrees of freedom
Multiple R-squared:  0.07923,   Adjusted R-squared:  0.05046
F-statistic: 2.753 on 3 and 96 DF,  p-value: 0.04674
```

Graph:



Significant Term: **Education**

Interpretation for the best fit model:

For our dataset of wages, the best fit model is the multiple regression model with **Wage** as independent variable, **education, and female** as independent variables. As we know that higher the adjusted R squared value, better the model is and hence we came to this conclusion based on adjusted R squared value which is 0.08202 for the multiple regression model. In all the models we tested, most of the models has "**education**" as significant term.The following is the summary of that multiple regression model race_B.out:

```
Call:
lm(formula = wage ~ education + female, data = new_data)
Residuals:
    Min      1Q  Median      3Q     Max
-16.719  -6.762  -2.063   4.244  33.570
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.4905     5.6175   0.977  0.33080
education     1.1743     0.4019   2.922  0.00433 **
```

```
female       -3.7590     2.0842  -1.804  0.07440 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 10.32 on 97 degrees of freedom
Multiple R-squared:  0.1006,  Adjusted R-squared:  0.08202
F-statistic: 5.423 on 2 and 97 DF,  p-value: 0.005856
```

Equation for the model:

Y= 20.98202 + 0.01295 *X2 – 3.12584 * X3

where β0(intercept)= 20.98202 and β1(slope)= 0.01295, β2(slope)= – 3.12584

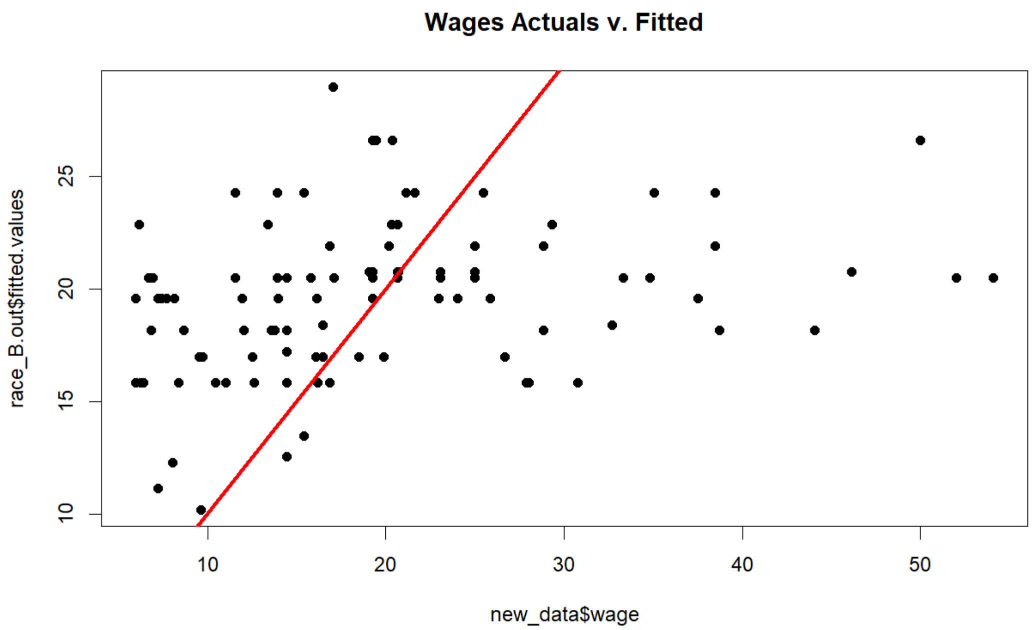
IV's and DV's

Here, the **independent variables are Education, and Female** whereas dependent variable is wage.

LINE INTERPRETATIONS:

LINEARITY:

```
##LINEARITY
plot(new_data$wage,
     race_B.out$fitted.values,
     pch=19,main="Wages Actuals v. Fitted")
abline(0,1,lwd=3,col="red")
```



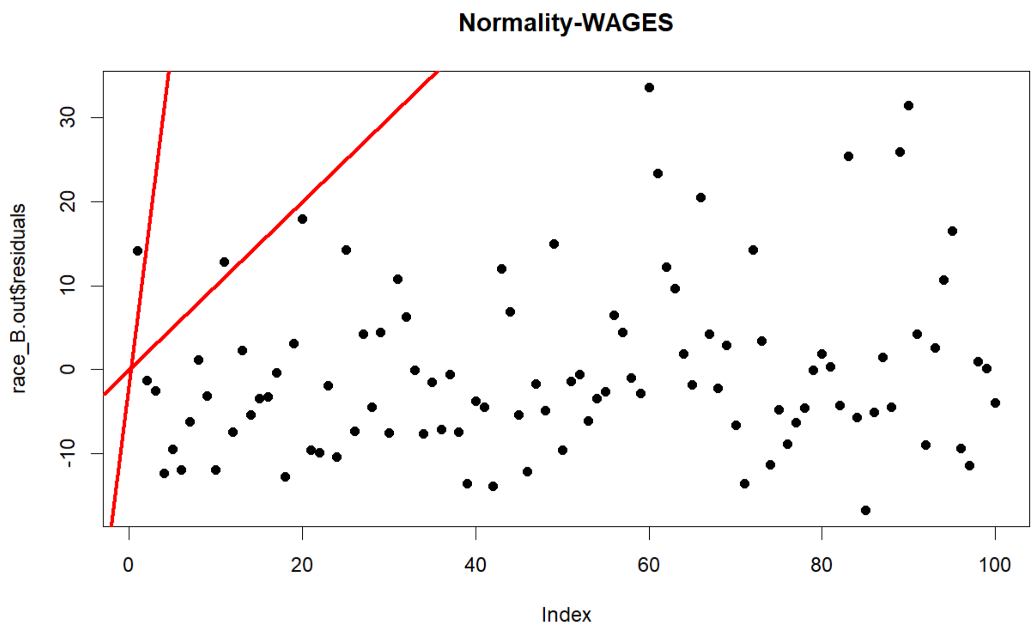
The data points are not in conformity with the abline ,hence we say that our data is not in conformity with linearity.

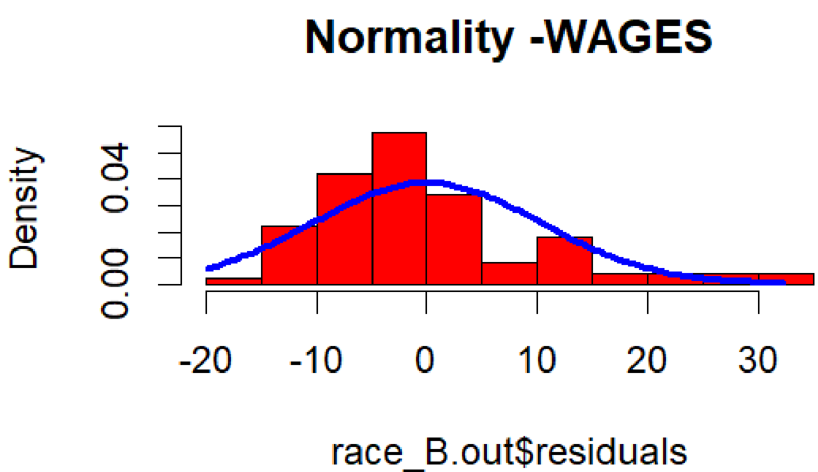
NORMALITY:

CODE:

```
##Normality
plot(race_B.out$residuals,
     pch=19,main="Normality-WAGES")
abline(0,1,lwd=3,col="red")
qqline(race_B.out$residuals,lwd=3,col="red")
hist(race_B.out$residuals,col="red",probability = TRUE,
     main="Normality -WAGES")
curve(dnorm(x,0,sd(race_A.out$residuals)),
      from=min(race_A.out$residuals),
      to=max(race_A.out$residuals),
      lwd=3,col="blue",add=TRUE)
```

RESULT:





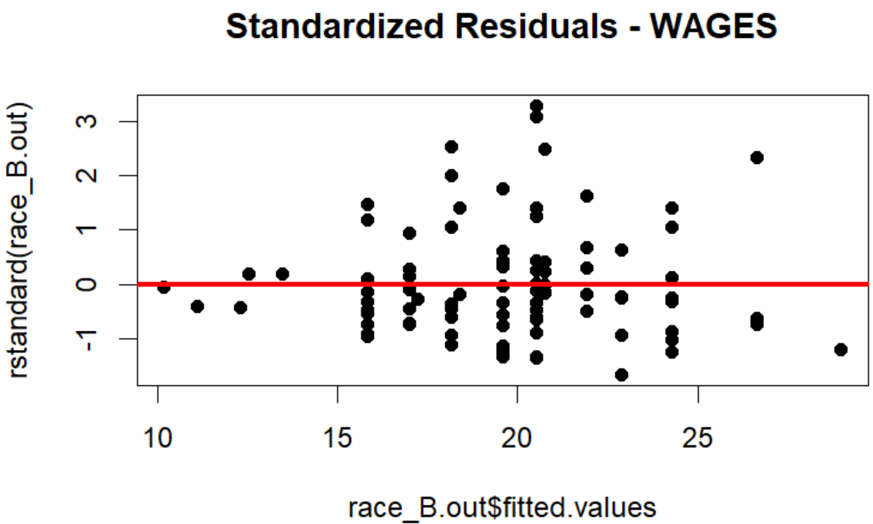
The data points are not ideally normally distributed from the ggplot, but from the above histogram we can say that the data points are partially in conformity with normality.

EQUALITY OF VARIANCES:

```
##EQUALITY OF VARIANCES
plot(race_B.out$fitted.values,rstandard(race_B.out),pch=19,
     main="Standardized Residuals - WAGES")
```

UNIVERSITY OF SOUTH FLORIDA

```
abline(0,0,col="red",lwd=3)
```



The data has some scattering but is good in terms of equality of variances.Hence the data is in conformity with Equality of variances, hence we are seeing homoscedasticity. We don't have a line spread.

Two types of prediction confidence intervals resulting from independent variable values

A prediction interval captures the uncertainty around a single value whereas confidence interval captures uncertainty around the mean predicted values.

```
updata=data.frame(education=50,female=1)
predict(race_B.out,updata,interval="predict")
```

```
  fit        lwr     upr
1 60.44612 25.0484 95.84383
```

Interpretation :From the above predicted values in the interval="predict", we can say that if an employee has an education of 50 years and is "female" ,then she can earn upto $60.44612 per hour and least and highest pay in present being $25.0484 and $95.84383 respectively.

```
updata=data.frame(education=50,female=0)
predict(race_B.out,updata,interval="confidence")
```

```
     fit        lwr       upr
1 64.20511   34.8893    93.52093
```

UNIVERSITY OF SOUTH FLORIDA

Interpretation :From the above predicted values in the interval=”confidence”, we can say that if an employee has an education of 50 years and is “male” ,then he can earn up to $64.20511 per hour and least and highest pay being $34.8893 and $93.52093 respectively during some time in the future.

CODE:

```
# QMB FINAL PROJECT
# SUBMITTED BY
#Karna Haritha                  U46427632
#Sri Venkata Likhitha Duggi     U37870238
#Rahul Gomedhikam               U10270893


rm(list=ls())
library(rio)
```

UNIVERSITY OF SOUTH FLORIDA

```
library(moments)
Datas=import("C:/Users/KUSHAL/Downloads/CPS2 (1).xlsx")
names(Datas)
set.seed(100)
new_data=Datas[sample(1:nrow(Datas),100),]
new_data

str(new_data)

new_data$female=as.factor(new_data$female)
str(new_data)

#_______________________SIMPLE REGRESSION__________________________
race1.out=lm(wage~education,data=new_data)
summary(race1.out)
plot(new_data$wage,race1.out$fitted.values,
    pch=15,
    xlim=c(0,80),ylim=c(0,80),
    main="Wage vs Education")
abline(0,1,lwd=3,col="red")

race2.out=lm(wage~experience,data=new_data)
summary(race2.out)
plot(new_data$wage,race2.out$fitted.values,
    pch=15,
    xlim=c(0,80),ylim=c(0,80),
    main="Wage vs Experience")
abline(0,1,lwd=3,col="red")

race3.out=lm(wage~female,data=new_data)
summary(race3.out)
plot(new_data$wage,race3.out$fitted.values,
    pch=15,
    xlim=c(0,80),ylim=c(0,80),
    main="Wage vs Gender(Female)")
abline(0,1,lwd=3,col="red")

#_____________MULTIPLE REGRESSION_____________
race_A.out=lm(wage ~ education+experience,data=new_data)
summary(race_A.out)
plot(new_data$wage,race_A.out$fitted.values,
    pch=15,
    xlim=c(0,80),ylim=c(0,80),
    main="Wage vs (Education, Experience)")
abline(0,1,lwd=3,col="red")
```

```
race_B.out=lm(wage ~ education+female,data=new_data)
summary(race_B.out)
plot(new_data$wage,race_B.out$fitted.values,
    pch=15,
    xlim=c(0,80),ylim=c(0,80),
    main="Wage vs (Education, female)")
abline(0,1,lwd=3,col="red")

race_C.out=lm(wage ~ experience+female,data=new_data)
summary(race_C.out)
plot(new_data$wage,race_C.out$fitted.values,
    pch=15,
    xlim=c(0,80),ylim=c(0,80),
    main="Wage vs (Experience, female)")
abline(0,1,lwd=3,col="red")

#_______________FULL REGRESSION MODEL_________________
full.out=lm(wage ~ education+experience+female,data=new_data)
summary(full.out)
plot(new_data$wage,full.out$fitted.values,
    pch=15,
    xlim=c(0,80),ylim=c(0,80),
    main="Wage vs (Experience, Education, female)")
abline(0,1,lwd=3,col="red")



#________________MULTIPLE REGRESSION WITH INTERACTION______________

employ1_int= lm(wage ~ education+experience+I(education*experience),data=new_data)
summary(employ1_int)

plot(new_data$wage,employ1_int$fitted.values,
    pch=15,
    xlim=c(0,80),ylim=c(0,80),
    main="Multiple regression with Interaction")
abline(0,1,lwd=3,col="red")

#_________________SIMPLE REGRESSION USING SQUARED TERMS________________
races_reg1.out= lm(wage ~ education+experience+I(education^2),data=new_data)
summary(races_reg1.out)
plot(new_data$wage,races_reg1.out$fitted.values,
    pch=15,
    xlim=c(0,80),ylim=c(0,80),
    main="Simple regression model using squared terms")
abline(0,1,lwd=3,col="red")
```

UNIVERSITY OF SOUTH FLORIDA

```
races_reg2.out= lm(wage ~ education+experience+I(experience^2),data=new_data)
summary(races_reg2.out)
plot(new_data$wage,races_reg2.out$fitted.values,
    pch=15,
    xlim=c(0,80),ylim=c(0,80),
    main="Simple regression model using squared terms")
abline(0,1,lwd=3,col="red")
#________________________-LINE_____________________
##LINEARITY
plot(new_data$wage,
    race_B.out$fitted.values,
    pch=19,main="Wages Actuals v. Fitted")
abline(0,1,lwd=3,col="red")

##Normality
plot(race_B.out$residuals,
    pch=19,main="Normality-WAGES")
abline(0,1,lwd=3,col="red")
qqline(race_B.out$residuals,lwd=3,col="red")
hist(race_B.out$residuals,col="red",probability = TRUE,
    main="Normality -WAGES")
curve(dnorm(x,0,sd(race_A.out$residuals)),
     from=min(race_A.out$residuals),
     to=max(race_A.out$residuals),
     lwd=3,col="blue",add=TRUE)
##EQULAITY OF VARIANCES
plot(race_B.out$fitted.values,rstandard(race_B.out),pch=19,
    main="Standardized Residuals - WAGES")
abline(0,0,col="red",lwd=3)


#______________________PREDICT___________
updata=data.frame(education=50,female=1)
predict(race_B.out,updata,interval="predict")

updata=data.frame(education=50,female=0)
predict(race_B.out,updata,interval="confidence")
```

UNIVERSITY OF
SOUTH FLORIDA