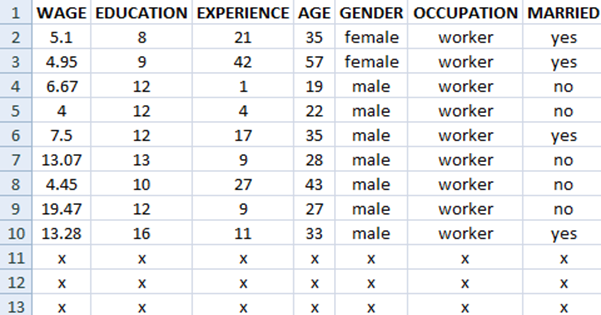
Introduction

Wage is a monetary compensation (or remuneration, personnel expenses, labor) paid by an employer to an employee in exchange for work done. Payment may be calculated as a fixed amount for each task completed, or at an hourly or daily rate, or based on an easily measured quantity of work done.

The data used in ourprojectis a **Cross-section data originating from the May 1985 Current Population Survey by the US Census Bureau**.

It consists of a random sample of 534 persons, with information on wages and other characteristics of the workers, including gender, number of years of education, years of work experience, occupational status and marital status. The data talks about how the aforementioned factors determine wage.

Data is of the form:



Here, **Wage** is the **response** variable.

**Education, Experience and Age** are **numeric** regressors and **Gender, Occupation and Married** are **categorical** regressors..

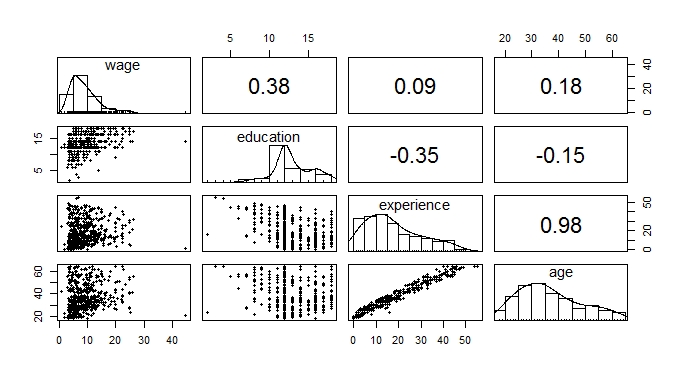
1. Gender has 2 categories, namely, Male and Female.
2. Occupation has 6 categories namely, Management, Office, Sales, Services, Technical and Worker.
3. Married has two categories, Yes and No.

Our **goal** in this project is to observe how these regressors are able to explain wage by considering a linear model of wage on them.

**Exploratory Analysis**

Exploratory data analysis is an approach to analyze data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily it is for seeing what the data can tell us other than the formal modelling or hypothesis testing task.

First, we graphically explore the **numeric variables.**

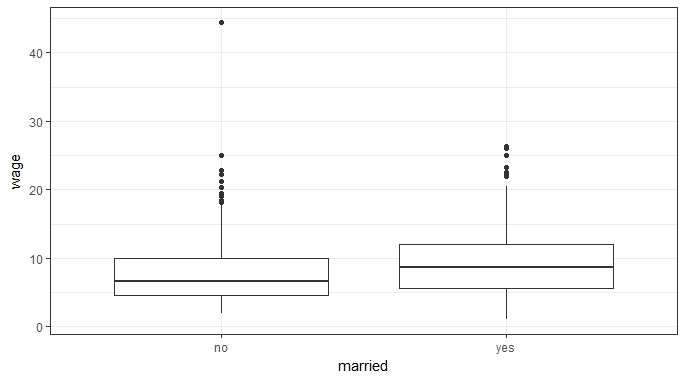
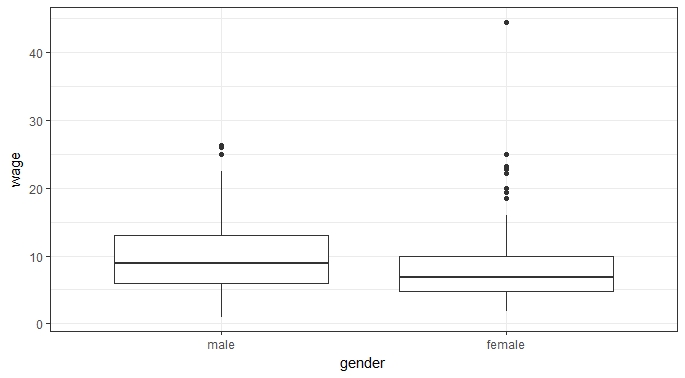


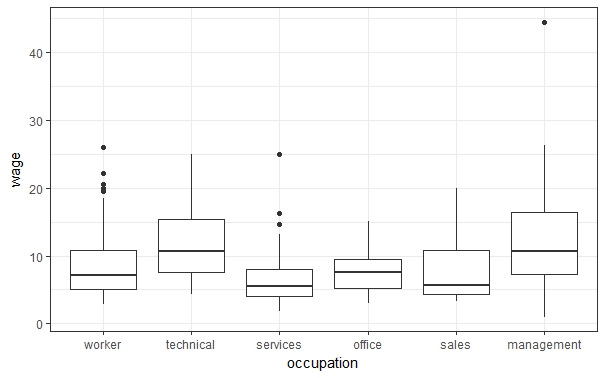
We observe that:

1. Wage is positively skewed, which implies presence of outliers
2. Age, education, experience are also positively skewed, which implies there might be some high leverage points
3. None of the individual correlations between the regressors and the response are very high
4. Age and experience are highly correlated.
5. Education has low correlation with age and experience.
6. The scatter plot of wage with other regressors doesn’t show any apparent relationship.

Next we explore the **categorical variables**:

In order to explore their relationship with wage, we consider boxplots:





We observe that

1. Median wage of males and married persons are higher than females and unmarried persons resp.
2. Some occupations are paying more than the others.

There is a female, unmarried, working in management position who is seems to be an outlier.

The variable occupation has six levels. Reducing the number of levels might lower the complexity of the problem. We consider a linear model of wage on occupation; the summary table is as follows:

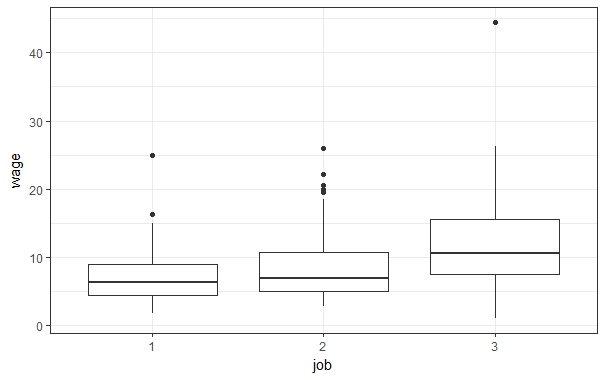
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients** | **Estimate** | **Standard Error** | **t value** | **Pr(> | t | )** |
| Intercept | 8.4265 | 0.3743 | 22.513 | < 2e-16 \*\*\* |
| occupationtechnical | 3.5210 | 0.5901 | 5.967 | 4.44e-09 \*\*\* |
| occupationservices | -1.8890 | 0.6351 | -2.974 | 0.00307 \*\* |
| occupationoffice | -1.0039 | 0.6045 | -1.661 | 0.09735 |
| occupationsales | -0.8338 | 0.8457 | -0.986 | 0.32459 |
| occupationmanagement | 4.2775 | 0.7331 | 5.835 | 9.40e-09 \*\*\* |

Thus,

People in ‘sales’ aren’t earning significantly different from workers. Peoples in ‘Services’ and ‘office’ are earning significantly less than ‘workers’ whereas ‘management’ and ‘technical’s are earning significantly more than workers.

We can combine them in groups of two as job category 1, 2 & 3 which might be interpreted as low, medium and high paying jobs. We call this new variable as ‘job’.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients** | **Estimate** | **Standard Error** | **t value** | **Pr(> | t | )** |
| Intercept | 7.0144 | 0.3486 | 20.12 | <2e-16 \*\*\* |
| job2 | 1.2487 | 0.4840 | 2.58 | 0.0102 \* |
| Job3 | 5.1931 | 0.5082 | 10.22 | < 2e-16 \*\*\* |

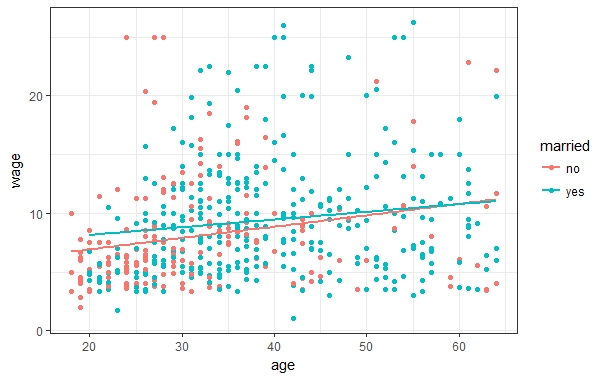
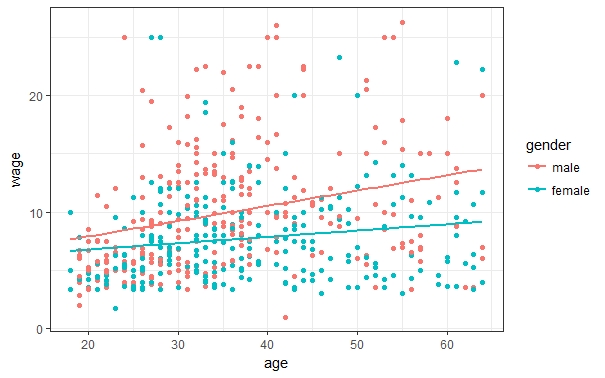


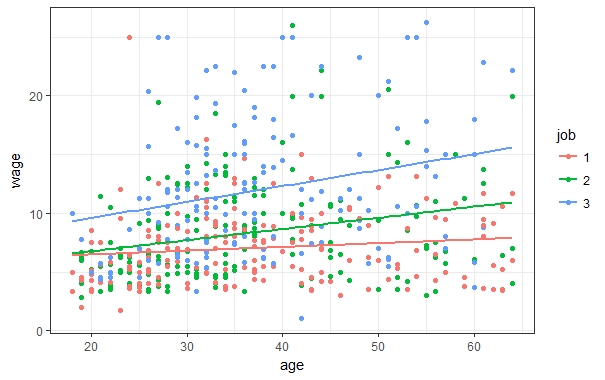
Thus from six groups we lowered to three groups all of which are intuitive and significant.

There is a female, unmarried, working in management position who is seems to be an outlier. We remove this point by making a subjective judgement.

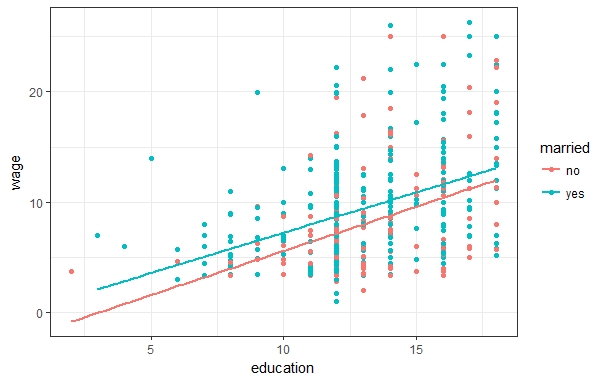
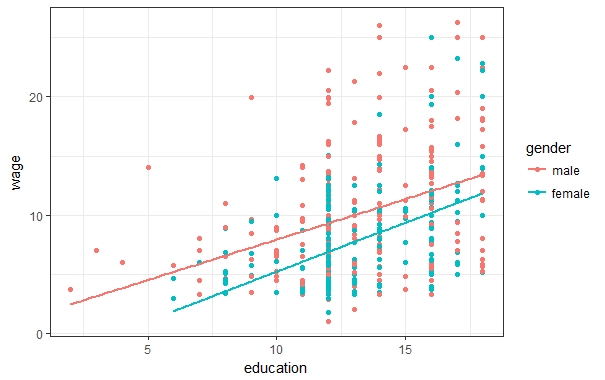
Next we explore the possible **interactions** between the variables

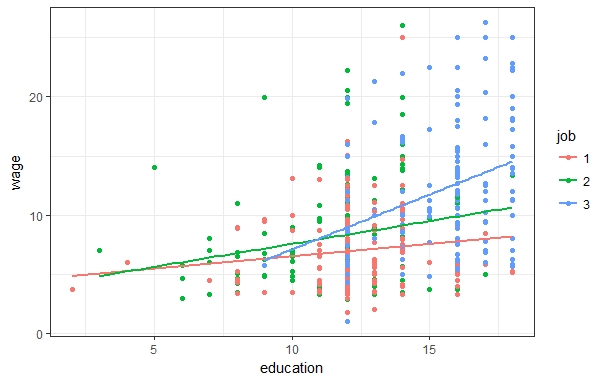
1. **Gender, marital status and job with age:**





1. **Gender, marital status, job with education:**

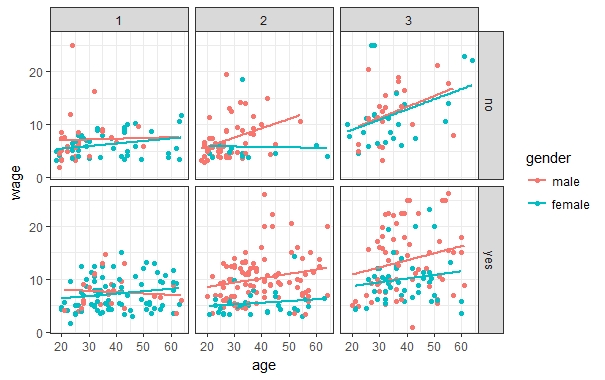


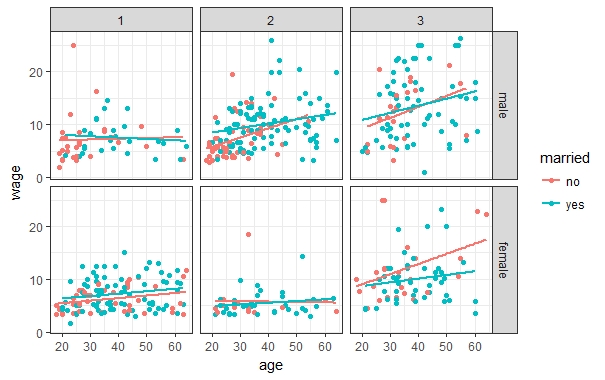


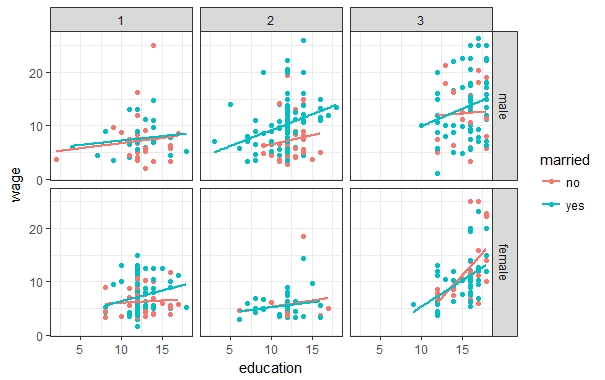
**Observations:**

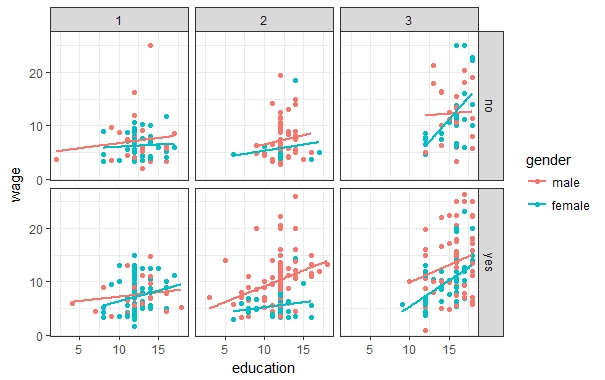
1. Thus although being of same age and education level, males seem to earn more than females and the growth of wage for males with age seem to be faster than females.
2. Married and unmarried peoples having same age does not seem to earn differently although across same education levels, married peoples seem to earn more.
3. High paying jobs seem to pay a lot more and their growth with rising education levels and age is also clear.

In order to further understand these interactions properly we consider another set of plots by making more finer groups.









**Observations:**

1. The gender gap seems to be very prominent in job category 2 where males are earning higher than females. Also in job 3, married males are earning much higher than married females.
2. Unmarried females in job category 3 are earning higher than married females of similar age, although with similar education level, this gap is not prominent.
3. Married males in job category 2 are earning more than unmarried ones with similar educational level although this is not apparent for similar age groups.
4. There isn’t any gender gap or difference between wages between married and unmarried individuals in job category 1.

Thus most of the interactions doesn’t seem to significant, but some of them are seems to be highly significant.

Before modelling wage on these variables we check for the presence of multicollinearity

**Multicollinearity**

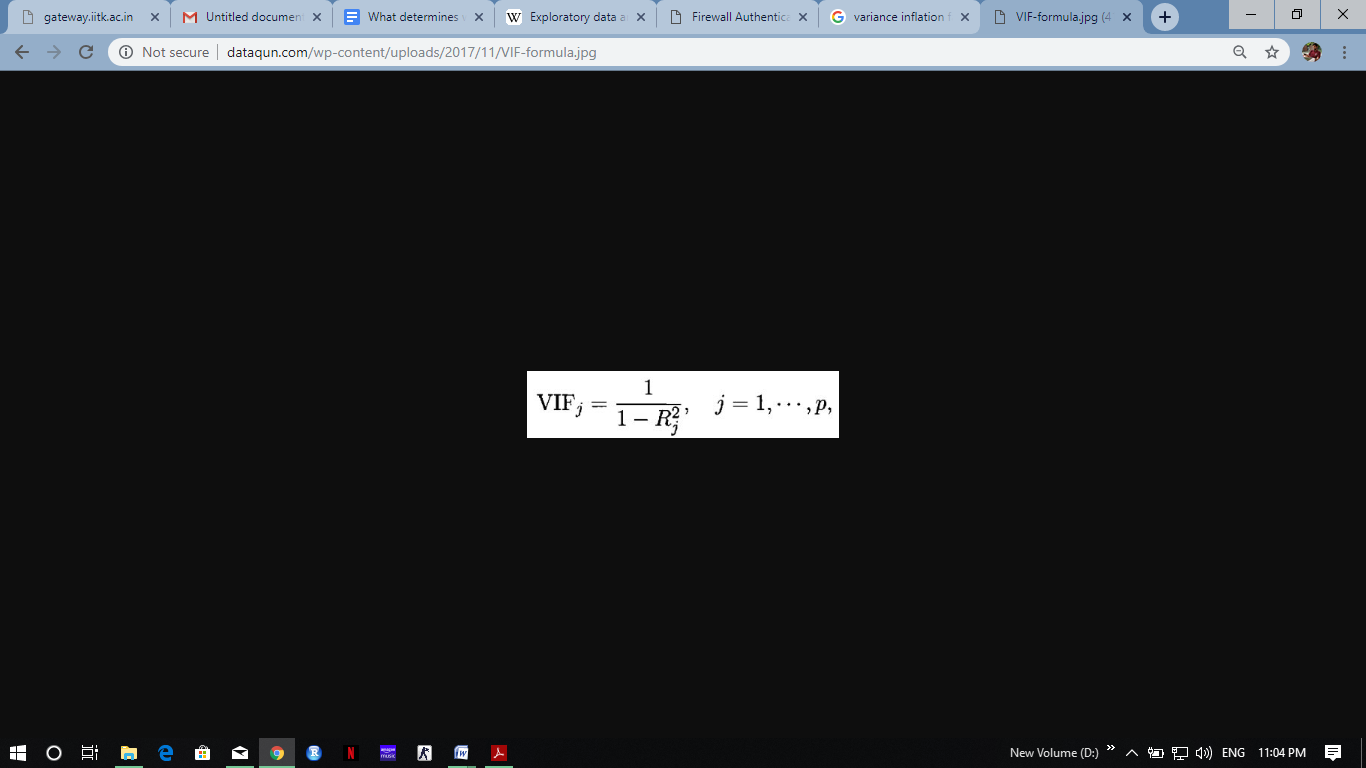
Multicollinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted by the other predictor variables in the model. In some situations the predictors are nearly perfectly linearly related and in such cases the inferences based on the regression model can be misleading and erroneous.

To check multicollinearity in the data, we have used two methods namely:

1. Variance Inflation Factor (VIF) technique: It is used to identify regressors causing multicollinearity.

2. Variance Decomposition Proportion (VDP) Method: It is used to identify the subsets of regressors associated with multicollinearity problem.

**Variance Inflation Factor:**



Rj2 is the coefficient of multiple determination obtained from regressing Xj on other regressors. The VIF value of 5 or more is an indicator of multicollinearity. Large values of VIF indicate multicollinearity leading to poor estimates of associated regression coefficients.

Now we consider the model as,

Wage = β0 + β 1\*age + β2\*experience + β3 \*education + Є,

where **Є** follows usual error assumptions

Table: VIFs

|  |  |  |
| --- | --- | --- |
| **Age** | **Experience** | **Education** |
| 4595.73 | 5130.25 | 229.52 |

It is evident from the table above that the VIF values are **very high.**

**Variance Decomposition Method:**

It is a method to identify subsets of regressors involved in multicollinearity.

Variance decomposition proportions are defined as under:

𝜋𝑖𝑗= ((𝑣𝑖𝑗\*𝑣𝑖𝑗) 𝜆𝑖𝑗) /Σ ((𝑣𝑖𝑗\*𝑣𝑖𝑗) 𝜆𝑖)

Σ𝜋𝑖𝑗=1 𝑓𝑜𝑟 𝑎𝑙𝑙 𝑗

Corresponding to the largest condition index (C.I. > 30), the regressors having high variance decomposition proportions indicates multicollinearity.

Results yielded by Variance Decomposition Method are given in table below:

Table: Variance Decomposition Proportions

|  |  |  |  |
| --- | --- | --- | --- |
| **Condition Index** | **Age** | **Experience** | **Education** |
| 1 | 0 | 0 | 0 |
| 3.436 | 0 | 0.004 | 0.007 |
| 80.913 | 1 | 0.996 | 0.993 |

Large proportions in a row indicate presence of multicollinearity among corresponding regressors but should be checked in consumption with a high(>30) condition index.

It is evident from the table that all of them are linearly related. We remove ‘experience’ from the model as it has the highest VIF and we have the model,

**Wage = β0 + β 1\*age + β2\*education + Є**

Table: VIFs

|  |  |
| --- | --- |
| **Age** | **Education** |
| 1.0228 | 1.022 |

Since the VIFs are less than 5, we have been able to remove multicollinearity in the model.

**Residual Analysis and Model Adequacy checking**

We consider the full model involving main effects and all the plausible interactions.

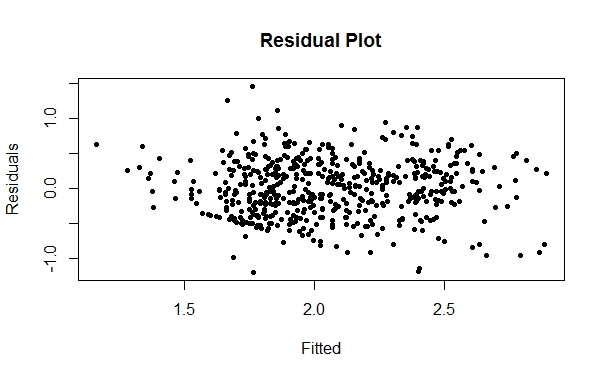
**Wage = β0 + β1\*age + β2\*education + β3\*gender + β4\*married + β5\*job + β 6\*(gender:married) +**

**Β7\*(gender:job) + β8\*(married:job) + β 9\*(gender:age) + β10\*(gender:education) +**

**β11\*(married:age) + β12\*(married:education) + β13\*(job:age) + β14\*(job:education) +**

where Є follows usual error assumptions.

First, we plot the residuals against the fitted values.

**Observations:**

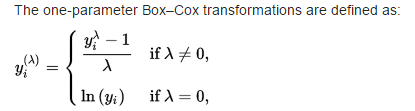
1. The plot doesn’t show any systematic pattern and appears random.
2. Indicates presence of mild heteroscedasticity.

Also, we have previously found that wage is positively skewed, thus, prior to removing any influential observations from the dataset, we consider transforming the response since an outlier in this model might not be an outlier in the transformed model.

We consider transforming the response using Box Cox method.

**Box Cox Method**

Suppose that we wish to transform y to correct non normality and non constant variance . A useful class of transformation is the power transformation where the lambda parameter is needed to be determined. It is defined as-

****

Here, we get the estimated value of λ=0.

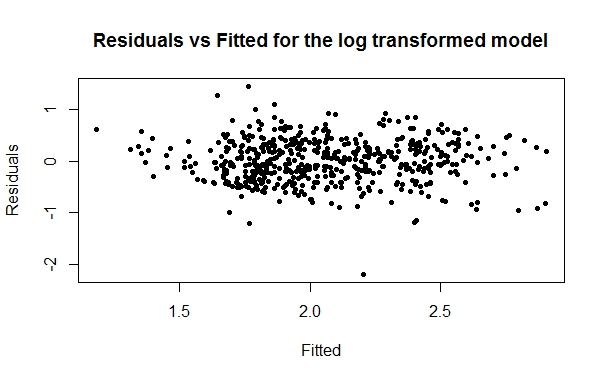
Thus, we apply log transformation to the response and the model becomes,

**Log(wage) = β0 + β1\*age + β2\*education + β3\*gender + β4\*married + β5\*job + β 6\*(gender:married) +**

**Β7\*(gender:job) + β8\*(married:job) + β 9\*(gender:age) + β10\*(gender:education) +**

**β11\*(married:age) + β12\*(married:education) + β13\*(job:age) + β14\*(job:education) +**

We plot the residuals against the fitted values again for this model.



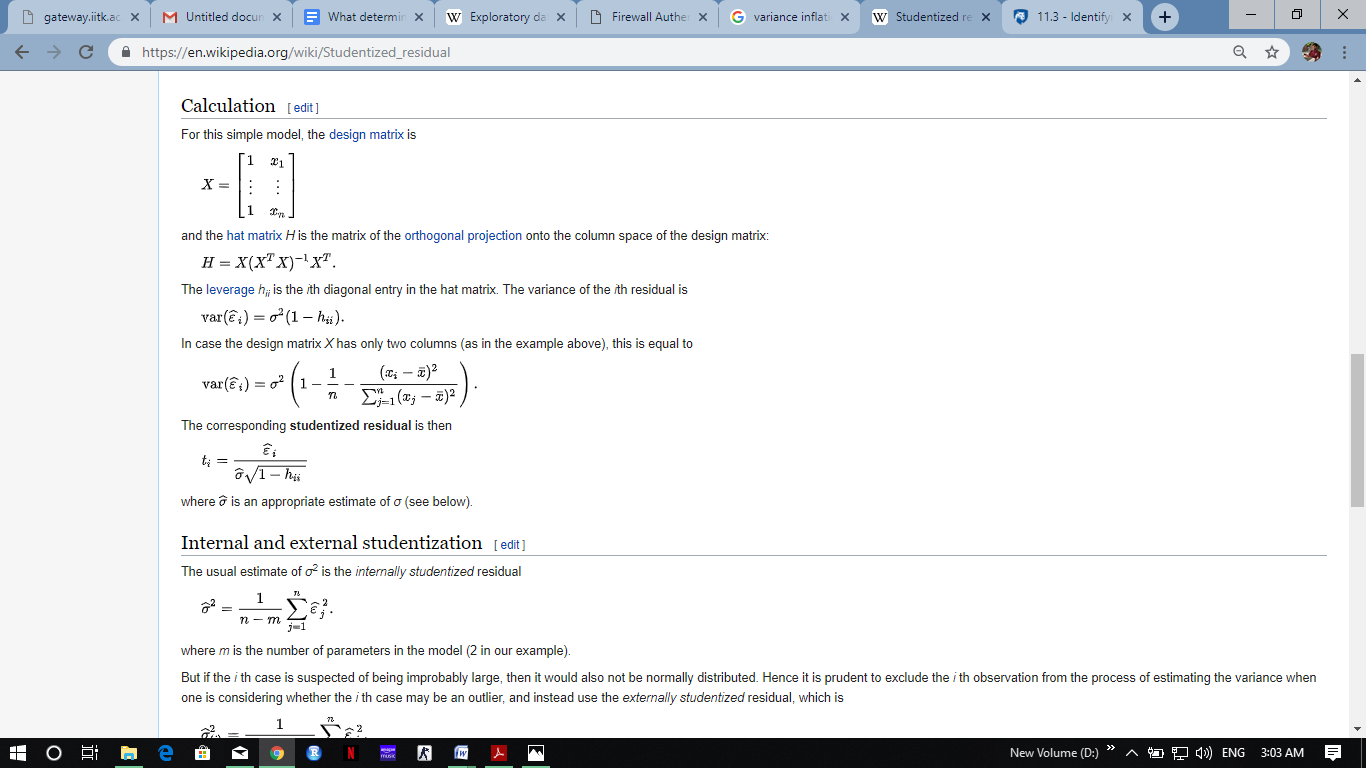
Again the plot doesn’t show any systematic pattern in the plot, hence, heteroscedasticity seems to be removed.

**Detection of Outliers**

An **outlier** is an observation point that is distant from other observations. An **outlier** may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.

In order to find outliers, we consider the studentized residuals.

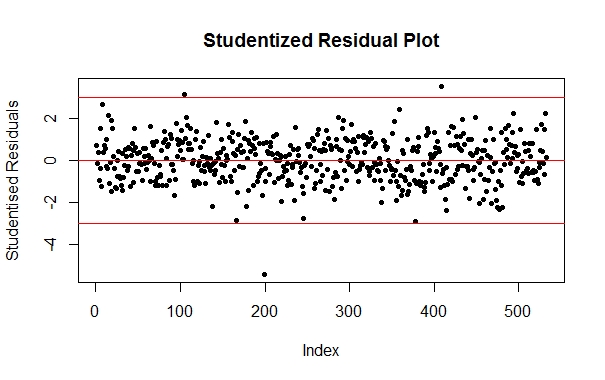
The Standardized Residual is defined as the residual (i) divided by its standard error (), where the residual is the difference between the data response and the fitted response.



where= Mean Square (Res),

hii= element of

where i=1,2,...,n. |ti| > 3 indicate the presence of an outlier.



**Observation:** The above graph indicates that there are 3 points outside the range (-3,3). Thus, we notice the presence of 3 outliers.

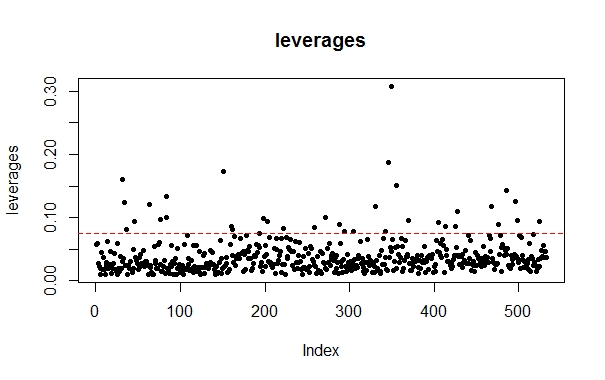
**Detection of leverage points**

A point is determined as a leverage point based on the location of the point on the x-space and hence remote points impart more on the parameters of the model.

Generally, a point is considered to be a leverage point if,

hii > 2p/n

where hii is the diagonal element of the matrix and p denotes the number variables (independent and dependent) in the model.



A substantial number of points seem to have very high leverages.

In order to understand which of these outliers and leverage points are influential, we consider the Cook’s distance of each residual.

**Detection of Influential points**

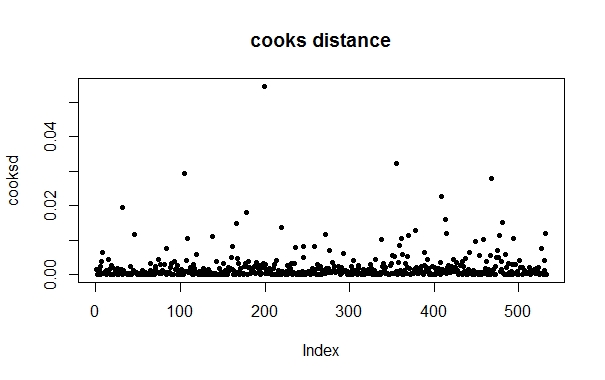
Cook's distance or Cook's *D* is a commonly used estimate of the [influence](https://en.wikipedia.org/wiki/Influential_observation) of a data point Data points with large [residuals](https://en.wikipedia.org/wiki/Residual_(statistics)) ([outliers](https://en.wikipedia.org/wiki/Outlier)) and/or high [leverage](https://en.wikipedia.org/wiki/Leverage_(statistics)) may distort the outcome and accuracy of a regression. Cook's distance measures the effect of deleting a given observation. Points with a large Cook's distance are considered to merit closer examination in the analysis.

Cook's distance {\displaystyle D\_{i}}Di of observation {\displaystyle i\;({\text{for }}i=1,\dots ,n)}i (i=1,2,...n)  is defined as the sum of all the changes in the regression model when observation {\displaystyle i}i is removed from it

C:\Users\Asif\Pictures\Screenshots\Screenshot (18).png

where j(i) is the fitted response value obtained when excluding i, and s2 is the MSE.

A point is said to be influential if Di > 1.



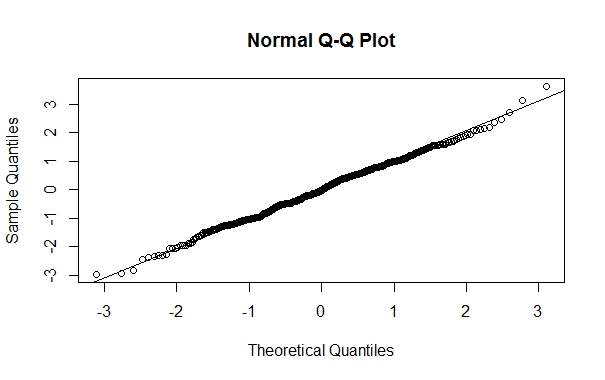
Although none of the cooks distances is higher than 1, the point which has the highest value of cooks distance also turns out to be an outlier. We consider removing this point.

**QQ Plot**

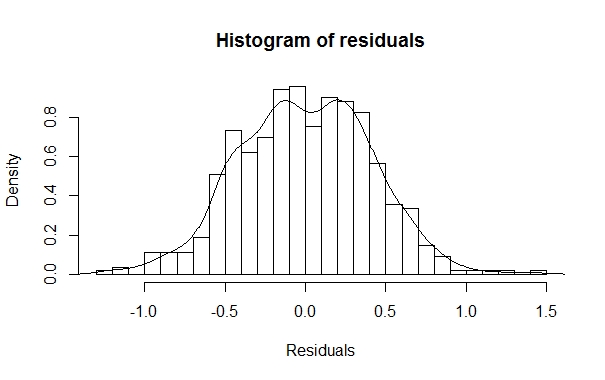
In our data, after controlling for heteroscedasticity and removing potential influential points we check whether the normality assumption is true.

Q-Q plot is used to check normality assumption of error.

Under this, we plot ordered standardized/studentized residuals against the theoretical quantiles. A 45° line is considered to be an ideal situation.



**Histogram of Residuals**



**Observation:** We observe that almost all the points fall on the 45 degree straight line in the QQ plot. Combined with the histogram these justifies the normality assumption.

**Model Building**

Consider again the full model,

**Log(wage)= β0\*age + β1\*education + β2\*gender + β3\*married + β4\*job + β 5\*(gender:married) +**

**β6\*(gender:job) + β7\*(married:job) + β 8\*(gender:age) + β9\*(gender:education) +**

**β10\*(married:age) + β11\*(married:education) + β 12\*(job:age) + β13\*(job:education) +**

The summary table is as follows:

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

(Intercept) 1.195313 0.317740 3.762 0.000188 \*\*\*

age 0.009242 0.004354 2.122 0.034282 \*

education 0.024476 0.020494 1.194 0.232904

genderfemale -0.223506 0.282786 -0.790 0.429677

marriedyes 0.045082 0.297897 0.151 0.879773

job2 -0.352434 0.338327 -1.042 0.298043

job3 -0.039431 0.372631 -0.106 0.915769

genderfemale:marriedyes -0.089378 0.087966 -1.016 0.310086

genderfemale:job2 -0.377931 0.099153 -3.812 0.000155 \*\*\*

genderfemale:job3 -0.111442 0.111198 -1.002 0.316725

marriedyes:job2 0.049712 0.103268 0.481 0.630446

marriedyes:job3 -0.212751 0.116844 -1.821 0.069219 .

age:genderfemale -0.001723 0.003603 -0.478 0.632653

education:genderfemale 0.019475 0.017829 1.092 0.275216

age:marriedyes -0.002593 0.003537 -0.733 0.463845

education:marriedyes 0.016535 0.019318 0.856 0.392426

age:job2 0.006291 0.004239 1.484 0.138379

age:job3 0.005599 0.004271 1.311 0.190439

education:job2 0.025325 0.021830 1.160 0.246549

education:job3 0.024624 0.021995 1.120 0.263424

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4147 on 512 degrees of freedom

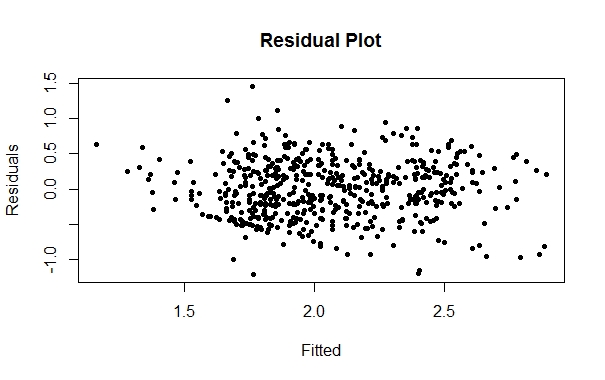
Multiple R-squared: 0.3763, Adjusted R-squared: 0.3531

F-statistic: 16.26 on 19 and 512 DF, p-value: < 2.2e-16

**Observations:**

1. The full model (consists of all the main effects and all the plausible interaction effects) only explains about 37% of the variation in wage.
2. Most of the terms in the model seem to be insignificant when considered together.
3. The intercept term is highly significant, which indicates that the chosen regressors are not sufficient to explain wage.

We again plot the residuals against the fitted values.



The residual plot does not show any systematic pattern.

Since except two, most of the interactions seem to be insignificant, in order to reduce the complexity of the model we retain the main effects and significant interactions and remove all the insignificant interactions. These two interactions also seemed significant in the beginning when the exploratory analysis was conducted.

The two interactions considered are indicators of females which are at job2 and married people who are in job3 resp.

Next with those variables, we build our optimum model by **backward selection.**

**Log(wage)=wage\*= β0 + β1\*age + β2\*education + β3\*(genderfemale) + β4\*(marriedyes) + β5\*(job2) +β6\*(job3) + β7\*(female:job2) + β8\*(married:job3) +**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

(Intercept) 0.705887 0.139724 5.052 6.05e-07 \*\*\*

age 0.010796 0.001644 6.567 1.24e-10 \*\*\*

education 0.061695 0.008610 7.165 2.66e-12 \*\*\*

genderfemale -0.146229 0.047459 -3.081 0.00217 \*\*

marriedyes 0.146621 0.046675 3.141 0.00178 \*\*

job2 0.189667 0.057162 3.318 0.00097 \*\*\*

job3 0.423173 0.077697 5.446 7.92e-08 \*\*\*

female:job2 -0.343211 0.084130 -4.080 5.22e-05 \*\*\*

married:job3 -0.195488 0.085521 -2.286 0.02266 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4153 on 523 degrees of freedom

Multiple R-squared: 0.3612, Adjusted R-squared: 0.3514

F-statistic: 36.96 on 8 and 523 DF, p-value: < 2.2e-16

All the regressors in the above model are coming out to be significant.

Thus our final model is-

**Log(wage)=wage\*= β0 + β1\*age + β2\*education + β3\*(genderfemale) + β4\*(marriedyes) + β5\*(job2) +β6\*(job3) + β7\*(female:job2) + β8\*(married:job3) +**

**Interpretation:**

1. Wage seems to increase with age and education
2. Females are earning lesser than males with similar age, education level and marital status, particularly females in job category 2 are earning even lesser than other females.
3. Married persons are earning more than unmarried ones having same age,education level and gender, although married persons in job category 3 are earning significantly lesser than other married persons.
4. Technical and management positions are earning much higher than the other categories
5. The intercept term is highly significant, which implies the chosen regressors aren’t sufficient.

**Conclusion:**

This model has only been able to explain 36% of the variation in wage. Thus this isn’t a very suitable model to determine wages. Our choice of regressors might have been poor and the model confirms that wage isn’t explained properly by age, education, gender, marital status and occupation.