REGRESSION TECHNIQUES PROJECT

AUTHORS : Adhiraj Mandal,Bhargob Kakoty and Pritam Dey

PROJECT GUIDE : Swagata Nandi

DATA

Data on racial composition, income and age and value of residential units for each ZIP code in Chicago

In a study of insurance availability in Chicago, the U.S. Commission on Civil Rights attempted to examine charges by several community organizations that insurance companies were redlining their neighborhoods, ie. cancelling policies or refusing to insure or renew. First the Illinois Department of Insurance provided the number of cancellations, non-renewals, new policies, and renewals of homeowners and residential fire insurance policies by ZIP code for the months of December 1977 through February 1978. The companies that provided this information account for more than 70% of the homeowners insurance policies written in the City of Chicago. The department also supplied the number of FAIR plan policies written an renewed in Chicago by zip code for the months of December 1977 through May 1978. Since most FAIR plan policyholders secure such coverage only after they have been rejected by the voluntary market, rather than as a result of a preference for that type of insurance, the distribution of FAIR plan policies is another measure of insurance availability in the voluntary market.

Secondly, the Chicago Police Department provided crime data, by beat, on all thefts for the year 1975. Most Insurance companies claim to base their underwriting activities on loss data from the preceding years, i.e. a 2-3 year lag seems reasonable for analysis purposes. the Chicago Fire Department provided similar data on fires occurring during 1975. These fire and theft data were organized by zip code.

Finally the US Bureau of the census supplied data on racial composition, income and age and value of residential units for each ZIP code in Chicago. To adjust for these differences in the populations size associated with different ZIP code areas, the theft data were expressed as incidents per 1,000 population and the fire and insurance data as incidents per 100 housing units.

The first few rows of our data is given below:

```
kable(head(project_data))
```

...1	race	fire	theft	age	volact	involact	income
60626	10.0	6.2	29	60.4	5.3	0.0	11744
60640	22.2	9.5	44	76.5	3.1	0.1	9323
60613	19.6	10.5	36	73.5	4.8	1.2	9948
60657	17.3	7.7	37	66.9	5.7	0.5	10656
60614	24.5	8.6	53	81.4	5.9	0.7	9730
60610	54.0	34.1	68	52.6	4.0	0.3	8231

race : racial composition in percent minority

fire : fire incidents per 100 housing units

theft : theft per 1000 population

age : percent of housing units built before 1939

volact : new homeowner policies plus renewals minus cancellations and non renewals per 100 housing units

involact: new FAIR plan policies and renewals per 100 housing units

income : median family income

ANALYSIS OF THE PROBLEM

Homeowner's insurance is a form of property insurance that covers losses and damages to an individual's house and to assets in the home. Homeowner's insurance also provides liability coverage against incidents in the home or on the property.
These policies might get rejected under various circumstances, such as,

- If the building is very old.(High values in the **age** column)
- If the building is prone to catch fire. (High values in the **fire** column)
- If the locality is prone to theft incidents. (High values in the **theft** column)

The **volact** column represents the net increase in the number of homeowner policies in different ZIP Codes. If the claim of the Community organizations is true, the percentage minority of a particular race in a region should impact the **volact** negatively. We would like to check if the other factors such as **age, fire, theft, income** depend on **volact**. The **involact** column which gives the number of FAIR plan policies (both new and renewal) can only reveals the number of rejected homwowner policies, but fails to give any information regarding the new policies. Therefore **volact** appears as the best candidate for response variable.

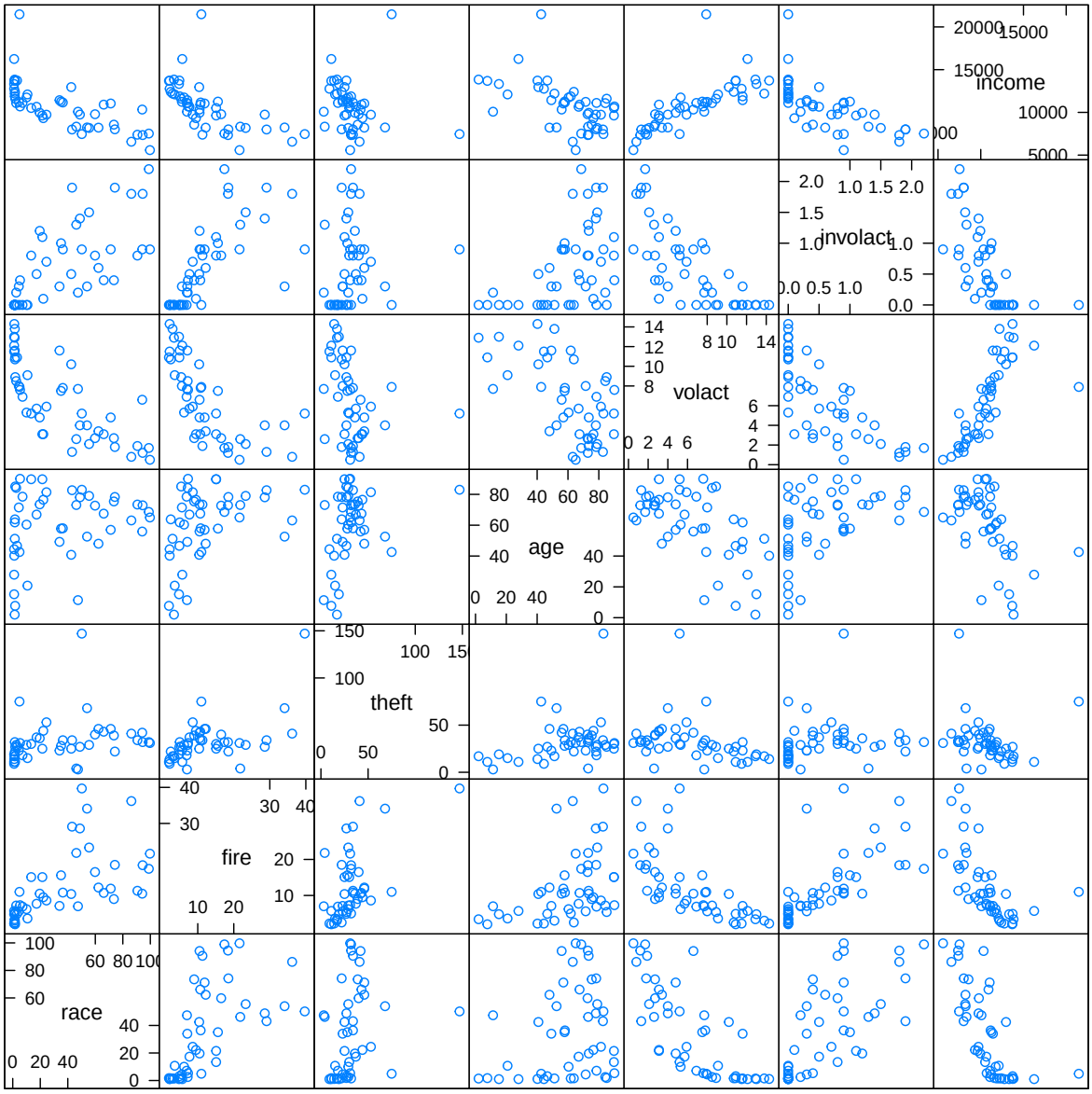
OBJECTIVE

To carry out a Statistical test to investigate if the claim of the Community Organizations is valid.
The **Null Hypothesis** is that there is no significant impact of the race in availing policy from the voluntary market, against the **Alternative Hypothesis** that there is a negative impact of the race in availing the same.

GRAPHICAL REPRESENTATION OF THE DATA

Before going into the analysis, we try to have some overall idea about the data by looking at the **Scatterplot**.

```
library(lattice)
splom(project_data[-1])
```



Scatter Plot Matrix

- From the Scatter Plot Matrix, it can be seen that **theft** has no visible relationships with the other variables.

- Apparently **Volact** is *negatively* related with all other variables except **income** and **theft**.It has a strong *positive* relationship with **income**.
- **Involact** shows exactly the opposite picure of **volact**.
- There is a positive relationship between **race** and **fire**.

- **fire** has a slight positive relationship with **age**. It clearly incdicates that as the houses get old, the chance of catching fire incidents increases.
- **income** shows a linear relationship with every other variable except for **theft**.

SELECTION OF APPROPRIATE MODEL

We shall try to fit various linear models to the given data set by taking ***volact*** as response variable and selecting different subsets of explanatory variables from *race*, *age*, *fire*, *theft*, *income*, *involact*. The ultimate model selection will be based on some popularly used criteria such as ***Residual Sum of Squares (RSS), Adjusted*** R^2***, Akaike Information Criterion (AIC)*** and ***Mallow's CP***. Since the scatter plot shows relationships between most pairs of variables, we will also consider **condition number**(κ) as one of the model selection criteria.

We shall investigate the measures of all such criteria for all possible subsets ($2^6 - 1 = 63$) of these explanatory variables. Then we will sort different models based on **adjusted** R^2 and choose such a model that has lower values of **AIC** and **kappa**.

```
kable(head(subsets[, -1]))
```

predictors	adjr	cp	aic	sbic	kappa
race age fire theft	0.7719758	3.541960	200.1342	68.29620	8.546039
race age fire	0.7707816	2.724635	199.4857	67.08496	7.657315
race age fire income	0.7704256	3.817591	200.4527	68.54124	24.802400
race age fire income theft	0.7687606	5.134724	201.6597	70.27367	28.217151
race age fire involact	0.7687154	4.121664	200.8015	68.80994	9.289235
race age fire income involact	0.7678592	5.291182	201.8426	70.40504	26.859473

The model with explanatory variables **race**, **age**, **fire** and **theft** has largest value of *adjusted* R^2. But the **AIC** and **kappa** value corresponding to this model is less than that of the second model in the table. We also see that the *adjusted* R^2 differ only by a very little amount. Therefore, we shall consider the second model with explanatory variables **race**, **age** and **fire** as our final model. If Y denote the values of the response variable **volact**, the considered model is given below:-

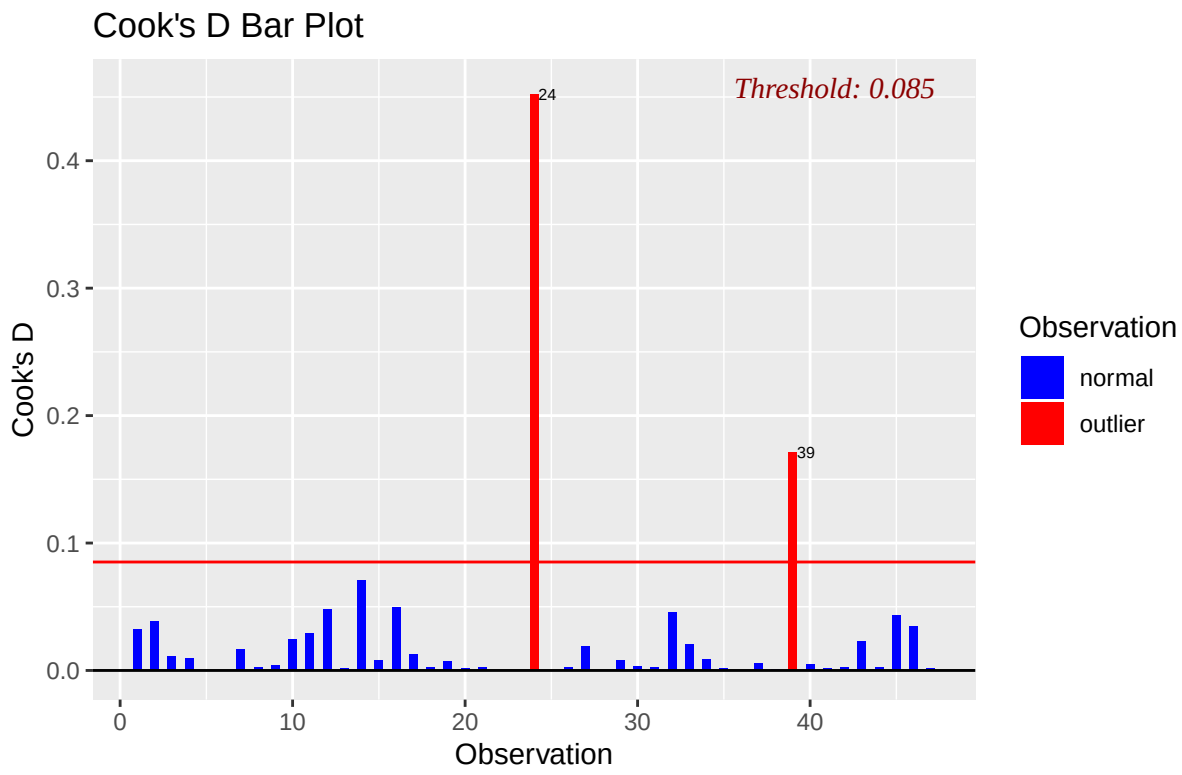
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Here $ X_1,X_2,X_3$ denote the values corresponding to the columns *race, age, fire* respectively and, we assume that, ϵ has the following properties :
* it has 0 expectation * variance covariance matrix of ϵ is $\sigma^2 I_n$ * It is normally distributed

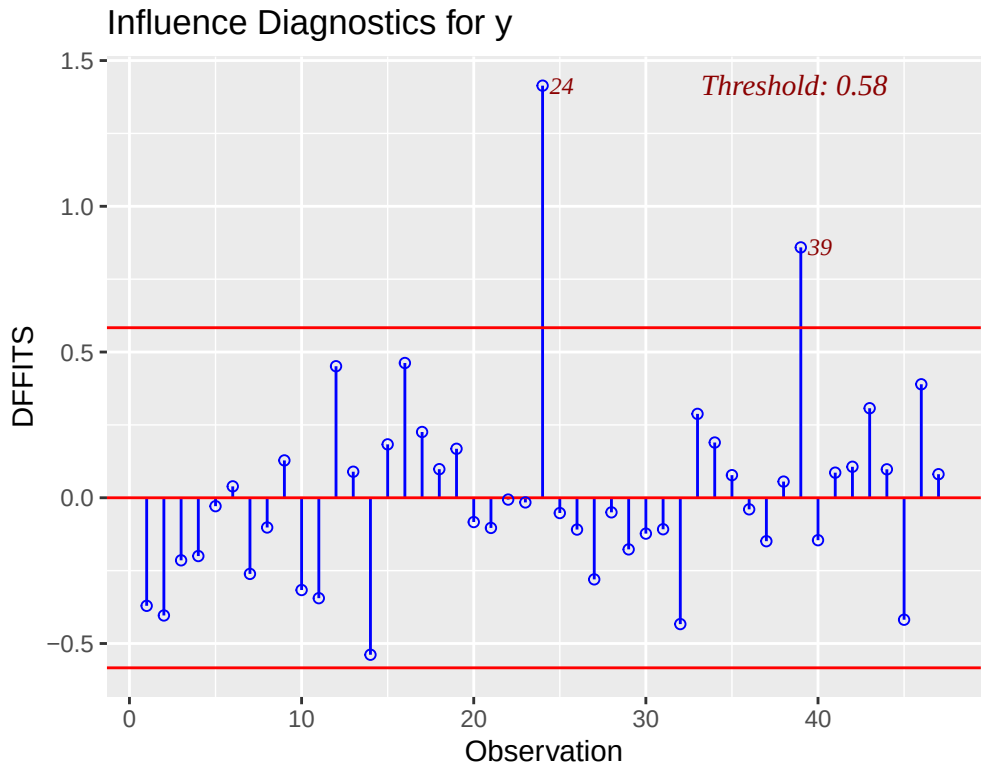
PRELIMINARY ANALYSIS OF THE SELECTED MODEL

We will try to detect influential points in our model with the help of *Cook's D Bar Plot* and *DFFIT plot.*

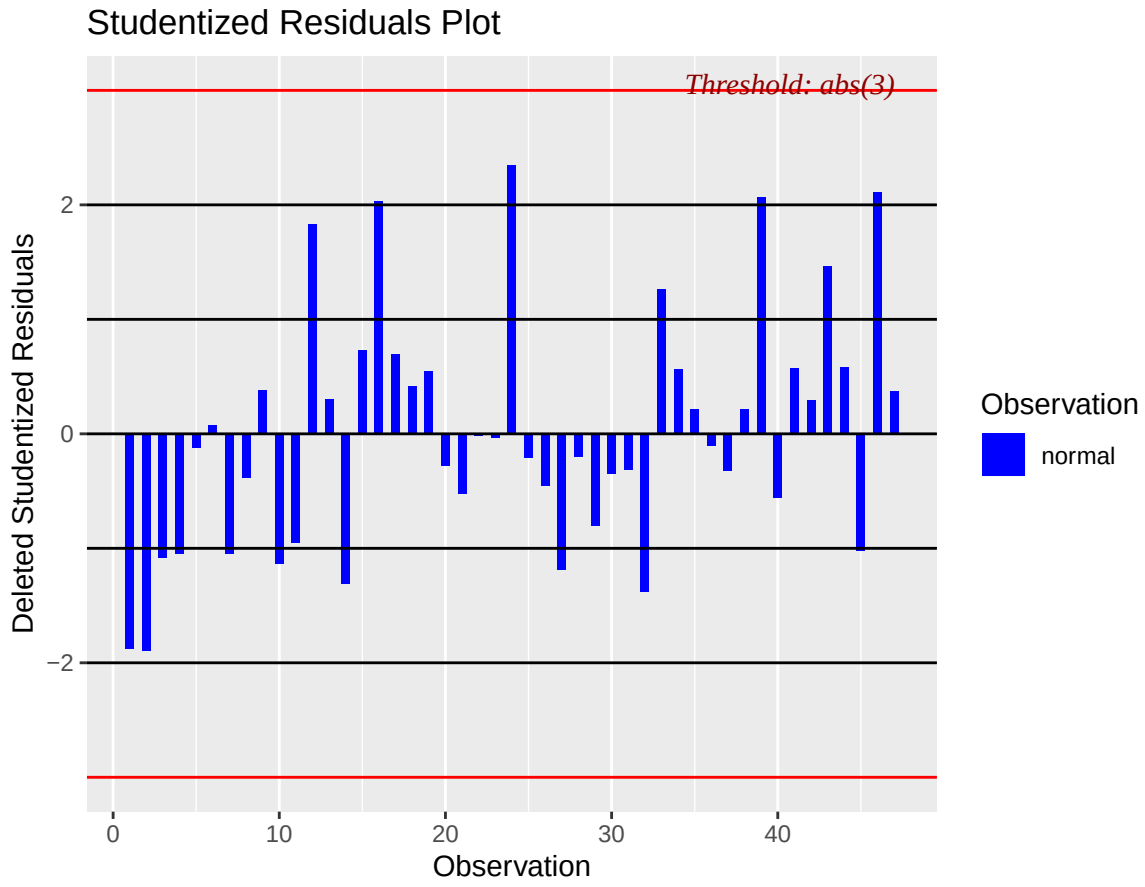
```
X <- model.matrix(~ 0 + race + age + fire, data = data)
y <- data[["volact"]]
M <- lm(y ~ X)
ols_plot_cooksd_bar(M)
```




```
ols_plot_dffits(M)
```



```
ols_plot_resid_stud(M)
```



We have seen that observations$ 24$ and 39 appear as influential points in both *Cook's D bar plot* and *DFFIT plot*. The studentized residual plot also suggests that the points 24 and 39 are outliers in both directions. So we will remove these two points and check if the fit improves anymore. We compare adjusted R^2 and $\hat{\sigma}^2$ values for these two models with and without outliers respectively.

These are the values for the original model

```
summary(M)$adj.r.squared
```

```
## [1] 0.7707816
```

```
summary(M)$sigma^2
```

```
## [1] 3.606295
```

Here are the values after removing outliers

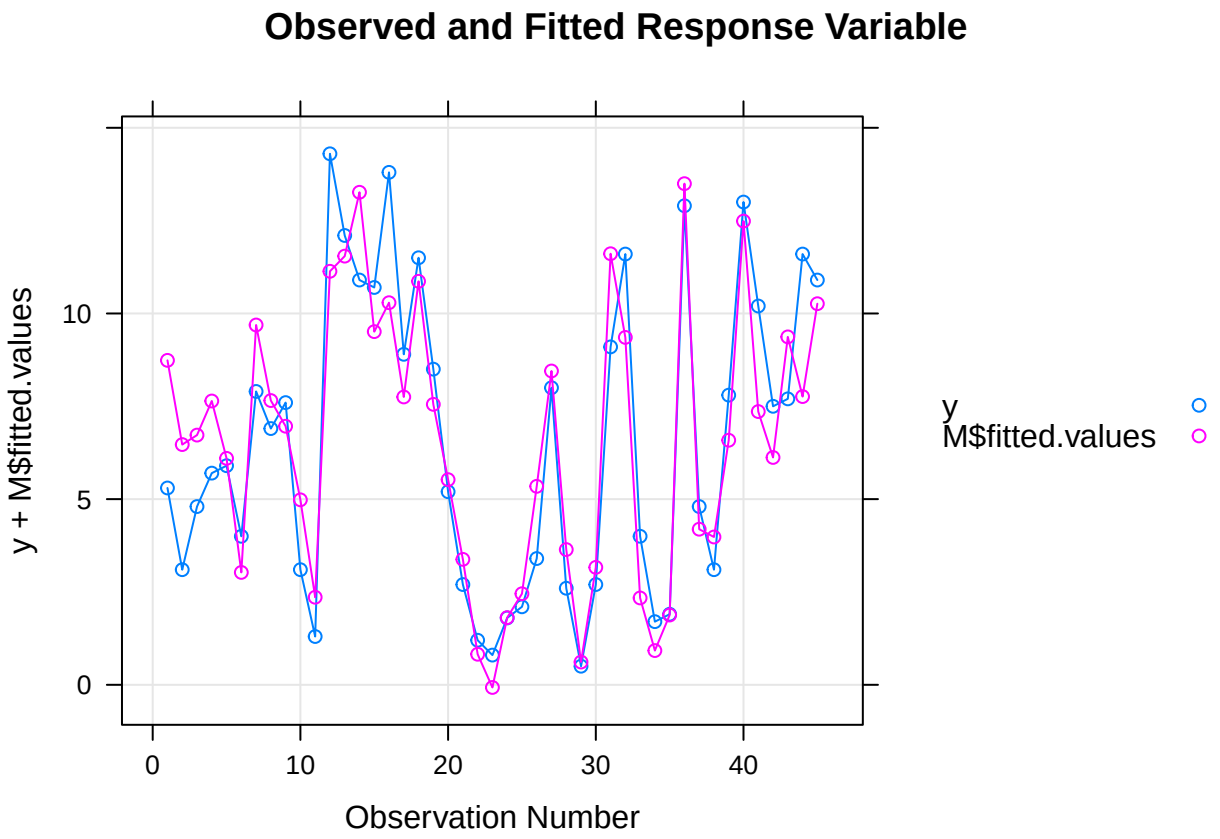
```
X <- model.matrix(~ 0 + race + age + fire, data = data[-c(24, 39), ])
y <- data[["volact"]][-c(24, 39)]
M <- lm(y ~ X)
summary(M)$adj.r.squared
```

```
## [1] 0.8135019
```

```
summary(M)$sigma^2
```

```
## [1] 3.059881
```

```
xyplot(y + M$fitted.values ~1:45, auto.key = list(space = "right"),
       grid = TRUE, main = "Observed and Fitted Response Variable",
       xlab = "Observation Number", type = "b" )
```



We can see from the *Observed vs Fitted* plot that the fit is pretty good. The $\hat{\sigma}^2$ and adjusted R^2 values have also improved.

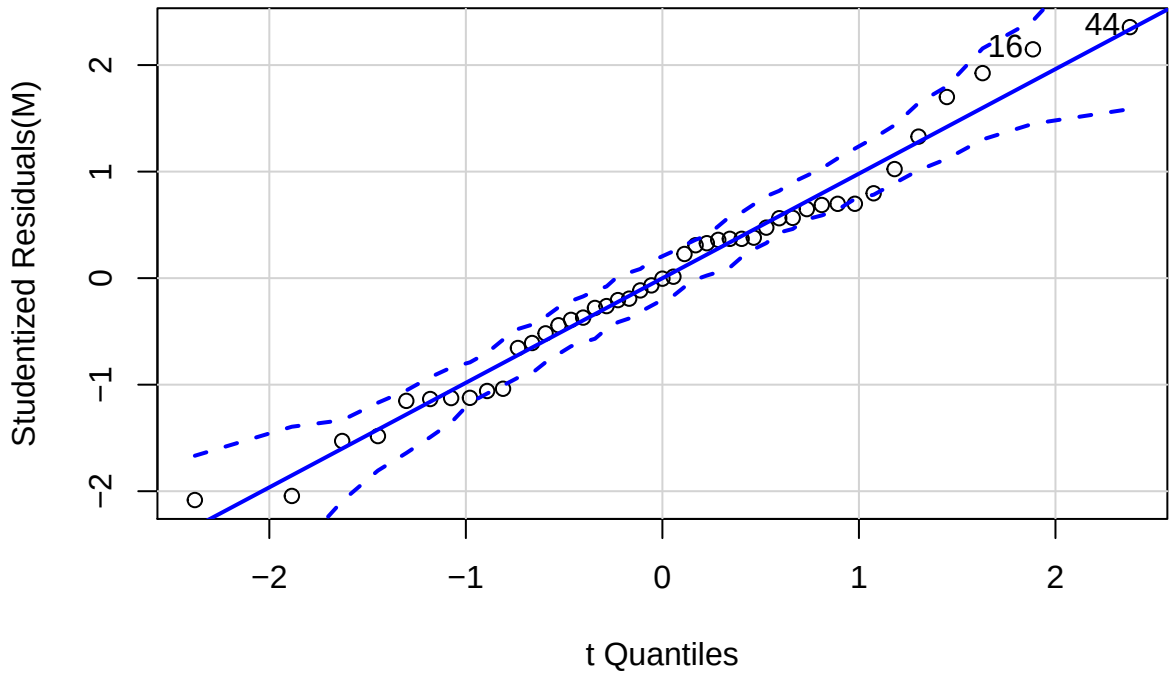
Next we will like to check the normality and homoscadasticity assumptions for our model.

CHECKING FOR ASSUMPTIONS

NORMALITY:

The normal qqplot of the model after removal of the outliers is presented below

```
qqPlot(M)
```



From the qqplot it can be observed that almost all the points are within the 95% confidence band, but still we will perform **Shapiro-Wilk test** for normality to have a better idea.

```
ols_test_normality(M)$shapiro
```

```
##
##  Shapiro-Wilk normality test
##
## data:  y
## W = 0.98026, p-value = 0.6301
```

It shows that we do not have any strong evidence for rejecting the null hypothesis that the errors are normally distributed.

HOMOSCEDASTICITY:

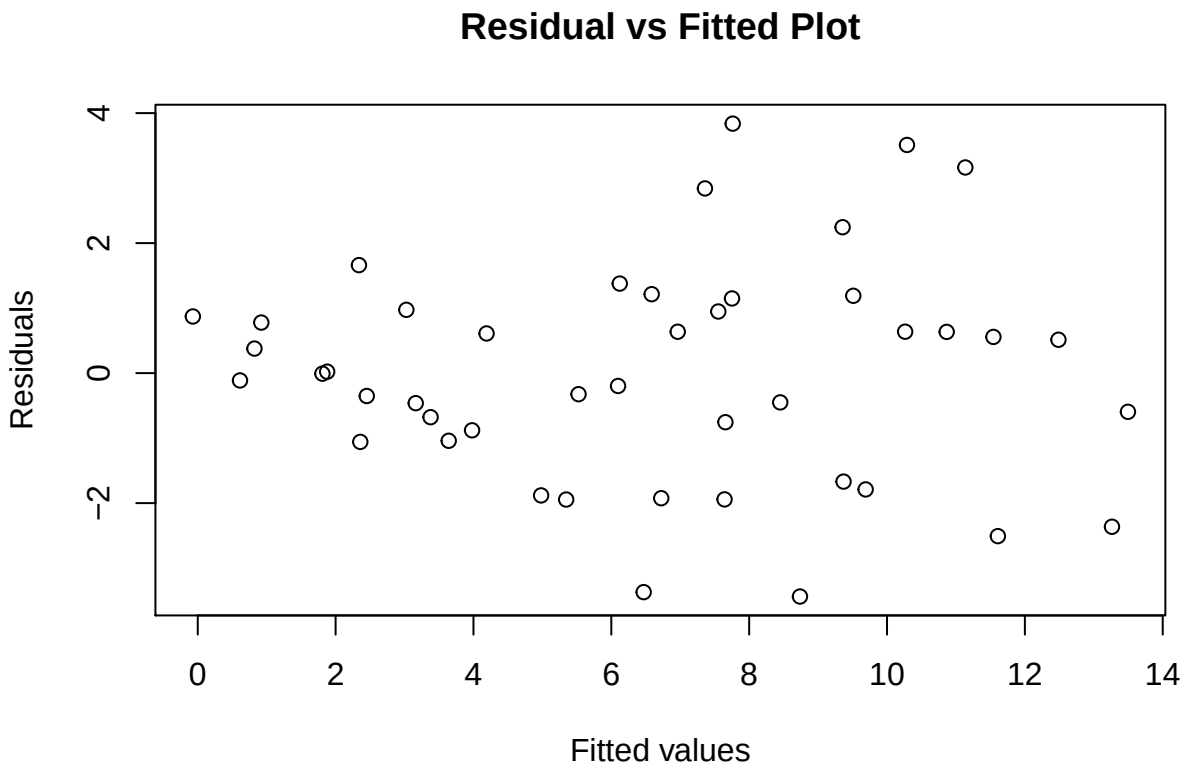
We will perform *Breusch-Pagan* test to check for homoscedasticity. The p-value for this test statistic is given below.

```
ols_test_breusch_pagan(M)$p
```

```
## [1] 0.02593968
```

The p-value is very small which indicates the presence of heteroscedasticity in our data. We present the **Residual vs Fitted** plot which will enable us to visualize heteroscedasticity more clearly.

```
plot(M$fitted.values, M$residuals, xlab = "Fitted values", ylab = "Residuals", main = "Residual vs Fitte
```



The points do not lie in a single horizontal band. That also indicates the presence of heteroscedasticity in the data. Now, we are going to make various transformations on the response variable, to make the data homoscedastic. We will try a series of transformations using the *trafo* package in *R*. We will select that transformation which gives the lowest value of $\hat{\sigma}^2$.

The different transformations used in this project are listed below

Table 2: Data-driven transformations.

Transformation	Source	Formula	Support	N	H	L
Box-Cox (shift)	Box and Cox (1964)	$\begin{cases} \frac{(y+s)^\lambda-1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y+s) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	✗
Log-shift opt	Feng, Hannig, and Marron (2016)	$\log(y+\lambda)$	$y \in \mathbb{R}$	✗	✗	✗
Bickel-Docksum	Bickel and Doksum (1981)	$\frac{\lvert y\rvert^\lambda \text{Sign}(y)-1}{\lambda}$ if $\lambda > 0$	$y \in \mathbb{R}$	✗	✗	
Yeo-Johnson	Yeo and Johnson (2000)	$\begin{cases} \frac{(y+1)^\lambda-1}{\lambda} & \text{if } \lambda \neq 0, y \geq 0; \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0; \\ \frac{(1-y)^{2-\lambda}-1}{\lambda-2} & \text{if } \lambda \neq 2, y < 0; \\ -\log(1-y) & \text{if } \lambda = 2, y < 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	
Square Root (shift)	Medina *et al.* (2018)	$\sqrt{y+\lambda}$	$y \in \mathbb{R}$	✗	✗	
Manly	Manly (1976)	$\begin{cases} \frac{e^{\lambda y}-1}{\lambda} & \text{if } \lambda \neq 0; \\ y & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	
Modulus	John and Draper (1980)	$\begin{cases} \text{Sign}(y)\frac{(\lvert y\rvert+1)^\lambda-1}{\lambda} & \text{if } \lambda \neq 0; \\ \text{Sign}(y)\log(\lvert y\rvert+1) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗		
Dual	Yang (2006)	$\begin{cases} \frac{(y^\lambda-y^{-\lambda})}{2\lambda} & \text{if } \lambda > 0; \\ \log(y) & \text{if } \lambda = 0. \end{cases}$	$y > 0$	✗		
Gpower	Kelmansky, Martínez, and Leiva (2013)	$\begin{cases} \frac{(y+\sqrt{y^2+1})^\lambda-1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y+\sqrt{y^2+1}) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗		

Table 1: Transformations without transformation parameter.

Transformation	Source	Formula	Support	N	H	L
Log (shift)	Box and Cox (1964)	$\log(y+s)$	$y \in \mathbb{R}$	✗	✗	✗
Glog	Durbin *et al.* (2002)	$\log(y+\sqrt{y^2+1})$	$y \in \mathbb{R}$	✗	✗	✗
Neglog	Whittaker *et al.* (2005)	$\text{Sign}(y)\log(\lvert y\rvert+1)$	$y \in \mathbb{R}$	✗	✗	
Reciprocal	Tukey (1977)	$\frac{1}{y}$	$y \neq 0$	✗	✗	

```
abp_models(M)
```

transf	$\hat{\sigma}^2$	adjusted R^2
boxcox	0.392533983400502	0.844690804896685
bickeldoksum	0.392542263660149	0.84469073599327
logshiftopt	0.0322775273743844	0.845775593107643
yeojohnson	0.193158702455696	0.845128277358805
sqrtshift	0.102309110555102	0.842665331030252
manly	0.862633918402704	0.841536277834518
modulus	0.193158702454934	0.845128277358819
dual	0.301305212779837	0.840926657790193
gpower	0.482662677877516	0.845529147090464
log	0.122766241167932	0.819631633760822
glog	0.0956508240768469	0.835858386257998
neglog	0.0626816745949788	0.842286764937687
reciprocal	0.0598542147478041	0.519946364857671

We can see from the table that *logshiftopt* transformation gives the most satisfactory result in case of both $\hat{\sigma}^2$

and adjusted R^2. Hence we will adopt this transformation.

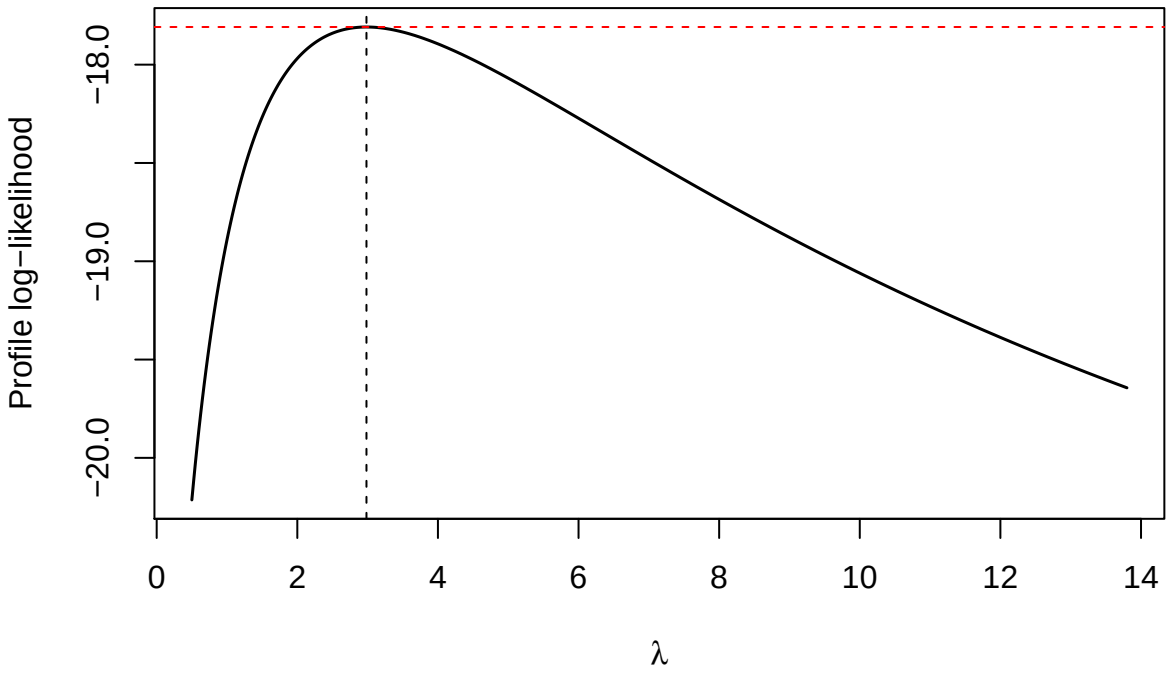
So, our new response variable will be changed from Y to $Y^\star$ such that

$$Y^\star = log(Y + \lambda) \tag{1}$$

The value of lambda which is estimated using maximum likelihood method has been extracted from R and shown below

```
y_tr <- logshiftopt(M)
```

```
## The default lambdarange for the Log shift opt transformation is calculated dependent on the data rang
```




```
y_star <- y_tr$zt
y_tr$lambdahat
```

```
## [1] 2.983805
```

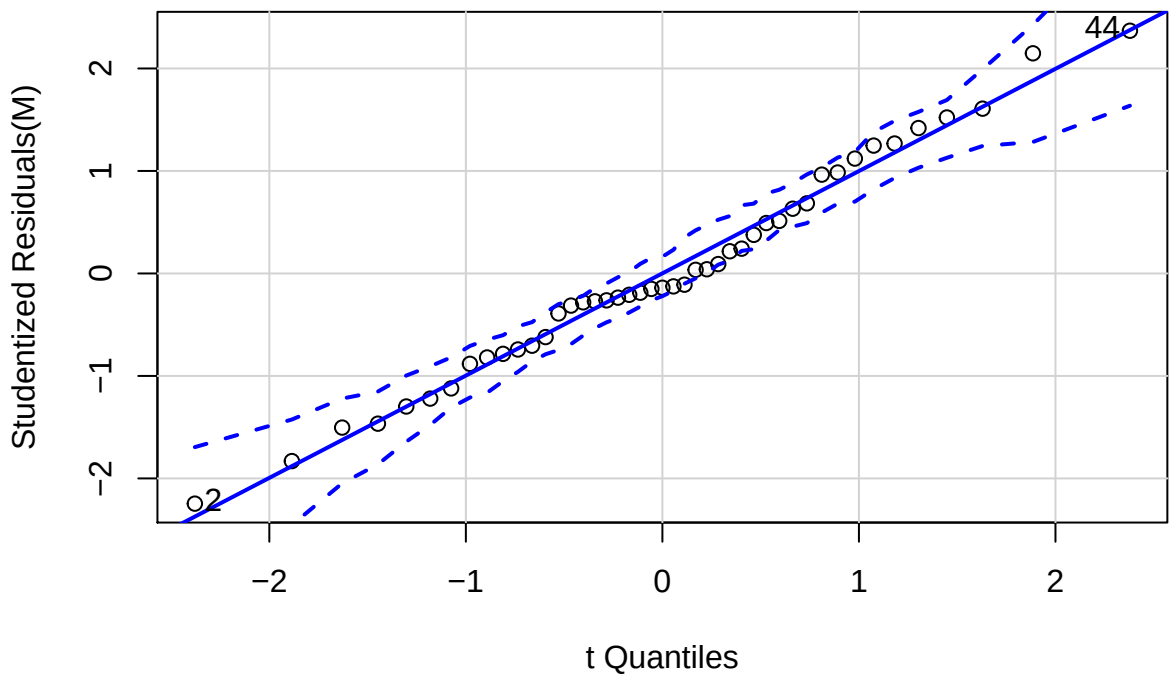
TRANSFORMED MODEL

Thus, after the transformation our model will be as follows:-

$$Y^\star = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

where $Y^* = \log(Y + 2.983805)$ Now we test for all the assumptions for this transformed model. First, we check for normality using the *qqplot* and *Breusch-Pagan* test.

```
M <- lm(y_star ~ X)
qqPlot(M, envelope = 0.96)
```




```
ols_test_normality(M)$shapiro
```

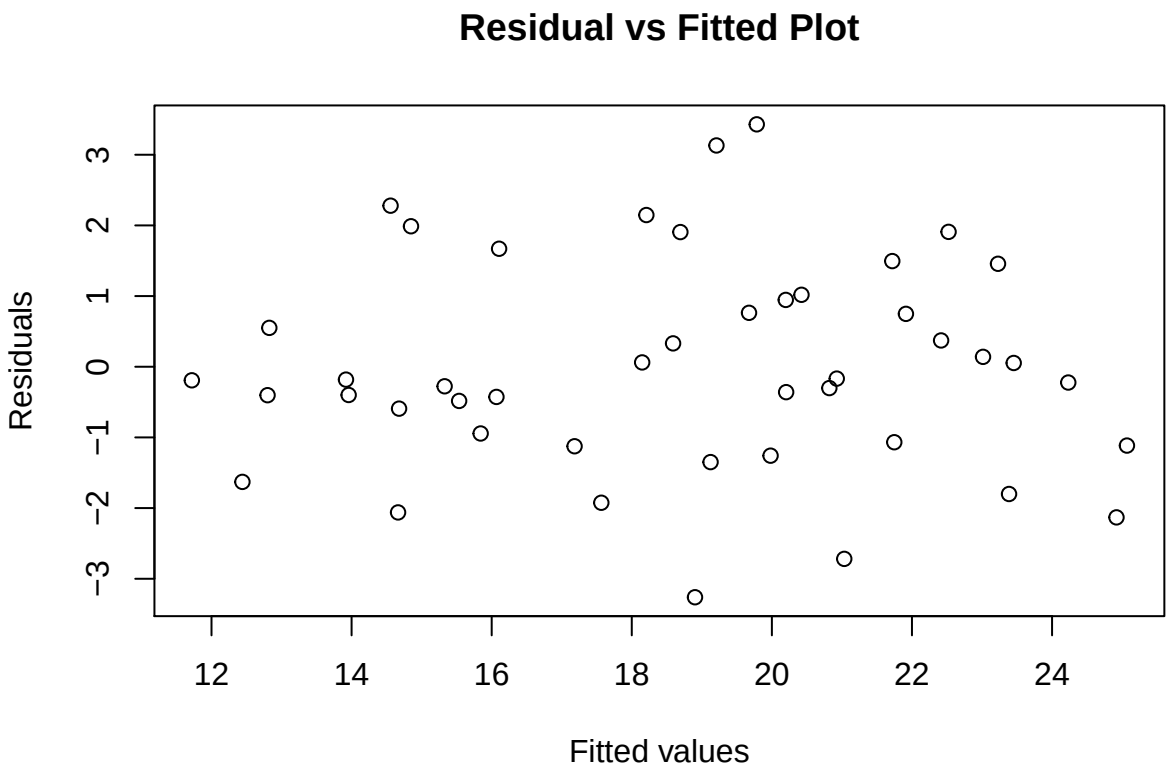
```
##
##  Shapiro-Wilk normality test
##
## data:  y
## W = 0.98577, p-value = 0.8475
```

Neither the plot nor the p-value of the test statistic give strong evidence to reject the null hypothesis of normality. Next we will check for homoscedasticity.

```
ols_test_breusch_pagan(M)$p
```

```
## [1] 0.6331837
```

```
plot(M$fitted.values, M$residuals, xlab = "Fitted values", ylab = "Residuals", main = "Residual vs Fitt
```

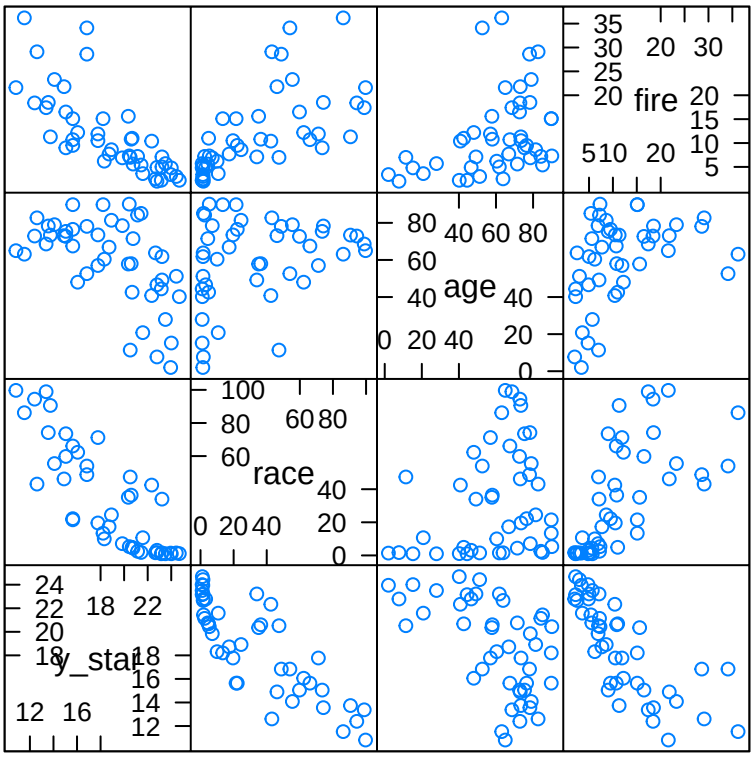


Now it can be seen that this transformation has reduced the heteroscedasticity in our data.

TESTING FOR COLLINEARITY IN OUR TRANSFORMED MODEL:

Firstly we present the scatter plot matrix of our transformed variable against all the explanatory variables.

```
splom(cbind(y_star, X))
```



Scatter Plot Matrix

The scatter plot depicts a possibility of having linear relationship between the two explanatory variables **race** and **fire**. We would like to formally check for collinearity using *variance inflation factor*s and *conditional number*(κ).

```
omcdiag(X, y_star)$odiags[6][1]
```

```
## [1] 7.474837
```

```
M <- lm(y_star ~ race + age + fire, data = data[-c(24, 39), ])
vif(M)
```

```
##     race      age     fire
## 1.764425 1.179141 1.939215
```

Both κ and *vif* values are sufficiently lower than their respective cut off points(30 and 5 respectively). So there is no indication of severe multi collinearity in our data.

SUMMARY OF THE TRANSFORMED MODEL

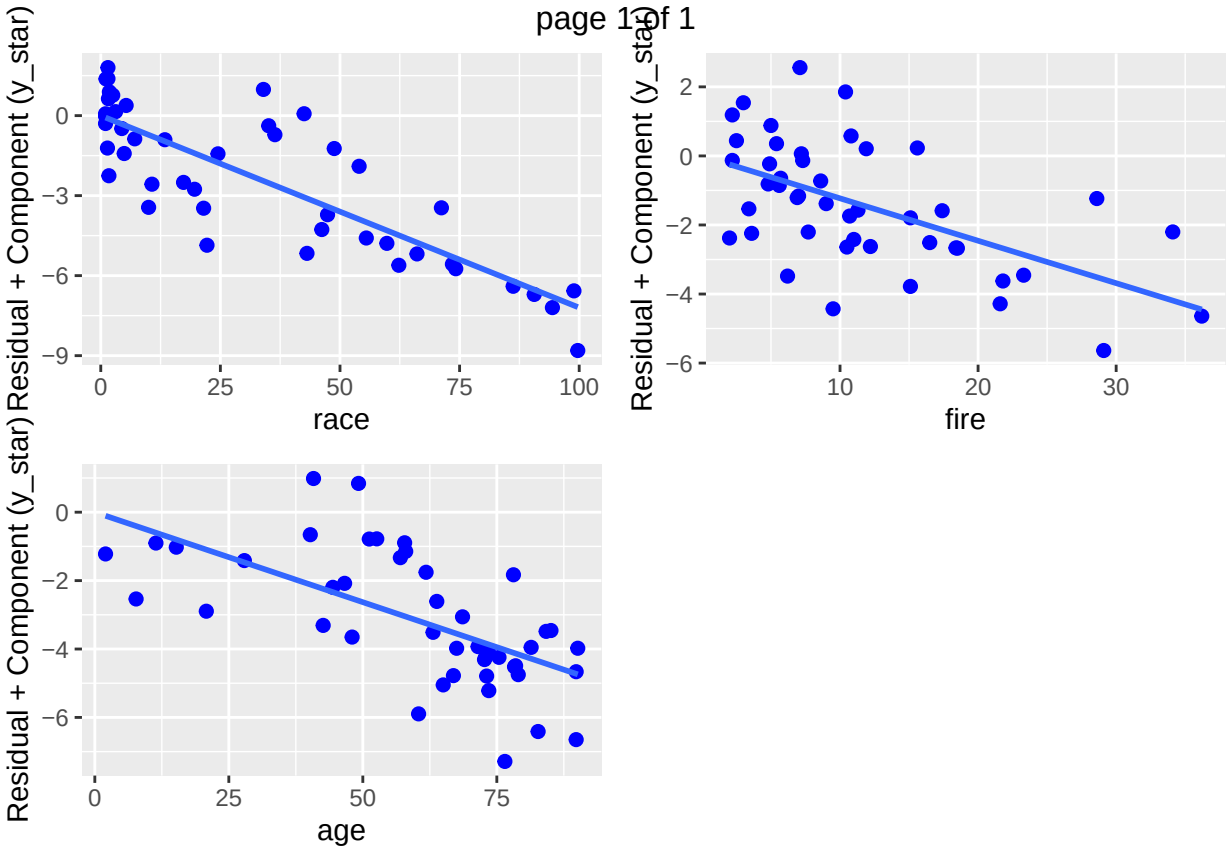
```
summary(M)
```

```
##
## Call:
```

```
## lm(formula = y_star ~ race + age + fire, data = data[-c(24, 39),
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2616 -1.0681 -0.1921  0.9459  3.4301
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.693009   0.665591  38.602  < 2e-16 ***
## race        -0.071998   0.009745  -7.388 4.69e-09 ***
## age         -0.052614   0.011167  -4.712 2.82e-05 ***
## fire        -0.122859   0.038246  -3.212  0.00256 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.556 on 41 degrees of freedom
## Multiple R-squared:  0.8563, Adjusted R-squared:  0.8458
## F-statistic: 81.43 on 3 and 41 DF,  p-value: < 2.2e-16
```

We have seen that all the coefficients are significant and the p-value of the F statistic is very low. The adjusted R^2 value is pretty satisfactory. Now we will make a partial residual plot in order to check for possible curvature in the explanatory variables.

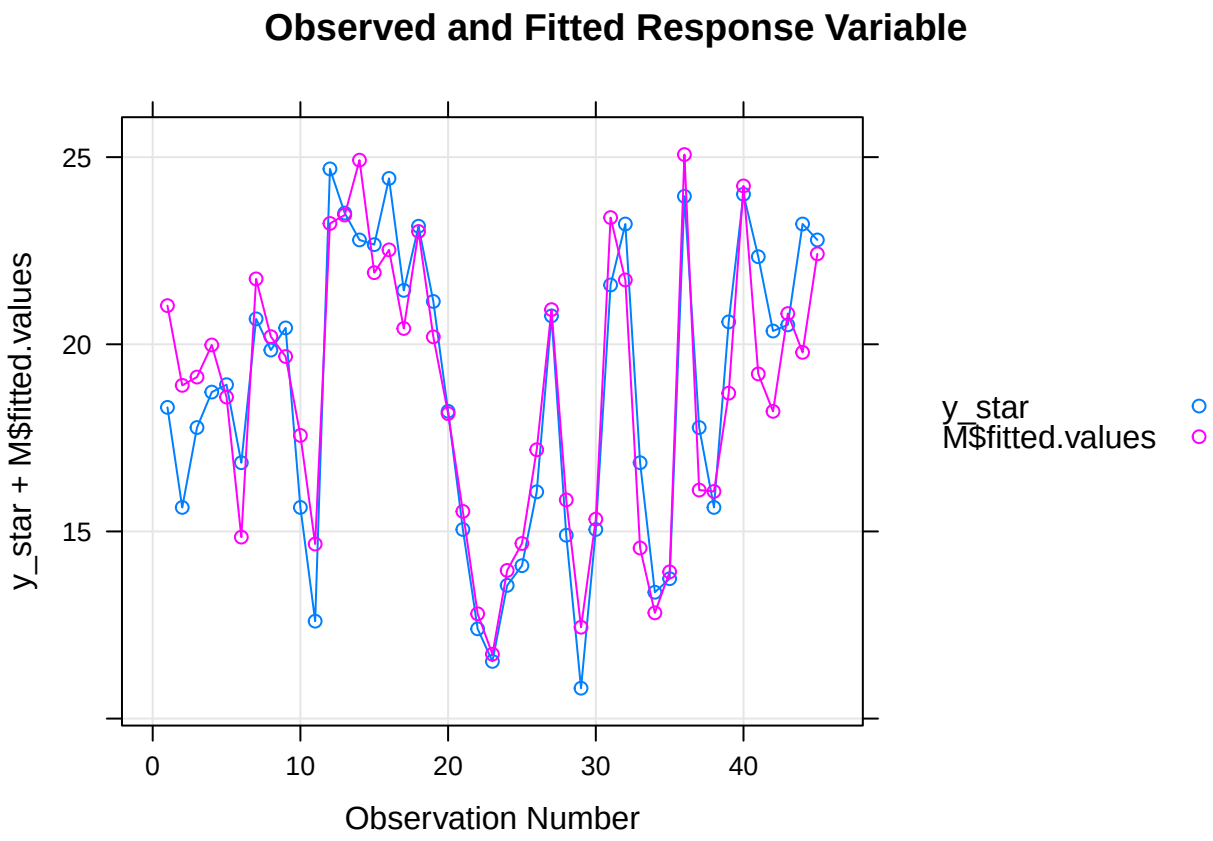
```
ols_plot_comp_plus_resid(M)
```



The partial residual plots do not show any specific curvature.

Now we are plotting the observed and fitted values of the response variable in the same graph, which will show how good our fit is.

```
xyplot(y_star + M$fitted.values ~1:45, auto.key = list(space = "right"),
       grid = TRUE, main = "Observed and Fitted Response Variable",
       xlab = "Observation Number", type = "b" )
```



This shows our fit is very good. Once we have got our model we head forward to the testing of hypothesis.

CONCLUSION:

We can clearly see from the summary that coefficient of race is significant(p-value: 4.69e-09) for predicting volact. We note that the coefficient is negative, hence race has a negative impact over volact. So the claim of the organisation that insurance companies are redlining neighbourhood of this particular race is correct. We also note that the areas having old houses and the areas more prone to fire incidents are getting redlined by different insurance companies in Chicago.