Indian Institute of Technology Bombay

DATA ANALYSIS GROUP PROJECT

REPORT

Dataset – "INSURANCE"
Instructor – MONIKA BHATTACHARJEE

Group Members: Subhojit Kayal
Priya Kandare
Chiranjib Mazumder

Data

The data set contains information on 1338 families with 7 attributes, namely, Insurance charges, Age, Sex, BMI, Number of Children, Smoking, and Region.

Notations: Taking Insurance Charge as the response variable, we define the variables to be used as

$y = Insurance\ charges$
$x_1 = Age$
$x_2 = Sex$
$x_3 = BMI$
$x_4 = Number\ of Children$
$x_5 = 1\ \ ,if\ Person\ is\ Smoker$
$\quad\ 0\ \ ,if Person\ is\ Non-Smoker$
$x_6 = 1\ \ ,if\ the\ Person\ is\ from\ NorthWest, SouthWest, SouthEast\ Region$
$\quad\ 0\ \ ,if\ the\ Person\ is\ from\ SouthWest\ Region$

The observed cases, n = 1338

We observe the tuples

$$\{(\boldsymbol{y_i}, \boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{i6}}): \boldsymbol{i = 1, 2, \dots, 1338}\}$$

on the aforementioned variables.

Observation vector for Response variable: $Y = (y_1, y_2, \dots, y_{1338})'$
Observation vector for j-th predictor: $X_j = \left(x_{1j}, x_{2j}, \dots, x_{1338j}\right)'$

Design matrix: Consider the 1338 × 7 matrix X whose first column is 1_n and other columns are $\{X_j: j = 1, 2, \dots, 6\}$,i.e.

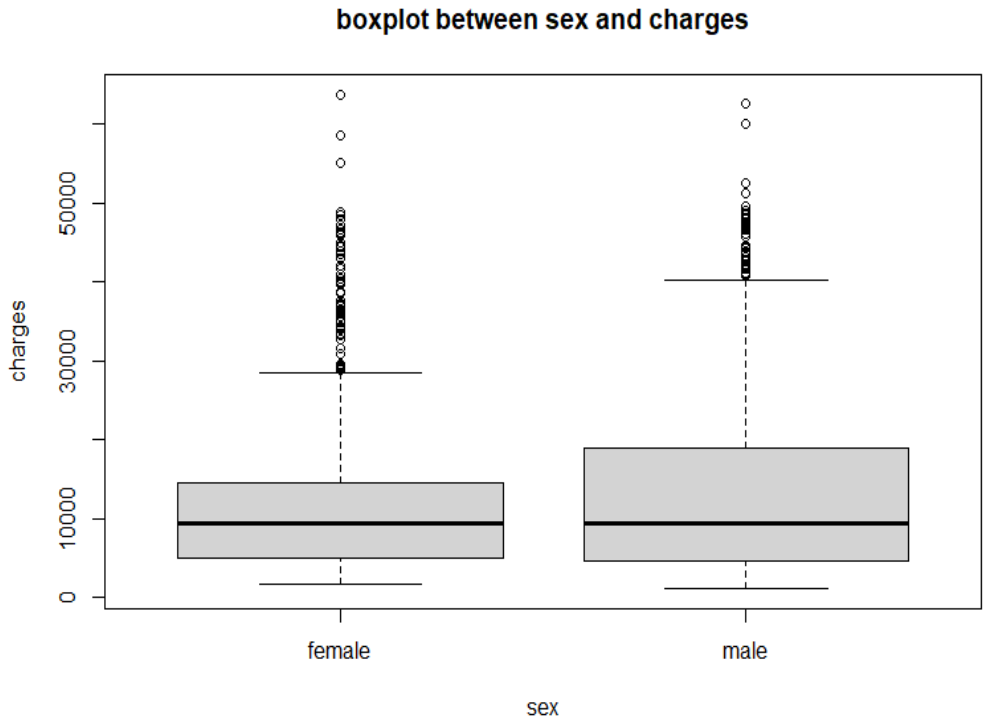
$$X = \{1_n, X_1, X_2 \dots, X_6\}$$

Plots

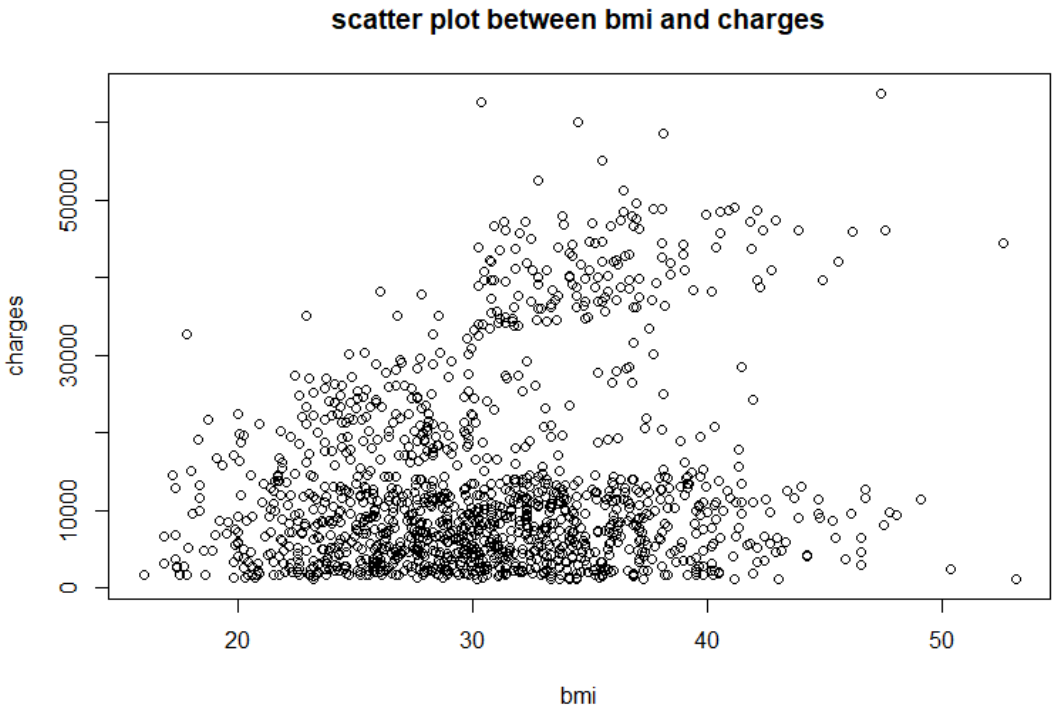
Here are the pairwise scatter diagrams and boxplots between each predictor variable and the Response variable to determine the relationship between them

- **Age and Insurance Charges**

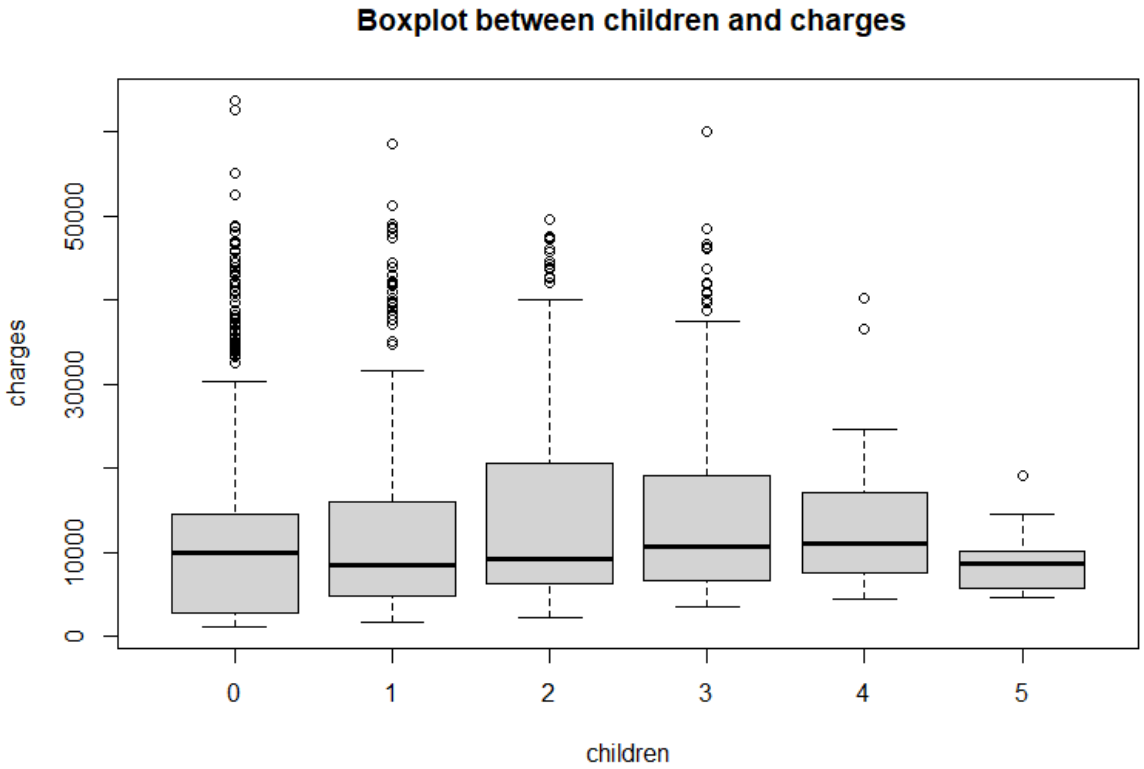
- **Sex and Insurance Charges**



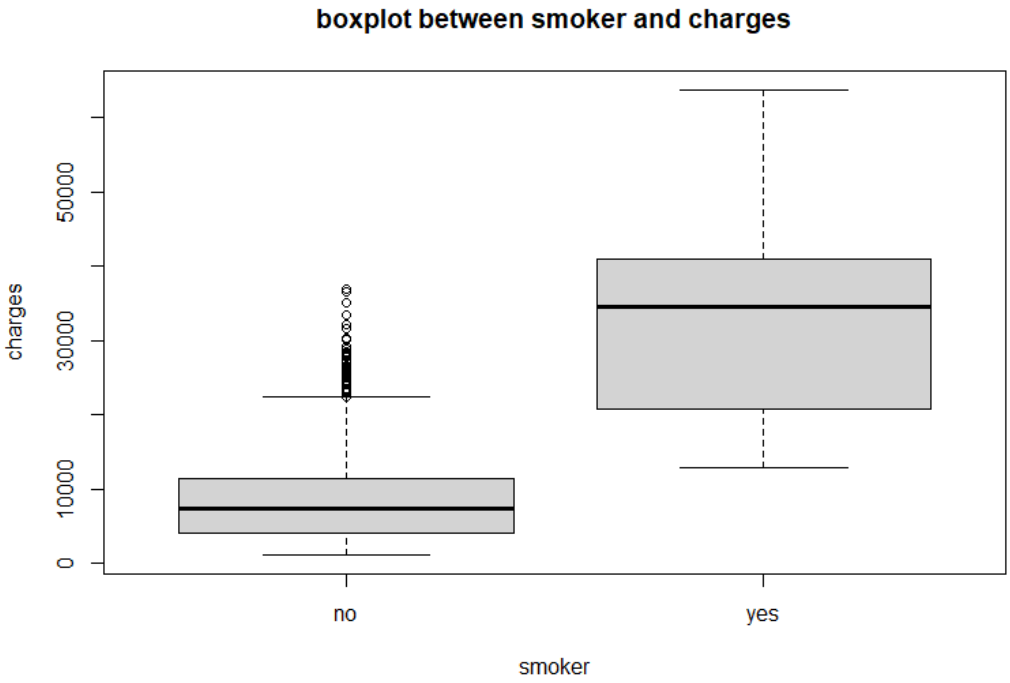
- **BMI and Insurance Charges**



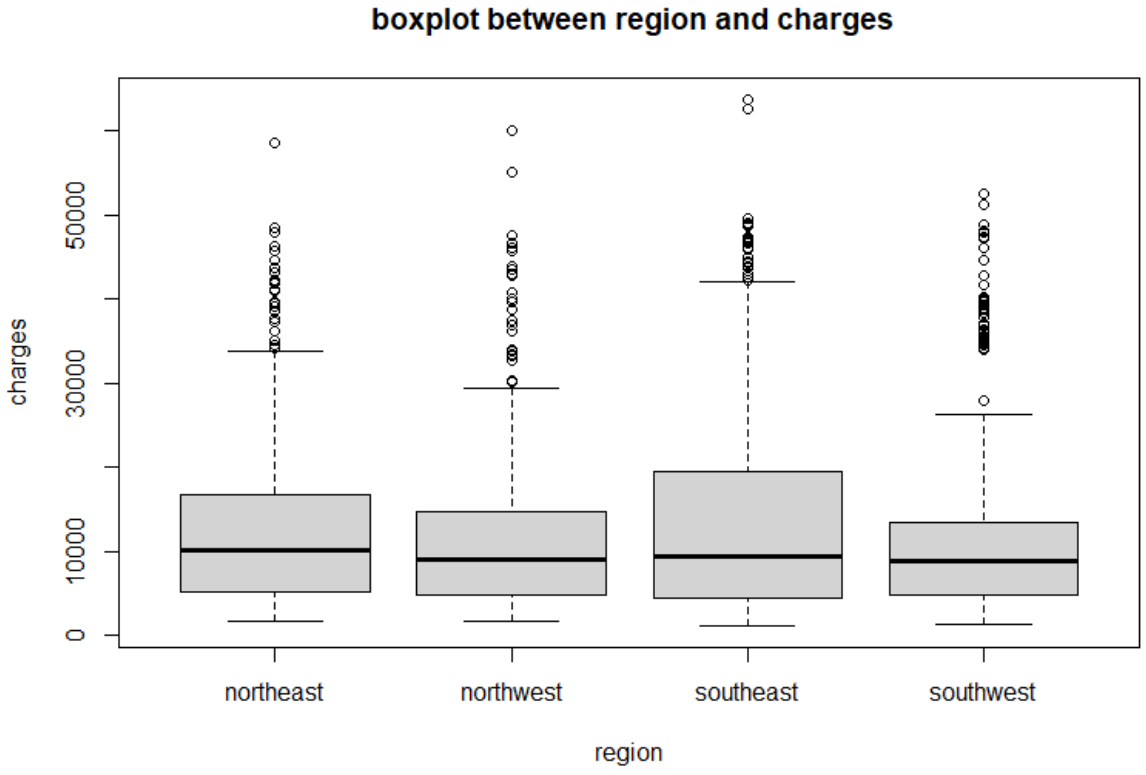
- **Number of Children and Insurance Charges**



- **Smoking and Insurance Charges**



- **Region and Insurance Charges**



Observations

- From the scatter plot we can see that there is strong positive relation between Age and Charges. Hence, Age has **Significant effect** on Insurance Charges.

- Boxplots of Insurance charges for Male and Female are significantly different. Hence, Sex has **Significant effect** on Insurance Charges.

- A **strong positive relation** between BMI and Charges is evident from the Scatterplot.

- Boxplots of Insurance Charges for different number of children are Significantly different. Hence, Number of children has **Important effect** on Insurance Charges.

- Boxplots of Insurance Charges for smoker and non-smoker are highly different. Hence, Smoker has **Vital effect** on Insurance Charges.

- Boxplots of Insurance Charges for different regions are different. Hence, Different regions has **Significant effect** on Insurance Charges.

Model

Multiple Linear Regression model: We propose to fit the MLRM to the given dataset with all the appropriate assumptions.

$$\boldsymbol{Y = X\beta + E} \qquad \textbf{(0.1)}$$

where $E = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$is the random error vector, and $\beta = (\beta_0, \beta_1, \dots, \beta_8)'$ is the unknown vector of regression coefficients.

Assumptions:

- $\boldsymbol{E \sim N(0_n, \sigma^2 I_n)}$
- We assume that the predictor variables are non-stochastic.

For the given dataset, $|X'X| = 3.86 \times 10^{24}$. Therefore, the design matrix is of full column rank.

Fitting: The least-squares estimator of β is given as

$$\hat{\beta} = (X'X)^{-1}X'Y$$

For the given data, it turns out to be

$$\hat{\beta} = (-11938.5 \;\; 256.9 \;\; -131.3 \;\; 339.2 \;\; 475.5 \;\; 23848.5 \;\; -353.0 \;\; -1035.0 \;\; -960.1)'$$

and therefore, the fitted multiple linear regression model is

$$\hat{Y} = X(-11938.5 \quad 256.9 \quad -131.3 \quad 339.2 \quad 475.5 \quad 23848.5 \quad -353.0 \quad -1035.0 \quad -960.1)'$$

Fitted values and Residuals: Tabulated below

Individuals	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Fitted values	**25293.7**	**3448.6**	**6706.9**	**3754.8**	**5592.5**	**3719.8**	**10660**	**8047.9**	**8503**
Residuals	**-8408.8**	**-1723.1**	**-2257.5**	**18229.6**	**-1725.6**	**36.8**	**-2419.4**	**-766.4**	**-2096.6**

……..

Individuals	**992**	**993**	**994**	**995**	**996**	**997**	**998**	**999**	**1000**
Fitted values	**8214.5**	**11613.7**	**6720.2**	**26717.8**	**7400.1**	**10111.8**	**15701.7**	**10273.6**	**6074.5**
Residuals	**-1069.6**	**-1495.3**	**-1235.7**	**-10297.3**	**586.4**	**-2693.3**	**-1819.7**	**-3721.8**	**-806.7**

……

Significance

We define out hypotheses as:
$\boldsymbol{H_0: \beta_1 = \beta_2 = \cdots = \beta_8 = 0}$
$\boldsymbol{H_1}$: at least one of $\boldsymbol{\beta_1, \beta_2, \ldots, \beta_8}$ is non-zero

We use the test statistic

$$\boldsymbol{F_0 = \frac{MSR}{MSE}}$$

where MSR and MSE are respectively the mean of squares due to the model, and mean of residual squares.

We reject H_0 at $100\alpha\%$ level if $\boldsymbol{P\left(F_{8,n-9} > (F_0)_{observed}\right) < \alpha}$

For our data $\boldsymbol{P\left(F_{8,n-9} > (F_0)_{observed}\right) = 2.2 \times 10^{-16} < 0.05}$. Therefore, we reject H_0 at 5% level. So, we can conclude that, **the proposed multiple linear regression model is significant**.

R^2 and Adjusted R^2: For the model, the value of **R^2** and **adjusted R^2** turns out to be **0.7509** and **0.7494**. This can be interpreted as, **75.09%** of the total variation in the response variable is explained by the above least-squares fitted multiple linear regression model.

Significance of each predictor variable:

Here, we wish to test $H_{0j} : \beta_j = 0$ against $H_{1j} : \beta_j \neq 0$. We use the test statistic

$$T_j = \hat{\beta} / \sqrt{MSE\ e'_{j+1,p+1}(X'X)^{-1}\ e_{j+1,p+1}}.$$

We know that

$$T_j \overset{H_{0j}}{\sim} t_{n-9} \ and \ \overset{H_{1j}}{\sim} t'_{n-9,\partial}$$

$where\ \partial = \beta_j / \sqrt{\sigma^2\ e'_{j+1,p+1}\ (X'X)^{-1} e_{j+1,p+1}}$. Thus, we reject $H_{0j}\ at\ 100\alpha\%$ level of significance if the p-value $p_j = \mathbb{P}\left(|t_{n-9}| > \left|(T_j)_{observed}\right|\right) < \alpha$.

For the given data

$$p_{age} = 2 \times 10^{-16}, \quad p_{male} = 0.7, \quad p_{bmi} = 2 \times 10^{-16},$$
$$p_{children} = 0.0004,$$
$$, \quad p_{smoker} = 2 \times 10^{-16},$$
$$p_{northwest} = 0.4588, \quad p_{southeast} = 0.0308, \ p_{southwest} = 0.0448.$$

This implies that, for the given data, p_{age}, p_{bmi}, $p_{children_2}$, $p_{children_4}$, p_{smoker}, $p_{southeast}$, $p_{southwest}$ < 0.05. Thus, we reject $H_{0j}\ at\ 100\alpha\%$ level of significance for j=1,3,4,5,6 i.e. the predictor variables Age, BMI, Children, Smoker, Region are significant. p_{male} > 0.05, so we fail to reject $H_{0j}\ at\ 100\alpha\%$ level of significance for j=2.

Therefore, all the predictor variables are not significant in the multiple linear regression model. Since, we fail to reject $H_{0j}\ at\ 100\alpha\%$ level of significance ∀j

Variable Selection Process: Since, all the predictor variables are not significant in the multiple linear regression model. We want to fit a multiple linear regression model with an appropriate subset of these predictors which can significantly explain the response Y. For this purpose, we need the following notation, definitions and discussion on the various criteria for evaluating regression models.

Some Notations and definitions:

- Let X be a random variable and $f(\cdot)$ be a measurable function. Then $f(X)_{observed}$ is the observed value of $f(X)$. Often, we abuse notation by writing $f(X)$ instead $f(X)_{observed}$. It is suggested to understand the difference between the random variable $f(X)$ and the observed value of $f(X)$ from the context. For example, when $f(X)$ is compared with some real numbers or expected to be contained in some set, then we mean the observed value of $f(X)$. On the other hand, if probability and expectation are involved with $f(X)$, then we mean the random variable $f(X)$.
- The total sum of squares is $TSS = \sum_{i=1}^{n}((y_i - \overline{y})^2)$.
- A regression model with a subset of p-many predictor variables is called a subset regression model.
- Let $\mathcal{M}_k$ be the set of all subsets regression model corresponding to subsets of k−1 predictor variables.
- The subset regression model corresponding to a subset of predictor variables $\mathbb{A}$ is $\mathbb{M}_{\mathbb{A}}$.
- Let $\mathcal{F} = \{1, x_1, x_2, \dots\dots, x_p\}$ be the set of all predictors. $\mathbb{M}_{\mathcal{F}}$ is the full regression model. Also, note that $\mathcal{M}_{p+1} = \{\mathbb{M}_{\mathcal{F}}\}$.
- For an $\mathbb{M} \in \mathcal{M}_k$, let $\hat{y}_i(\mathbb{M})$ be the fitted values corresponding to the i-th observation of y.
- The residual sum of squares $SSE_k(\mathbb{M})$ and the sum of squares due to regressor $SS_{R,k}(\mathbb{M})$ are defined respectively by

$$SSE_k(\mathbb{M}) = \sum_{i=1}^{n}\left(\left(y_i - \hat{y}_i(\mathbb{M})\right)^2\right) \;\&\; SS_{R,k}(\mathbb{M}) = \sum_{i=1}^{n}\left(\left(\hat{y}_i(\mathbb{M}) - \overline{y}\right)^2\right)$$

- The residual mean squares/mean squares error is given by,

$$MSE_k(\mathbb{M}) = \frac{SSE_k(\mathbb{M})}{n-k}$$

- MSE_{p+1} denotes the mean squares error corresponding to the full regression model.
- Two subset regression models $\mathbb{M}_1$ & $\mathbb{M}_2$ satsify $\mathbb{M}_1 \subset \mathbb{M}_2$ if $\mathbb{M}_2$ contains all the predictor variables of $\mathbb{M}_1$.
- For the subset of regression models $\mathbb{M}_1$ & $\mathbb{M}_2$ satisfying $\mathbb{M}_1 \in \mathcal{M}_k, \mathbb{M}_2 \in \mathcal{M}_{k+r}$ & $\mathbb{M}_1 \subset \mathbb{M}_2$, Define

$$\tilde{F}(\mathbb{M}_1, \mathbb{M}_2) = \frac{\left(SS_{R,k+r}(\mathbb{M}_2) - SS_{R,k}(\mathbb{M}_1)\right)/r}{MSE_{k+r}(\mathbb{M}_2)}$$

- Let $\left(F_{a,b}\right)_{\alpha}$ be the $100(1-\alpha)$-$th\ percentile$ of the F-$distribution$ with degrees of freedom a & b.

- Define $\mathcal{C}_\alpha(\mathbb{M}_1, \mathbb{M}_2) = \left(F_{r,n-k-r-1}\right)_\alpha \; \forall \mathbb{M}_1 \in \mathcal{M}_k, \mathbb{M}_2 \in \mathcal{M}_{k+r}, \alpha \in (0,1)$. In this problem set, $\mathcal{C}_\alpha(\mathbb{M}_1, \mathbb{M}_2)$ is referred as cut-off for $\tilde{F}(\mathbb{M}_1, \mathbb{M}_2)$. In this problem set, n=1338.
- Let $\mathbb{A}$ be a subset of predictor variables which does not contain x_j. Partial correlation coefficient between y & x_j $given$ $\mathbb{A}$ is denoted by $r_{yx_j,\mathbb{A}}$. The correlation coefficient between y & x_j is nothing but $r_{yx_j,\{1\}}$, which we also denote r_{yx_j}.

Criteria for evaluating subset regression models: Here we need the following criteria for evaluating and comparing subset regression models.

- *Coefficient of determination*: The coefficient of determination corresponding to the model $\mathbb{M} \in \mathcal{M}_k$ is defined as
$$R_k^2(\mathbb{M}) = \frac{SS_{R,k}(\mathbb{M})}{TSS} = 1 - \frac{SSE_k(\mathbb{M})}{TSS}$$

- *Adjusted R^2:* The adjusted R^2 corresponding to the model $\mathbb{M} \in \mathcal{M}_k$ is $R^2_{adjusted,k}(\mathbb{M}) = 1 - \frac{MSE_k(\mathbb{M})}{TSS/(n-1)}$. One criterion for selection of an optimum subset model is to choose the model that has a maximum adjusted R^2 value.

- *Mallow's statistic:* Mallow's statistic corresponding to the model $\mathbb{M} \in \mathcal{M}_k$ is $\mathrm{C_k}(\mathbb{M}) = \frac{\mathrm{SSE_k}(\mathbb{M})}{\mathrm{MSE_{p+1}}} - \mathrm{n} + 2\mathrm{k}$. A Model with smaller value of Mallow's statistic is preferred.

- *Akaike's information criterion (AIC):* In case of multiple linear regression model, AIC turns out to be $AIC_k(\mathbb{M}) = n\log\left(\frac{SSE_k(\mathbb{M})}{n}\right) + 2k$. A Model with smaller value of AIC is preferable.

- *Bayesian information criterion (BIC)*: In case of multiple linear regression model, BIC turns out to be $BIC_k(\mathbb{M}) = n\log\left(\frac{SSE_k(\mathbb{M})}{n}\right) + 2k$. A Model with smaller value of BIC is preferable.

i) *Evaluating all possible subset regression models*: This procedure requires that the analyst fit all the regression equations involving one

candidate predictor variable, two candidate predictor variables, three candidate predictor variables, and so on. These equations are evaluated according to some suitable criterion (take one of the above criteria) and the appropriate regression model is selected. The output for the criteria adjusted R^2, Mallow's statistic, AIC and BIC are reported in Rows 2-5 of Table 1.

ii) *Forward selection procedure:* This methodology assumes that there is no predictor variable in the model except an intercept term. It adds variables one by one and test the fitted model at each step using some suitable criterion. Here we use the following algorithm.

- Consider only the intercept term.
- Choose any predictor variable from $argmax_{j\in\{1,2,\ldots,p\}} r_{yx_j}$. Suppose it is x_1.
- If $\tilde{F}\left(\mathbb{M}_{\{1\}}, \mathbb{M}_{\{1,x_1\}}\right) > \mathcal{C}_{0.05}\left(\mathbb{M}_{\{1\}}, \mathbb{M}_{\{1,x_1\}}\right)$, then x_1 enters into the model.
- Next consider any predictor variable from $argmax_{j\in\{2,3,\ldots,p\}} r_{yx_{j.\{1,x_1\}}}$. Suppose it is x_2.
- If $\tilde{F}\left(\mathbb{M}_{\{1,x_1\}}, \mathbb{M}_{\{1,x_1,x_2\}}\right) > \mathcal{C}_{0.05}\left(\mathbb{M}_{\{1,x_1\}}, \mathbb{M}_{\{1,x_1,x_2\}}\right)$, then x_2 enters into the model.
- Next consider any predictor variable from $argmax_{j\in\{3,4,\ldots,p\}} r_{yx_{j.\{1,x_1,x_2\}}}$. Suppose it is x_3.
- If $\tilde{F}\left(\mathbb{M}_{\{1,x_1,x_2\}}, \mathbb{M}_{\{1,x_1,x_2,x_3\}}\right) > \mathcal{C}_{0.05}\left(\mathbb{M}_{\{1,x_1,x_2\}}, \mathbb{M}_{\{1,x_1,x_2,x_3\}}\right)$, then x_3 enters into the model.
- Continue with such selection as long as either at particular step, the observed value of $\tilde{F}$-statistic does not exceed its cut-off value or when the last predictor variable is added to the model.

Output of this procedure is reported in Row 6 of Table 1.

iii) *Backward elimination procedure:* The backward elimination methodology begins with all predictor variables and keep on deleting one variable at a time until a suitable model is obtained. Here we use the following algorithm.

- Consider all p predictor variables and fit the model.
- Consider any predictor variable from $argmin_{j\in\{1,2,\ldots,p\}} \tilde{F}\left(\mathbb{M}_{\mathcal{F}-\{x_j\}}, \mathbb{M}_{\mathcal{F}}\right)$. Suppose it is x_1.

- If $\tilde{F}\left(\mathbb{M}_{\mathcal{F}-\{x_1\}}, \mathbb{M}_{\mathcal{F}}\right) < \mathcal{C}_{0.05}\left(\mathbb{M}_{\mathcal{F}-\{x_1\}}, \mathbb{M}_{\mathcal{F}}\right)$, then remove x_1 from the model.
- Next consider any predictor variable from $argmin_{j\in\{2,3,\ldots,p\}}\tilde{F}\left(\mathbb{M}_{\mathcal{F}-\{x_1,x_j\}}, \mathbb{M}_{\mathcal{F}-\{x_1\}}\right)$. Suppose it is x_2.
- If $\tilde{F}\left(\mathbb{M}_{\mathcal{F}-\{x_1,x_2\}}, \mathbb{M}_{\mathcal{F}-\{x_1\}}\right) < \mathcal{C}_{0.05}\left(\mathbb{M}_{\mathcal{F}-\{x_1,x_2\}}, \mathbb{M}_{\mathcal{F}-\{x_1\}}\right)$, then remove x_2 from the model.
- Repeat this procedure.
- Stop the procedure when smallest $\tilde{F}$-statistic exceeds its cut-off value.

Output of this procedure is reported in Row 7 of Table 1.

iv) *Stepwise selection method:* Stepwise regression is a modification of forward selection procedure in which at each step all predictor variables entered into the model previously are regressed via their $\tilde{F}$-statistic. A predictor variable added at an earlier step may now be redundant be-cause of the relationships between it and the predictor variables now in the equation. If the $\tilde{F}$-statistic for a variable is less than its cut-off value, that variable is dropped from the model. Output of this procedure is reported in Row 8 of Table 1.

v) Define $g(\mathbb{M}) = \frac{R_k^2(\mathbb{M})}{R_6^2(\mathbb{M}_{\mathcal{F}})}$ $\forall \mathbb{M} \in \mathcal{M}_k$. Note that $0 \leq g \leq 1$. If $g(\mathbb{M})$ is close to 1, then the subset regression model $\mathbb{M}$ performs equally good as the full regression model $\mathbb{M}_{\mathcal{F}}$ and the predictor variables, which are absent in $\mathbb{M}$, are not significant. On the other hand, if $g(\mathbb{M})$ is far away from 1, then the model $\mathbb{M}$ does not contain all the significant predictor variables and we should search for better subset regression model. Based on the reported g-value in Table 1, we can reasonably conclude that all the variable selection methods perform equally good as the full regression model.

Table 1: Output of different variable section methods.

Methods	Criterion	Selected Variables	g-value
All possible subset models	**Adjusted R^2**	x_1, x_3, x_4, x_5, x_6	1
All possible subset models	**Mallow's Statistic**	x_1, x_3, x_4, x_5, x_6	1
All possible subset models	**AIC**	x_1, x_3, x_4, x_5, x_6	1
All possible subset models	**BIC**	x_1, x_3, x_4, x_5, x_6	1
Forward selection	$\tilde{F}$**-Statistic**	x_1, x_3, x_4, x_5, x_6	1
Backward Elimination	$\tilde{F}$**-Statistic**	x_2	0.004
Stepwise Selection	**Partial** $\tilde{F}$**-Statistic**	x_1, x_3, x_4, x_5	1

Conclusion of Variable selection procedure: Since the g-value for Forward selection method is close to 1, so we consider the subset of predictor variables x_1, x_3, x_4, x_5, x_6 for the updated Multiple linear regression model, which is obtained in Forward selection method.

Checking Multi-collinearity for the updated model: Though determinant of $X'X$ is much larger than 0, it may be mis-leading to conclude about the multi-collinearity at this stage of analysis. The summary table of the regression coefficients corresponding to this multiple linear regression fit is as follows. (*)

Regression coefficients	Estimated value	Standard Error	t-value	p-value
β_0	**-11938.5**	**987.8**	**-12.086**	**$<2\times10^{-16}$**
β_1	**256.9**	**11.9**	**21.587**	**$<2\times10^{-16}$**
β_2	**-131.3**	**332.9**	**-0.394**	**0.69334**
β_3	**339.2**	**28.6**	**11.860**	**$<2\times10^{-16}$**
β_4	**475.5**	**137.8**	**3.451**	**0.000577**
β_5	**23848.5**	**413.1**	**57.723**	**$<2\times10^{-16}$**
β_6	**-353.0**	**476.3**	**-0.741**	**0.458769**
β_7	**-1035.0**	**478.7**	**-2.162**	**0.030782**
β_8	**-960.0**	**477.9**	**-2.009**	**0.044765**

Here we note that except the variable Sex, all other predictor variables are significant under .05 level of significance. So we cannot say strongly that there is a presence of multicollinearity.

Notation for centred and scaled variables: To deal with multi-collinearity, it is suggested to work with the centred and scaled variables. Define

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n}(y_i) \ , \overline{x}_j = \frac{1}{n}\sum_{i=1}^{n}(x_{ij}) \ , S_{yy} = \sum_{i=1}^{n}((y_i - \overline{y})^2) \, , S_{x_j x_j} = \sum_{i=1}^{n}\left((x_{ij} - \overline{x}_j)^2\right),$$

$$y_{i,cs} := \frac{y_i - \overline{y}}{\sqrt{S_{yy}}} \, , x_{ij,cs} := \frac{x_{ij} - \overline{x}_j}{\sqrt{S_{x_j x_j}}}, Y_{cs} :=$$

$$X_{j,cs} := \left(x_{1j,cs}, x_{2j,cs}, \dots\dots, x_{nj,cs}\right)', X_{R,cs} := \left[X_{1,cs} \ X_{2,cs} \ \dots. \ X_{p,cs}\right],$$

$$\Re := X'_{R,cs}X_{R,cs}, C_{cs} := \left(\left(C_{ij,cs}\right)\right)_{1\le i,j\le p} := \Re^{-1} \, , \mathfrak{D} := Det(\Re).$$

Multiple linear regression model with the centred and scaled variables: It is given by

$$Y_{cs} = X_{R,cs}\beta_{R,cs} + E \qquad (0.2)$$

Where E, the predictor variables $x_1, x_2, \ldots, x_p$ are as in the Model and $\beta_{R,cs} = (\beta_{1,cs}, \beta_{2,cs}, \ldots\ldots, \beta_{p,cs})'$ is the parameter vector.

Least squares estimator of $\beta_{R,cs}$: $\hat{\beta}_{R,cs} = \left(\hat{\beta}_{0,cs}, \hat{\beta}_{1,cs}, \ldots\ldots, \hat{\beta}_{p,cs}\right)' = \left(X'_{R,cs}X_{R,cs}\right)^{-1}X'_{R,cs}Y_{cs}$ provided $X'_{R,cs}X_{R,cs}$ is non-singular.

Fitted values: Fitted value corresponding to the i-th observation is given by

$$\hat{y}_{i,cs} = \hat{\beta}_{1,cs}x_{i1,cs} + \hat{\beta}_{2,cs}x_{i2,cs} + \ldots\ldots + \hat{\beta}_{p,cs}x_{ip,cs}$$

For our dataset, $X'_{R,cs}X_{R,cs}$ is non-singular as its determinant is 0.94987. Hence by the above formula of $\hat{\beta}_{R,cs}$, we get the least estimator of the regression coefficients $\hat{\beta}_{R,cs}$= (-11938.5 , 256.9 , -131.3 , 339.2 , 475.5 , 23848.5 , -353.0 , -1035.0 , -960.1).Thus the fitted values can be obtained by $\hat{y}_{i,cs}$.

The summary table of $\hat{\beta}_{R,cs}$ provides same p-values as we report in the above table. Therefore, fitted values provides same set of significant predictor variables as mentioned in (*). This happens because the Models (0.1) and (0.2) are equivalent in the sense that $\hat{\beta}_{R,cs}$ can be recovered from $\hat{\beta}$ and vice versa.

Determinant of $\mathfrak{R}$ i.e. $\mathfrak{D}$: If $\mathfrak{D}$ is close to 0, then multi-collinearity is suspected. We know that$0 \leq \mathfrak{D} \leq 1$. Any value of $\mathfrak{D}$ between 0 and 1 gives an idea of the degree of multi-collinearity. But it gives no information about the number and location of linear dependencies among predictor variables.

For the given dataset, $\mathfrak{D} = 0.94987$. As this value is close to 1, we can conclude that there is no multi-collinearity.

Inspection of $\mathfrak{R}$: The sample correlation coefficient between X_j and X_k is given by

$$r_{x_j x_k} = \frac{S_{x_j x_k}}{\sqrt{S_{x_j x_j} S_{x_k x_k}}} \; where \; S_{x_j x_k} = \sum_{i=1}^{n} \left(\left(x_{ij} - \overline{x}_j\right)(x_{ik} - \overline{x}_k)\right)$$

Note that $\mathfrak{R} = \left(\left(r_{x_j x_k}\right)\right)_{1 \le i,j \le p}$. If $\left|r_{x_i x_j}\right|, i \neq j$, is close to 1, then multi-collinearity is suspected and consequently X_i and X_j are suspected to be nearly linearly dependent. When more than two predictor variables are considered and if they are involved in near-linear dependency, then it is not necessary that any of the off-diagonal entries will be close to±1. Generally, pairwise inspection of sample correlation coefficients is not sufficient for detecting multi-collinearity in the data.

For the given dataset, the sample correlation matrix for (X_1, X_2, . . ., X_6) is given by

$$\mathfrak{R} = \begin{bmatrix} 1.00 & -0.02 & 0.11 & 0.04 & -0.03 & 0.00 \\ -0.02 & 1.00 & 0.05 & 0.02 & 0.08 & 0.00 \\ 0.11 & 0.05 & 1.00 & 0.01 & 0.00 & 0.16 \\ 0.04 & 0.02 & 0.01 & 1.00 & 0.01 & 0.02 \\ -0.03 & 0.08 & 0.00 & 0.01 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.16 & 0.02 & 0.00 & 1.00 \end{bmatrix}$$

This shows that the pairwise correlation between (Age and sex, BMI, children, smoker, region), (Sex and BMI, children, smoker, region), (BMI and children, smoker, region), (Children and smoker, region) and (Smoker and region) are all close to 0. Hence, we can conclude that there is **no Multi-collinearity**.

(0.1)

Variance inflation factors: The variance inflation factor corresponding to the regression coefficient of the predictor variable x_j is given by

$$VIF_j = C_{jj,cs} \; \forall j \; = \; 1,2, \ldots \ldots, p.$$

In practice, usually $VIF_j > 5 \; or \; 10$ for one or more $j \in \{1,2, \ldots, p\}$ indicates that the associate regression coefficients are poorly estimated because of multi-collinearity. The number of variance inflated factors that are large, say more than 5 or 10, indicate the number of regression coefficients suffering from

multi-collinearity. However, it sheds no light on the number of dependencies among the predictor variables. Moreover, the cut-off 5 or 10 is just a rule of thumb which may differ depending on the situation.

For our given dataset, we have

$$VIF_{age(1)} = 1.02, VIF_{sex(2)} = 1.01, VIF_{bmi(3)} = 1.04,$$
$$VIF_{children(4)} = 1.00, VIF_{smoker(5)} = 1.01, VIF_{region(6)} = 1.03,$$

Here, $VIF_j < 5\ \forall j = 1,2,\ldots,7$. Hence, there is **no Multi-collinearity** for our dataset. (0.2)

Condition indices: Let $\lambda_1(\Re) \geq \lambda_2(\Re) \geq \cdots \ldots \geq \lambda_p(\Re)$ be the eigenvalues of $\Re$. The condition indices of $\Re$ are defined as

$$C_j = \frac{\lambda_1(\Re)}{\lambda_j(\Re)}, j = 1,2,\ldots,p.$$

Multi-collinearity causes large value of the condition indices. Conventionally, if $C_j > 50\ or\ 100$ for one or more $j \in \{1,2,\ldots,p\}$, then multi-collinearity is suspected. The number of condition indices that are large, say more than 50 or 100, indicates the number of near-linear dependencies among the predictor variables. Moreover, large value of C_j indicates that the j-th principal component of the predictor variables is suspected to be responsible for multi-collinearity. However, this cut-off varies with different situations.

For our given dataset, $C_1 = 1.00, C_2 = 1.10, C_3 = 1.18, C_4 = 1.24, C_5 = 1.31, C_6 = 1.52$. Here, $C_j \ll 50$ i.e., C_j are much less than 50 $\forall j = 1,2,\ldots,6$. Hence, no predictor variable is suspected to be responsible for multi-collinearity. (0.3)

Conclusion on Multi-collinearity: From the above results **Inspection of $\Re$** (0.1), **Variance inflation factors (**0.2**)**, **Condition indices (0.3),** We can conclude that, there is **no Multi-collinearity** on our model.

Outliers Detection: Since the g-value for Forward selection method is close to 1, so we consider the subset of predictor variables x_1, x_3, x_4, x_5, x_6 for the updated Multiple linear regression model, which is obtained in Forward selection method. Also recall that n=1338 in this dataset.

Some important notations:

- $X_j = \left(x_{1j}, x_{2j}, \ldots, x_{nj}\right)', j = 1,3,4,5,6.$
- 1_n is the n × 1 vector whose all entries are 1.
- $X_{back} = [1_n, X_1, X_3, X_4, X_5, X_6]$
- $H_{back} = X_{back}(X'_{back}X_{back})^{-1}X'_{back} = \left(\left(h_{ij,back}\right)\right)_{1\le i,j\le n}$.
- $(X'_{back}X_{back})^{-1} = \left(\left(C_{ij,back}\right)\right)_{1\le i,j\le p}$
- $Y = (y_1, y_2, \ldots, y_n)'$
- $\hat{\beta}_{back} = (X'_{back}X_{back})^{-1}X'_{back}Y = \left(\hat{\beta}_{0,back}, \hat{\beta}_{1,back}, \hat{\beta}_{3,back}, \hat{\beta}_{4,back}, \hat{\beta}_{5,back}, \hat{\beta}_{6,back}\right)'$.
- $\hat{y}_{i,back} = \hat{\beta}_{0,back} - \hat{\beta}_{1,back}x_{i1} - \hat{\beta}_{3,back}x_{i3} - \hat{\beta}_{4,back}x_{i4} - \hat{\beta}_{5,back}x_{i5} - \hat{\beta}_{6,back}x_{i6}\ \forall\ 1 \le i \le n.$
- $MSE = \frac{1}{n-6}\sum_{i=1}^{n}\left(\left(y_i - \hat{\beta}_{0,back} - \hat{\beta}_{1,back}x_{i1} - \hat{\beta}_{3,back}x_{i3} - \hat{\beta}_{4,back}x_{i4} - \hat{\beta}_{5,back}x_{i5} - \hat{\beta}_{6,back}x_{i6}\right)^2\right)$.
- Let $\hat{y}_{j,(i)}$ be the predicted value of y at $x_1 = x_{j1}$, $x_3 = x_{j3}$, $x_4 = x_{j4}$, $x_5 = x_{j5}$, $x_6 = x_{j6}$ when fitting of the multiple linear regression model is done after deleting the i-th observation.
- $\hat{Y}_{(i)} = \left(\hat{y}_{1,(i)}, \hat{y}_{2,(i)}, \ldots, \hat{y}_{n,(i)}\right)'$.
- Let $\hat{\beta}_{j,(i)}$ be the least squares estimator of β_j when fitting of the multiple linear regression model is done after deleting the i-th observation.
- $\hat{\beta}_{(i)} = \left(\hat{\beta}_{0,(i)}, \hat{\beta}_{1,(i)}, \hat{\beta}_{3,(i)}, \hat{\beta}_{4,(i)}, \hat{\beta}_{5,(i)}, \hat{\beta}_{6,(i)}\right)'$.
- The residual sum of squares from the fitting of multiple linear regression model, after deleting the i-th observation, turns out to be

$$S^2_{(i)} = \sum_{j\neq i}\left(\left(y_j - \hat{\beta}_{0,(i)} - \hat{\beta}_{1,(i)} - \hat{\beta}_{3,(i)} - \hat{\beta}_{4,(i)} - \hat{\beta}_{5,(i)} - \hat{\beta}_{6,(i)}\right)^2\right)$$

i) The large value of $h_{ii,back}$ indicates that the i-th observation is a leverage point. The plot of $\{ h_{ii,back} : 1 \leq i \leq n \}$ is given in Figure 1. It shows that the observations present on the fourth quadrant are all leverage points.

ii) Cook's distance statistic corresponding to the i-th observation is given by

$$D_i = \frac{(\beta_{back}-\beta_{(i)})'(X'_{back}X_{back})(\beta_{back}-\beta_{(i)})}{6MSE}.$$

We usually consider the i-th observation to be an influential point if the corresponding D_i is large. The plot of $\{D_i : 1 \leq i \leq n\}$ are given in Figure 2. It shows that all the observations (marked as Red line on the X-axis) in Figure 2 are all influential points.

iii) DFFITS statistic corresponding to the i-th observation is given by

$$DFFITS_i = \frac{\hat{y}_{i,back} - \hat{y}_{i,(i)}}{\sqrt{S^2_{(i)}h_{ii,back}}}.$$

We usually consider the i-th observation to be an influential point if the corresponding $DFFITS_i$ is large. The plot of $\{DFFITS_i : 1 \leq i \leq n\}$ are given in Figure 3. It shows that all the observations (the lines crossed the upper and lower boundary as shown in Figure 3) are all influential points.

iv) DFBETAS statistic corresponding to the (j, i)-th observation is given by

$$DFBETAS_{j,i} = \frac{\hat{\beta}_{j,back}-\hat{\beta}_{j,(i)}}{\sqrt{S^2_{(i)}C_{jj,back}}}$$

Large magnitude of $DFBETAS_{j,i}$ indicates that i-th observation has considerable influence on the j-th regression coefficient. Plots of DFBETAS statistics are given in Figures 4 and 5.

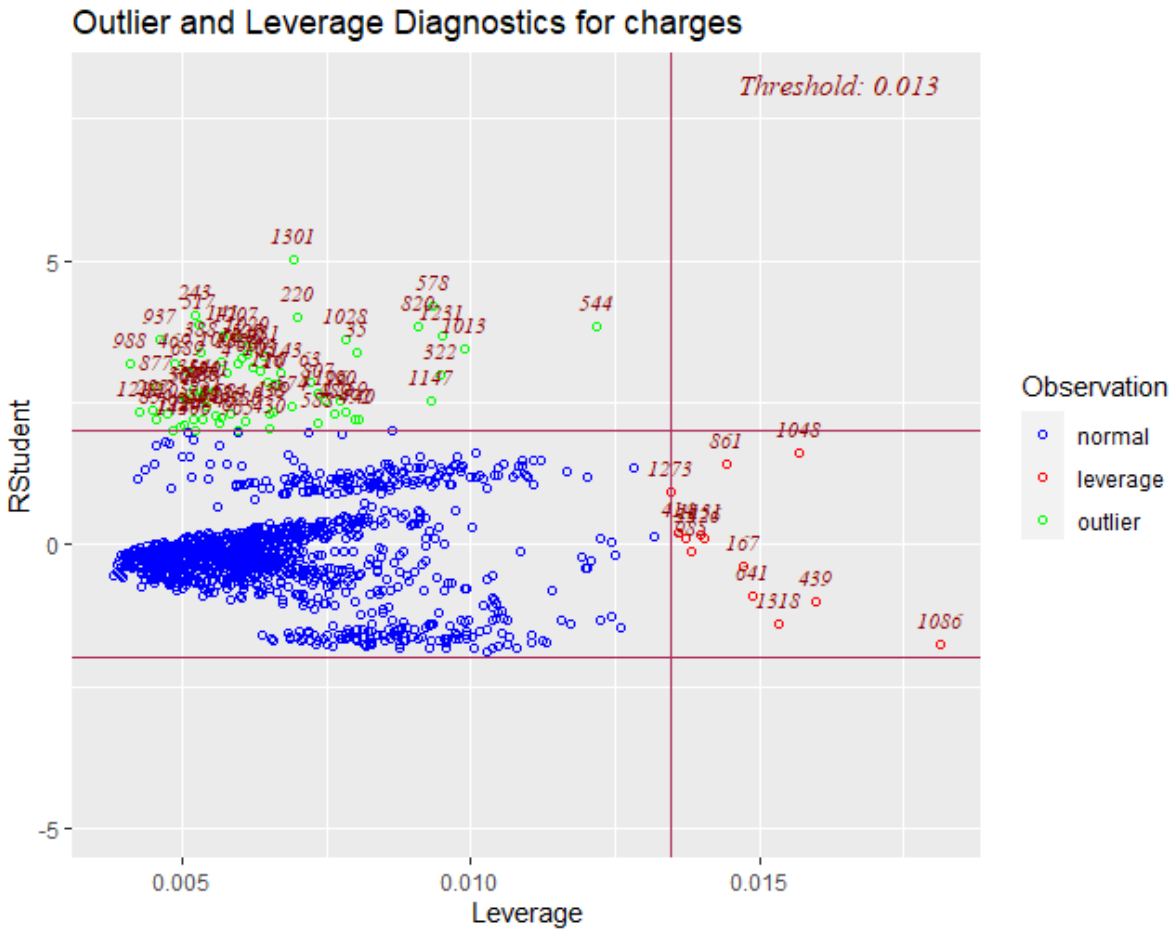


Figure 1: Leverage diagnostic for "Insurance" dataset

Here we have noted there several outliers and leverage values, though no observation which is both leverage and outliers is present. As an exact count we noted there 70 outliers and 13 leverage values where the threshold is taken as 0.013. Here we note among the observations 1301th observation is an significant outlier (>5)and 1086th observation is a noticeable leverage point.

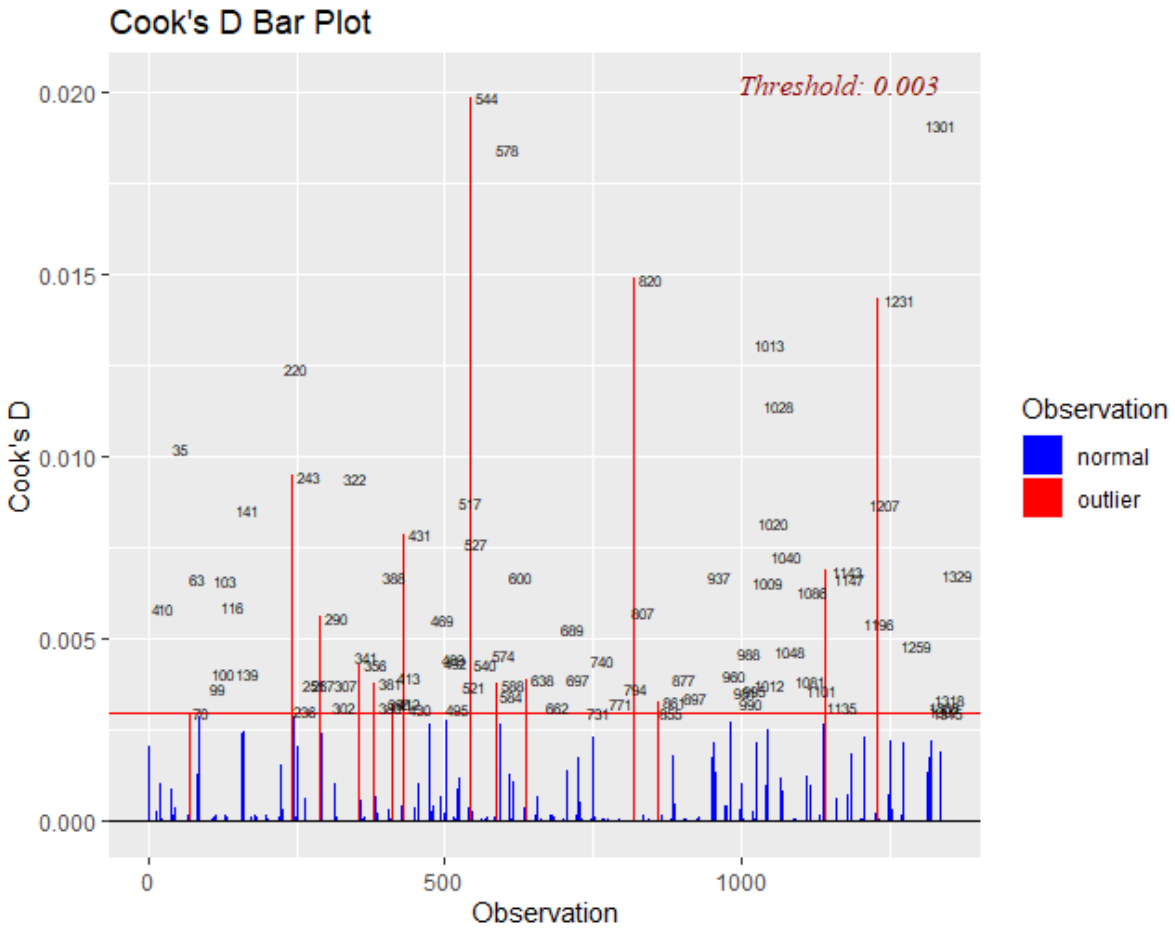


Figure 2: Plot of Cook's distance statistic for "Insurance" dataset

On the other hand the based on Cook's distance statistic, under threshold .003 we have obtained that there are 87 outliers by this model among which observations with index 544, 578 and 1301 are noticeably significant.

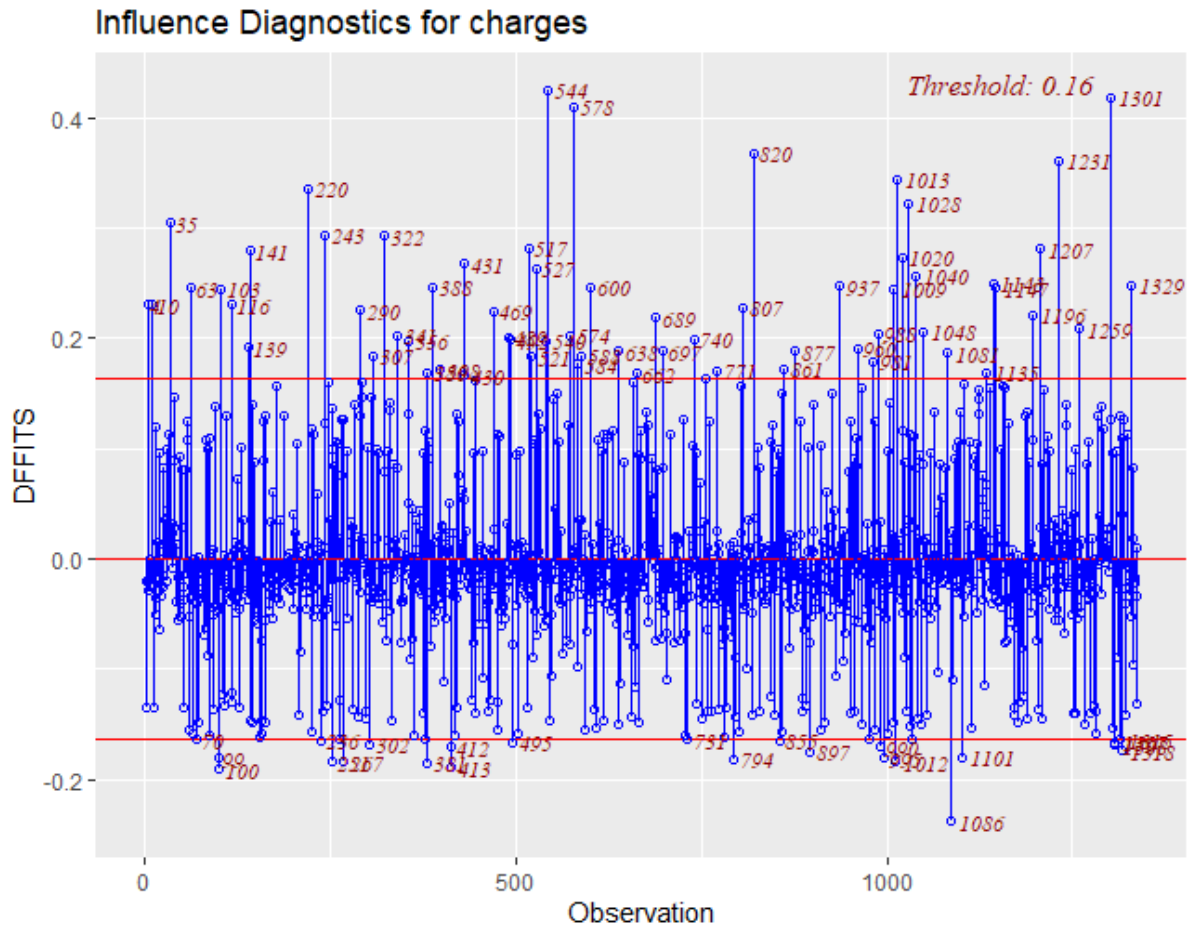


Figure 3: Plot of DFFITS for "Insurance" dataset

Also by the DFFITS value under the threshold 0.16 it is noted the same points obtained before which are observations with index 544,578,1301,1086 are influential points. Also by this method we have obtained the same 87 observations as obtained by Cook's distance as the influential observations.

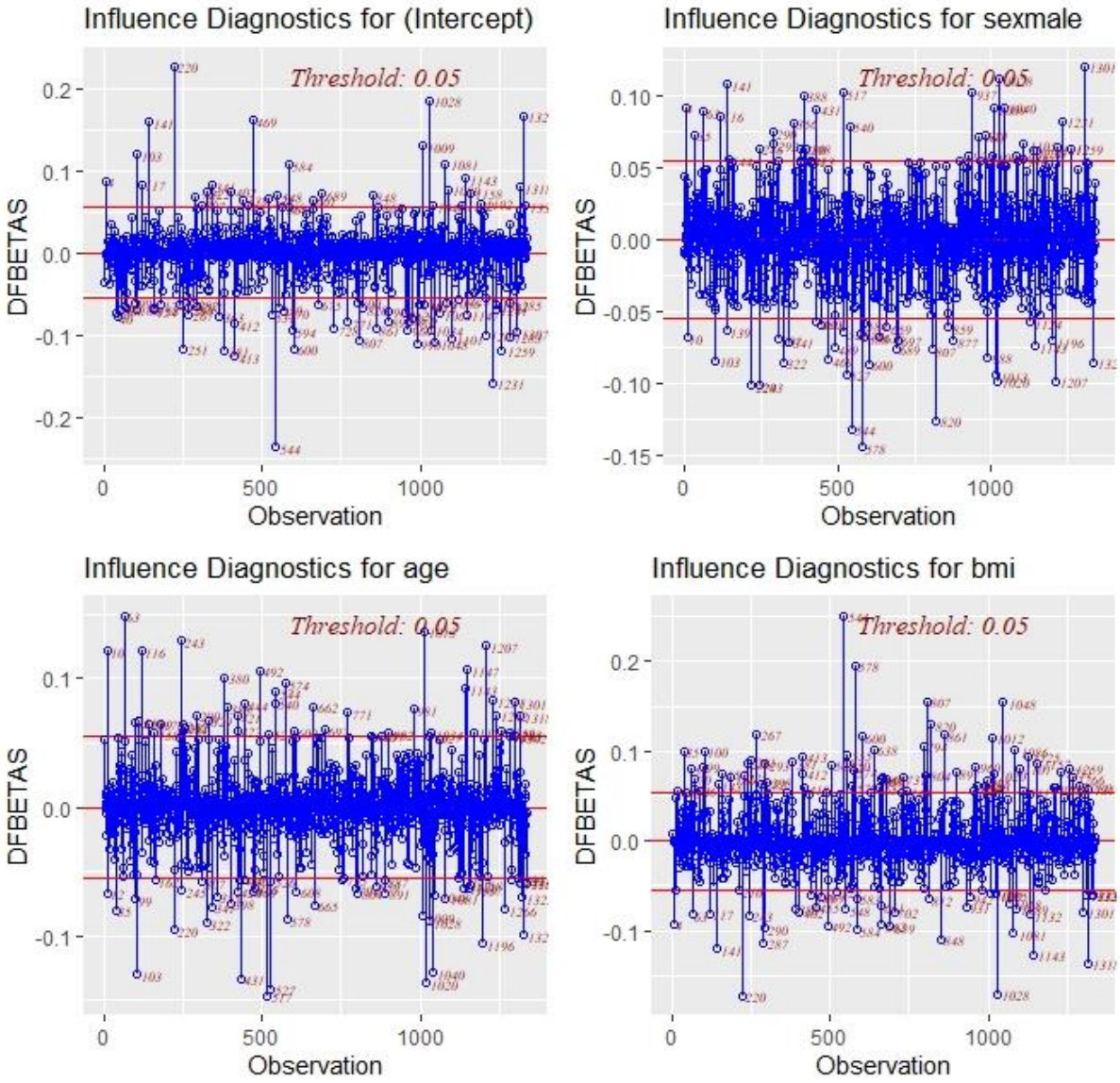


Figure 4: Plot of DFBETAS for the intercept parameter and the regression coefficients to sex , age and bmi

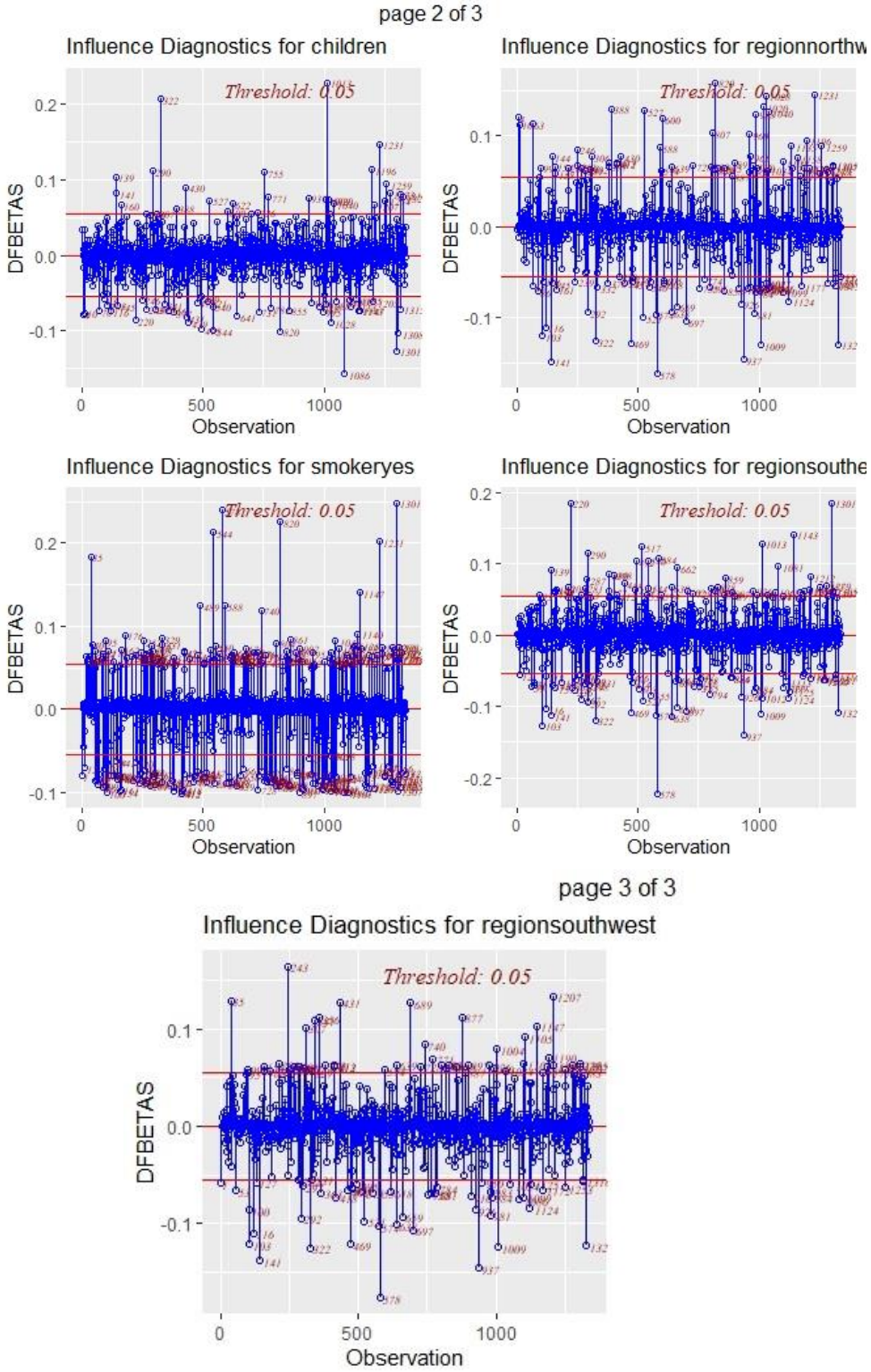


Figure 5: Plot of DFBETAS for the regression coefficient corresponding to number of children and region

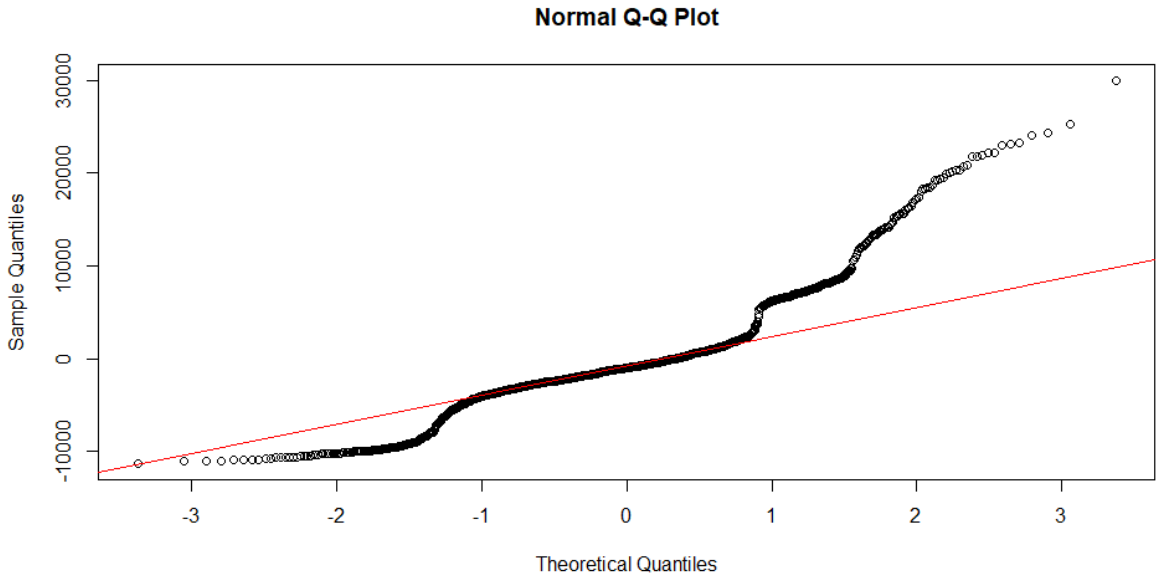
Now by the Influence Diagnostics by DFBETA we have observed as below (here from each plot we noticed there are many points which affect the

coefficients but we will only mention thee four observations that we obtained from the previous diagnostics as influential values):

- 544th observation influences the intercept parameter
- 544, 578 and 1301th observation influences regression coefficient corresponding to variable sex.
- 544 and 578th observation influences regression coefficient corresponding to variable bmi.
- 1086th observation influences regression coefficient corresponding to variable children
- 544, 578 and 1031th observation influences regression coefficient corresponding to variable region.

As a remedy we can suggest to remove such observations. We can also check the Normal QQ plot which is for checking whether the normality assumption of the error term is satisfied or not.

The plot is as below:



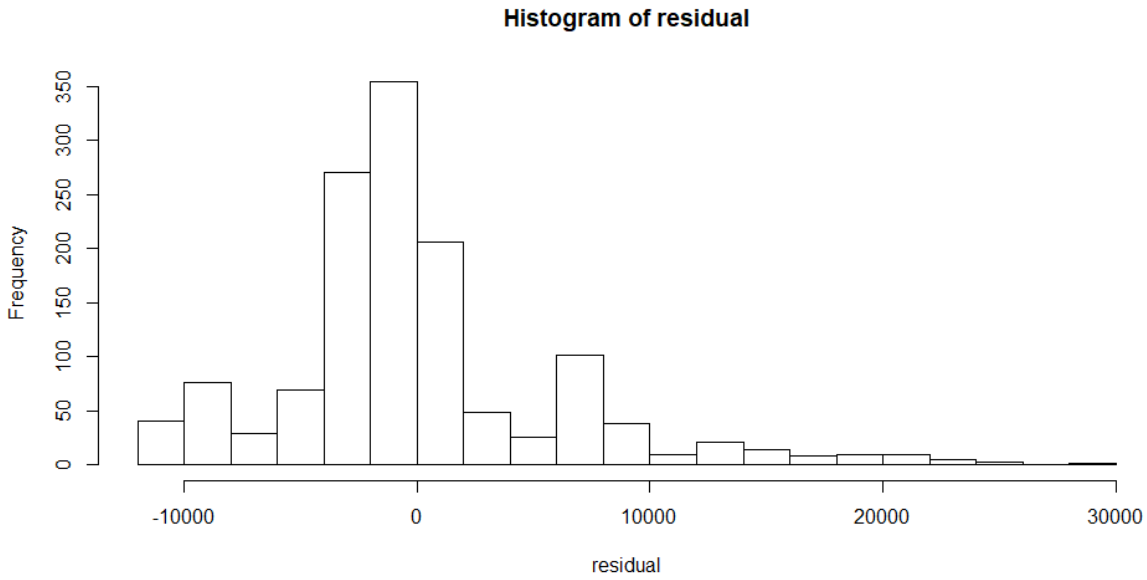


Figure 6: Normal QQ plot and Histogram of the residuals obtained from full model.

Here also from the QQ plot it is evident there is a presence of significant number of outliers and influential points which creates a longer tail for the distribution of residuals which is not supporting the Normality assumption of the model.

Final Model:

Now lastly we want to mention the final model which is based on all covariates except the variable sex (due to non significant p value as well as by using several selection methods). The final estimated coefficients are as below:

Regression coefficients	Estimated value	Standard Error	t-value	p-value
β_0	**-11990.27**	**987.76**	**-12.250**	**$<2\times10^{-16}$**
β_1 (age)	**256.97**	**11.89**	**21.610**	**$<2\times10^{-16}$**
β_2 (bmi)	**338.66**	**28.56**	**11.858**	**$<2\times10^{-16}$**
β_3 (children)	**474.57**	**137.74**	**3.445**	**0.000588**
β_4 (smoker yes)	**23836.30**	**411.86**	**57.875**	**$<2\times10^{-16}$**
β_5 (region north west)	**-352.18**	**476.12**	**-0.740**	**0.459618**

β_6 (region south east)	**-1034.36**	**478.54**	**-2.162**	**0.030834**
β_7 (region south west)	**-959.37**	**478.78**	**-2.008**	**0.044846**

The multiple R square value for this model is 0.7509 and adjusted R square value is 0.7496 which is very close implying the higher parsimony of the model. Also the value 0.7509 implies by this model 75.09 percent of total variability is explained which is considerably high proportion of explained variability. So we can say apart from some violation of assumption this model satisfactorily estimates the charge of insurance of a particular person.

Conclusion:

This whole model set up and estimation procedure of model showing in real life the charges of insurance is significantly and positively related to age, bmi, number of children and with the fact that the insurer is a smoker, which is justifiable and acceptable under the scenario that all these parameters are more or less direct representation of the goodness of health of a person. In simple words the higher age , or higher number of children, or higher bmi all are more or less represents the higher risk factor of the person being ill and hence the person's charges also increases. The smoking habit also adds additional positive effect in these charges. Beside that we also note, among all regions Northeast region people tend to pay higher charges than any other region (as all other region coefficients are negative) whereas the southeast region on average (and significantly) pay lesser charges than other region people, keeping other covariates fixed. There might be a geo economical reason behind this difference which is for further research. Additionally it is also to be mentioned the factor sex does not have any differentiating effect on charges. Lastly we can say, since charge is a economic variable hence there are opportunities for future assessment if instead of charge values we can build model based on log values (natural logarithm) of charge and assess whether that model is better than this one and also more efficient to reduce the influential and outlier observations.