Group Project on Regression Analysis

Instructor : Prof. Swagata Nandi

By Group 1 :

Bhushan Suresh Malgunkar, Hemant Banke, Sayantan Deb Barman

Date : December 22, 2021

Indian Statistical Institute, New Delhi

ABSTRACT

This project analyzed the Diabetes data containing 144 observations of Normal, Chemically diabetic and Overt. Diabetic patients. We aimed to find the change in Plasma Insulin due to different factors such as Glucose levels in blood and Relative weight for the three types of patients.

To achieve this we fitted Multiple Linear Regression Models for each type of patient and eventually compared the predicted values of plasma insulin against plasma glucose and relative weight for all three models. We noticed the presence of Multicollinearity between two explanatory variables for 'Overt Diabetic' patients. To deal with this employed Ridge Regression to get estimates and study significance of correlated variables. In 'Normal' patients, we faced the issue of non normality in residuals. We used Box-Cox transformation on the response variable to ensure the assumptions of linear models are not violated. In 'Chemically Diabetic' patients we observed presence of curvature from partial residual plots, for which we transformed the explanatory variables to ensure linearity with the response variable.

This project was also aimed at practically applying the concepts learned in 'Regression Techniques' Course under the guidance of Prof. Swagata Nandi.

ACKNOWLEDGEMENT

We wish to offer our heartiest gratitude to our supervisor, Professor Swagata Nandi, without whose invaluable expertise, mentorship and guidance this project would not have been a success. We shall forever be indebted to her for her constant support throughout the project in every possible way.

The completion of this project could not have been possible without the participation and assistance of individuals contributing to this project. However, we would like to express our deep appreciation and indebtedness to our teachers and supervisors for their endless support, kindness, and understanding during the project duration.

TABLE OF CONTENT

1. PROJECT OVERVIEW & EXPECTATIONS

Insulin is a hormone that helps move blood sugar, known as glucose, from our bloodstream into cells. Insulin plays a key role in keeping glucose at the right levels. If glucose levels are too high or too low, it can cause serious health problems. Diabetes is the most common cause of abnormal glucose levels. In ‘Type 1 Diabetes' the body makes little to no insulin, this causes an increase in blood glucose levels as it can’t get into cells. In ‘Type 2 Diabetes' the body is able to make insulin, but the cells don’t respond well to insulin and can’t easily take up enough glucose from blood.

Overt diabetes is the most advanced stage, characterized by elevated fasting blood glucose concentration. Preceding overt diabetes is the latent or chemical diabetic stage, with no symptoms of diabetes but demonstrable abnormality in glucose tolerance.

We aim to observe the dependence of Insulin in blood plasma on relative weight, fasting glucose level, random glucose level and steady state glucose level for Normal, Chemically Diabetic and Overt Diabetic patients. The difference in predicted insulin levels in the three cases can give us an insight on how insulin helps in transferring glucose to cells and how this process is affected during Diabetes.

In both Type 1 and Type 2 Diabetes, glucose is not transferred to body cells causing high glucose levels in blood plasma. Hence we expect higher plasma glucose concentration in Overt Diabetic patients and slightly higher for Chemically Diabetic patients.

2. GENERAL DISCUSSION ON REGRESSION

A major activity in statistics is the building of statistical models that hopefully reflect the important aspects of the object of study with some degree of realism. In particular, the aim of ***regression analysis*** is to construct mathematical models which describe or explain relationships that may exist between variables.

Statistical models are fitted for a variety of reasons. One important reason is that of trying to uncover causes by studying relationships between variables. Usually, we are interested in just one variable, called the ***response variable***, and we want to study how it depends on a set of variables called the ***explanatory variables.***

Important assumptions in regression analysis:

1. There should be a ***linear relationship*** between response variable and explanatory variable(s).
2. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as ***Autocorrelation***.
3. The independent variables should not be correlated. Absence of this phenomenon is known as ***multicollinearity***.
4. The error terms must have constant variance. This phenomenon is known as ***homoscedasticity***. The presence of non-constant variance is referred to as heteroscedasticity.
5. The error terms must be ***normally distributed***.

3. DATA DESCRIPTION

The Diabetes Data contains 144 observations and 7 columns. A Data Dictionary was provided with the data and is explained below :

1	patient	Patient ID
2	relwt	Relative weight
3	glufast	Fasting Plasma Glucose (mg/dl)
4	glutest	Test Plasma Glucose
5	sspg	Steady State Plasma Glucose
6	instest	Plasma Insulin during Test
7	group	Clinical Group

Clinical group is a categorical variable containing three levels : 1 = 'Overt Diabetic ', 2 = 'Chemically Diabetic', 3 = 'Normal'. These are the three patient groups we need to study Insulin levels for. The column ‘relwt’ contains the relative weight of every individual compared to the average weight of the population. The columns ‘glufast’, ‘glutest’ and ‘sspg’ are measures of glucose in blood plasma after fasting, randomly and in steady-state after insulin infusion respectively. The column ‘instest’ is a measure of Insulin in blood plasma.

The data does not have any missing or null values. But the data does have 52.7% (i.e. 76 out of 144 individuals) ‘Normal’ individuals, this leaves us only with 32 patients with Overt Diabetes and 36 patients that are Chemically Diabetic.

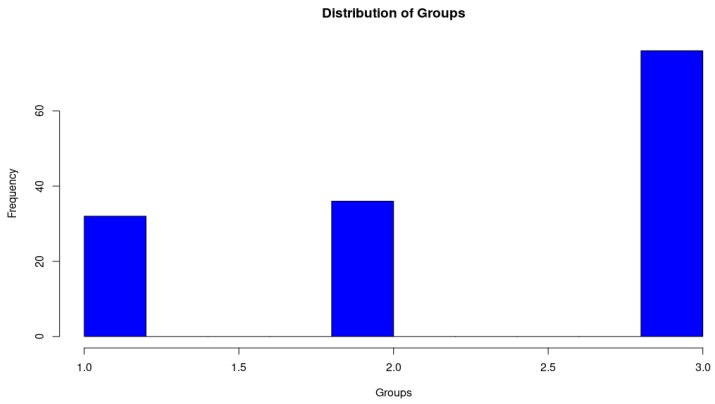


Fig 3.1; Distribution of Groups

4. PRELIMINARY ANALYSIS

Before predicting and modelling we will first try to find information already present in the given data. For this we can create correlation plots and pair plots for every variable. This will give us an intuition about relation between different variables and also distribution of all variables.

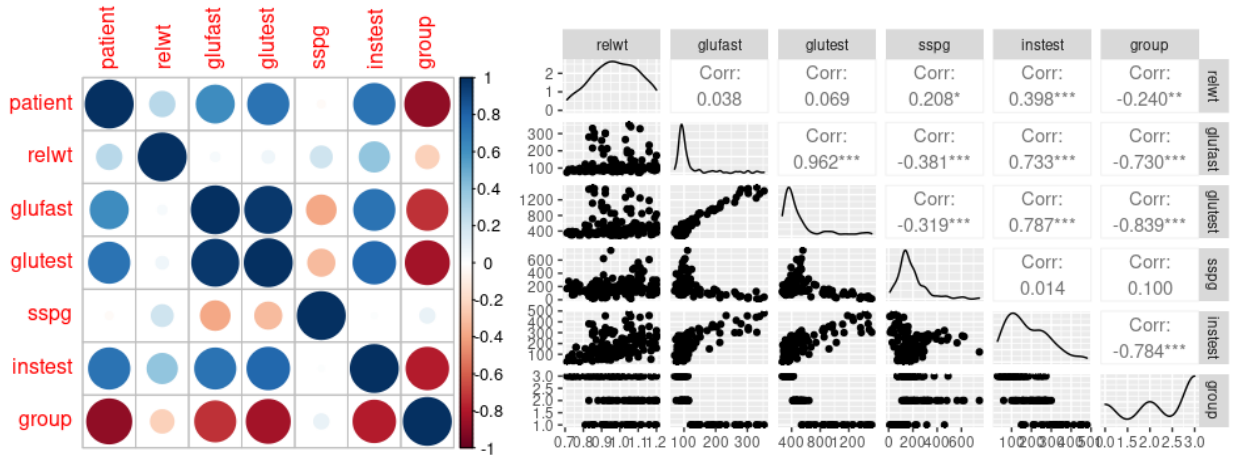


Fig 4.1; Correlation Plot (Left). Pair plot of all variable except patient ID (Right)

Clearly there is a significant linear relation between ‘glufast’ and ‘glutest’, but from pair plots we can also observe that more observations lie in the lower end of glucose levels in all ‘glufast’, ‘glutest’ and ‘sspg’.

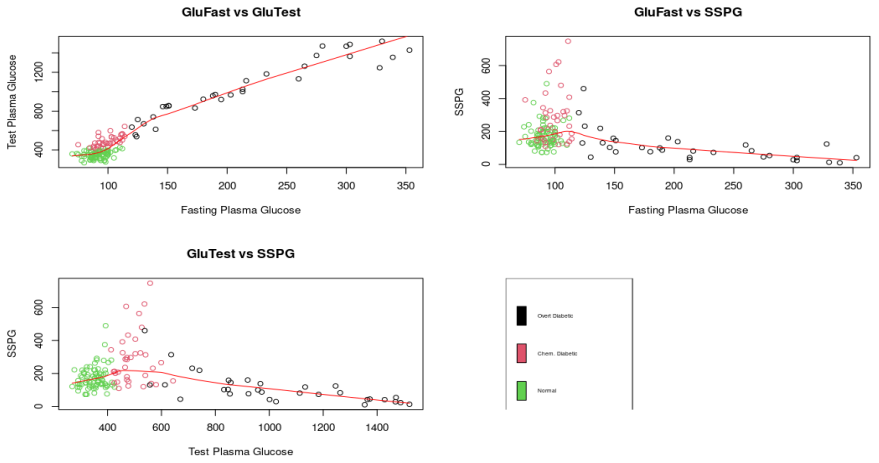


Fig 4.2; Plot of glufast vs glutest, glufast vs sspg, glutest vs sspg coloured according to group (black : Overt Diabetic , red : Chem Diabetic, green : Normal)

From Figure 4.2, we observe there is no significant relation between these measures of glucose for Normal and Chem Diabetic patients. But for Overt Diabetic patients, there is collinearity between glutest and glufast, also SSPG is decreasing for higher values of glufast and

glutest. The higher number of observations in lower glucose levels corresponds to Normal and Chem Diabetic patients, which is expected as in Normal patients pancreas secrete enough insulin and body cells respond well by taking up glucose from blood plasma.

We will now observe the relationship of relative weight and all measures of glucose against Insulin levels.

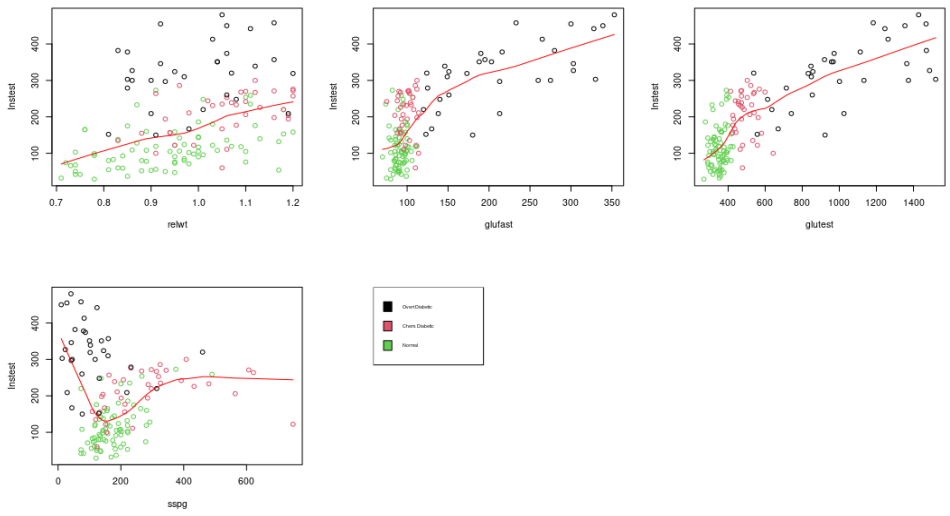


Fig 4.3; Plot of relwt, glufast, glutest and sspg against instest coloured according to group (black : Overt Diabetic , red : Chem Diabetic, green : Normal)

We observe that relative weight and Insulin levels have a good linear relationship in Normal and Chem Diabetic patients, which can be explained as the body needs to secrete more insulin to absorb higher amounts of glucose resulting from a heavier diet. Both glufast and glutest don't have a significant relation with Insulin levels in case of Normal and Chem Diabetic patients but there is almost linear relation in case of Overt Diabetic patients, which is expected as relatively more insulin will be secreted by body if higher amounts of glucose is present in blood plasma and majority of diabetic cases are Type 2 cases which means body is able to secrete insulin but cells don't respond well to it.

SSPG seems to have an interesting relation with Insulin levels. For Overt Diabetic patients SSPG is lowest and the Insulin levels decrease with increase in SSPG. Normal individuals lie in the depression of the curve with Insulin levels slightly increasing with increase in SSPG. For Chemically Diabetic patients, with increase in SSPG, insulin levels start to increase but gradually level off at 244 units.

As the data is divided into three groups and our goal is to study the behaviour of insulin levels for each group of individuals, we will fit separate models for each group and then eventually compare the predictions and findings to get more insight on Insulin levels in blood plasma.

5. Fitting Model For Normal Group

We have 76 observations corresponding to the Normal group and 6 variables. As we need to predict Insulin levels, we will choose ‘instest’ as our response variable and 'relwt', 'glufast', 'glutest', 'sspg' as explanatory variables. We will first observe the pair plots for the Normal group to eyeball presence of linearity with response and any multicollinearity.

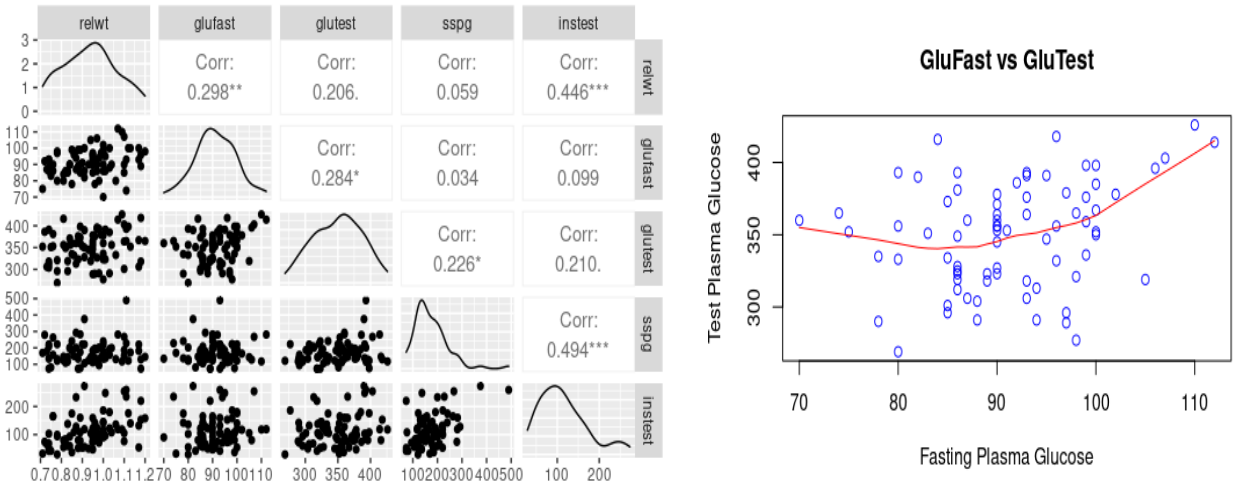


Fig 5.1; Pair plot of all variable except patient ID for Normal Group (Left) ;
GluFast vs GluTest for Normal Group (Right)

From Figure 5.1, we notice presence of good correlation of variables relwt and sspg with response instest. Also the relationship between glufast and glutest does not have as significant correlation as was in combined data. This hints towards very low multicollinearity which we will check later using Variance Inflation Factors (VIF’s) and Condition number.

5.1. Distribution of Response

From Figure 5.2, we can conclude that our response variable ‘instest’ is not normally distributed but rather skewed as is also suggested by following metrics of ‘instest’ :

Summary :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
29.00	73.75	105.00	114.00	142.25	273.00

SD : 57.53283
Skewness : 0.9379977
Kurtosis : 3.439338

Shapiro-Wilk normality test

data: x
W = 0.92671, p-value = 0.0002933

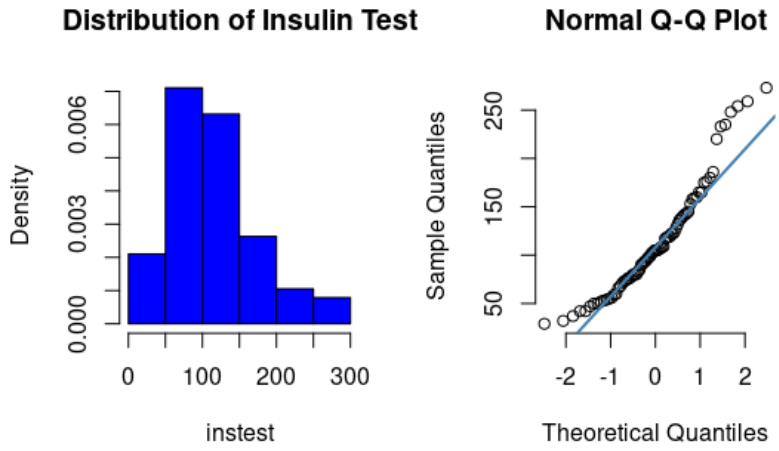


Fig 5.2; Density and Normal QQ plot for instest

As the p-value of Shapiro-Wilk test is < 0.05, we can reject the null hypothesis of instest being normally distributed.

5.2. Initial Linear Model

Fitting a Linear model using the lm function in R gives us the following model summary:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-115.88300	69.51525	-1.667	0.0999 .
relwt	191.43336	42.73489	4.480	2.80e-05 ***
glufast	-0.37262	0.68172	-0.547	0.5864
glutest	0.05035	0.15211	0.331	0.7416
sspg	0.38704	0.07754	4.992	4.12e-06 ***

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

Residual standard error: 45 on 71 degrees of freedom
Multiple R-squared: 0.4208, Adjusted R-squared: 0.3882
F-statistic: 12.9 on 4 and 71 DF, p-value: 6.056e-08

As suggested by adjusted R^2, this linear model does not explain variation in 'instest' very well, but as the p-value of the F test is < 0.05 we can reject the null hypothesis of all explanatory variables being insignificant in predicting response (i.e each coefficient $\beta_i = 0$ corresponding to explanatory variables).

5.3. Analyzing Residuals

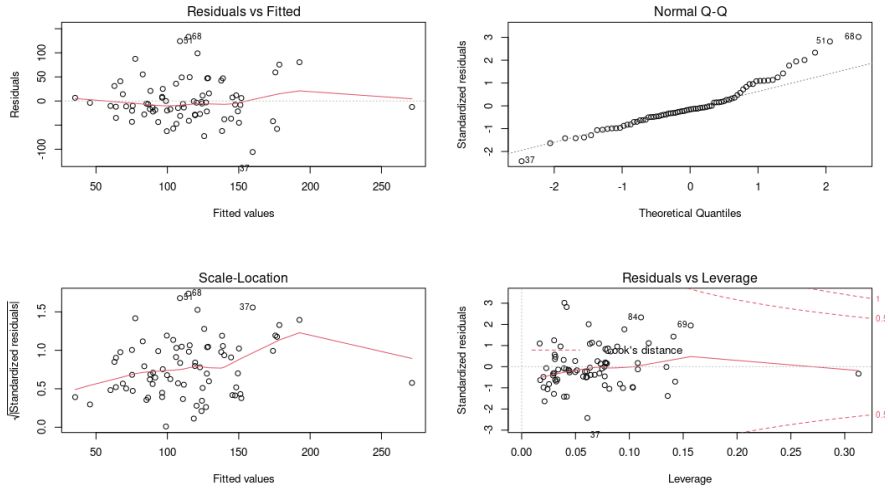


Fig 5.3; Residual Plots 1. Residual vs Fitted plot, 2. Normal QQ plot, 3. Scale Location plot, 4. Cook's Plot

From the Normal QQ plot we observe that the distribution of residuals is not normal, as is also indicated by following metrics :

Summary :

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -105.752  -27.080   -6.976    0.000   17.104  133.223
SD : 43.78449
Skewness : 0.7630806
Kurtosis : 3.960185
```

Shapiro-Wilk normality test

```
data:  x
W = 0.95307, p-value = 0.006875
```

As the p-value of Shapiro-Wilk test is < 0.05, we can reject the null hypothesis of residuals being normally distributed.

Checking Homoscedasticity

From the residual vs fitted values plot, we observe there is no significant increase in spread of residuals with increase in fitted values. For graphically checking for homoscedasticity we can plot b_i vs fitted values where,

$$b_i = \frac{e_i^2}{1-h_i}$$, where e_i is i'th residual and h_i is i'th hat matrix diagonal

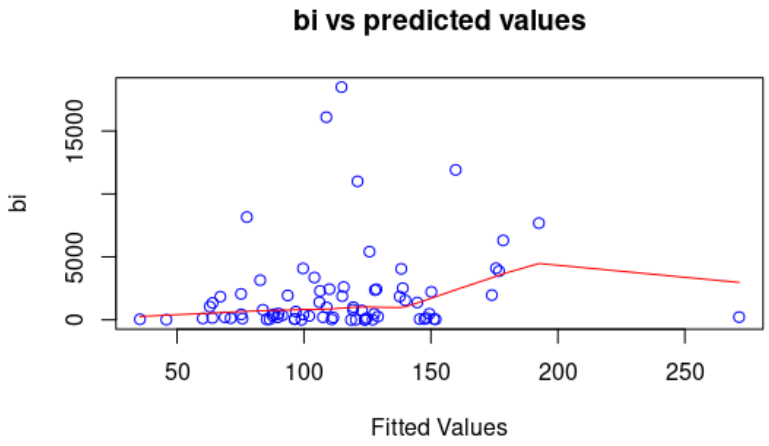


Fig 5.4; bi vs fitted values plot

As the plot does not have significant wedge shape and Residual vs Fitted values plot is not fan shaped, we cannot conclude the presence of heteroscedasticity. A more formal test for homoscedasticity is the Breusch Pagan Test which tests for linear relationship between error variances with explanatory variables.

statistic	p.value	parameter	method	alternative
2.68	0.612	4	Koenker (studentised)	Heteroskedasticity

As the p-value of the Breusch Pagan test is significant (>> 0.05), we can not reject the null hypothesis of presence of homoscedasticity.

Checking Multicollinearity

We initially observed no significant correlation between the explanatory variables for the Normal group. A result of multicollinearity is higher variance of parameter estimates. Variance Inflation Factor (VIF) is a measure of how much the variance of the estimated regression coefficient is inflated by the existence of correlation among the predictor variables in the model.

$$VIF_i = \frac{1}{1 - R_i^2}$$

is the VIF for i'th estimate $\widehat{\beta}_i$, where R_i^2 is the multiple R^2 for the regression of i'th variable on the other explanatory variables. The VIF's of variables in our model are :

relwt	glufast	glutest	sspg
1.117648	1.165242	1.164857	1.055634

The values of VIF > 4 need to be handled carefully but in our case the values are fairly close to 1. Another check is the condition number being considerably large (>30) but in our case it is only 1.571748. Hence we can safely conclude there is no multicollinearity in this model.

Checking Autocorrelation

Autocorrelation (or Serial Correlation) is the presence of correlation between successive residuals that are sequenced in time. Since our data is not indexed in time, we shall follow the natural index of patient Id's. The Durbin Watson test looks for a specific type of serial correlation, the AR(1) process in which the current value is based on immediately preceding value.

lag	Autocorrelation	D-W	Statistic p-value
1	-0.1031975	2.135519	0.61

Alternative hypothesis: rho != 0

As p-value is significant (>> 0.05), we can not reject the null hypothesis of no serial correlation.

5.4. Box-Cox Transformation

As neither the residuals nor the response is normally distributed, we will transform the response using a box-cox transformation.

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln y_i & \text{if } \lambda = 0, \end{cases}$$

To find the appropriate value of λ, we will choose the λ that maximizes the maximum value of the log likelihood function $l_{max}(\lambda)$.

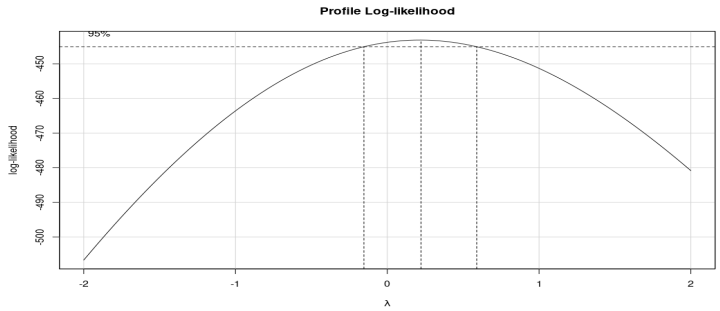


Fig 5.5; max log likelihood plot against λ

This plot is maximized at $\lambda = 0.22$. After transforming response using the chosen λ, from Fig 5.6 we observe that the linearity of variables with response has not been affected much while our response variable has become significantly normally distributed.

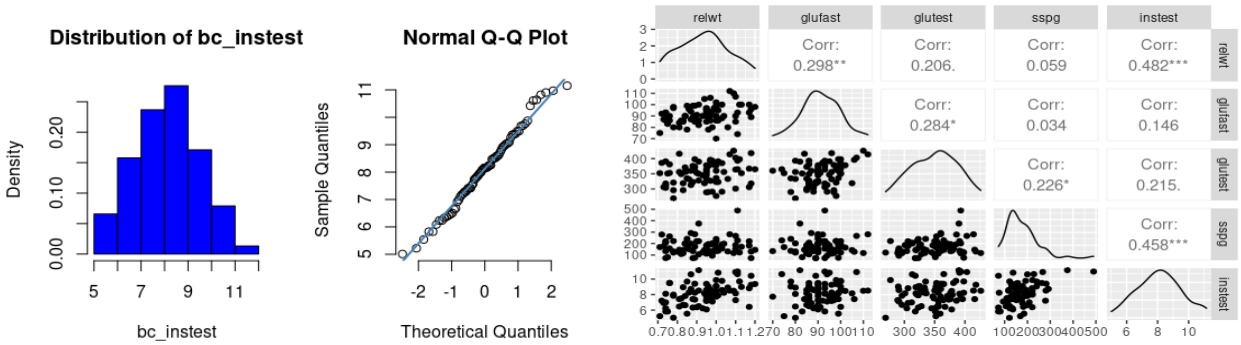


Fig 5.6; Distribution and Normal QQ plot of instest after box cox transformation (Left) ;
Pair plot of all variable except patient ID for Normal Group (Right)

Shapiro-Wilk test on transformed response gives a significant p-value (>> 0.05), hence we can not reject null hypotheses of response being normal.

Summary :

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.010   7.202   8.158   8.118   9.042  11.152
```

Shapiro-Wilk normality test

data: x

W = 0.98819, p-value = 0.7089

Fitting the linear model after this transformation gives us the residual plot in Fig 5.7. The Normal QQ plot does hint that residuals are normally distributed which can again be confirmed by the Shapiro-Wilk test which gives a significant p-value = 0.5015 >> 0.05.

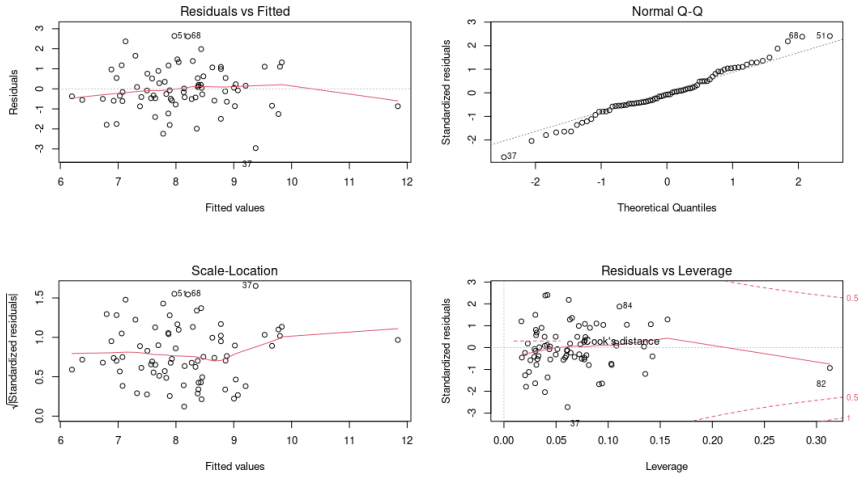


Fig 5.7; Residual Plots 1. Residual vs Fitted plot, 2. Normal QQ plot, 3. Scale Location plot, 4. Cook's Plot

As we have only transformed the response the VIF values remain the same but we do need to recheck homoscedasticity. The Breusch Pagan Test gives us a p-value = 0.997 which is significant, hence we cannot reject the null hypothesis of homoscedasticity.

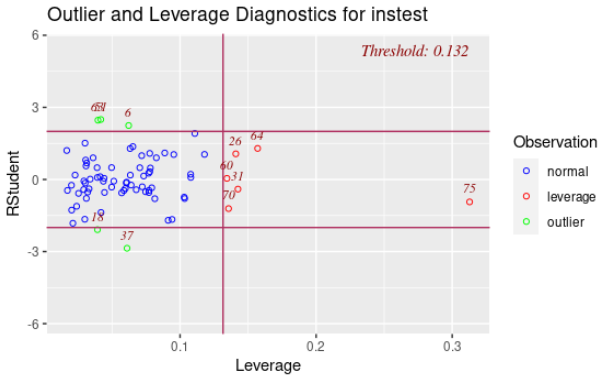
5.5. Dealing with Outliers

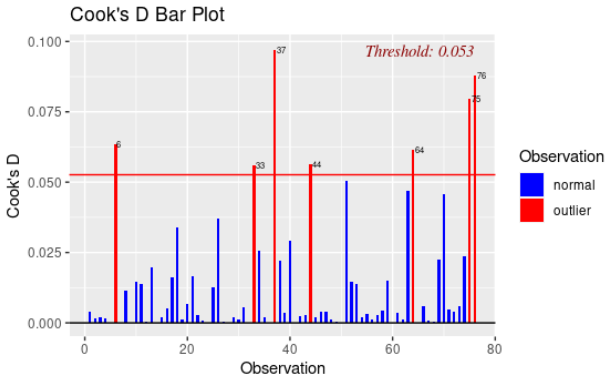
Type of Outliers

- **High Leverage Point** : A data point has high leverage if it has high "distance" from the center of the explanatory data. Hat matrix diagonals (h_i) are reasonable measures of leverage. $h_i > 2\frac{p}{n}$ acts as a good measure for detecting high leverage points, where p is the number of explanatory variables including intercept and n is the number of observations.
- **Error Outliers** : These points have large errors from the true regression plane. A reasonable estimate of error is residuals obtained from fitting the model. $|t_i| > 2$ is a good measure for detecting points with large errors, where t_i is the externally studentized residual corresponding to i'th observation.

Points that are both high leverage and have high error can highly influence the fitted regression plane and hence are called **influential points**. The measures we used to look for influential points are based on leave-one-out diagnostics and are listed as follows :

- DFBETAS (Difference in Betas) : Points having $|DFBETA_{i,j}| > \frac{2}{\sqrt{n}}$, for any j'th explanatory variable
- DFFITS (Difference in Fitted values) : Points having $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$, where p is number of explanatory including intercept
- Cook's Distance (Distance between parameter vector $\hat{\beta}$ and parameter vector removing i'th point $\hat{\beta}(i)$) : Points having $D_i > F^{0.1}_{p,n-p}$
- COVRATIO (Measures the effect of on the covariance matrix of parameter estimates) : Points having $|COVRATIO_i - 1| > 3\frac{p}{n}$





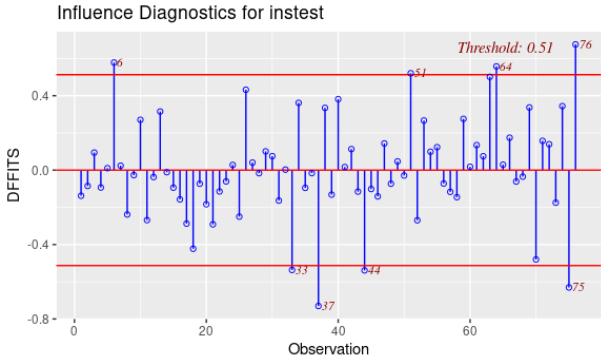


Fig 5.8; Outlier and Leverage Diagnostics (Plotting studentized residuals vs h_i) (Left); Cook's Distance Plot (Right)

We choose to remove points classified as outliers by more than 2 of these measures, there are 9 such points with patient ID's **6 18 33 37 44 51 68 82 84**. As these points do not arise because of measurement error or imputation we will choose to remove these points.

On fitting the regression model after removing these points, the adjusted R^2 increases considerably to **51%** and none of the assumptions of linear regression are violated when checked similar to Section 5.3.

5.6. Checking for Curvature

Figure 5.6 does reveal that 'relwt' and 'sspg' have linear relationship with response variable, for further checking the relationship of explanatory variables with response we can produce Partial Residual Plots (or Component + Residual plot) and Added Variable Plots.

Partial Residual Plots :

It is a plot of partial residual $e_i^* = e_i + \hat{\beta}_j x_{i,j}$ vs $x_{i,j}$, where $\boldsymbol{x_j}$'s are explanatory variables. This plot can reveal the shape of the relationship between $\boldsymbol{x_j}$ and the response variable.

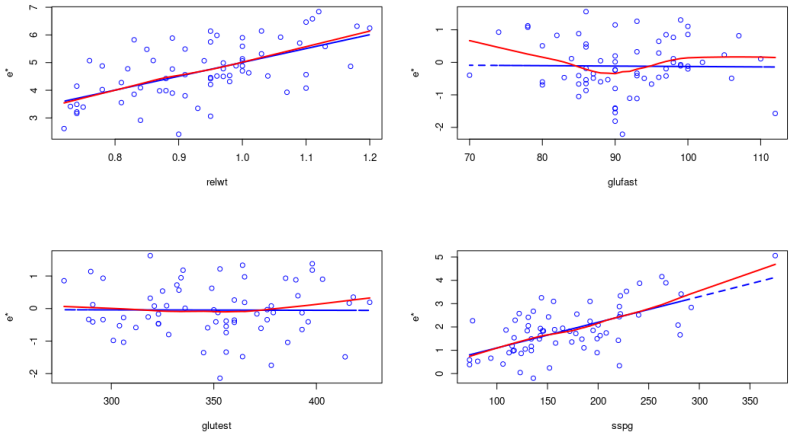


Fig 5.9; Partial Residual Plot

The dotted lines have slope of the corresponding parameter estimate; as the smoothed lines are almost in line with fitted lines, we can conclude the relationship of explanatory variables with response is linear.

Added Variable Plots :

The Added Variable plots provide information about the marginal importance of a predictor variable given the other variables already in the model. The plot for a variable $\boldsymbol{x_j}$ is made by plotting residuals from regressing response on all predictors other than $\boldsymbol{x_j}$ against residuals from regressing $\boldsymbol{x_j}$ on all other predictors (i.e. plot of $e^{(j)} = (I_n - P_j)Y$ against $(I_n - P_j)x_j$ where $P_j = X^{(j)}(X^{(j)'}X^{(j)})^{-1}X^{(j)'}$ and $X^{(j)}$ is a regression matrix excluding variable $\boldsymbol{x_j}$). A strong linear relationship indicates that adding $\boldsymbol{x_j}$ does explain the variability in response that was not explained by other explanatory variables, indicating the importance of adding the variable to our model.

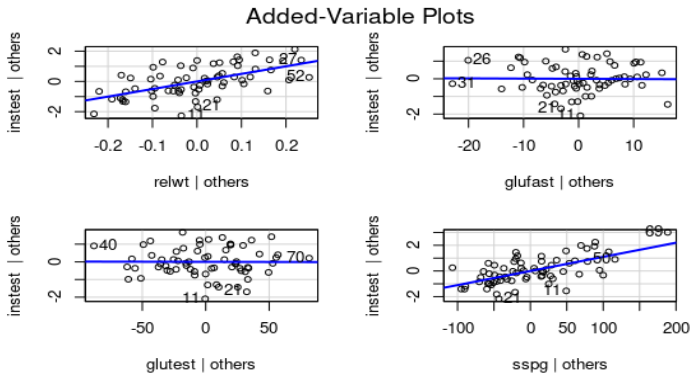


Fig 5.10; Added Variable Plot

As the plots in Figure 5.10 are spread linearly we can safely conclude that the variables chosen should be included in the model but the parameters of glufast and glutest are very close to 0.

5.7. Model Selection

Model selection helps us in deciding the significant variables out of all chosen variables. We used All Possible Regression (APR) to find the best model, in which we considered every possible model and used Mallow's C_p and AIC as our criterias for model selection.

Mallow's C_p :

This criteria is based on minimizing Model error. C_p is calculated as,

$C_p = \frac{RSS_p}{\hat{\sigma}^2} + (2p - n)$, $where\ \hat{\sigma}^2 = \frac{RSS_{k+1}}{n-k-1}$ and k is the number of explanatory variables excluding intercept and p is the number of variables in the subset model chosen.
Expected value of C_p is close to p hence one criteria to select the best subset is to choose a model with minimum value of C_p - p.

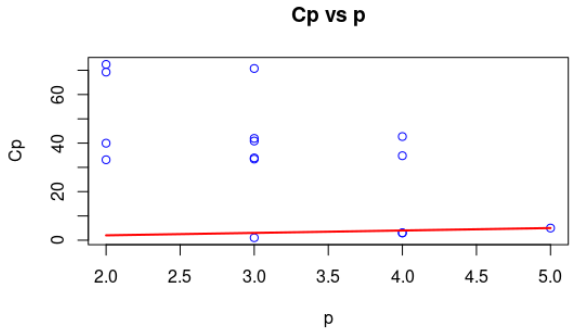


Fig 5.11; Mallow-s C_p vs p plot

The model with minimum C_p - p is model with only relwt and sspg.

	formula	Cp	p	AIC
7	relwt+sspg	1.013937	3	169.5739

AIC :

AIC (Akaike Information Criterion) estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.

$AIC_c = nlog(2\pi\hat{\sigma}^2) + \frac{n(n+p)}{n-p-2}$ can be used to calculate the AIC's for every subset model.
We found the minimum value of AIC (= 169.57) is also for the model with only relwt and sspg.

As the model with only relwt and sspg is the best model according to both AIC and Mallow's C_p criteria, we will choose to go further with this model. Fitting a model gives us an adjusted R^2 of **53.3%** which is an improvement from last. Performing similar analysis as Section 5.3 tells us the model still follows all assumptions of linear regression. But on analyzing outliers again, we notice the presence of 5 more outliers corresponding to patient ID's **11 25 34 69 76**. One possible reason for the more than usual number of outliers is higher amounts of variability in response that is not being explained by the explanatory variables, as was also observed in preliminary analysis.

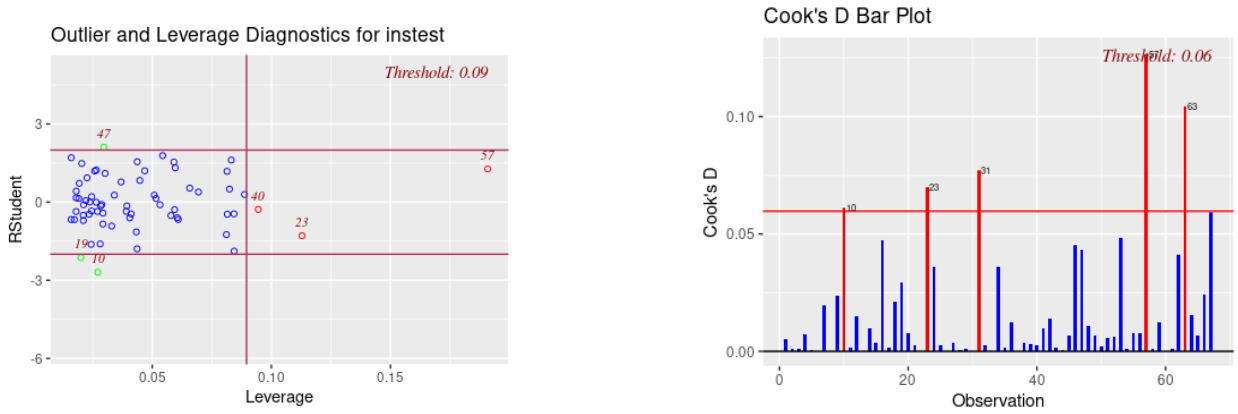


Fig 5.12; Outlier and Leverage Diagnostics (Plotting studentized residuals vs h_i) (Left);
Cook's Distance Plot (Right)

5.8. Final Predictions

After removing these outliers we are ready to fit the final model, which gives us this summary :

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.7222 -0.4469 -0.1052  0.5016  1.6624
Coefficients:
             Estimate     Std. Error     t value Pr(>|t|)
(Intercept)  1.386839     0.802952       1.727   0.0894 .
relwt        5.112417     0.812245       6.294   4.18e-08 ***
sspg         0.011514     0.001913       6.017   1.21e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.7546 on 59 degrees of freedom
Multiple R-squared:  0.5871,  Adjusted R-squared:  0.5731
F-statistic: 41.95 on 2 and 59 DF,  p-value: 4.646e-12
```

We notice the adjusted R^2 = **57.31%** is considerably better than our initial model and as the p-value of the F test is < 0.05, we can reject the null hypothesis of all explanatory variables being insignificant in predicting response. The predictive R^2 is a leave-one-out cross validation measure used to gauge the model fit on each observation when they are not included in the model. Predictive R^2 of our model is **0.54**. This model has relatively lower R^2 implying there can be other variables causing random variations in response along with the variables accounted for in the model.

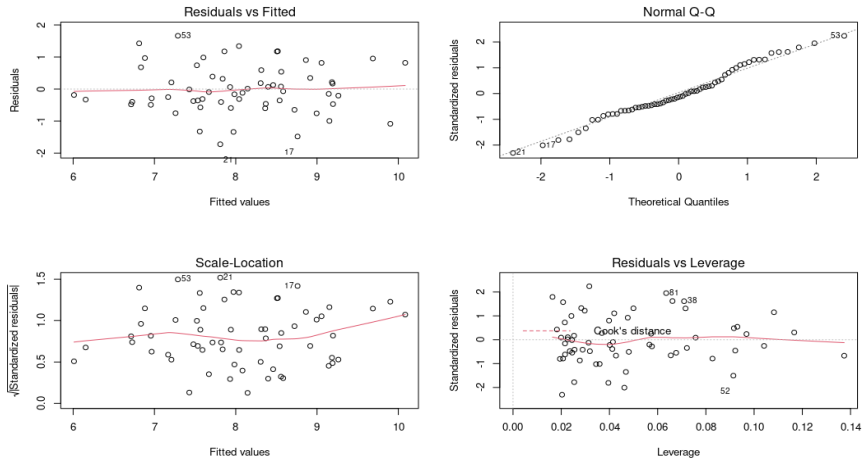


Fig 5.13; Residual Plots 1. Residual vs Fitted plot, 2. Normal QQ plot, 3. Scale Location plot, 4. Cook's Plot

From the Normal QQ plot and p-value of Shapiro Wilk test for residuals = 0.445 (>> 0.05), we can conclude that residuals are normally distributed. In Residuals vs Fitted plot, the residuals are spread implying evenly around the x axis implying homoscedasticity which can also be confirmed by p-value of Breusch Pagan Test = 0.781 (>> 0.05). The VIF for both variables are ≈ 1.009 (<< 4) and condition number is 1.403206 (< 30) implying the data does not have multicollinearity. The p-value for Durbin Watson Test is 0.352 (> 0.05) signifying no autocorrelation. From Fig. 5.14, the partial residuals for both variables are also very much linear, so they are linearly related with response.

As all the assumptions of linear regression are satisfied, we shall use this model for final analysis of Normal Group.

$$\frac{(instest)^{0.22} - 1}{0.22} = 1.3868 + 5.1124\,(relwt) + 0.0115\,(sspg)$$

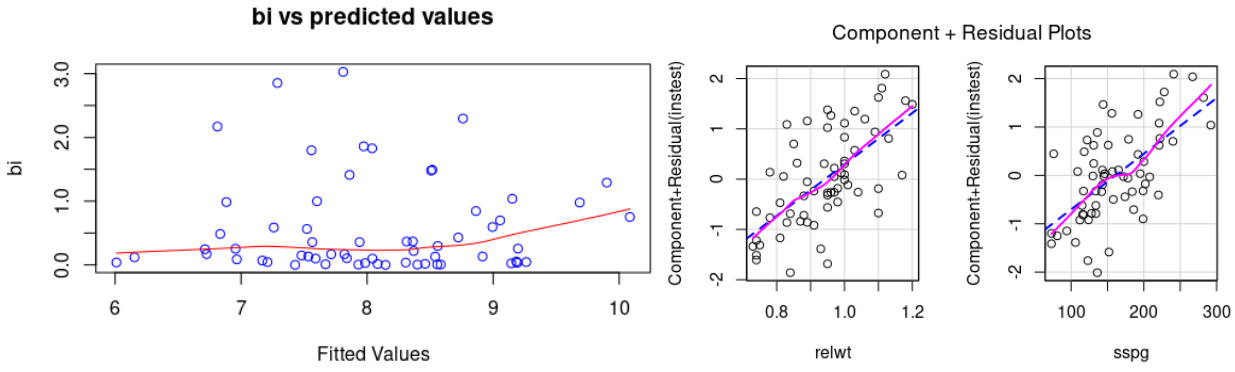


Fig 5.14; bi vs fitted values (Left);
Partial Residual Plot for final model (Right)

6. Fitting Model For Overt Diabetic Patient Group

We have 32 observations corresponding to the Normal group and 6 variables. As we need to predict Insulin levels, we will choose ‘instest’ as our response variable and 'relwt', 'glufast', 'glutest', 'sspg' as explanatory variables.
We will first plot each explanatory variable vs. another to find if there is any multicollinearity between the explanatory variables.

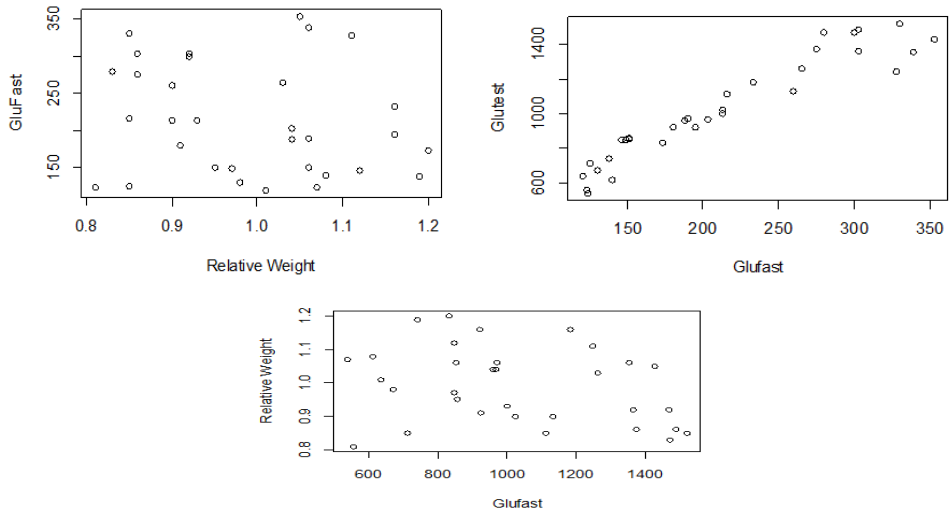


Fig 6.1 shows the plots of one the explanatory variables against each other.

From the plots, it is evident that there seems to be high multicollinearity between Glufast and Glutest. This will be further confirmed by using measures like VIF and Condition Number.

6.1 Distribution of Response

We try to draw a summary of the distribution of the response variable by looking at some basic measures like skewness, kurtosis, etc and doing a test for normality.

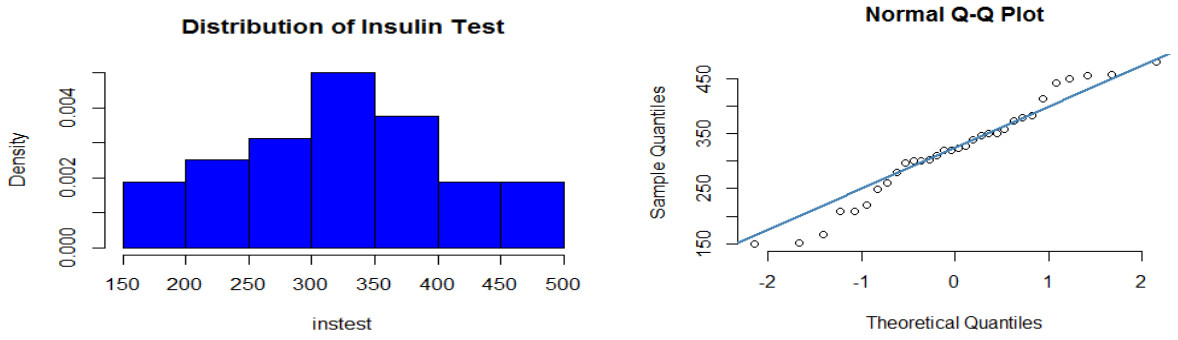


Fig:6.2; Density and Normal QQ plot for Instest

```
Summary :
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  150.0   274.2   322.0   320.9   375.0   480.0
SD : 88.92728
Skewness : -0.1689691
Kurtosis : 2.45469
        Shapiro-Wilk normality test
data:  x
W = 0.96717, p-value = 0.4251
```

From the histogram in Fig 6.2 it seems that Instest is almost symmetrically distributed. From the QQ plot, it is apparent that the data is normally distributed. Also from the Shapiro Wilks test, the p value is 0.4251 (>0.05), therefore we accept the null hypothesis (i.e. the data is normally distributed).

6.2 Initial Linear Model

Fitting a linear model on the data (using instest as the response) using the lm() function gives the following summary:

```
Residuals:
    Min      1Q  Median      3Q     Max
-129.563  -28.392    9.326   36.005   91.570

Coefficients:
                        Estimate Std.  Error t value  Pr(>|t|)
(Intercept)     -350.18201  120.94667  -2.895  0.007414 **
relwt            349.33166   94.51418   3.696  0.000983 ***
glufast            0.09743    0.44827   0.217  0.829570
glutest            0.26536    0.12047   2.203  0.036327 *
sspg               0.28750    0.14471   1.987  0.057192 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 54.89 on 27 degrees of freedom
Multiple R-squared:  0.6682,  Adjusted R-squared:  0.619
F-statistic: 13.59 on 4 and 27 DF,  p-value: 3.41e-06
```

For this model, the adjusted R^2 value comes out as **0.619** which suggests this model is fairly good in explaining the variability in instest. As the p-value of the F test is < 0.05 we can reject the null hypothesis of all explanatory variables being insignificant in predicting response (i.e each coefficient $\beta_i = 0$ corresponding to explanatory variables).

6.3 Checking for Multicollinearity

We use the Variance Inflation Factors (VIF's) as a measure for multicollinearity between the explanatory variables. As a thumb rule, we will say there is significant multicollinearity if the VIF value is >4.

The VIF values are as follows:

relwt	glufast	glutest	sspg
1.190872	11.379238	13.390247	1.882148

We see that the VIF values for glufast and glutest are significantly high, indicating high multicollinearity. This is in accordance with what we had suspected after observing Fig 6.1.
So we need some methods to deal with multicollinearity. This may include:

i> Disregard the variables until the data is not collinear. However for this we need to ensure that the deleted variables do not have a significant relationship with the response.

ii> Use a Ridge Regression.

We will use both methods and then compare to find the better model.

Ridge Regression

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.
In ordinary least squares, the regression coefficients are estimated using the formula

$$\hat{\underline{\mathbf{B}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{\mathbf{Y}}$$

Here the variables are standardized. Therefore X'X = R, where R is the correlation matrix of independent variables.
Ridge regression proceeds by adding a small value, k, to the diagonal elements of the correlation matrix.
The estimate of the coefficients of a ridge regression is given by:

$$\tilde{\underline{\mathbf{B}}} = (\mathbf{R} + k\mathbf{I})^{-1}\mathbf{X}'\underline{\mathbf{Y}}$$

where k is a positive quantity less than 1.
The value of k is chosen so that the MSE is minimized.
We now fit a Ridge regression model for our data using R. The summary of the fitted model is as follows:

Coefficients: for Ridge parameter K= 0.03

	Estimate	Estimate (Sc)	StdErr (Sc)	t-value (Sc)	Pr(>|t|)
Intercept	-2.9762e+02	-3.8444e+05	1.2306e+05	-3.1240	0.0042 **
relwt	3.2485e+02	2.0587e+02	5.5312e+01	3.7220	0.0009 ***
glufast	2.9510e-01	1.2188e+02	1.1011e+02	1.1069	0.2779
glutest	2.0120e-01	3.3545e+02	1.1735e+02	2.8585	0.0080 **
sspg	2.4530e-01	1.2762e+02	6.8235e+01	1.8704	0.0721 .

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

Ridge Summary

R2	adj-R2	DF ridge	F	AIC	BIC
0.61880	0.57790	3.47349	14.02305	258.22313	374.21790

Ridge minimum MSE= 52444.14 at K= 0.03
P-value for F-test (3.47349 , 28.18126) = 4.239818e-06

(The ridge regression model is run for all values of k ranging from 0.01 to 1 incrementing by 0.01. The minimum MSE is attained at 0.03).
The VIF values are as follows:

	relwt	glufast	glutest	sspg
k=0.03	1.04763	4.15145	4.7156	1.59432

There seems to have been a significant improvement in terms of fixing the multicollinearity problem (significantly lower VIF than LS model). However, the adjusted R^2 value drops significantly to 0.57790, indicating this model can be bettered.

Excluding Variables

In our initial linear model, after observing the p values, it was apparent that the glufast variable did not have a significant relationship with the response.
We will try using all possible regression models and find the combination of explanatory variables used which gives the lowest Mallow’s C_p - p value, here p is the number of regression coefficients in the model used.

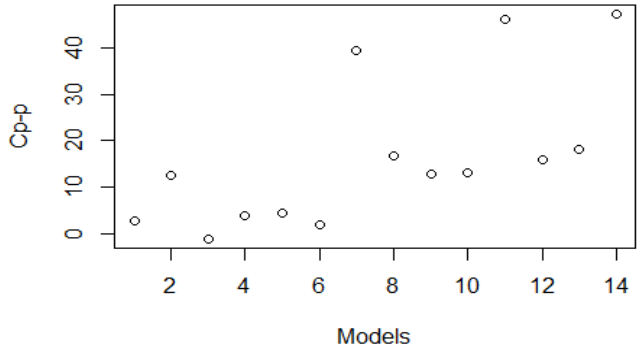


Fig 6.3; Cp-p values vs All possible regression models

We see that the lowest value of C_p - p is for model 3. Model3 is the model when we discard glufast from the explanatory variables of the full model. In other words, model 3 is the regression model of instest on relwt, glutest and sspg.
We also use AICc to find out the best model.

$AIC_c = n\log(2\pi\hat{\sigma}) + n(n+p)/(n-p-2)$

The model with the lowest AIC_c values is the best model. We tabulate the AIC_c values for all possible models using R.

Model selection based on AICs:

	K	AICc	Delta_AICc	AICcWt	Cum.Wt	LL
Mod3	5	354.08	0.00	0.57	0.57	-170.88
Mod6	4	355.76	1.68	0.24	0.81	-173.14
Mod5	4	358.17	4.09	0.07	0.88	-174.34
Mod1	5	358.39	4.31	0.07	0.95	-173.04
Mod4	5	359.31	5.23	0.04	0.99	-173.50
Mod9	4	365.29	11.21	0.00	0.99	-177.91
Mod10	4	365.53	11.45	0.00	1.00	-178.02
Mod12	3	365.54	11.46	0.00	1.00	-179.34
Mod2	5	367.12	13.05	0.00	1.00	-177.41
Mod13	3	367.18	13.10	0.00	1.00	-180.16
Mod8	4	368.10	14.03	0.00	1.00	-179.31
Mod7	4	380.86	26.78	0.00	1.00	-185.69
Mod11	3	381.72	27.64	0.00	1.00	-187.43
Mod14	3	382.14	28.07	0.00	1.00	-187.64

Here also we see that for model 3 the AIC_c value is the smallest indicating that it is the best model.

We will now draw a summary of this model using R.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-360.17668	109.94429	-3.276	0.002807 **
relwt	354.87556	89.44557	3.968	0.000458 ***
glutest	0.28964	0.04436	6.529	4.46e-07 ***
sspg	0.29093	0.14138	2.058	0.049021 *

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

Residual standard error: 53.95 on 28 degrees of freedom
Multiple R-squared: 0.6676, Adjusted R-squared: 0.632
F-statistic: 18.75 on 3 and 28 DF, p-value: 7.23e-07

The VIF values for this model are as follows:

relwt	glutest	sspg
1.104140	1.879600	1.859675

The VIF values are quite low (in fact close to 1), indicating there is no multicollinearity in the data now. Also now all the explanatory variables seem significant i.e. $\boldsymbol{\beta}_i \neq 0$
Also there is an improvement in the adjusted R^2 value further indicating that this is the right model.

6.4 Analyzing Residuals

Checking Homoscedasticity

We will use the following 2 plots to check for heteroscedasticity:

i> plot of b_i vs $\hat{y}_i$: if this plot is roughly wedge shaped, we may say there is presence of heteroscedasticity. A smoother passed through the plot may reveal the relationship between the means and the variances.

ii> plot of e_i vs $\hat{y}_i$: a fan-shaped pattern indicates variances increasing with the mean

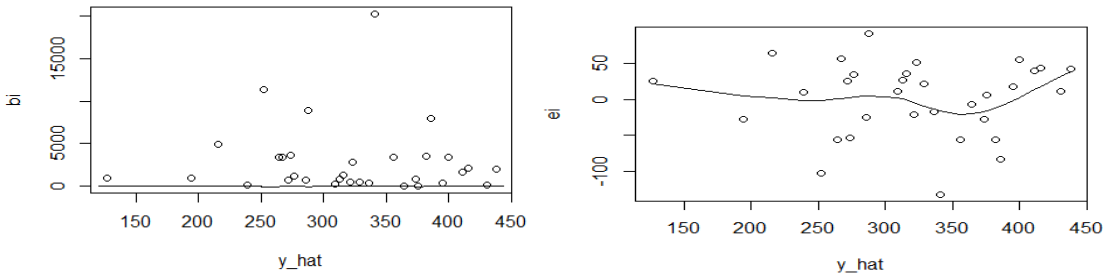


Fig 6.4: plot of b_i vs $\hat{y}_i$ (left) and plot of e_i vs $\hat{y}_i$ (right)

From Fig 6.4, the plot on the left shows no wedge shaped pattern. This is further confirmed by the fitted smoother which is a line almost parallel to the X axis. And the plot on the right shows no fan shaped pattern.

To formally test that the errors are homoscedastic, we use the Breusch Pagan test.

Studentized Breusch-Pagan test

data: mod3

BP = 0.25676, df = 3, p-value = 0.9679

We see that the p value is 0.9679 (>>0.05). Thus we accept the null hypothesis that the errors are homoscedastic.

Checking Autocorrelation

The Durbin Warson test serves as a test for the presence of autocorrelation in the residuals. The null hypothesis for the Durbin Watson test is there is no autocorrelation against the alternative hypothesis that there is presence of autocorrelation.

We run the Durbin Watson test using R and the results are as follows:

lag	Autocorrelation	D-W Statistic	p-value
1	0.02234218	1.916141	0.762

Alternative hypothesis: rho != 0

The p value =0.762 (>0.05) indicating that we can accept the null hypothesis (i.e. there is no autocorrelation in the residuals).

Checking Normality

We check the assumption that the residuals are normally distributed using a QQ plot.

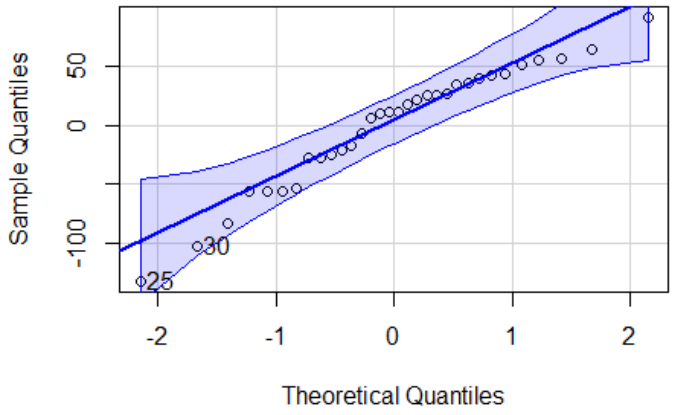


Fig 6.5: QQ plot of Sample Quantiles vs Theoretical Quantiles

From the QQ plot, it seems that the residuals are normally distributed but we will test this formally using the Shapiro Wilks test.

Shapiro-Wilk normality test
data: res
W = 0.95453, p-value = 0.1936

From the Shapiro Wilks test, we observe that the p value=0.1936 (>0.05). Thus we accept the null hypothesis i.e. the residuals are normally distributed.

6.5 Checking for Curvature

To check for the presence of curvature, we use Partial Residual Plots (PRP's). The PRP's are essentially a plot of $e_i^*=e_i+\hat{\beta}_j x_{ij}$ vs x_{ij} where x_j's are explanatory variables. This plot approximately is equivalent to plotting $\hat{\beta}_j g(x_{ij})$ vs x_{ij} where g is the function whose curvature we

are interested in studying. If the plot reveals the shape of g as more or less a straight line, then we can conclude there is no curvature. The shape of the plot is revealed after fitting a smoother through it.

The PRP's obtained from R are as follows:

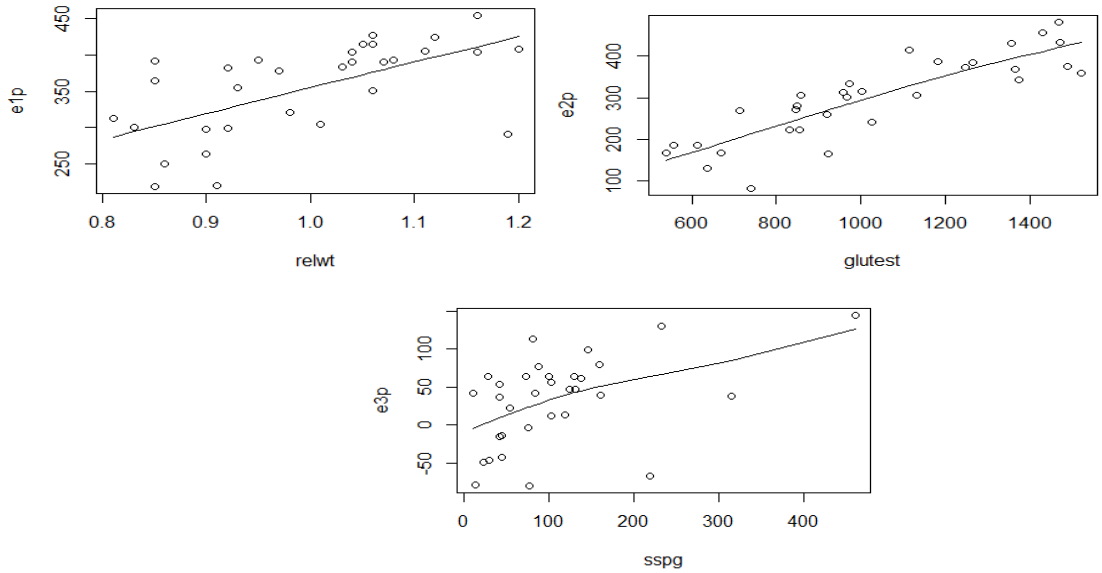


Fig 6.6: Partial Residual Plots

Fig 6.6 reveals that the smoothers in all the PRP's are more or less straight lines. This shows that there is no significant presence of curvature.

6.6 Dealing with Outliers

First we would need to check the externally studentized residuals (t_i) to detect the presence of any outlier. We follow the following rule:

If $|t_i|>2$ for a particular point then the i'th point may be treated as an outlier.

The t_i values are calculated using R.

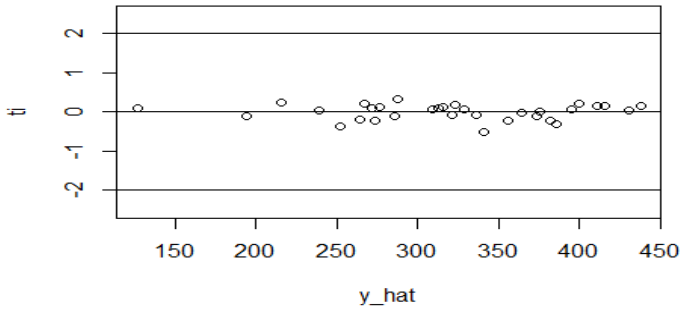


Fig. 6.7: t_i vs $\hat{y}_i$ plot

It is evident from Fig 6.7 that none of the $|t_i|$ values are greater than 2.

However this works well only in the absence of high leverage points. In the presence of high leverage points, we cannot expect it to reveal the presence of an outlier.
Now we would want to check if there are any high leverage points. For checking high leverage points ,we use the rule $h_i > 2\frac{p}{n}$ where h_i are the hat matrix diagonals, p is the number of explanatory variables including intercept and n is the number of data points.
The calculated value of 2p/n in this case is 0.25. We use R to find that the 19th and the 22nd data points are high leverage points.
Influential points are the points that actually have significant impact on the regression. To find the influential point(s) we use the measures DFFITS and Cook's Distance.

- DFFITS (Difference in Fitted values) : Points having $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$, where p is number of explanatory including intercept
- Cook's Distance (Distance between parameter vector β and parameter vector removing i'th point $\beta(i)$) : Points having $D_i > F^{0.1}_{p, n-p}$

We use R to calculate the DFFITS and the Cook's Distance and hence find the influential points.

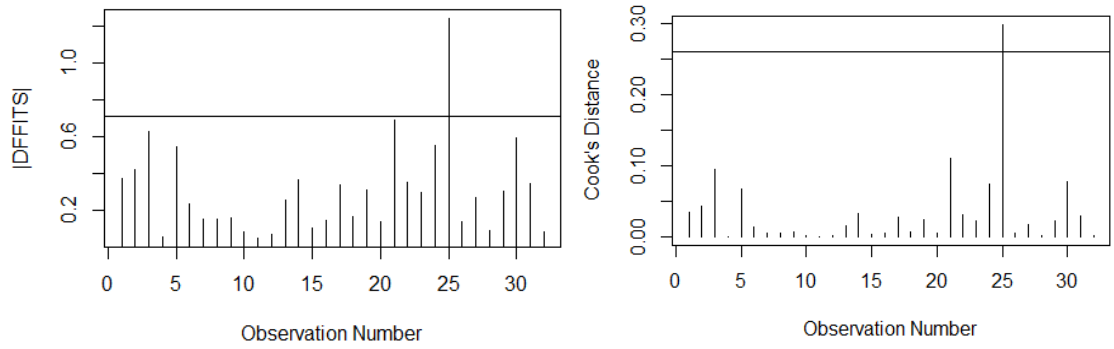


Fig 6.8: Plot of |DFFITS| vs Observation Number (Left);
plot of Cook's Distance and Observation Number (Right)

From both the plot, it is apparent that the **25th** observation is an influential point. We eliminate this point and fit the regression model again.

6.7 Final Predictions

After eliminating the influential point and refitting the model, we get the following summary:

Coefficients:

	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	-426.38367	99.55331	-4.283	0.000209 ***
relwt_	421.97947	82.09182	5.140	2.09e-05 ***
glutest_	0.28970	0.03916	7.398	5.87e-08 ***
sspg_	0.33154	0.12554	2.641	0.013579 *

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 47.62 on 27 degrees of freedom
Multiple R-squared:  0.7363,  Adjusted R-squared:  0.707
F-statistic: 25.13 on 3 and 27 DF,  p-value: 5.661e-08
```

We see that the adjusted R^2 value has increased to **0.707** (from 0.632 in the model including the influential point), which is a significant improvement. Also now all the explanatory variables seem to have a significant relationship with the response (p value<0.05). The predictive R^2 of the model is **0.657**.

We rerun the other diagnostics too to check if the assumptions are fulfilled.

```
The VIF values are as follows:
 relwt         glutest         sspg
1.072655       1.822359       1.794672
```

All the VIF values are less than 4, indicating no multicollinearity.

We plot the residuals vs the fitted values to see if there is any visible pattern in the plot. This will perform a visual test for the presence of any heteroscedasticity, autocorrelation, etc.

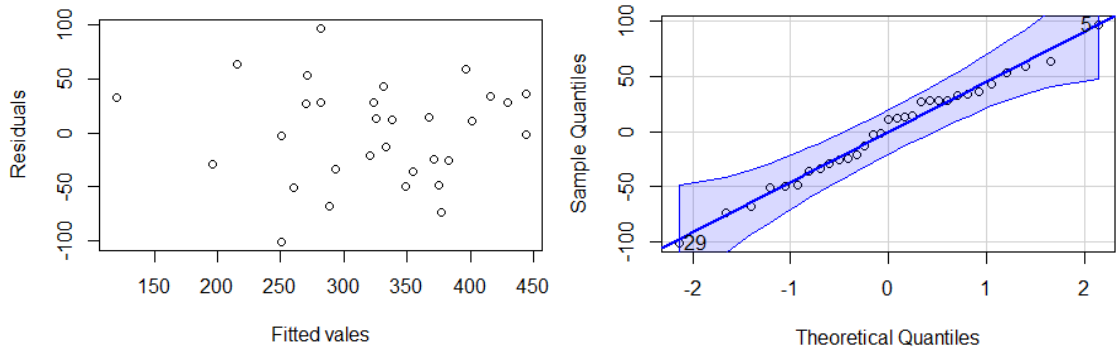


Fig 6.9: Residuals vs Fitted values plot (Left);
QQ plot of Sample Quantiles vs Theoretical Quantiles (Right)

From the residuals vs fitted values plot, it seems apparent that there is no presence of any heteroscedasticity or autocorrelation in the residuals. However, we will also perform formal tests to ensure these.

Performing the Shapiro Wilks test the value comes out as 0.9137, strongly indicating the residuals are normally distributed. This is further ensured by the use of a QQ plot in Figure 6.9.

The Breusch Pagan test for testing heteroscedasticity gives a p value of 0.1031 (>0.05) indicating there is no heteroscedasticity in the residuals. For checking autocorrelation the Durbin Watson test statistic gives a p value of 0.884 (>0.05) indicating no presence of autocorrelation.

We use the partial residual plots to detect the presence of any curvature.

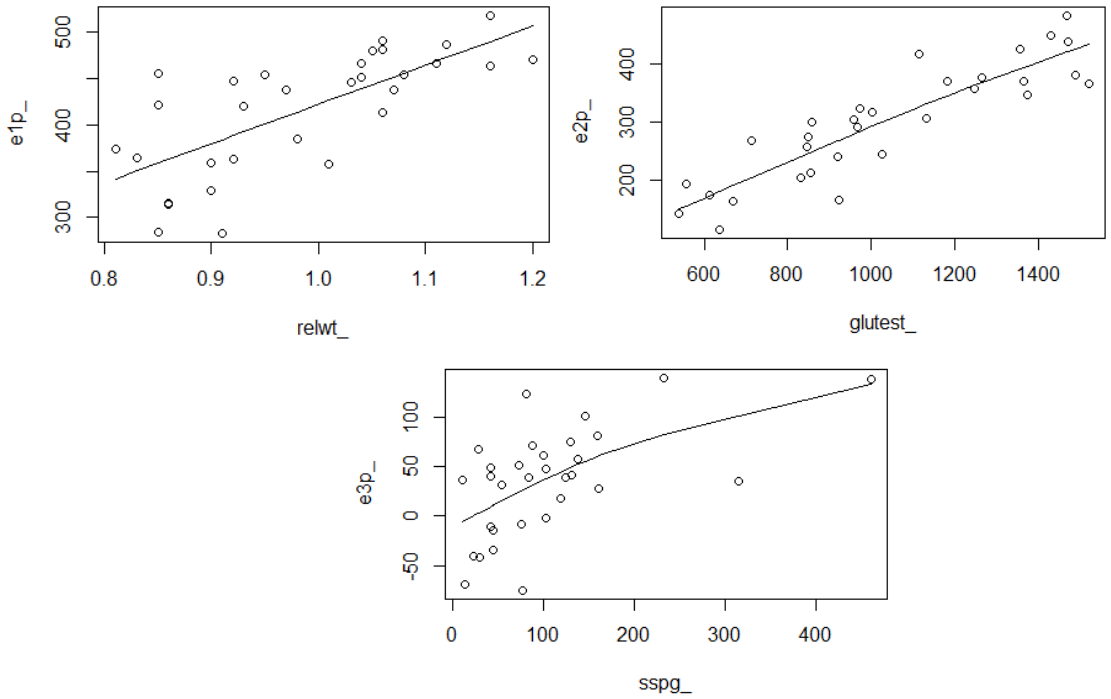


Fig 6.10: The Partial Residual plots

Smoothers fit through the partial residual plots are almost straight lines, indicating no presence of curvature.

We use Cook's distance and DFFITS to check for any influential points.

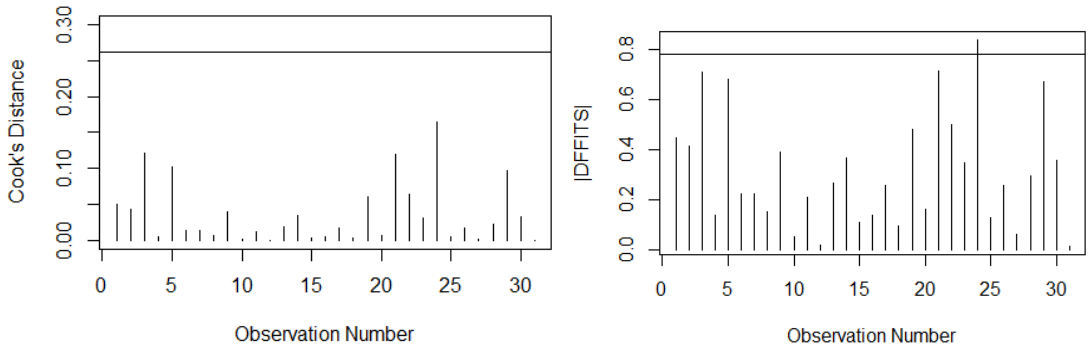


Fig 6.11: Cook's Distance vs Observation Number (Left);
|DFFITS| vs Observation Number (Right)

Cook's Distance indicates the presence of no influential points, but DFFITS indicates the 24th observation is influential. However since the |DFFITS| for the 24th observation crosses the threshold value only by a small amount, we will not consider the 24th point as influential.

Hence this model satisfies all the assumptions and we will therefore use this model for our final predictions.

$$(instest) = -426.3836 + 421.9794\,(relwt) + 0.2897(glutest) + 0.3315\,(sspg)$$

The fitted values are plotted against the actual values.

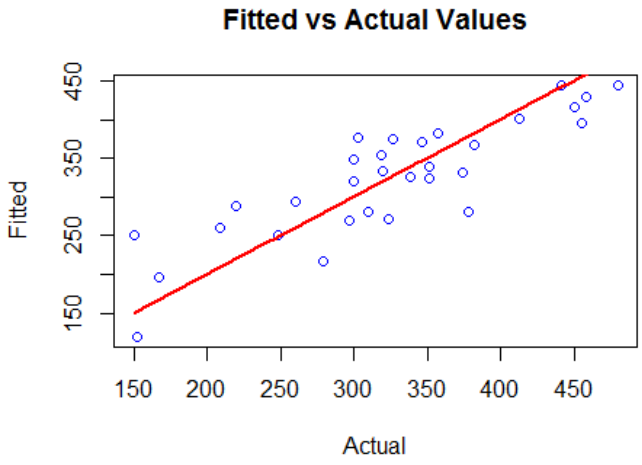


Fig 6.12: Fitted values vs Actual values

From Fig 6.12 it seems that the points are along the y=x line indicating that the fit is good.

7. Fitting Model For Chemically Diabetic Patient Group

We have 36 observations corresponding to the Chemically Diabetic group and 6 variables. As we need to predict Insulin levels, we will choose ‘instest’ as our response variable and 'relwt', 'glufast', 'glutest', 'sspg' as explanatory variables. We will first observe the pair plots for the Chemically Diabetic group to visualize presence of any linearity with response and any multicollinearity within explanatory variables.

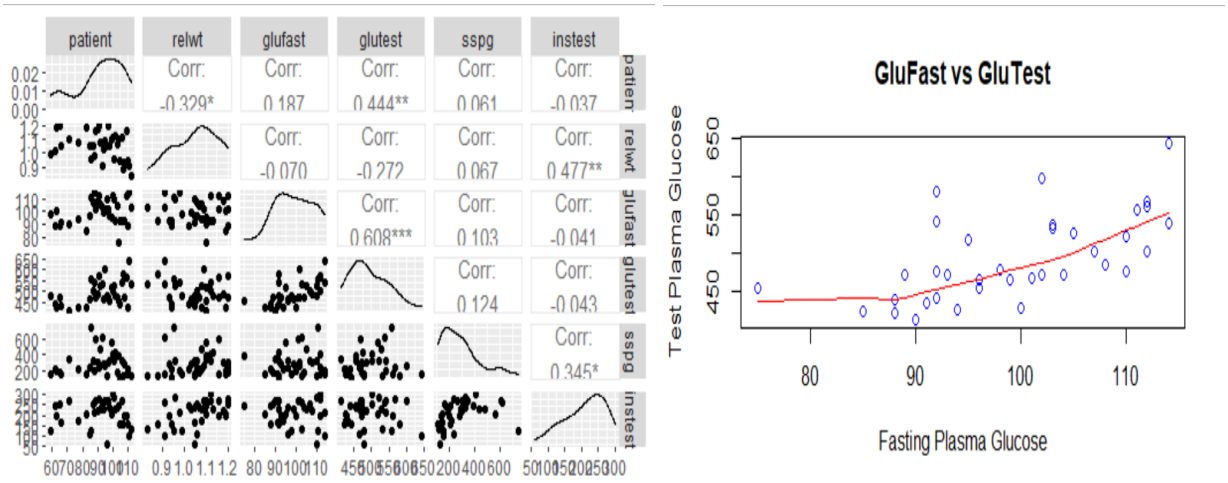


Fig 7.1; Pair plot of all variable except patient ID for Group 2 (Left) ;
GluFast vs GluTest for Group 2 (Right)

From Figure 7.1, we notice presence of good correlation of variables relwt and sspg with response instest. Also the relationship between glufast and glutest is as significant as in Figure 4.1. This hints towards multicollinearity which we will check later using Variance Inflation Factors (VIF’s) and Condition Number.

7.1. Distribution of Response

From Figure 7.2, we can conclude that our response variable ‘instest’ is normally distributed but rather skewed as is also suggested by following metrics of ‘instest’ :

Summary :

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   60.0   164.5   223.0   209.0   258.8   300.0
SD : 59.95927
Skewness : -0.6284339
Kurtosis : 2.530267
        Shapiro-Wilk normality test
data:  x
W = 0.9458, p-value = 0.07716
```

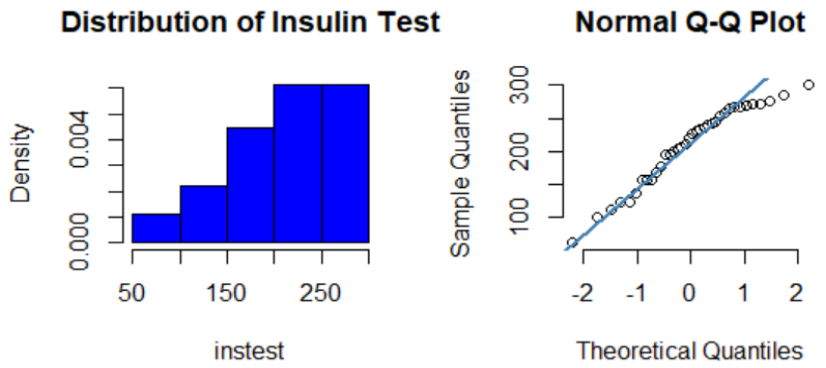


Fig 7.2; Density and Normal QQ plot for instest

As the p-value = 0.07716 > 0.05, we fail to reject the null hypothesis that the population is normally distributed. Hence we can say our response variable is normally distributed. We notice from QQ-plot that due to the presence of some outliers the p-value is relatively smaller.

7.2. Initial Linear Model

Fitting a Linear model using the lm function in R gives us the following model summary:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-119.22548	147.51117	-0.808	0.42511
relwt	285.21631	91.46641	3.118	0.00391 **
glufast	-0.69697	1.17423	-0.594	0.55712
glutest	0.12635	0.20953	0.603	0.55089
sspg	0.11757	0.05637	2.085	0.04535 *

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

Residual standard error: 51.93 on 31 degrees of freedom
Multiple R-squared: 0.3356, Adjusted R-squared: 0.2499
F-statistic: 3.915 on 4 and 31 DF, p-value: 0.01096

As suggested by adjusted R^2, this linear model does not explain variation in 'instest' very well, but as the p-value of the F test is < 0.05 we can reject the null hypothesis of all explanatory variables being insignificant in predicting response (i.e each coefficient $\beta_i = 0$ corresponding to explanatory variables).

7.3. Analyzing Residuals

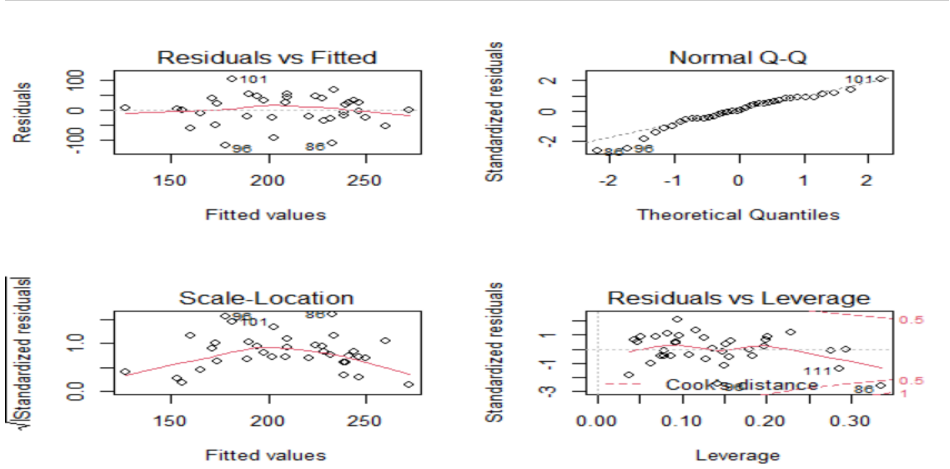


Fig 7.3; Residual Plots 1. Residual vs Fitted plot, 2. Normal QQ plot, 3. Scale Location plot, 4. Cook's Plot

From the Normal QQ plot we observe that the distribution of residuals is normal, as is also indicated by following metrics :

Summary :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-118.433	-24.288	1.364	0.000	34.582	103.920

SD : 48.87245
Skewness : -0.5382696
Kurtosis : 3.215754

Shapiro-Wilk normality test

data: x
W = 0.96552, p-value = 0.3161

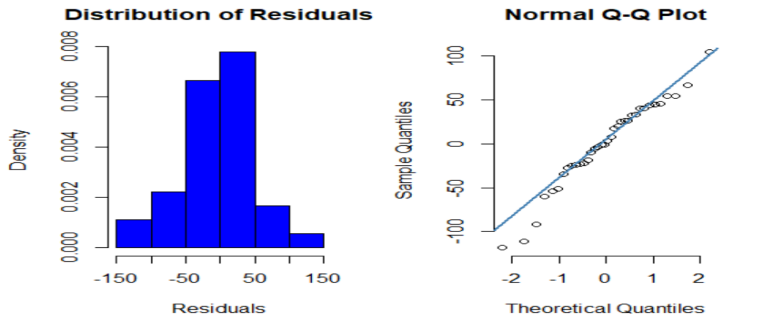


Fig 7.4; Density and Normal QQ plot for residuals.

As from Figure 7.4 it can be easily seen & the p-value of Shapiro-Wilk test is > 0.05 , we conclude that we fail to reject the null hypothesis of residuals being normally distributed. Hence the 5th assumption from Section 2 is satisfied.

Checking Homoscedasticity

From the residual vs fitted values plot, in Figure 7.3, we observe there is no significant increase in spread of residuals with increase in fitted values. For graphically checking for homoscedasticity we can plot b_i vs fitted values.

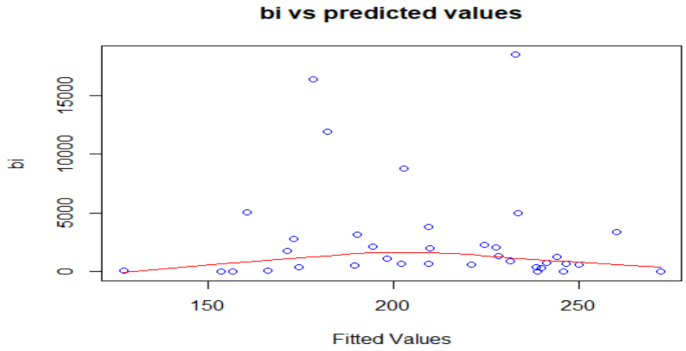


Fig 7.5; bi vs fitted values plot

As, in Figure 7.5, the plot does not have significant wedge shape, so we cannot conclude the presence of heteroscedasticity.

The Breusch Pagan Test.

Statistic	p.value	parameter	method	alternative
9.47	0.0503	4	Koenker (studentised)	greater

As the p-value of the Breusch Pagan test is significant (> 0.05), we fail to reject the null hypothesis of presence of homoscedasticity. Hence the 4th assumption from Section 2 is also satisfied.

Checking Multicollinearity

We initially observed no significant correlation between the explanatory variables for group 2 except for glufast & glutest. A result of multicollinearity is higher variance of parameter estimates.

The **Variance Inflation Factors**(VIF) of variables in our model for group 2 are :

relwt	glufast	glutest	sspg
1.108387	1.611430	1.749634	1.027458

The values of VIF greater than 4 need to be handled carefully but in our case the values are fairly close to 1. Also the Condition Number is 2.197 (< 30). Hence we can conclude the presence of no multicollinearity in this model. Hence the 3rd assumption from Section 2 is also satisfied.

Checking Autocorrelation

The Durbin-Watson statistic states that null hypothesis: rho = 0.

lag	Autocorrelation	D-W Statistic	p-value
1	-0.05315425	2.074481	0.846

Alternative hypothesis: rho != 0

As p-value is significant (> 0.05), we fail to reject the null hypothesis of no serial correlation. Hence the 2nd assumption from Section 2 is also satisfied.

7.4. Dealing with Outliers

Same measures as used earlier, in Section 5.5 & 6.6, are used here.

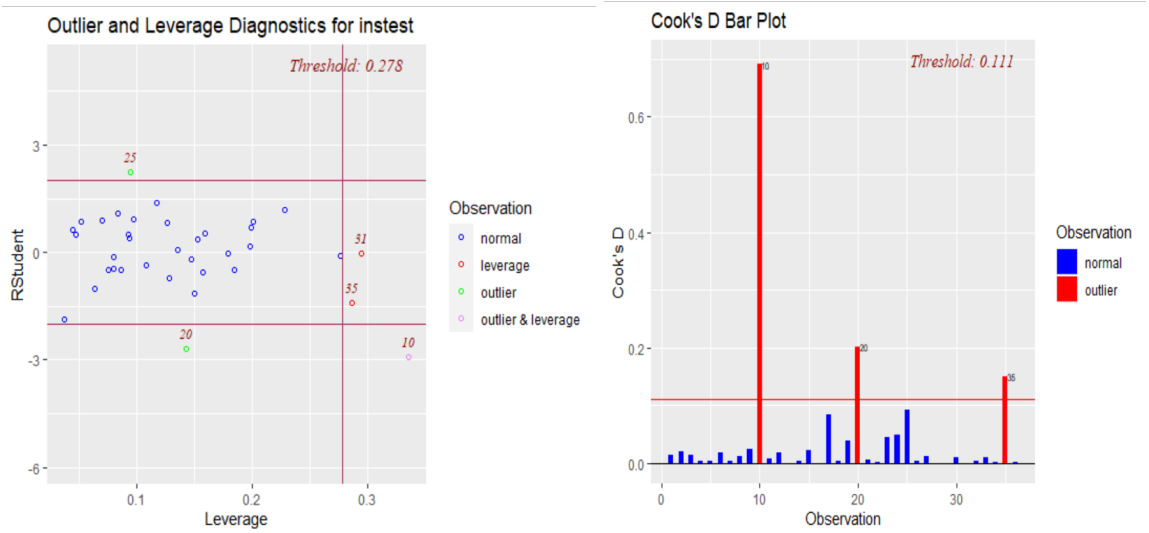


Fig 7.6; Outlier and Leverage Diagnostics (Plotting studentized residuals vs h_i) (Left); Cook's Distance Plot (Right)

In Fig. 7.6, we find that point 10(Patient ID: **86**) is both an outlier & high leverage point. Similarly Patient ID: **96** is also an outlier & high leverage point. We choose to remove points classified as outliers by more than 2 of the influence measures. On fitting the regression model after removing these points, the adjusted R^2 increases considerably to **39.78%** and again none of the assumptions of linear regression are violated when checked similar to Section 7.3. We also noticed the p-value for Breusch-Pagan Test is now 0.876 (0.0503 previously) and the p-value for the F-test of the model is 0.0007697 (0.01096 previously). Hence after removing outliers our conclusions are following more strictly.

7.5. Checking for Curvature

Same plots as discussed in Section 5.6 are drawn.

Partial Residual Plots :

This plot can reveal the shape of the relationship between $\boldsymbol{x_j}$ and the response variable.

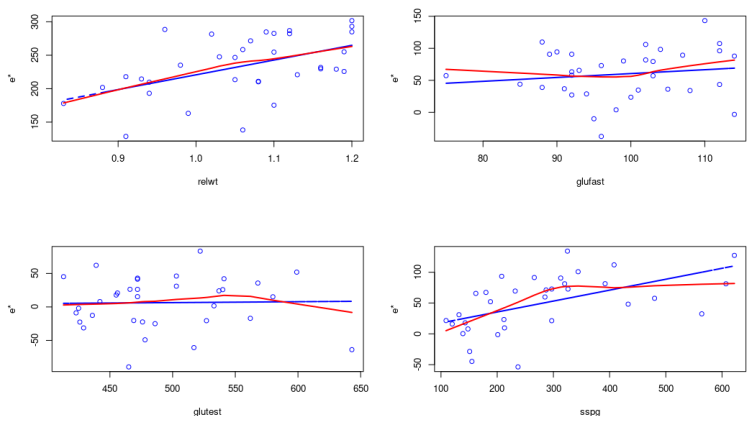


Fig. 7.6; Partial Residuals Plot

The dotted blue lines have slope of the corresponding parameter estimate, as the smoothed lines are almost in line with fitted lines, we can conclude the relationship of explanatory variables with response is linear except for the column sspg where a curvature is apparent.

Added Variable Plots :

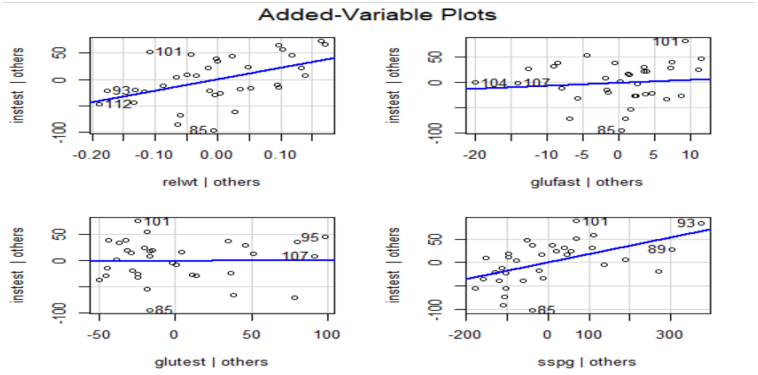


Fig.7.7; Added Variables Plot

As the residual points in plots of Figure 7.7 are spread linearly except sspg, we can conclude that the variables chosen should be kept in the model but the parameters of glufast and glutest are very close to 0 and some transformation is needed for sspg to deal with curvature.

Transforming the variable "sspg"

By visualizing the 4th plot in Figure 7.6, we can say that a logarithmic or square root transformation is required. We fitted two different models using these two transformations & checked whether all the assumptions are satisfied or not & compared which transformations is better. We found out logarithmic transformation is better & we fitted our next model.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-376.12085	127.83156	-2.942	0.006347 **
relwt	197.85291	73.24770	2.701	0.011419 *
glufast	0.51492	0.95204	0.541	0.592735
glutest	0.02922	0.16432	0.178	0.860124
sspg	57.46925 1	5.11513	3.802	0.000683 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.86 on 29 degrees of freedom
Multiple R-squared: 0.5122, Adjusted R-squared: 0.4449
F-statistic: 7.613 on 4 and 29 DF, p-value: 0.0002542

On fitting the regression model after removing these points, the adjusted R^2 increases considerably to **44.49%** and again none of the assumptions of linear regression are violated when checked similar to Section 7.3.

7.6. Model Selection

Model selection helps us in deciding the significant variables out of all chosen variables. We will use Mallow's C_p and AIC as our criteria for checking significant variables, same as discussed in Section 5.7.

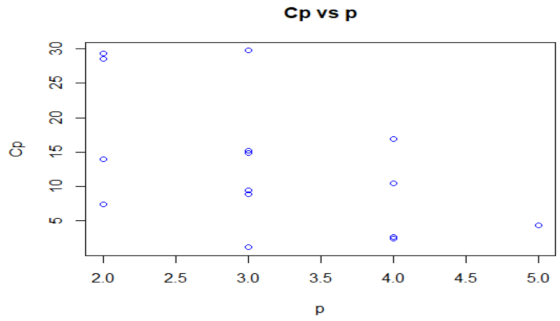


Fig. 7.8; Mallow-s C_p vs p plot

The model with minimum C_p - p is model with only relwt and sspg.

	formula	Cp	p	AIC	diff
7	relwt+log(sspg)	1.494676	3	350.2523	-1.505324

AIC:

We found the minimum value of AIC (= 350.25) is also for the model with only relwt and sspg.

As the model with only relwt and sspg is the best model according to both AIC and Mallow's C_p criteria, we will choose to go further with this model. Fitting a model gives us an adjusted R^2 of **46.77%** which is an improvement from last. Performing similar analysis as Section 7.3 tells us the model still follows all assumptions of linear regression. But on analyzing outliers again, we notice the presence of 2 outliers corresponding to patient IDs **85** &**101**.

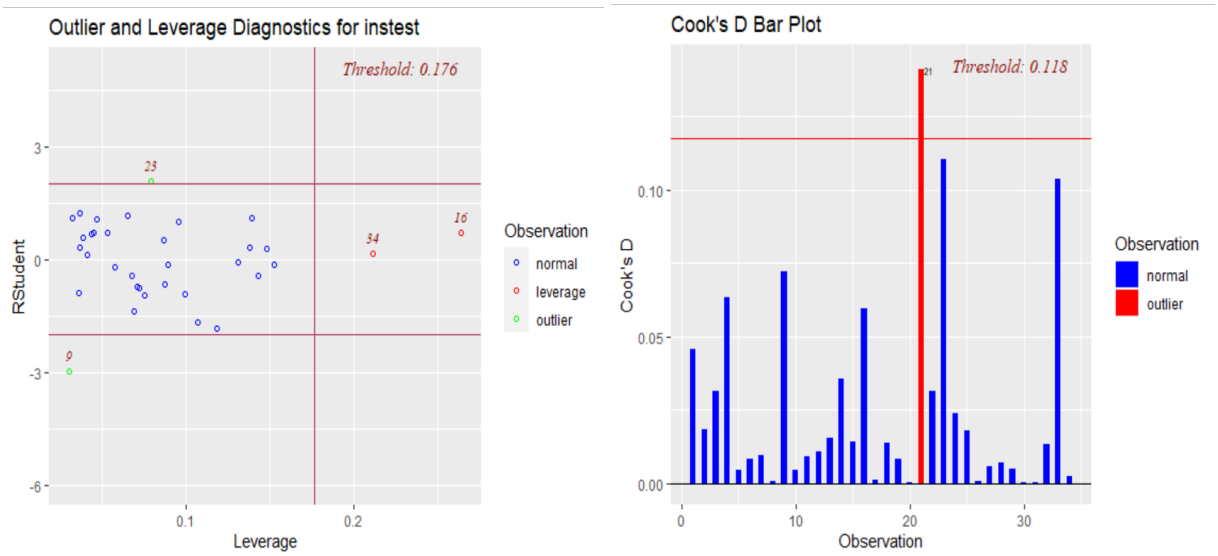


Fig 7.9; Outlier and Leverage Diagnostics (Plotting studentized residuals vs h_i) (Left); Cook's Distance Plot (Right)

7.7. Final Predictions

After removing these outliers we are ready to fit the final model, which gives us this summary :

Residuals:

Min	1Q	Median	3Q	Max
-62.923	-23.642	6.973	26.990	45.694

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-309.81	79.31	-3.906	0.000516 ***
relwt	220.04	58.81	3.741	0.000803 ***
log(sspg)	53.15	12.53	4.243	0.000206 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.7 on 29 degrees of freedom
Multiple R-squared: 0.6048, Adjusted R-squared: 0.5775
F-statistic: 22.19 on 2 and 29 DF, p-value: 1.427e-06

We notice the adjusted R^2 is considerably better(i.e. **57.75%**) than our initial model(**24.99%**) and as the p-value of the F test is < 0.01, we can reject the null hypothesis of all explanatory

variables being insignificant in predicting response. The predictive R^2 of the model is **52.051%**. This model has relatively lower R^2 implying there can be other variables causing random variations in response along with the variables accounted for in the model.

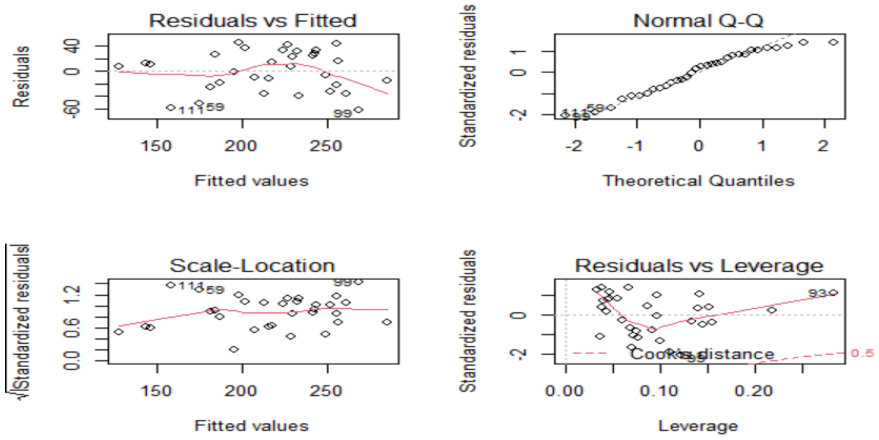


Fig 7.10; Residual Plots 1. Residual vs Fitted plot, 2. Normal QQ plot, 3. Scale Location plot, 4. Cook's Plot

From the Normal QQ plot and p-value of Shapiro Wilk test for residuals (= 0.1408 >> 0.05), we can conclude that residuals are normally distributed. In Residuals vs Fitted plot, the residuals are spread implying evenly around the x axis implying homoscedasticity which can also be confirmed by p-value of Breusch Pagan Test (= 0.833 > 0.05). The VIF for both variables are $\approx$ 1. 08 (<< 4) and condition number is 2.2073(<< 30) implying the data does not have multicollinearity. The p-value for Durbin Watson Test is 0.06 (> 0.01) signifying no autocorrelation. From Figure 7.11, the partial residuals for both variables are also very much linear, so they are linearly related with response. Hence we shall use this model for final analysis of the Chemically Diabetic Group.

$$(instest) = -309.81 + 220.04\,(relwt) + 53.15\,log(sspg)$$

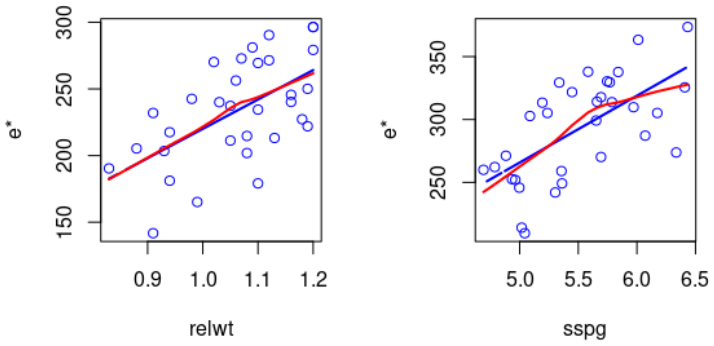


Fig 7.11; Partial Residual Plot for final model

8. CONCLUSION

First we shall compare the combined predictions made by our three final models against the actual values to compare the performance of predictions.

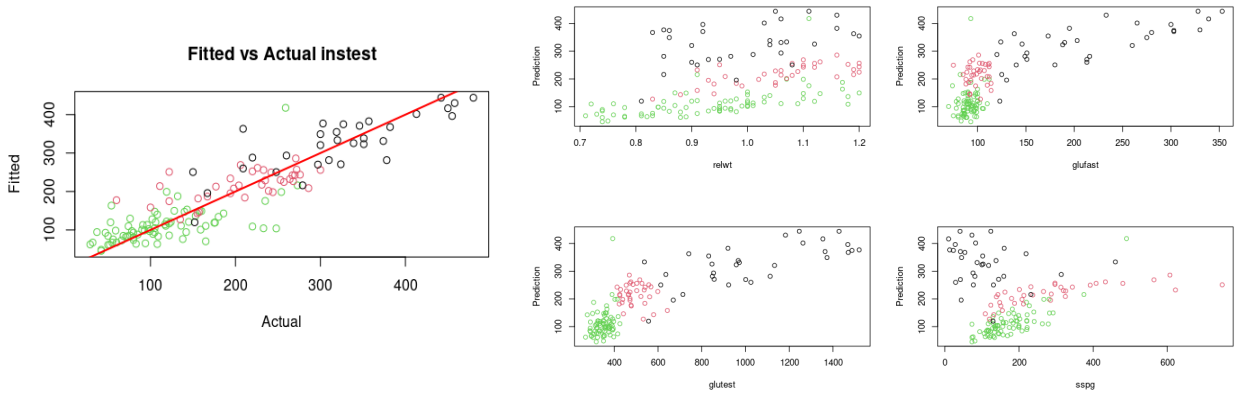


Fig 8.1; Fitted vs Actual plot (Left); Plot of explanatory variables against instest (Right)
Coloured according to group (Black : Overt Diabetic , Red : Chem Diabetic, Green : Normal)

The plot of predicted values of instest against all explanatory variables resembles that of the actual values and also in Fitted vs Actual plot the fitted values are spread evenly along line $y = x$, signifying the predicted values are close to the actual values and that our combined model does not have a bias. Using these actual and predicted values we can also find the corresponding residuals and the R^2 for this combined model as $R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})}{SST} = \mathbf{0.7927}$ where $\widehat{y_i}$ is the fitted value of i'th observation from the model corresponding to its group. Thus the three models combined do explain the variability in whole data very well.
The summary of predicted insulin levels in the three cases is :

Overt Diabetic Insulin Levels :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
119.9	278.9	332.4	325.8	375.4	444.1

Chem. Diabetic Insulin Levels :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
127.3	193.1	221.1	216.5	245.2	286.0

Normal Insulin Levels :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
45.48	83.53	103.06	111.52	123.34	417.67

We observe that Insulin levels are increasing as the seriousness of diabetes is increasing. This is expected as this is seen in patients with hyperinsulinemia observed in Type 2 diabetes patients which is the majority of diabetes cases.

We observed in Normal and Chemically diabetic patients the Insulin levels only significantly depended on relative weight and sspg and not much on plasma glucose levels as it was already sufficiently low. But in case of Overt Diabetic patients as plasma glucose levels are abnormally higher, the body keeps secreting more Insulin (Type 2 patients) and hence we observe a linear relationship between them.

The conclusions from our models match our expectations, hence the combination of our models can be safely used for prediction of Insulin levels given we already know the state of diabetes in patients.

9. SCOPE OF IMPROVEMENT

- One of the first drawbacks in the models is the requirement to already know the state of diabetes in patients in order to predict the Insulin levels using the appropriate model. This condition can be solved by also predicting the state of diabetes in patients i.e the ‘group’ variable, but as it is a categorical variable we will need to use techniques like Logistic Regression for prediction.

- The need for creation of three models can also be solved by considering ‘group’ as a factor column and then creating dummy variables for two out of its three levels. This model will then directly predict the Insulin levels rather than having to use an appropriate model by manually checking the group.

- As we observed the means of the Insulin levels are very distinct for the three models, this can also hint us towards creating a size alpha test for the state of diabetes using the predicted Insulin levels and possibly also other explanatory variables.