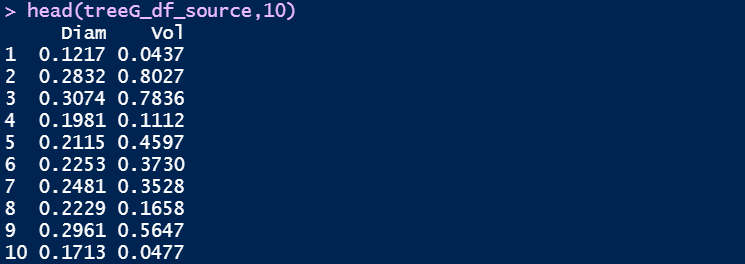
**Analysis of Tree Dataset**

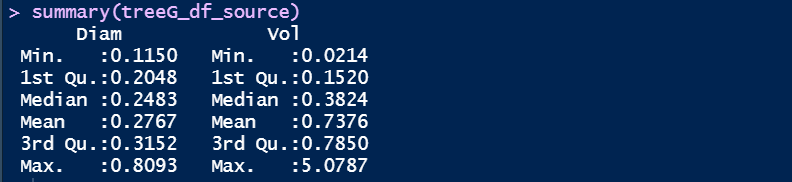
**a) Make a numerical and graphical summary of the data.**

It is given that the dataset consists of n observations of 2 variables, Diam and Vol which measure the diameter (m) and the volume (m3) of n tree trunks from the same species.

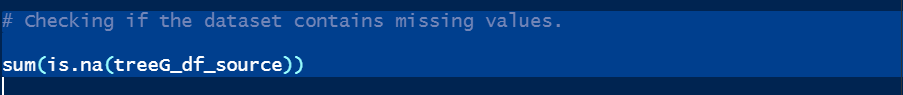
**First 10 observations from the Dataset.**



**Summary of the dataset.**

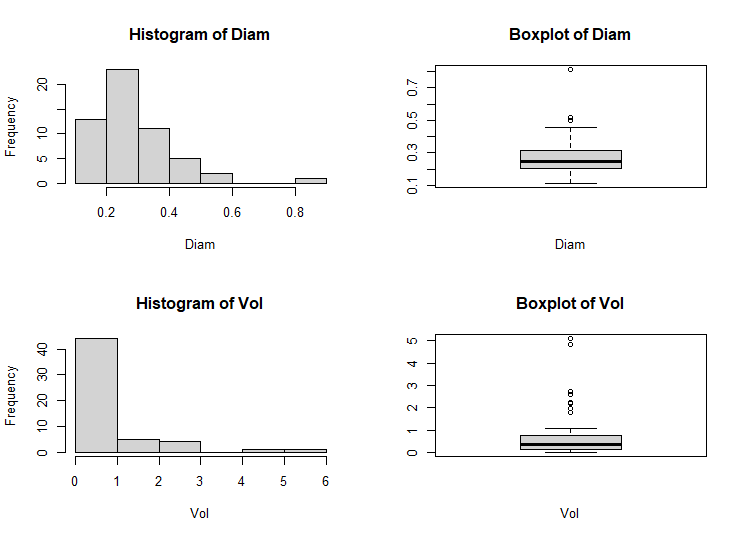


There are no missing values in the dataset.



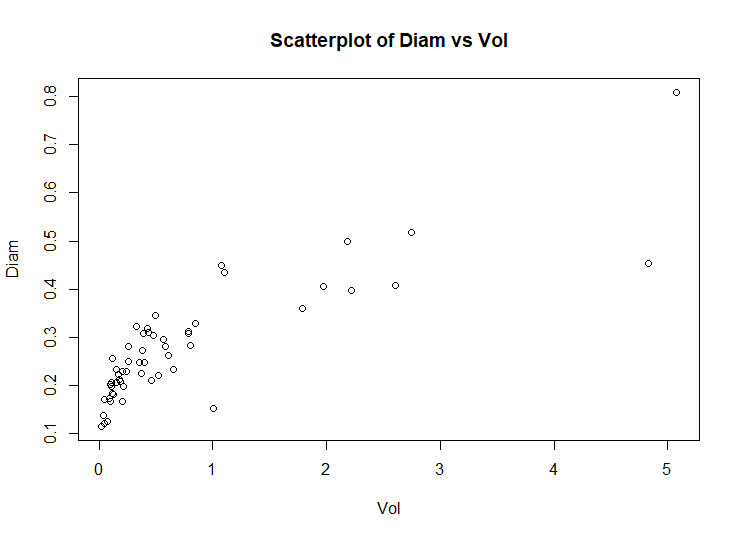


Here, univariate plots are created to understand the behavior of each field.



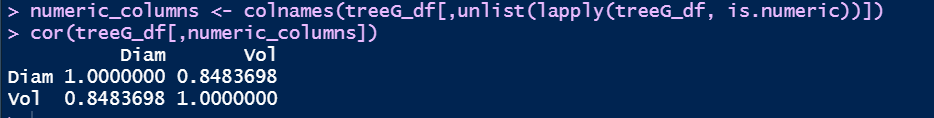
The histograms of Diam and Vol predictors resemble a right-skewed normal distribution. From the boxplots, we can see that there are three outliers in the Dam variable and seven outliers in the Vol variable.

Since there are only two numeric columns, a scatterplot would be used for the bi-variate analysis.



We can observe an increasing pattern between Vol and Diam variables.

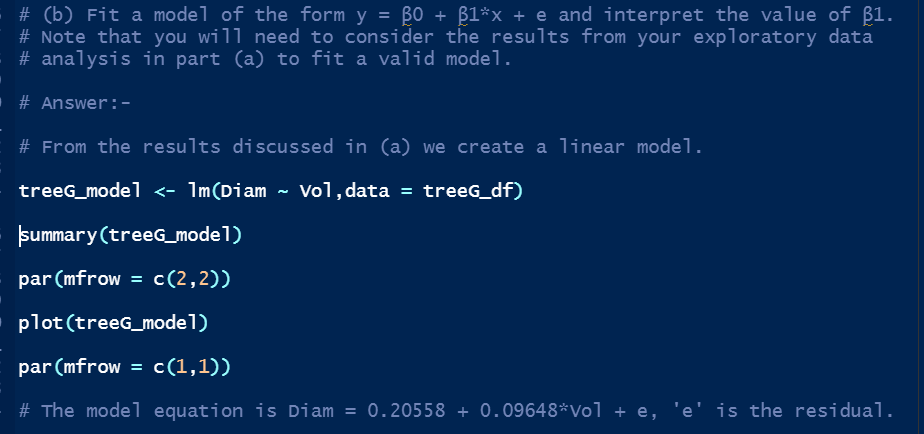
The correlation matrix can be used to validate this pattern.



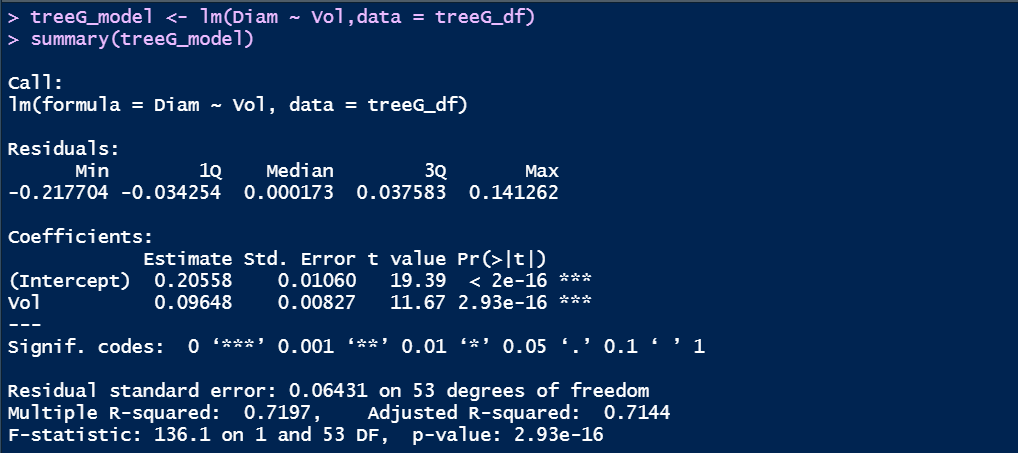
Since the correlation coefficient between the variables Diam and Vol is 0.843698, we can say that there exists a very strong linear relationship between them. Hence, we can fit a linear model.

**b) Fit a model of the form and interpret the value of .**

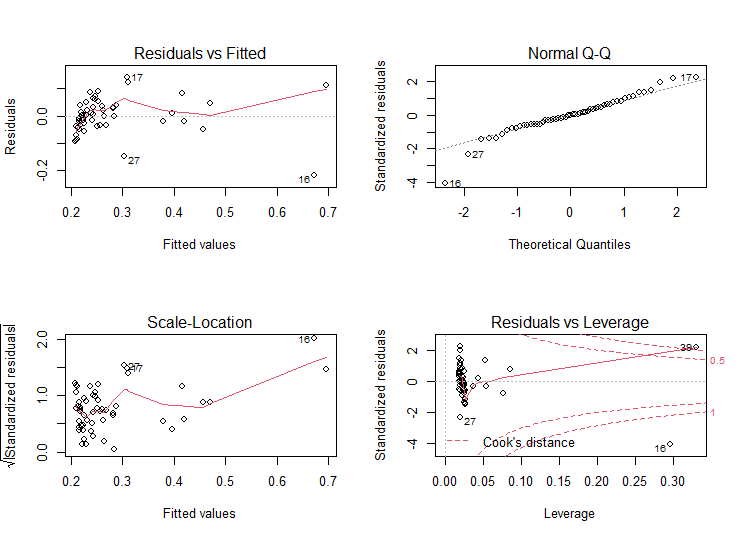
From the results discussed in (a) we create a linear model.



**Model summary**



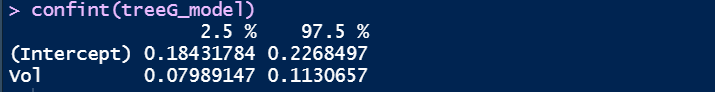
**Diagnostic Plots of the Model**



The model equation is Diam = 0.20558 + 0.09648\*Vol + e, where 'e' is the residual. Here, β1 = 0.09648.

**c) Calculate a 95% confidence interval for the coefficient.**

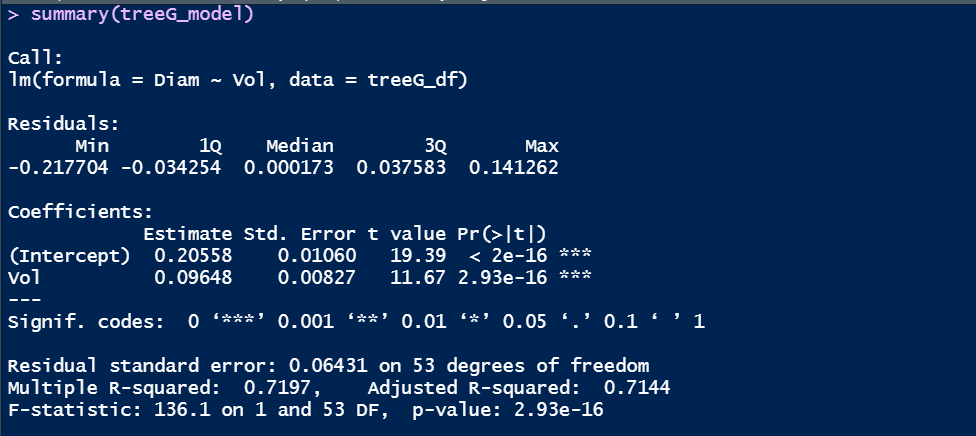
The 95% confidence interval for the coefficient can be calculated using the confint () function in R.



The 95% confidence intervals for the coefficient is [0.07989147, 0.11306572]. The coefficient of β1 = 0.09648 lies well within the confidence intervals.

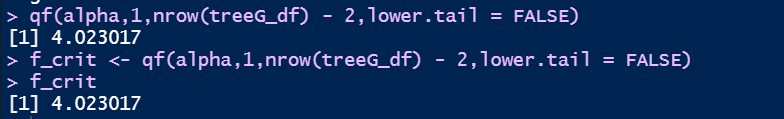
**d) Test the hypothesis for the regression model.**

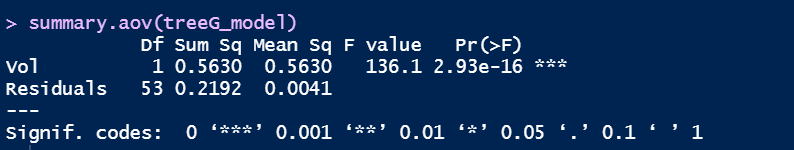
We can see the summary of the model to get the p-values.



Since p-value (2.93e-16) is less than alpha (0.05), we reject the null hypothesis. Also, the slope (0.20558) is non-zero which implies that the variable Vol can be used for modeling the Diam variable.

We can also test the hypothesis by viewing the ANOVA table.



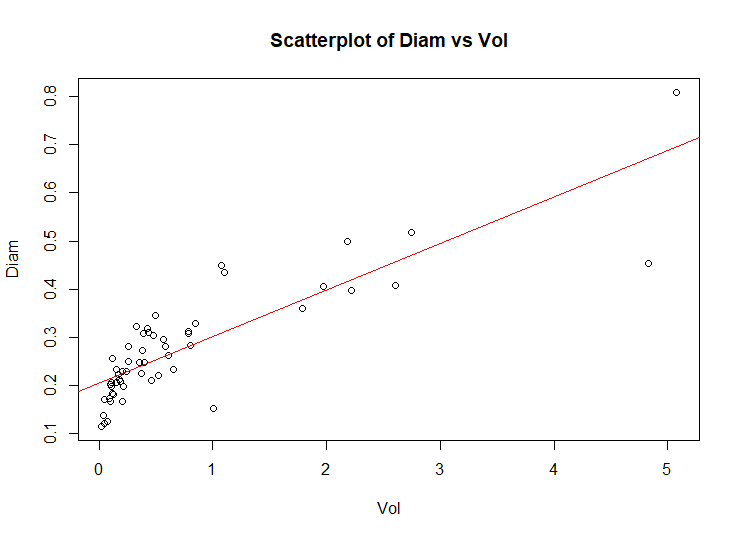


The F-statistic is 136.1 with (1,53) degrees of freedom and p < 0.001.

Since, the F-statistic (136.1) is greater than F-critical (4.023017), we conclude that the slope is non-zero and Vol is of value in explaining the variability of the Diam.

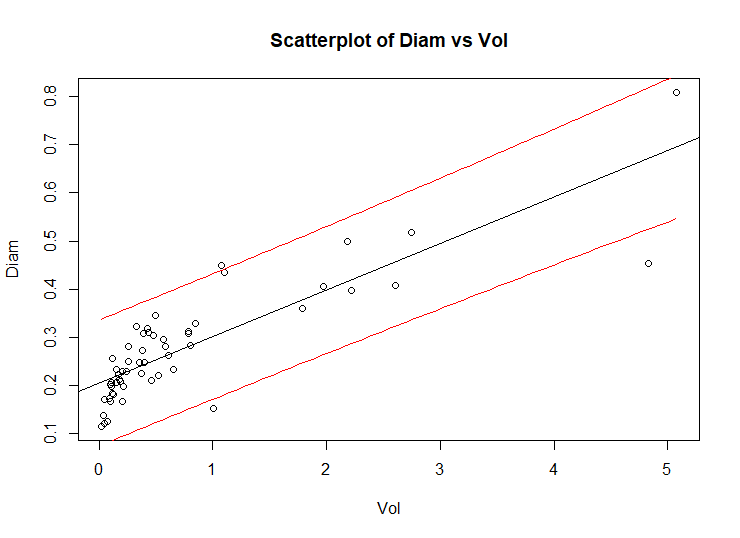
**e) Plot the regression line onto a scatterplot of the data and plot a 95% prediction band.**

**Plotting the regression line on the scatterplot**



The regression line shows a very strong linear relationship between Vol and Diam.

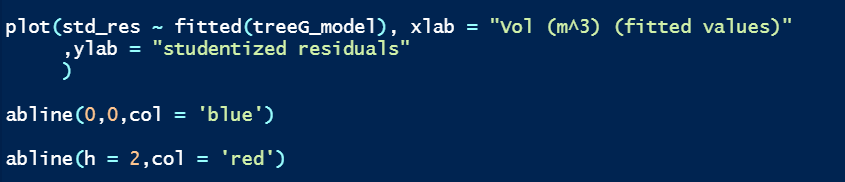
**Plotting the 95% prediction band**

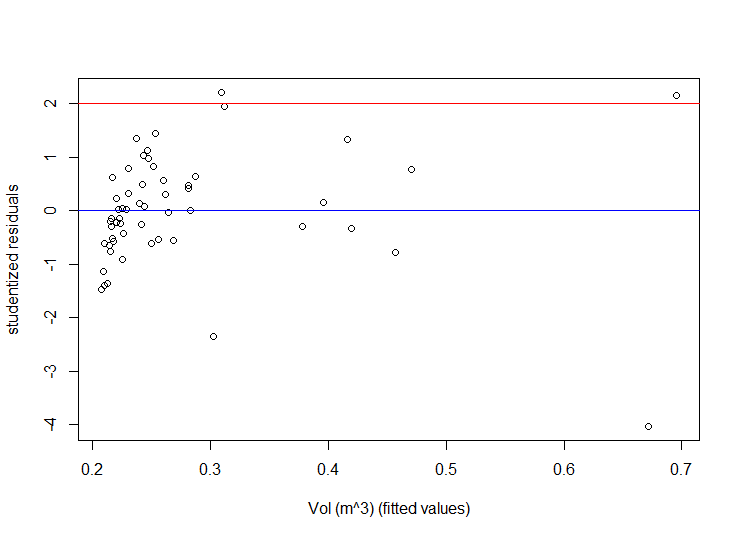


We can see that the data lies well within the 95% prediction band save for a couple of observations.

**f) Plot the studentized residuals against the fitted values and identify any outliers.**

The studentized residuals can be viewed with the help of plot command.



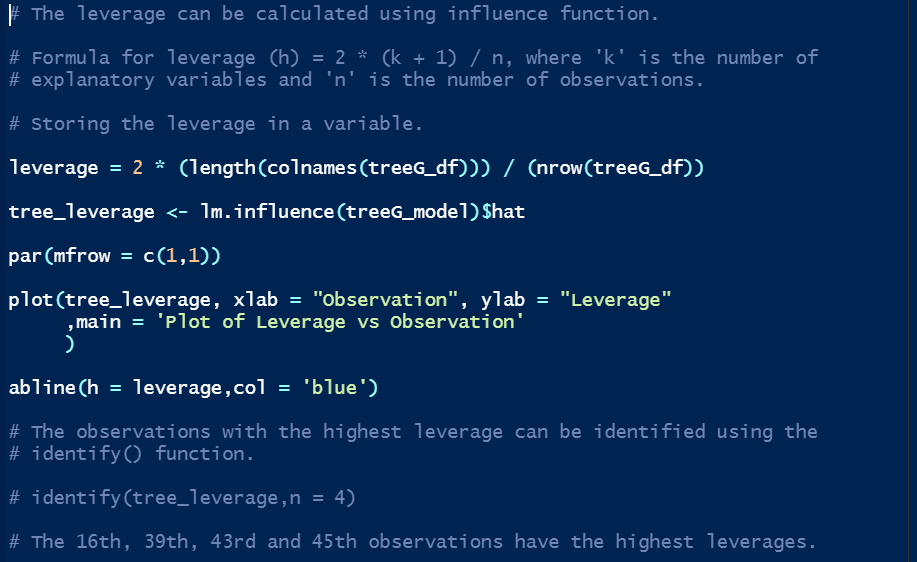


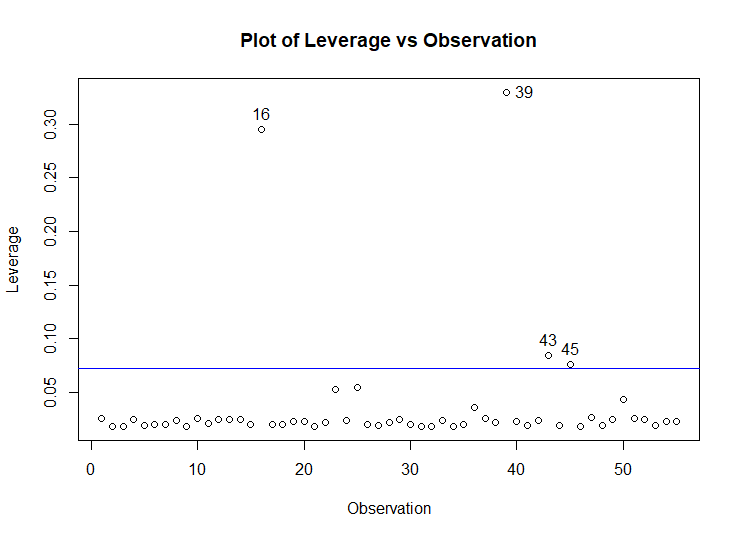
An observation may be considered an outlier if the magnitude of its studentized residual is greater than 2. From the plot we can say that there are very few outliers. However, there seems to be no pattern in the variance of the residuals i.e. there is no decrease or increase in the variance with respect to the mean.

**g) Plot the leverage of each case and identify any observations that have high leverage.**

The leverage can be calculated using influence function. Formula for leverage (h) = 2 \* (k + 1) / n, where 'k' is the number of explanatory variables and 'n' is the number of observations.

The observations with the highest leverage can be identified using the identify () function.

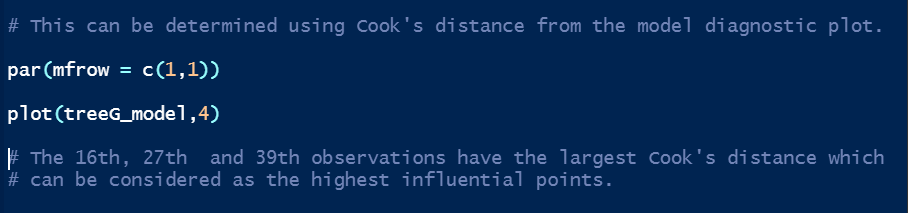


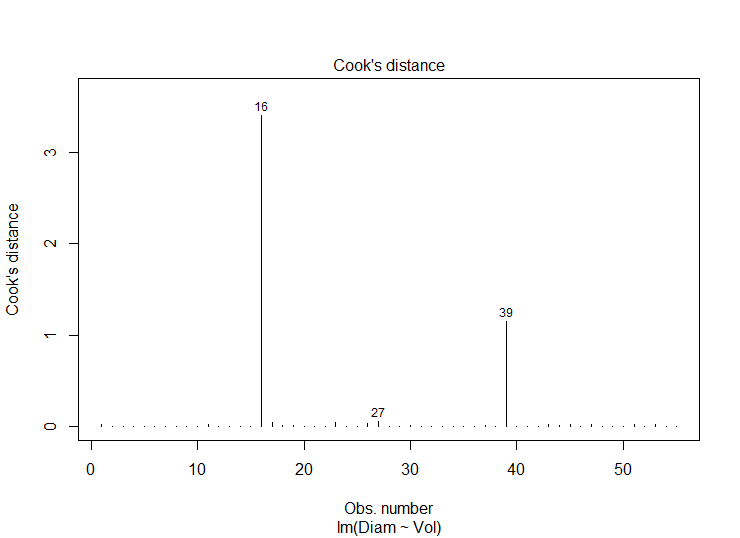


The 16th, 39th, 43rd and 45th observations have the highest leverages.

**h) Identify the observation that has the largest influence on the estimate of the coefficient.**

The influential points can be determined using Cook's distance from the model diagnostic plot.





The 16th, 27th and 39th observations have the largest Cook's distance which can be considered as the highest influential points.

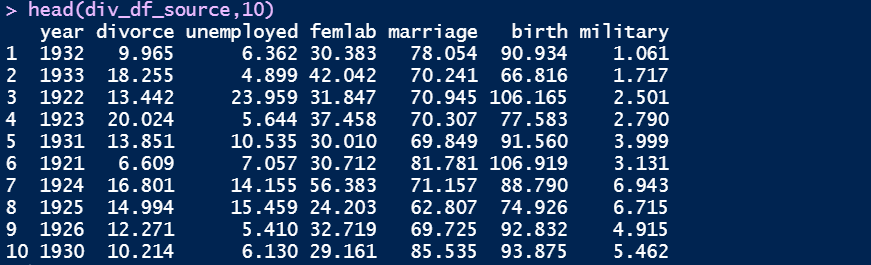
**Analysis of Divorce Dataset**

**a) Make a numerical and graphical summary of the data.**

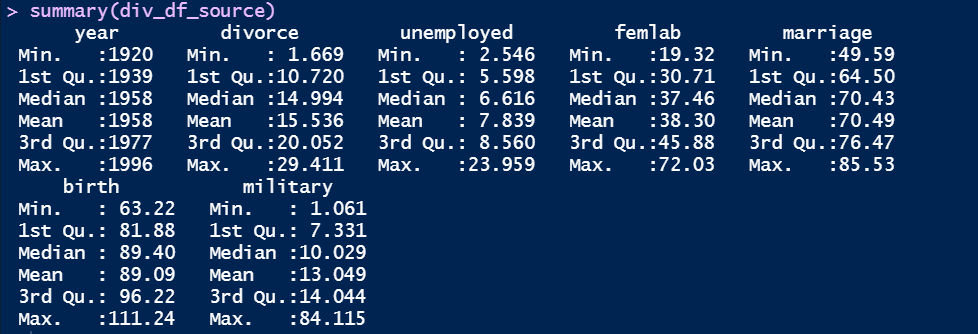
It is given that the dataset consists of 77 observations and 7 variables.

1. year the year from 1920-1996
2. divorce per 1000 women aged 15 or more
3. unemployed unemployment rate
4. femlab percent female participation in labour force aged 16+
5. marriage marriages per 1000 unmarried women aged 16+
6. birth births per 1000 women aged 15-44
7. military - military personnel per 1000 population

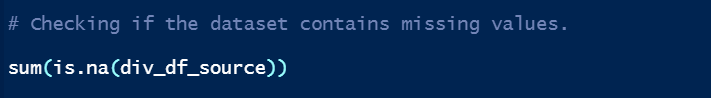
**First 10 observations from the Dataset.**



**Summary of the dataset.**



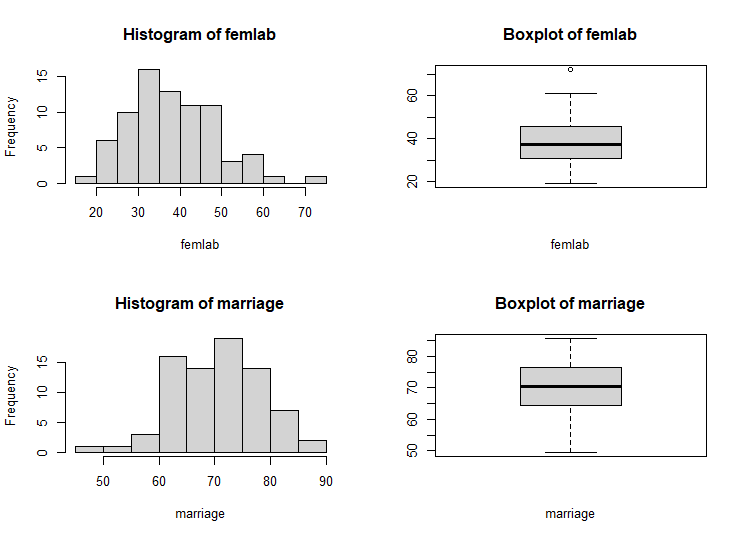
There are no missing values in the dataset.

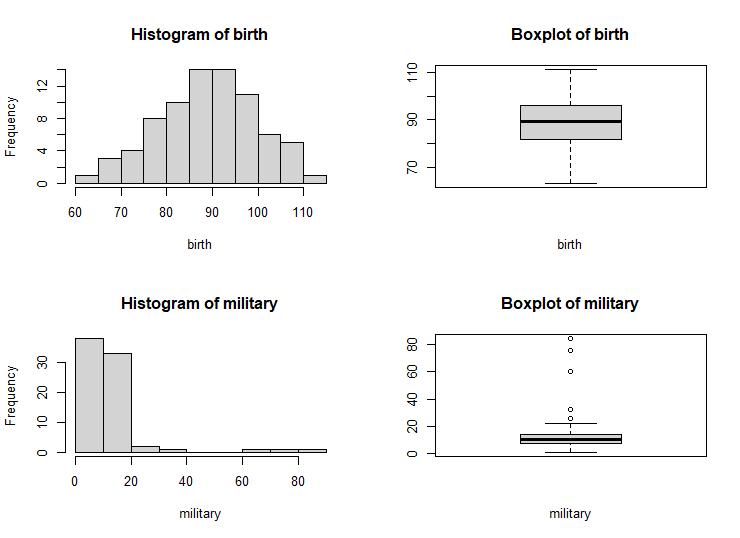




Here, univariate plots are created to understand the behavior of each field.

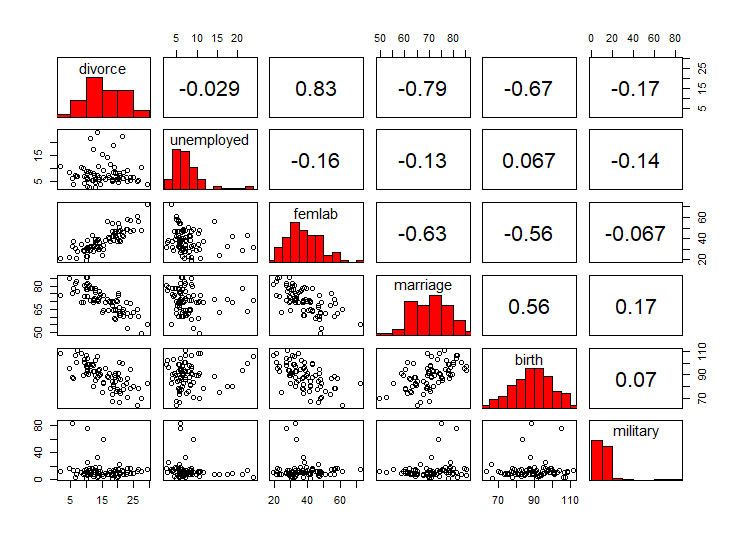






The histograms of divorce, unemployed, femlab and military predictors resemble a right-skewed normal distribution whereas the histograms of marriage and birth predictors resemble a normal distribution with a minor left skew. From the boxplots, we can see that there are very few outliers in the unemployed and military variables and only one outlier in the femlab variable. The other attributes have no outliers.

Here, pairwise plot would be used for Bi-Variate analysis.



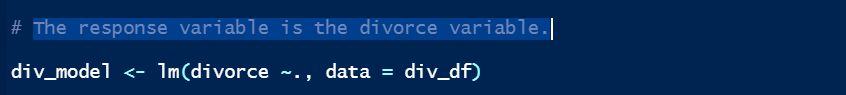
We can see that there is a very strong positive linear relationship between divorce and femlab variables, while there is a very strong negative linear relationship between variables divorce and marriage.

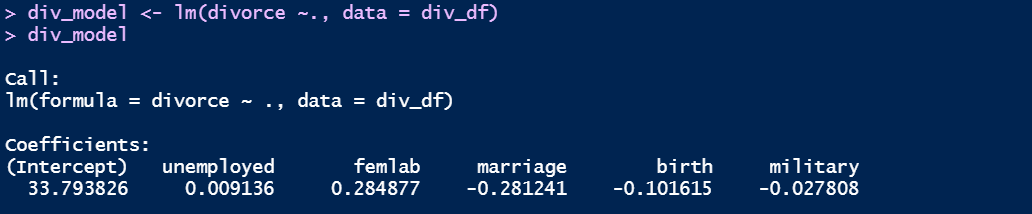
Moderate relationships are found among the variable pairs (divorce, birth), (femlab, marriage), (femlab, birth) and (marriage, birth).

Weak relationships are found among the remainder of the variable pairs.

**b) Fit the model:**

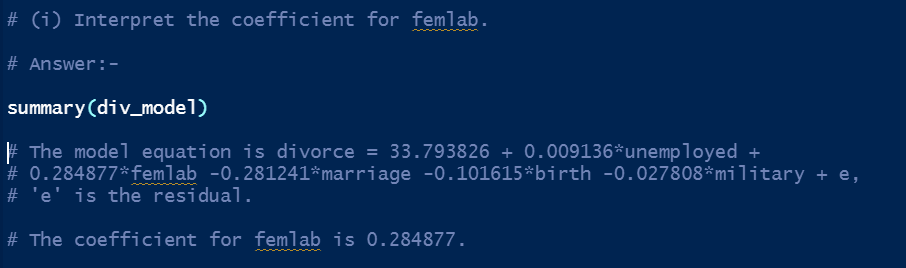
The response variable is the divorce variable.

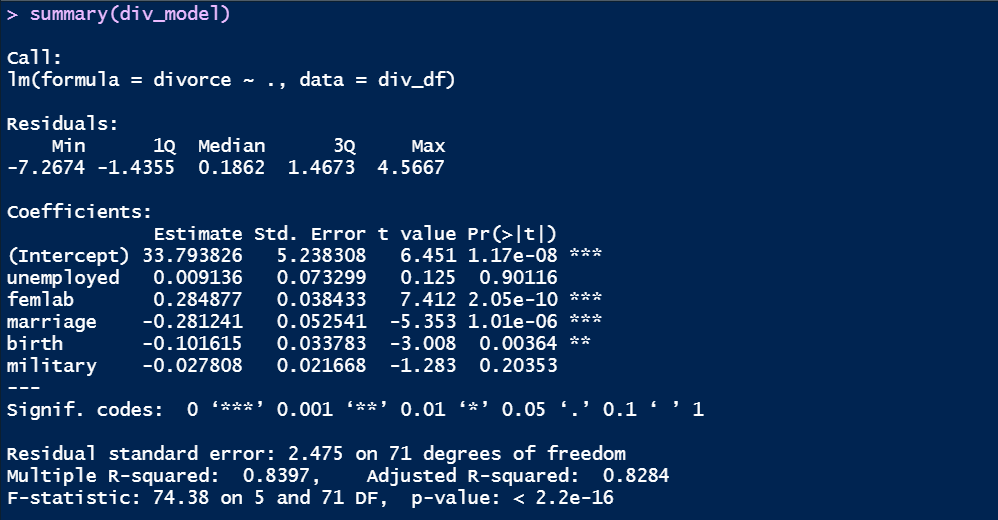




**b) (i) Interpret the coefficient for femlab.**

We can check the model summary for interpreting the coefficients.

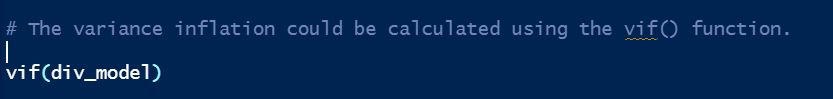


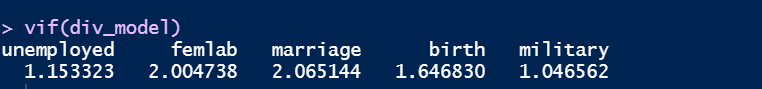


The model equation is divorce = 33.793826 + 0.009136\*unemployed + 0.284877\*femlab -0.281241\*marriage -0.101615\*birth -0.027808\*military + e, where 'e' is the residual. The coefficient for femlab is 0.284877.

**b) (ii) Calculate the variance inflation factors for this model and discuss their implications for collinearity in the model.**

The variance inflation could be calculated using the vif () function.



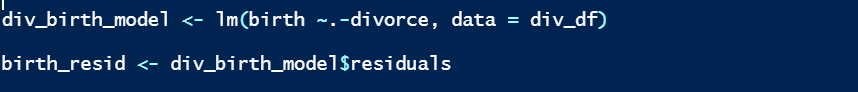


Collinearity can be detected from the variance inflation factor (VIF) associated with each coefficient. If the maximum VIF exceeds 5 then this is an indication of serious collinearity between the explanatory variables.

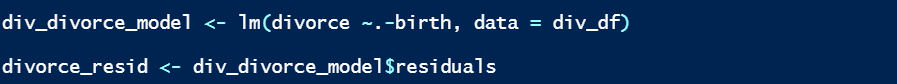
Since the VIF is less than five for all the variables, we can say that there exists a very weak collinearity between them.

**b) (iii) Create a partial regression plot to examine relationship between birth and divorce adjusted for unemployed, femlab, marriage and military.**

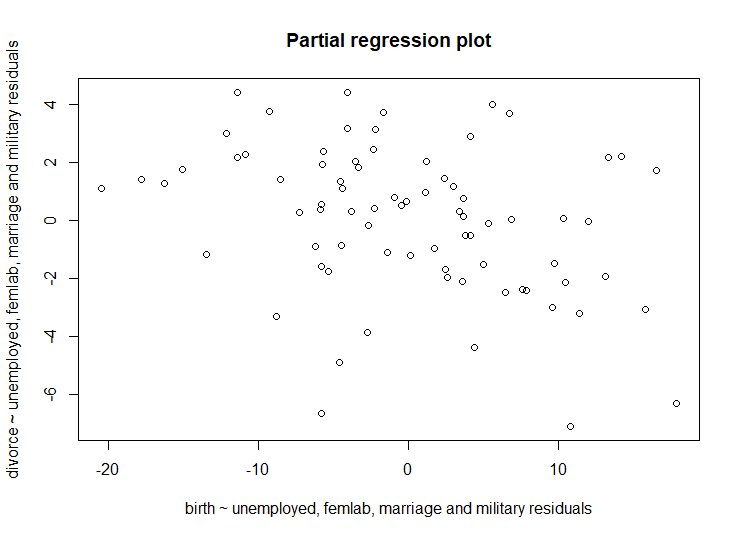
For creating a partial regression plot, first we need to calculate the residuals for the model birth ~ unemployed, femlab, marriage and military.



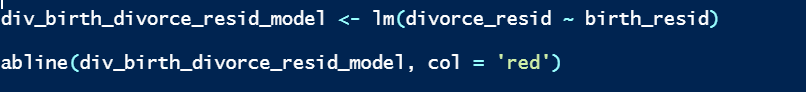
Next, we need to calculate the residuals for the model divorce ~ unemployed, femlab, marriage and military.



We then plot the residuals from the second model (y) against the residuals from the first model (x).

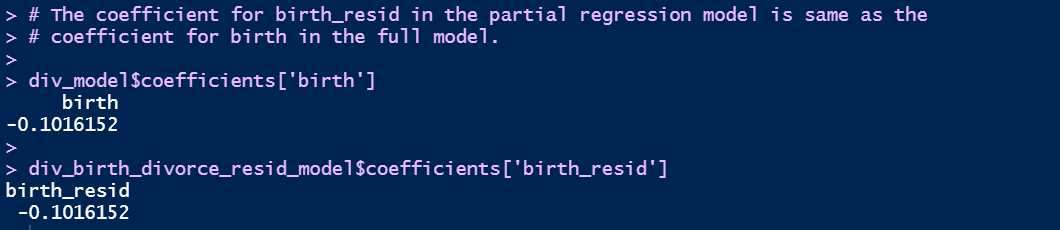


Finally, we can fit a regression line to the two sets of residuals, the slope of the regression line measures the effect of birth on divorce adjusted for unemployed, femlab, marriage and military.





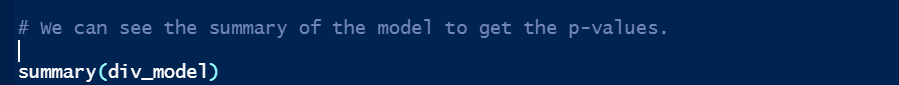
We can see a negative relationship between the two sets of residuals which represents the relationship between birth and divorce adjusted for unemployed, femlab, marriage and military.

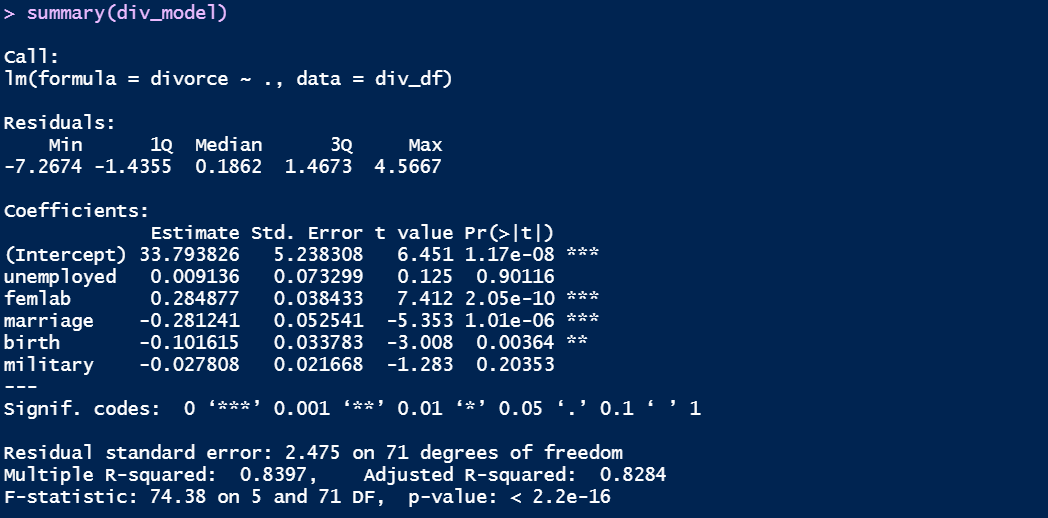


The coefficient for birth\_resid in the partial regression model is same as the coefficient for birth in the full model.

**b) (iv) Test the hypothesis of the regression model.**

We can see the summary of the model to get the p-values.



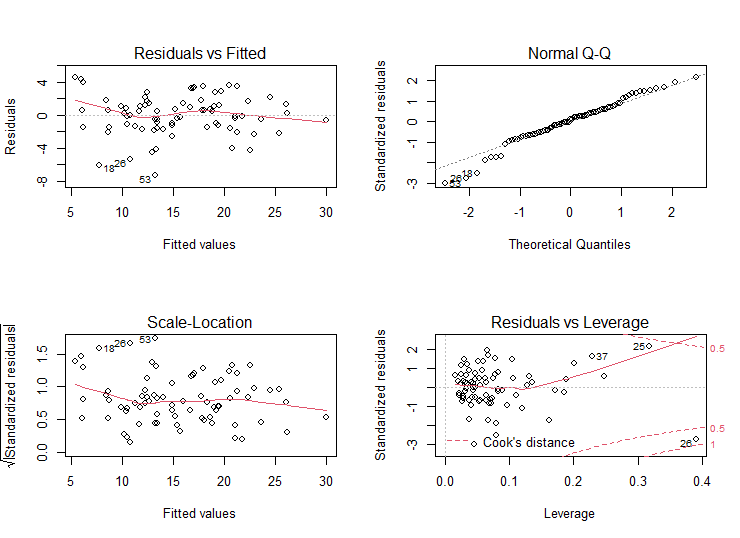


From the summary we can see that the F-Statistic is F (5,71) = 74.38 and p-value < 2.2e-16. This F-Statistic compares the intercept-only model (null model) to the fitted model and here we may reject the null hypothesis at the 1% confidence level and conclude that at least one of the predictors is associated with divorce.

**b) (v) Assess the fit of the model using diagnostic plots, commenting on the assumptions of the regression model and influential points.**

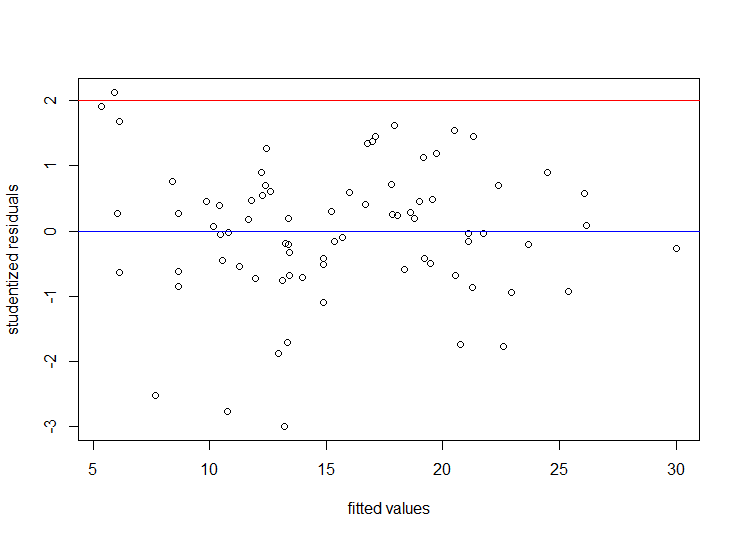
By plotting the model, we can perform the model diagnosis.

First, we need to check the homoscedasticity of the residuals.



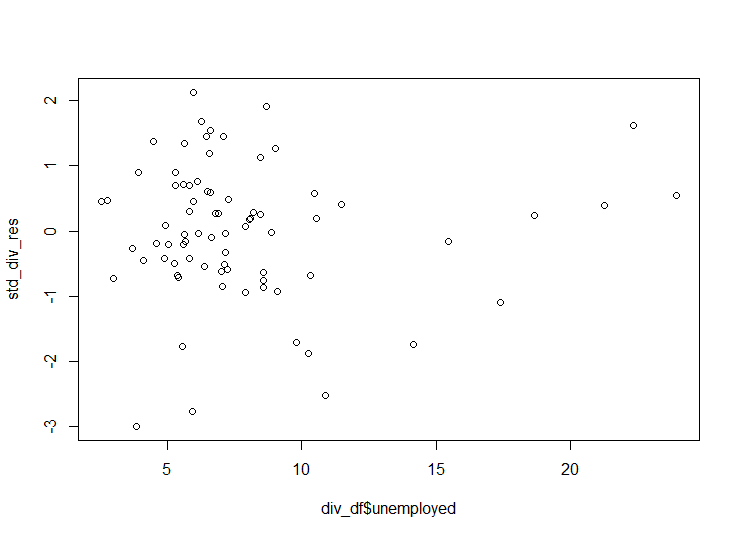
From the Residuals vs Fitted values plot we can see that there is no pattern in the variance of the residuals i.e. there is no decrease or increase in the variance with respect to the mean. The residuals are randomly distributed with zero mean and standard deviation (sigma).

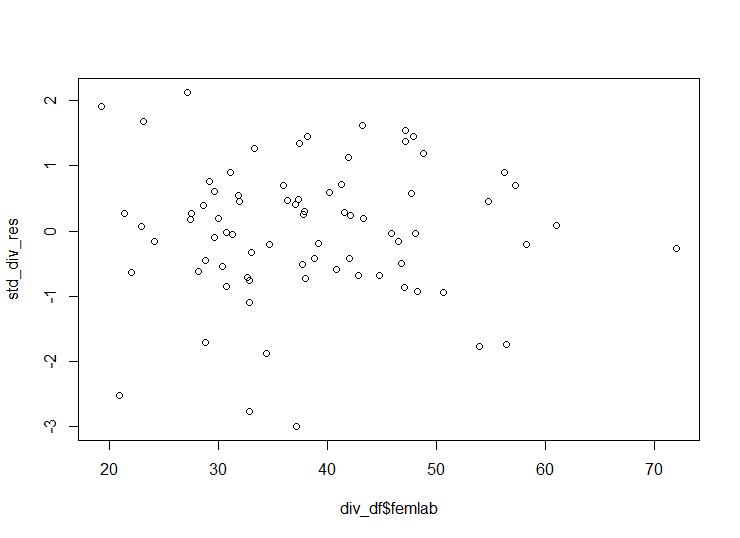
We then check the studentized residuals for the fitted values.

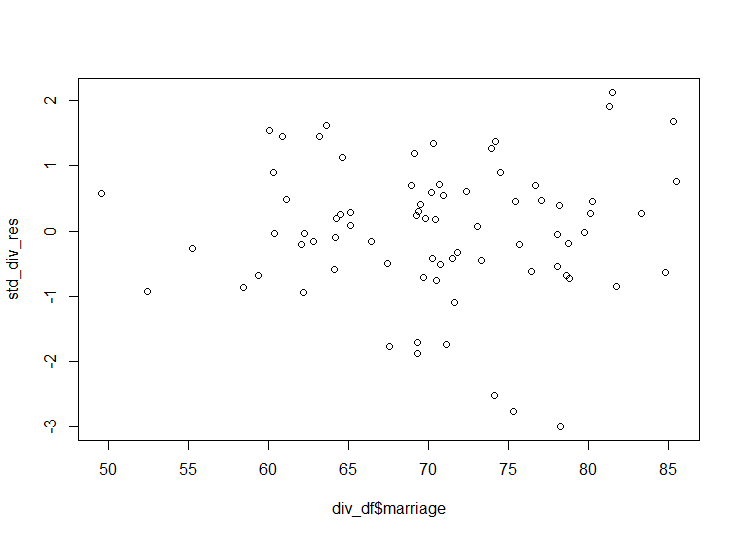


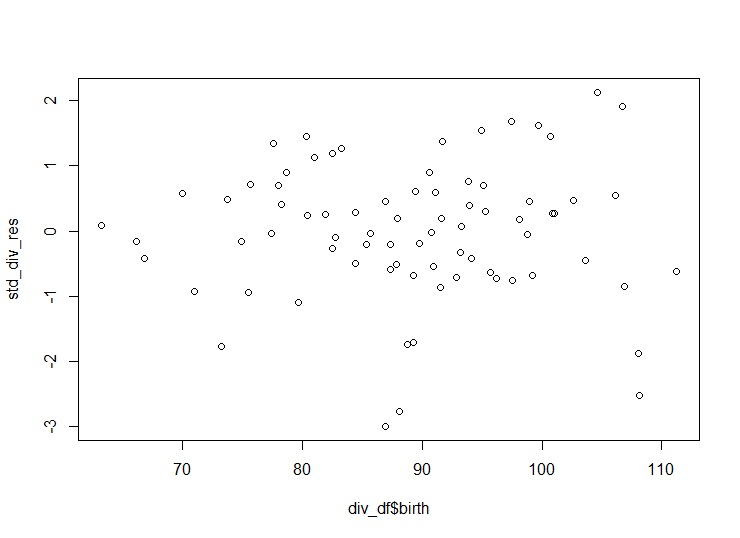
An observation may be considered an outlier if the magnitude of its studentized residual is greater than 2. From the plot we can say that there are very few outliers. However, there seems to be no pattern in the variance of the residuals i.e. there is no decrease or increase in the variance with respect to the mean.

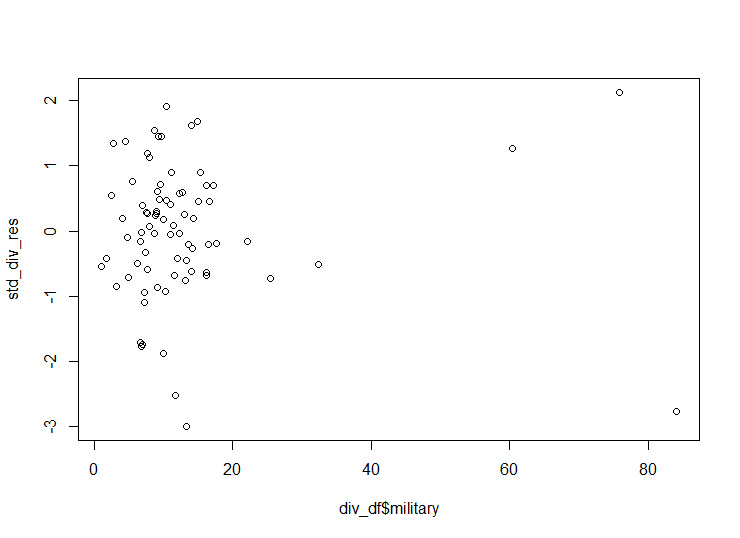
We then check the studentized residuals for each predictor in the dataset.







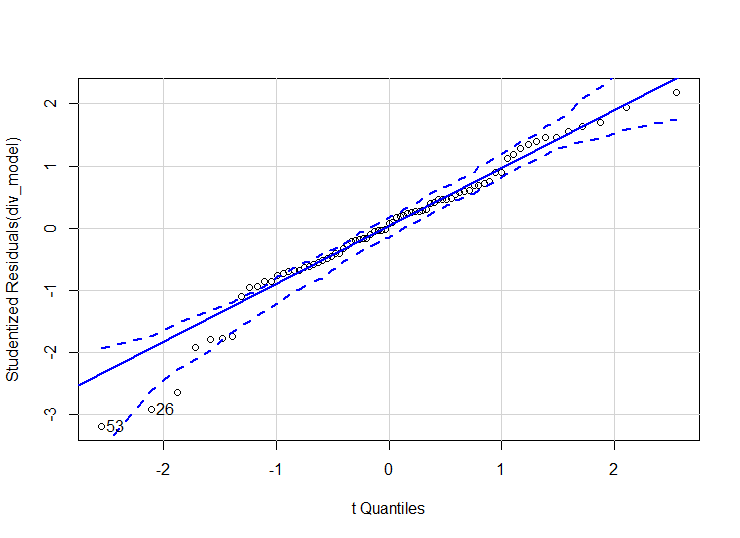


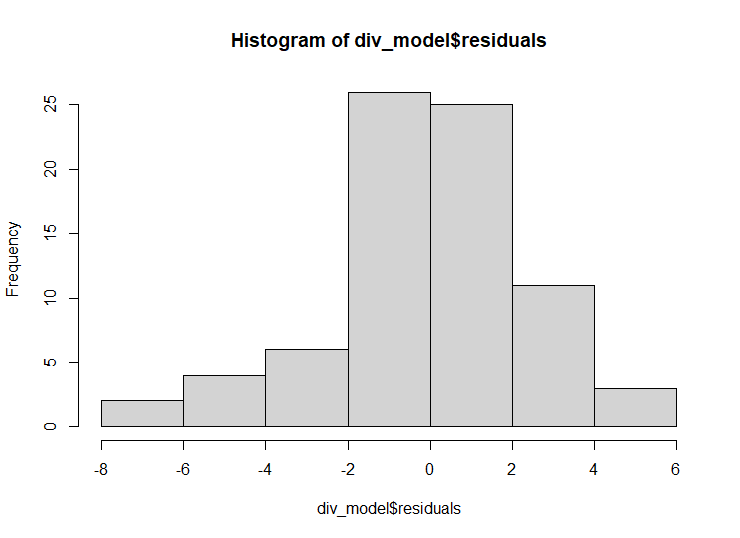


In the first plot, the residuals appear to be split into two groups, with one group having a larger variance with the other having a smaller variance.

In the second plot there could be one possible outlier in the variable femlab. The residuals in the next couple of plots show no sign of heteroscedasticity. The last plot of the military variable shows three possible outliers.

Next, we need to check the normality of the residuals.

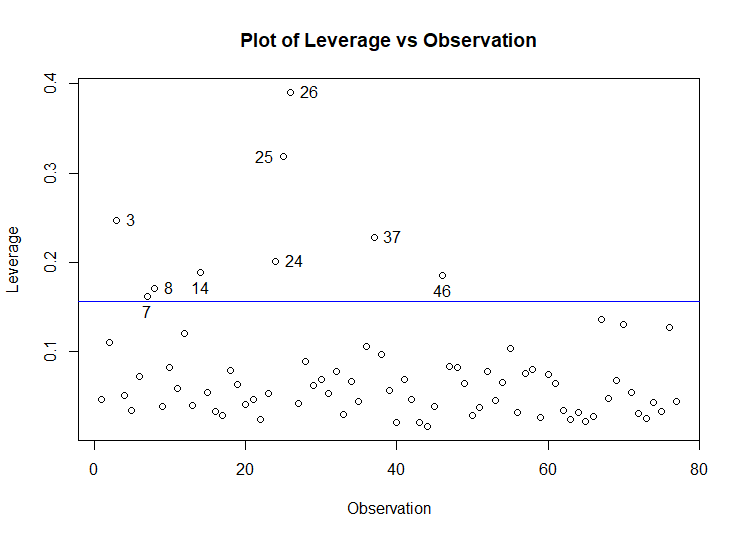


From the Normal Q-Q plot we can see that the residuals are almost normally distributed except for few observations.

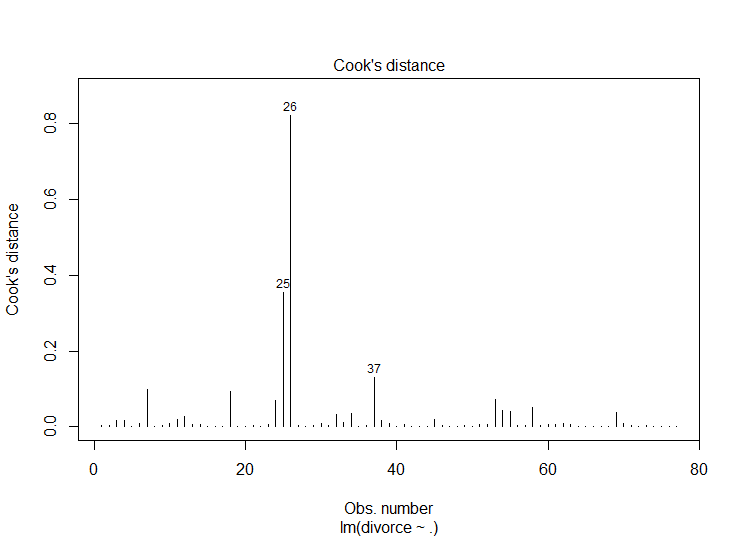
This can be observed in the histogram as well.

Finally, we need to find the influential points. The leverage can be calculated using influence function. Formula for leverage (h) = 2 \* (k + 1) / n, where 'k' is the number of explanatory variables and 'n' is the number of observations.

The observations with the highest leverage can be identified using the identify () function.



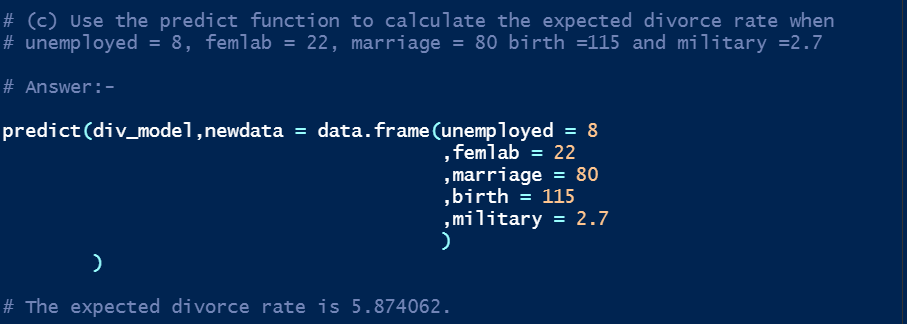
This can be determined using Cook's distance from the model diagnostic plot.



The 25th, 26th and 37th observations have the largest Cook's distance which can be considered as the highest influential points.

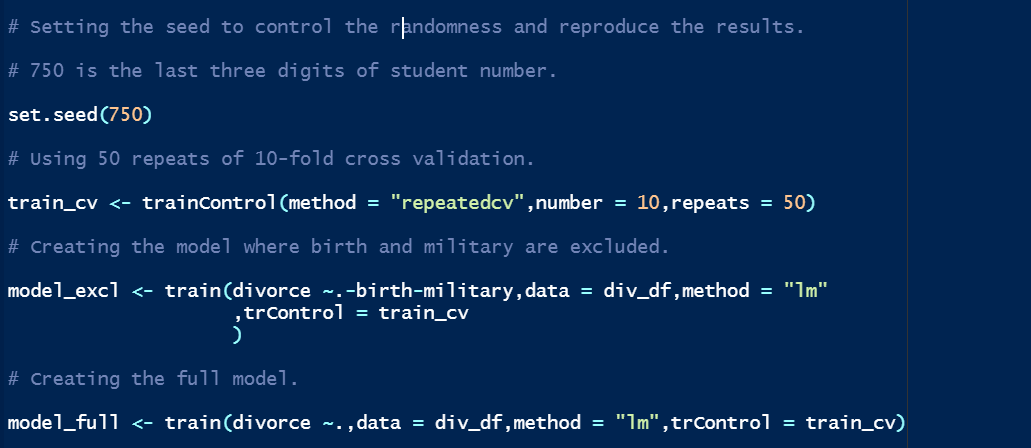
**c) Use the predict function to calculate the expected divorce rate when unemployed = 8, femlab = 22, marriage = 80 birth =115 and military =2.7**

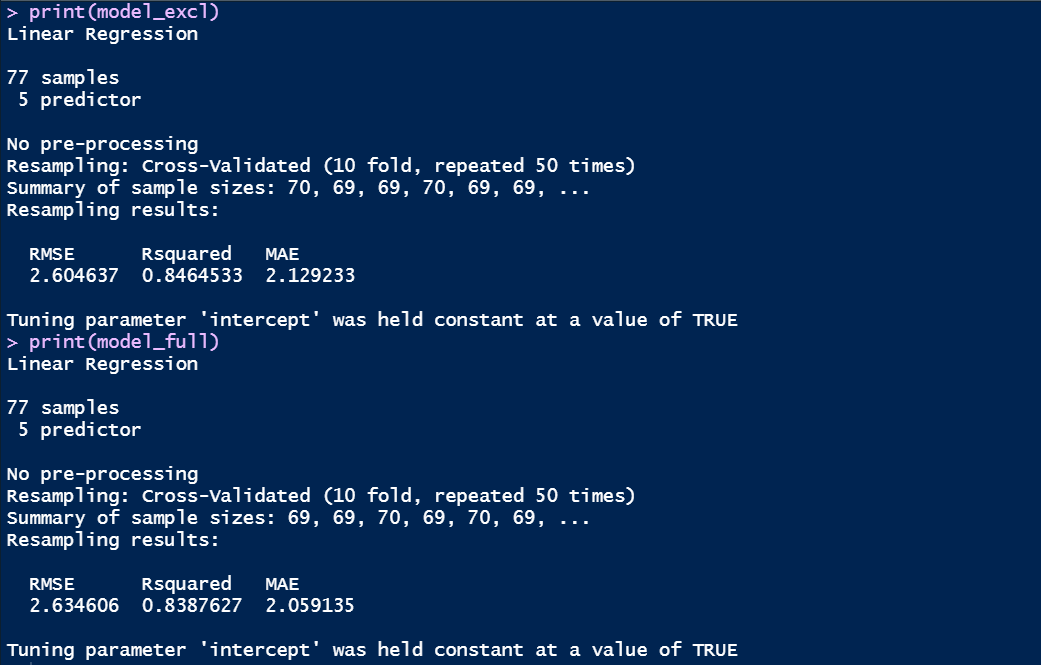
The expected divorce rate is 5.874062 for the given input values.



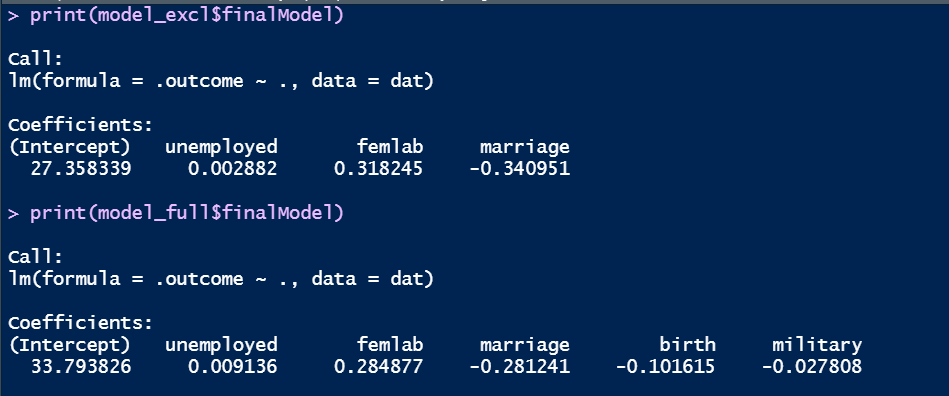
**d) Compare the full model to the model where birth and military are excluded using 50 repeats of 10-fold cross validation. Which model would you choose to predict crime rate?**

Creating the full and reduced models using 50 repeats of K = 10 cross-fold validation.





**Displaying the final models in both the full and reduced models.**



The RMSE of the excluded model is less than the RMSE of the full model.

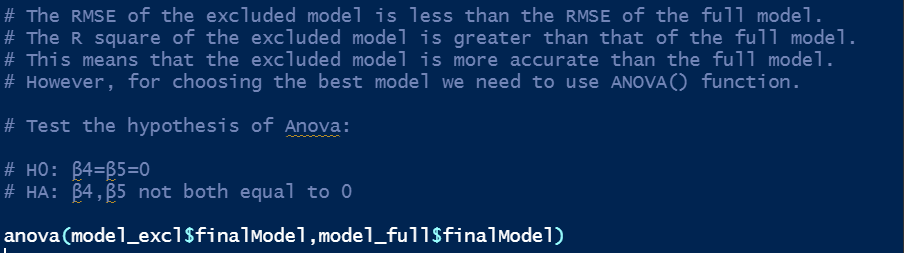
The R square of the excluded model is greater than that of the full model.

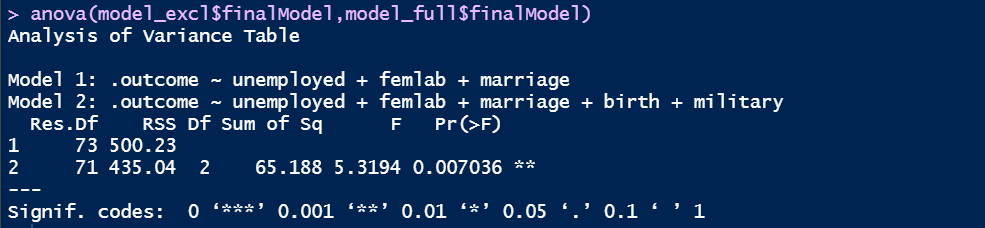
This means that the excluded model is more accurate than the full model. However, for choosing the best model we need to use ANOVA () function.

Test the hypothesis of ANOVA:

H0: β4=β5=0

HA: β4, β5 not both equal to 0





From the ANOVA we can see that the F-statistic is F (2, 71) = 5.3194 and p-value = 0.007036. This F-statistic compares the fit of reduced model to the full model at the 1% confidence level. In this case, we fail to reject H0 and conclude that there is not enough evidence to prove that the variables birth and military are associated with divorce (when they are included in the model). Hence, we can use the reduced model for prediction.

**Study on Step-wise Regression**

While building multiple linear regression models, the difficulty arises during the variable selection (Flom et al., 2007). Since there are multiple explanatory variables involved, it is hard to find the best combination of variables that would yield the optimal result. To overcome this, stepwise regression is a common method used while building the regression models.

Stepwise regression is an automated procedure used for assessing the inclusion or deletion of a variable based on predefined criteria such as Akaike’s Information Criterion (AIC), R2 (coefficient of determination), etc.

There are three types of stepwise regression models namely forward, backward, and bi-directional regression models (Smith, 2018).

In the backwards model, all the explanatory variables are used to build the initial regression model and slowly the variables are dropped after assessing them. In the forwards model, a null model (intercept-only model) is created, and variables are added after assessment. In the bi-directional model, the initial model is created like in the forward model and the variables are added after assessment. In case the addition of variable does not yield the desired result, the variable is dropped, and another variable is added. There is a chance that the same variable could be added, deleted, and once again added in the subsequent steps. Whichever type is used, the stepwise model always returns only one model (best fit).

The most used criteria for assessing the impact of a variable in a regression model are Akaike’s Information Criterion (AIC), R2 (coefficient of determination), Akaike’s Corrected Information Criterion (AIC), R2 (coefficient of determination), Bayesian Information Criterion (BIC), and Mallow’s Cp (Lindsey and Sheather, 2010).

However, despite the automation in variable selection there are multiple problems with the stepwise methods such as high bias in the R2 values, too much susceptibility to collinearity, getting very low p-values and so on.

One such problem is in the p-values. Usually models use test statistics such as F-statistic for assessing the hypothesis associated with them.

However, they are not suitable when multiple steps are involved during the variable selection. This would hamper the standard errors of the model coefficients and narrow down the confidence intervals.

As a result, this would increase the test-statistic values thereby reducing the p-values. This could cause overfitting and induce false confidence intervals in the final model (Smith, 2018).

**Step-wise Regression using R**

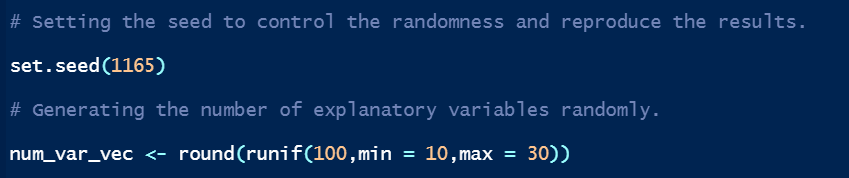
Stepwise regression can be performed in the statistical software R using functions such as stepwise () function from the Package ‘StepReg’ and stepAIC () function from the package ‘MASS’. The stepAIC () function takes arguments such as the model object or formula, type of regression to be used, etc. It evaluates the explanatory variables using Akaike’s Information Criterion (AIC).

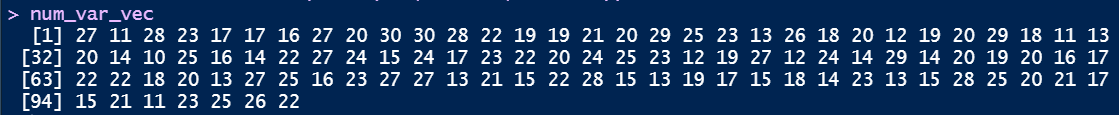
Basic programming algorithms (matrix manipulations, conditional expressions, etc.) are implemented inside the function. This function supports all types of stepwise regression namely forward, backward, and bi-directional models (Venables and Ripley, 2002). When the linear model is passed and backward direction is assumed, the function creates a linear model using all the explanatory variables and captures the AIC associated to each variable. The variable with the lowest AIC is removed from the full model and a new model is created with the new AIC values. The ANOVA () function is used internally in the stepAIC () function for comparing the models. If the AIC (old model) < AIC (new model), then the old model is retained. The new model would be considered only when the AIC of the new model is less than the old model.

As AIC decreases, the desirability of the model increases. This process is automatic and will stop only when the model with the best parameters is obtained i.e. until the model with the lowest AIC is generated.

**How does varying the number of predictors affect the performance of stepwise regression?**

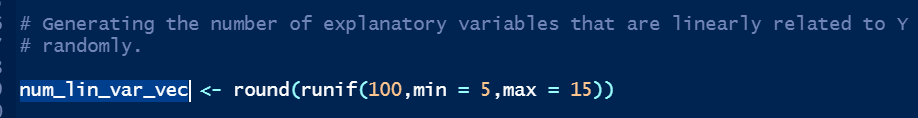
For assessing the relationship between the performance of stepwise regression and the number of predictors, we generate 100 random numbers which indicates the number of explanatory variables to be used for each model.

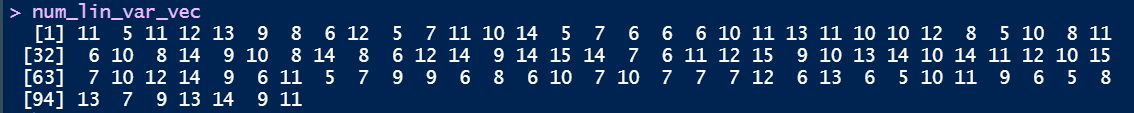




For instance, the first model would be having 27 explanatory variables.

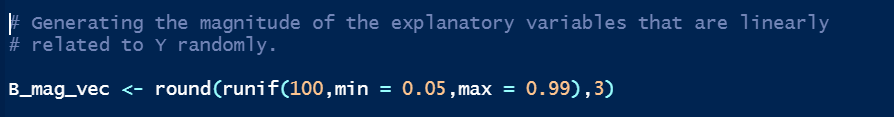
Next, we generate 100 random numbers which indicates the number of explanatory variables that are linearly related to Y which is to be used for each model.

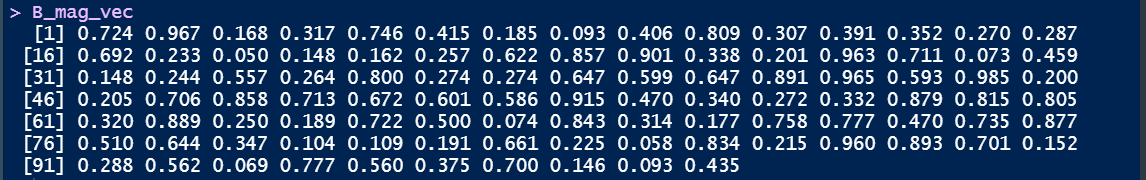




For instance, the first model would be having 11 explanatory variables that are linearly related to Y. The other variables would be considered as noise.

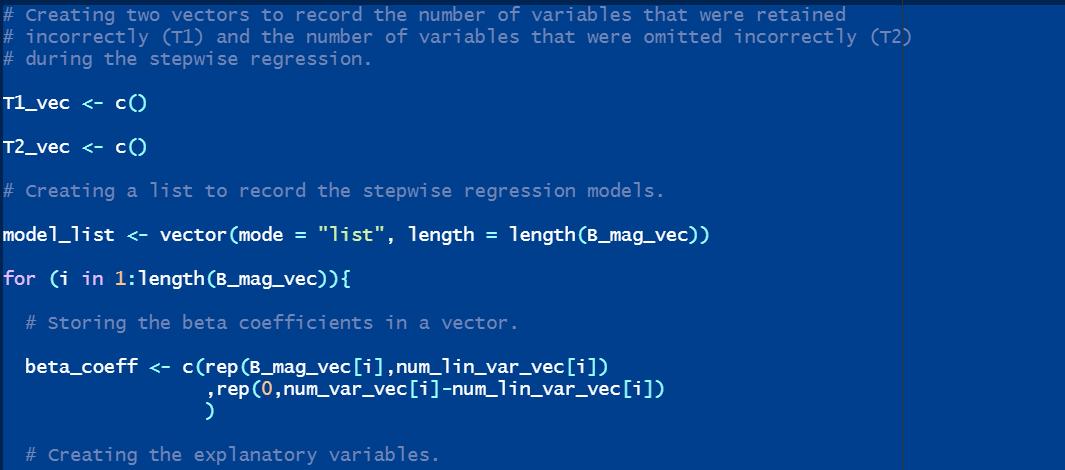
Similarly, we generate 100 random numbers that would be the model coefficients (β) or effect size for each model.

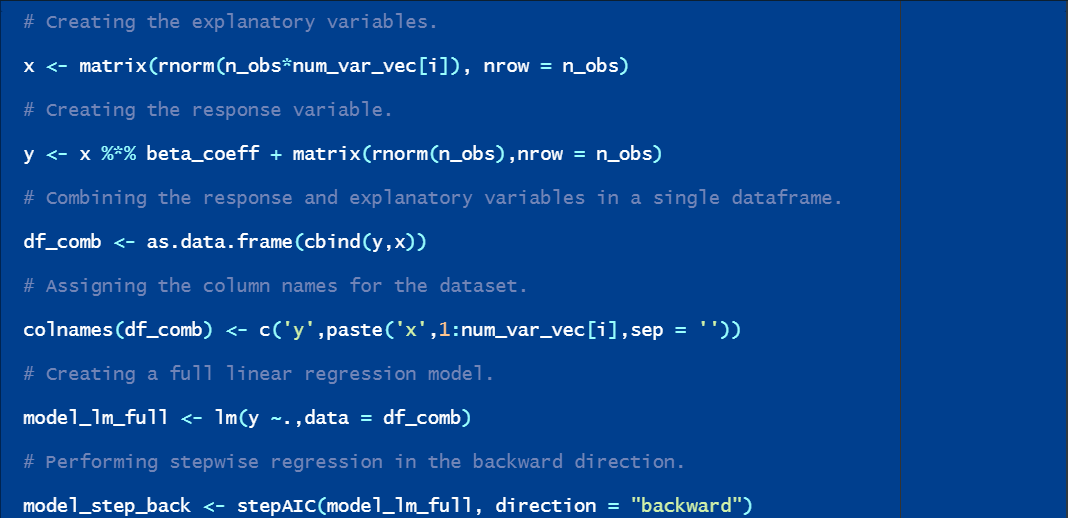


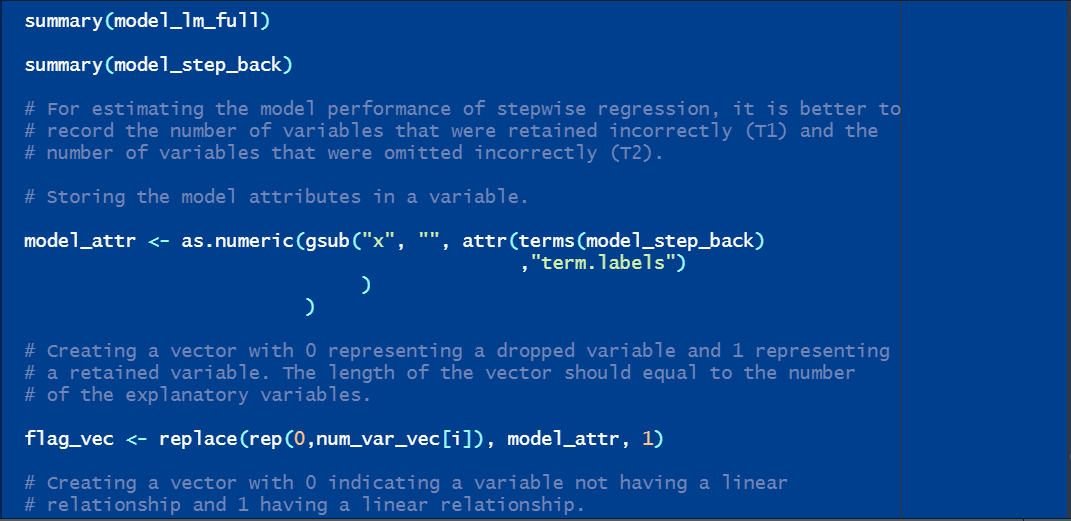


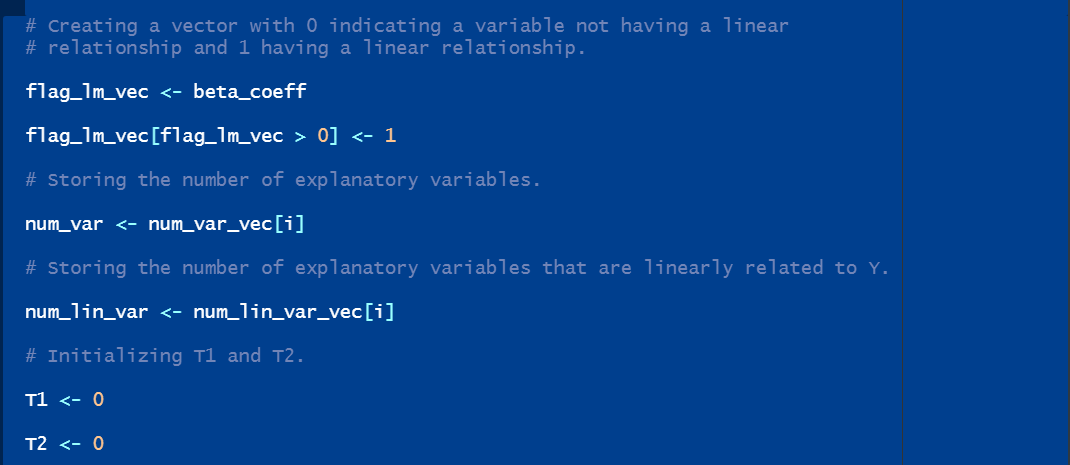
For instance, the first model would be having 11 explanatory variables that are linearly related to Y. The other variables would be considered as noise.

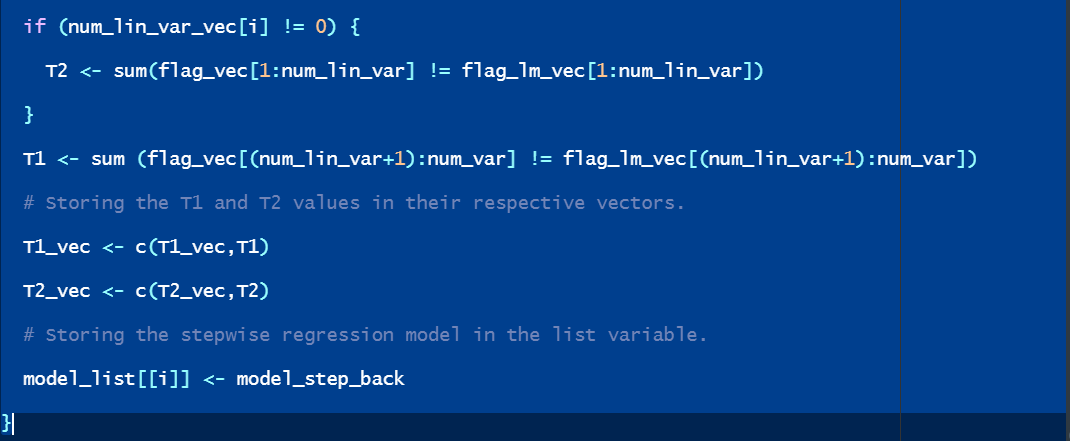
Next, we would be generating 100 datasets with 100 observations.



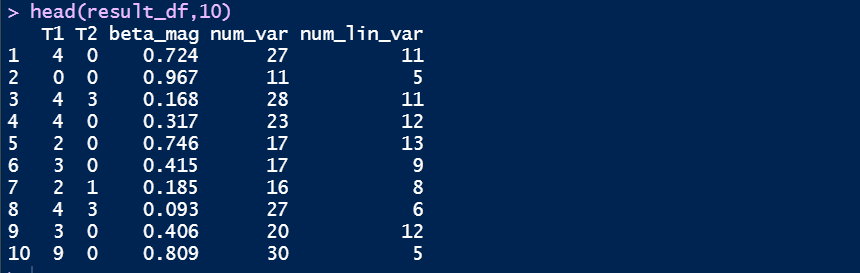




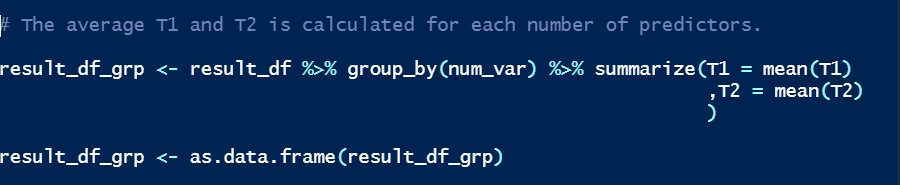


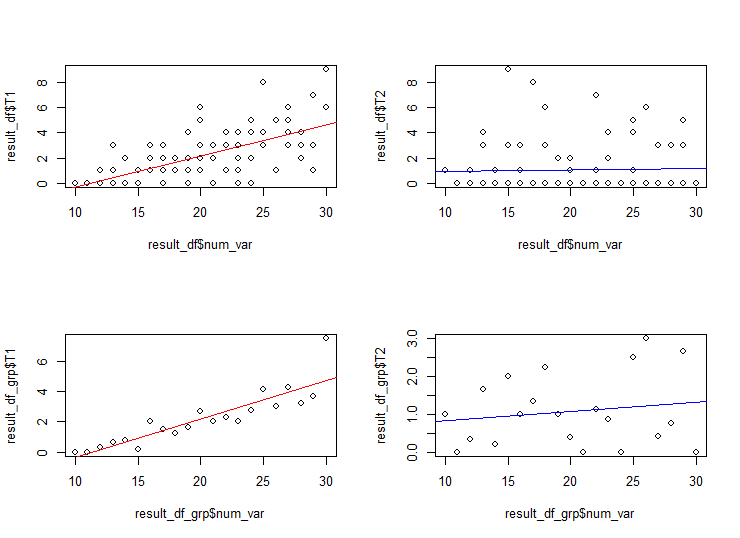


For estimating the model performance of stepwise regression, we record the number of variables that were retained incorrectly (T1) and the number of variables that were omitted incorrectly (T2). The models are also stored in a list variable for assessing the results. The simulation results are shown for the first ten regression models.



The average T1 and T2 is calculated for each number of predictors and stored in result\_df\_group variable.





The top-left scatterplot describes the relationship between number of predictors and T1. A regression line is created using T1 and the number of predictors and fitted into the graph. The correlation coefficient is 0.67 which suggests that there is a strong positive linear relationship between number of predictors and T1.



The top-right scatterplot describes the relationship between number of predictors and T2. A regression line is created using T2 and the number of predictors and fitted into the graph. The correlation coefficient is 0.028 which suggests that there is a very weak linear relationship between number of predictors and T2.



The bottom-left scatterplot describes the relationship between number of predictors and T1 in result\_df\_group variable. A regression line is created using T1 and the number of predictors and fitted into the graph. The correlation coefficient is 0.885 which suggests that there is a very strong positive linear relationship between number of predictors and T1.



The bottom-right scatterplot describes the relationship between number of predictors and T2 in result\_df\_group variable. A regression line is created using T2 and the number of predictors and fitted into the graph. The correlation coefficient is 0.16 which suggests that there is a weak linear relationship between number of predictors and T2.



Due to computation issues, multiple simulations were not performed to fully understand the relationship between performance of stepwise regression model and number of predictors.

**References**

Flom, P.L., Development, N., Institutes, R., York, N., 2007. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use 7.

Lindsey, C., Sheather, S., 2010. Variable Selection in Linear Regression. Stata J. 10, 650–669. https://doi.org/10.1177/1536867X1101000407

Smith, G., 2018. Step away from stepwise. J. Big Data 5, 32. https://doi.org/10.1186/s40537-018-0143-6

Venables, B., Ripley, B., 2002. Modern Applied Statistics With S, in: Springer. https://doi.org/10.1007/b97626