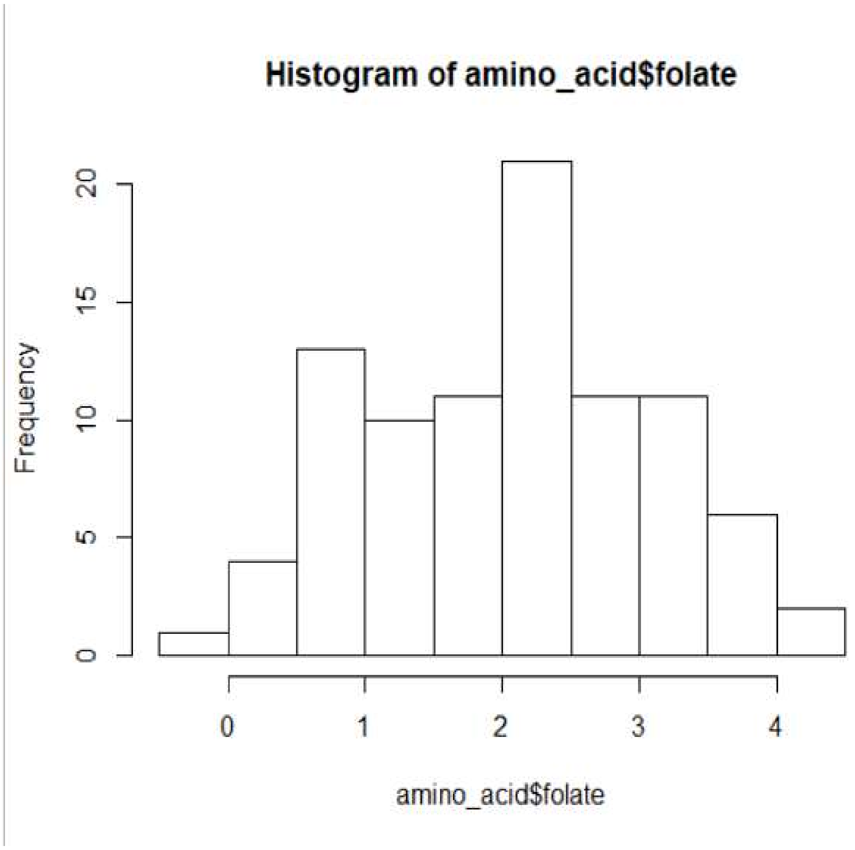
Name: Sumeet Shivgand
Student ID: R00182850
Subject: Statistical Data Analysis

Assignment - 1

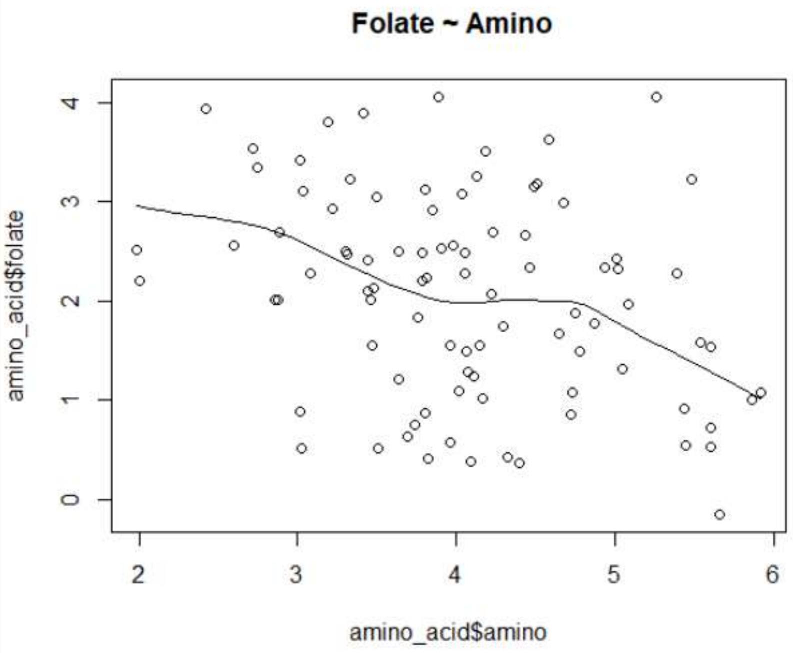
Question 1:

a. **Numerical and Graphical summary of the data.**



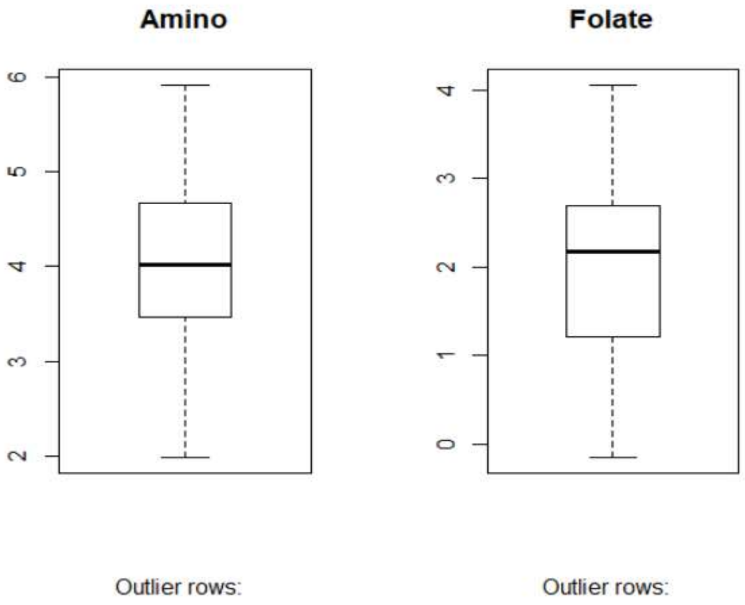
Histogram

The observations are roughly bell-shaped (more observations in the middle of the distribution, fewer on the tails), so we can proceed with the linear regression.



Scatterplot

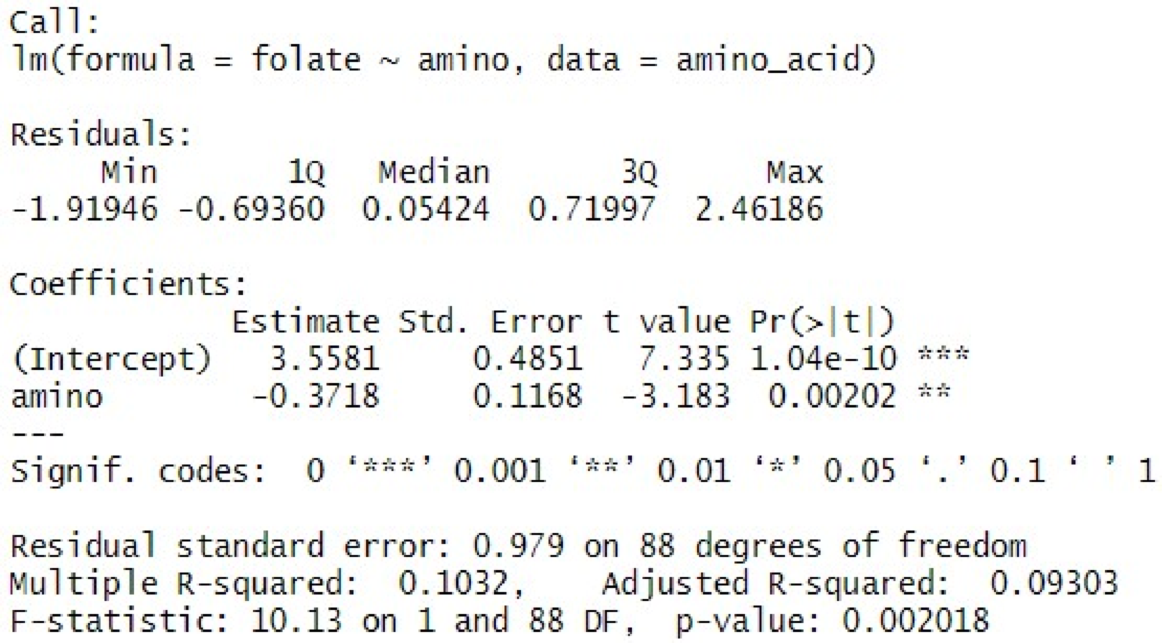
The scatter plot along with the smoothing line above suggests a linearly decreasing relationship between the 'folate' and 'amino' variables.



Boxplot

There is no outlier in the 'amino' and 'folate' variable.

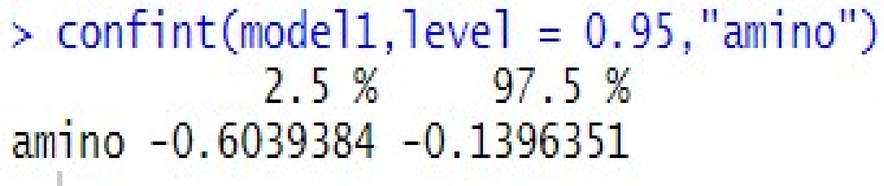
b. Fit a linear model



```
Call:
lm(formula = folate ~ amino, data = amino_acid)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.91946 -0.69360  0.05424  0.71997  2.46186 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.5581     0.4851   7.335 1.04e-10 ***
amino        -0.3718     0.1168  -3.183  0.00202 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.979 on 88 degrees of freedom
Multiple R-squared:  0.1032,	Adjusted R-squared:  0.09303 
F-statistic: 10.13 on 1 and 88 DF,  p-value: 0.002018
```

The fitted model is
$\hat{y} = 3.5581 - 0.3718x$
Interpretation of $\beta1$ coefficient: - $\beta1$ = - 0.3718
Each percentage increase in the ‘amino’ corresponds to a reduction of 0.3718 in ‘folate’.
The coefficient of determination $R^2 = 0.1032$ which tells us that 10.32% of the variability in ‘folate’ can be explained by ‘amino’.
The estimate for the error variance $\hat{\sigma}^2 = 0.979^2$

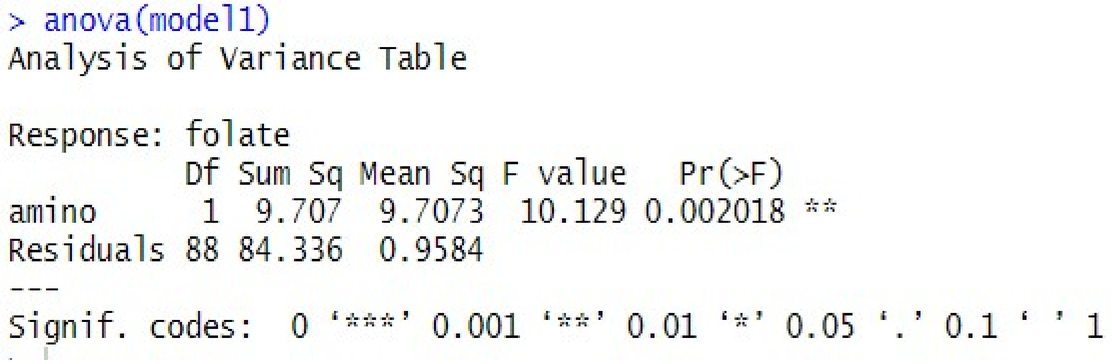
c. 95% confidence interval for the $\hat{\beta}_1$ coefficient.



```
> confint(model1,level = 0.95,"amino")
           2.5 %     97.5 %
amino -0.6039384 -0.1396351
```

Since this interval excludes 0, so we conclude that, there is a significant relationship between amino and folate.

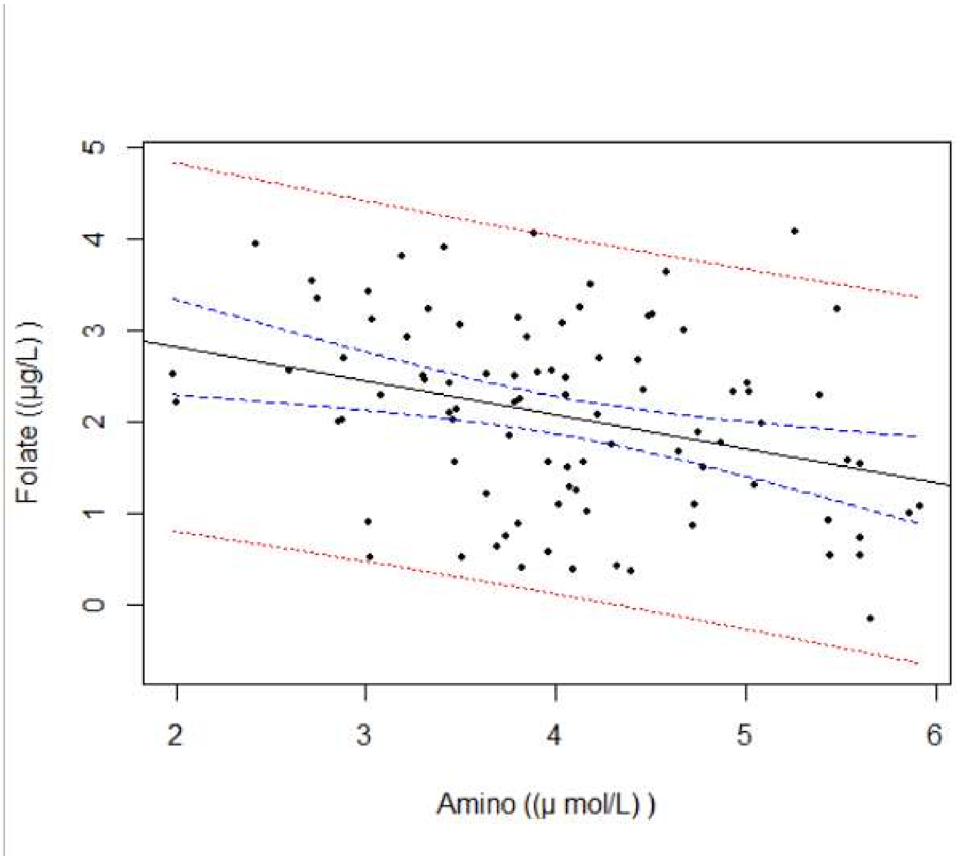
d. **Test the Hypothesis**
H0: $\beta_1 = 0$
HA: $\beta_1 \neq 0$



```
> anova(model1)
Analysis of Variance Table

Response: folate
          Df Sum Sq Mean Sq F value   Pr(>F)
amino      1  9.707  9.7073  10.129 0.002018 **
Residuals 88 84.336  0.9584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

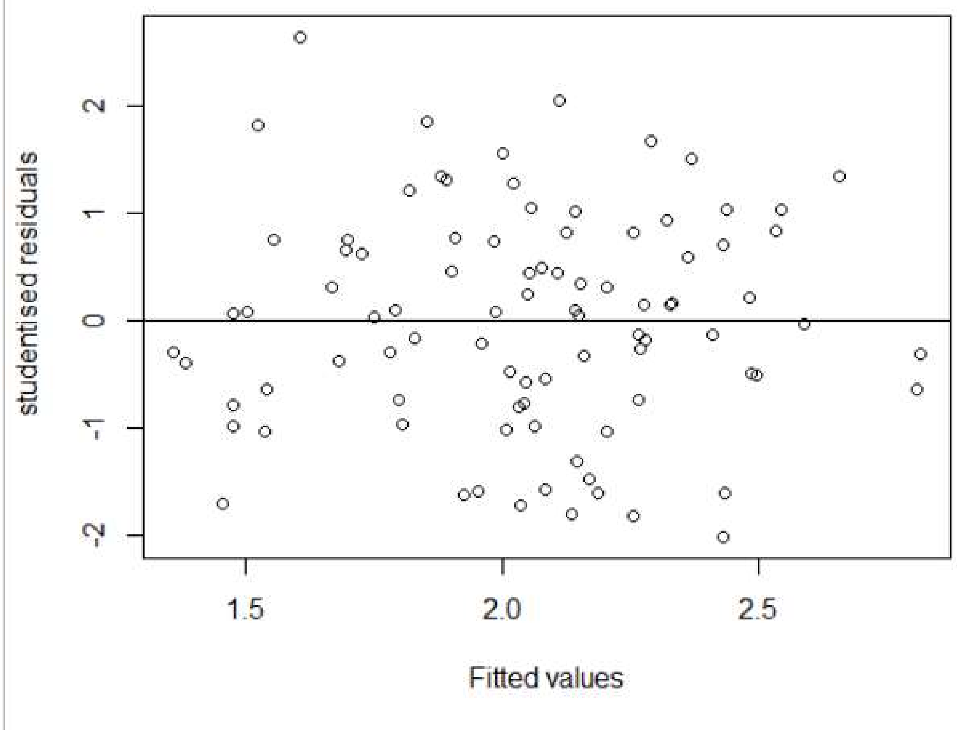
Examining the output for anova(model1), we see that the F-statistic is F (1, 88) = 10.129 and p =0.002. In this instance, we fail to reject the null hypothesis at the 5% confidence level and conclude that the slope β1 is not equal to zero.

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



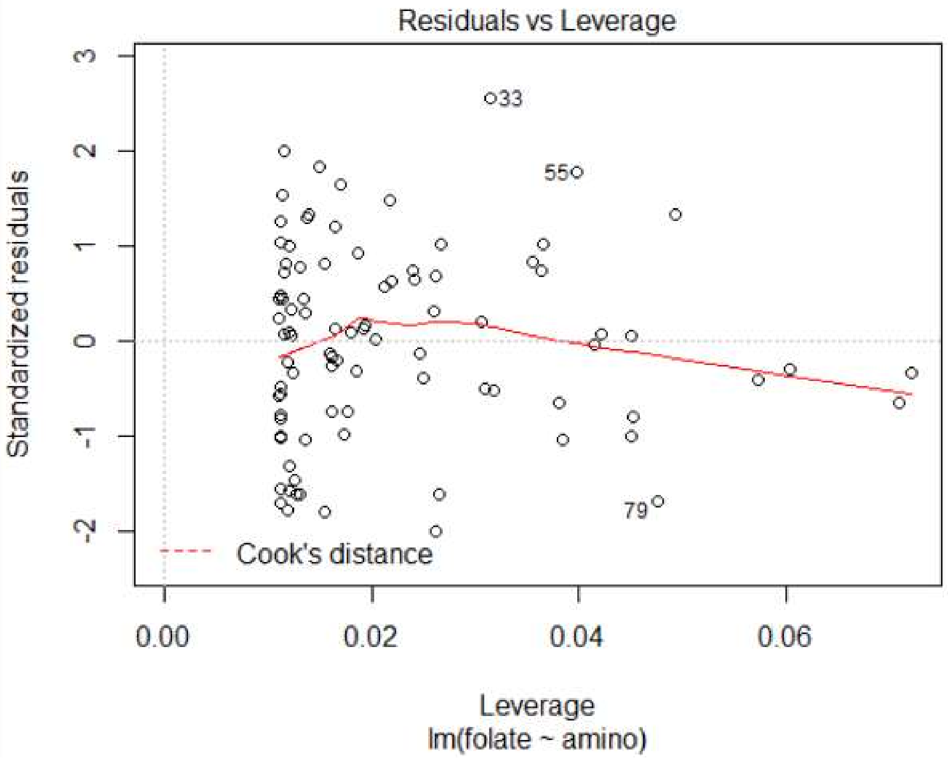
Here, we notice that the bands are a little bit narrowest at the center of the regression line.

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



Here, we see that points are scattered and then the horizontal line is passed through it. In the above plot, we see a non-linear pattern. There is no outlier or influence point.

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

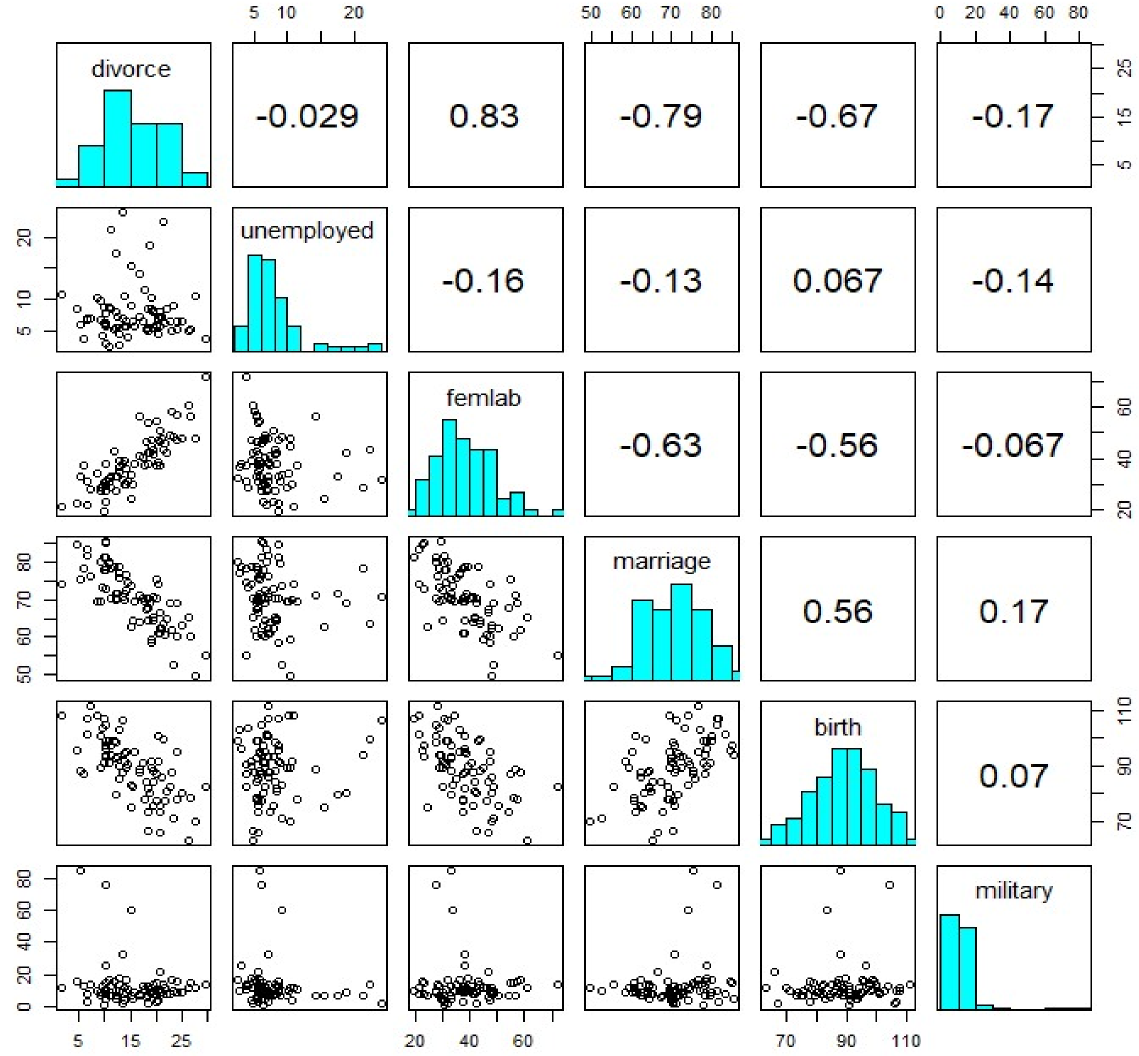


In this plot the dotted red lines are cook's distance and the areas of interest for us are the ones outside the dotted line. If any point falls in that region, we say that the observation has high leverage or potential for influencing our model.
Here we see that "Cook's distance" dashed curves don't even appear on the plot. So, none of the points come close to having both high residual and leverage.

h. Identify the observation that has the largest influence on the estimate of the $\hat{\beta}_1$ coefficient. Explain why this observation has a large influence.
We didn't find any observation that has the largest influence on the β1 coefficient. The plot above highlights the top 3 most extreme points (#33, #55, and #79), with a standardized residual between 1 & 3 and below -1.

Question 2:

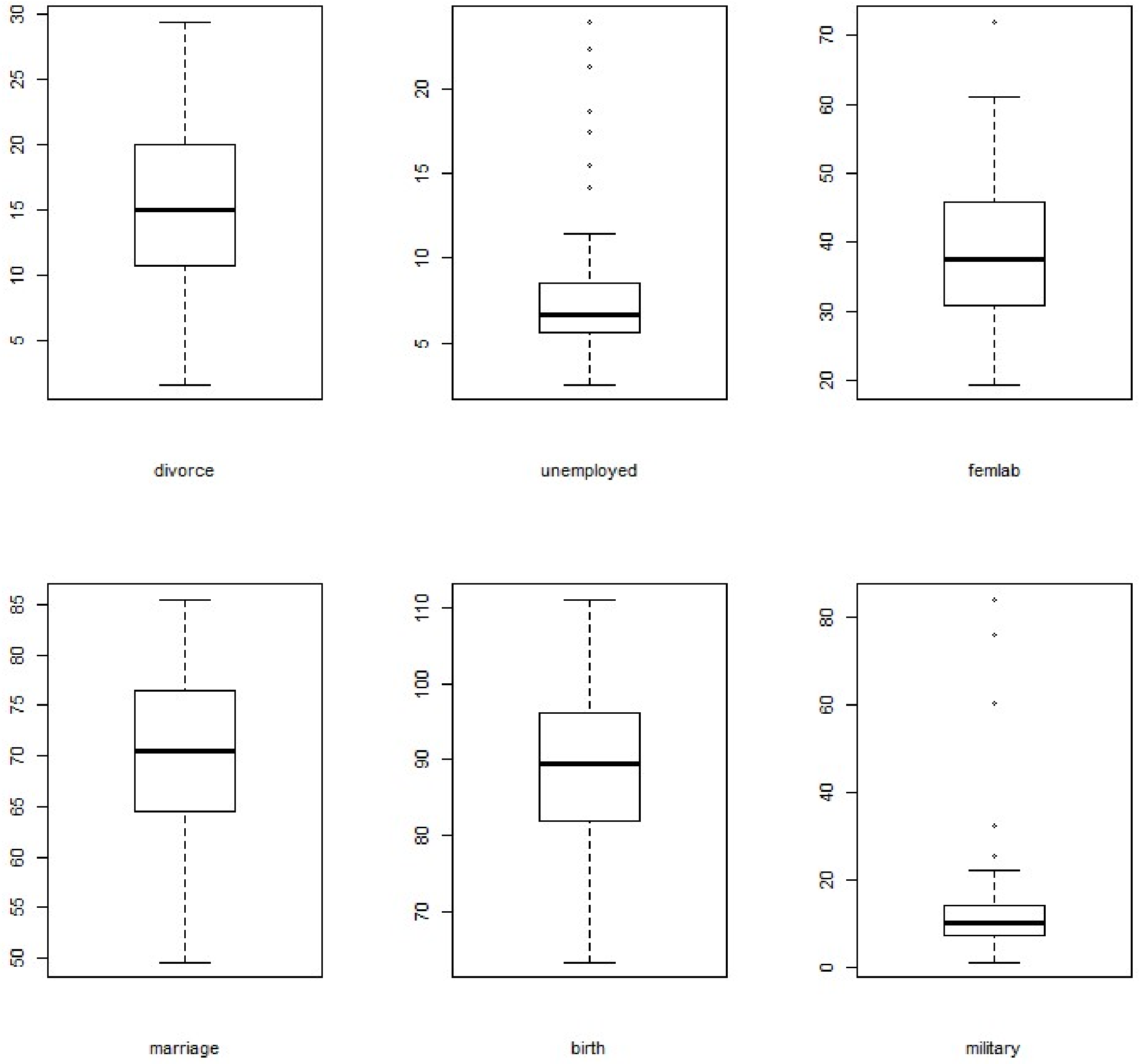
a. Numerical and Graphical summary of the data.



Pair Plots

Pair Plot:

- Note that the unemployed variable and military variables are skewed
- There is a positive correlation between divorce and femlab (r =0.83)
- There is a positive correlation between unemployed and birth (r =0.67)
- There is a positive correlation between marriage and birth (r =0.56)



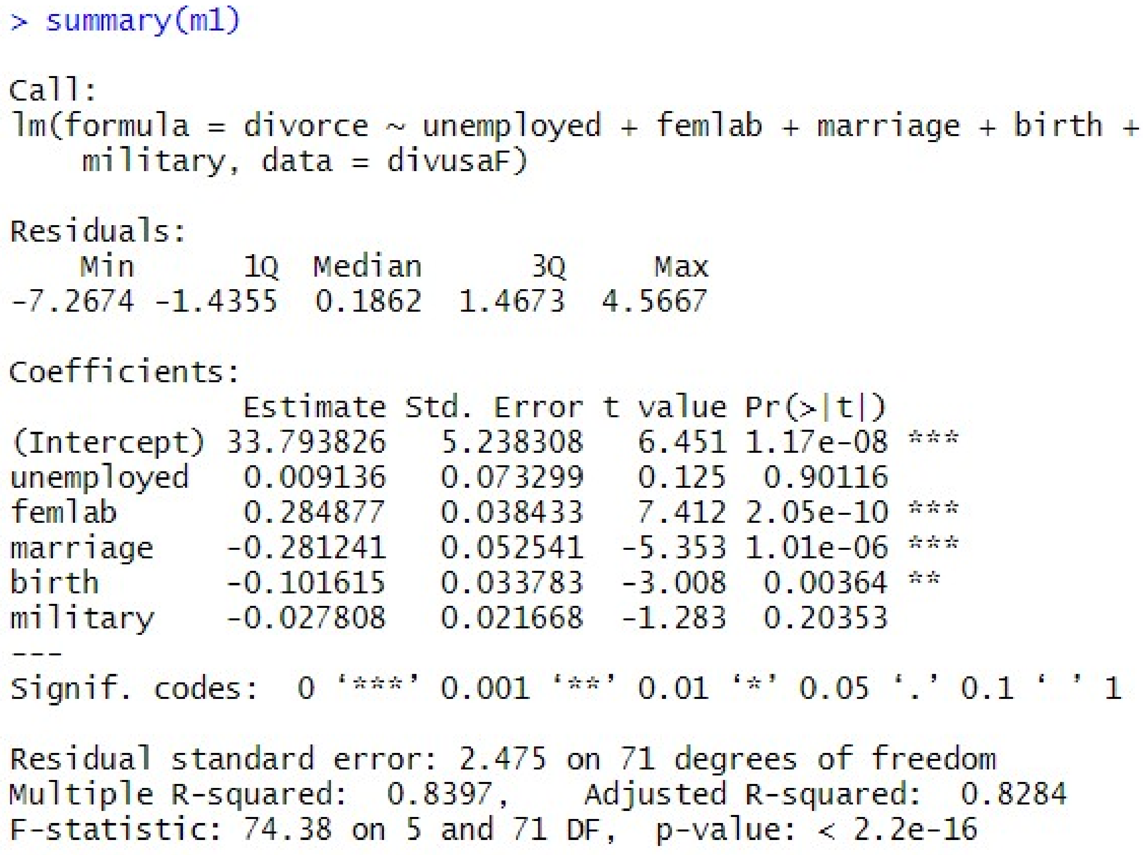
Boxplot

Boxplots:

- There are many outliers in the unemployed variable (this may be down to skew).
- There is one potential outlier in the femlab variable.
- There are number of outliers in the military variable (this may be down to skew)

b. **Fit the model :**

$$y = \beta_0 + \beta_1 unemployed + \beta_2 femlab + \beta_3 marriage + \beta_4 birth + \beta_3 military + e$$



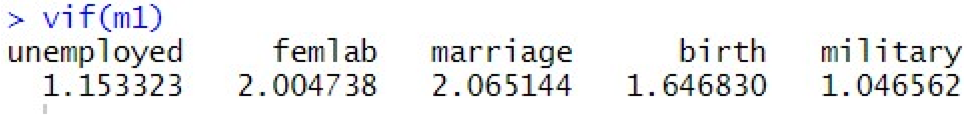
```
> summary(m1)

Call:
lm(formula = divorce ~ unemployed + femlab + marriage + birth +
    military, data = divusaF)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2674 -1.4355  0.1862  1.4673  4.5667

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.793826   5.238308   6.451 1.17e-08 ***
unemployed   0.009136   0.073299   0.125  0.90116
femlab       0.284877   0.038433   7.412 2.05e-10 ***
marriage    -0.281241   0.052541  -5.353 1.01e-06 ***
birth       -0.101615   0.033783  -3.008  0.00364 **
military    -0.027808   0.021668  -1.283  0.20353
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.475 on 71 degrees of freedom
Multiple R-squared:  0.8397,	Adjusted R-squared:  0.8284
F-statistic: 74.38 on 5 and 71 DF,  p-value: < 2.2e-16
```

i. **Interpret the coefficient for `femlab`.**

y = 33.793826 + 0.009136 unemployed + 0.284877 femlab - 0.281241 marriage - 0.101615 birth - 0.027808 military + e

coefficient for 'femlab'is 0.284877

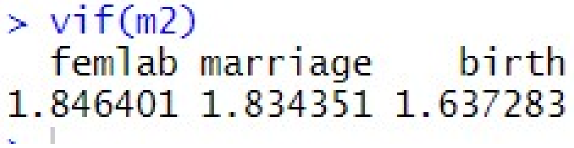
ii. **Calculate the variance inflation factors for this model and discuss their implications for collinearity in the model.**



```
> vif(m1)
unemployed     femlab   marriage      birth   military
  1.153323   2.004738   2.065144   1.646830   1.046562
```

The VIFs lie between 1 and 2.1 indicating that collinearity is not having a large impact on the coefficient estimates for this model.

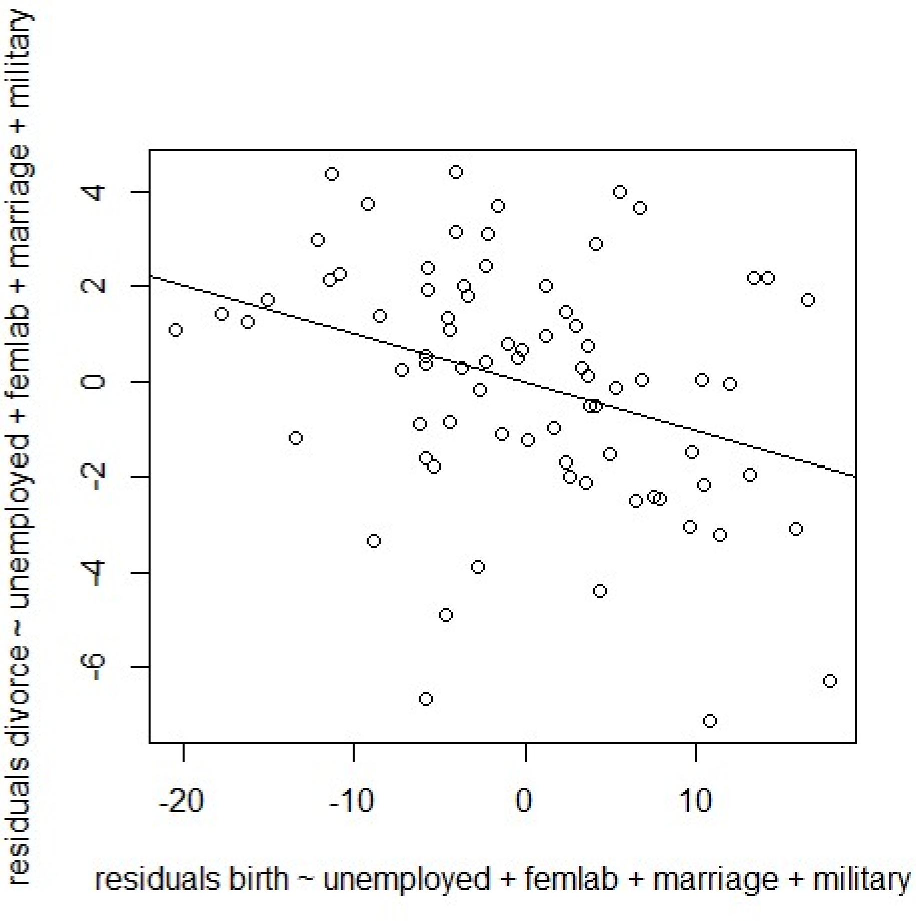
iii. **By fitting alternative models, determine whether collinearity has an impact on the coefficient estimates.**



```
> vif(m2)
  femlab marriage    birth
1.846401 1.834351 1.637283
```

For an alternative model, the VIFs lie between 1 and 2 indicating that collinearity is not having a large impact on the coefficient estimates for this model.

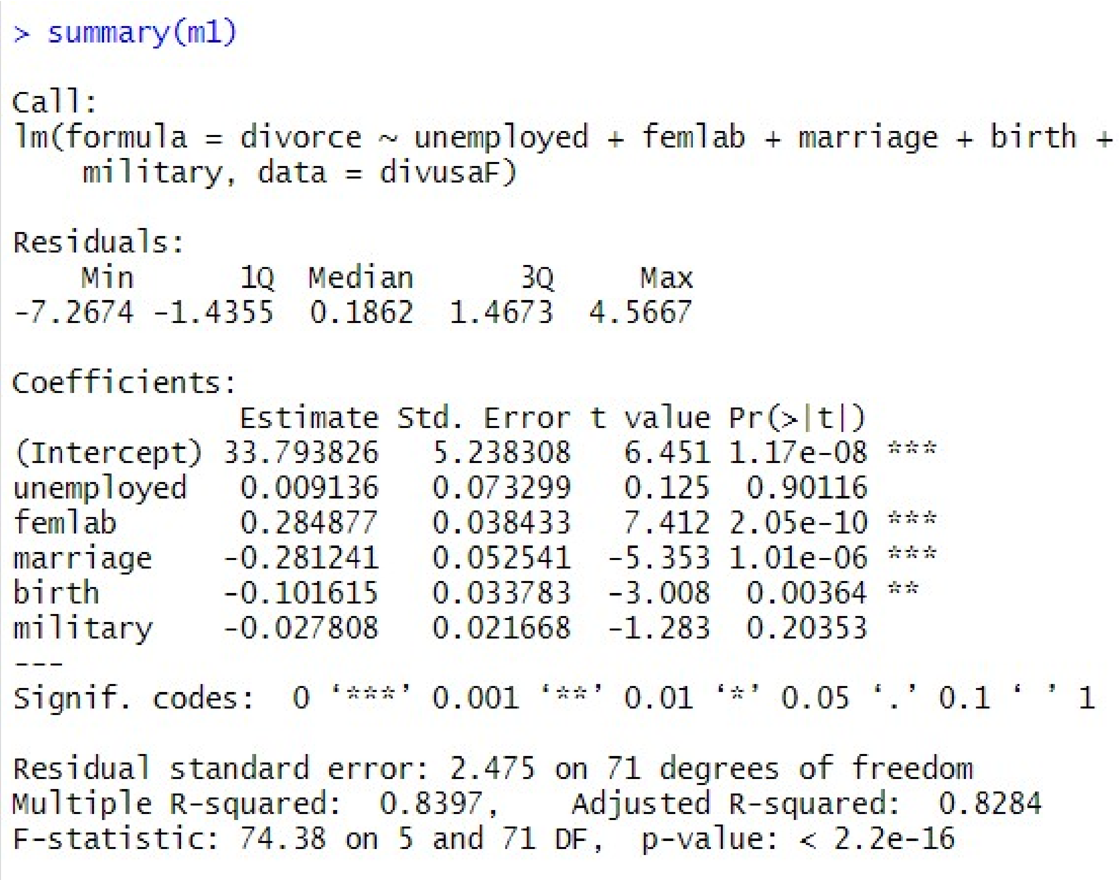
iv. **Create a partial regression plot to examine the relationship between birth and divorce adjusted for *unemployed, femlab, marriage,* and *military*.**



The slope of the regression line is the estimate for the beta coefficient associated with birth in the model containing unemployed, femlab, marriage, and military and birth.

v. **Test the hypothesis:**

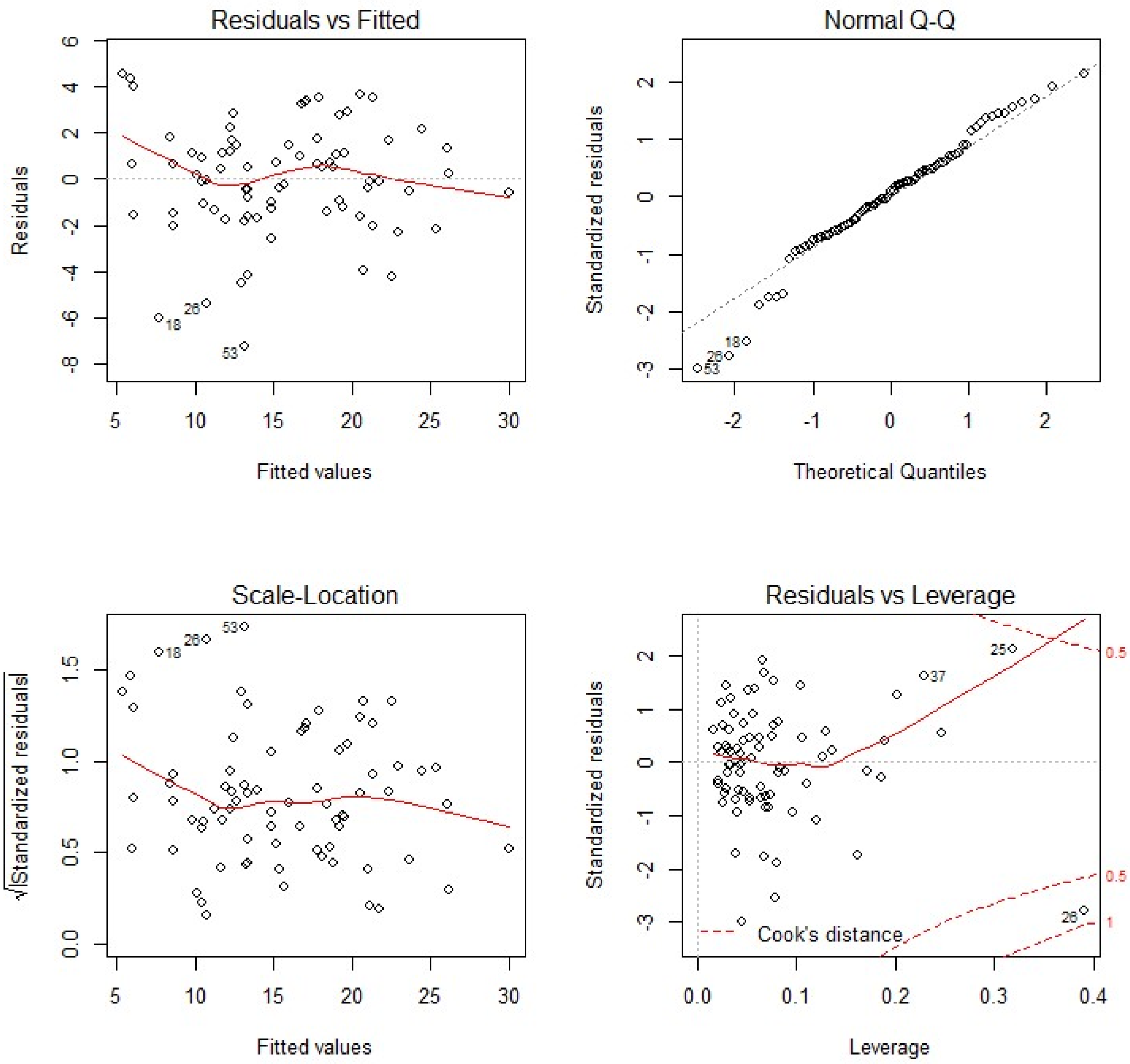
H$_0$: $\boldsymbol{\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0}$
H$_A$: **at least one of the** $\boldsymbol{\beta_i \neq 0}$
What do the results of the hypothesis test imply for the regression model?



```
> summary(m1)

Call:
lm(formula = divorce ~ unemployed + femlab + marriage + birth +
    military, data = divusaF)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2674 -1.4355  0.1862  1.4673  4.5667

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.793826   5.238308   6.451 1.17e-08 ***
unemployed   0.009136   0.073299   0.125  0.90116
femlab       0.284877   0.038433   7.412 2.05e-10 ***
marriage    -0.281241   0.052541  -5.353 1.01e-06 ***
birth       -0.101615   0.033783  -3.008  0.00364 **
military    -0.027808   0.021668  -1.283  0.20353
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.475 on 71 degrees of freedom
Multiple R-squared:  0.8397,	Adjusted R-squared:  0.8284
F-statistic: 74.38 on 5 and 71 DF,  p-value: < 2.2e-16
```

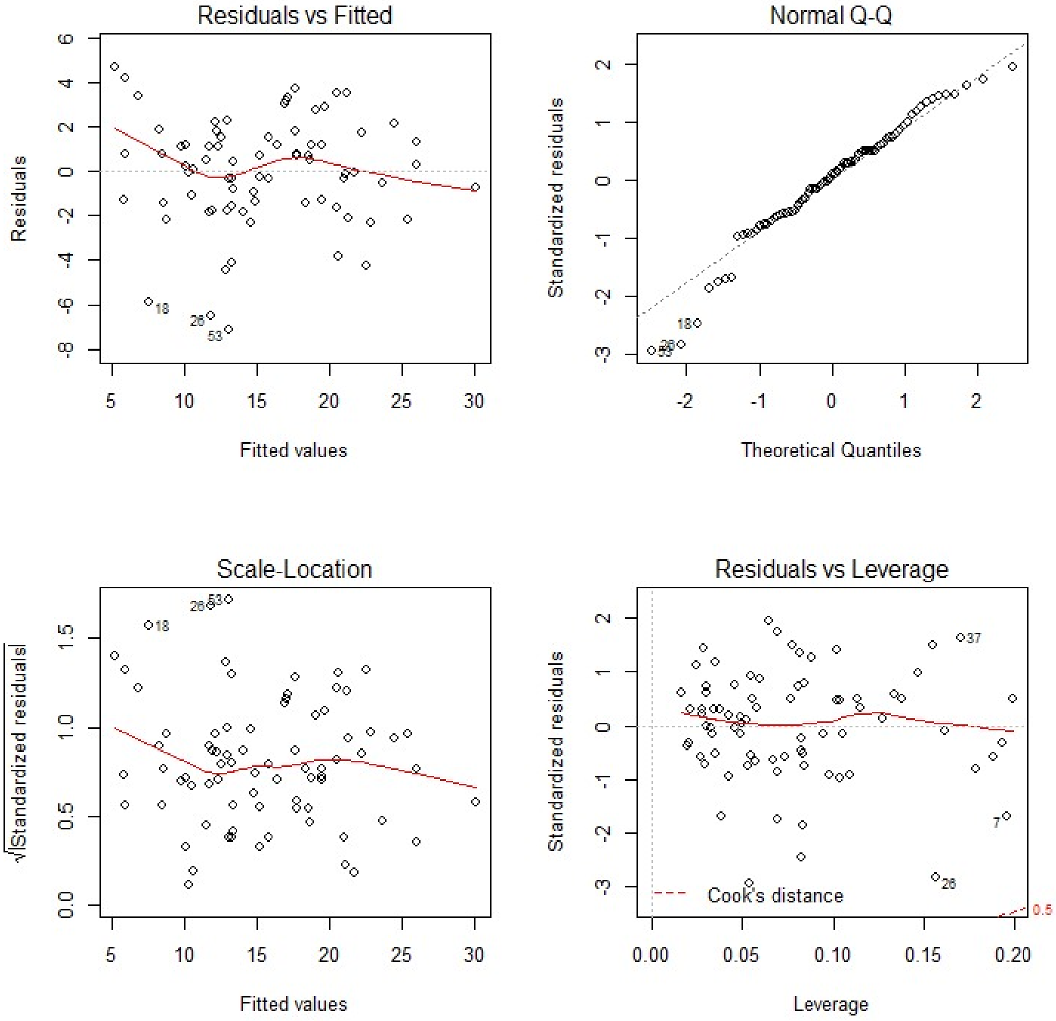
Here, we see that the global F-statistic is F (5, 71) = 74.38 and p = (< 2.2e-16), this F-statistic compares the fitted model to the null model (also called the intercept only model). In this instance, we may reject the null hypothesis at the 1% confidence level and conclude that at least one of the predictors is associated with divorce.

vi. Assess the fit of the model using diagnostic plots, commenting on the assumptions of the regression model and influential points.



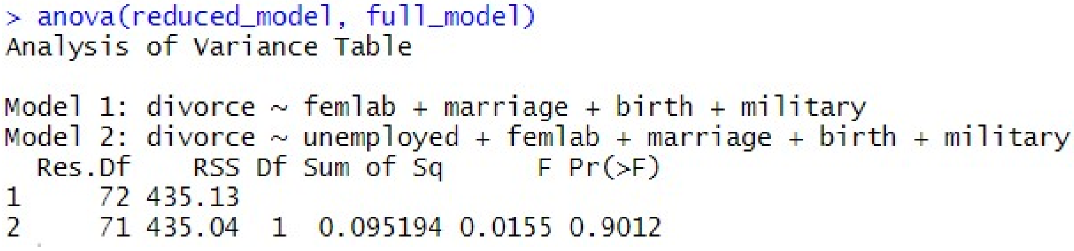
The plot of residuals vs. fitted shows that the residuals do not seem to be random, many of the residuals are clustered together in the middle-fitted values - we should probably address the skew we observed when examining the histograms and scatterplots. Also, we note some observations with high leverage.

Transforming the unemployed and military variable as there is skewness.



Now, the plot of residuals vs. fitted shows that the residuals seem to be random and there is no skewness in the data. Further, we also see that there is no observation with high leverage.

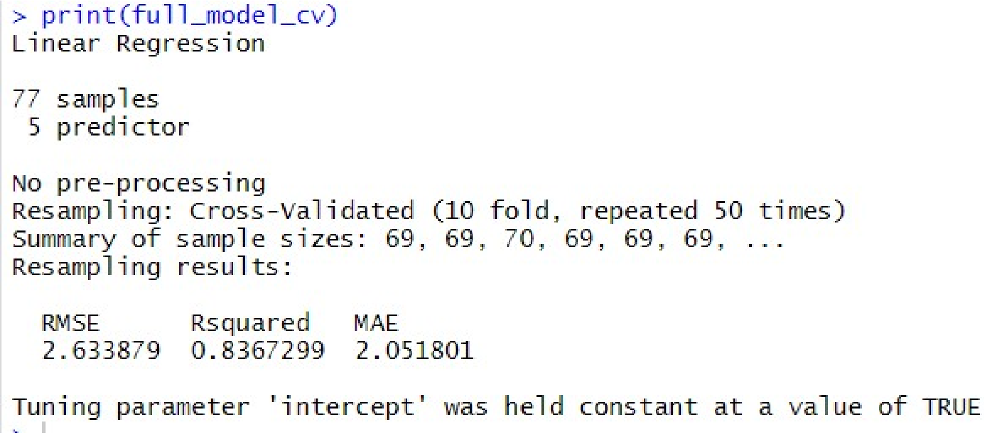
c. **F-test to compare the full model to the model including all variables except unemployment.**



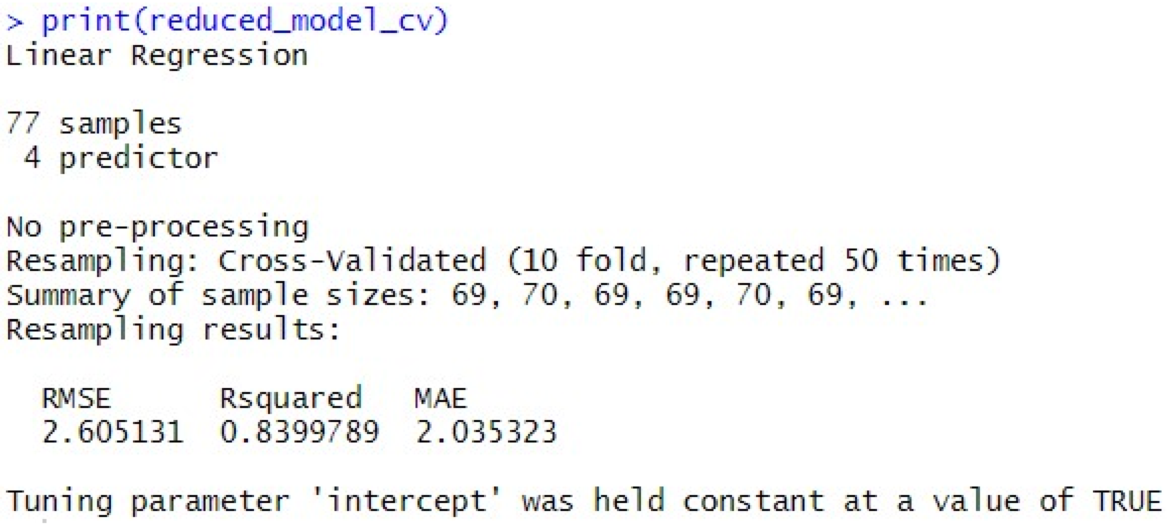
```
> anova(reduced_model, full_model)
Analysis of Variance Table

Model 1: divorce ~ femlab + marriage + birth + military
Model 2: divorce ~ unemployed + femlab + marriage + birth + military
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     72 435.13
2     71 435.04  1  0.095194 0.0155 0.9012
```

From above output, we say that the F-statistic is F(1, 71) = 0.155 and p =0.9012. In this instance we fail to reject the null hypothesis at the 5% confidence level and conclude that the data does not provide evidence.

d. **Compare the predictive accuracy of the above two models using 50 repeats of 10-fold cross-validation.**



```
> print(full_model_cv)
Linear Regression

77 samples
 5 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 50 times)
Summary of sample sizes: 69, 69, 70, 69, 69, 69, ...
Resampling results:

  RMSE      Rsquared   MAE
  2.633879  0.8367299  2.051801

Tuning parameter 'intercept' was held constant at a value of TRUE
```



```
> print(reduced_model_cv)
Linear Regression

77 samples
 4 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 50 times)
Summary of sample sizes: 69, 70, 69, 69, 70, 69, ...
Resampling results:

  RMSE      Rsquared   MAE
  2.605131  0.8399789  2.035323

Tuning parameter 'intercept' was held constant at a value of TRUE
```

We had performed the cross-validation on two models i.e. full model and reduced model. Here, we see that, RMSE for the reduced model is lower than the full model. So, we can say that predictive accuracy for the 'reduced model' is better than 'full model'.

e. Can you suggest any improvements to the model?

To improve the model, we must check if we are missing any assumptions. If yes, then that might be contributing to the deterioration of performance.

Question 3

1. **Stepwise regression**

Stepwise regression is a popular data-mining tool that uses statistical significance to select the explanatory variables to be used in a multiple-regression model. Stepwise regression is a type of regression technique that builds a model by adding or removing the predictor variables, generally via a series of T-tests or F-tests. The variables, which need to be added or removed are chosen based on the test statistics of the coefficients estimated. Unlike other regression models, stepwise regression needs proper attention and only a skilled researcher who is familiar with statistical testing should perform it. There are three different types of stepwise regression such as forward selection, backward elimination, and bi-directional elimination.

The **forward selection** method is simple to define. A forward selection rule starts with no explanatory variables and then adds variables, one by one in the model based on a variable that is the most statistically significant. Select the variable that has the highest R-Squared. Stop adding variables when none of the remaining variables are significant. Note that once a variable enters the model, it cannot be deleted.

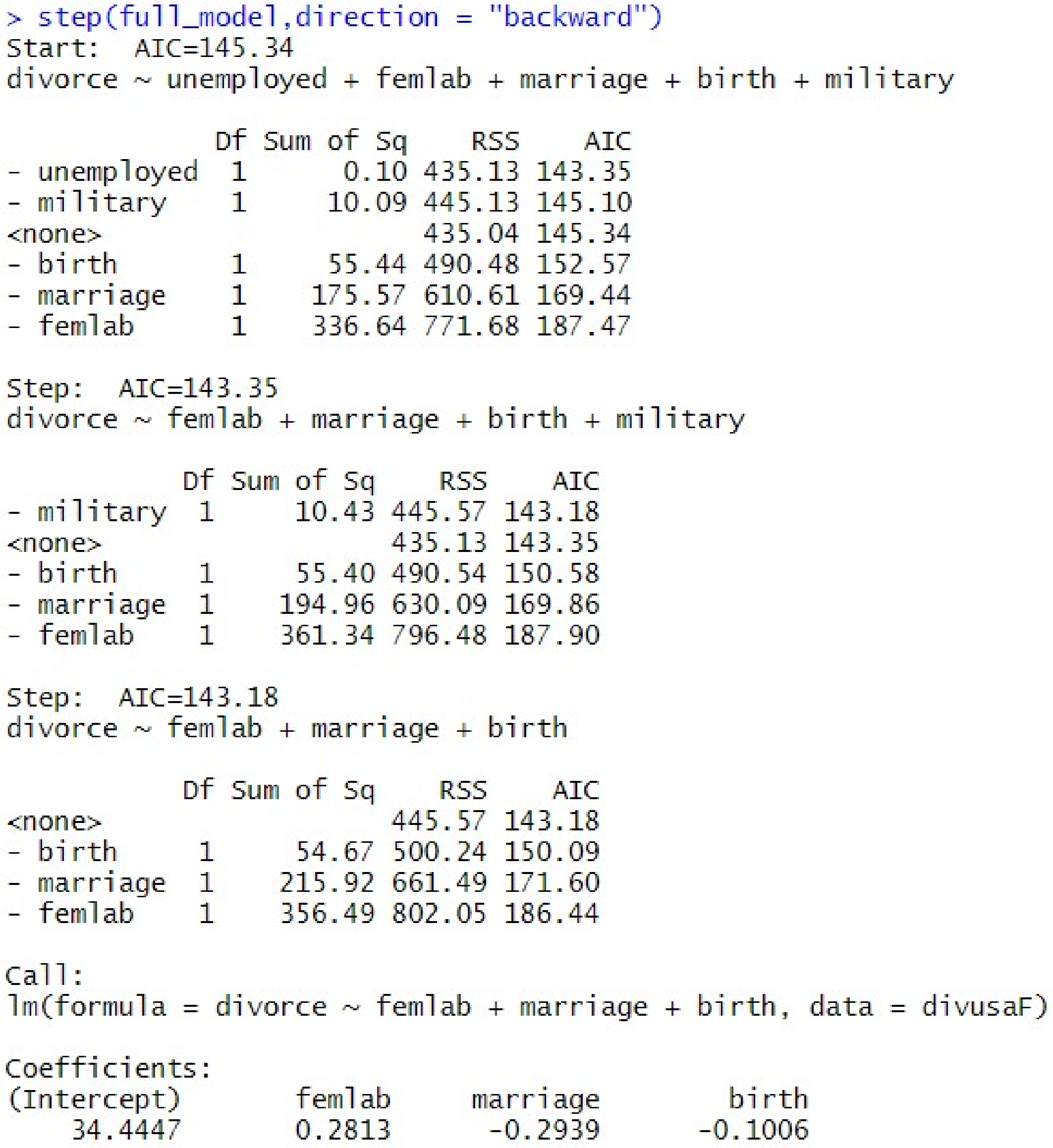
A **backward elimination** is less popular because it begins with a model in which all explanatory variables have been included. The user sets the significance level at which variables can enter the model. At each step, the variable that is the least significant is removed. This process continues until no nonsignificant variables remain. This method is challenging if there are many candidate variables and impossible if the number of candidate variables is larger than the number of observations. The user sets the significance level at which variables can be removed from the model.

A **bi-directional elimination** procedure is a combination of forward selection and backward elimination. It is essentially a forward selection procedure but with the possibility of deleting a selected variable at each stage, as in the backward elimination, when there are correlations between variables. It is often used as a default approach.

One problem associated with stepwise regression is ‘Overfitting’. Overfitting is where your model is too complex for your data — it happens when your sample size is too small. If you put enough predictor variables in your regression model, you will nearly always get a model that looks significant.

The easiest way to avoid overfitting is to increase your sample size by collecting more data. If you can’t do that, the second option is to reduce the number of predictors in your model — either by combining or eliminating them. An overfit model result in misleading regression coefficients, p-values, and R-squared statistics. Factor Analysis is one method you can use to identify related predictors that might be candidates for combining.

2. **Stepwise regression function**



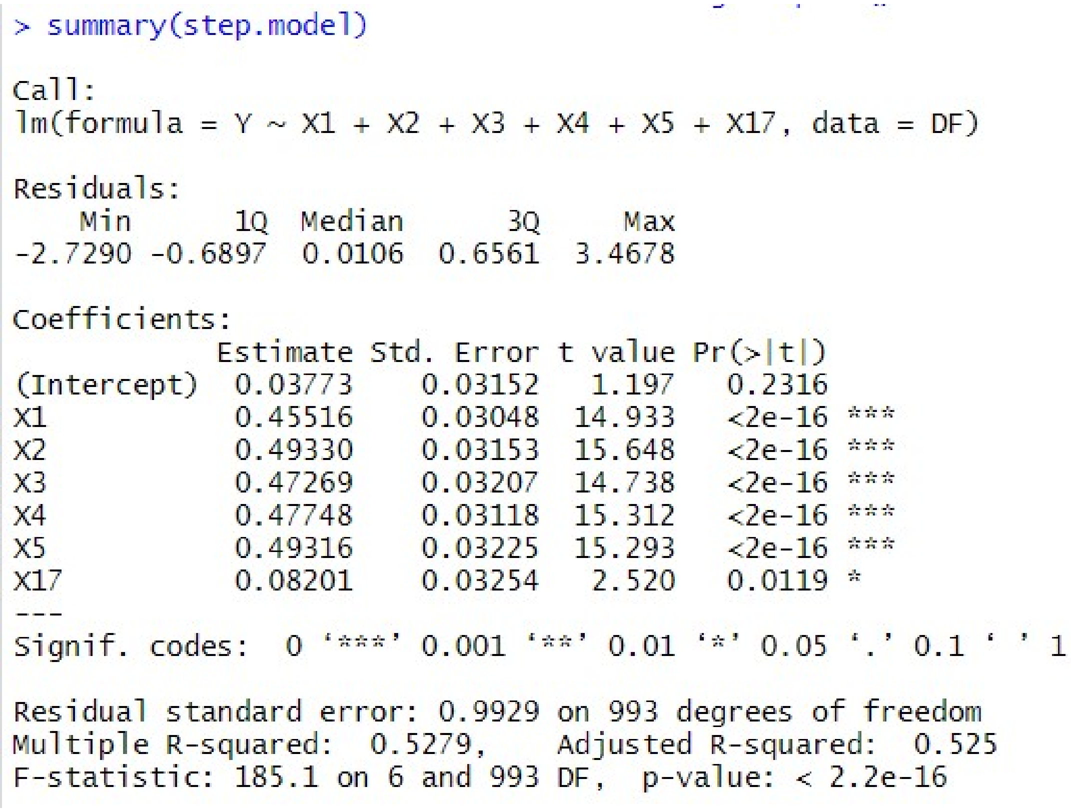
```
> step(full_model,direction = "backward")
Start:  AIC=145.34
divorce ~ unemployed + femlab + marriage + birth + military

             Df Sum of Sq    RSS    AIC
- unemployed  1      0.10 435.13 143.35
- military    1     10.09 445.13 145.10
<none>                    435.04 145.34
- birth       1     55.44 490.48 152.57
- marriage    1    175.57 610.61 169.44
- femlab      1    336.64 771.68 187.47

Step:  AIC=143.35
divorce ~ femlab + marriage + birth + military

           Df Sum of Sq    RSS    AIC
- military  1     10.43 445.57 143.18
<none>                  435.13 143.35
- birth     1     55.40 490.54 150.58
- marriage  1    194.96 630.09 169.86
- femlab    1    361.34 796.48 187.90

Step:  AIC=143.18
divorce ~ femlab + marriage + birth

           Df Sum of Sq    RSS    AIC
<none>                  445.57 143.18
- birth     1     54.67 500.24 150.09
- marriage  1    215.92 661.49 171.60
- femlab    1    356.49 802.05 186.44

Call:
lm(formula = divorce ~ femlab + marriage + birth, data = divusaF)

Coefficients:
(Intercept)       femlab     marriage        birth  
    34.4447       0.2813      -0.2939      -0.1006  
```

Here, we are using backward elimination as all the explanatory variables are included. The variable that is least significant is removed and this process continues until there will be no nonsignificant variables remain. We are going to select ‘AIC’ (Akaike Information Criterion) as this will be a key thing in this model. Lower the AIC value better would be the model.

Working of algorithm:

i. Fit the model with all possible predictors.
ii. Consider the predictor with the lowest AIC value. If AIC is less, then go to point (iii).
iii. Remove the predictor.
iv. Fit the model without this variable and repeat the step (ii) until the condition becomes false.

3. **Simulations to explore the performance of stepwise regression for model selection.** Here, we create a data set with 20 predictors, out of which 5 are linearly related to the outcome Y and 15 of which are noise. In regression analysis, the signal to noise ratio is an important consideration i.e. the magnitude of the effect size (the β coefficients) to the variance. To start, set the β coefficients that are linearly related to the outcome to 0.5 and set the variance of the error term to 1. We create a data set with 1000 observations.



```
> summary(step.model)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X17, data = DF)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7290 -0.6897  0.0106  0.6561  3.4678 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.03773    0.03152   1.197   0.2316    
X1           0.45516    0.03048  14.933   <2e-16 ***
X2           0.49330    0.03153  15.648   <2e-16 ***
X3           0.47269    0.03207  14.738   <2e-16 ***
X4           0.47748    0.03118  15.312   <2e-16 ***
X5           0.49316    0.03225  15.293   <2e-16 ***
X17          0.08201    0.03254   2.520   0.0119 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9929 on 993 degrees of freedom
Multiple R-squared:  0.5279,	Adjusted R-squared:  0.525 
F-statistic: 185.1 on 6 and 993 DF,  p-value: < 2.2e-16
```

The model has correctly retained variables X1...X5 but has incorrectly included variables X17. If we examine the estimated coefficients, we see that the estimates for the β coefficients (β1 to β5) are close to their true value of 0.5 and the estimate for β17 is 0.08201. If we let T1 denote the number of variables that were retained incorrectly and T2 represents the number of variables that were omitted incorrectly then in this example, T1 = 1 and T2 = 0.

Compare the performance of stepwise regression with LASSO regression

Stepwise regression:

Stepwise regression is a combination of the forward and backward selection techniques and variable selection is carried out by automatic method. All predictor variables in the model are checked to see if their significance has been reduced below the specified tolerance level, it added in the final model. If a nonsignificant variable is found, then it is removed from the model Stepwise regression assumes that the predictor variables are not highly correlated.

One of the main issues with stepwise regression is that it searches a large space of possible models. Hence it is prone to overfitting the data. Another issue is that when estimating the degrees of freedom, the number of the candidate independent variables from the best fit selected may be smaller than the total number of final model variables, causing the fit to appear better than it is when adjusting the r^2 value for the number of degrees of freedom.

Lasso regression:

Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces.

Regularization techniques are used to prevent statistical overfitting in a predictive model. In general statistics, the number of observations should be greater than the number of variables, where the number of variables is bigger than the number of observations, the selection of variable gains greater importance. In this situation, Lasso regression can perform with some limitations as compared to stepwise regression.

The Lasso sets a constraint on the sum of the absolute values of the model parameters; the sum must be less than a fixed value (upper bound). To do so the method applies a shrinking process where it penalizes the coefficients of the regression variables shrinking some of them to zero. During the variable selection process the variables that still have a non-zero coefficient after the shrinking process are selected to be part of the model. The purpose of this process is to minimize the prediction error.