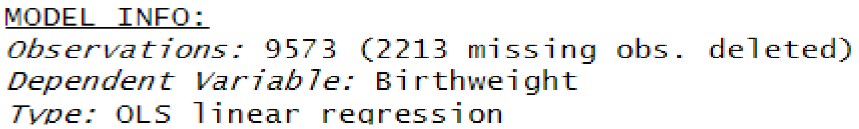
1. Multiple Linear Regression

Aim: To study the impact of maternal characteristics and lifestyle on infant weight using regression modeling.

1.1 Model Summary:

The dependent variable in the linear regression model is Birthweight, and the independent variables are Age, Social class, BMI, Folate, VitC, VitB6, VitB12, and VitE.

- **Model Info:**

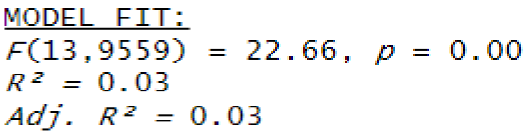


```
MODEL INFO:
Observations: 9573 (2213 missing obs. deleted)
Dependent Variable: Birthweight
Type: OLS linear regression
```

Fig.1

The initial dataset had 9573 observations. However, 2213 observations with missing values were deleted, leaving 7360 observations for analysis.

- **Model Fit:**



```
MODEL FIT:
F(13,9559) = 22.66, p = 0.00
R² = 0.03
Adj. R² = 0.03
```

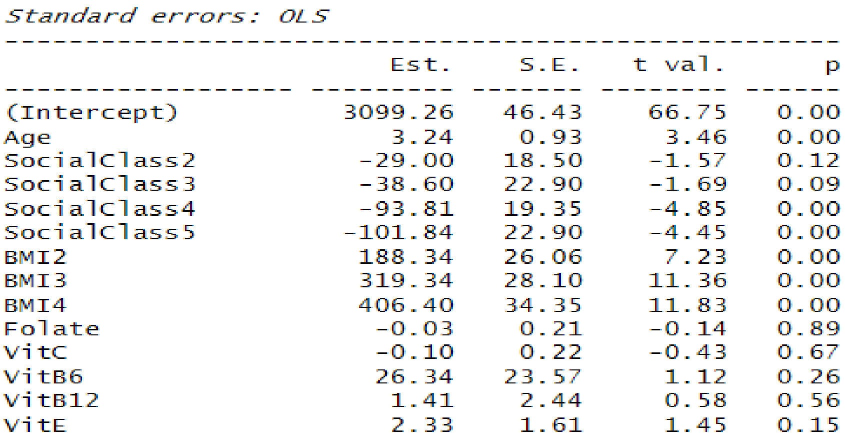
Fig.2

The F-statistic (13,9559) = 22.66 tests the overall significance of the regression model. The low p-value (0.00) indicates that the independent variable is significantly related to the dependent variable.

R-squared (R^2) = 0.03 is the percentage of the dependent variable's variance that can be predicted based on the independent variables. In this model, only 3% of the variance in birthweight is explained by the independent variables.

Adjusted R-squared = 0.03 is similar to R^2 but adjusted for the number of predictors in the model.

- **Standard errors:**



```
Standard errors: OLS
--------------------------------------------------
                          Est.     S.E.    t val.      p
------------------ --------- ------- -------- ------
(Intercept)             3099.26    46.43     66.75   0.00
Age                        3.24     0.93      3.46   0.00
SocialClass2             -29.00    18.50     -1.57   0.12
SocialClass3             -38.60    22.90     -1.69   0.09
SocialClass4             -93.81    19.35     -4.85   0.00
SocialClass5            -101.84    22.90     -4.45   0.00
BMI2                     188.34    26.06      7.23   0.00
BMI3                     319.34    28.10     11.36   0.00
BMI4                     406.40    34.35     11.83   0.00
Folate                    -0.03     0.21     -0.14   0.89
VitC                      -0.10     0.22     -0.43   0.67
VitB6                     26.34    23.57      1.12   0.26
VitB12                     1.41     2.44      0.58   0.56
VitE                       2.33     1.61      1.45   0.15
```

Fig.3

- When all independent variables are zero, the estimated birthweight is represented by the intercept. In this instance, the standard error is 46.43 and the intercept is 3099.26. Given the extremely high t-value (66.75) and the p-value of 0.00 indicates a very significant predictor.
- For each one-unit increase in Age, the Birthweight is expected to increase by 3.24 units. The standard error (0.93) indicates the precision of this estimate. The t-value (3.46) is relatively high, and the p-value (0.00) is less than 0.05, indicating that Age is a significant predictor of birth weight.
- Social class 2 and 3 have a coefficient of -29.00 and -38.60 which has a negative impact and is associated with the decrease in the birthweight, it is not a significant predictor of the birthweight as the p-value > 0.05. Social class 4 and 5 have the maximum negative impact on the birthweight with coefficients -93.81 and -101.84 and with p-value<0.05 indicating that it is a significant predictor of the birthweight.
- For BMI2, each unit increase corresponds to a significant positive impact on birthweight (coefficient 188.34, t-value 7.23, p < 0.05). BMI3 and BMI4 exhibit even larger positive impacts on birthweight, with coefficients of 319.34 and 406.40, respectively. Both BMI3 and BMI4 are highly significant predictors because of low p-values (0.00), suggesting that it is highly significant.
- Folate shows a small negative impact on birthweight with a coefficient of -0.03. However, the standard error (0.21) is relatively low, indicating a good estimate. The t-value (-0.14) and p-value (0.89) suggest that Folate is not a significant predictor of birth weight.
- Vitamin C (VitC) exhibits a negative coefficient (-0.10), indicating a decrease in birthweight with higher VitC levels. However, the t-value (-0.43) and p-value (0.67) suggest that this relationship is not statistically significant. Vitamin B6 (VitB6) has a positive coefficient (26.34), but the relatively high standard error (23.57) results in a non-significant predictor (t-value 1.12, p > 0.05).
- Vitamin B12 (VitB12) shows a positive impact on birthweight with a small coefficient of 1.41. However, the t-value (0.58) and p-value (0.56) suggest that this relationship is not statistically significant. Vitamin E (VitE) has a positive coefficient (2.33) but is not significantly different from zero, as indicated by the t-value (1.45) and p-value (0.15). Both VitB12 and VitE do not emerge as significant predictors of birthweight in this model.

1.2 Assumptions of the model:

Linearity:

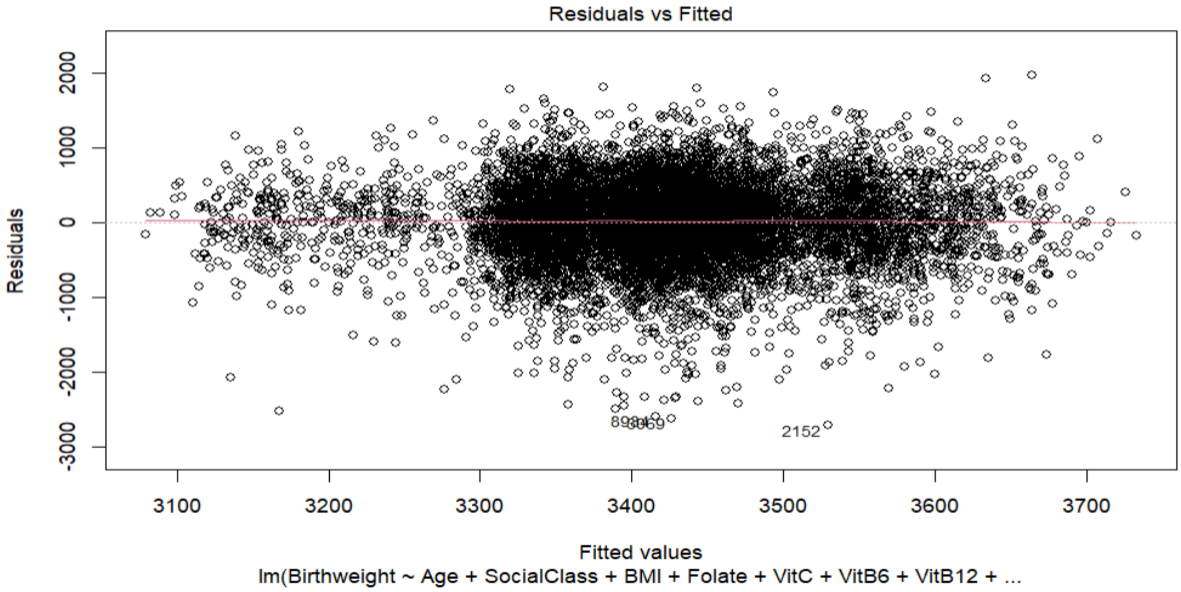


Fig.4

The assumption of linearity is met as the horizontal red line is straight and is not deviating at any point along the regression line. The points in the plot are randomly distributed along the line without deviating from the line which meets the linearity assumptions.

Normality (Q-Q Plot):

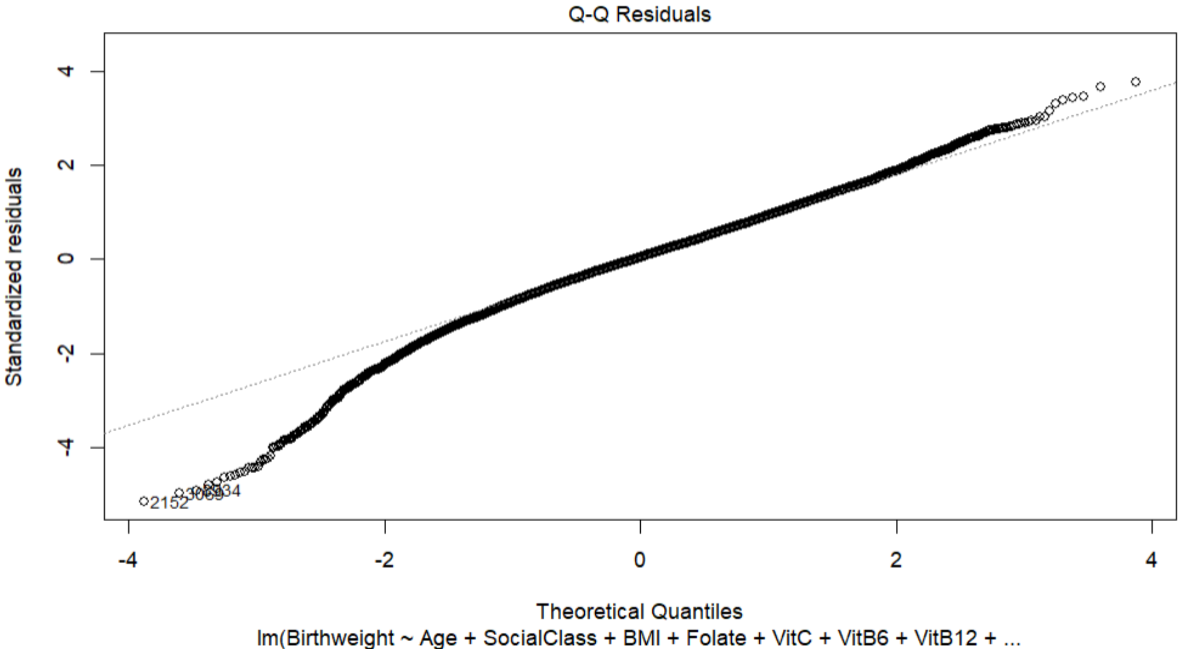


Fig.5

It can be observed that not all of the data points follow along the diagonal line with many outliers, proving that the model does not have a normal distribution.

Homoscedasticity and Heteroscedasticity:

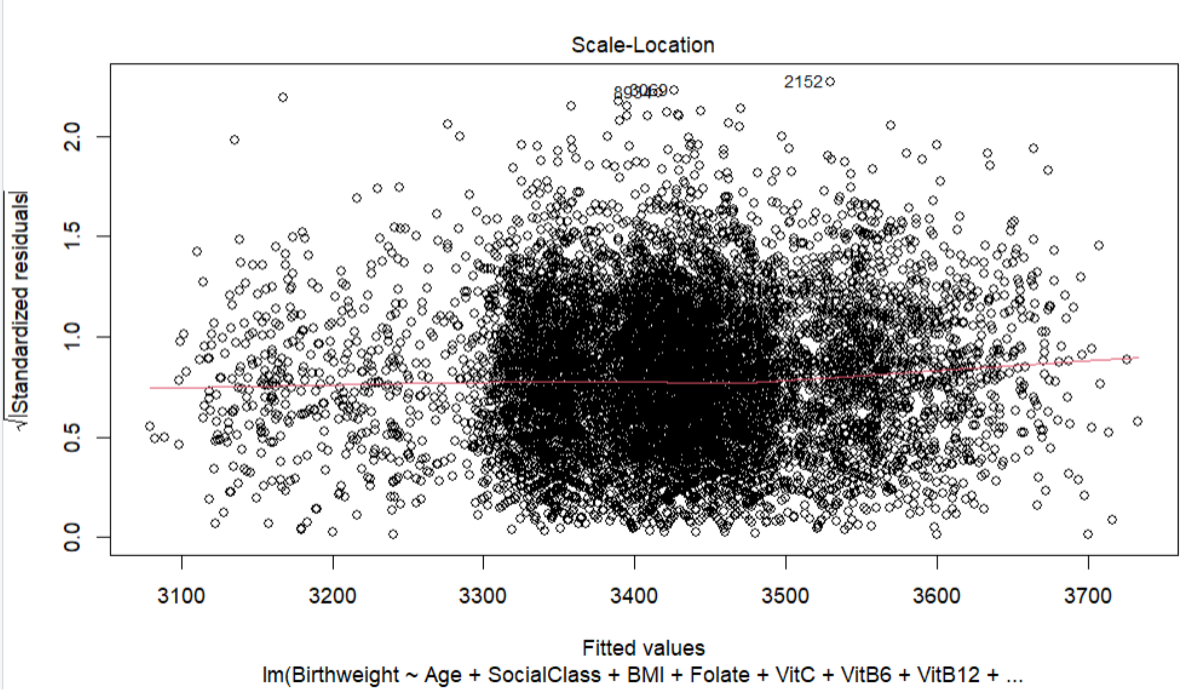


Fig.6

- The spread of the observations is not constant across all levels of fitted values along the regression line, which means that the plot does not follow homoscedasticity.
- Since there is no funnel shape formed during the spread of residuals, this suggests that the conditions of heteroscedasticity are not met as well.

1.3 Influential observations

Cooks distance:

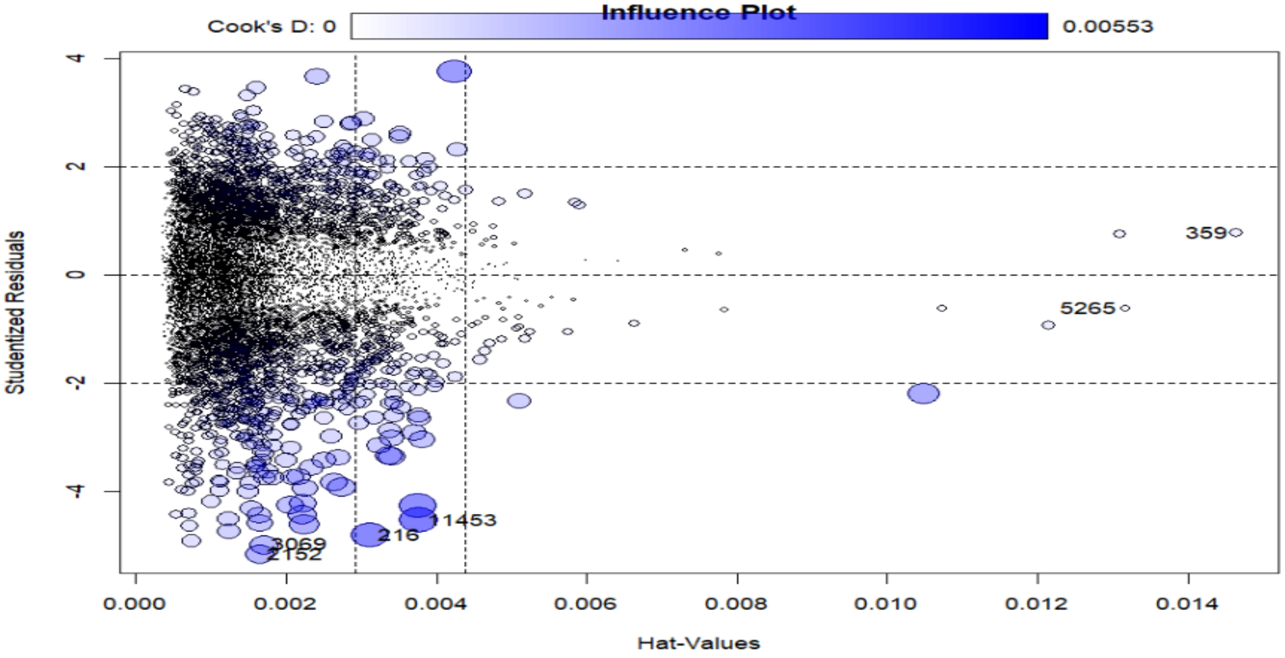


Fig.7

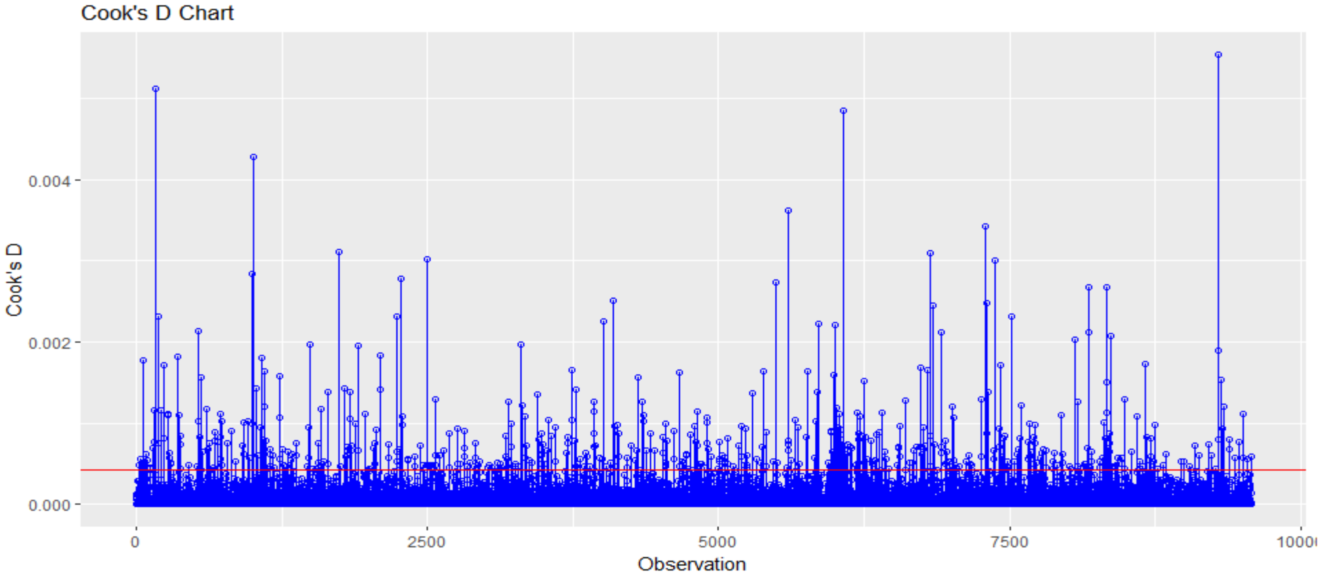


Fig.8

The influence plot (Fig 7) provides a visual representation of the Cook's D statistic. The horizontal line in the plot represents the regression line without the influence of any particular data point. Due to this, the closer a data point is to this line, the less influence it has on the regression line. The blue points on the plot correspond to the actual data points. If a data point significantly influences the regression line it will lie far from the regression line.

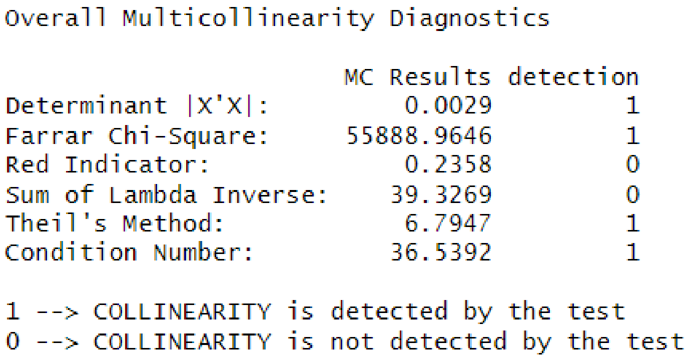
In the plot, observation 1453 has the highest influence on the regression line as data point 1453 lies significantly far from the regression line, indicating a strong influence on the regression line. The values closer to 0 are less influential and the values are more influential when they get apart from zero.

A larger hat-value for a data point suggests that this data point has a strong influence on the estimated regression coefficients. The large hat-value for data point 1453 also suggests that this data point plays a crucial role in the influence.

Overall, the influence of data point 1453 on the regression line is considerable, which can affect the robustness and reliability of the regression results.

1.4 Multicollinearity

Diagnostics:



```
Overall Multicollinearity Diagnostics

                       MC Results detection
Determinant |X'X|:         0.0029         1
Farrar Chi-Square:     55888.9646         1
Red Indicator:             0.2358         0
Sum of Lambda Inverse:    39.3269         0
Theil's Method:            6.7947         1
Condition Number:         36.5392         1

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Fig.9

From these results, we can see that the Farrar Chi-Square, Theil's Method, and Condition number suggest the presence of multicollinearity, while the Red Indicator and Sum of Lambda do not detect collinearity.

Multicollinearity Plots:

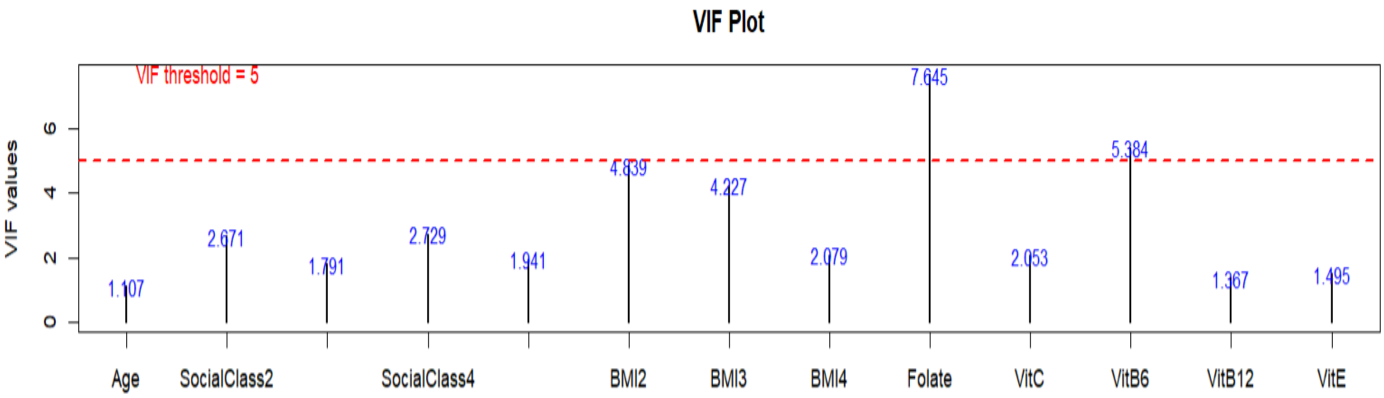


Fig.10

The Variance Inflation Factor (VIF) of a predictor is obtained by calculating the ratio of the variance of the predictor with the variance of the residuals. VIF threshold of 5 is a commonly used value to identify multicollinearity. (Fig.10)

Additionally, the high VIF values may not always mean wrong. It is crucial to inspect the model for signs of multicollinearity, even if the VIF values do not exceed the threshold.

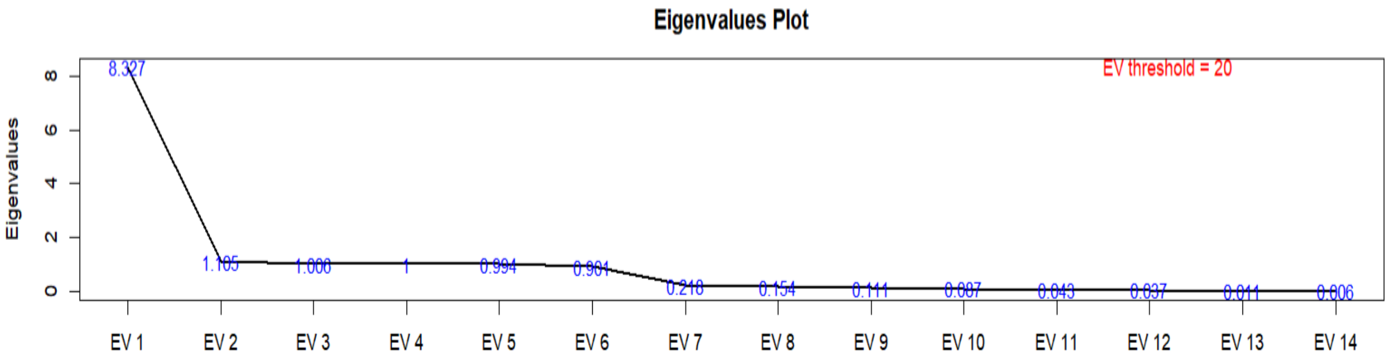
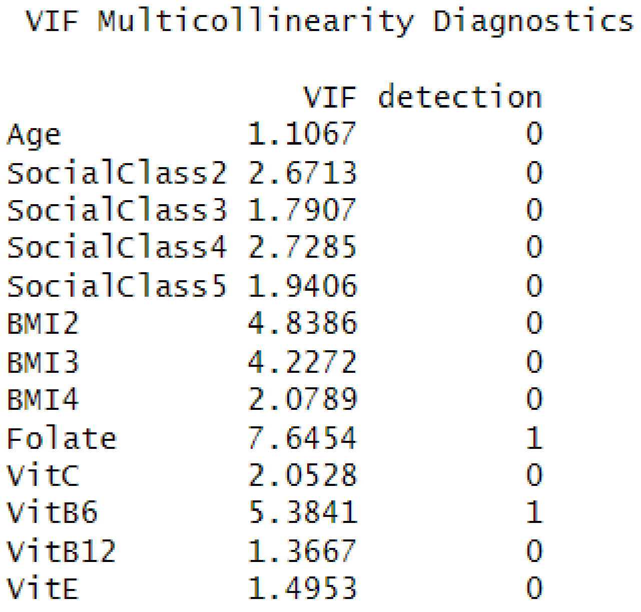


Fig.11

In the graph above, we can see the eigenvalues sorted in descending order. These eigenvalues represent the variability in the data, where a larger eigenvalue indicates a larger variability.

The graph suggests that the variance of the data decreases as we move from left to right, indicating a reduction in the variability of the data.

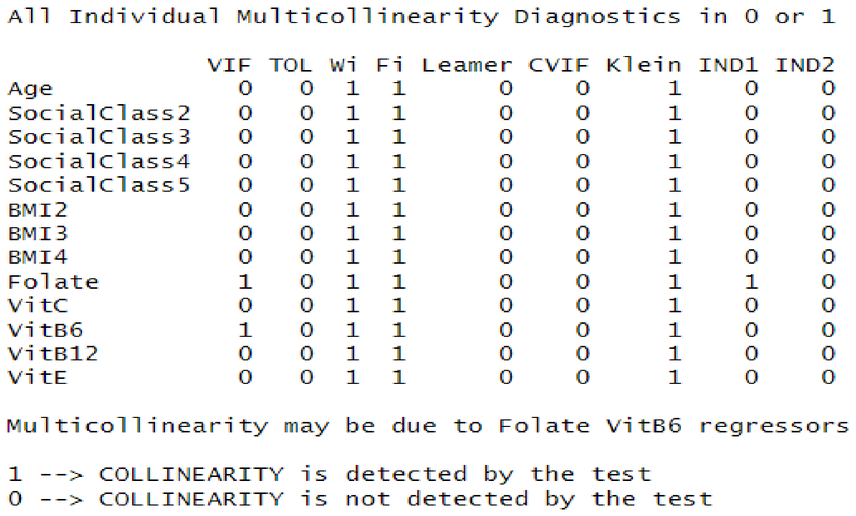
VIF multicollinearity diagnostics:



```
  VIF Multicollinearity Diagnostics

                 VIF detection
Age           1.1067         0
SocialClass2 2.6713         0
SocialClass3 1.7907         0
SocialClass4 2.7285         0
SocialClass5 1.9406         0
BMI2          4.8386         0
BMI3          4.2272         0
BMI4          2.0789         0
Folate        7.6454         1
VitC          2.0528         0
VitB6         5.3841         1
VitB12        1.3667         0
VitE          1.4953         0
```

Fig.12

The VIF values for each variable are presented in Fig12. The VIF greater than 5 is considered an indication of significant multicollinearity. Only Folate and VitB6 exceed this threshold.



```
All Individual Multicollinearity Diagnostics in 0 or 1

             VIF TOL Wi Fi Leamer CVIF Klein IND1 IND2
Age            0   0  1  1      0    0     1    0    0
SocialClass2   0   0  1  1      0    0     1    0    0
SocialClass3   0   0  1  1      0    0     1    0    0
SocialClass4   0   0  1  1      0    0     1    0    0
SocialClass5   0   0  1  1      0    0     1    0    0
BMI2           0   0  1  1      0    0     1    0    0
BMI3           0   0  1  1      0    0     1    0    0
BMI4           0   0  1  1      0    0     1    0    0
Folate         1   0  1  1      0    0     1    1    0
VitC           0   0  1  1      0    0     1    0    0
VitB6          1   0  1  1      0    0     1    0    0
VitB12         0   0  1  1      0    0     1    0    0
VitE           0   0  1  1      0    0     1    0    0

Multicollinearity may be due to Folate VitB6 regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Fig.13

The table in Fig13 shows additional individual diagnostics for each variable. The VIF, Tolerance (TOL), Wi, Fi, Leamer, CVIF, Klein, IND1, and IND2 are provided. These are various measures to detect multicollinearity. All of them are indicated by either 1 (Collinearity is detected by the test) or 0 (Collinearity is not detected by the test)

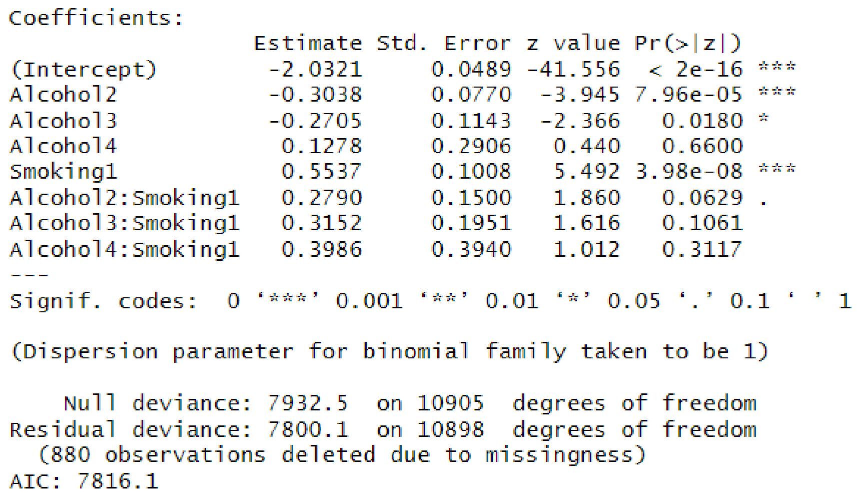
The output suggests that there may be an issue of multicollinearity due to Folate and VitB6. This implies that these two variables are highly correlated, which could lead to instability in estimating their individual effects in the regression model.

2. Binary Logistic Regression

Aim: Run a binary logistic regression to investigate the effects of Alcohol and smoking and their interaction on the response.

2.1 Binary Logistic Regression

Coefficients:

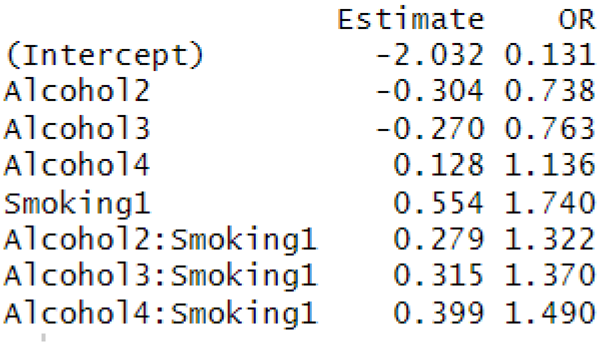


```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -2.0321     0.0489 -41.556  < 2e-16 ***
Alcohol2           -0.3038     0.0770  -3.945 7.96e-05 ***
Alcohol3           -0.2705     0.1143  -2.366   0.0180 *  
Alcohol4            0.1278     0.2906   0.440   0.6600    
Smoking1            0.5537     0.1008   5.492 3.98e-08 ***
Alcohol2:Smoking1   0.2790     0.1500   1.860   0.0629 .  
Alcohol3:Smoking1   0.3152     0.1951   1.616   0.1061    
Alcohol4:Smoking1   0.3986     0.3940   1.012   0.3117    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7932.5  on 10905  degrees of freedom
Residual deviance: 7800.1  on 10898  degrees of freedom
  (880 observations deleted due to missingness)
AIC: 7816.1
```

Fig.14

- Intercept is the estimated variable when all predictors are zero. The intercept is -2.0321.
- One unit increase in Alcohol2 is associated with a 0.3038 decrease in the response variable, similarly, one unit increase in alcohol3 is responsible for a 0.2705 decrease. Alcohol4 is associated with the 0.1278 increase in the responsible variable.
- A one-unit increase in Smoking1 is associated with a 0.5537 increase of the response variable.
- When Alcohol2:Smoking1, Alcolol3:Smoking1, and Alcohol4:Smoking1 are increased by one unit, there is a positive increase in the response variable.
- Deviance: Null Deviance is the deviance for the null model (model with no predictors). In this case, it's 7932.5 on 10905 degrees of freedom. Residual Deviance is the deviance for the models with predictors. A lower residual deviance indicates a better fit. In this case, it's 7800.1 on 10898 degrees of freedom.

Intercept:



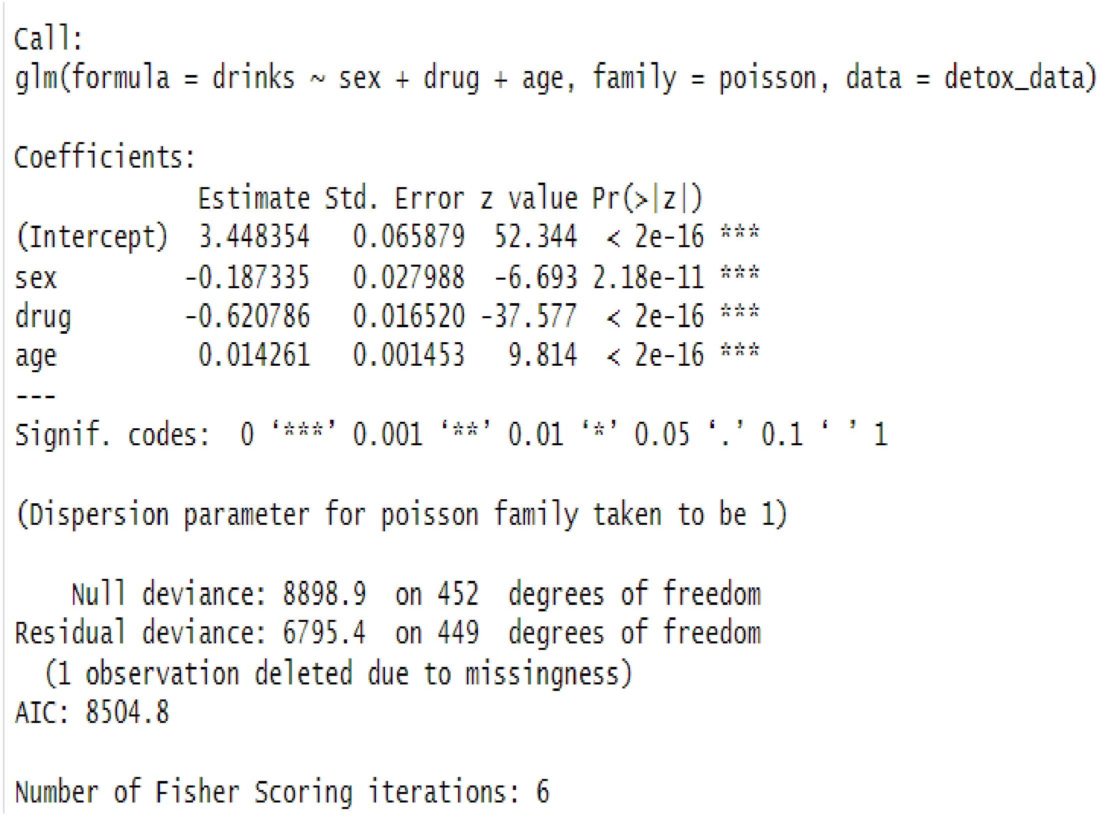
```
                  Estimate    OR
(Intercept)         -2.032 0.131
Alcohol2            -0.304 0.738
Alcohol3            -0.270 0.763
Alcohol4             0.128 1.136
Smoking1             0.554 1.740
Alcohol2:Smoking1    0.279 1.322
Alcohol3:Smoking1    0.315 1.370
Alcohol4:Smoking1    0.399 1.490
```

Fig.15

- The intercept is -2.032, and the possibility of the event occurring decreases by a factor of 0.131 when all other predictors are zero.
- Alcohol2, Alcohol3, and Alcolol4 represent different levels of the Alcohol variable. As the Alcohol level increases from 2 to 4, the possibility of the event occurring increases.
- The Smoking1 variable has a positive coefficient of 0.554, suggesting that it is responsible for the increased chances of the event taking place.
- Alcohol2:Smoking1, Alcohol3:Smoking1, Alcohol4:Smoking1 represent the interaction between Alcohol and Smoking1. All three have positive coefficients, indicating that the combination of higher Alcohol levels and Smoking1 is associated with an increase in the possibility of the event.

3.

3.1 Standard Poisson regression model

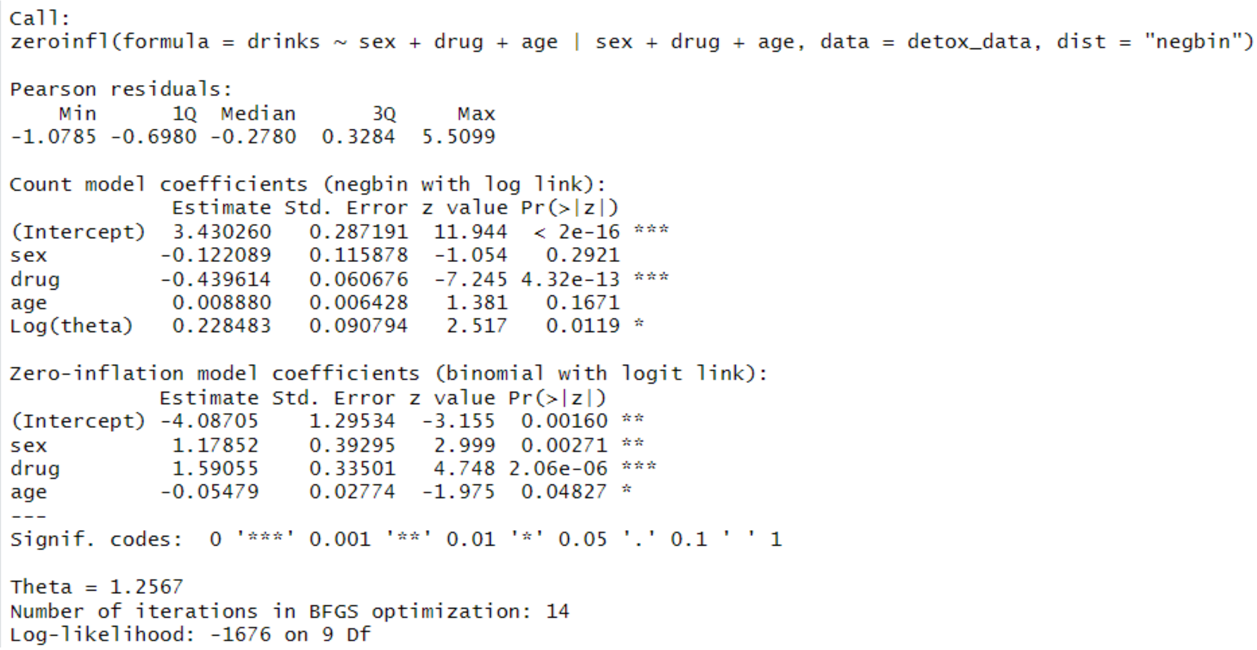


```
Call:
glm(formula = drinks ~ sex + drug + age, family = poisson, data = detox_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  3.448354   0.065879  52.344  < 2e-16 ***
sex         -0.187335   0.027988  -6.693 2.18e-11 ***
drug        -0.620786   0.016520 -37.577  < 2e-16 ***
age          0.014261   0.001453   9.814  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 8898.9  on 452  degrees of freedom
Residual deviance: 6795.4  on 449  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 8504.8

Number of Fisher Scoring iterations: 6
```

Fig.16

- The output has a highly significant intercept which gives us the count of drinks when the other predictors are null. As the sex and drug are negative, the increase in the coefficient of sex and drug results in a decrease in the count of drinks. The positive age indicates the increase in the count of drinks as age increases.
- The null deviance (no predictors), is 8898.9 on 452 degrees of freedom, and the residual deviance (with predictors) is 6795.4 on 449 degrees of freedom. In this case, the model with predictors has a lower deviance suggesting a better fit.
- Lower AIC values indicate better-fitting models. In this case, the AIC is 8504.8, which can be used to compare this model with others.

3.2 Zero-inflated negative binomial

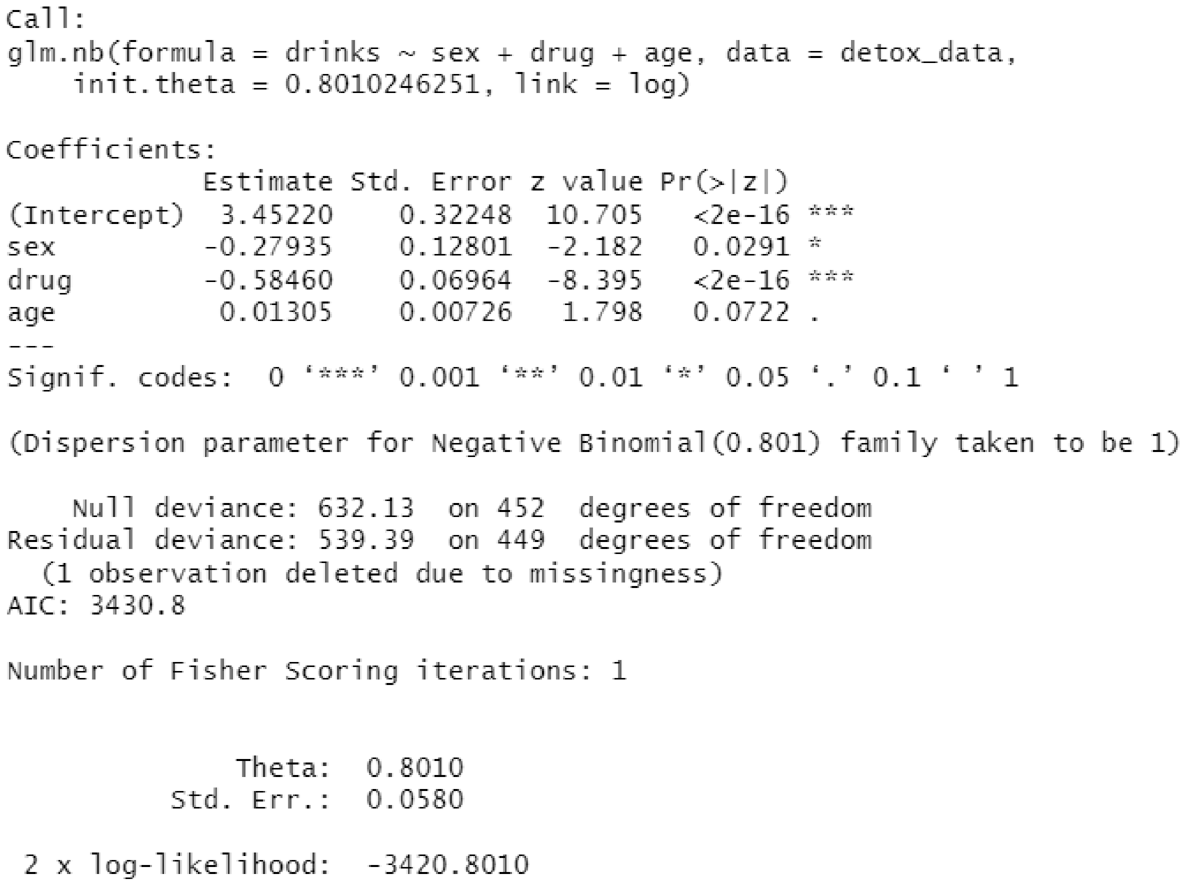


```
Call:
zeroinfl(formula = drinks ~ sex + drug + age | sex + drug + age, data = detox_data, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.0785 -0.6980 -0.2780  0.3284  5.5099 

Count model coefficients (negbin with log link):
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  3.430260   0.287191  11.944  < 2e-16 ***
sex         -0.122089   0.115878  -1.054   0.2921    
drug        -0.439614   0.060676  -7.245 4.32e-13 ***
age          0.008880   0.006428   1.381   0.1671    
Log(theta)   0.228483   0.090794   2.517   0.0119 *  

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.08705    1.29534  -3.155  0.00160 ** 
sex          1.17852    0.39295   2.999  0.00271 ** 
drug         1.59055    0.33501   4.748 2.06e-06 ***
age         -0.05479    0.02774  -1.975  0.04827 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Theta = 1.2567 
Number of iterations in BFGS optimization: 14 
Log-likelihood: -1676 on 9 Df
```

Fig.17

- Pearson residuals give an idea of how well the model fits the data. Large residuals may indicate areas where the model doesn't fit well.
- Count model coefficients:
  The significant intercept indicates the expected count of drinks when the predictors are zero. Sex and Drug are significant, with negative coefficients. This suggests that an increase in sex or drug is associated with a decrease in the expected count of drinks. Age is not statistically significant in the count model due to low value.
- Zero-Inflation Model Coefficients:
  The intercept is not so significant due to its negative value.
  Sex and drug are significant possible risers of the excess zeros.
- Theta is the dispersion parameter in the negative binomial distribution. The estimated value of theta is 1.2567, indicating overdispersion compared to a Poisson distribution.
- The log-likelihood is -1676, and the number of iterations taken in optimization is 14. A higher number of iterations means complexity in the optimization process.

3.3 Standard negative binomial

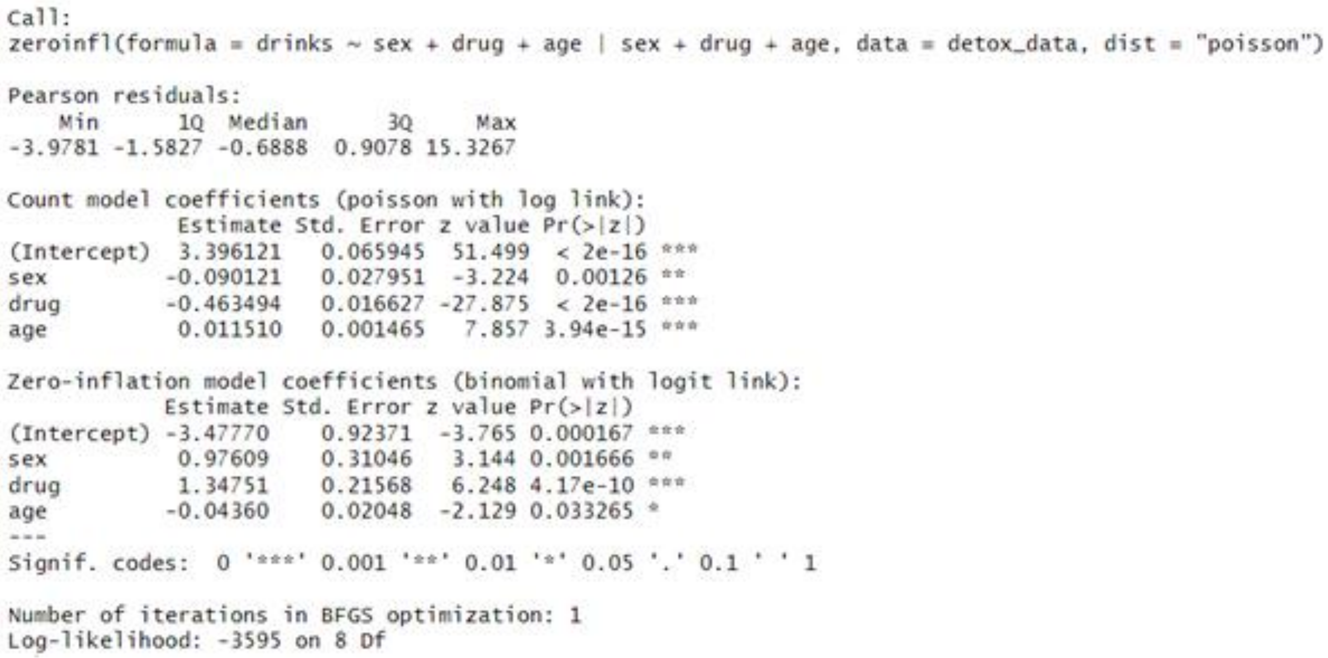


```
Call:
glm.nb(formula = drinks ~ sex + drug + age, data = detox_data,
    init.theta = 0.8010246251, link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.45220    0.32248  10.705   <2e-16 ***
sex         -0.27935    0.12801  -2.182   0.0291 *
drug        -0.58460    0.06964  -8.395   <2e-16 ***
age          0.01305    0.00726   1.798   0.0722 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.801) family taken to be 1)

    Null deviance: 632.13  on 452  degrees of freedom
Residual deviance: 539.39  on 449  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 3430.8

Number of Fisher Scoring iterations: 1


              Theta:  0.8010
          Std. Err.:  0.0580

 2 x log-likelihood:  -3420.8010
```

Fig.18

- A highly significant intercept indicates that the expected log count of drinks when all other predictors are zero.
- Sex and Drug are significant with negative coefficients. This suggests that an increase in sex or drug is responsible for a decrease in the expected count of drinks. Age is not significant in the model.
- In this case, the residual deviance has a lower deviance, suggesting an improvement in fit compared to the null model.
  The difference in degrees of freedom (3) between null and residual deviance is the number of predictors in the model.
- The estimated value of theta is 0.8010, indicating overdispersion compared to a Poisson distribution. The standard error of the estimated theta is 0.0580.

3.4 Zero-inflated Poisson model

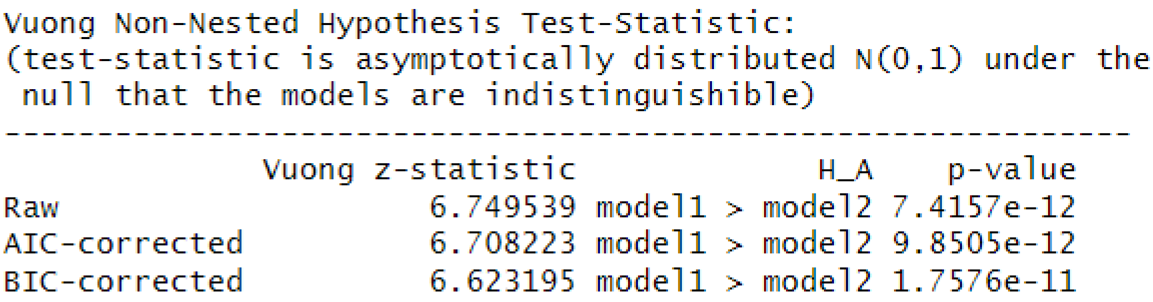


```
Call:
zeroinfl(formula = drinks ~ sex + drug + age | sex + drug + age, data = detox_data, dist = "poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-3.9781 -1.5827 -0.6888  0.9078 15.3267 

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  3.396121   0.065945  51.499  < 2e-16 ***
sex         -0.090121   0.027951  -3.224  0.00126 ** 
drug        -0.463494   0.016627 -27.875  < 2e-16 ***
age          0.011510   0.001465   7.857 3.94e-15 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.47770    0.92371  -3.765 0.000167 ***
sex          0.97609    0.31046   3.144 0.001666 ** 
drug         1.34751    0.21568   6.248 4.17e-10 ***
age         -0.04360    0.02048  -2.129 0.033265 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Number of iterations in BFGS optimization: 1 
Log-likelihood: -3595 on 8 Df
```

Fig.19

Considering a zero inflated model is optimal when the data contains a high number of zero values that cannot be explained by the Poisson distribution. In our model, the zero-inflation model suggests that the probability of observing zero drinks is influenced by sex, drug use, and age. The zero-inflation model is justified if there is evidence of excess zeros in the data.

Comparison of Zero-inflated Poisson Model and Standard Poisson Model:



```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                    6.749539 model1 > model2 7.4157e-12
AIC-corrected          6.708223 model1 > model2 9.8505e-12
BIC-corrected          6.623195 model1 > model2 1.7576e-11
```

Fig.20

- model1 - zero-inflated Poisson model
  model2 - standard Poisson model

The p-value is extremely small (<0.05) suggesting strong evidence against the null hypothesis. In all the 3 tests, the model1 is greater than the model2 which provides strong evidence that the Zero-inflated Poisson model is a better fit for the data compared to the Standard Poisson Regression Model. It supports the decision to use the ZIP model based on its ability to handle excess zeros in the data.