Homework 2

Snigdha Peddi

Question 1: (Ex. 7.2 in HSAUR, modified for clarity) Collett (2003) argues that two outliers need to be removed from the **plasma** data. Try to identify those two unusual observations by means of a scatterplot. (Hint: Consider a plot of the residuals from a simple linear regression.)

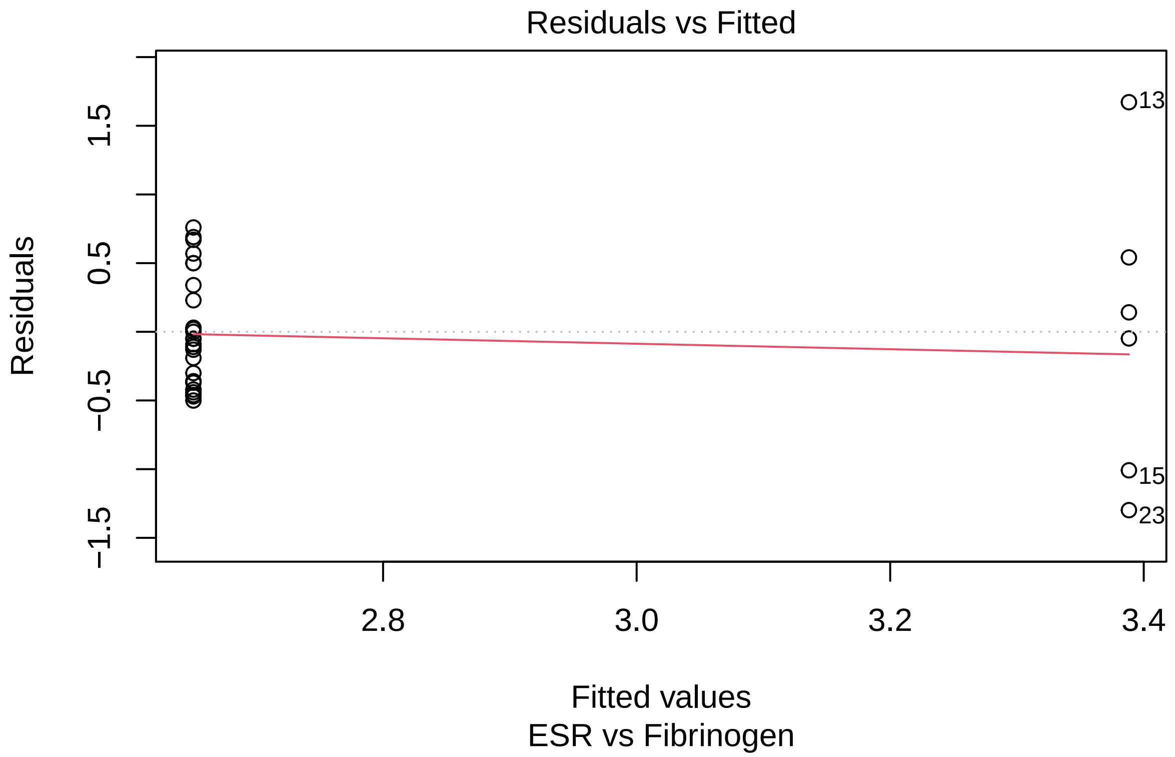
Answer 1: The p values of the simple linear regression model obtained from the plasma date with ESR as dependent variable show that fibrinogen is significant at 95% confidence level.Upon fitting a model between ESR and fibrinogen the plot of residuals indicates that the two outliers are the ones with residuals of about 1.5 and -1.2. These are the outliers with record numbers 13(fibrinogen 5.06) and record 23(fibrinogen=2.09) which are the minimum and maximum values of the feature.Similar trend is observed in QQ plots. These are the two unusual observations that have to be removed. .

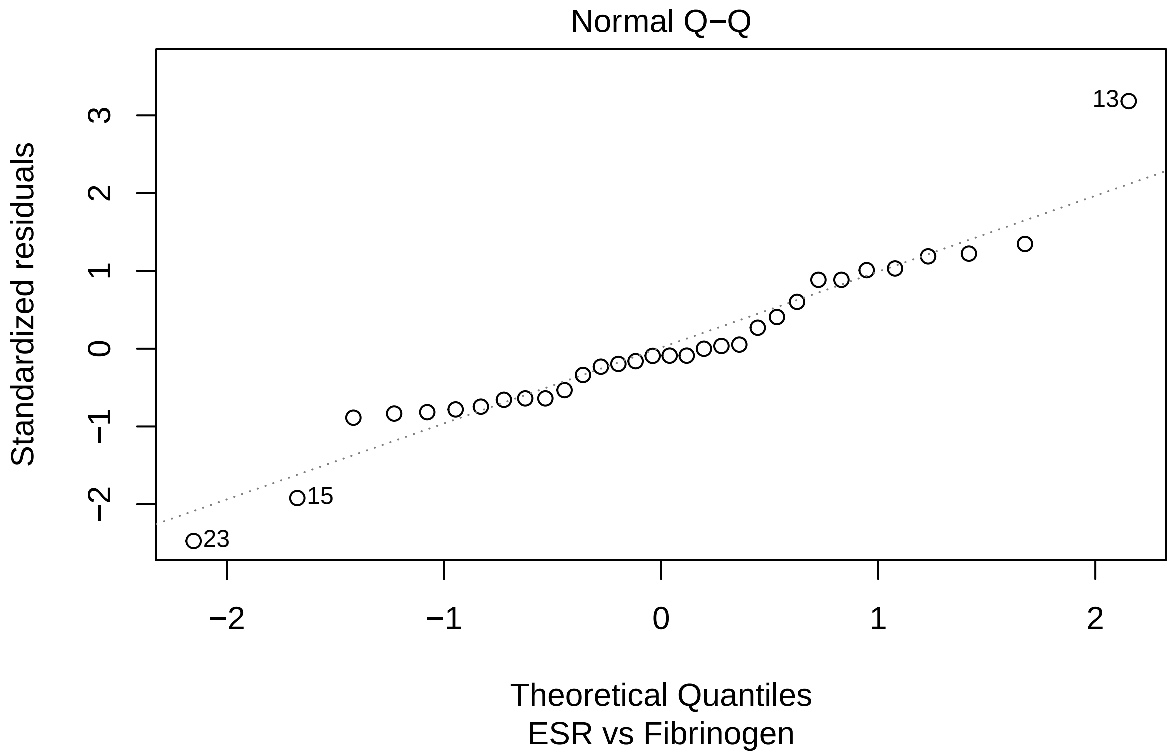
```
##
## Call:
## lm(formula = fibrinogen ~ ESR, data = plasma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29833 -0.36288 -0.05038  0.37962  1.67167
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.6504     0.1128  23.494  < 2e-16 ***
## ESRESR > 20   0.7379     0.2605   2.833  0.00818 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5752 on 30 degrees of freedom
## Multiple R-squared:  0.211,  Adjusted R-squared:  0.1847
## F-statistic: 8.023 on 1 and 30 DF,  p-value: 0.008175
```

```
##
## Call:
## lm(formula = globulin ~ ESR, data = plasma)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1154 -4.1154  0.8846  2.8846 10.8846
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.1154     0.8848  39.687   <2e-16 ***
## ESRESR > 20   2.8846     2.0434   1.412    0.168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.512 on 30 degrees of freedom
## Multiple R-squared:  0.06229,    Adjusted R-squared:  0.03103
## F-statistic: 1.993 on 1 and 30 DF,  p-value: 0.1683
```

```
## Summary: 2.09 2.29 2.6 2.78875 3.1675 5.06
```





Question 2. (Ex. 6.6 in HSAUR, modified for clarity) (Multiple Regression) Continuing from the lecture on the **hubble** data from **gamair** library:

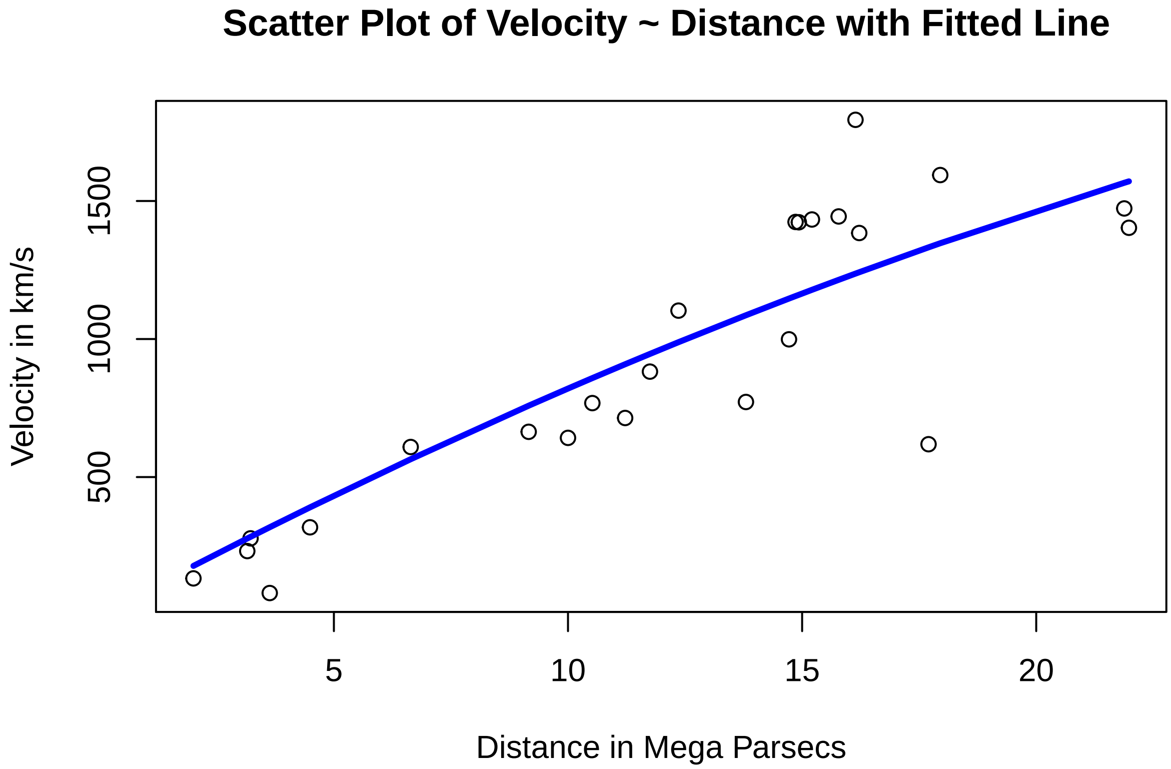
a) Fit a quadratic regression model, i.e., a model of the form

$$\text{Model 2: } velocity = \beta_1 \times distance + \beta_2 \times distance^2 + \epsilon$$

```
## [1] "Summary of Quadratic Model:"

##
## Call:
## lm(formula = y ~ x + x2 - 1, data = hubble.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -713.15 -152.76  -54.85  163.92  557.01
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## x   90.9046    16.5726   5.485 1.64e-05 ***
## x2  -0.8837     0.9925  -0.890    0.383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 260.1 on 22 degrees of freedom
## Multiple R-squared:  0.944,  Adjusted R-squared:  0.9389
## F-statistic: 185.3 on 2 and 22 DF,  p-value: 1.715e-14
```

b) Plot the fitted curve from Model 2 over the scatterplot of the data.

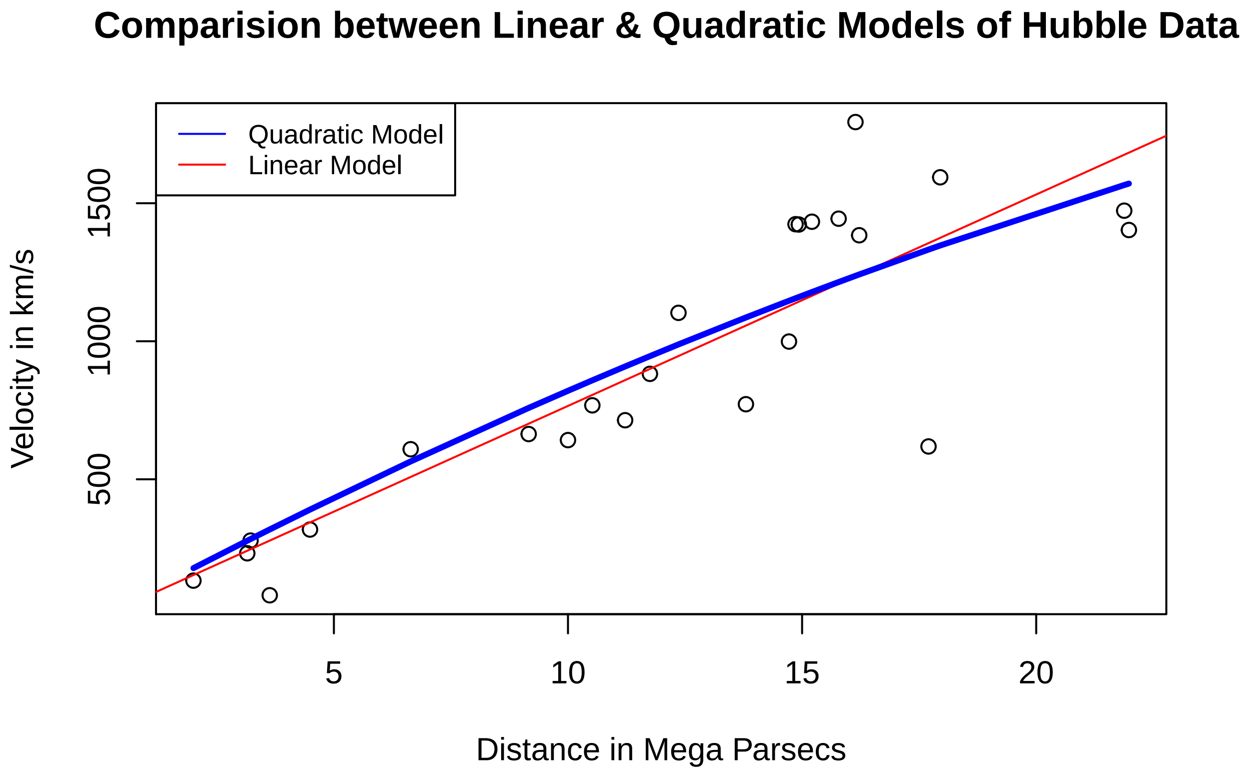


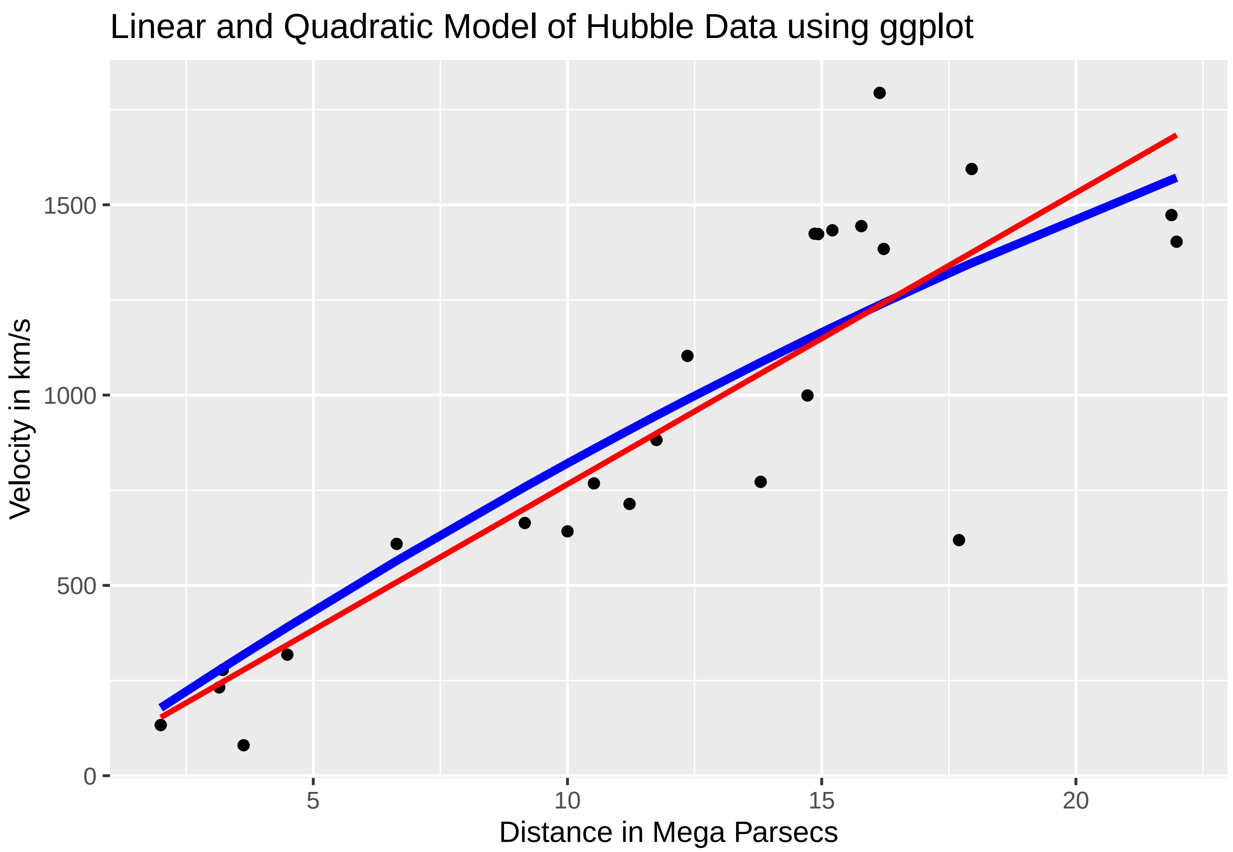
c) Add a simple linear regression fit over this plot. Use the relationship between *velocity* and *distance* to determine the constraints on the parameters and explain your reasoning. Use different color and/or line type to differentiate the two and add a legend to differentiate between the two models.

Answer 2.c: The added parameter of the quadratic model is increasing the variance and residual error. It is clearly visible in the plots that this added parameter is not very significant.

```
## [1] "Summary of Linear Model:"

##
## Call:
## lm(formula = y ~ x - 1, data = hubble)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -736.5 -132.5  -19.0  172.2  558.0
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   76.581      3.965   19.32 1.03e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 258.9 on 23 degrees of freedom
## Multiple R-squared:  0.9419, Adjusted R-squared:  0.9394
## F-statistic: 373.1 on 1 and 23 DF,  p-value: 1.032e-15
```





d) Examine the plot, which model do you consider most sensible?

Answer 2.d: After examining the plot and both the regression lines I believe that the simple linear model is most sensible than the Quadratic model as residual error and variance are less.The x^2 variable of the quadratic model is not very significant.

e) Which model is better? Provide a statistical justification for your choice of model.

Answer 2.e: Comparing the the summaries of two models clearly shows that the added variable of quadratic model is not significant as its p value is higher. Though the Residual errors and Adjusted R-squared of both models are similar the standard error of linear model(3.965) is smaller than the quadratic model(16.5726) indicating that for Hubbles data simple linear regression model is good fit. The comaprision between the two models using Chi square test also shows that the Quadratic model is not very significant and X2 variable is not improving the model.

```
## Analysis of Variance Table
##
## Model 1: y ~ x - 1
## Model 2: y ~ x + x2 - 1
##   Res.Df     RSS Df Sum of Sq Pr(>Chi)
## 1     23 1542066
## 2     22 1488429  1     53636   0.3733
```

Note: The quadratic model here is still regarded as a `linear regression` model since the term `linear` relates to the parameters of the model and not to the powers of the explanatory variables.

Question 3. (Ex. 7.4 in HSAUR, modified for clarity) The **leuk** data from package **MASS** shows the survival times from diagnosis of patients suffering from leukemia and the values of two explanatory variables, the white blood cell count (wbc) and the presence or absence of a morphological characteristic of the white blood cells (ag).

a) Define a binary outcome variable according to whether or not patients lived for at least 24 weeks after diagnosis. Call it *surv24*.

Answer 3.a: Using **dplyr** package **mutate** function a new variable "serv24" is added to leuk.dat data set where all observations with time >= 24 were assigned "1" and others "0".

b) Fit a logistic regression model to the data with *surv24* as the response variable. If regression coefficients are close to zero, then apply a log transformation to the corresponding covariate. Write the model for the fitted data (see Exercise 2a for an example of a model.)

Answer 3.b: A Logistic Regression model is fit with wbc and ag as independent variables and serv24 as dependent variable.As the regression coefficients of wbc are close to zero another model with natural log of wbc is fit.

```
## [1] "Summary of Model.serv24:"
```

```
##
## Call:
## glm(formula = surv24 ~ wbc + ag, family = "binomial", data = leuk.dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5357  -0.8207  -0.7475   0.8677   1.9390
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.706e-01  6.469e-01  -1.346    0.178
## wbc         -8.436e-06  1.150e-05  -0.733    0.463
## agpresent    1.733e+00  7.785e-01   2.226    0.026 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45.475  on 32  degrees of freedom
## Residual deviance: 39.516  on 30  degrees of freedom
## AIC: 45.516
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] "Coefficients of Model.serv24:"
```

```
##   (Intercept)           wbc     agpresent
## -8.706387e-01 -8.435736e-06  1.732867e+00
```

```
## [1] "Summary of Model.serv24.log - after log transformation:"
```

```
##
## Call:
## glm(formula = surv24 ~ log(wbc) + ag, family = "binomial", data = leuk.dat)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6310  -0.9056  -0.6258   0.8592   2.1032
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.4556     2.9821   1.159   0.2466
## log(wbc)     -0.4822     0.3149  -1.531   0.1257
## agpresent     1.7621     0.8093   2.177   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45.475  on 32  degrees of freedom
## Residual deviance: 37.498  on 30  degrees of freedom
## AIC: 43.498
##
## Number of Fisher Scoring iterations: 3
```

```
## [1] "Coefficients of Model.serv24.log:"
```

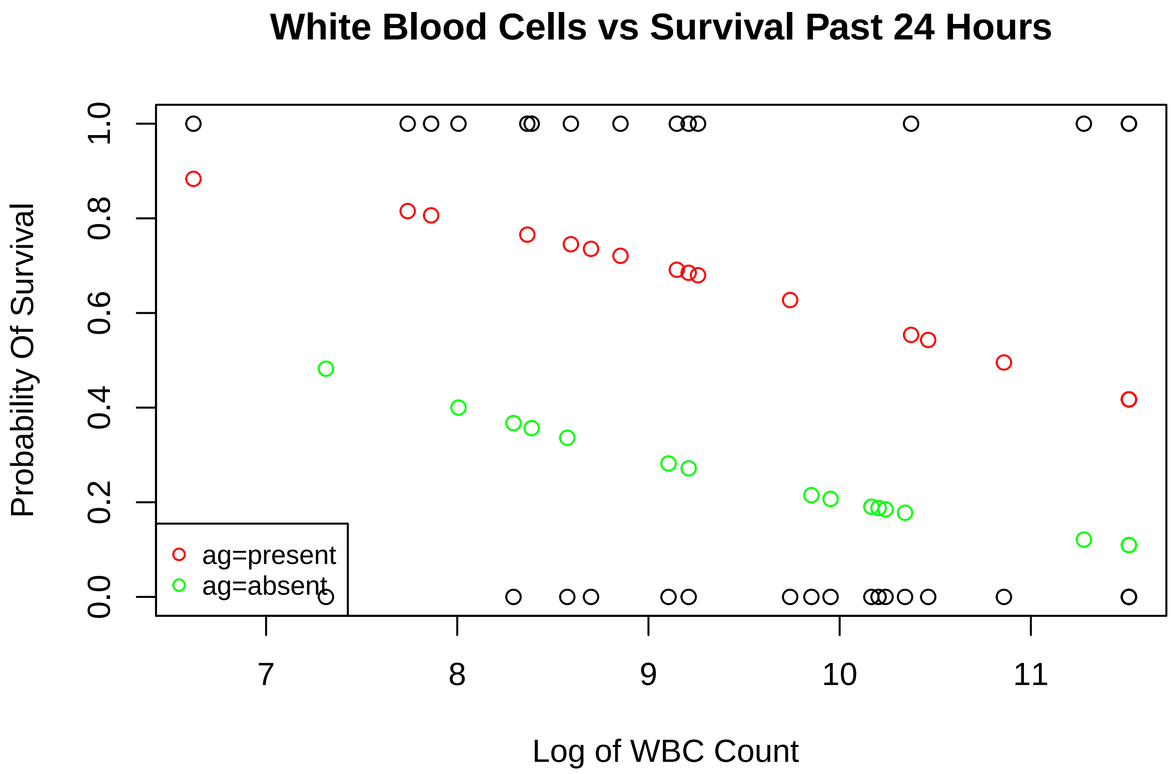
```
## (Intercept)    log(wbc)   agpresent
##   3.4555870  -0.4821891   1.7621259
```

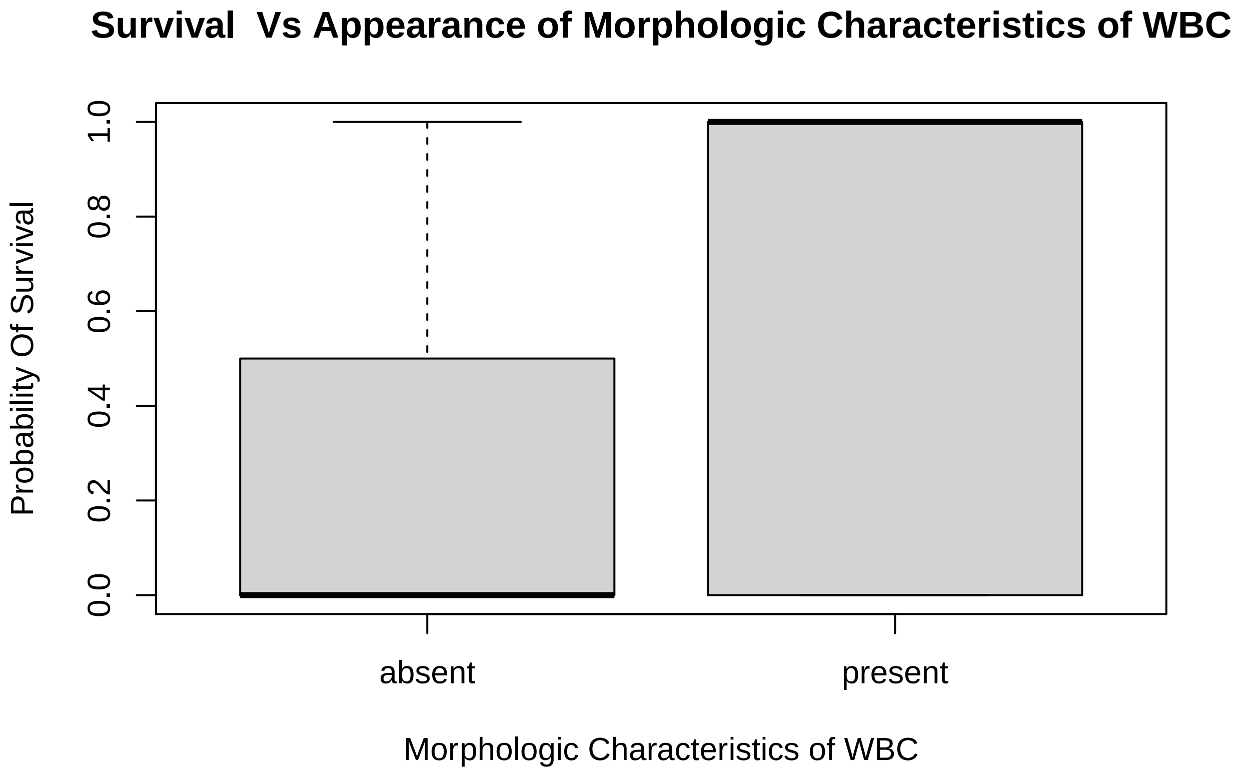
c) Interpret the final model you fit. Provide graphics to support your interpretation.

Answer 3.c: The Confusion matrix shows that the fit model is having an accuracy of 76% and an error rate of 24%. The first plot, 'White Blood Cells vs Survival Past 24 Hours', shows a high probability of living more than 24 weeks for patients with ag 'present' test result even with low wbc count and high probability of death if ag is absent. The second plot " Survival Vs Appearance of Morphologic Characteristics of WBC" shows that the presence of ag results in survival of patients in almost every case and absence of ag results in death in every case.This atypical behavior have to be further investigated along with the significance of ag on survival.

```
## Confusion Matrix and Statistics
##
##      True
## pred  No Yes
##   No  15   5
##   Yes  3  10
##
##                Accuracy : 0.7576
##                  95% CI : (0.5774, 0.8891)
##     No Information Rate : 0.5455
##     P-Value [Acc > NIR] : 0.0101
##
##                   Kappa : 0.5056
##
##  Mcnemar's Test P-Value : 0.7237
##
##             Sensitivity : 0.8333
##             Specificity : 0.6667
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.7692
##              Prevalence : 0.5455
##          Detection Rate : 0.4545
##    Detection Prevalence : 0.6061
##       Balanced Accuracy : 0.7500
##
##        'Positive' Class : No
##

## Error Rate:  0.2424242
```





d) Update the model from part b) to include an interaction term between the two predictors. Which model fits the data better? Provide a statistical justification for your choice of model.

Answer 3.d: Second model has a lower AIC value than the model without interaction term (42.167 vs 43.498)indicating that this model better represents the data.

```
## [1] "Summary of Model.serv24:"

##
## Call:
## glm(formula = surv24 ~ log(wbc) + ag, family = "binomial", data = leuk.dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6310  -0.9056  -0.6258   0.8592   2.1032
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.4556     2.9821   1.159   0.2466
## log(wbc)     -0.4822     0.3149  -1.531   0.1257
## agpresent     1.7621     0.8093   2.177   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 45.475  on 32  degrees of freedom
## Residual deviance: 37.498  on 30  degrees of freedom
## AIC: 43.498
## 
## Number of Fisher Scoring iterations: 3
```

```
## [1] "Summary of Model.serv24_new:"
```

```
## 
## Call:
## glm(formula = surv24 ~ ag * log(wbc), family = "binomial", data = leuk.dat)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.9183  -0.7835  -0.6750   0.7310   1.7838  
## 
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)  
## (Intercept)          -2.5946     4.6583  -0.557   0.5775  
## agpresent            13.6306     7.0909   1.922   0.0546 .
## log(wbc)              0.1545     0.4746   0.326   0.7447  
## agpresent:log(wbc)  -1.2315     0.7182  -1.715   0.0864 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 45.475  on 32  degrees of freedom
## Residual deviance: 34.167  on 29  degrees of freedom
## AIC: 42.167
## 
## Number of Fisher Scoring iterations: 4
```

```
## Analysis of Deviance Table
## 
## Model 1: surv24 ~ log(wbc) + ag
## Model 2: surv24 ~ ag * log(wbc)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1        30     37.498                       
## 2        29     34.167  1   3.3315  0.06797 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Question 4. (Adapted from ISLR) Load the **Default** dataset from **ISLR** library. The dataset contains four features on 10,000 customers. We want to predict which customers will default on their credit card debt based on the observed features.

a) Select a class of models using appropriate summaries and graphics. **Do not overplot.**

Answer 4.a: -Initial exploratory analysis indicate that the mean salary of a student is around 17950 and mean balance on credit card is about 988 where as the mean income and balance of non-students are about 40012 and 772 respectively . -There is a negative correlation between balance and income. As the income increases by a unit the balance reduces by 15%. -As the balance increases customers tend to default more. -From the plot it is difficult to establish relationship between income and the default by customers.

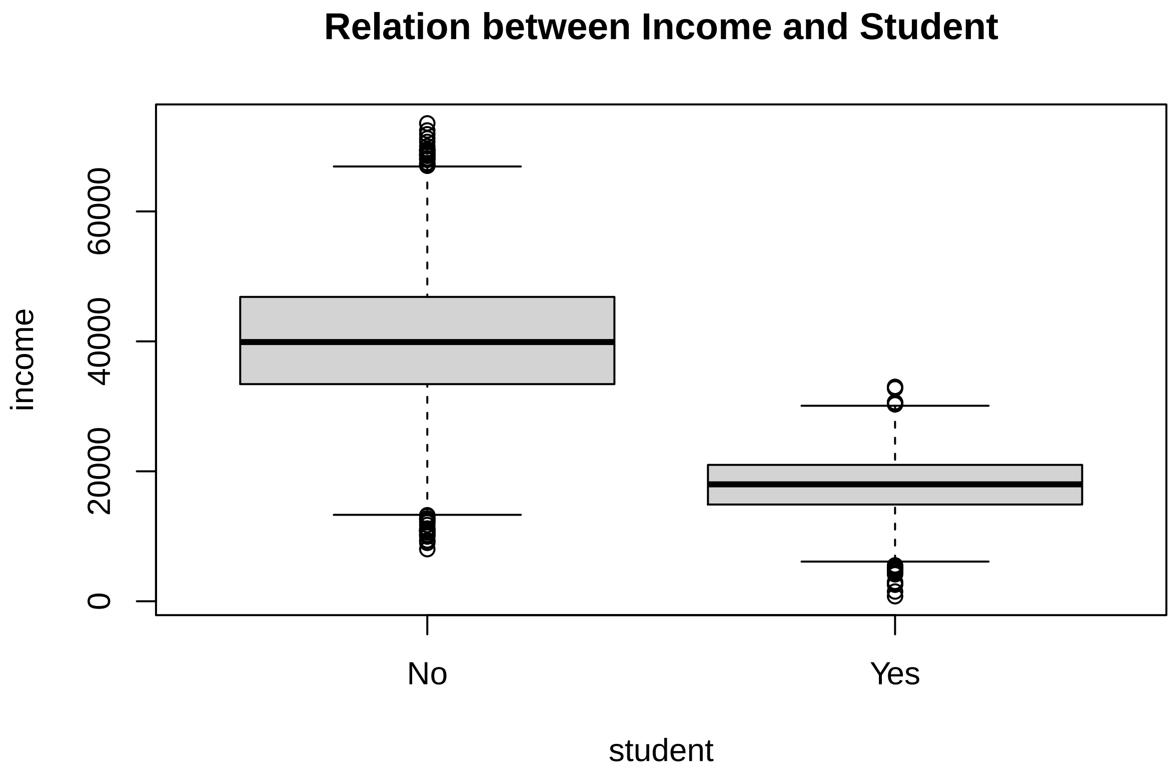
After thorough investigation I believe generalized linear regression model of family binomial (logistic regression) is good model for this data set.

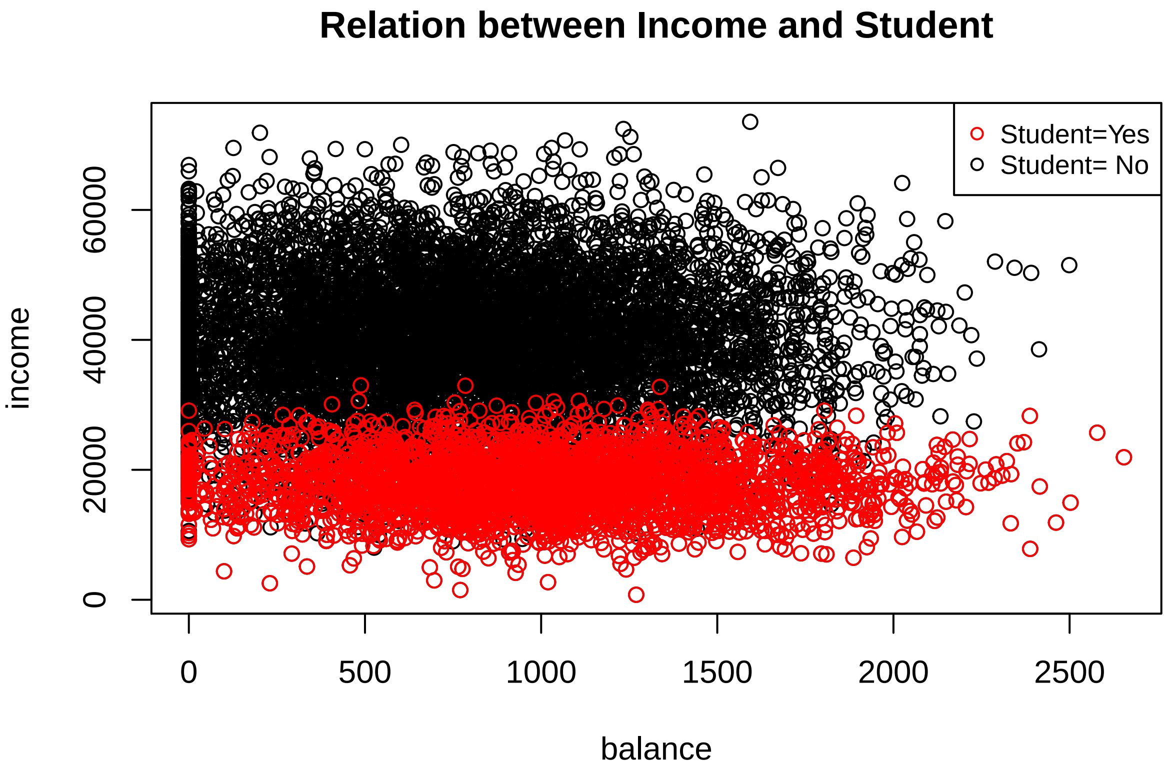
```
##  default    student       balance           income     
##  No :6850   No :7056   Min.   :   0.0   Min.   : 8018  
##  Yes: 206   Yes:   0   1st Qu.: 418.2   1st Qu.:33417  
##                        Median : 759.2   Median :39893  
##                        Mean   : 771.8   Mean   :40012  
##                        3rd Qu.:1093.3   3rd Qu.:46841  
##                        Max.   :2499.0   Max.   :73554  
```

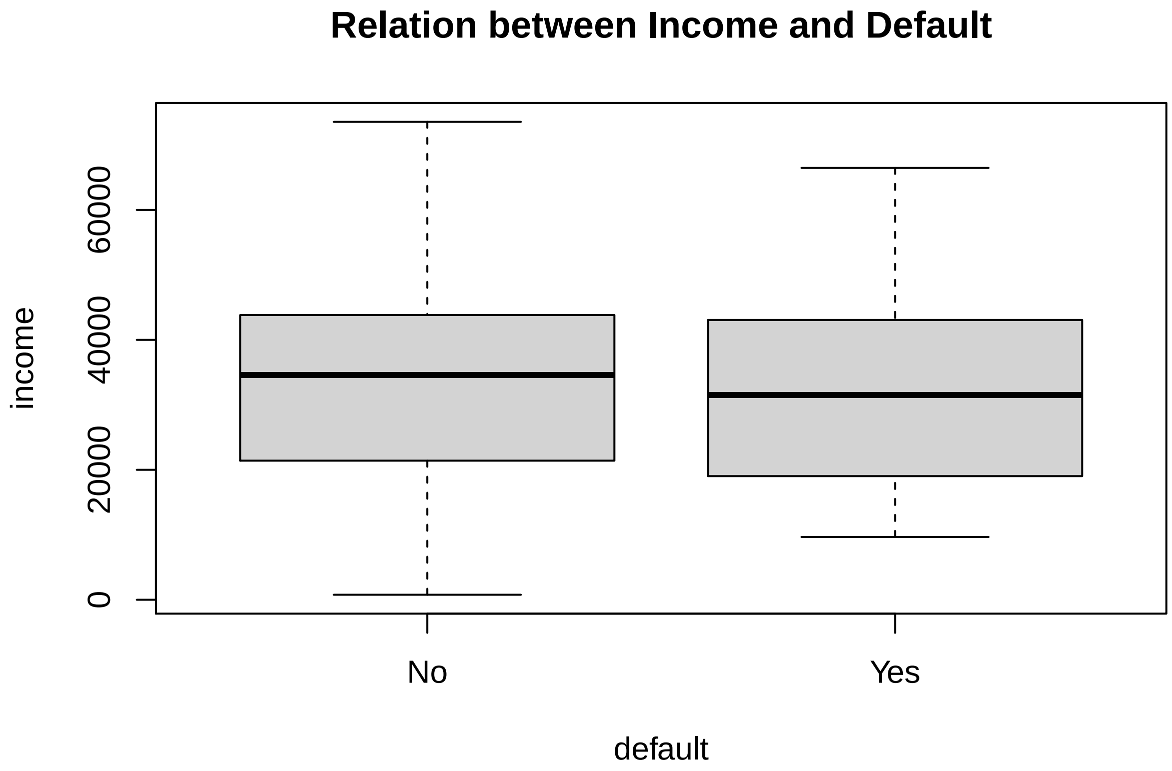
```
##  default    student       balance           income     
##  No :2817   No :   0   Min.   :   0.0   Min.   :  772  
##  Yes: 127   Yes:2944   1st Qu.: 655.6   1st Qu.:14887  
##                        Median : 980.0   Median :17994  
##                        Mean   : 987.8   Mean   :17950  
##                        3rd Qu.:1303.9   3rd Qu.:20986  
##                        Max.   :2654.3   Max.   :33003  
```

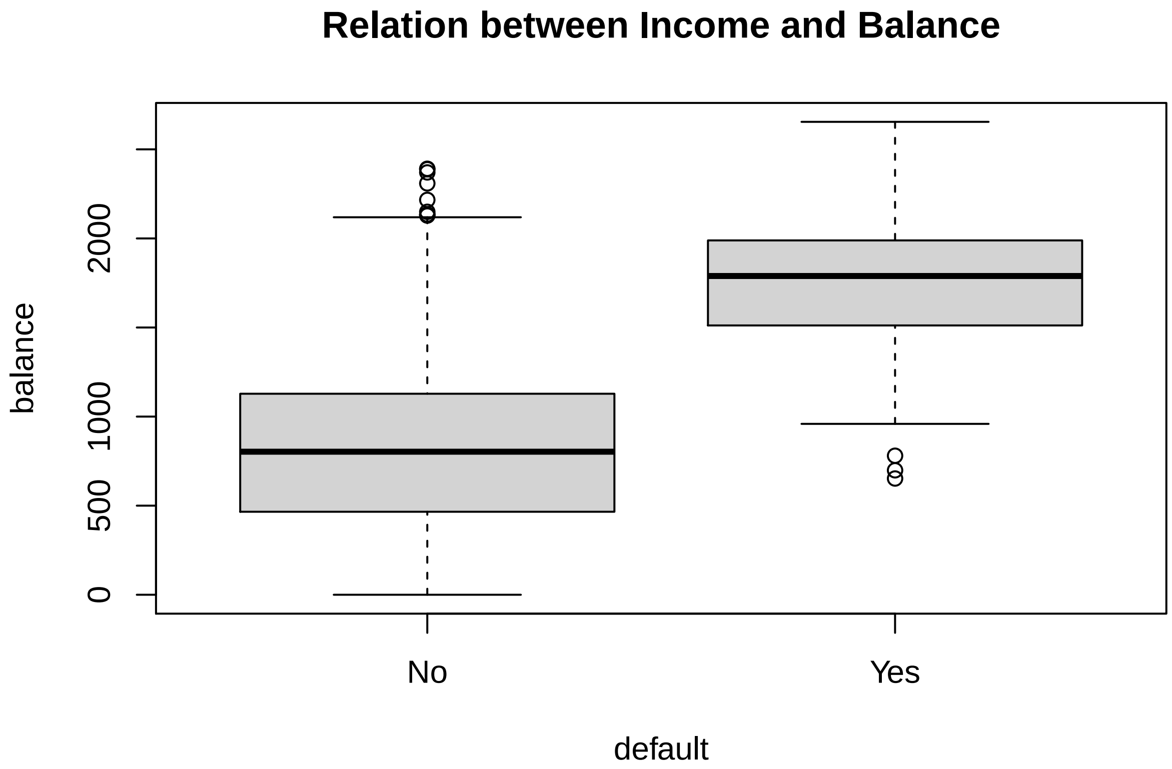
```
## [1] "Correlation between Balance and Income:"
```

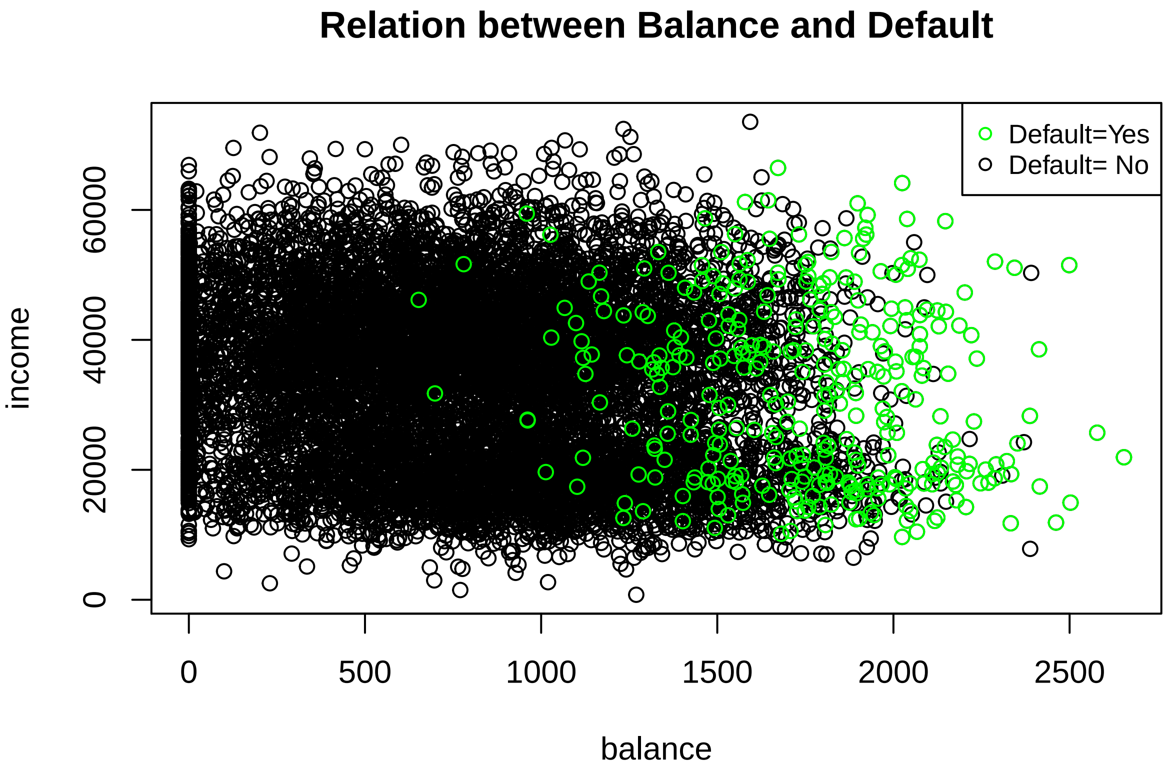
```
##         balance income
## balance    1.00  -0.15
## income    -0.15   1.00
```

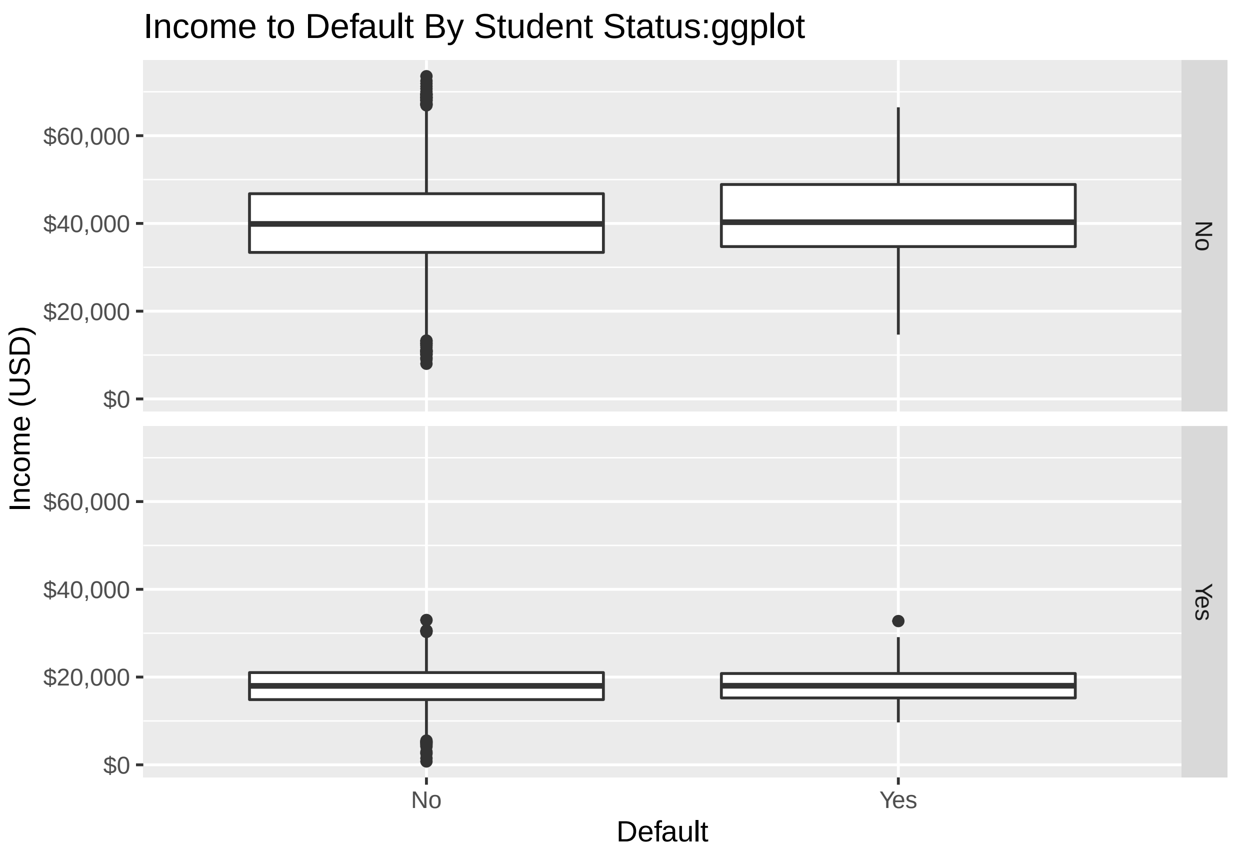


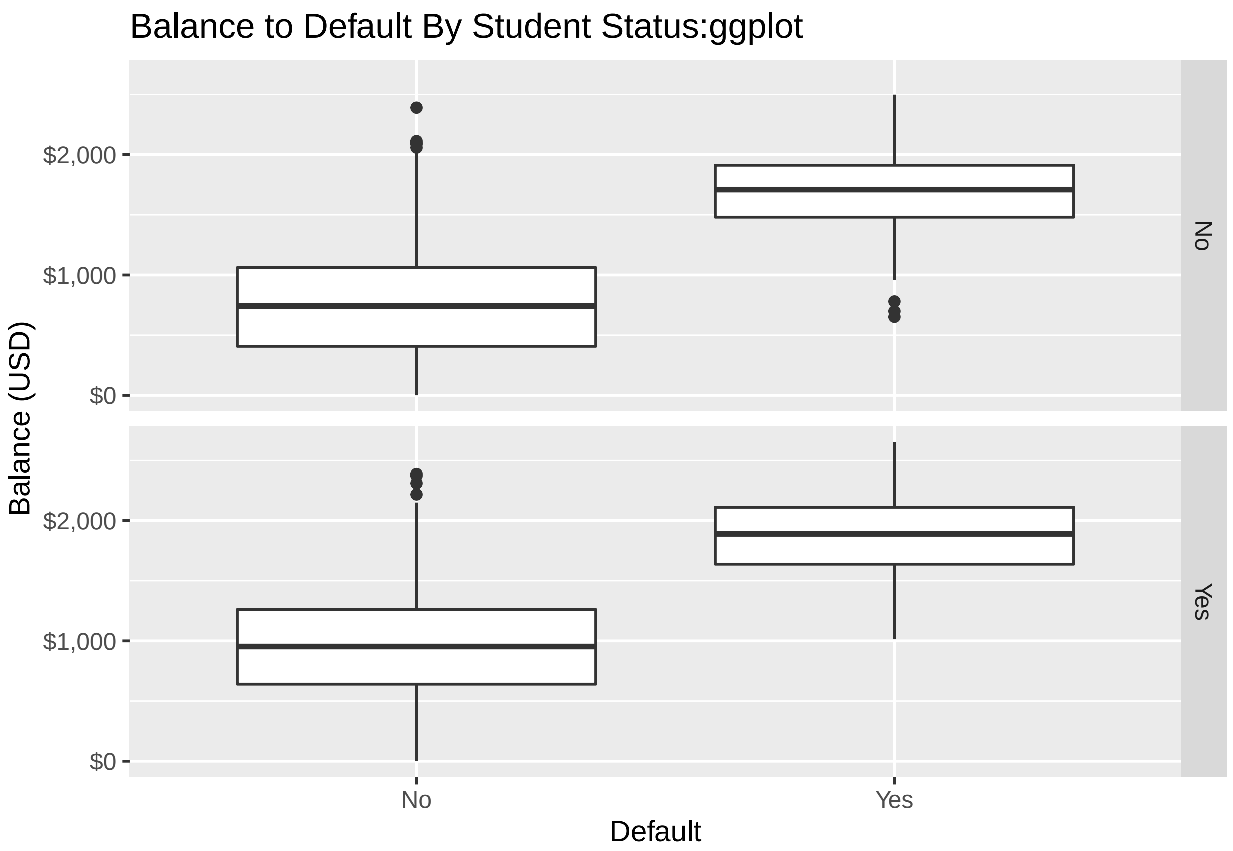

Relation between Income and Student
Student=Yes
Student= No
income
0
20000
40000
60000
0
500
1000
1500
2000
2500
balance


Relation between Income and Default


Relation between Income and Balance
balance
0
500
1000
2000
No
Yes
default


Relation between Balance and Default
Default=Yes
Default= No
income
0
20000
40000
60000
balance
0
500
1000
1500
2000
2500


Income to Default By Student Status:ggplot
Income (USD)
$60,000
$40,000
$20,000
$0
$60,000
$40,000
$20,000
$0
No
Yes
No
Yes
Default



The analysis of variance between income and default shows that there is a correlation between these features at 5% significance level and confirms that there is also a strong correlation between balance and default at 95% confidence interval.

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## default         1 7.023e+08 702276944    3.95 0.0469 *
## Residuals    9998 1.778e+12 177813503
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Tables of means
## Grand mean
##
## 33516.98
##
##  default
##        No   Yes
##     33566 32089
## rep  9667   333
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## default         1 2.868e+08 286792390    1397 <2e-16 ***
## Residuals    9998 2.053e+09    205319
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Tables of means
## Grand mean
##
## 835.3749
##
##  default
##       No  Yes
##      804 1748
## rep 9667  333
```

b) State the class of models. Fit the appropriate logistic regression model.

Answer 4.b: The class of both the models is glm (generalized linear model),lm(linear model)

Two logistic regression models were fit.

Model 1: default student+balance+income

Model 2: default student + balance + income + (student * income) + (balance * student) + (balance * income)

```
## Class of model with all features: glm lm
```

```
## Class of model with features and interaction terms: glm lm
```

c) Discuss your results, paying particular attention to which feature variables are predictive of the response. Are there meaningful interactions among the feature variables?

Answer 4.c: The student status and balance plays a significant role in predicting the customers who default. The interaction terms in the models are not significant at 95% confidance interval. The simpler logistic regression model with no interaction terms is best fit.Also, the lower Akaike information criterion (AIC) of the model confirms that simpler model is better than the model with interaction terms.

```
##
## Call:
## glm(formula = default ~ student + balance + income, family = "binomial",
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4691  -0.1418  -0.0557  -0.0203   3.7383
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
## AIC: 1579.5
##
## Number of Fisher Scoring iterations: 8
```

```
##
## Call:
## glm(formula = default ~ student + balance + income + student *
##     income + balance * student + balance * income, family = binomial(),
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4848  -0.1417  -0.0554  -0.0202   3.7579
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.104e+01  1.866e+00  -5.914 3.33e-09 ***
## studentYes         -5.201e-01  1.344e+00  -0.387    0.699
## balance             5.882e-03  1.180e-03   4.983 6.27e-07 ***
## income              4.050e-06  4.459e-05   0.091    0.928
## studentYes:income   1.447e-05  2.779e-05   0.521    0.602
## studentYes:balance -2.551e-04  7.905e-04  -0.323    0.747
## balance:income     -1.579e-09  2.815e-08  -0.056    0.955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.1  on 9993  degrees of freedom
## AIC: 1585.1
##
## Number of Fisher Scoring iterations: 8
```

d) How accurate is your model for predicting the response? What is the error rate?

Answer 4.d: The error rate for both the models is very low (~3%) and predicts the default customers with 97% accuracy. However, lower AIC of simpler logistic regression model (without interaction terms) suggests the simpler model is better in predicting the default customers ,avoids overfitting. This is further confirmed by the Chi square test which indicates the model with interaction terms is not significant at 0.05% significance level.

```
## Confusion Matrix and Statistics
##
##        True
## pred.4a   No  Yes
##     No  9627  228
##     Yes   40  105
##
##                Accuracy : 0.9732
##                  95% CI : (0.9698, 0.9763)
##     No Information Rate : 0.9667
##     P-Value [Acc > NIR] : 0.0001044
##
```

```
##                  Kappa : 0.4278
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##            Sensitivity : 0.9959
##            Specificity : 0.3153
##         Pos Pred Value : 0.9769
##         Neg Pred Value : 0.7241
##             Prevalence : 0.9667
##         Detection Rate : 0.9627
##   Detection Prevalence : 0.9855
##      Balanced Accuracy : 0.6556
##
##       'Positive' Class : No
##
```

```
## Error Rate of model without interaction terms: 0.0268
```

```
## Confusion Matrix and Statistics
##
##        True
## pred.4b   No  Yes
##     No  9627  229
##     Yes   40  104
##
##               Accuracy : 0.9731
##                 95% CI : (0.9697, 0.9762)
##     No Information Rate : 0.9667
##     P-Value [Acc > NIR] : 0.0001319
##
##                  Kappa : 0.4245
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##            Sensitivity : 0.9959
##            Specificity : 0.3123
##         Pos Pred Value : 0.9768
##         Neg Pred Value : 0.7222
##             Prevalence : 0.9667
##         Detection Rate : 0.9627
##   Detection Prevalence : 0.9856
##      Balanced Accuracy : 0.6541
##
##       'Positive' Class : No
##
```

```
## Error Rate of model with interaction terms: 0.0269
```

```
## Analysis of Deviance Table
##
## Model 1: default ~ student + balance + income
## Model 2: default ~ student + balance + income + student * income + balance *
##     student + balance * income
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      9996     1571.5
## 2      9993     1571.1  3  0.47911   0.9235
```

5. Go through Section 7.3.1 of HSAUR. Run all the codes (additional exploration of data is allowed) and write your own version of explanation and interpretation. *For this problem, please show the code of function you created as well as show the output. You can do this by adding* `echo = T` *to the code chunk header.*

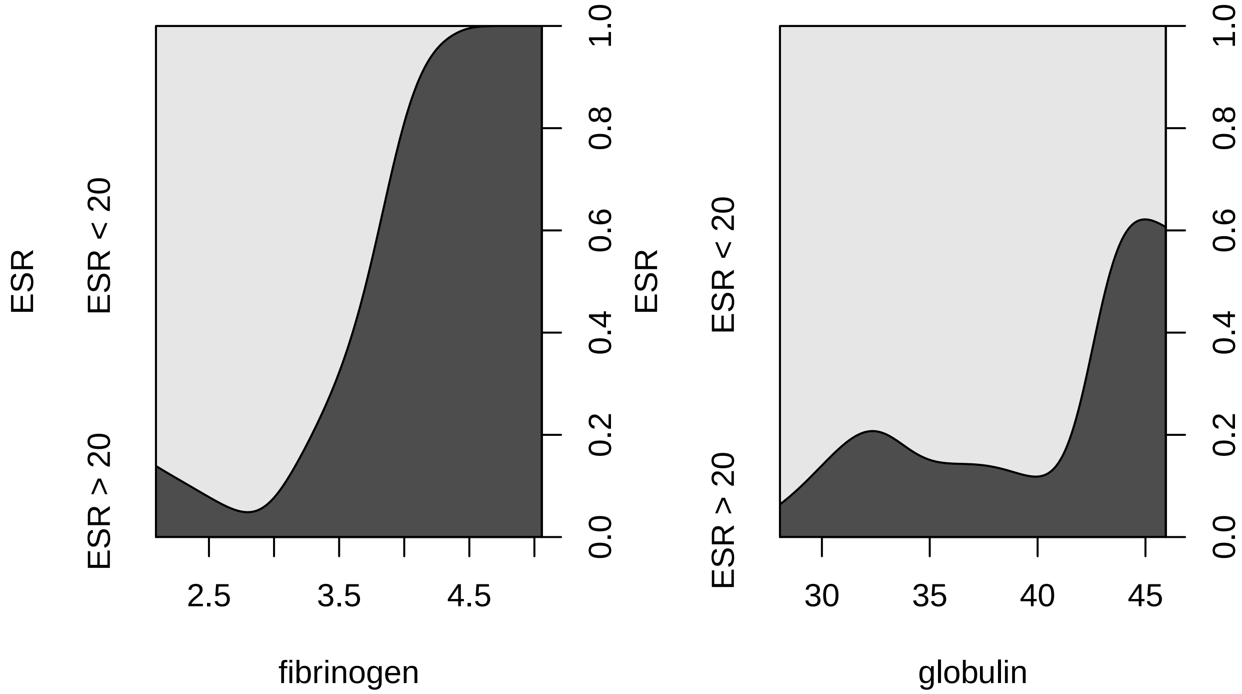
Answer 5: The density plots describe the distribution of variables fibrinogen and globulin.

```
# Exploring top data variables
head(plasma)
```

```
##   fibrinogen globulin      ESR
## 1       2.52       38 ESR < 20
## 2       2.56       31 ESR < 20
## 3       2.19       33 ESR < 20
## 4       2.18       31 ESR < 20
## 5       3.41       37 ESR < 20
## 6       2.46       36 ESR < 20
```

```
layout(matrix(1:2,ncol=2))

# Plotting density plot for ESR vs other features
cdplot(ESR ~ fibrinogen,data=plasma)
cdplot(ESR ~ globulin,data=plasma)
```



Fitting a Generalized linear model where family is binomial with a logit link function(logistic regression). ESR is the dependent variable and fibrinogen is the independent variable.The summary of the model shows that fibrinogen is significant at 0.05% significance level.The coefficient of fibrinogen indicates that a change from ESR<20 to ESR>20,increases the log odds in fav of ESR value greater than 20 by 1.83 times at 95% confidence interval.

```
# Fit the logistic regression model
plasma_glm_1 <-glm(ESR~fibrinogen,data=plasma,family=binomial())
# Summary of the model
summary(plasma_glm_1)
```

```
##
## Call:
## glm(formula = ESR ~ fibrinogen, family = binomial(), data = plasma)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9298  -0.5399  -0.4382  -0.3356   2.4794
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.8451     2.7703  -2.471   0.0135 *
## fibrinogen    1.8271     0.9009   2.028   0.0425 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30.885  on 31  degrees of freedom
## Residual deviance: 24.840  on 30  degrees of freedom
## AIC: 28.84
##
## Number of Fisher Scoring iterations: 5
```

Exponentiating the confidence intervals and coefficients will get the odds of the values which can me more helpful. The confidence interval is higher as number of observations with ESR value greater than 20 are less.

```
#confidence interval
confint(plasma_glm_1,parm="fibrinogen")
```

```
##     2.5 %    97.5 %
## 0.3387619 3.9984921
```

```
#Exponentiating the estimates
cat("odds of fibrinogen:",exp(coef(plasma_glm_1)["fibrinogen"]),"\n")
```

```
## odds of fibrinogen: 6.215715
```

```
# #Exponentiating the confidence intervels
exp(confint(plasma_glm_1,parm="fibrinogen"))
```

```
##     2.5 %    97.5 %
##  1.403209 54.515884
```

A logistic regression model with both the explanatory variables is fit and the summary of the model indicates that the variable,globulin is not significant at 95% confidence interval and the coefficint of globulin is almost zero and not very significant.

```
# Fitting logistic regression with both the explanatory variables.
plasma_glm_2 <-glm(ESR~fibrinogen+globulin,data=plasma,family=binomial())

# Summary of the new model
summary(plasma_glm_2)
```

```
##
## Call:
## glm(formula = ESR ~ fibrinogen + globulin, family = binomial(),
##     data = plasma)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9683  -0.6122  -0.3458  -0.2116   2.2636
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.7921     5.7963  -2.207   0.0273 *
## fibrinogen    1.9104     0.9710   1.967   0.0491 *
## globulin      0.1558     0.1195   1.303   0.1925
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30.885  on 31  degrees of freedom
## Residual deviance: 22.971  on 29  degrees of freedom
## AIC: 28.971
##
## Number of Fisher Scoring iterations: 5
```

The above fitted two nested models are compared by Chi Square test using the anova function.The AIC of the model with variable fibrinogen is less than the one with both variables(28.84 vs 28.971) indicating model with fibrinogen is better over later model. Subtracting residual deviance of model 2 from model 1 (24.84-22.97) is 1.87,with single degree of freedom and high p value,this added variable is not improving the model and is not significant at 5% level.So, we can conclude that the globulin is not significant in predicting the ESR.

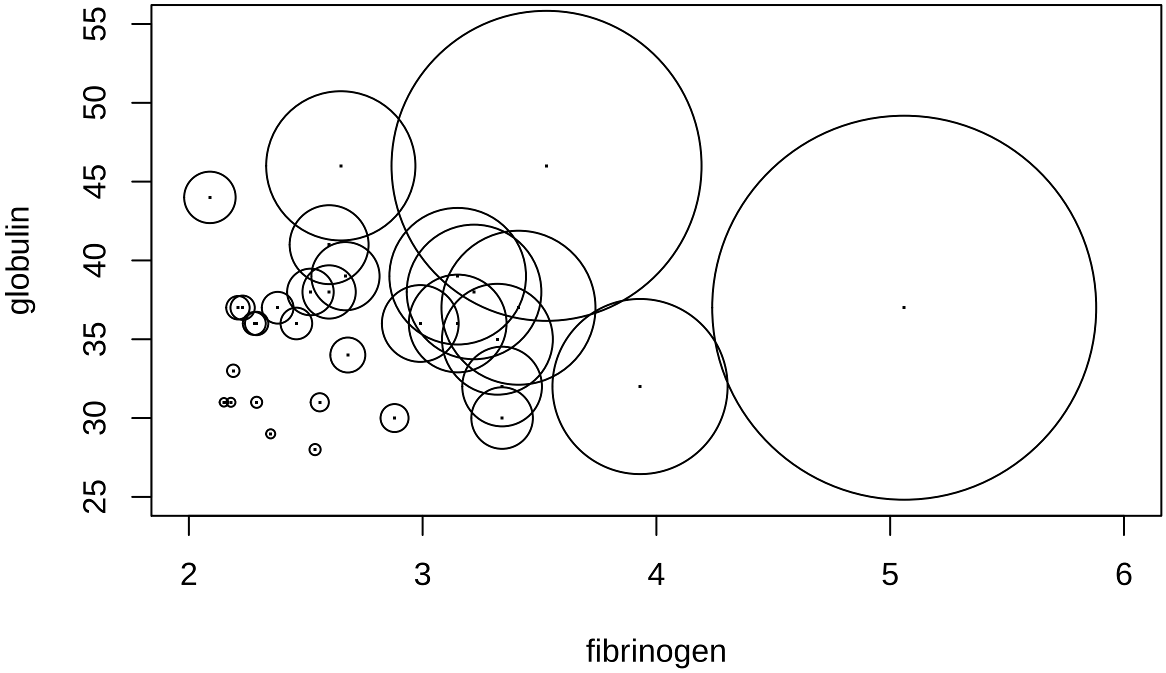
```
# comparing the models with anova function
anova(plasma_glm_1,plasma_glm_2,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: ESR ~ fibrinogen
## Model 2: ESR ~ fibrinogen + globulin
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        30     24.840
## 2        29     22.971  1   1.8692   0.1716
```

```
# Plotting the predicted values from second model against both explanatory variables

prob <- predict(plasma_glm_2,type="response")

plot(globulin~fibrinogen,data=plasma,xlim=c(2,6),ylim=c(25,55),pch=".")
symbols(plasma$fibrinogen,plasma$globulin, circles= prob,add=TRUE)
```


globulin
fibrinogen