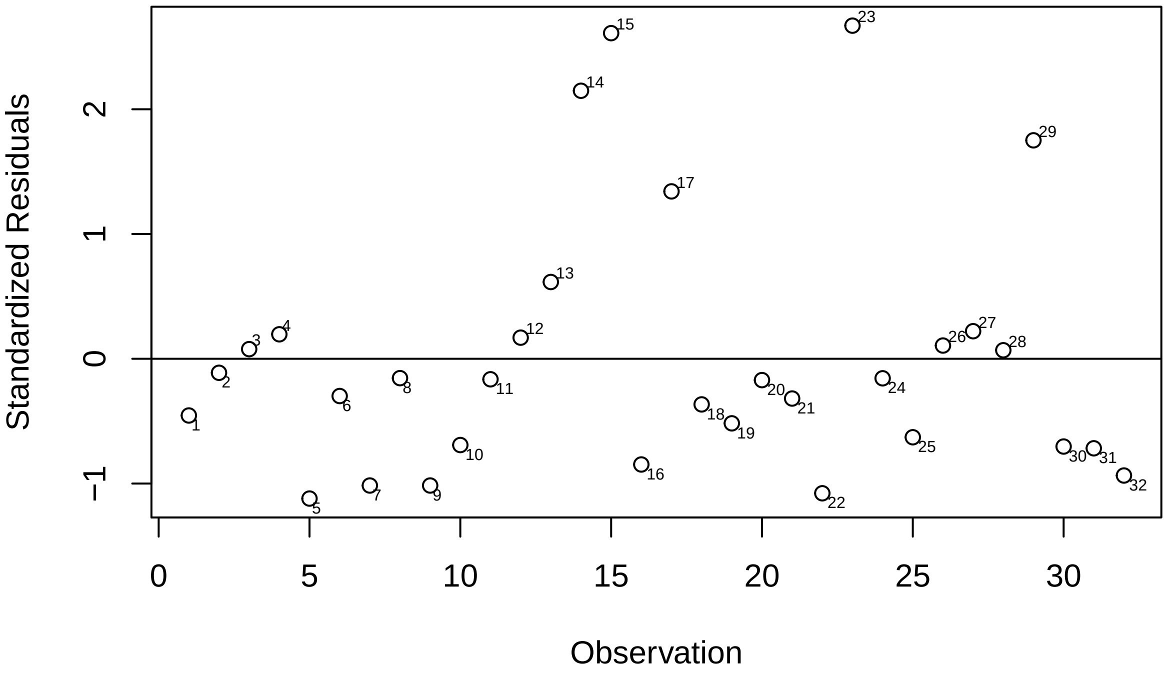
Data analysis from the HSAUR

Yamuna Dhungana

Following answers are from the text book R Handbook.

1. 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. (7.2 on Handbook)



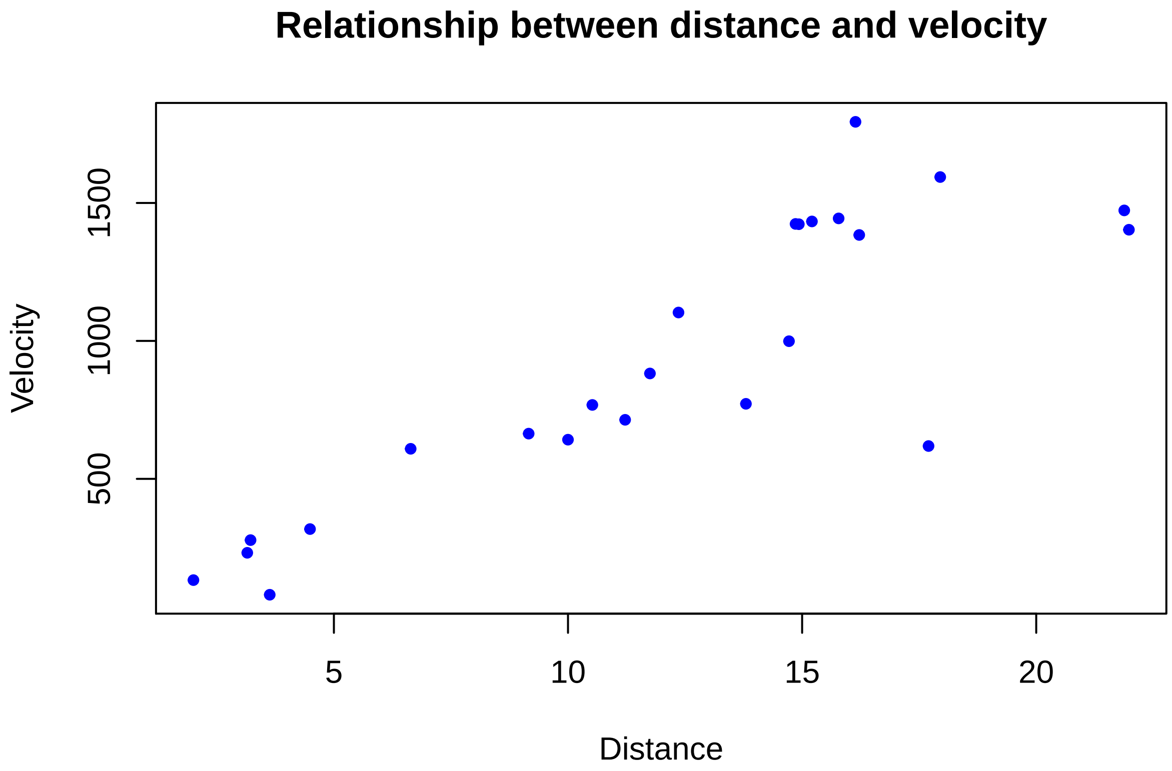
As we already know that the anything below first quartile + 1.5(IRQ) and anything above the third quartile - 1.5(IRQ) is considered a outliers.From the graphs, we came to know that point 15 and the point 23 lies farthest from point zero. Therefore, the point 15 and the point 23 is the outliers.

2. (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

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

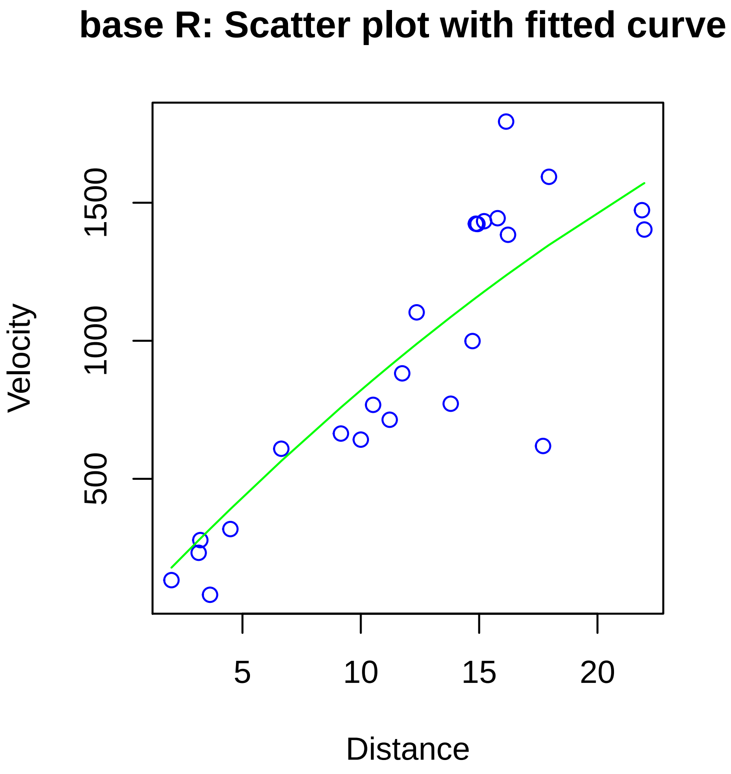
```
## Visualizing the data
```

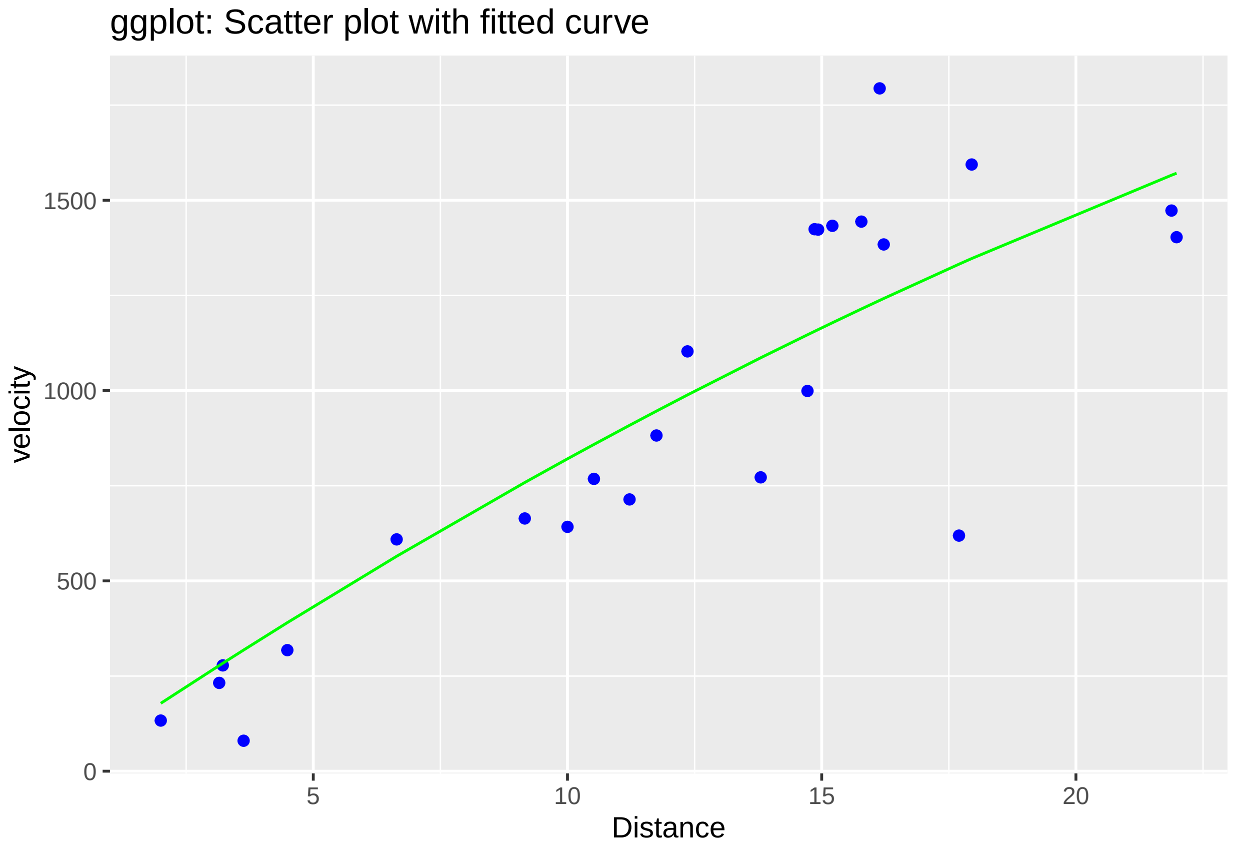


```
##
## Call:
## lm(formula = y ~ x + x2 - 1, data = hubble)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -713.15 -152.76  -54.85  163.92  557.01
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## x   90.9046    16.5726   5.485 1.64e-05 ***
## x2  -0.8837     0.9925  -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
```

I have performed the quadratic model whose basic formula is y = axˆ2 + b x + c where C = Y-intercept. In the model, -1 is used to exclude y-intercept because the intercept hasn't added anything to the model. When we remove -1 from the model, we get intercept -196.364 which means when x(distance) = 0 the velocity will be -196.364, which is not possible so, the intercept is removed. The residuals are the distance of data from the fitted line. From the model, we find out that the median of the data is 29.7. Data is slightly skewed towards left. We can find this by analyzing the data of q1 and q3, if the value of q1 and q3 is equidistance from the median, then the data will be normally distributed. The coefficients say the values for the least-squared estimates for the fitted lines. The standard error and the t-value identifies how P-value were calculated. The most important to view in summary is P-value. The model is identified as statistically significant if the value of P is less than or equal to 0.05. Here we will consider the value of intercept but will need the value of x and x-square. The value of ‘x’ is 0.002, which is statistically significant.The standard error of the model is 260.1 with 21 DF and the multiple R-squared value is 0.7651, which means the value x can explain 76.51% in the ‘y’.

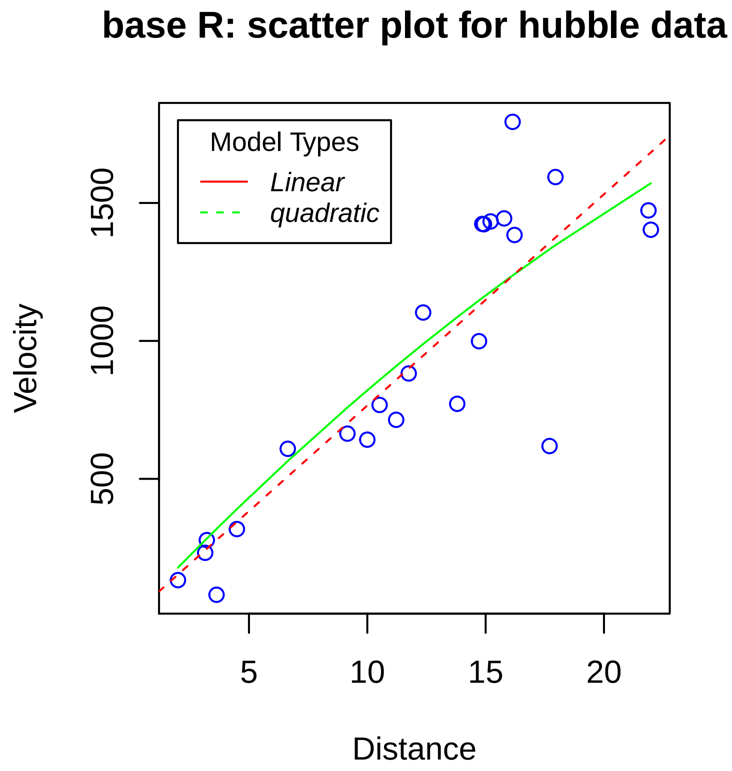
b) Plot the fitted curve from Model 2 on the scatterplot of the data





The above graphs is the base r and the ggplot graphs that explains the fitted model. The model we performed here is a quadratic model.

c) Add the simple linear regression fit (fitted in class) on this plot-use different color and line type to differentiate the two and add a legend to your plot.



The red color is the simple linear regression model and the green denotes the quadratic model of the data.

d) Which model do you consider most sensible considering the nature of the data -looking at the plot?

By looking at the graphs, linear regression fits the model better than the logistic regression because the linear regression includes the many points than the quadratic model. Also, the points seem to follow the line from bottom to top. However, there is no much difference between the two models.

e) Which model is better? - provide a statistic to support you claim. 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.

```
##
## Call:
## lm(formula = y ~ x + x2 - 1, data = hubble)
##
## 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
```

```
##
## 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
```

```
## Adjusted R-square
```

Table 1: Adjusted R-square

Quadratic	Linear
0.9388554	0.9394063

The statistic appears to support the simple linear regression as the better one. Here the simple linear regression's adjusted R is 0.9394, and the quadratic regression's adjusted R is 0.7428, which explains more variability in the data then the quadratic model.

3. 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).

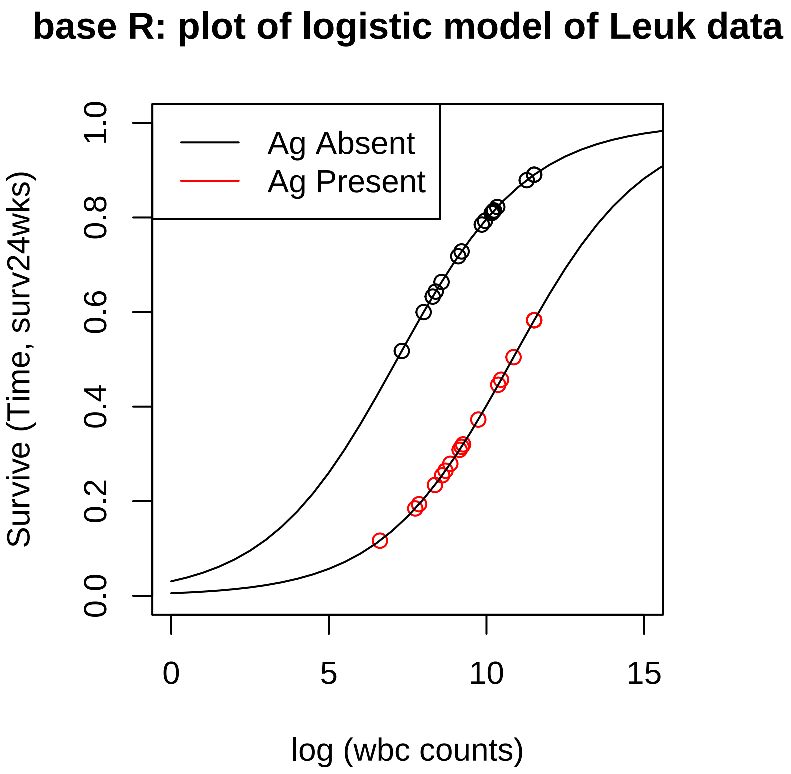
a) Define a binary outcome variable according to whether or not patients lived for at least 24 weeks after diagnosis. Call it *surv24*.

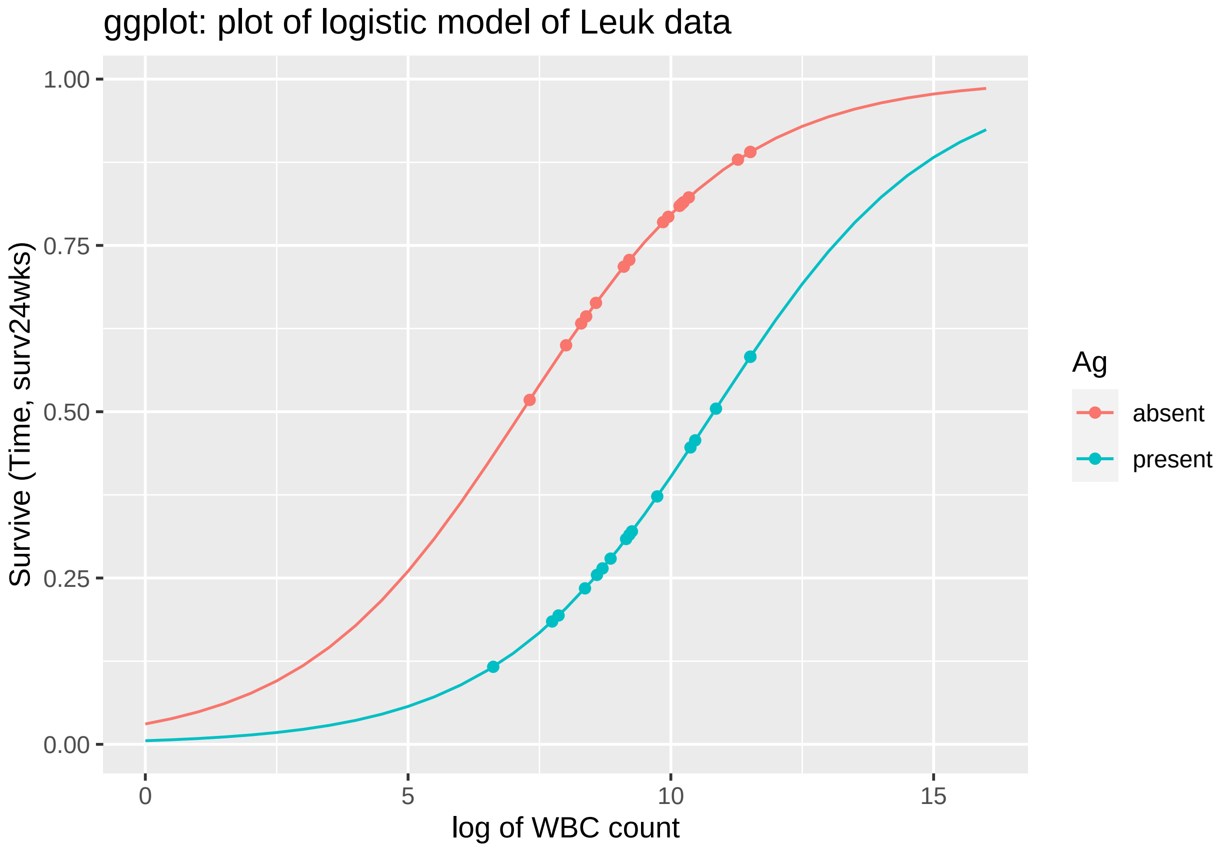
b) Fit a logistic regression model to the data with *surv24* as response. It is advisable to transform the very large white blood counts to avoid regression coefficients very close to 0 (and odds ratio close to 1). You may use log transformation.

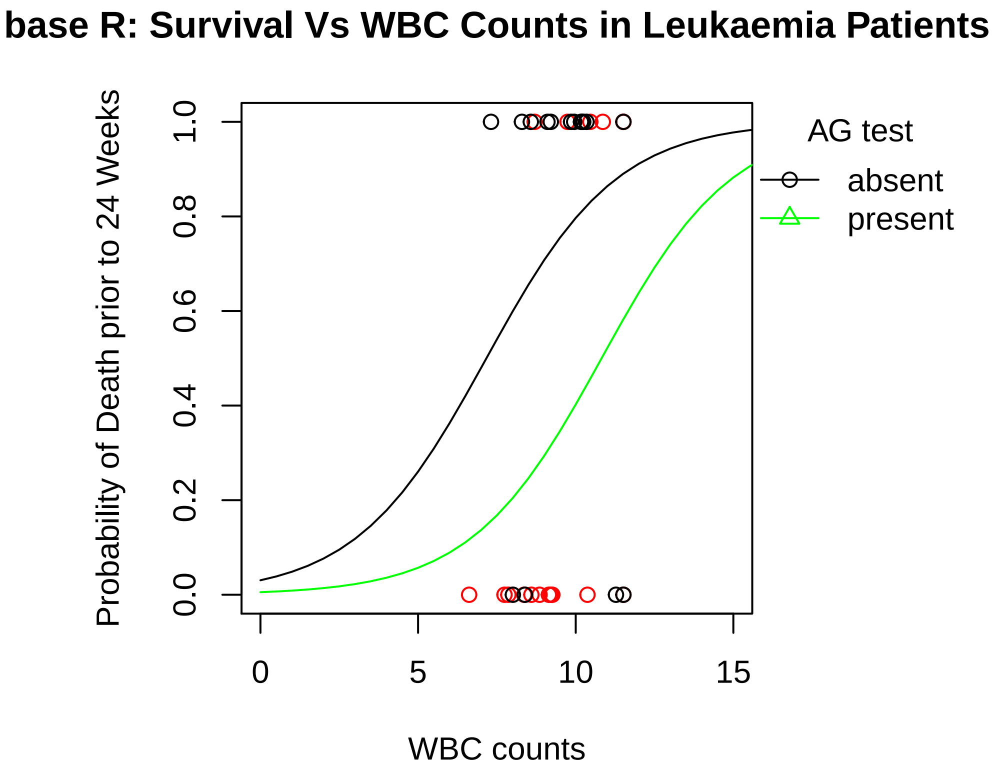
```
##
## Call:
## glm(formula = surv24 ~ log(wbc) + ag, family = binomial, data = leuk)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1032  -0.8592   0.6258   0.9056   1.6310
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.4556     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
```

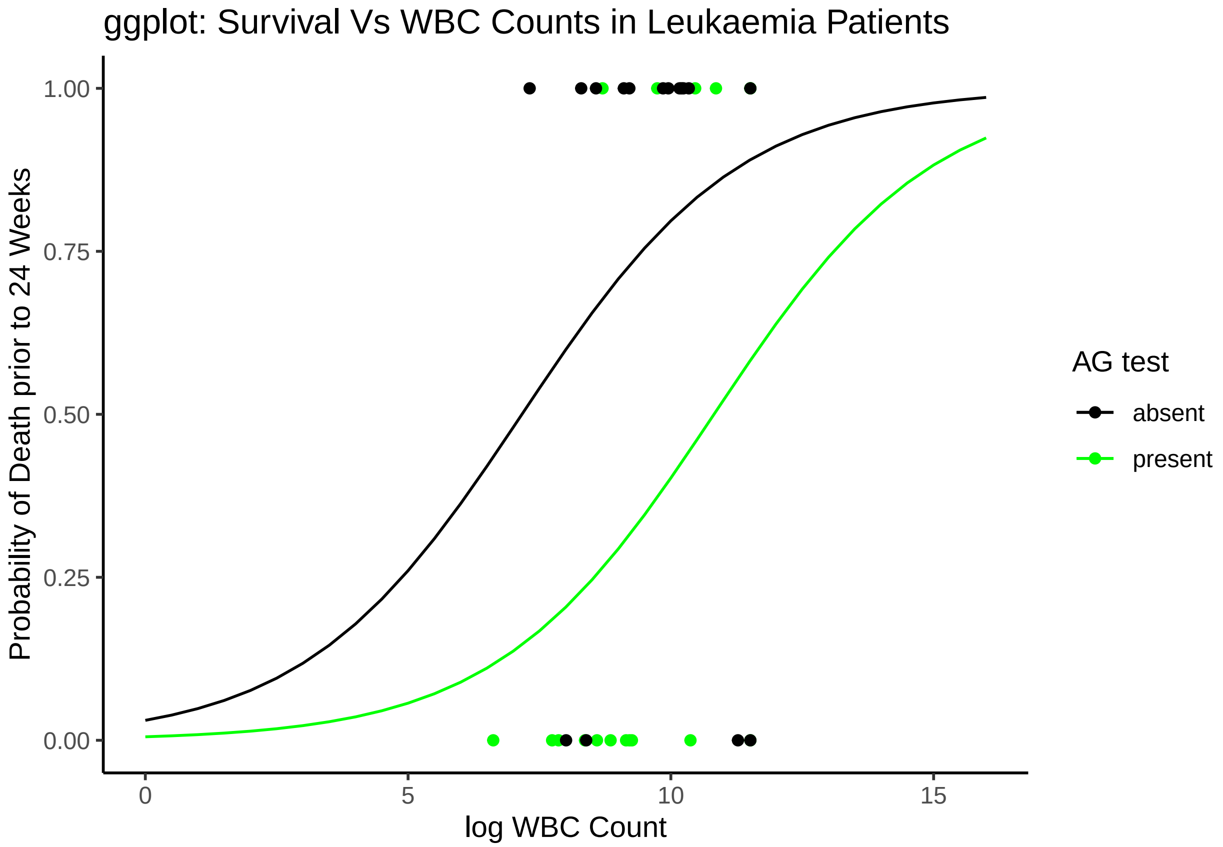
Here, I have performed the logistic model.The residuals are the distance of data from the fitted line. In the model, the median value is 0.6258. Data is normally distributed because the distance between q1 and the median is nearly equal to the distance between median and the q3. If the value of q1 and q3 is equidistance from the median, then the data is considered to be normally distributed.The coefficients say the values for the least-squared estimates for the fitted lines. The standard error and the t-value identifies how P-value is calculated. The most important value to view in summary its P-value. The model is identified as statistically significant if the value of P is less than or equal to 0.05. Here, we may consider the value of intercept but necessarily needed. The value of log(WBC), the p-value is 0.1531, which is not statistically significant. The value of the ag present is which is statistically significant with the P-value if 0.02, which means the Ag present can add the significant in the model.

c) Construct some graphics useful in the interpretation of the final model you fit.




ggplot: plot of logistic model of Leuk data
Survive (Time, surv24wks)
1.00
0.75
0.50
0.25
0.00
0
5
10
15
log of WBC count
Ag
absent
present


base R: Survival Vs WBC Counts in Leukaemia Patients
Probability of Death prior to 24 Weeks
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
WBC counts
AG test
absent
present



d) Fit a model with an interaction term between the two predictors. Which model fits the data better? Justify your answer.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.5946168	4.6583072	0.5569870	0.5775363
log(wbc)	-0.1545345	0.4745961	-0.3256127	0.7447174
agpresent	-13.6306111	7.0908703	-1.9222762	0.0545710
log(wbc):agpresent	1.2315041	0.7181873	1.7147394	0.0863930

	Adjusted R-square values
Linear model	0.1890705
Linear model with interation	0.2361365

Since the adjusted R square is higher for the model with interaction. Therefore the model with the interaction fits the model better.

4. Load the **Default** dataset from **ISLR** library. The dataset contains information on ten thousand customers. The aim here is to predict which customers will default on their credit card debt. It is a four-dimensional dataset with 10000 observations. The question of interest is to predict individuals who will default . We want to examine how each predictor variable is related to the response (default). Do the following on this dataset

a) Perform descriptive analysis on the dataset to have an insight. Use summaries and appropriate exploratory graphics to answer the question of interest.

b) Use R to build a logistic regression model.

c) Discuss your result. Which predictor variables were important? Are there interactions?

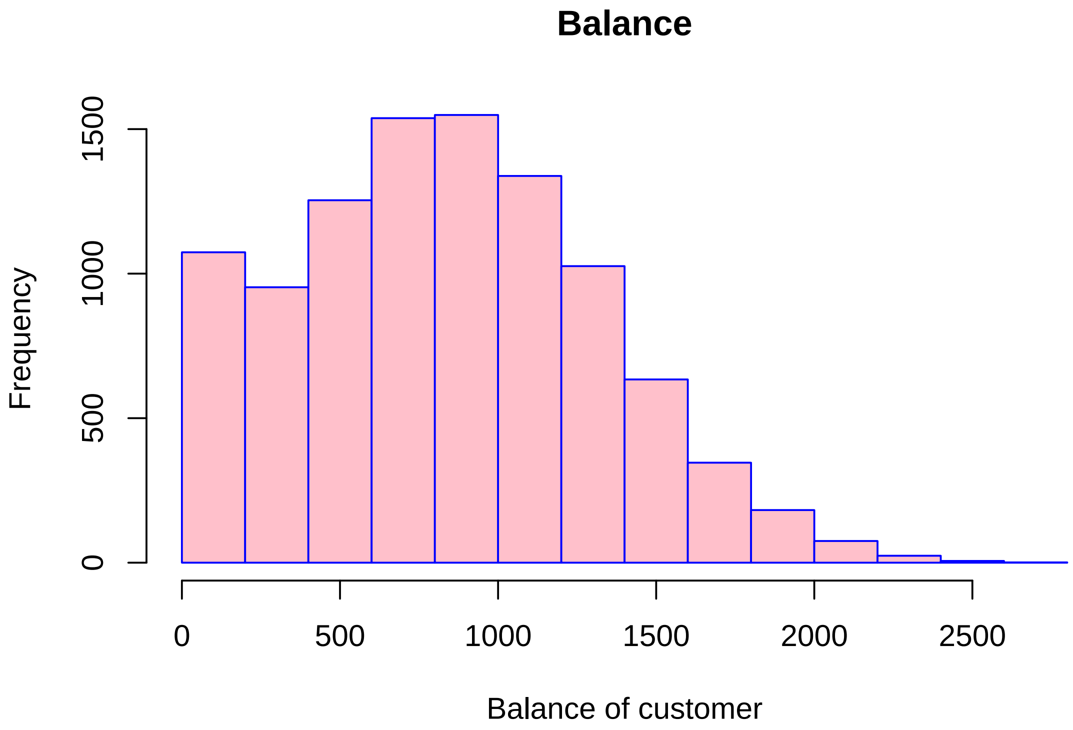
d) How good is your model? Assess the performance of the logistic regression classifier. What is the error rate?

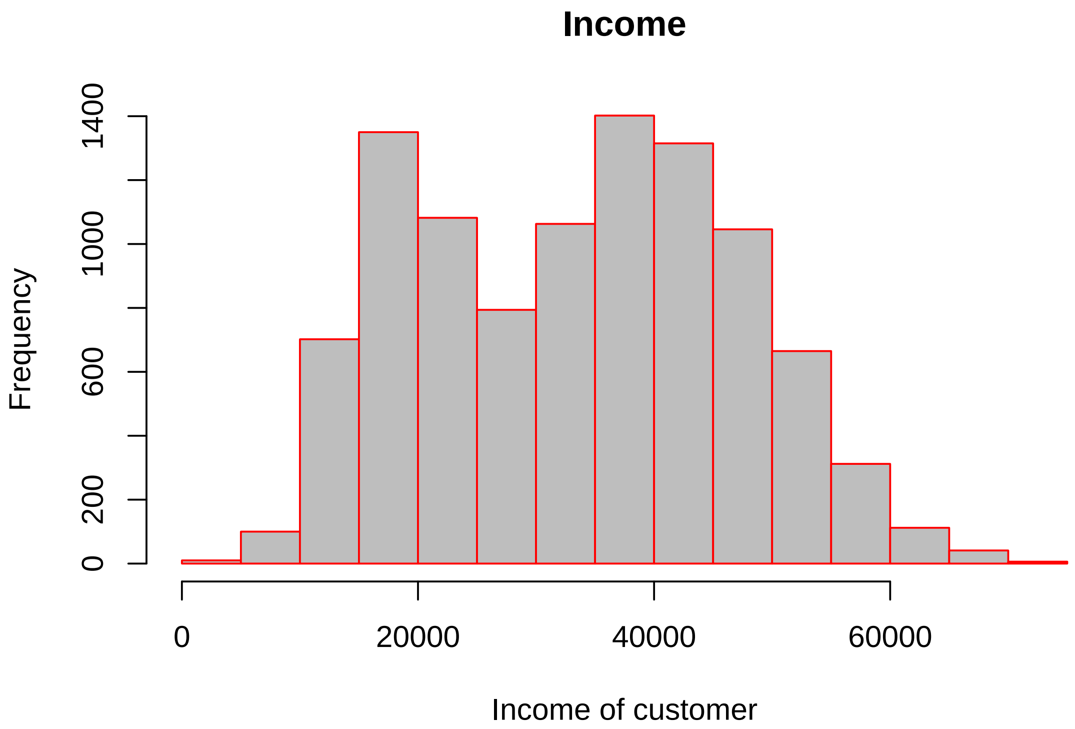
Table 4: summary of default and the student status

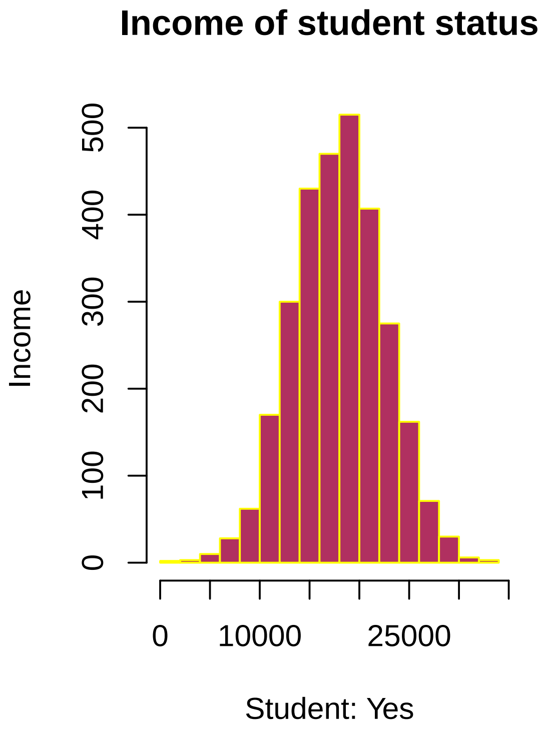
default	student
No :9667	No :7056
Yes: 333	Yes:2944

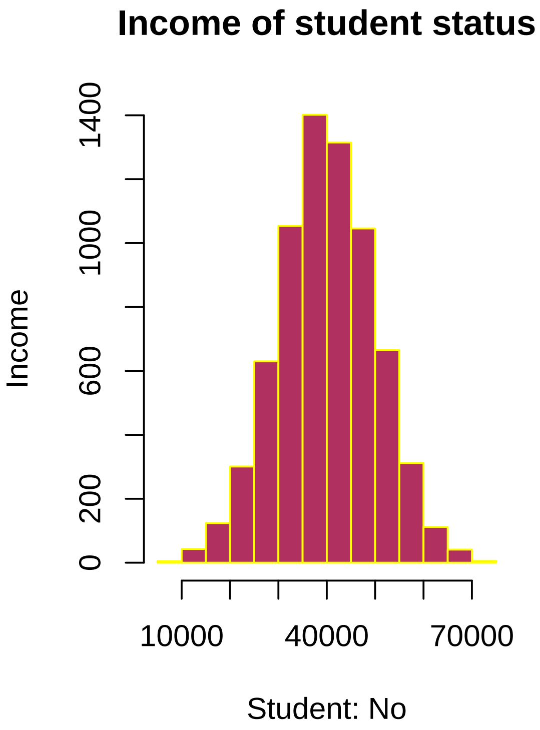
Table 5: Summary of Balance and Income

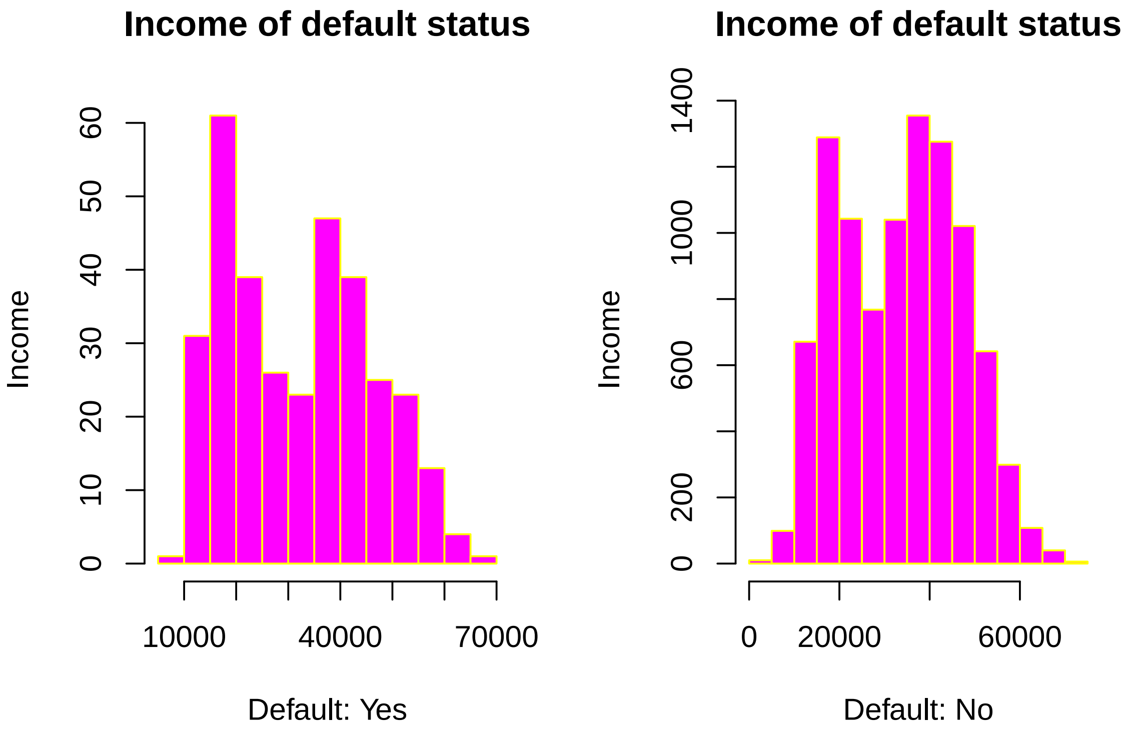
balance	income
Min. : 0.0	Min. : 772
1st Qu.: 481.7	1st Qu.:21340
Median : 823.6	Median :34553
Mean : 835.4	Mean :33517
3rd Qu.:1166.3	3rd Qu.:43808
Max. :2654.3	Max. :73554

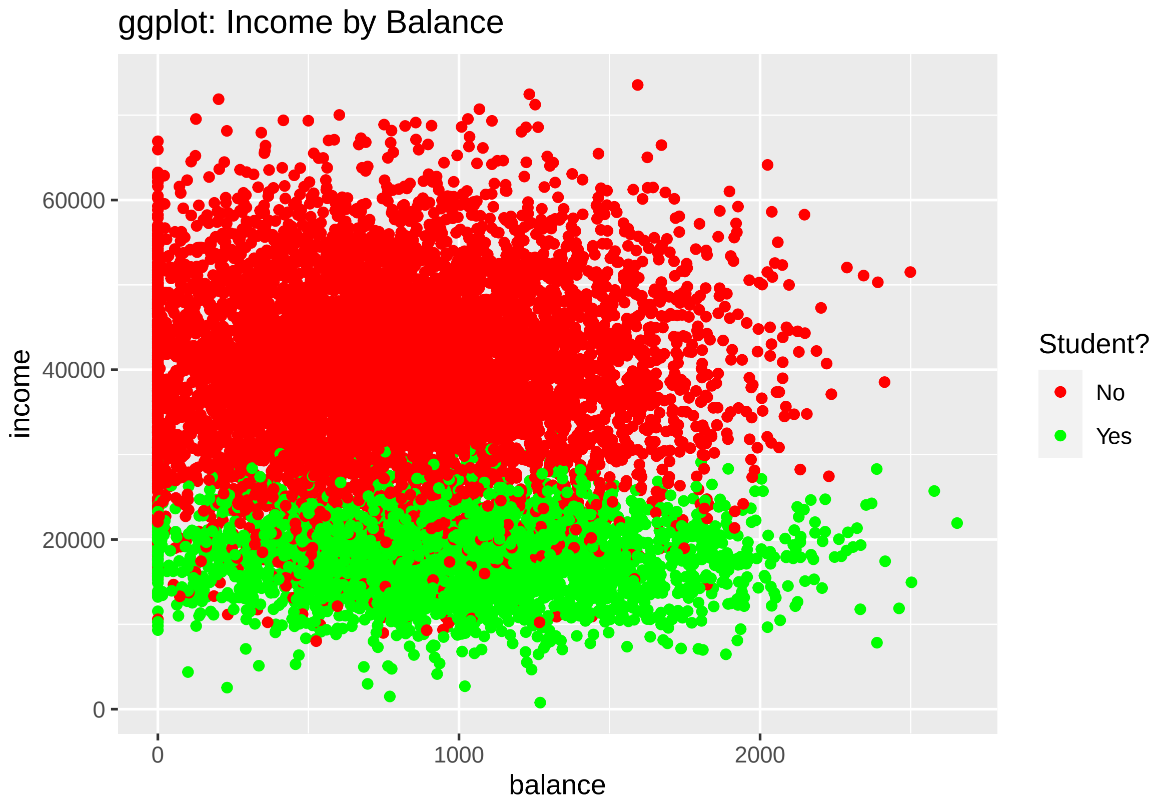


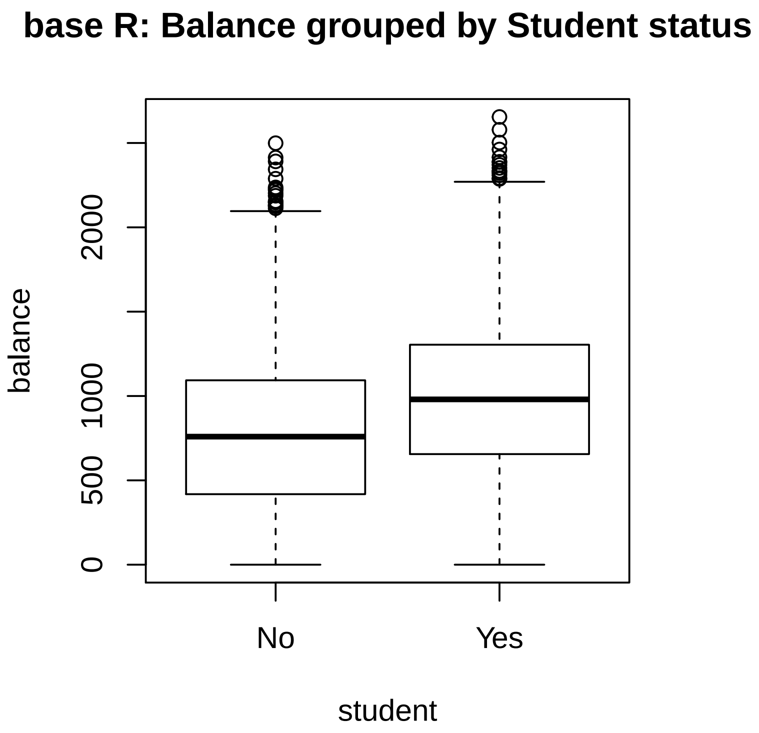

Income
Frequency
0
200
600
1000
1400
0
20000
40000
60000
Income of customer

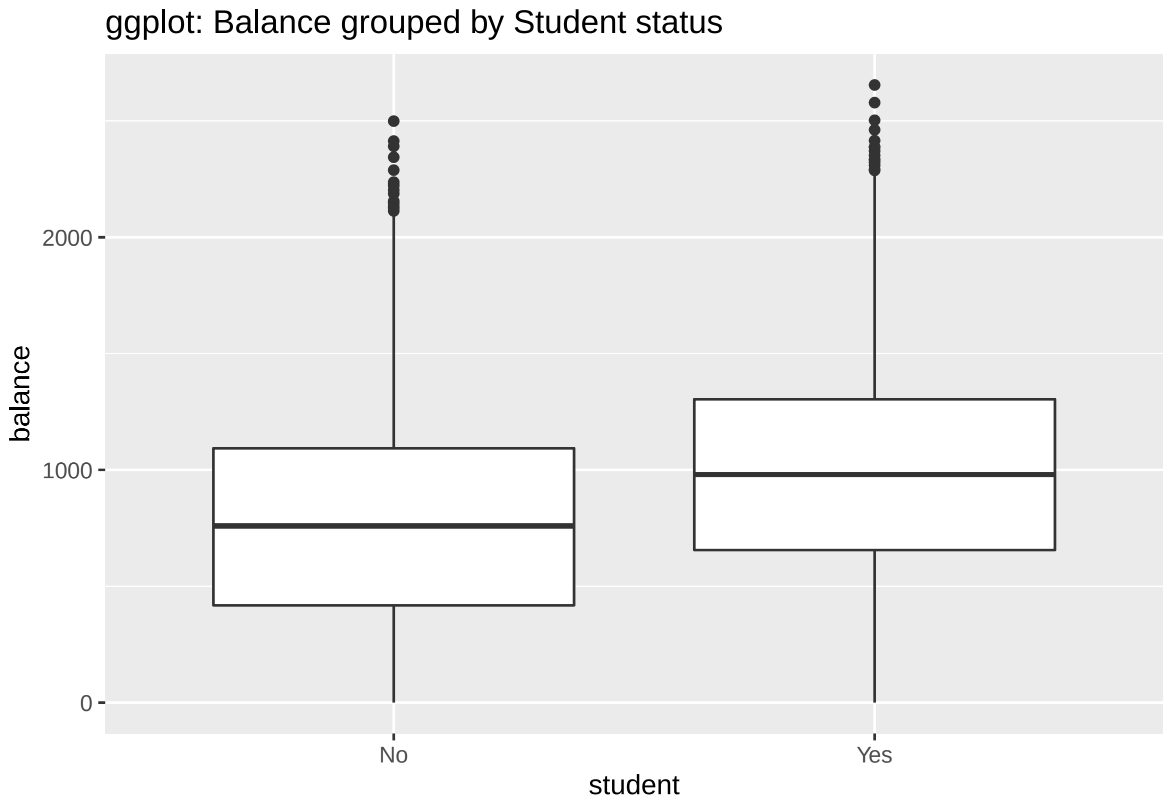

Income of student status
Income
0
100
200
300
400
500
0
10000
25000
Student: Yes

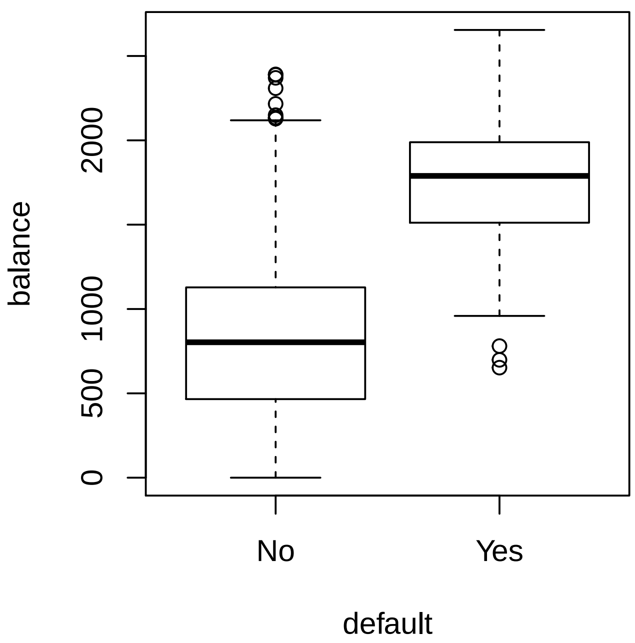

Income of student status
Income
0
200
600
1000
1400
10000
40000
70000
Student: No


Income of default status
Income
0
10
20
30
40
50
60
10000
40000
70000
Default: Yes
Income of default status
Income
0
200
600
1000
1400
0
20000
60000
Default: No

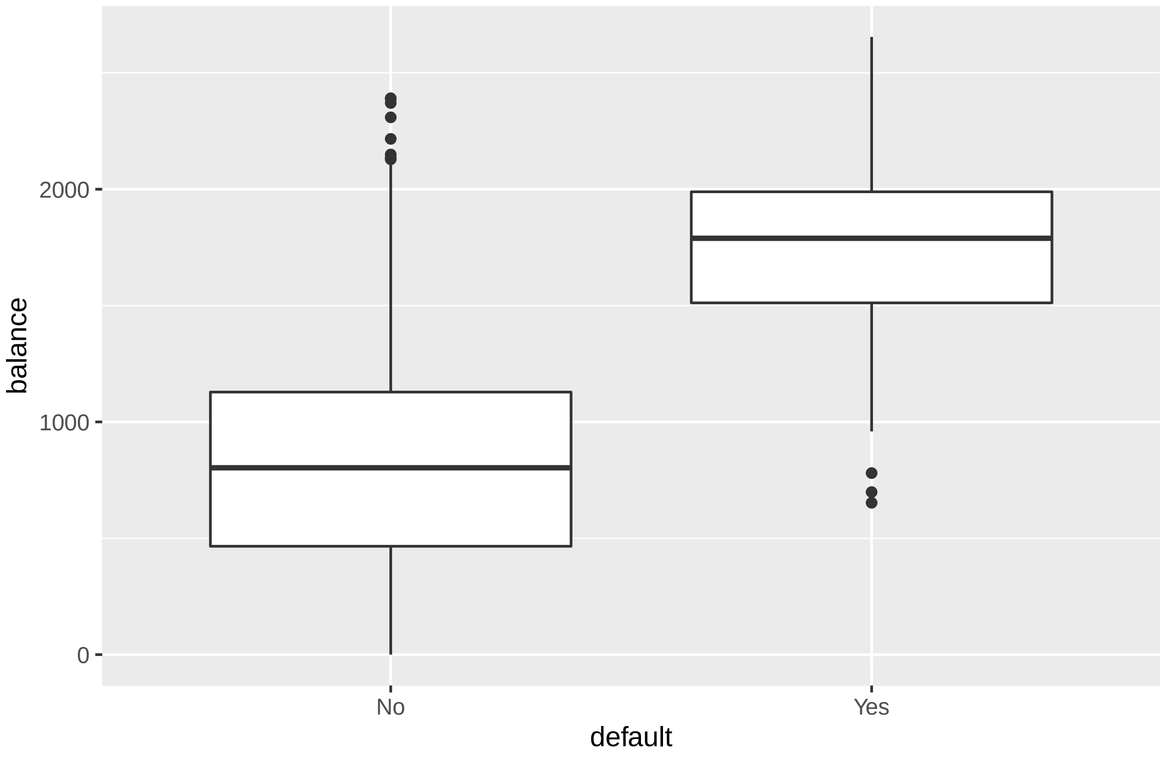

ggplot: Income by Balance
income
0
20000
40000
60000
0
1000
2000
balance
Student?
No
Yes


base R: Balance grouped by Student status
balance
0
500
1000
2000
No
Yes
student

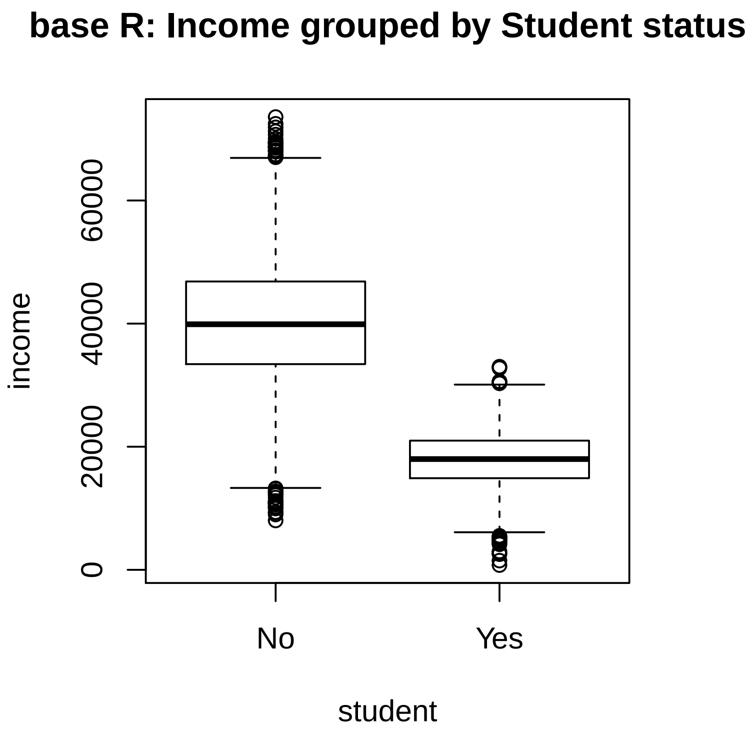

ggplot: Balance grouped by Student status
balance
0
1000
2000
No
Yes
student

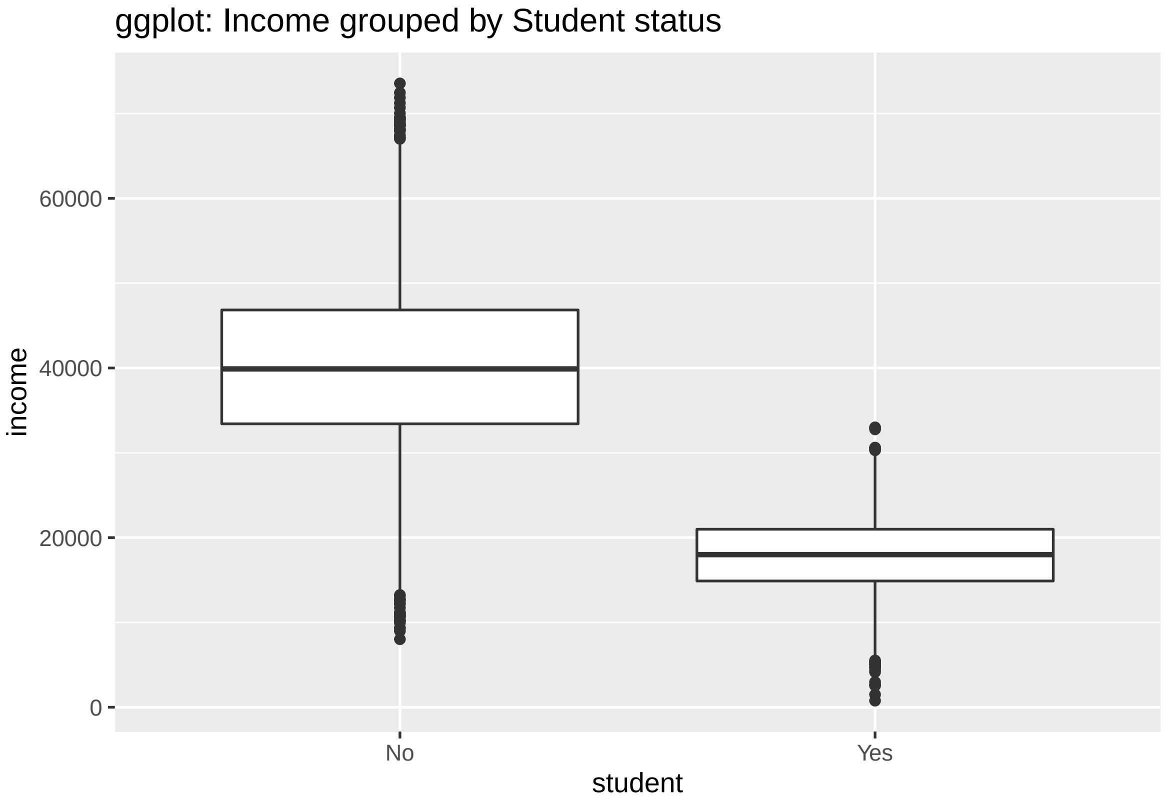


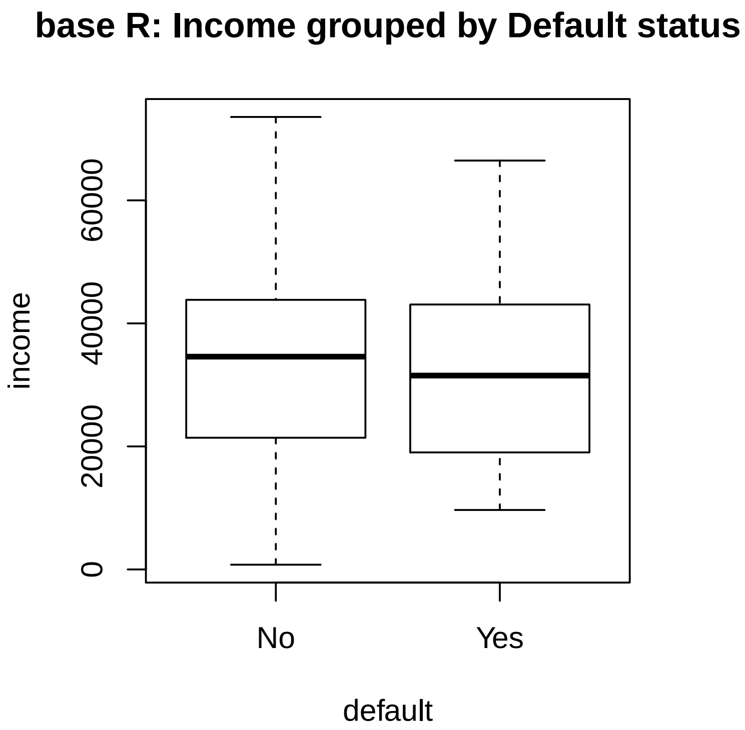
base R: Balance grouped by Default status

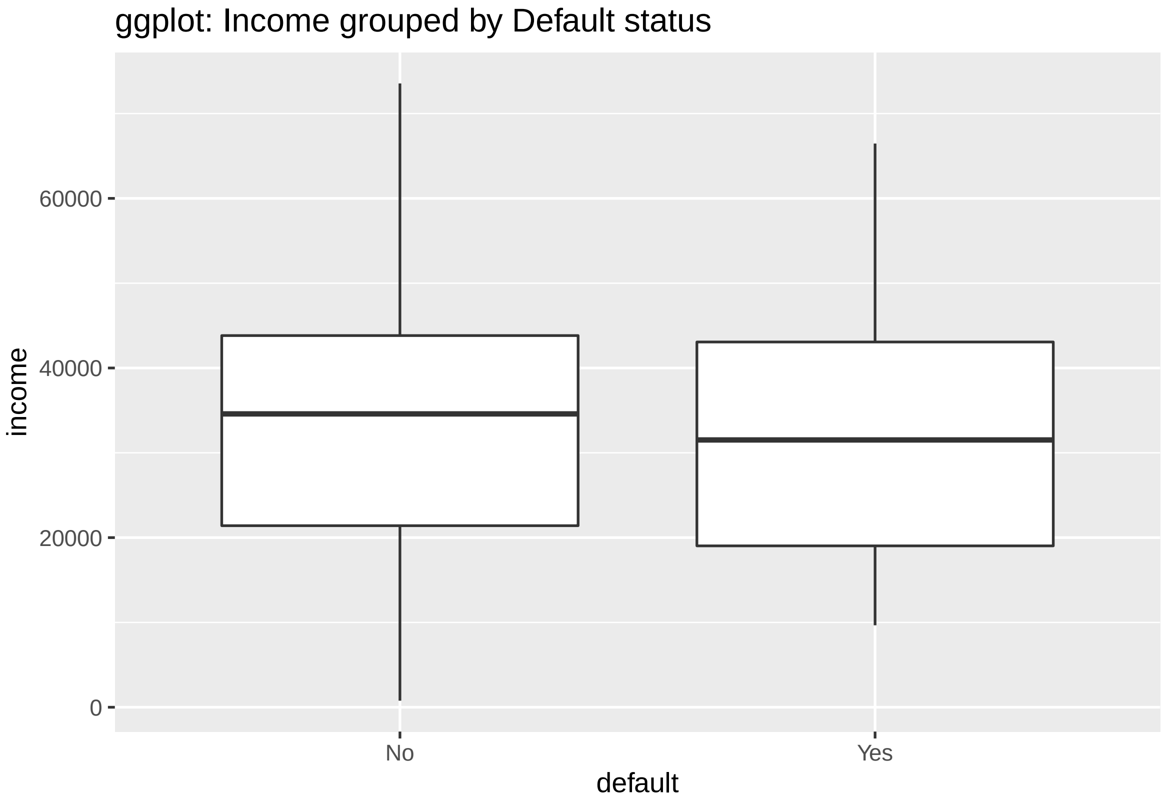


ggplot: Balance grouped by Default status


base R: Income grouped by Student status
income
0
20000
40000
60000
No
Yes
student


ggplot: Income grouped by Student status
income
0
20000
40000
60000
No
Yes
student


base R: Income grouped by Default status
income
0
20000
40000
60000
No
Yes
default


ggplot: Income grouped by Default status
income
0
20000
40000
60000
No
Yes
default

```
## $Defaulted
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9664   19028   31515   32089   43067   66466
##
## $‘Not Defaulted‘
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     772   21405   34589   33566   43824   73554
```

```
## $Defaulted
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   652.4  1511.6  1789.1  1747.8  1988.9  2654.3
##
## $‘Not Defaulted‘
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   465.7   802.9   803.9  1128.2  2391.0
```

```
## #B. Use R to build a logistic regression model
```

```
##
## Call:
## glm(formula = binary_def ~ std + blnc + incm, family = binomial())
##
## 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 ***
## std         -6.468e-01  2.363e-01  -2.738  0.00619 **
## blnc         5.737e-03  2.319e-04  24.738  < 2e-16 ***
## incm         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
```

```
## Then with interactions:
```

```
##
## Call:
## glm(formula = binary_def ~ std + blnc + incm + std * blnc + std *
##     incm + blnc * incm, family = binomial())
##
## 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 ***
## std         -5.201e-01  1.344e+00  -0.387    0.699
## blnc         5.882e-03  1.180e-03   4.983 6.27e-07 ***
## incm         4.050e-06  4.459e-05   0.091    0.928
## std:blnc    -2.551e-04  7.905e-04  -0.323    0.747
## std:incm     1.447e-05  2.779e-05   0.521    0.602
## blnc:incm   -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. Error Rate
```

```
##               True
## Predicted.o     Defaulted Not Defaulted
##   Defaulted           105            40
##   Not Defaulted       228          9627
```

```
## [1] 0.0268
```

```
##               True
## Predicted1      Defaulted Not Defaulted
##   Defaulted           104            40
##   Not Defaulted       229          9627
```

```
## [1] 0.0269
```

```
## analysis of variance
```

```
## Analysis of Deviance Table
##
## Model 1: binary_def ~ std + blnc + incm
## Model 2: binary_def ~ std + blnc + incm + std * blnc + std * incm + blnc *
##     incm
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      9996     1571.5
## 2      9993     1571.1  3  0.47911   0.9235
```

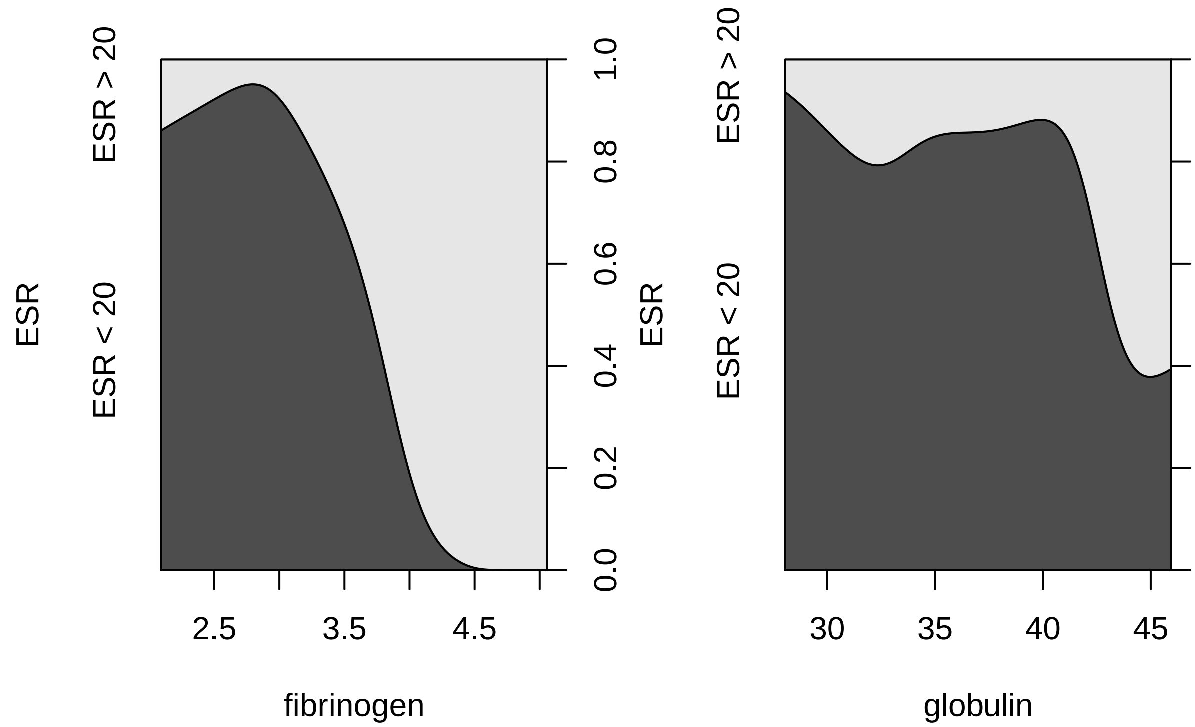
Based on the output of the data following result can be seen:

- Fewer people default than don’t default. -Defaulters and non-defaulters appear to have the same income range, given student status. -Defaulters appear to have higher balances. -If students default, they likely do it with over $1,000 balance. -If non-students default, they are likely do it with over $500 balance.

Without taking interactions into account, it appears that two predictors-student and balance are significant. With interactions involved, it appears that only balance predictor is important.

The model with interaction has the AIC 1585.1 and 1579.5 for the model without interaction. The value is slightly higher. Therefore, the model with AIC is better.Also,since analysis of deviance also shows that the chi-square test has no significance at 5% level, we can conclude that both models are almost the same as a working model

5. Go through Section 7.3.1 of the Handbook. Run all the codes (additional exploration of data is allowed) and write your own version of explanation and interpretation.



Form the graphs it appears as the value of fibrinogen to the ESR drops drastically. Likewise in globulin the value does not drops drastically.

Performing the logistic regression

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

```
##
## 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
```

The summary output indicates a 5% significance of fibrinogenand and increase of the log-odds of ESR > 20 by about 1.83 with confidence interval (CI) of 0.33 to 3.99.

```
## fibrinogen
##   6.215715
```

Fibrinogen might have value as a predictor of ESR. To make the results more readable, it is useful to apply an exponent function. This exponenetiates the log-odds of fibriogen and CI to correspond with the data.

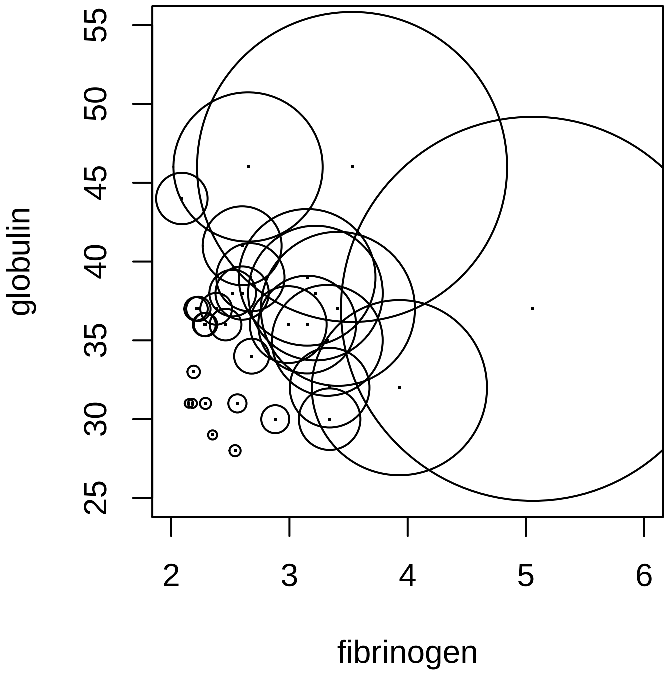
```
##     2.5 %    97.5 %
##  1.403209 54.515884
```

We can also perform logistic regression of both fibrinogen and globulin and text for the deviance.

```
##
## 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
```

```
## 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
```

Now, we can make the bubble plot of the predicted values of model II (plasma_glm_2). The plot shows that the probablity of ‘good’ ESR reading increases as fibrinogen increases. This is true of globulin only up to a point.



Bubble plot of the predicted values of model II