Generlized Linear Models-1

Mohamed Ahmed

9/22/2020

Load libraries

Instructions

Answer all questions stated in each problem. Discuss how your results address each question.

Submit your answers as a pdf, typeset (knitted) from an Rmd file. Include the Rmd file in your submission. You can typeset directly to PDF or typeset to Word then save to PDF In either case, both Rmd and PDF are required. If you are having trouble with .rmd, let us know and we will help you.

This file can be used as a template for your submission. Please follow the instructions found under "Content/Begin Here" titled . No code should be included in your PDF submission unless explicitly requested. Use the `echo = F` flag to exclude code from the typeset document.

For any question requiring a plot or graph, answer the question first using standard R graphics (See ?graphics). Then provide a equivalent answer using `library(ggplot2)` functions and syntax. You are not required to produce duplicate plots in answers to questions that do not explicitly require graphs, but it is encouraged.

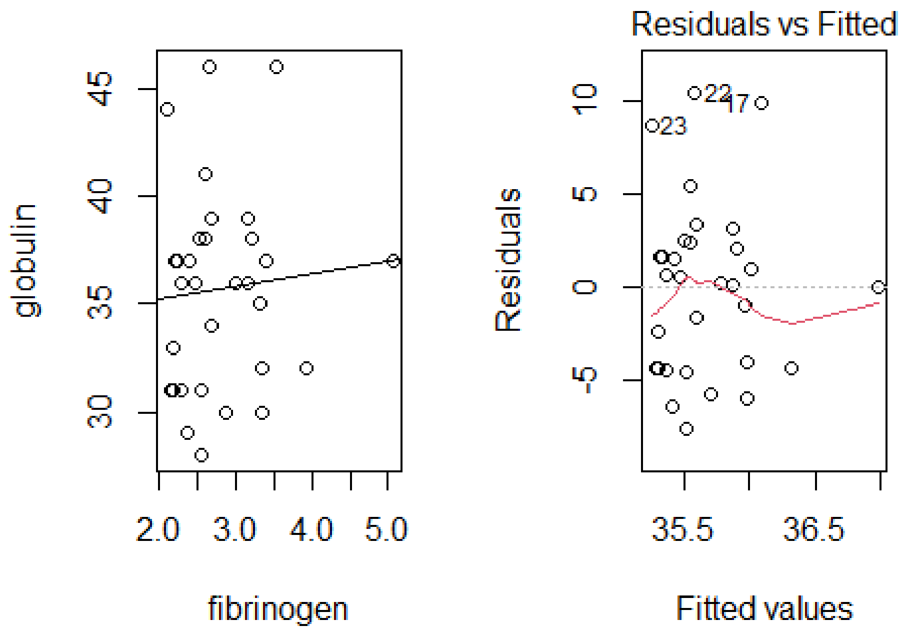
You can remove the `Instructions` section from your submission.

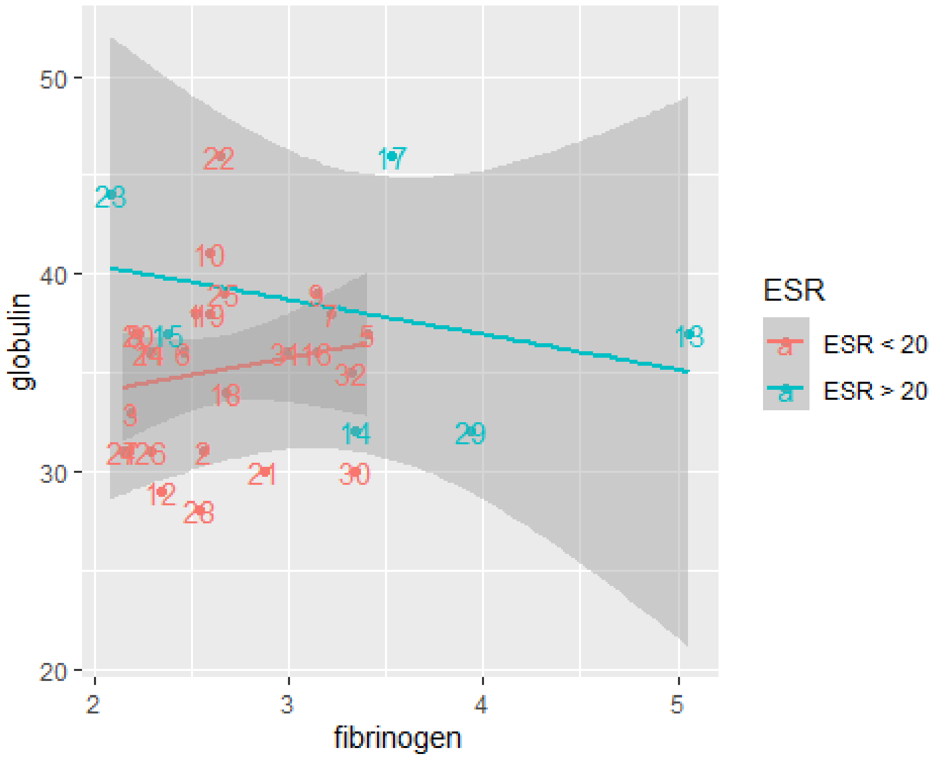
Exercises

Please answer the following questions from **Handbook of Statistical Analyses in R** (HSAUR) and the written questions. Refer to **R Graphics Cookbook or Modern Data Science with R** for any ggplots.

1. (Ex. 7.2 in HSAUR, modified for clarity) Collett (2003) argues that two outliers need to be removed from the 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: I have constructed a simple linear regression model to be able to visually inspect the residuals. I have constructed a scatter plot to see the relationship between the two variables and try to spot out outliers. Initially, I have noticed that there is a group of three data points at the top left of the scatter plot that are kind of separated from where the data points are. By inspecting the Residuals vs Fitted, I was able to confirm that point (17,22,23) are outliers.





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

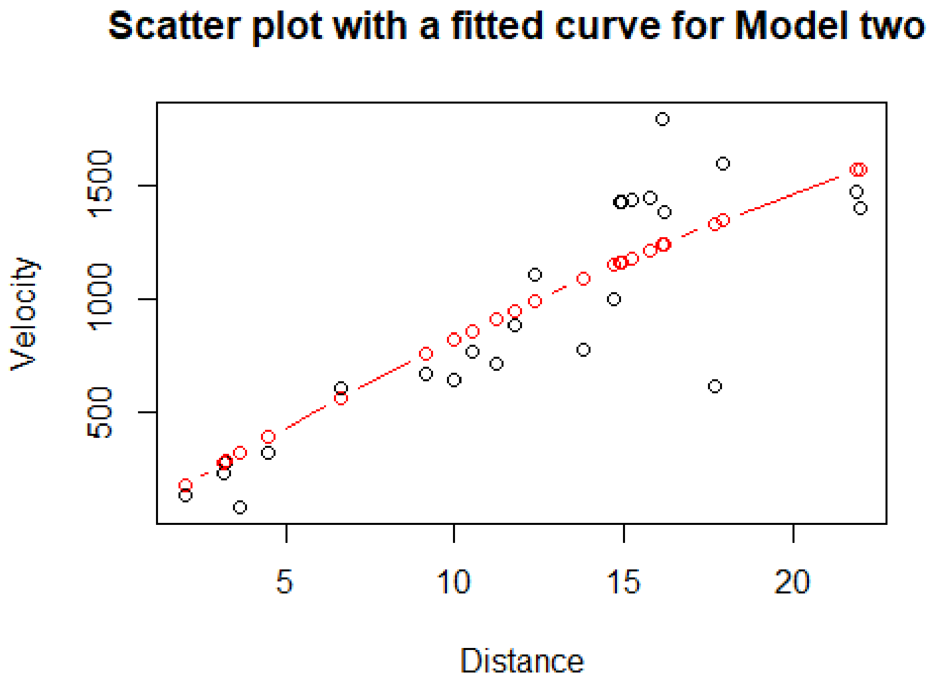
a) Fit a quadratic regression model, i.e., a model of the form

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

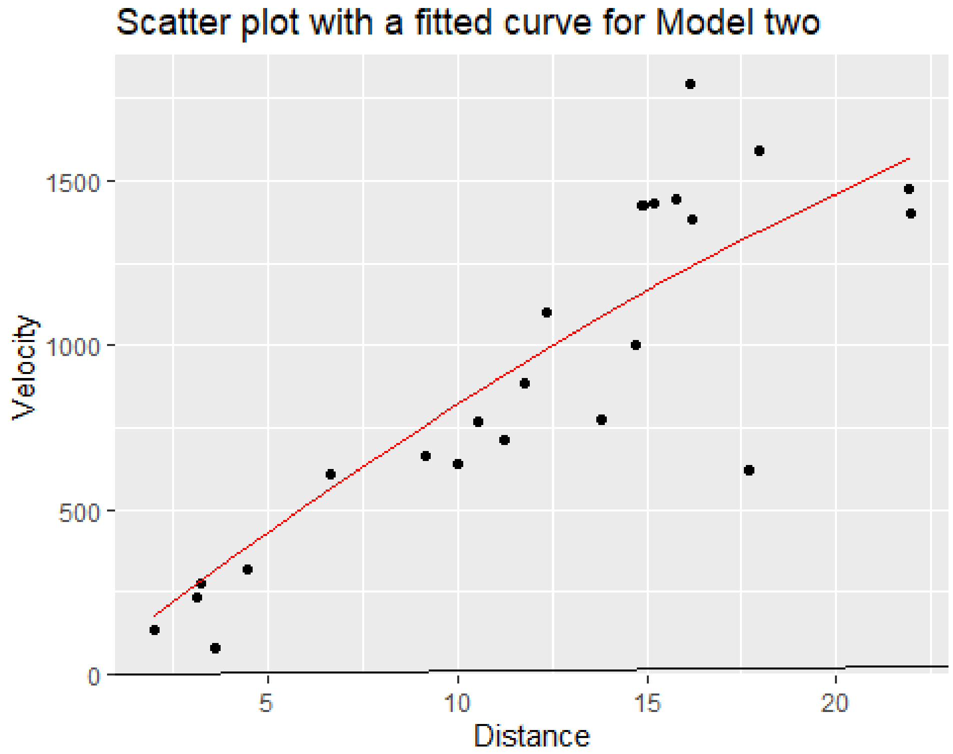
I have created a quadratic variable and appended it to the original data set to use it as an explanatory variable.

```
##
## Call:
## lm(formula = y ~ x + x2 - 1, data = hubble)
##
## Coefficients:
##       x       x2
## 90.9046  -0.8837

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

Used base R and ggplot to construct a scatter plot and fit the predicted values into it.



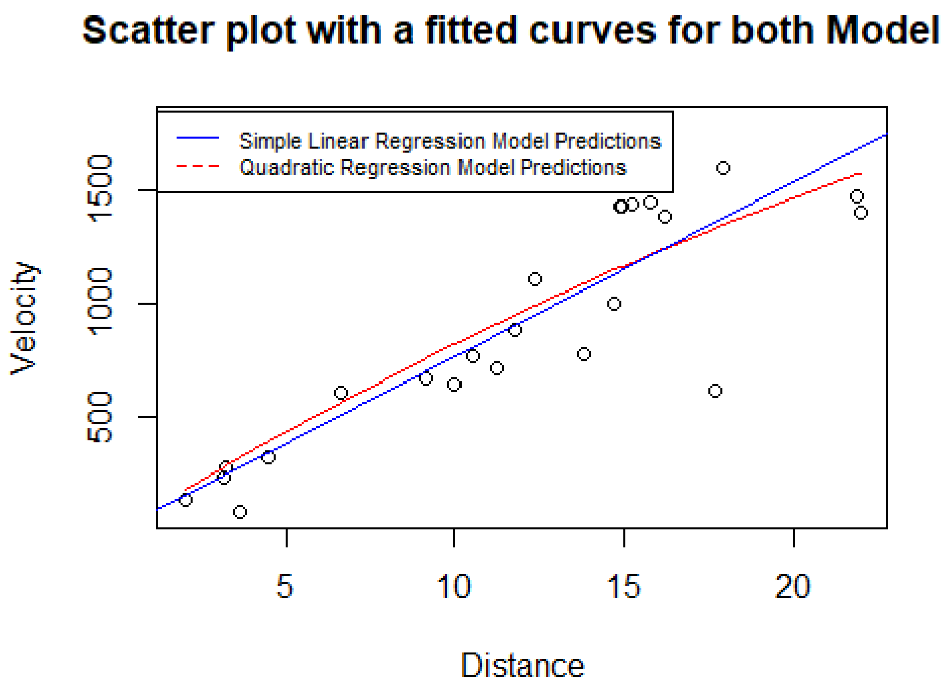
Scatter plot with a fitted curve for Model two

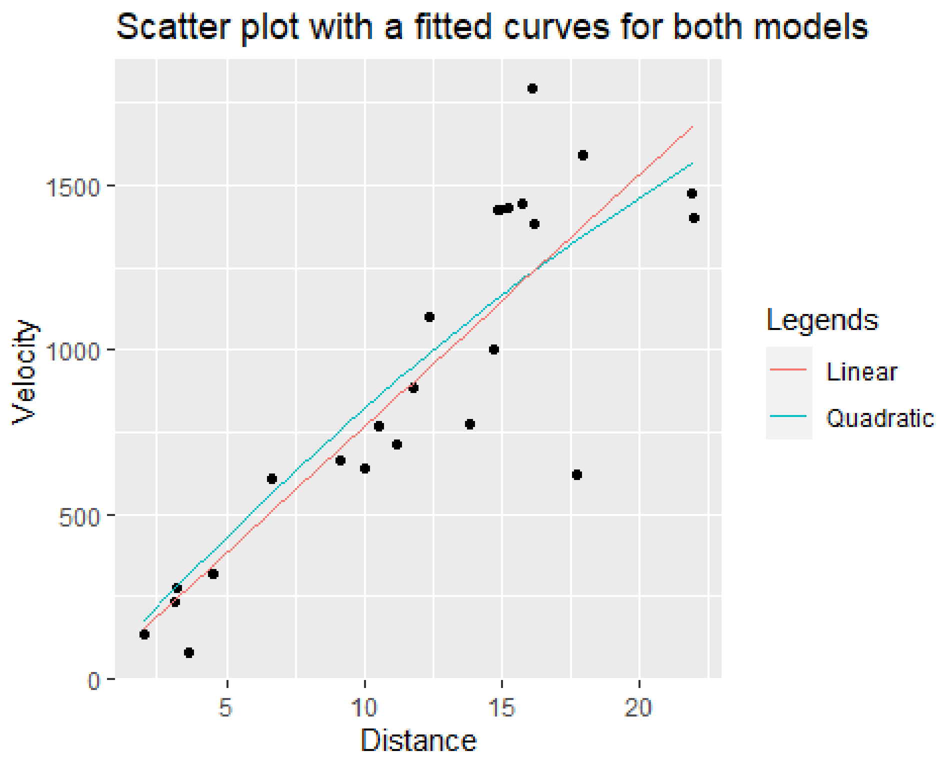


Scatter plot with a fitted curve for Model two

```
c) Add a simple linear regression fit over this plot. Use the relationship
between \textit{velocity} and \textit{distance} to determine the constraints
onthe 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.
```





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

By looking at the plot, we can see that the linear line fits the data points better than the polynomial line.The polynomial line does not fit the data well. Using simple linear regression model to fit a linear regression line is more sensible if want to achieve a minimum error.

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

## 
## 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
```

Answer: For the polynomial model, the the coefficient for x^2 is insignificant because the p-value is higher than 0.19494 > 0.05. However, the coefficient for x is significant because the P value is 1.64e-05 < 0.05. For the linear model the the coefficient for x is a lot more significant than the x for the previous model because the P value is 1.03e-15<0.05. if obt out the x^2 variable, the polynomial model would have identical results to the simple linear regression model. The P-value for the polynomial model is2.476e-07 > 1.032e-15 for the linear model which shows that the simple linear model is better. To pick the better model , we will assess the errors associated with each model. The polynomial model has an adjusted R-squared value of 0.7651 vs 0.9419 for the simple linear model, which shows that the linear model better highlights the variability in the predicted values.The F-statistics for the polynomial model is 34.2 < 373.1 for the simple linear regression model which indicates the simple linear regression performs better than the other model. Based on the statistical information, We can conclude that the linear regression model is a better model.

```
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.
```

3. (Ex. 7.4 in HSAUR, modified for clarity) The data from package 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 .

From the problems details, we can assume that the presence or the absence of morphological characteristic in white blood cells in would be a good indicator of if the patient would live more than 24 weeks or less than 24 weeks. This would mean that we will create binary variables 0 for patients that lived less than 24 weeks and 1 for patients that lived more than 24 hours.

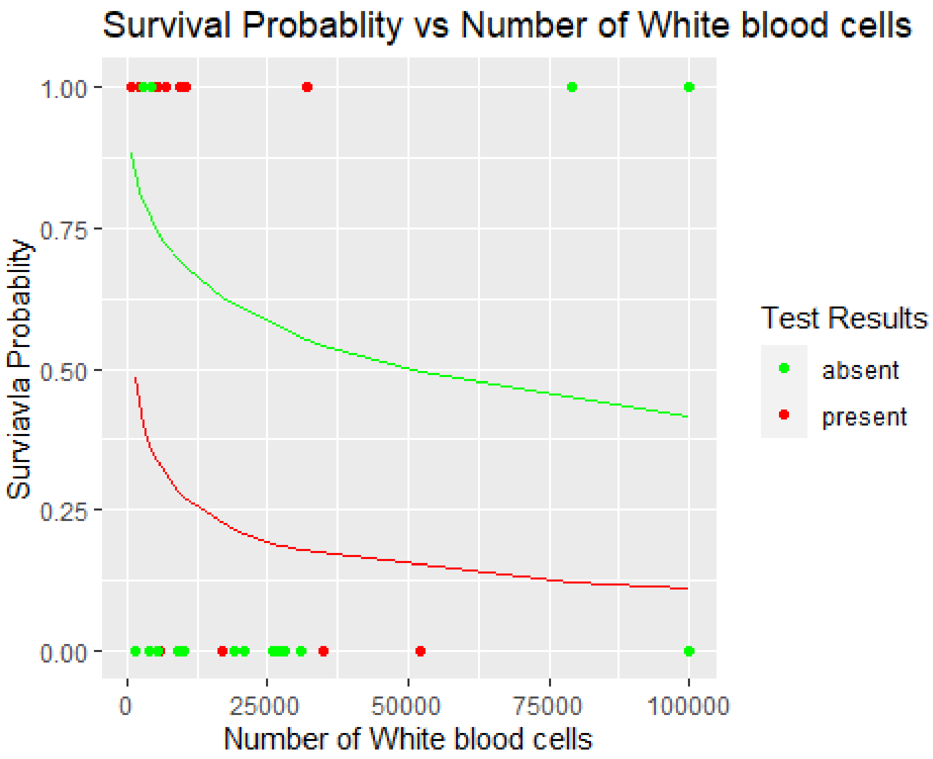
```
##    wbc      ag time surv24
## 1 2300 present   65      1
## 2  750 present  156      1
## 3 4300 present  100      1
## 4 2600 present  134      1
## 5 6000 present   16      0
```

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

In this step, we fit a logistic regression model to our data.

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

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



We can see from the graph that patients with more have a higher probability of living. Patients with present test results have higher probability to die within 24 weeks with most patients being above the 50% chance of dying. patients absent test results chances to live are better with most patients having a max 48% chance of dying for some patiesnts and less. Genarlly patiesnts with low white blood cells count have very high chance of dying.

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.

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

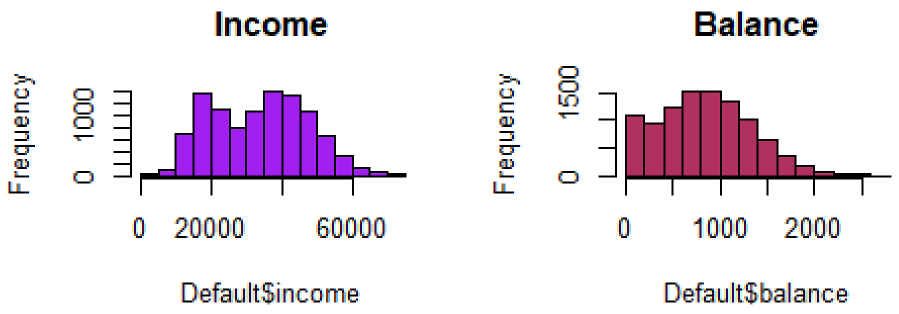
```
##
## Call:
## glm(formula = surv24 ~ log.wbc + ag + ag * log.wbc, family = "binomial",
##     data = leukemia.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9183  -0.7835  -0.6750   0.7310   1.7838
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.5946     4.6583  -0.557   0.5775
## log.wbc             0.1545     0.4746   0.326   0.7447
## agpresent          13.6306     7.0909   1.922   0.0546 .
## log.wbc:agpresent  -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
```

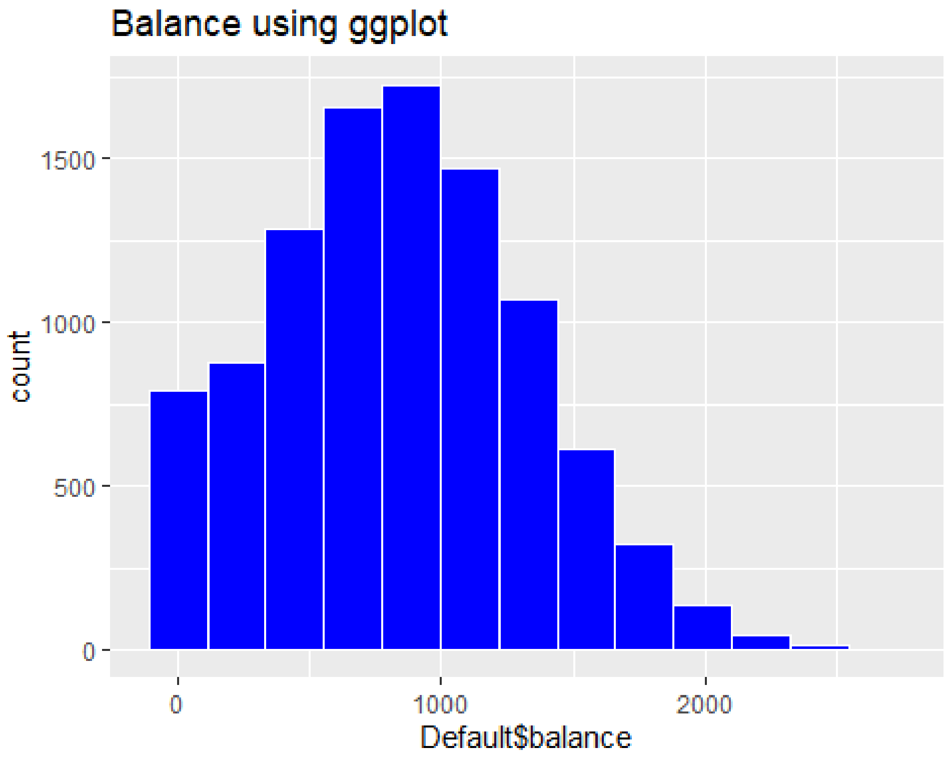
The AIC for the simpler model is 43.498 The AIC for the more complex model is 42.167 The model with the lower AIC is better which the more complex model, but there is not much difference between the AIC. It seems that the interaction between the explanatory variables is not significant with p-value higher than 0.05. by looking at the Adjusted R-square value the complex model has a higher adjusted R-squared. Therefore, we will choose the complex model that includes the interaction is the better model.

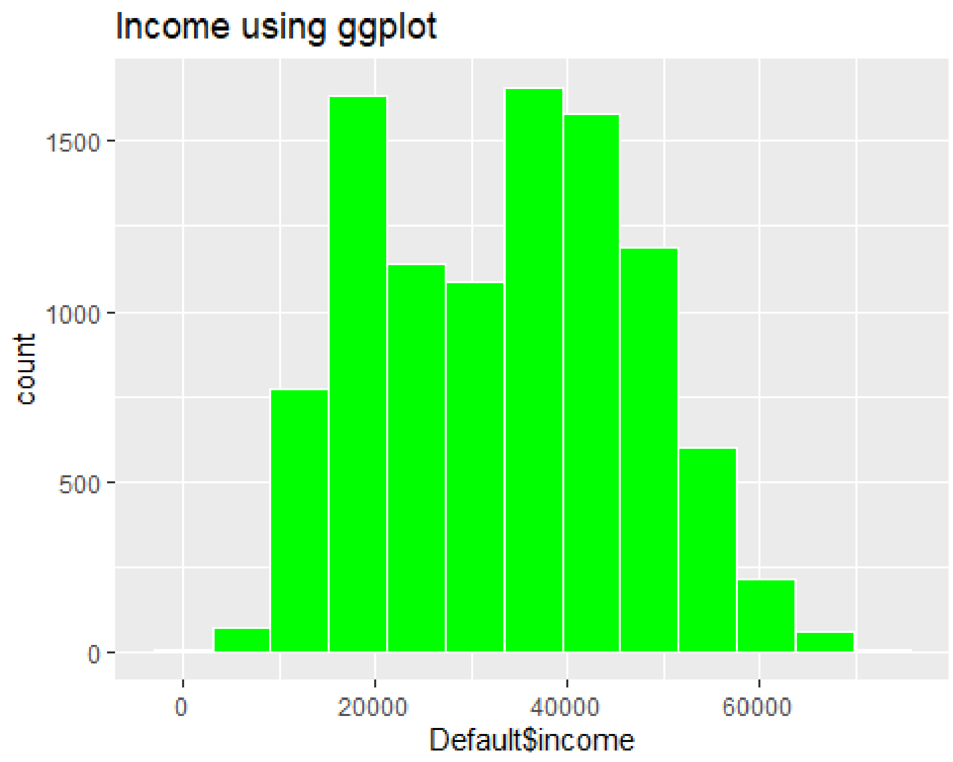
4. (Adapted from ISLR) Load the dataset from library. The dataset contains four features on 10,000 customers. We want to predict which customers will default on their credit card debt based on the observed features.

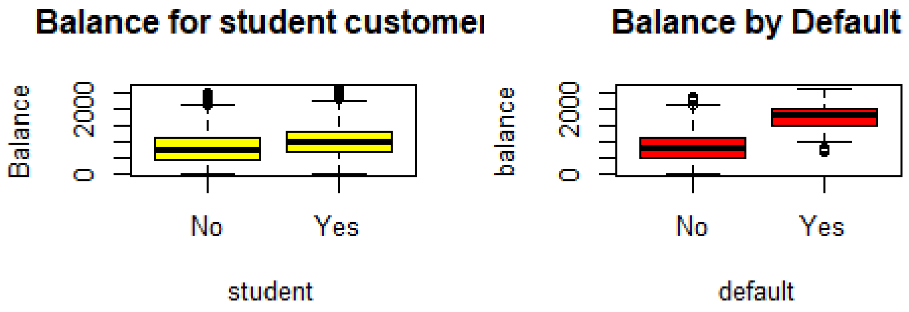
    a) Select a class of models using appropriate summaries and graphics. **Do not overplot.**

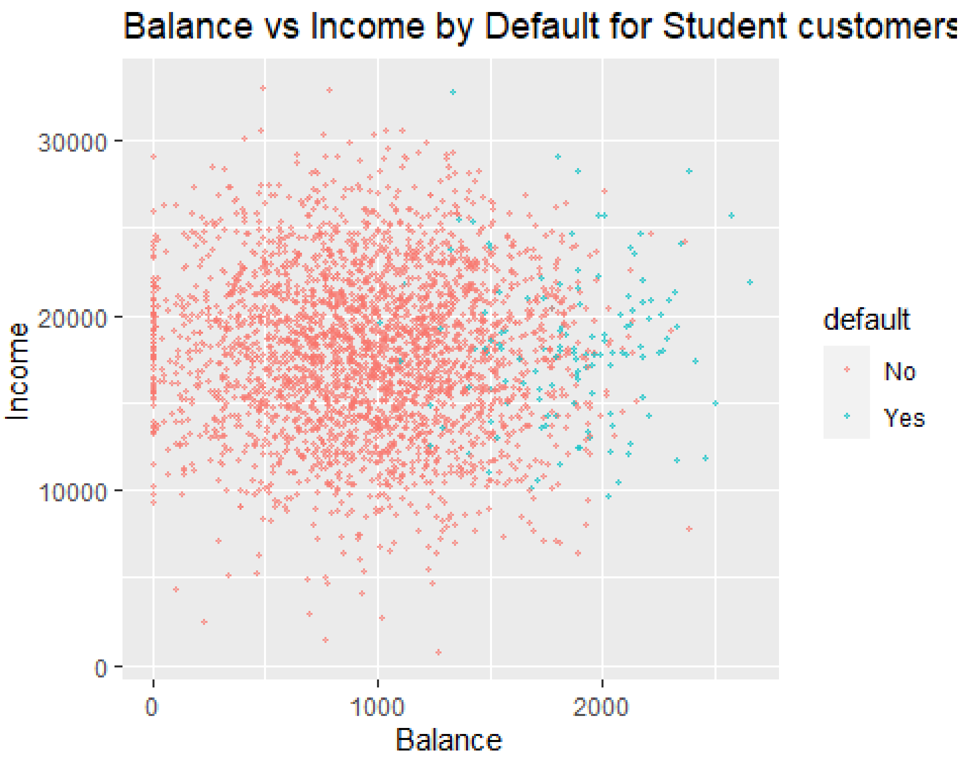
```
##  default    student       balance           income     
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772  
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340  
##                        Median : 823.6   Median :34553  
##                        Mean   : 835.4   Mean   :33517  
##                        3rd Qu.:1166.3   3rd Qu.:43808  
##                        Max.   :2654.3   Max.   :73554  

##  default   student      balance           income     
##  No :  0   No :206   Min.   : 652.4   Min.   : 9664  
##  Yes:333   Yes:127   1st Qu.:1511.6   1st Qu.:19028  
##                      Median :1789.1   Median :31515  
##                      Mean   :1747.8   Mean   :32089  
##                      3rd Qu.:1988.9   3rd Qu.:43067  
##                      Max.   :2654.3   Max.   :66466  

##  default    student       balance           income     
##  No :9667   No :6850   Min.   :   0.0   Min.   :  772  
##  Yes:   0   Yes:2817   1st Qu.: 465.7   1st Qu.:21405  
##                        Median : 802.9   Median :34589  
##                        Mean   : 803.9   Mean   :33566  
##                        3rd Qu.:1128.2   3rd Qu.:43824  
##                        Max.   :2391.0   Max.   :73554  
```

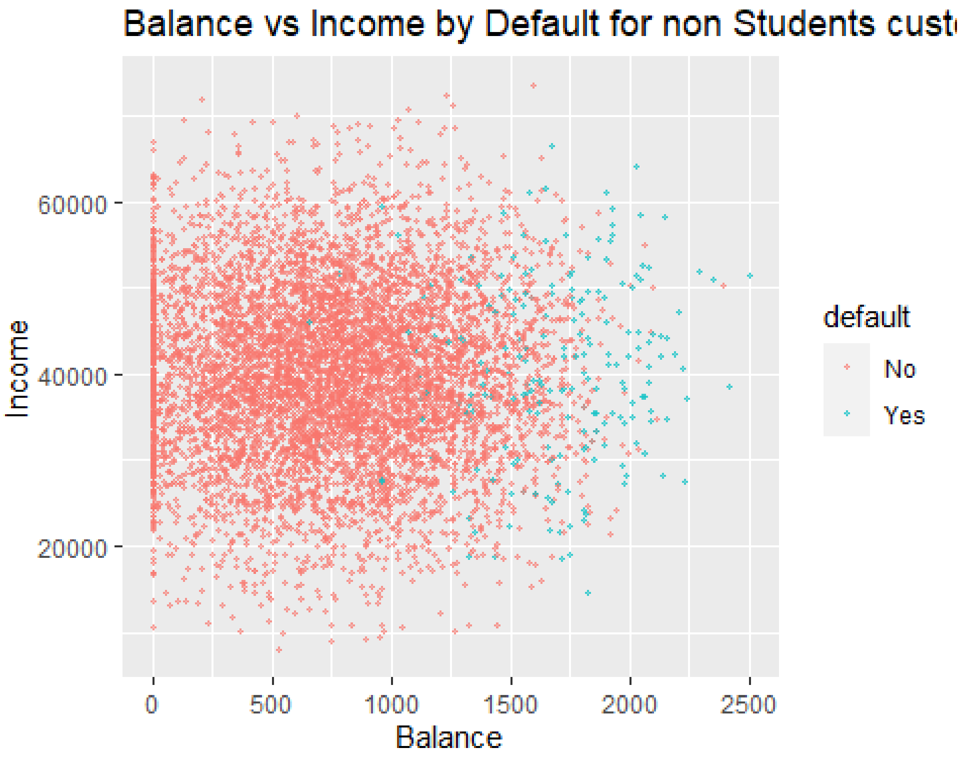

Income
Frequency
0
1000
0
20000
60000
Default$income
Balance
Frequency
0
1500
0
1000
2000
Default$balance


Balance using ggplot
count
1500
1000
500
0
0
1000
2000
Default$balance

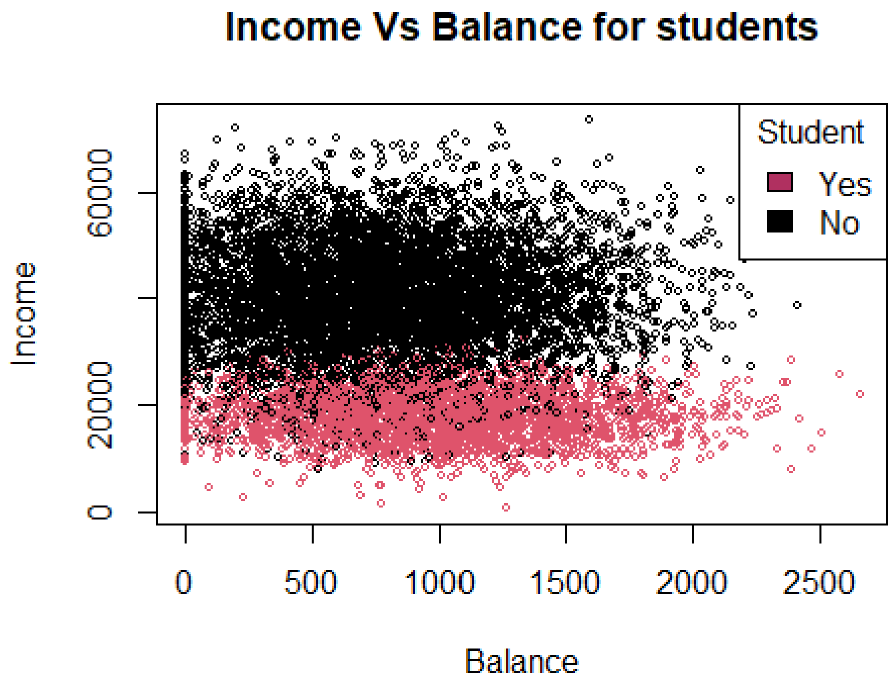

Income using ggplot
count
1500
1000
500
0
0
20000
40000
60000
Default$income


Balance for student customer
Balance
2000
0
No
Yes
student
Balance by Default
balance
2000
0
No
Yes
default





box plots, We can see that customers with higher balances have tend to default more. From scatter Plots, we can see that student and non-student customers with more balance tend to have more defaults.



It seems that customers that are not students have a higher income than customers that are students.

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

We will use logistic regression model because we are trying to answer the question of weather customers defaulted on their credit debt based on their income, weather they were a student, and their balance.

```
##
## Call:
## glm(formula = default1 ~ student1 + 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 ***
## student1    -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 = default1 ~ student1 + balance + income + student1 *
##     income + balance * student1 + 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 ***
## student1         -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
## student1:income   1.447e-05  2.779e-05   0.521    0.602
## student1: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
```

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

For Model 1 We noticed that the most significant variables for the model are student and balance since they very low p-values.for the second model, it seems that strudent and balance are the only ones that are significant and the interactions are not significant.

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

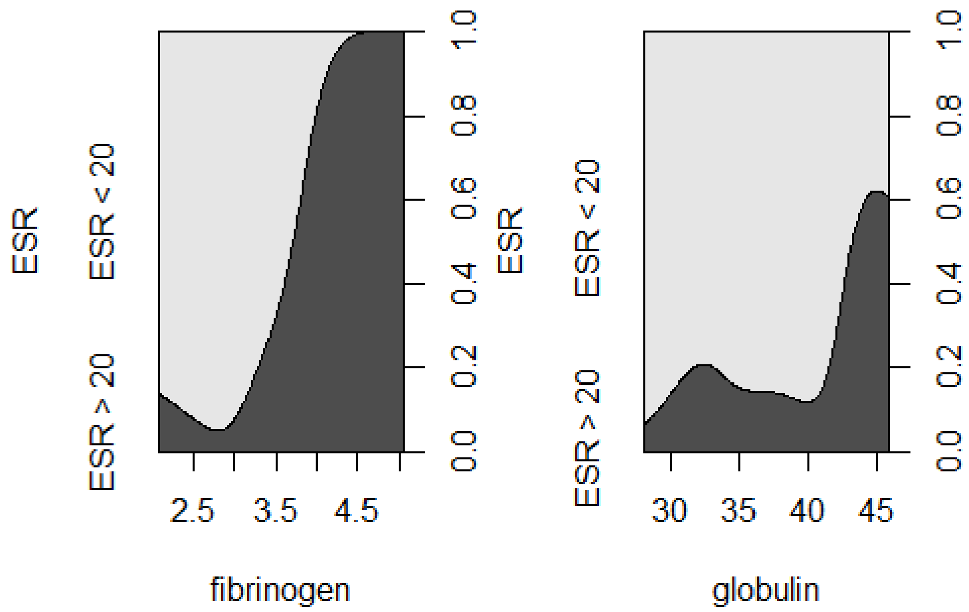
I performed an ANOVA Chi-square test to check the overall effect of variables on the dependent variable. For model one, we can see that the the the weather customers were students or not had a big effect on the their defaults so we can say that the student variable is significant. Also, the other variable that has a significant impact on the response variable is balance. Income is not significant For the second model, We can see that the extra variables we added are not significant with P-values that above 0.05. Similar to the first model, variables students and balance were the only variables that are significant for this model with a P-value below 0.05.

```
## Analysis of Deviance Table
##
## Model 1: default1 ~ student1 + balance + income
## Model 2: default1 ~ student1 + balance + income + student1 * income +
##     balance * student1 + balance * income
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      9996     1571.5
## 2      9993     1571.1  3  0.47911   0.9235

## [1] "Model one Confusion Matrix:"

##                     True
## model1.predications   No  Yes
##                  No  9627  228
##                  Yes   40  105

## Model Accuracy : 97.32

## [1] "Model two Confusion Matrix:"

##                     True
## Model2.predications   No  Yes
##                  No  9627  229
##                  Yes   40  104

## Model Accuracy : 97.31
```

The simpler model has an AIC of 1579.5 The more Complex model has an AIC of 1585.1 The more complex model has slightly higher AIC. Both models have an approximate error rate of 2.7%. both models are accurate in predicting the outcome of defaulting. I think the performance of both models is about the same with the simpler model have a tiny advantage over the complex model. The simpler the model the more generalized it will be.

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. `echo = T`

```
# conditional Density Plots
layout(matrix(1:2, ncol = 2))
cdplot(ESR ~ fibrinogen, data = plasma)
cdplot(ESR ~ globulin, data = plasma)
```



This two plots show how the explanatory variables vary with the factors of ESR. a small portion of fibrinogen is above ESR>20

```
plasma_glm_1 <- glm(ESR ~ fibrinogen, data = plasma,family = binomial())
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
```

It seems that the intercept and the explanatory variable fibrinogen is significant P-value<0.05.The difference in residual deviance from the first model is only 1.87

Apply the coef function to look at certain predictor

```
exp(coef(plasma_glm_1)["fibrinogen"])

## fibrinogen
##   6.215715
```

getting the confidence intervals

```
exp(confint(plasma_glm_1, parm = "fibrinogen"))

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

performing a logistic regression of both explanatory variables to see the difference.

```
plasma_glm_2 <- glm(ESR ~ fibrinogen + globulin,
data = plasma, family = binomial())
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
```

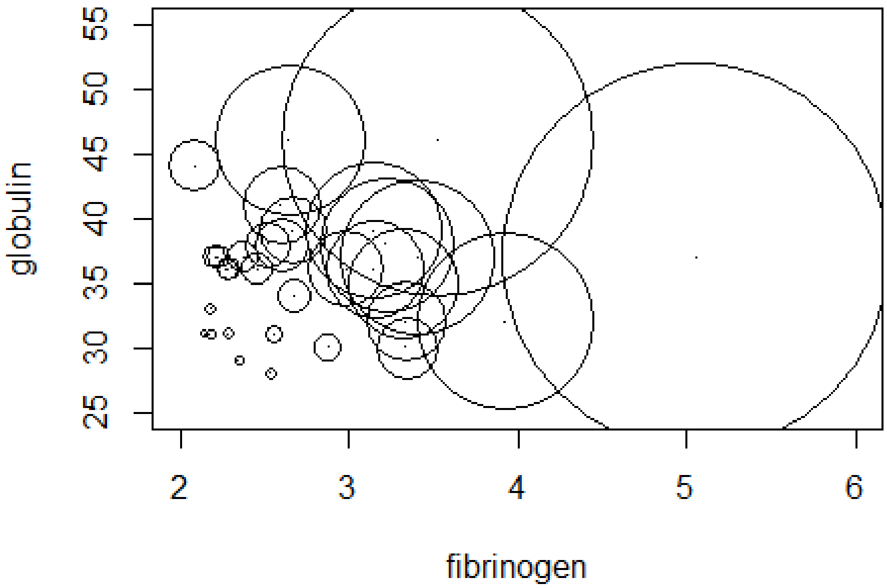
Globukin is not significant with p-value that is higher than 0.05. The fibrinogen variable is significant<0.05.

Run an Anova for both models

```
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
```

Bubble plot for model two.

```
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=T)
```



The following bubble plot shows the predicted values for the second model. We can see that as fibrinogen increases increases, the probability of getting a better ESR reading increases.