Homework 2

Amin Baabol

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.

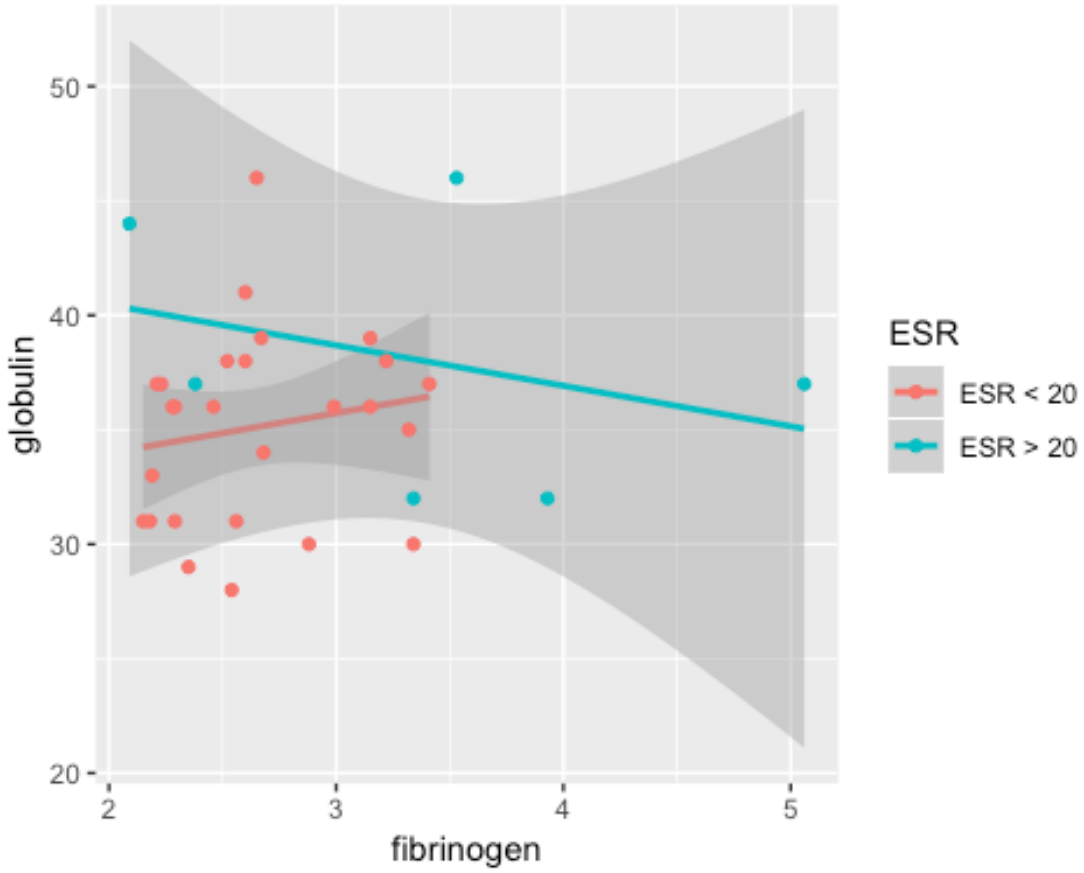
Exercises

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.

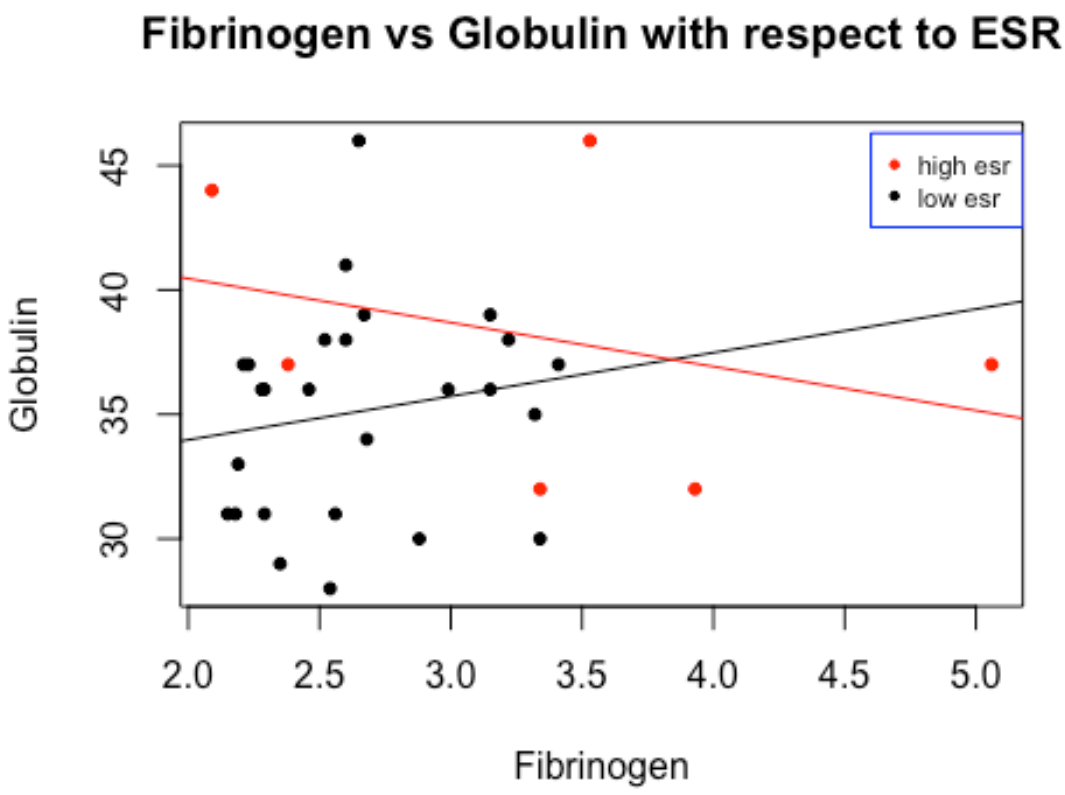
1. (Ex. 7.2 in HSAUR, modified for clarity) Collett (2003) argues that two outliers need to be removed from the data. Try to identify those two unusual observations by means of a scatterplot. (Hint: Consider a plot of the residuals from a simple linear regression.)

Assumptions: the relationship between the predictor variables(globulin&fibrinogen) is linear hence why we're using simple lineaer regression model.

```
##    fibrinogen       globulin           ESR
##  Min.   :2.090   Min.   :28.00   ESR < 20:26
##  1st Qu.:2.290   1st Qu.:31.75   ESR > 20: 6
##  Median :2.600   Median :36.00
##  Mean   :2.789   Mean   :35.66
##  3rd Qu.:3.167   3rd Qu.:38.00
##  Max.   :5.060   Max.   :46.00
```




```
## $x
## [1] "Figrinogen"
##
## $y
## [1] "Globulin"
##
## $title
## [1] "Fibrinogen vs Globulin with respect to ES"
##
## attr(,"class")
## [1] "labels"
```



Discussion: There are two outliers for when ESR is greater than 20 at fib ~c(3.8,5.1) and also other outliers for when esr is less than 20, which in my opinion suggests the data should be split into two groups. spliting the data into two binary groups indicates simple linear regression is not the correct model to be deployed.

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

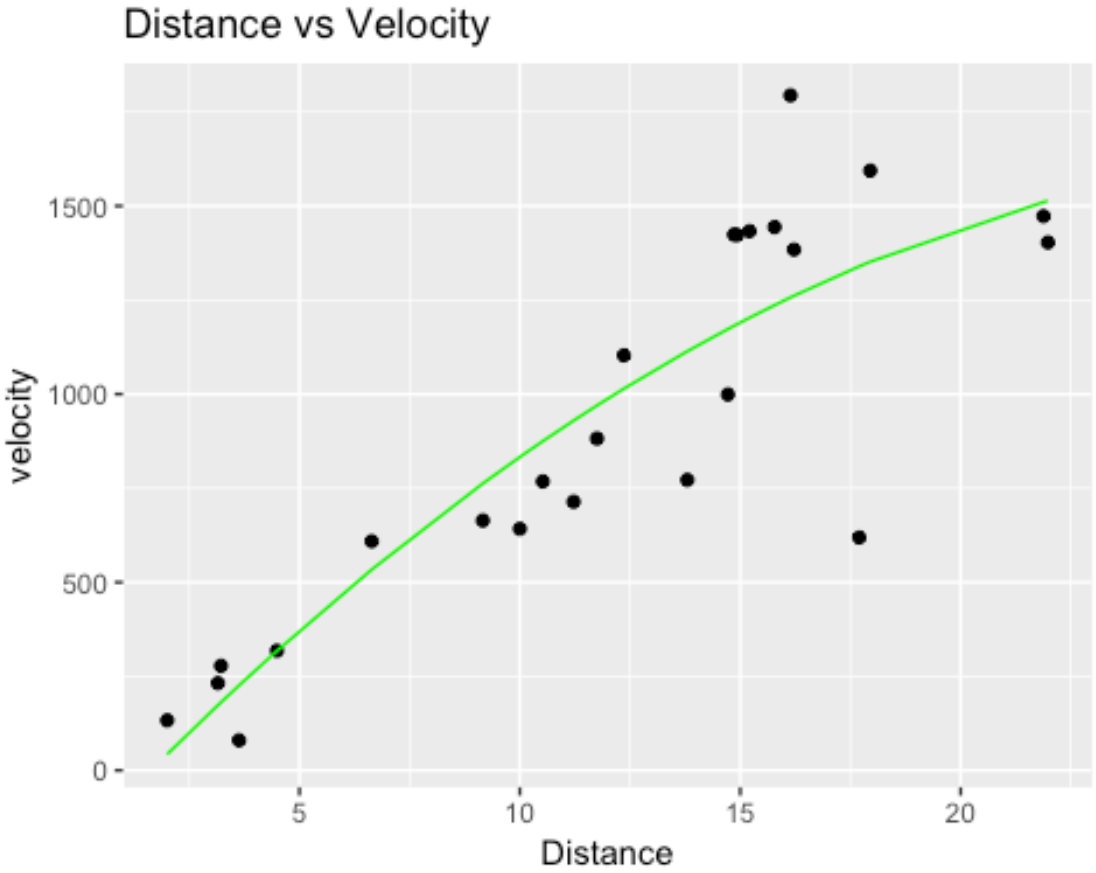
a) Fit a quadratic regression model, i.e., a model of the form

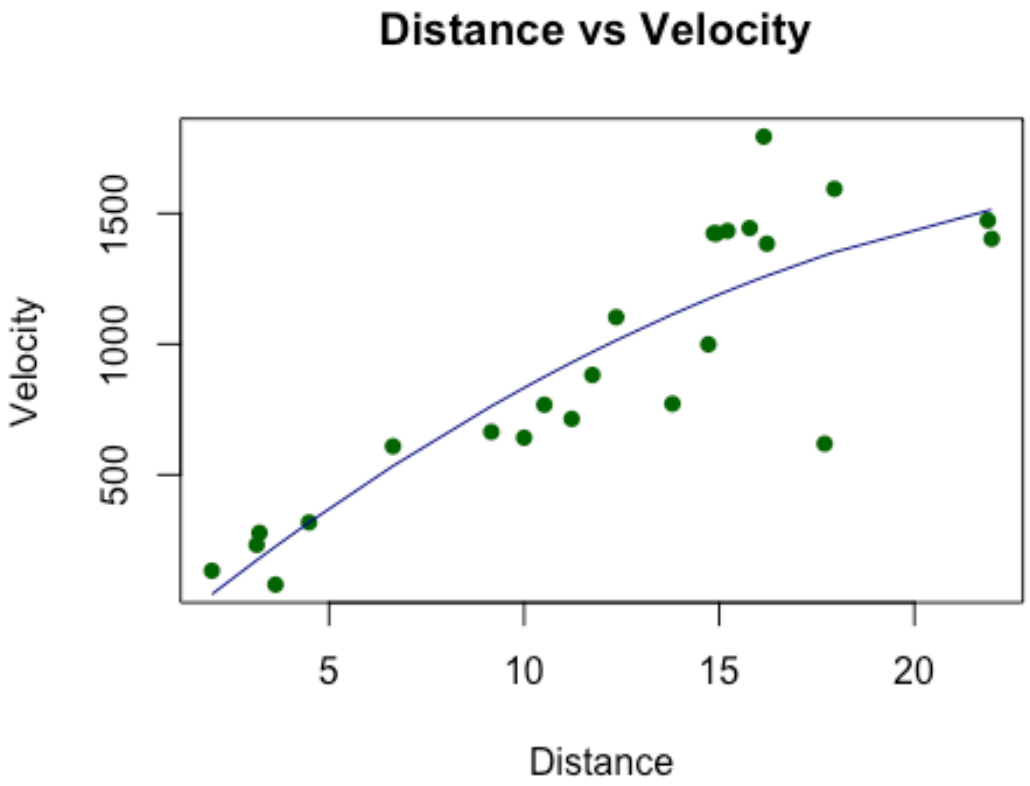
Model 2: $velocity = \beta_1 \times distance + \beta_2 \times distance^2 + \epsilon$

```
##
## Call:
## lm(formula = y ~ x + poly(x, 2), data = hubble)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -720.5 -119.5   29.7  143.8  537.1
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.696    124.338   0.054    0.958
## x             76.127      9.327   8.162 5.96e-08 ***
## poly(x, 2)1       NA         NA      NA       NA
## poly(x, 2)2 -348.217    260.093  -1.339    0.195
```

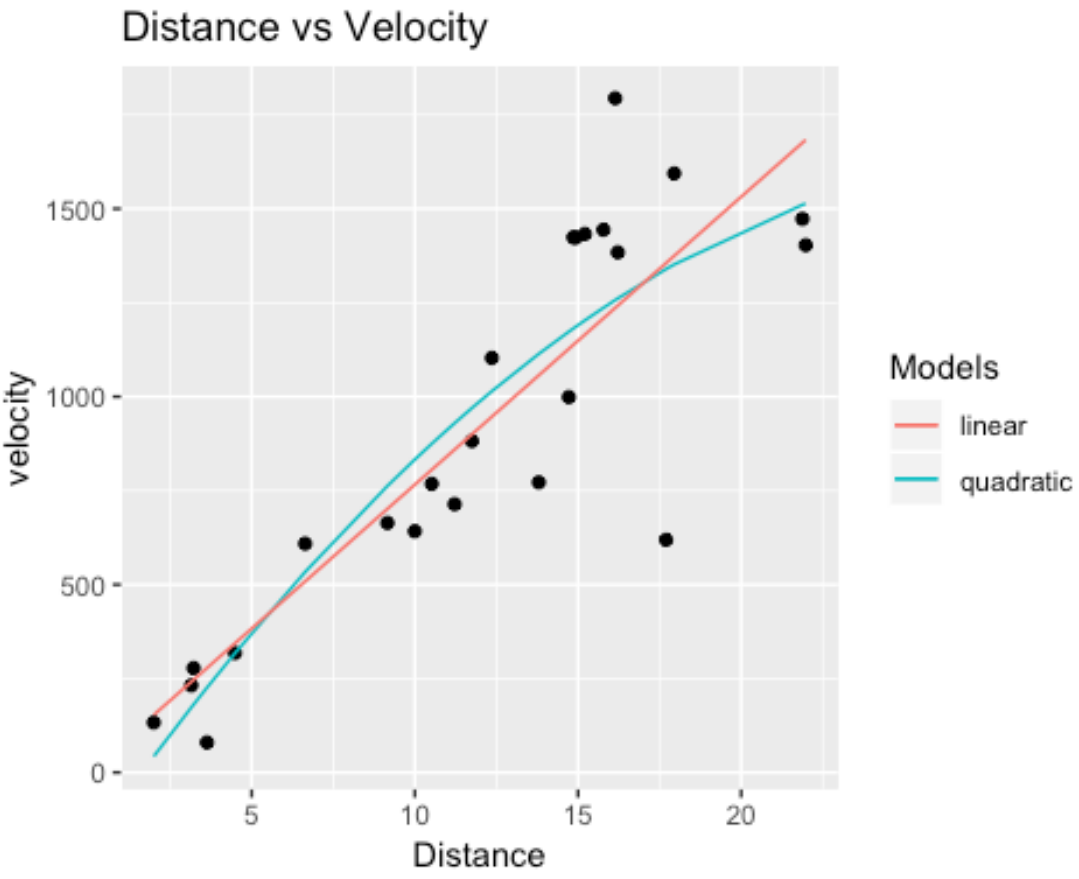
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 260.1 on 21 degrees of freedom
## Multiple R-squared:  0.7651, Adjusted R-squared:  0.7428
## F-statistic: 34.21 on 2 and 21 DF,  p-value: 2.476e-07
```

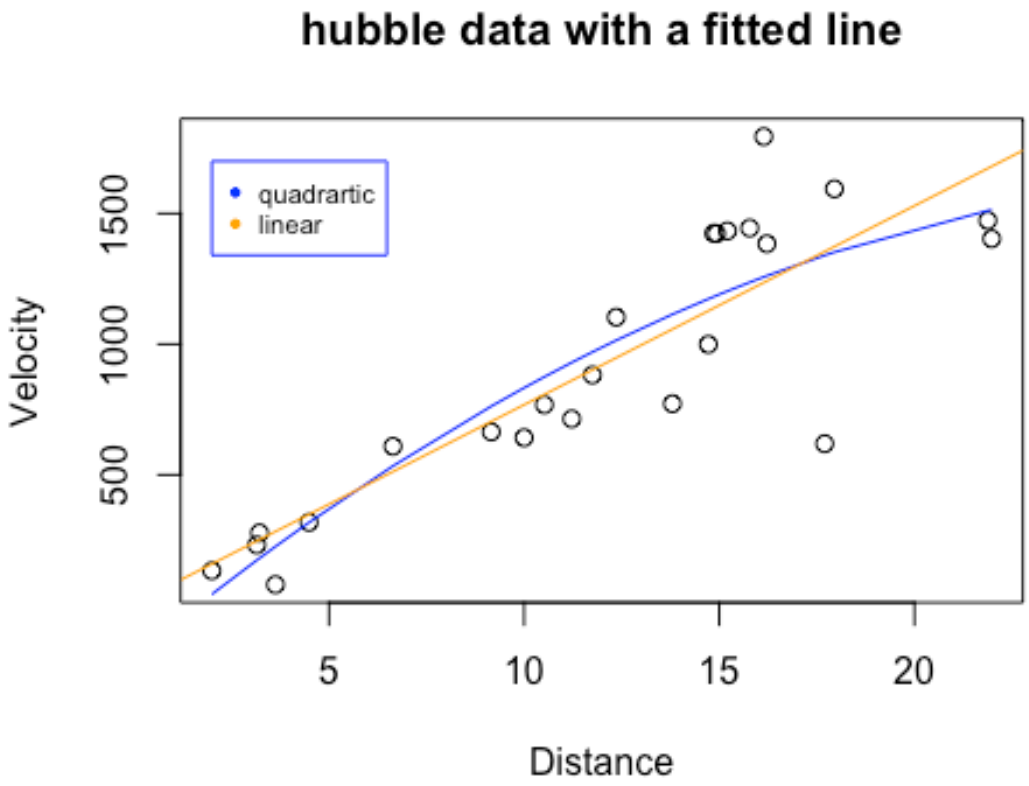
b) Plot the fitted curve from Model 2 over the scatterplot of the data.





c) Add a simple linear regression fit over this plot. Use the relationship between and 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.


Distance vs Velocity
velocity
1500
1000
500
0
5
10
15
20
Distance
Models
linear
quadratic



d) Examine the plot, which model do you consider most sensible? Although there is hardly a significant difference between the two models,however, given the small dataset a simple linear regression model seems to be better suited in the generalization of this data.

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

```
##
## Call:
## lm(formula = y ~ x + poly(x, 2), data = hubble)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -720.5 -119.5   29.7  143.8  537.1
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.696    124.338   0.054    0.958
## x             76.127      9.327   8.162 5.96e-08 ***
## poly(x, 2)1       NA         NA      NA       NA
```

```
## poly(x, 2)2 -348.217    260.093  -1.339    0.195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 260.1 on 21 degrees of freedom
## Multiple R-squared:  0.7651, Adjusted R-squared:  0.7428
## F-statistic: 34.21 on 2 and 21 DF,  p-value: 2.476e-07

##
## 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
```

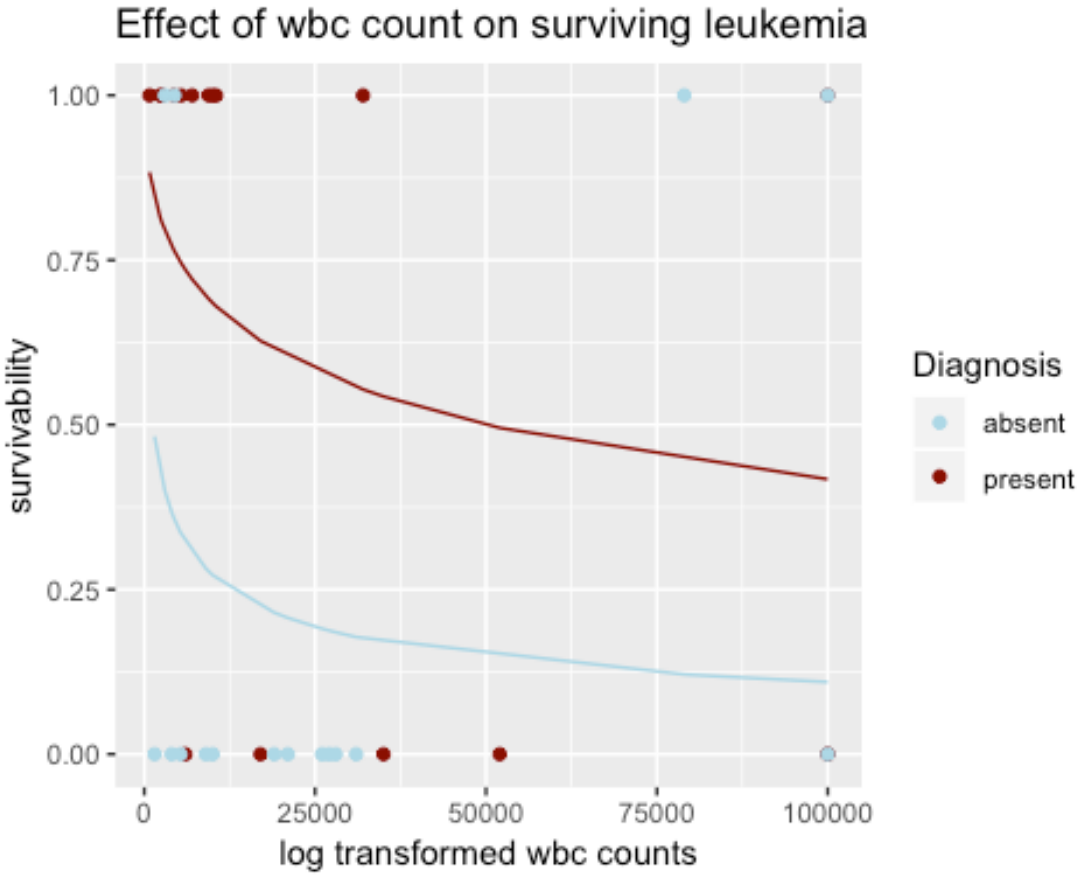
Dicussion:Looking at the R squared values, they quadratic model has an R squared value of 0.76 where as the simple regression has a value of 0.94. while those are both good R squared values,however the simple regression shows smaller variations between the observed data and the fitted values.This is further supported by the even smaller p-value of 1.032e-15 comparing it to that of 2.476e-07 for the quadratic model

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 .

b) Fit a logistic regression model to the data with 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.)

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



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 ~ ag + log.wbc, family = "binomial", data =
data1.log)
##
## 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
## agpresent     1.7621     0.8093   2.177   0.0295 *
## log.wbc      -0.4822     0.3149  -1.531   0.1257
## ---
## 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 ~ ag + log.wbc + ag * log.wbc, family = "binomial",
##     data = data1.log)
##
## 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
```

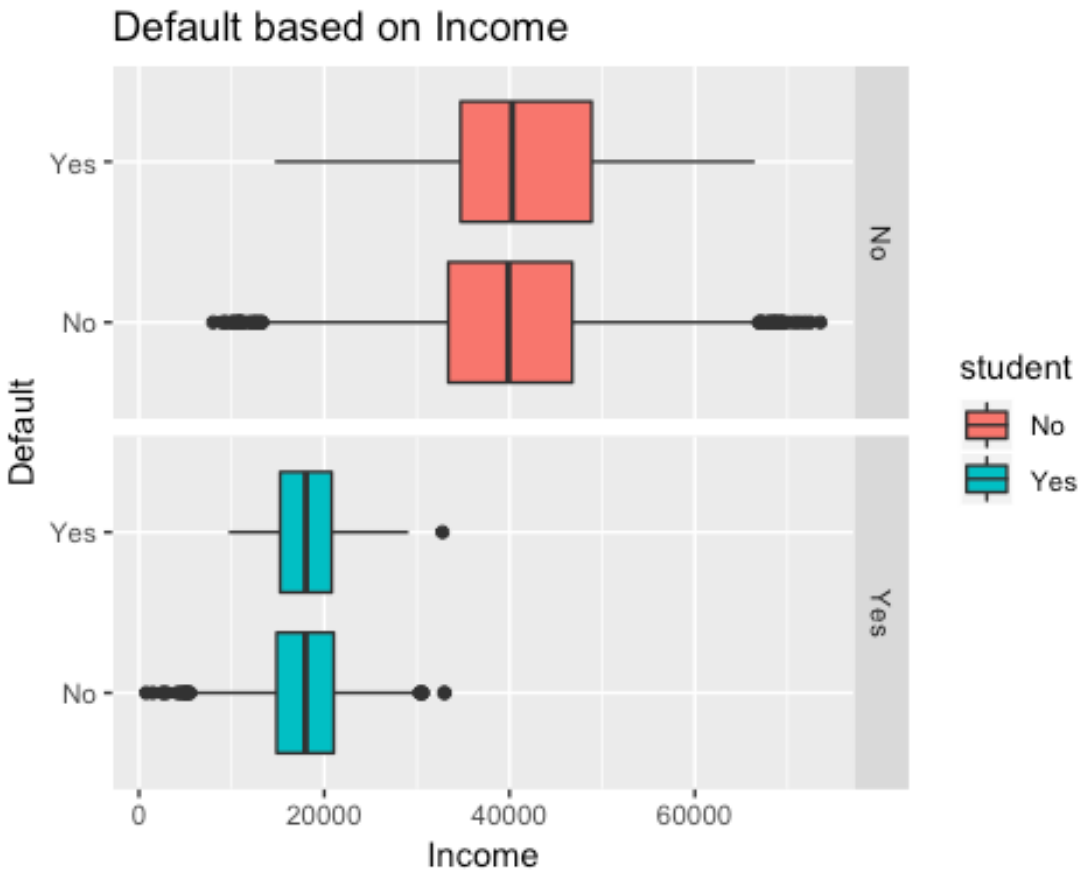
Discussion: The basic, simple regression model seems to be the better alternative because it offer lower p-value which indicates the independent variables have signifcant correlation with the response variable. Having said that, the more complex method does have it's own perks incldluing lower residuals.

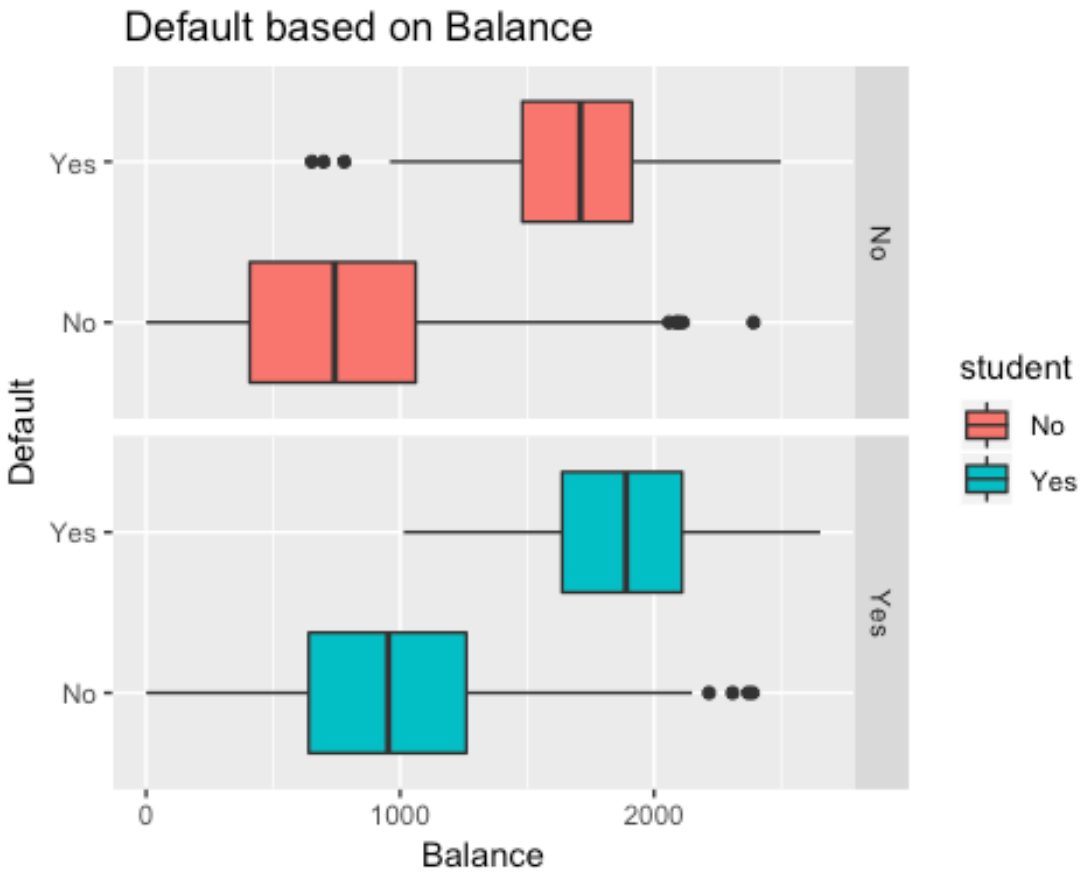
4. (Adapted from ISLR) Load the dataset from library. The dataset contains four features on 10,000 customers. We want to predict which customers will default on their credit card debt based on the observed features.

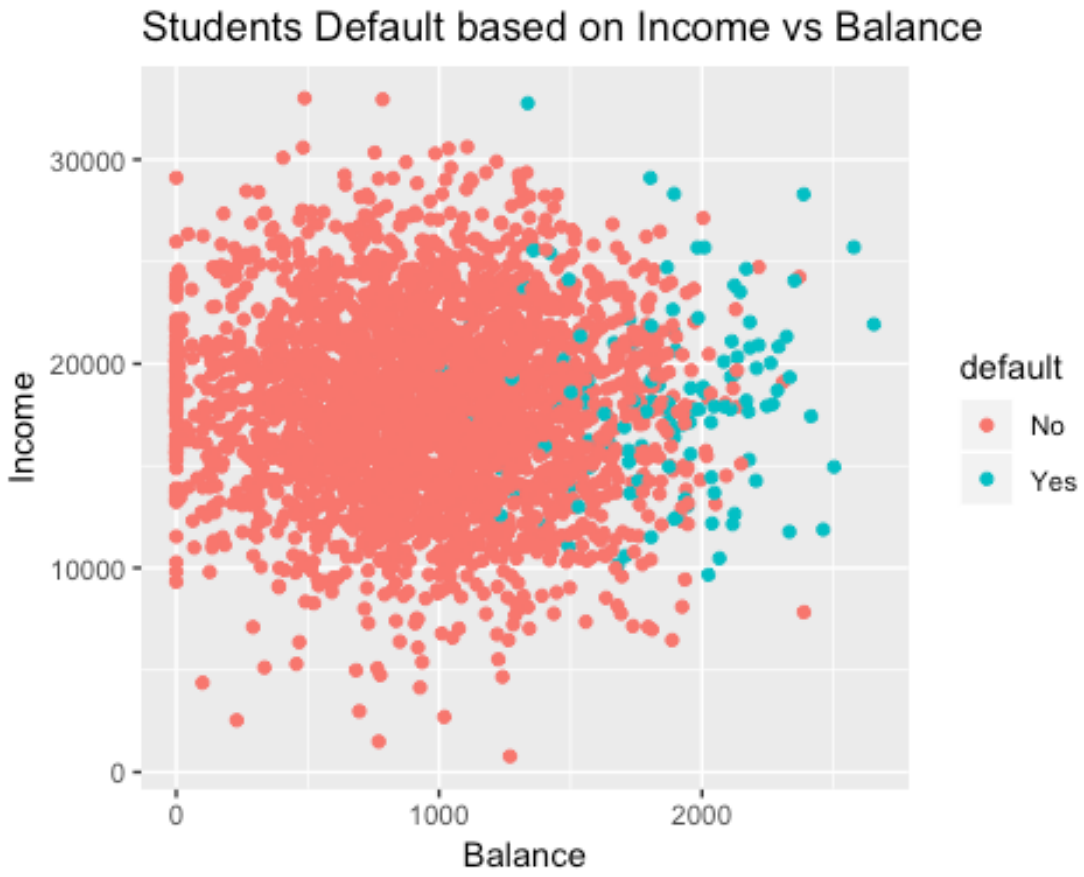
a) Select a class of models using appropriate summaries and graphics. **Do not overplot.**

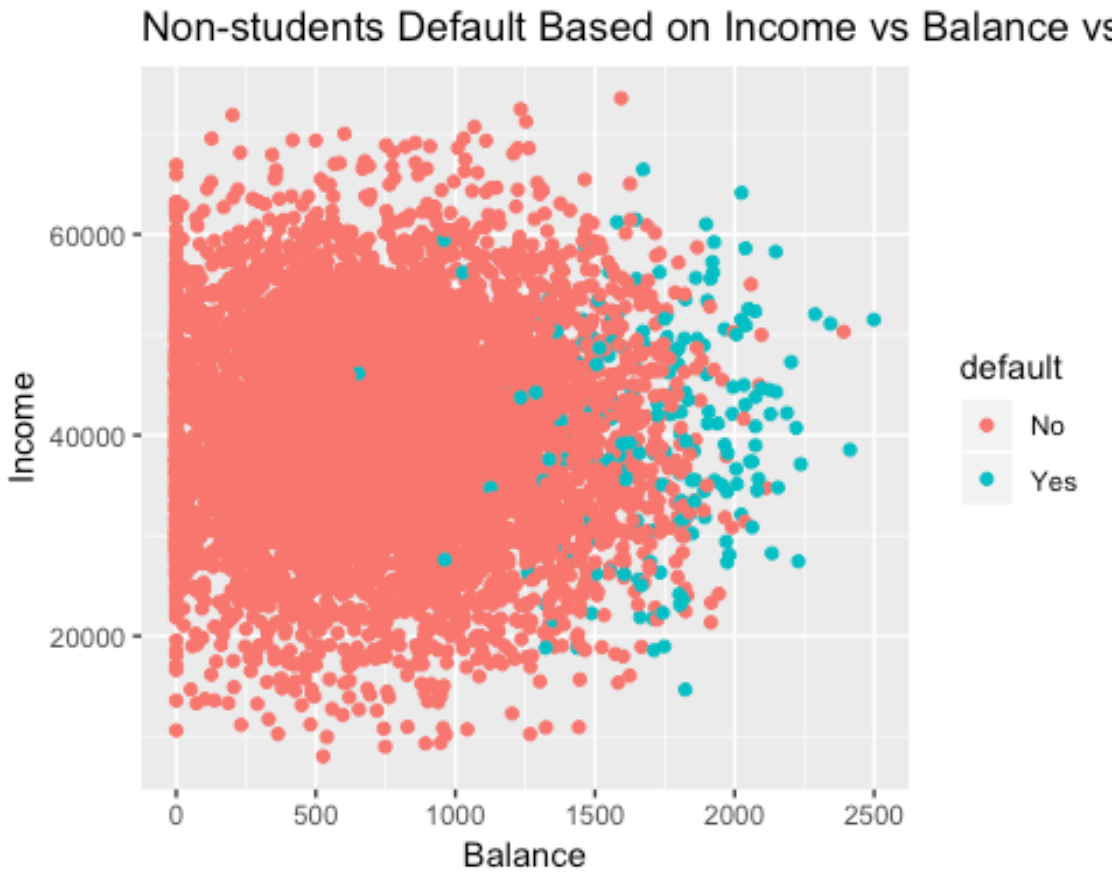
```
##  default   student      balance           income
##  No :  0   No :206   Min.   : 652.4   Min.   : 9664
##  Yes:333   Yes:127   1st Qu.:1511.6   1st Qu.:19028
##                      Median :1789.1   Median :31515
##                      Mean   :1747.8   Mean   :32089
##                      3rd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


Default based on Income
Yes
No
Default
Yes
No
No
Yes
0
20000
40000
60000
Income
student
No
Yes


Default based on Balance
Default
Yes
No
Yes
No
No
Yes
0
1000
2000
Balance
student
No
Yes


Students Default based on Income vs Balance
Income
30000
20000
10000
0
0
1000
2000
Balance
default
No
Yes



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

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

```
##
## 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 + student * balance + 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
```

Discussion: It seems to that the first simple model is better alternative with much lower standard error than the complicated model where all the variables are multiplied to one another. upon reading the summary of the simple model, balance plays the most significant effect on the response variable followed by the student category. The residual doesn't really change all that much.

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

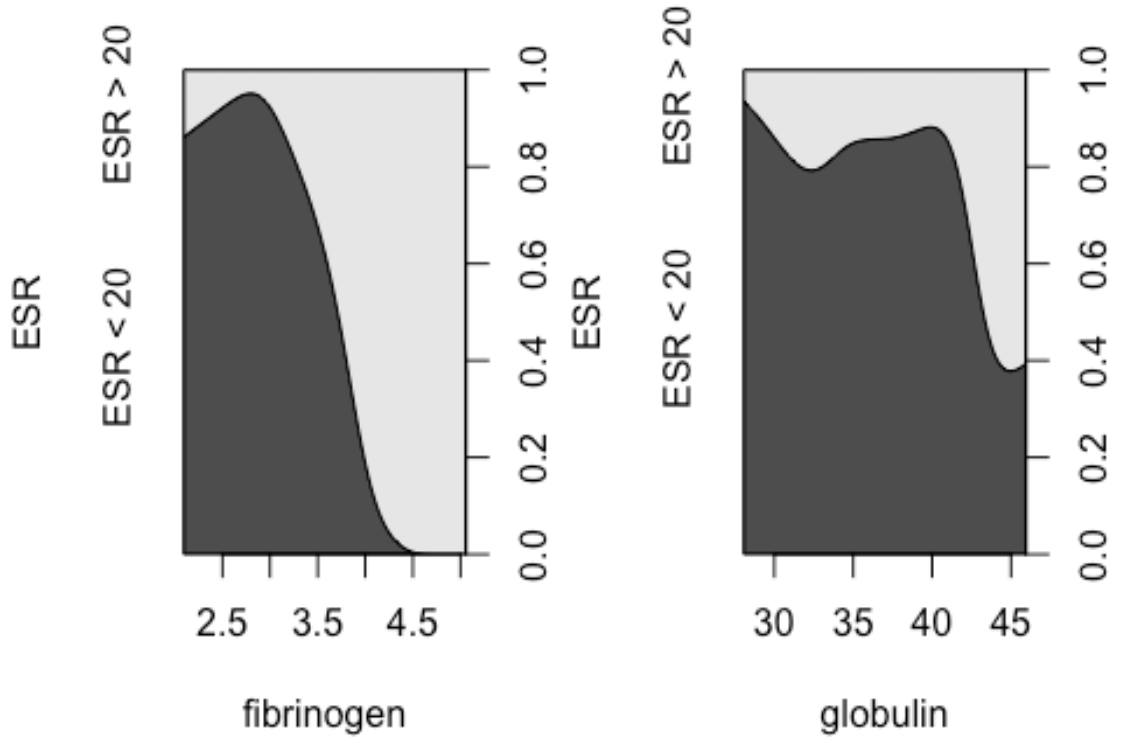
```
## [1] 2.68

## [1] 97.32
```

```
## [1] 2.69
```

```
## [1] 97.32
```

Discussion: Both models' prediction seems highly accurate.The simple model's error rate is a mere 2.68% which is very low. On the other hand, the complex model's error is also extremely low at 2.69%. Given the great accuracy of both models, i'd recommend the simple model.

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`



Discussion: The two plots indicate how the categorical variable esr behaves when the other predictor variables change. It seems that esr drasticall drops as fibrinogen approaches right around 4.5 while esr fluctuates as globulin continues.

```
##     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
```

Discussion: This part function begins by deploying a binomial logistic regression model to show fibrinogen as the predictor variable to potentially explain how ESR, the response variable would behave. The summary output shows a 5% signifance of finrinogen.

```
## fibrinogen
##   6.215715

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

Discussion: Exponentiating the function had undone the log odds of fibrinogen and its confidence interval to make the summary easier to read. This indicates that fibrinogen has considerable influence on esr.

complex logistic regression with multiple explanatory variables.

```
##
## 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
##
```

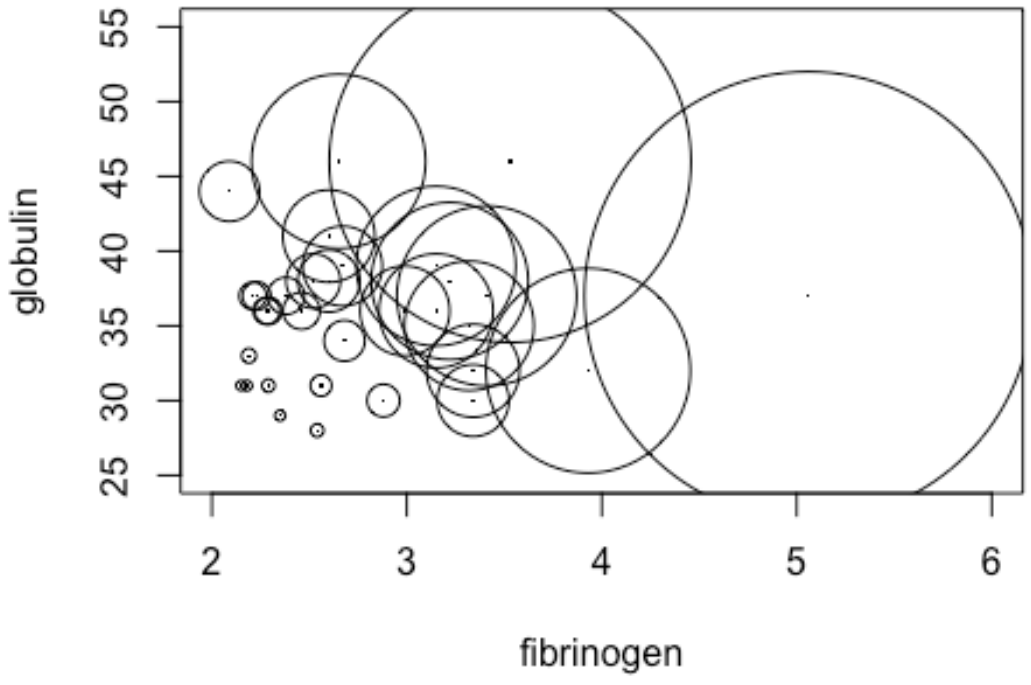
```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30.885  on 31  degrees of freedom
## Residual deviance: 22.971  on 29  degrees of freedom
## AIC: 28.971
##
## Number of Fisher Scoring iterations: 5
```

Discussion: In this model with both multiple explanatory variables, fibrinogen still shows the same significant level while globulin appears to be significant at the 0.05 confidence interval level. Hence it should be included in the model selction.

```
## 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
```

Discussion: The anova function is utilized when comparing two models.

The bubble plot is used to illustrate the interactions of all the three variables



Discussion: This bubble plot does a good job explaining how these explanatory and

response variables are behave when they interact with each other.The plot indicates that when fibrinogen increases esr value also increase. The same is also true for globulin but to a lesser exten, however globulin seems to have a cut off point.

#Citation

"A Handbook of Statistical Analyses Using R, third Edition" by Everitt and Hothorn

R Graphics Cookbook" by Winston Chang published through O'Reilly (Basic guide to Grammar of Graphics in R)

www.stackoverflow.com

http://r-statistics.co/Linear-Regression.html