Homework 2

Alex Soupir

September 05, 2019

Packages: HSAUR3, GGplot, gamair, MASS, ISLR

Collaborators:

Answer all questions specified on the problem and include a discussion on how your results answered/addressed the question.

Submit your **.rmd** file with the knitted **PDF** (or knitted Word Document saved as a PDF). If you are having trouble with .rmd, let us know and we will help you, but both the .rmd and the PDF are required.

This file can be used as a skeleton document for your code/write up. Please follow the instructions found under Content for Formatting and Guidelines. No code should be in your PDF write-up unless stated otherwise.

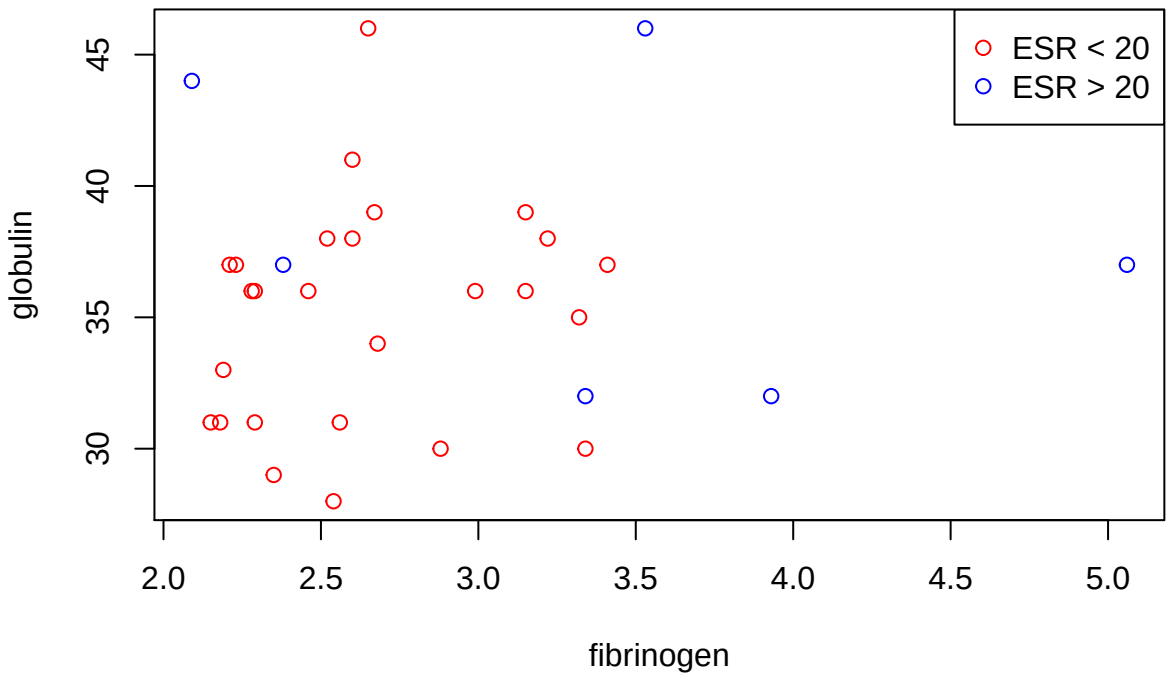
For any question asking for plots/graphs, please do as the question asks as well as do the same but using the respective commands in the GGPLOT2 library. (So if the question asks for one plot, your results should have two plots. One produced using the given R-function and one produced from the GGPLOT2 equivalent). This doesn't apply to questions that don't specifically ask for a plot, however I still would encourage you to produce both.

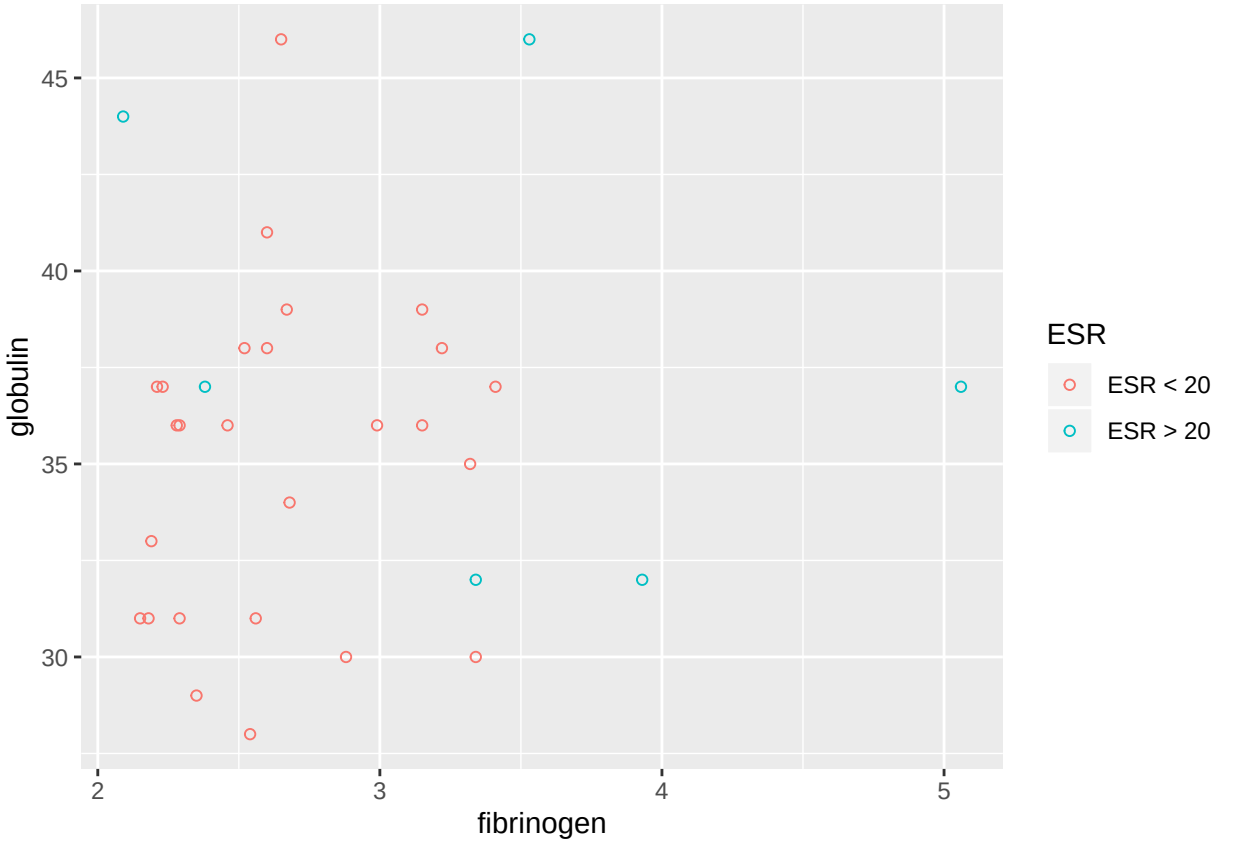
You do not need to include the above statements.

Please do the following problems from the text book R Handbook and stated.

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)

- Statement: Find two outliers that are argued to be removed from the **plasma** df.


ESR < 20
ESR > 20
globulin
30
35
40
45
fibrinogen
2.0
2.5
3.0
3.5
4.0
4.5
5.0



+ Discussion: The 2 observations that I think are the outliers by looking at the scatter plots from both *ggplot* and base R includes the observation with which fibrinogen is greater than 5 and fibrogen close to 4. These 2 points are located relatively by themselves whereas the observations that have globulin close to (and greater than) 45 have several observations at that level. + It's difficult to determine the outlier just by viewwing the data without a formal test that states emperically the bounds of an outlier, so the answer is mostly an educated guess based on the context of the plot. + I think it's important to note that the distribution of ESR < 20 has less variance than ESR > 20. Globulin is fairly evenly distributed with a few observations having higher levels in the blood, but the points are located too far away from the main body of the data.

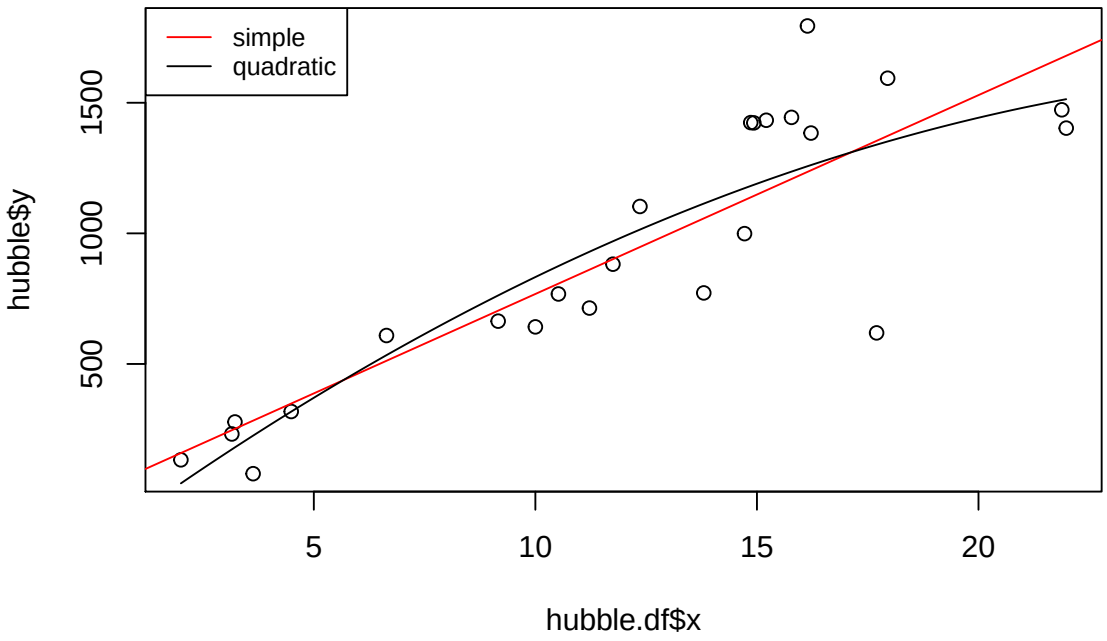
2. (Multiple Regression) Continuing from the lecture on the **hubble** data from **gamair** library;

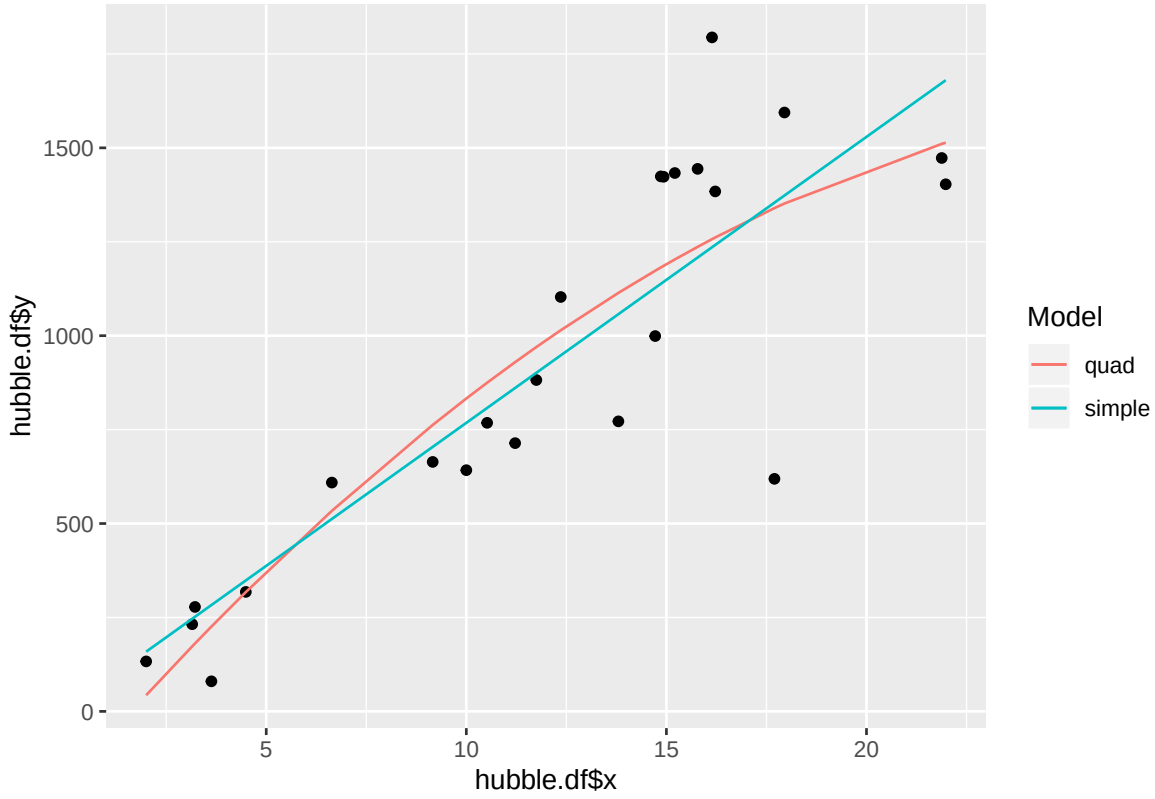
   a) Fit a quadratic regression model, i.e.,a model of the form

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

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

   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.





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

Looking at the plots, it seems as though both fit the data fairly well. I think that the quadratic model fits slightly better on the lower end but worse in the middle and the top end. Considering the plot, the linear model seems more sensible (but this may be incorrect without having looked at MSE yet).

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

The simple linear regression seems to perform slightly better than the quadratic regression model.

```
## Quadratic regression MSE =      1005787
## Simple linear regression MSE =  1020064
```

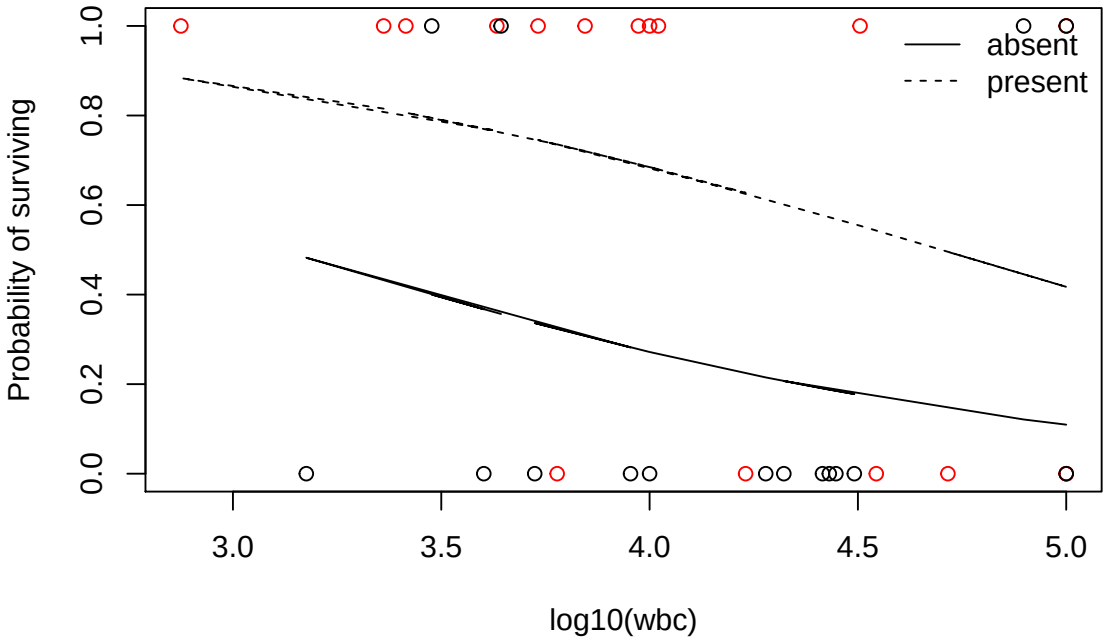
Note: The quadratic model here is still regarded as a `linear regression" model since the term`linear" relates to the parameters of the model and not to the powers of the explanatory variables.

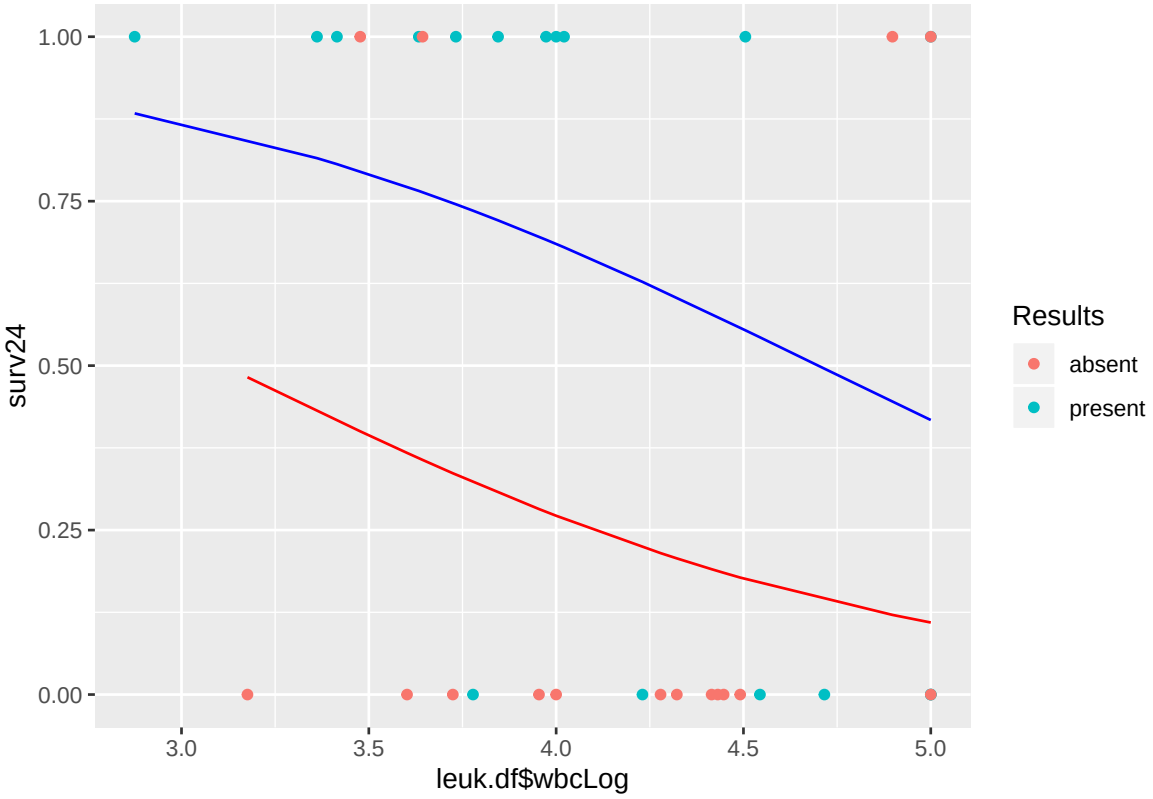
3. The **leuk** data from package **MASS** shows the survival times from diagnosis of patients suffering from leukemia and the values of two explanatory variables, the white blood cell count (wbc) and the presence or absence of a morphological characteristic of the white blood cells (ag).

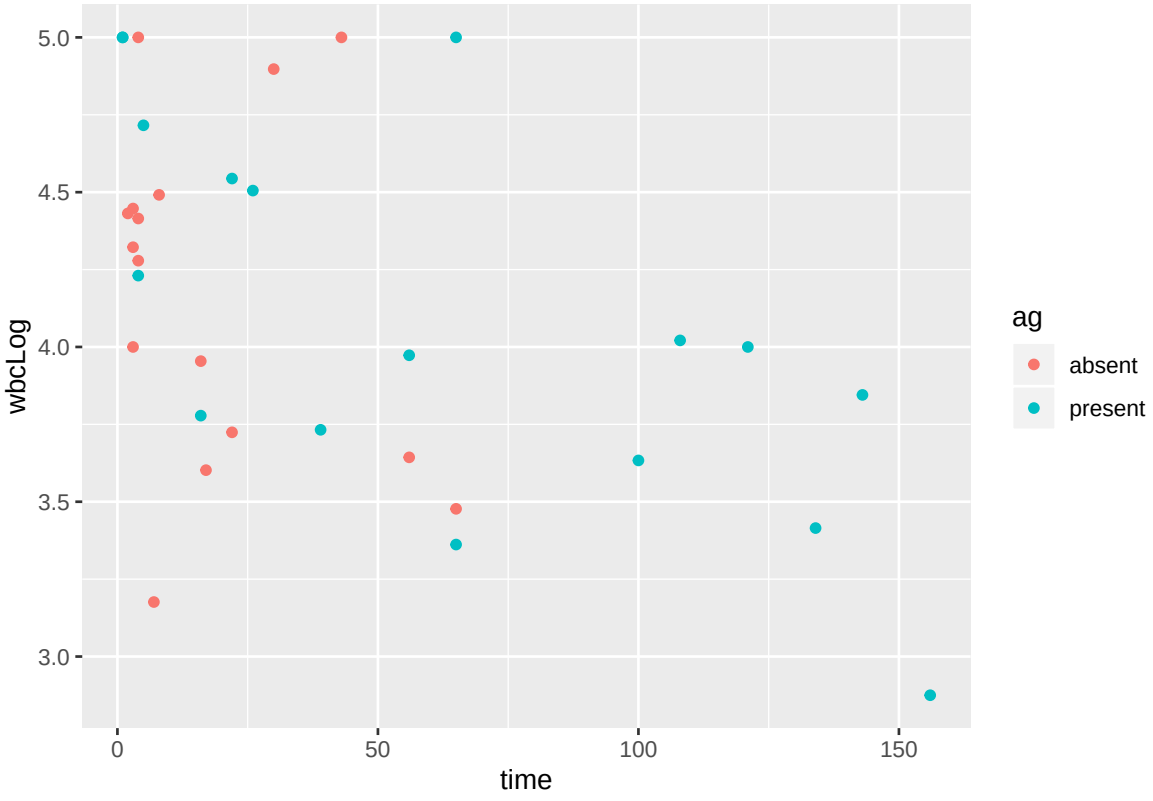
   a) Define a binary outcome variable according to whether or not patients lived for at least 24 weeks after diagnosis. Call it *surv24*.

   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.

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







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

```
## no interaction model MSE: 0.1864727

## interaction model MSE:    0.1712448
```

The interaction model has a lower MSE which is indicative of a better fit to the observations than the non-interaction model does. This means that the information together are better predictors of whether the patient will or will not survive past 24 weeks after diagnosis.

Dicussion: This exercise was difficult only being of the plotting. Trying to get the plots to look right was really difficult. The example given in the GLM_CH7.R for plotting the agree and disagree was not directly applicable here which caused some difficulties. The creation of the model wasn't awfully difficult to perform.

- **Looking at the graphs, up until about 3.5 to 4 log10 of the WBC do the data points begin to become mixed as to whether or not the patient lived longer than 24 weeks after diagnosis. However, the 2 groups of cellular morphology (absent and present) are rather distinct in the values that are predicted which can be viewed in the base R and ggplot graphs. The probability of surviving with cellular morphology present starts to decrease faster as the log10 of WBC increases whereas the probability of surviving past 24 weeks with the absence of cellular morphology levels off as log10 of WBC increases. This also shows the importance of cellular morphology in predicting whether the patient will survive past 24 weeks. The final scatter plot shows more clearly that those with ag present will survive longer than those without it.**

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

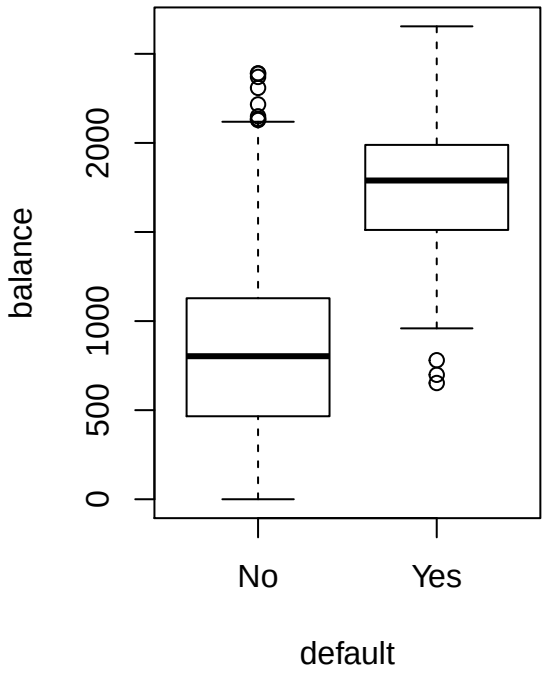
    a) Perform descriptive analysis on the dataset to have an insight. Use summaries and appropriate exploratory graphics to answer the question of interest.

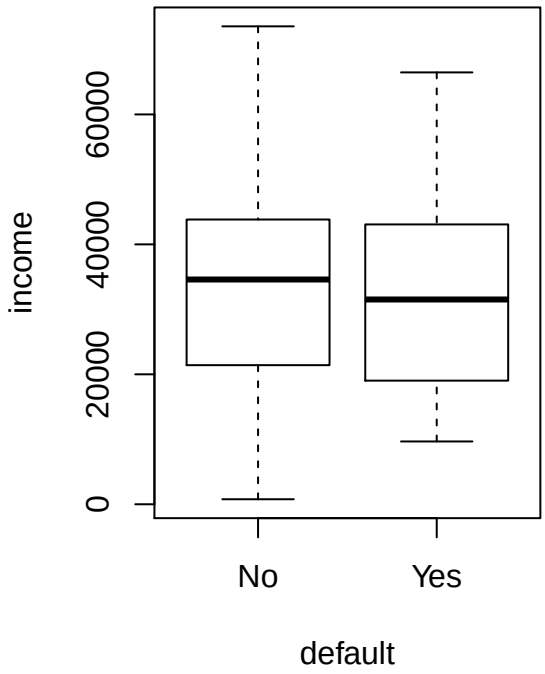
Summary Statistics

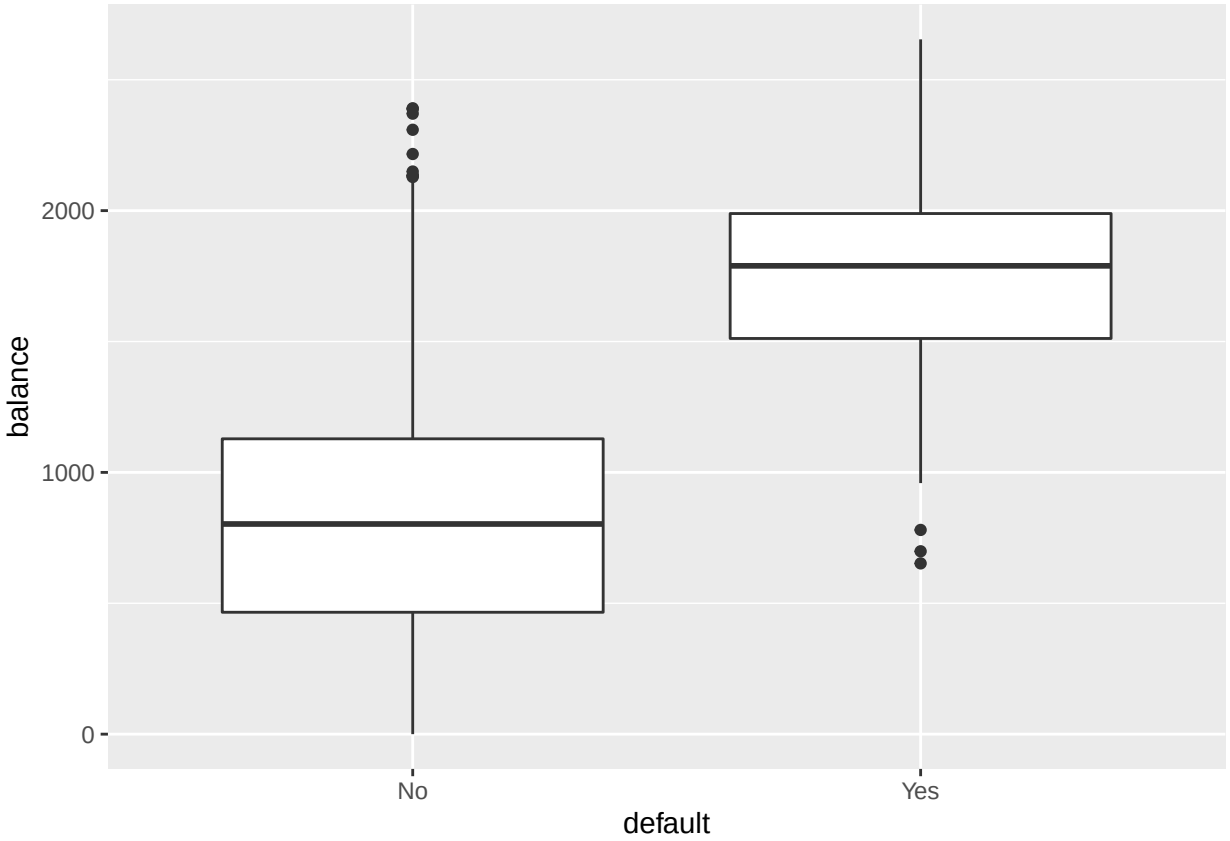
```
##  default   student      balance           income
##  No :  0   No :206   Min.   : 652.4   Min.   : 9664
##  Yes:333   Yes:127   1st Qu.:1511.6   1st Qu.:19028
##                      Median :1789.1   Median :31515
##                      Mean   :1747.8   Mean   :32089
##                      3rd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

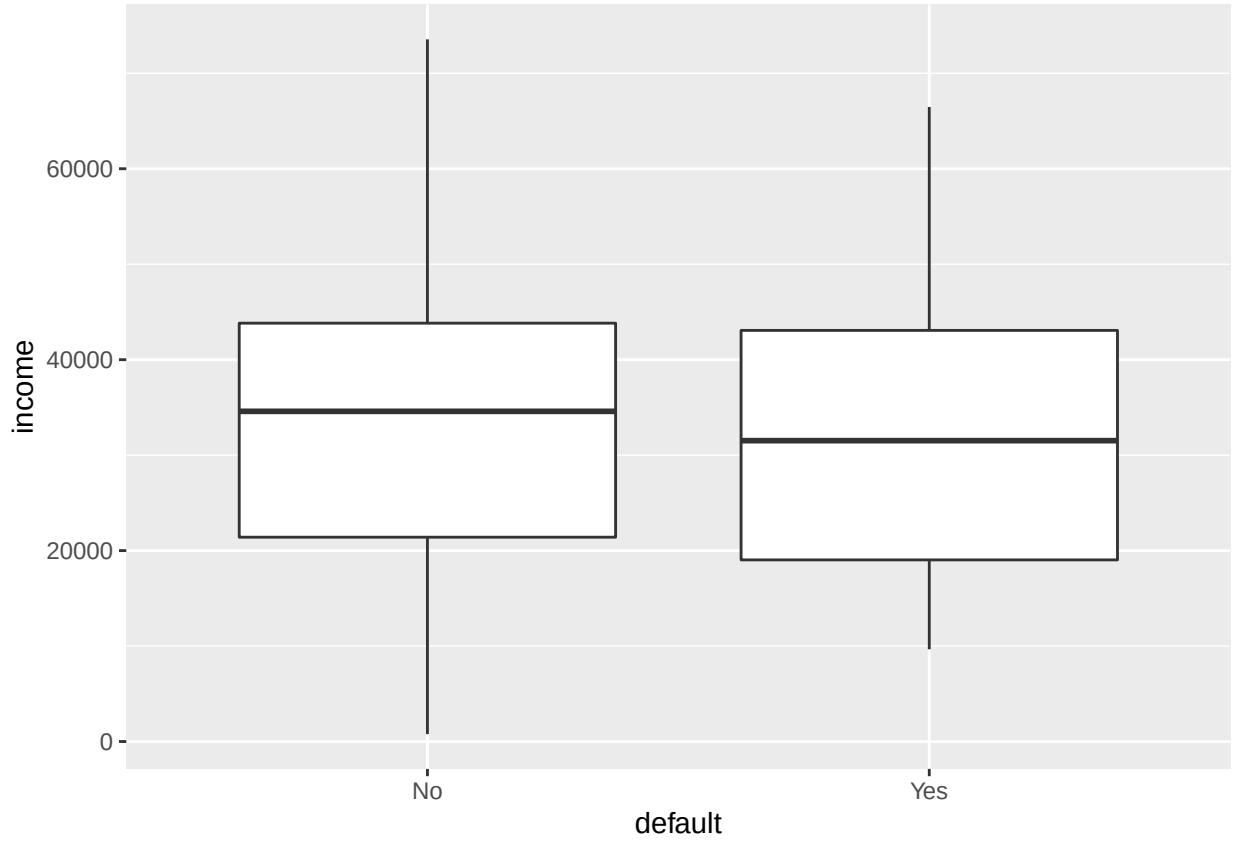
```
##  default    student       balance           income
##  No :9667   No :6850   Min.   :   0.0   Min.   :  772
##  Yes:   0   Yes:2817   1st Qu.: 465.7   1st Qu.:21405
##                        Median : 802.9   Median :34589
##                        Mean   : 803.9   Mean   :33566
##            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

Boxplots to show the difference in the balance and income of those that will default


balance
0
500
1000
2000
No
Yes
default


income
0
20000
40000
60000
No
Yes
default


balance
2000
1000
0
No
Yes
default



b) Use R to build a logistic regression model.

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

```
##
## Call:
## glm(formula = defs ~ student + balance + income + student * balance +
##     student * income + balance * income, family = binomial(),
##     data = default.df)
##
## 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:balance -2.551e-04  7.905e-04  -0.323    0.747
## studentYes:income   1.447e-05  2.779e-05   0.521    0.602
```

```
## 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
```

Including the interaction term coefficients in the model show that the only significant term is the balance. None of the interaction coefficients are significantly different than zero and therefore are not associated (together) in whether or not they will default.

```
##
## Call:
## glm(formula = defs ~ student + balance + income, family = binomial(),
##     data = default.df)
##
## 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
```

Removing the interaction terms in the model shows that balance is still a significant term. Whether or not the individual is a student is also significant in predicting whether the individual is going to default or not which is interesting. Balance is actually more significant in the model without interaction terms than it is with them.

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

```
## Logistic regression with interaction terms MSE =     0.02129659
```

```
## Logistic regression without interaction terms MSE =  0.02129814
```

The MSE of the model with interaction coefficients is slightly lower than the MSE from the

model without interactions. THey are almost identical, down to the 5th decimal place after which they are different.

```
## [1] "Accuracy of LRM with interaction: "
```

```
## [1] 0.9731
```

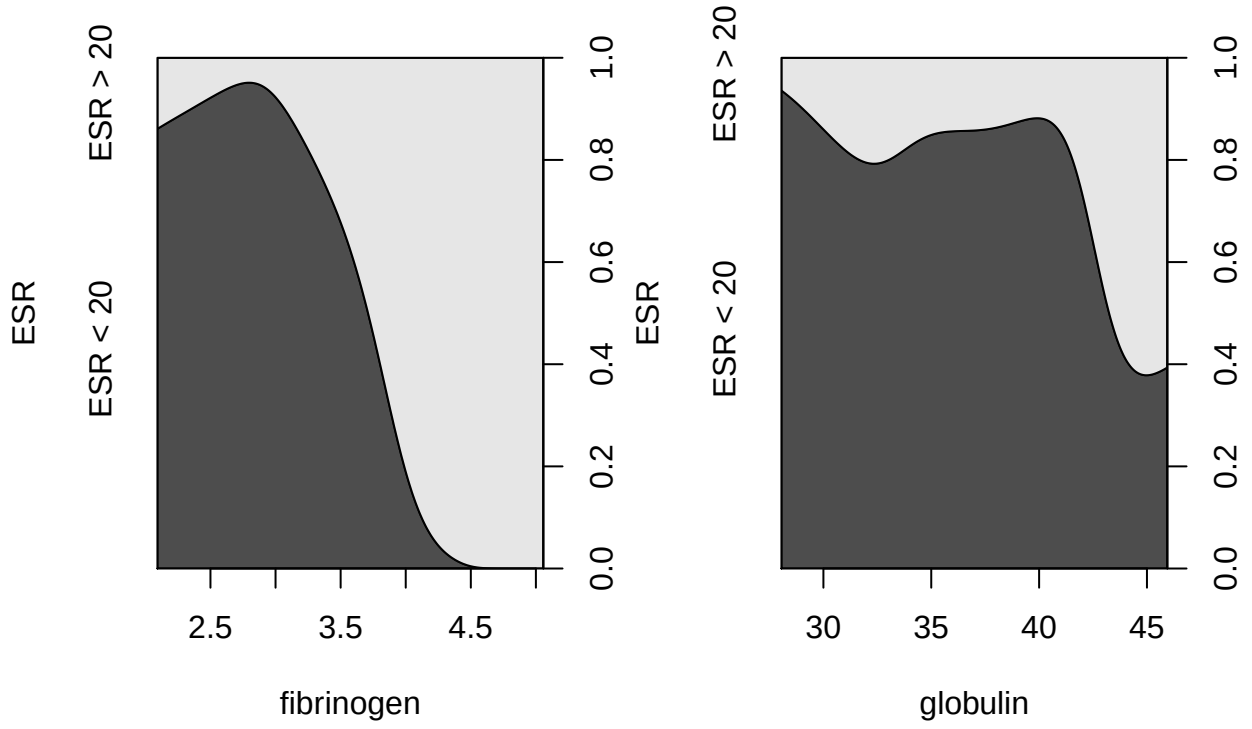
```
## [1] "Accuracy of LRM without interaction: "
```

```
## [1] 0.9732
```

Looking at confusion matrices of the predictions with the true results, and the accuracy of both models is greater than 97%. The model with interaction has a slightly lower accuracy (97.31% v 97.32%) but they are basically the same here too. It is interesting to note that the MSE of the model with interaction was lower than without and the accuracy was slightly better.

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.

Logistic Regression and Generalized Linear Models



These are the distributions of fibinogen and globulin based on the ESR value being lower or higher than 20. The ESR changes greatly as fibinogen increases. The ESR distibution changes a decent amount as globulin increases, but the change is less than that of fibrinogen.

```
##     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
```

Creating a logistic regression model for ESR using fibrinogen shows that as fibrinogen increases, ESR also increases. The fibrinogen coefficient is significant with a p-value of 0.0425. The estimated coefficient is 1.827 and has a 95% confidence interval between 0.339 and 3.998.

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

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

These values are the exponents of the coefficient and confidence interval which are the odds themselves? This is the odds that the response variable, ESR, increases by 1 when fibinogen increases by 1 conditional on all other variables remaining constant (which are none in this case). The confidence interval is exceptionally large, possibly due to only 6 of the 32 data points having an ESR greater than 20.

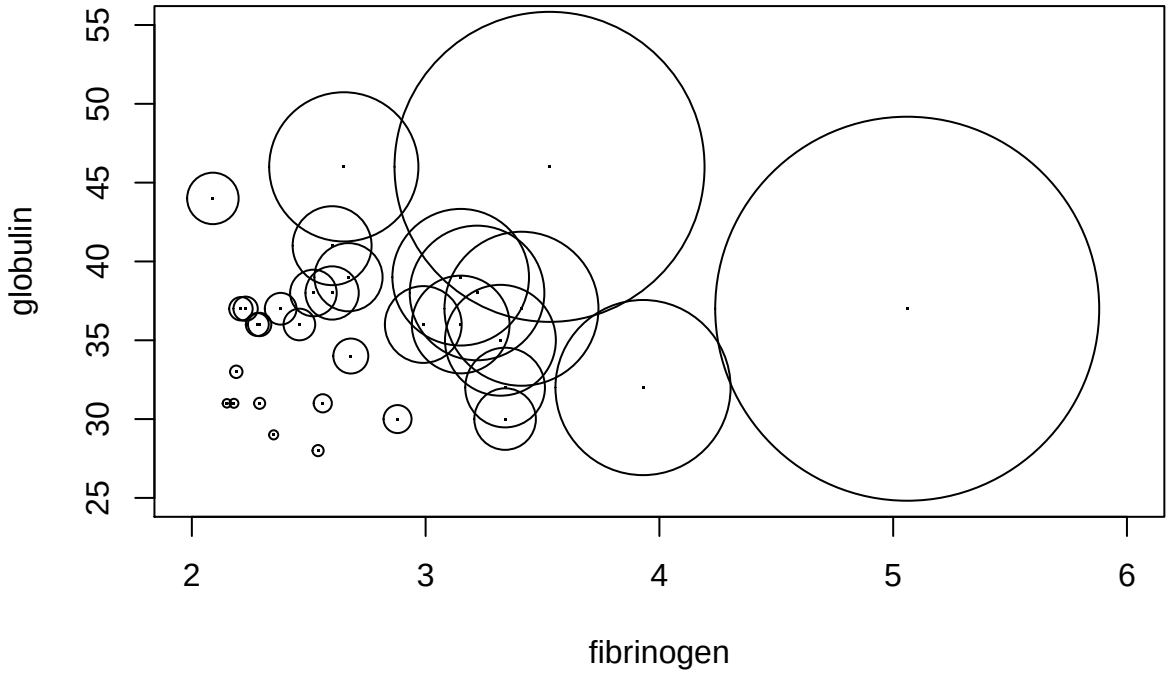
Creating a logistic regression model to predict ESR based in independent variables fibrinogen and globulin, without looking at any interactions.

```
##
## 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
```

From the summary of the logistic regression created using both fibrinogen and globulin as input variables, shows that the fibrinogen term coefficient is significant just like the previous model using only fibrinogen as an input variable. The coefficient for globulin isn't significantly different than 0.

Comparing the two different models, we can see that addinng the globulin term doesn't make the model significantly different from the model without it, and therefore can see that globulin doesn't really contribute to the ESR value.



As seen with the chi-squared test and the second regression model, globulin doesn't contribute greatly to the ESR. Fibrinogen contributes a lot more to the ESR and the plot shows that as fibrinogen increases, the probability of ESR being greater than 20 increases drastically (larger circle indicates a greater probability of having an ESR greater than 20).

Resources Used:

- StackOverflow
- community.rstudio.com
- stats.idre.ucla.edu