Homework #2

Justin Robinette

September 4, 2018

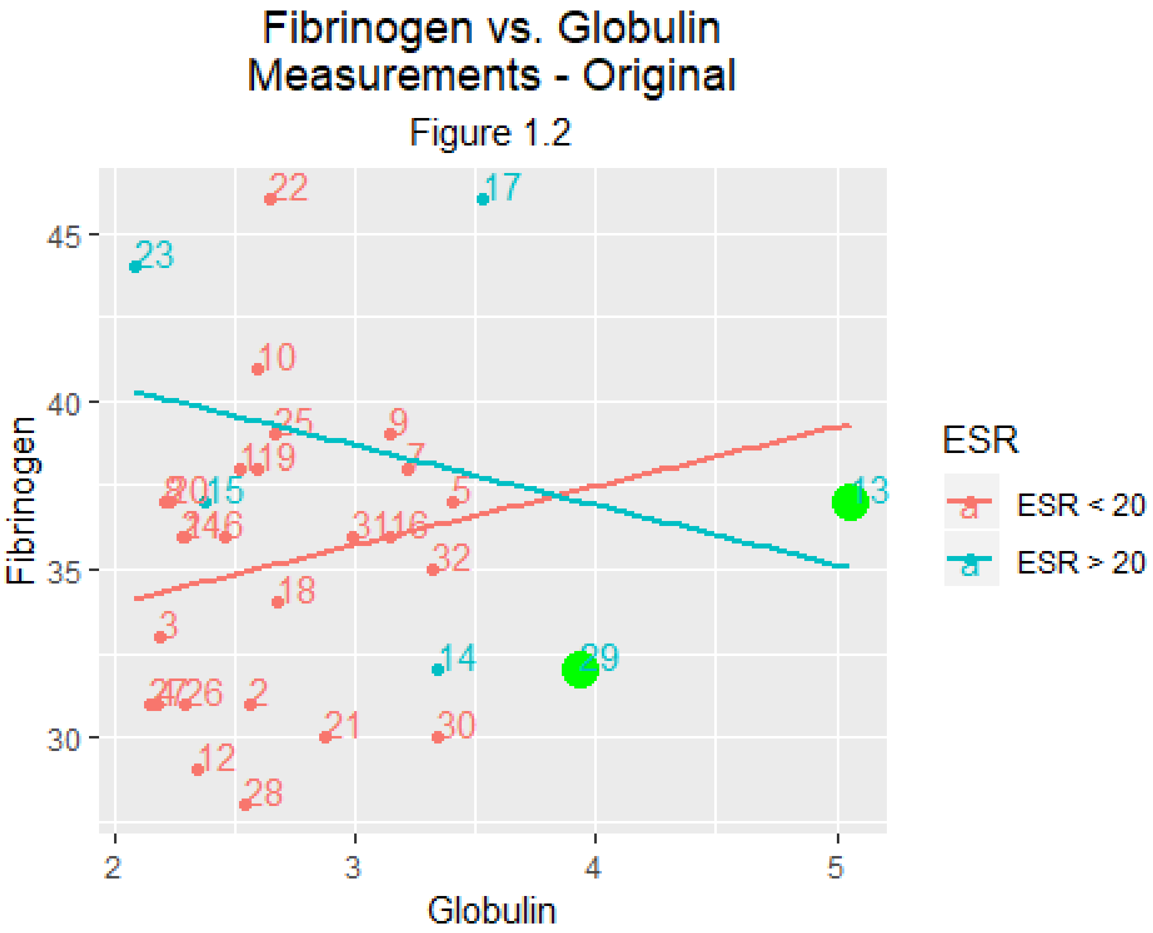
No collaborators for any problem

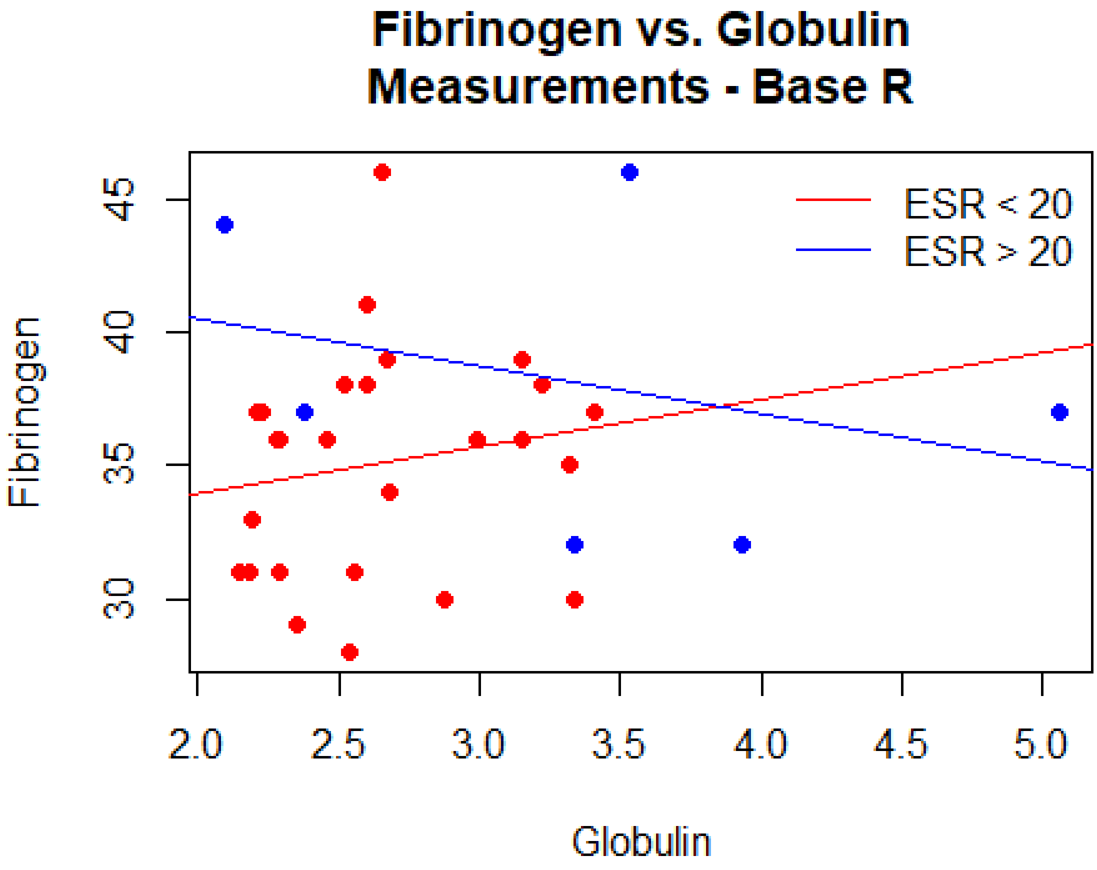
Problem #1: Collett (2003) argues that two outliers need to be removed from the *plasma* data. Try to identify those two unusual observations by means of a scatterplot.

Results: First, per the homework rule, I plotted a scatterplot both in base R (Figure 1.1) and ggplot2 (Figure 1.2). These plots show a few candidates for removal.

Next, I chose to remove observation id numbers 13 and 29 from Figure 1.2. These two points are highlighted in green. These observations are unusual because they contain Fibrinogen levels that are far greater, than that of the remaining data, without an expected increase in the corresponding Globulin levels.

The last plots, Figure 1.3(Base R) and Figure 1.4(ggplot2), provides a look at the scatterplot with these two outliers removed.



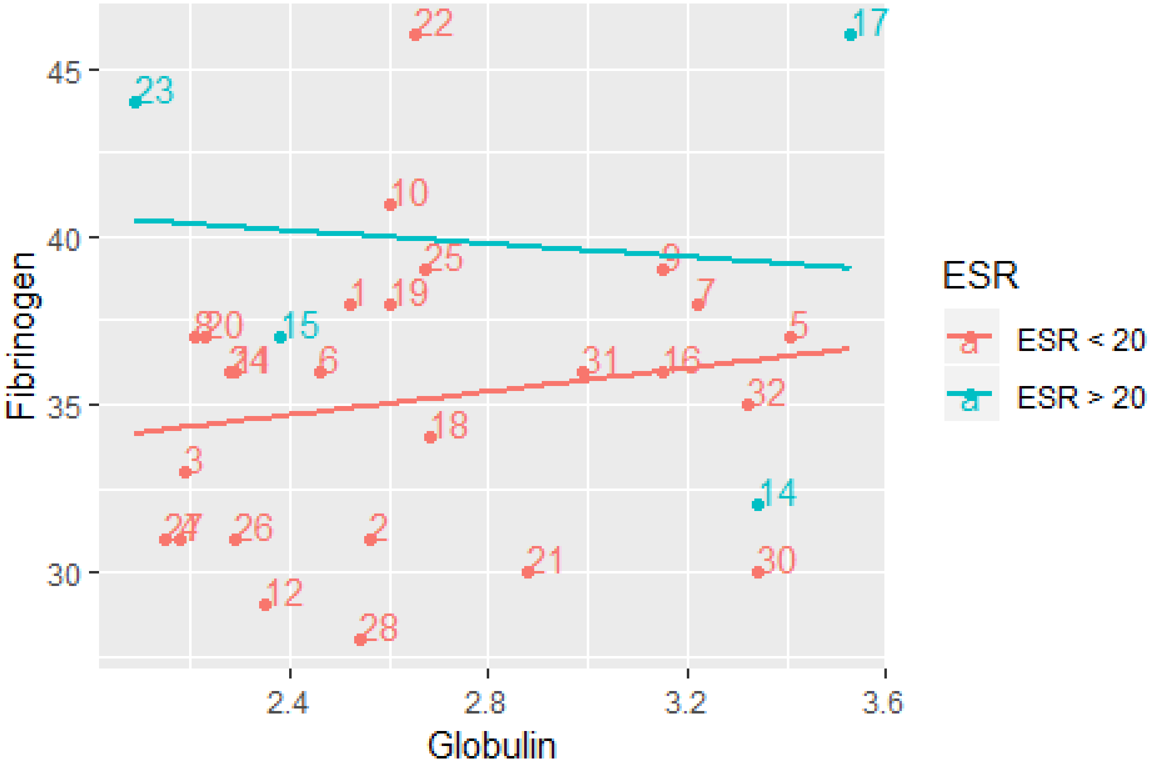


Fibrinogen vs. Globulin
Measurements - Base R

Figure 1.1

Fibrinogen vs. Globulin
Measurements - Updated

Figure 1.4



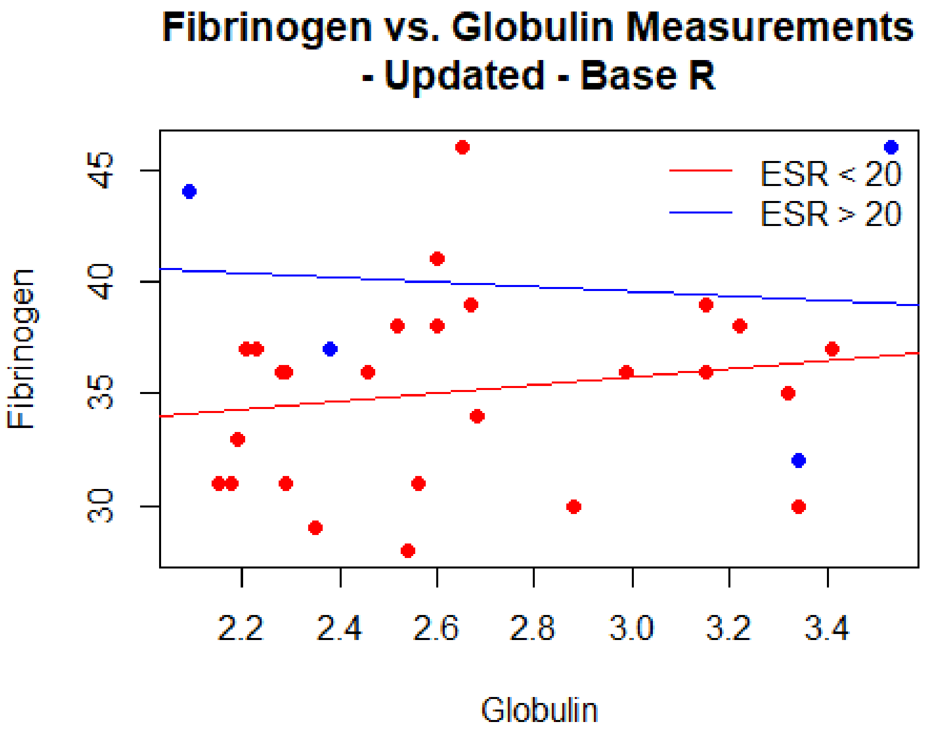


Figure 1.3

Problem #2, Part A: Continuing from the lecture on the *hubble* data from *gamair* library, fit a quadratic regressional model.

Results: I created a square of the distance (x), per the exercise instructions. I then fit a quadratic regression model and included the summary below. I did not subtract '1' from this equation, even though the simple linear model does subtract '1', for two reasons. For one, we were never instructed to use the 'hmod' formula from the text, let alone to subtract '1' from the quadratic regression model; we were only asked to *fit a quadratic regression model*. Additionally, with or without subtracting '1', the simple linear model fits the data much better.

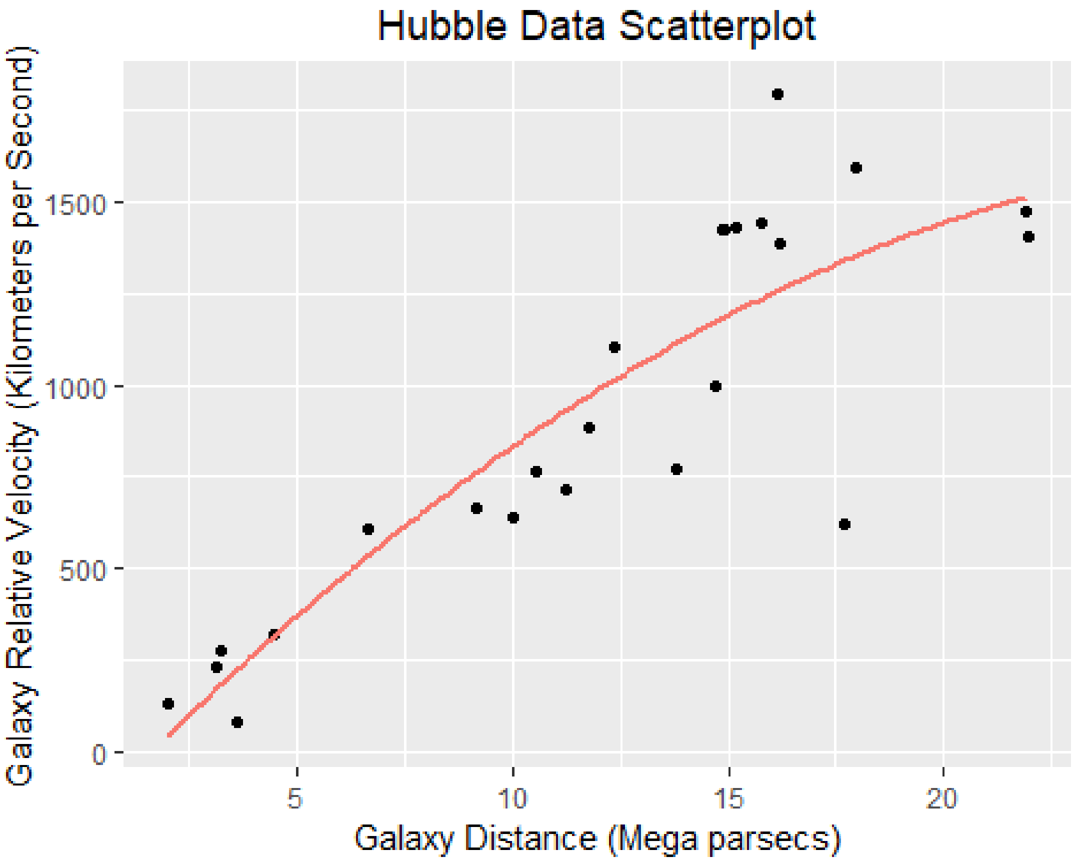
The first thing I notice from the summary, is that the x^2 value is not statistically significant in the relationship with 'y'.

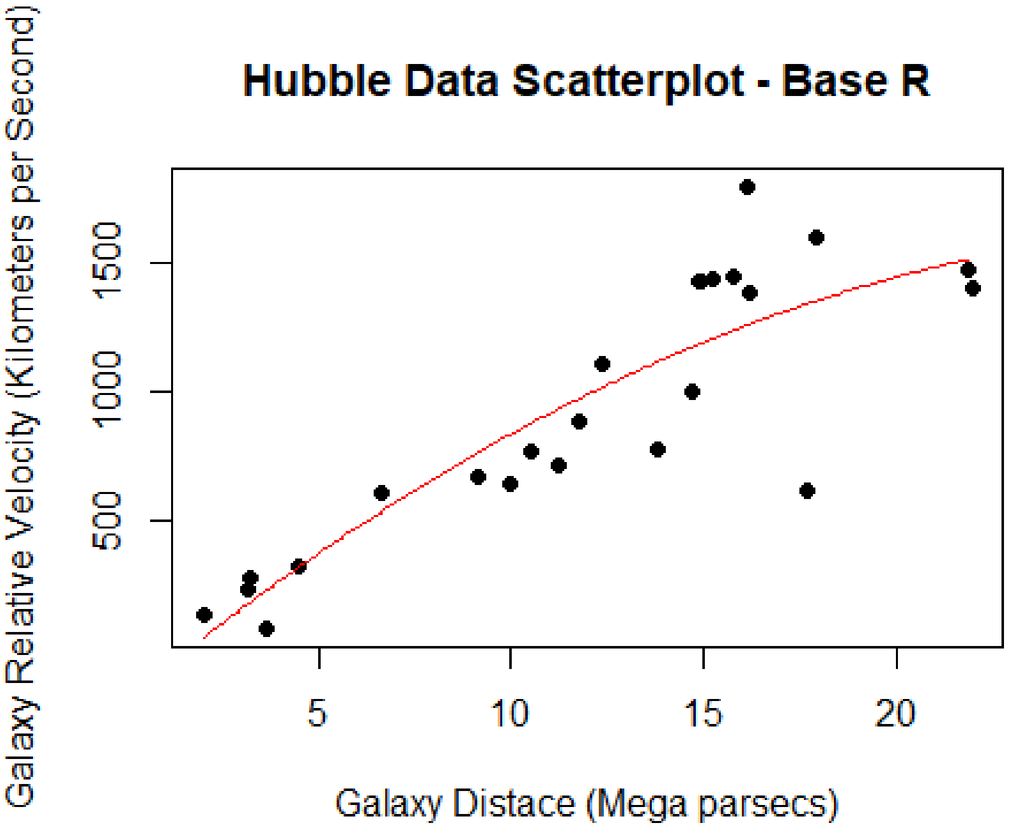
Next, I created a sequence of x values from the hubble data set, incrementing by 0.01 from the min(x) to the max(x). Then, I used predict() function to get y values using the incremented 'x_values' and squared 'x_values'.

```
##
## Call:
## lm(formula = y ~ x + x2, data = hubble)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -720.5 -119.5   29.7  143.8  537.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -196.364    196.122  -1.001  0.32811
## x            123.871     36.861   3.361  0.00296 **
## x2            -2.096      1.565  -1.339  0.19494
## ---
## 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
```

Problem #2, Part B: Plot the fitted curve from Model 2 on the scatterplot of the data.

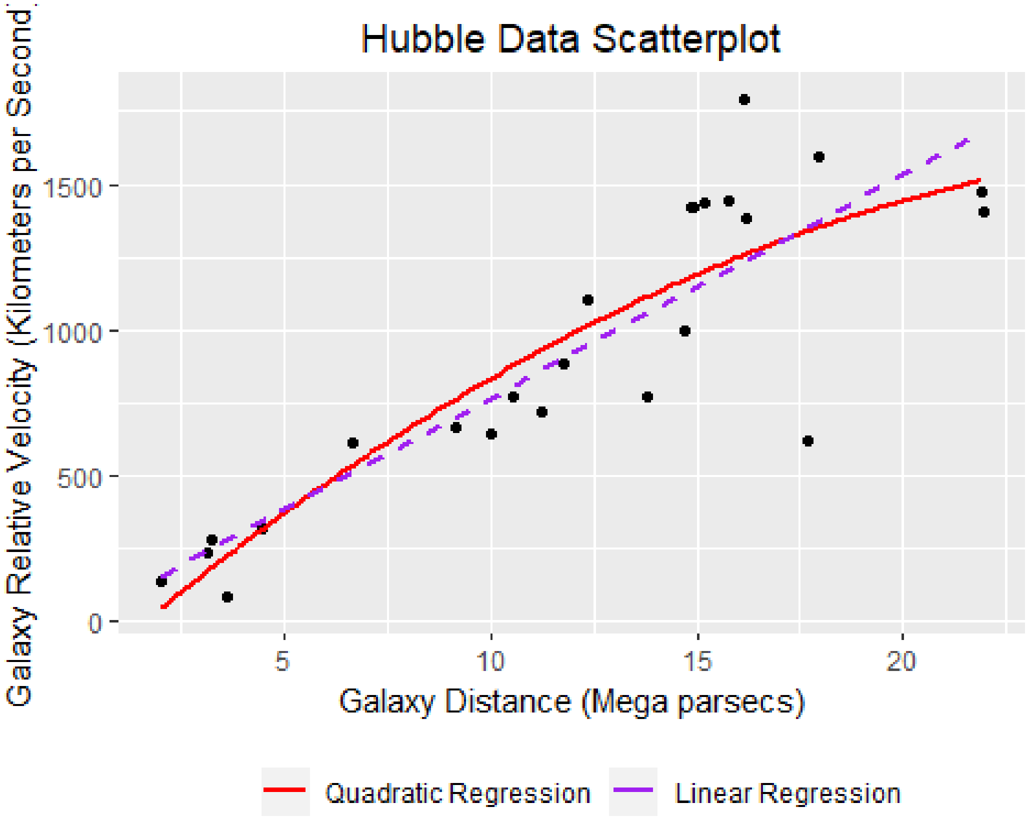
Results: Using the x and y values as a data frame, we are able to plot the fitted curve, in red, on the scatterplot of hubble data. Using the quadratic regression model, this curve attempts to minimize the vertical displacement between the points and the curve.

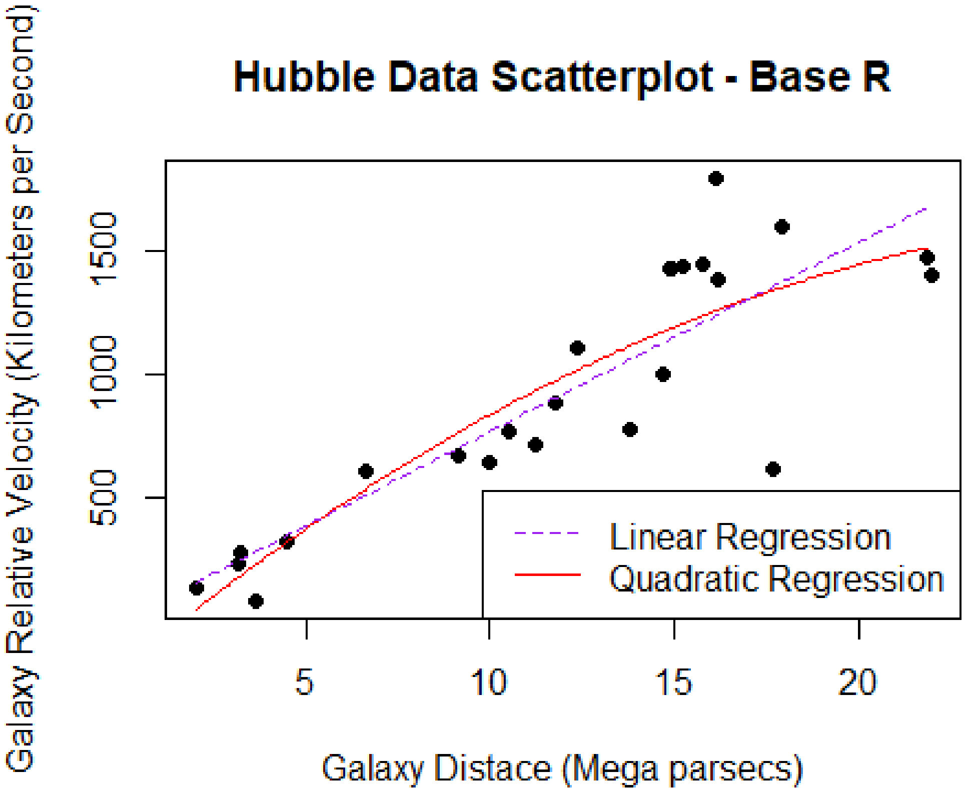




Problem #2, Part C: Add the simple linear regression fit (fitted in class) on this plot - use different color and line type to differentiate the two and add a legend to your plot.

Results: Here I've added the simple linear regression line to the previous plot. To differentiate it from the fitted curve, and per the homework instructions, I've added it as a purple dashed line.





Problem #2, Part D: Which model do you consider most sensible considering the nature of the data - looking at the plot?

Results: Looking at the plot, it appears the simple linear regression line is most sensible. There is a cluster of data points beginning near x=15 y=1400 and the fitted curve is moving in the opposite direction from that cluster.

Problem #2, Part E: Which model is better? Provide a statistic to support your claim.

Results: The simple linear model is better than the quadratic model in this instance. The multiple and adjusted R-squared values are both higher in the simple linear regression model.

The adjusted R-squared in the simple linear model is 0.9394 and in the quadratic model it is 0.7428. This value is a modified version of the R-squared value that has been altered to take into account the number of predictors in the model. It represents the percentage of the variation in the response variable that can be explained by the independent variables.

Additionally, the F value for the simple linear model is more than 10 times higher than that of the quadratic model, indicating that the simple linear model is superior.

```
## 
## 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
```

```
## 
## Call:
## lm(formula = y ~ x + x2, data = hubble)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -720.5 -119.5   29.7  143.8  537.1 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -196.364    196.122  -1.001  0.32811   
## x            123.871     36.861   3.361  0.00296 **
## x2            -2.096      1.565  -1.339  0.19494   
## ---
## 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
```

Problem #3, Part A: 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).

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

Results: I created a factor column in *leuk* called *surv24* that is “Yes” if the patient survived at least 24 weeks and equals “No” if the patient did not. Based on the ‘Details’ portion using ‘?leuk’ I also changed ‘ag’ to “positive” and “negative” to represent a positive or negative result on the test.

```
##     wbc       ag time surv24
## 1  2300 positive   65    Yes
## 2   750 positive  156    Yes
## 3  4300 positive  100    Yes
## 4  2600 positive  134    Yes
## 5  6000 positive   16     No
## 6 10500 positive  108    Yes
```

Problem #3, Part 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 ration close to 1). You may use log transformation.

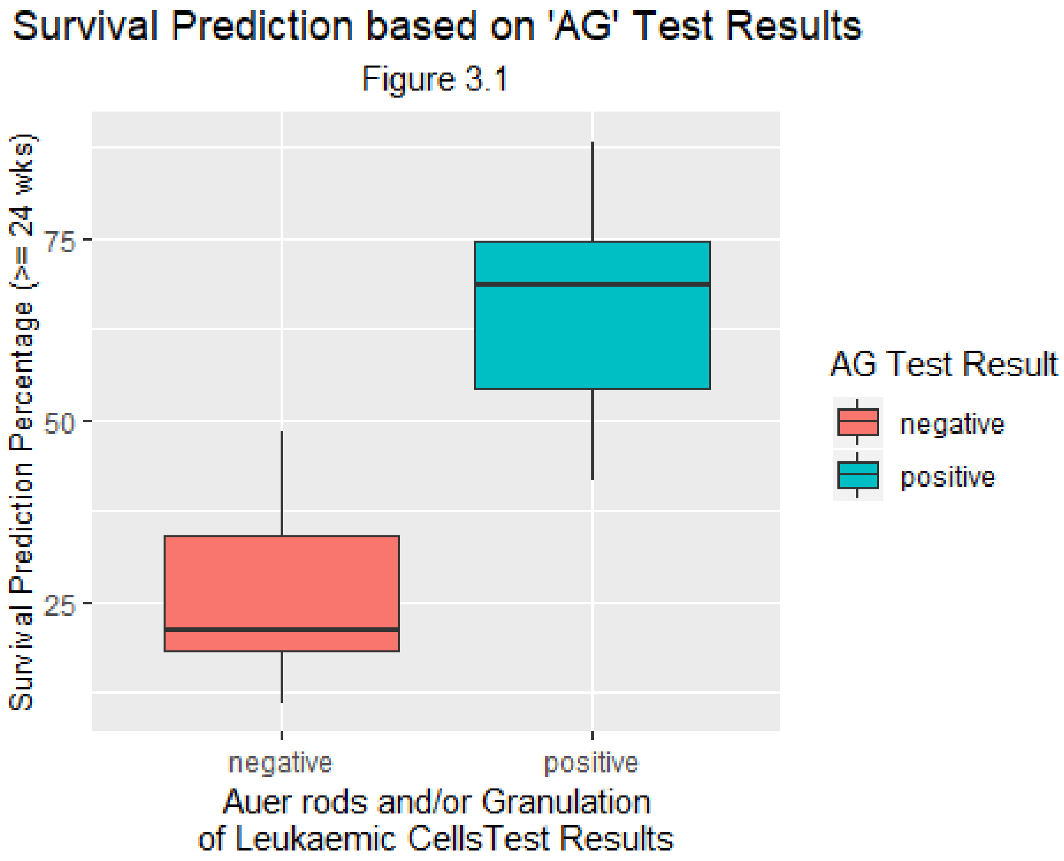
Results: From the summary of our fitted logistic regression model, we can see that the ‘ag’ test results is statistically significant at alpha = 0.05. This means that there is a statistically significant relationship between the test result for ‘ag’ and surviving for at least 24 weeks after leukemia diagnosis.

Under the heading of ‘Coefficients’, we can see that ‘agpositive’ has a positive Estimate. From this we can deduce that having a ‘positive’ value in the ‘ag’ column makes survival of at least 24 weeks after diagnosis more likely. Conversely, we can see that the ‘log(wbc)’ Estimate is negative. That means that as the white blood count increases, the likelihood of surviving at least 24 weeks after diagnosis decreases.

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

Problem #3, Part C: Construct some graphics useful in the interpretation of the final model you fit.

Results: There are four important illustrations for this section. Figure 3.1 shows the likelihood of surviving 24+ weeks, as predicted by the model, based on the results of the ‘ag’ test. As we can see, the percentage chance of surviving 24 or more weeks increases if the ‘ag’ test result is positive.



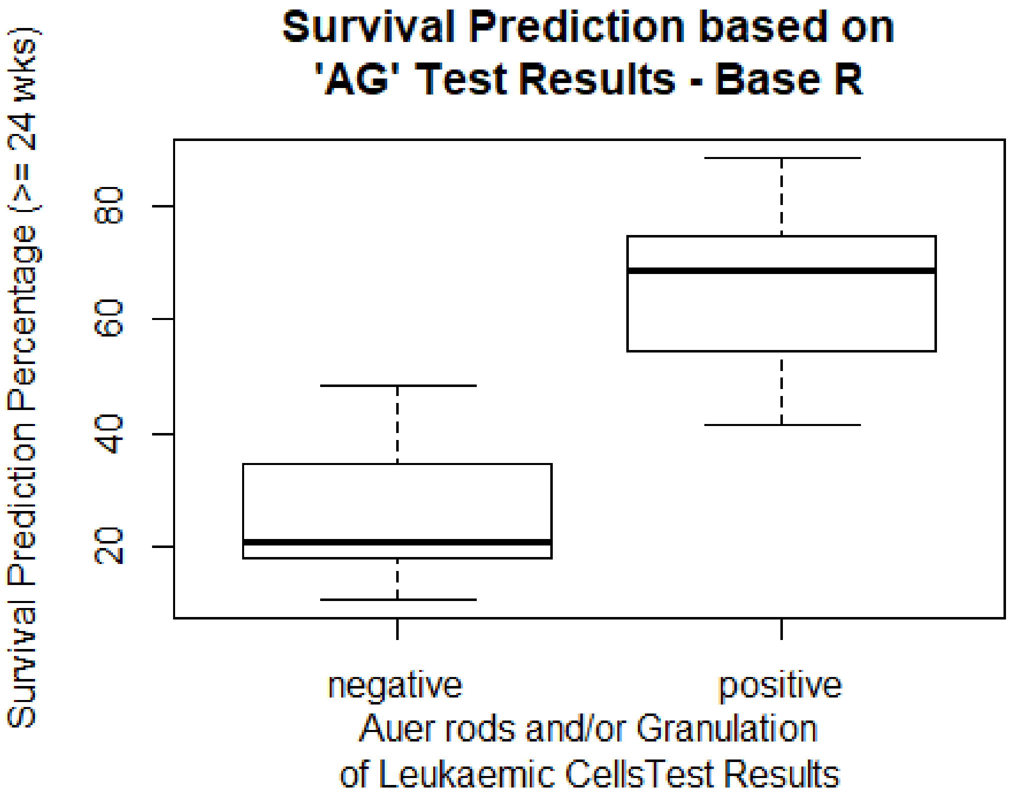


Figure 3.2 shows the likelihood of surviving 24+ weeks, as predicted by the model, based on the patient's white blood count. As that count increases, we see the model predicts a lesser percentage chance of surviving at least 24 weeks.

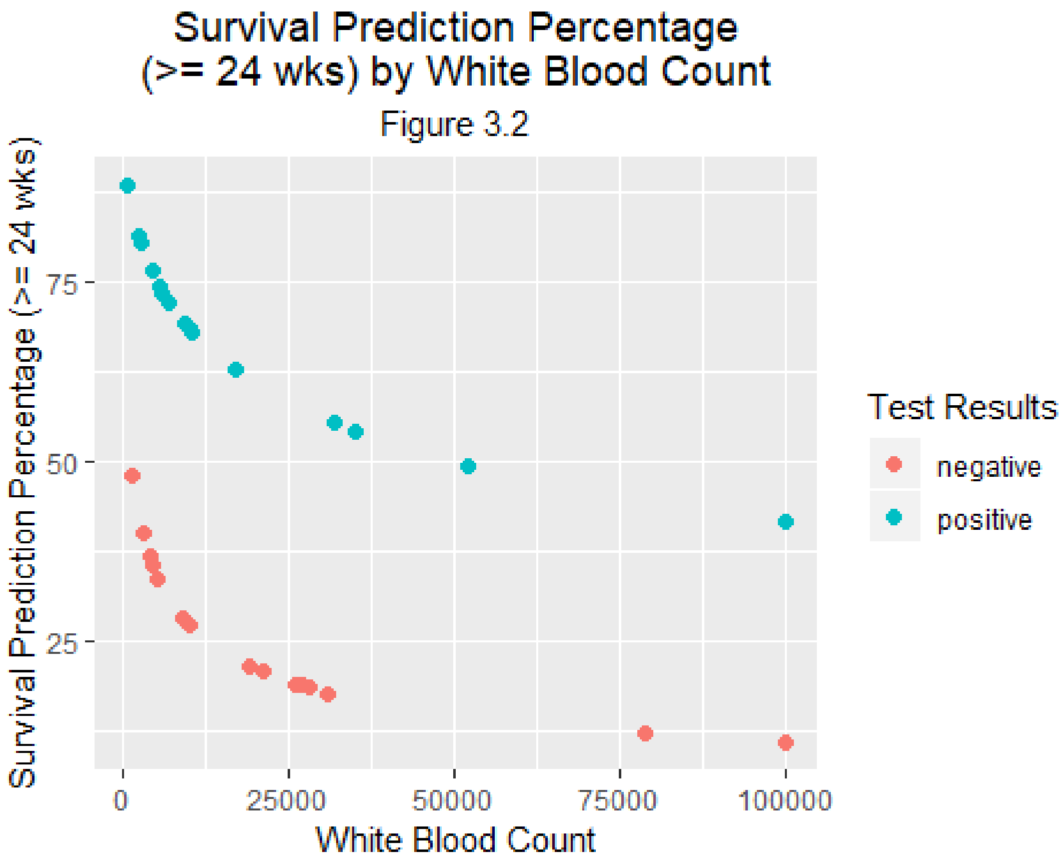

Survival Prediction Percentage (>= 24 wks) by White Blood Count

Figure 3.2

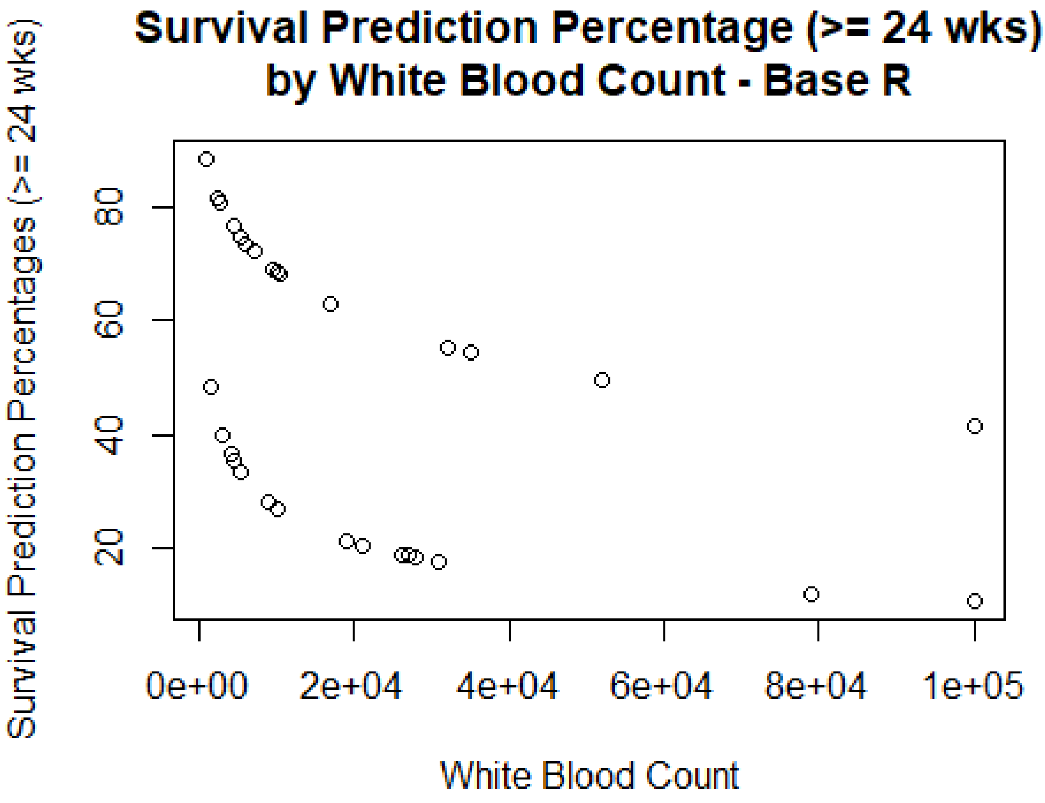
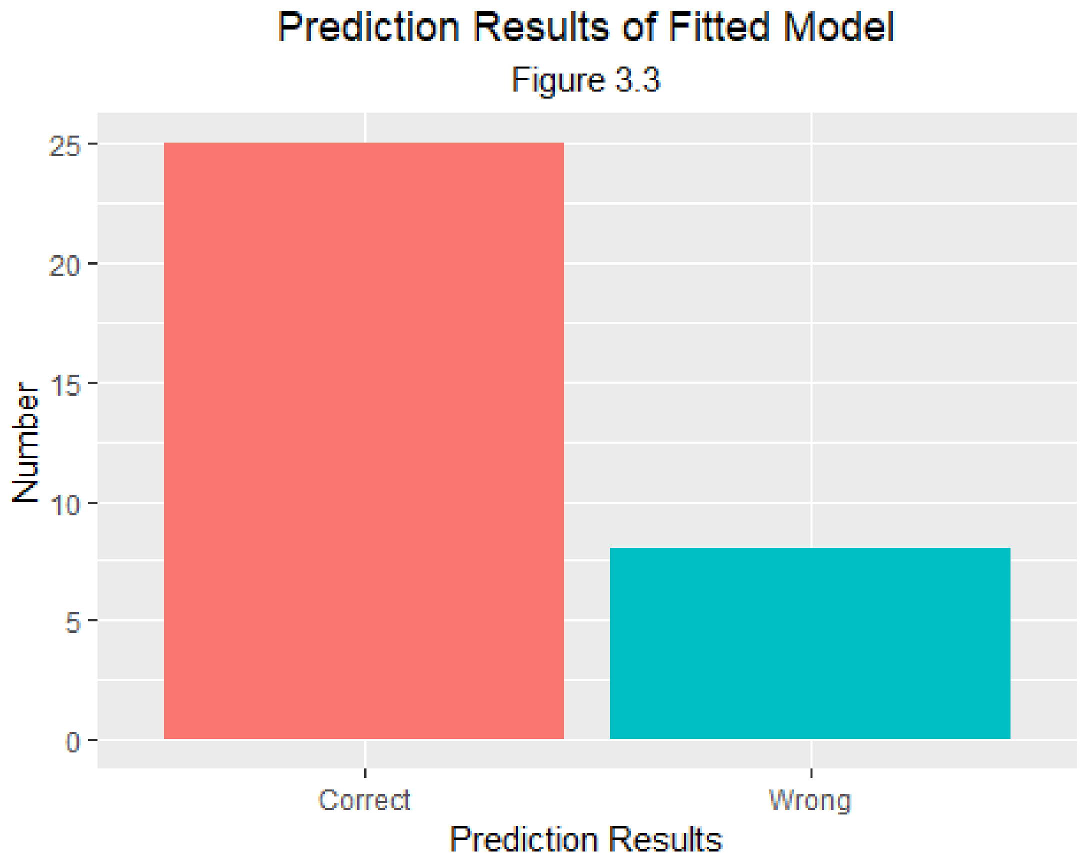
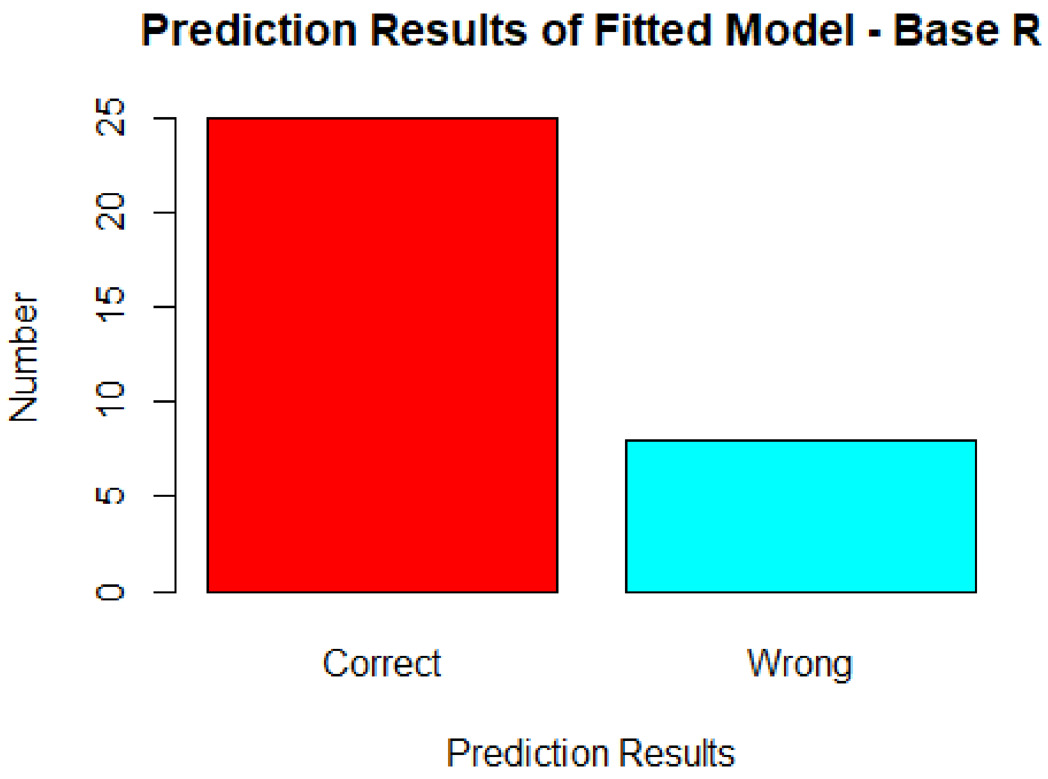

Survival Prediction Percentage (>= 24 wks) by White Blood Count - Base R

Figure 3.3 shows the accuracy of our model in predicting the actual results of the study. As we can see, the model was correct 25 times out of 33 observations.



Prediction Results of Fitted Model

Figure 3.3



Prediction Results of Fitted Model - Base R

Finally, we can see a Confusion Matrix summarizing the data in Figure 3.3. Also, we can deduce that the model is more accurate in predicting that a patient will not survive 24 weeks. The model is less accurate when predicting that the patient will survive at least 24 weeks.

```
##           Observed
## Predicted No Yes
##        No  15   3
##        Yes  5  10
```

Problem #3, Part D: Fit a model with an interaction term between the two predictors. Which model fits the data better? Justify your answer.

Results: The confusion matrices show the same level of accuracy between the two models. The histograms (Figure 3.3 and 3.4) provide the same conclusion.

Even though the number of correct predictions did not change, the model with the interaction is a better *predictive* model due to difference in the AIC values. The AIC table shows the respective AICs. The Alkaike information criterion (AIC) is an estimator of the relative quality of statistical models. The AIC is used as a predictor of the success of the model for future use.

The following formula can be interpreted as being proportional to the probability that one model minimizes the (estimated) information loss.

$$e^{(AIC\ min - AIC\ i)/2}$$

Using this formula, we see that the model without the interaction is only 0.5139 as probable to minimize the (estimated) information loss.

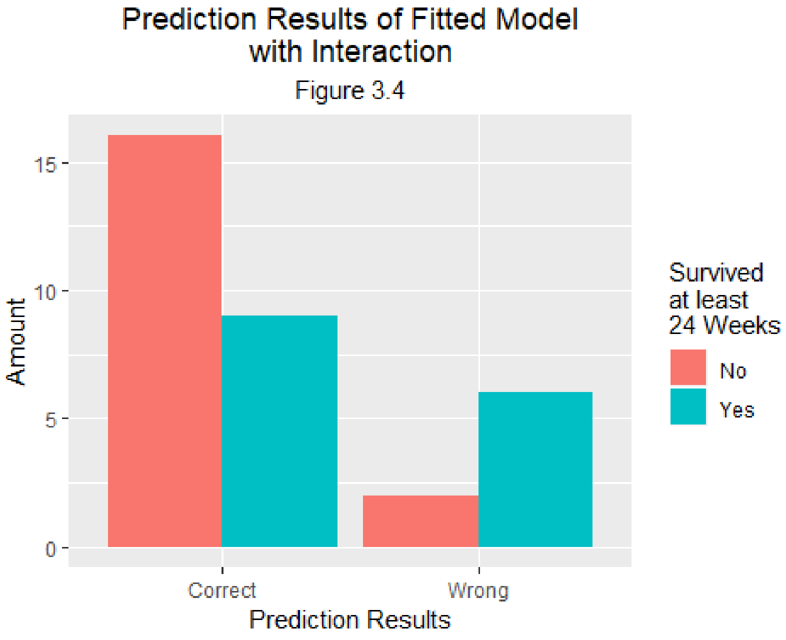
$$e^{(42.16667 - 43.49815)/2} = 0.5139$$

Despite the model with an interaction being a better predictive model, the model that best fits the data is the one with the lowest *p-value*. As we see, the model without the interaction has a lower *p-value*.

To summarize, the model with the interaction is a better predictive model but **the model without the interaction better fits the data** we have.

Credit to *Burnham, K.P.; Anderson, D.R. (2002): Model Selection and Multimodel Inference: A practical information-theoretic approach (2nd ed.)*

No base R plots were included since Part D did not explicitly request plots.



Prediction Results of Fitted Model with Interaction

Figure 3.4

```
##          Observed
## Predicted No Yes
##       No  16   2
##       Yes  6   9
```

```
## 
## Call:
## glm(formula = surv24 ~ ag + log(wbc), family = binomial, data = leuk)
## 
## 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  
## agpositive    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 = leuk)
## 
## 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  
## agpositive           13.6306     7.0909   1.922   0.0546 .
## log(wbc)              0.1545     0.4746   0.326   0.7447  
## agpositive: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
```

```
##   AIC without Interaction AIC with Interaction
## 1                43.49815             42.16667
```

Problem #4, Part A: 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 10,000 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).

Perform descriptive analysis on the dataset to have an insight.

Results: Here I've provided an overall summary, a summary where default and student values were "yes", a summary where the default was "yes" and the balance was greater than the mean, and a summary where the default was "yes" and the income was less than the mean.

I've also provided a table showing the 'MeanIncome' and 'MeanBalance' for observations, grouped by 'default' and 'student' status.

Finally, there are 4 plots. Figure 4.1 shows that students appear to have a slightly fewer number of defaults, than non-students, despite there being far less students in the study.

Figure 4.2 shows that having a balance that is above average leads to more defaults.

Figure 4.3 shows that having below average income makes one more likely to default.

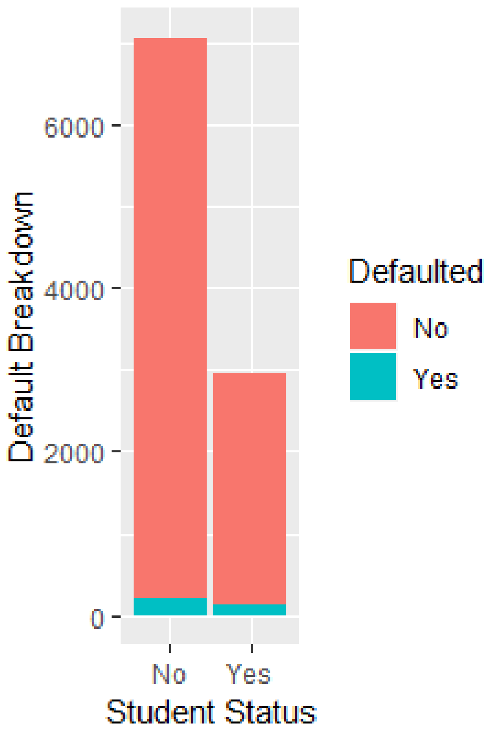
Figure 4.4 shows that students with below average income and above average balance have a similar number of defauls despite having approximately 6,000 fewer observations.

No base R plots were included since the question did not explicitly request plots.

```
##  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 :  0   Min.   :1013   Min.   : 9664  
##  Yes:127   Yes:127   1st Qu.:1638   1st Qu.:15241  
##                      Median :1889   Median :18021  
##                      Mean   :1860   Mean   :18244  
##                      3rd Qu.:2110   3rd Qu.:20809  
##                      Max.   :2654   Max.   :32761  

##  default   student      balance           income     
##  No :  0   No :203   Min.   : 959.2   Min.   : 9664  
##  Yes:330   Yes:127   1st Qu.:1530.4   1st Qu.:18981  
##                      Median :1789.9   Median :31416  
##                      Mean   :1757.3   Mean   :31988  
##                      3rd Qu.:1991.0   3rd Qu.:43008  
##                      Max.   :2654.3   Max.   :66466  

##  default   student      balance           income     
##  No :  0   No : 48   Min.   : 698.6   Min.   : 9664  
##  Yes:175   Yes:127   1st Qu.:1554.1   1st Qu.:16870  
##                      Median :1809.4   Median :19336  
##                      Mean   :1800.2   Mean   :20599  
##                      3rd Qu.:2029.7   3rd Qu.:24685  
##                      Max.   :2654.3   Max.   :33453  
## # A tibble: 4 x 4
## # Groups:   default [?]
##   default student MeanIncome MeanBalance
##   <fct>   <fct>        <dbl>       <dbl>
## 1 No      No          39994.        745.
## 2 No      Yes         17937.        948.
## 3 Yes     No          40625.       1678.
## 4 Yes     Yes         18244.       1860.
```

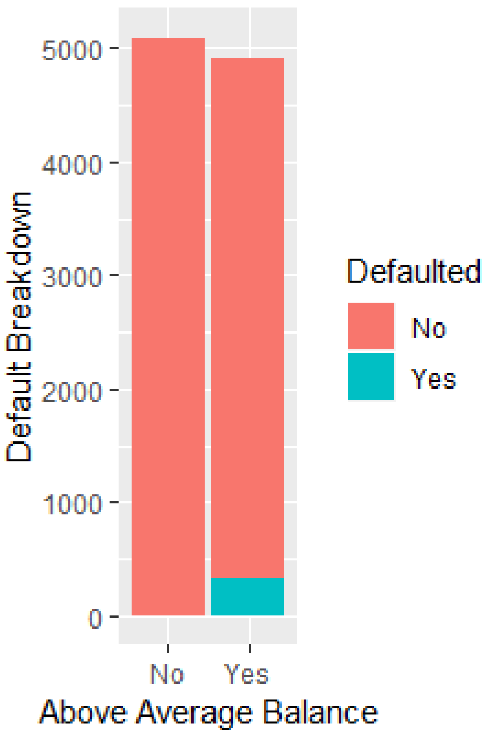
Defaults by Students

Figure 4.1



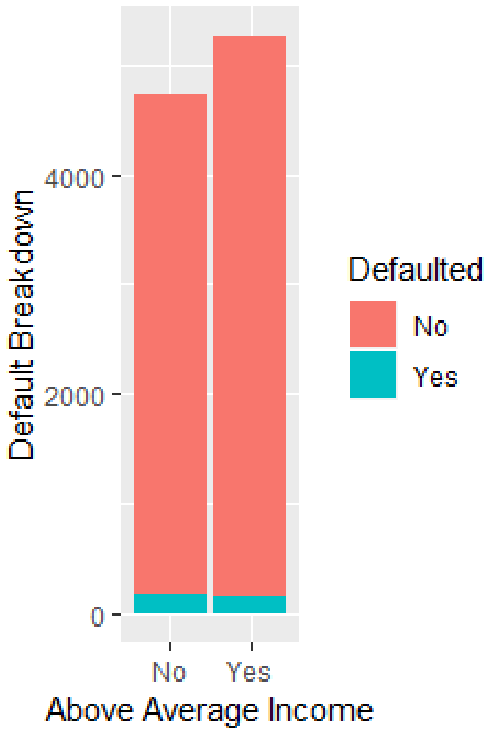
Defaults by Balance

Figure 4.2



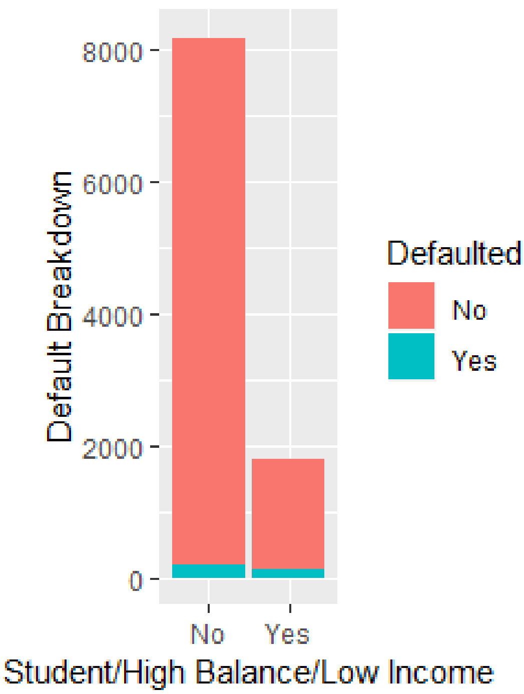
Defaults by Income

Figure 4.3



Defaults by Low-Income Student with High Balance

Figure 4.4



Problem #4, Part B: Use R to build a logistic regression model.

Results: I'll refer to the first model as the **'default'** model. I also did an alternative model, referenced herein as the **'alternative'** model, that removed the 'income' treatment and added 'PoorStudHighBalance' as a treatment. The **alternative** model has a slightly lower AIC score but I will retain both going forward. I will calculate the error rates of both and decide on the best model.

```
##
## 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 ~ balance + student + PoorStudHighBal,
##     family = binomial, data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4498  -0.1431  -0.0565  -0.0132   3.7376
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.071e+01  3.716e-01 -28.824   <2e-16 ***
## balance             5.714e-03  2.336e-04  24.456   <2e-16 ***
## studentYes         -1.233e+01  3.017e+02  -0.041    0.967
## PoorStudHighBalYes  1.162e+01  3.017e+02   0.039    0.969
## ---
## 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: 1570.8  on 9996  degrees of freedom
## AIC: 1578.8
##
## Number of Fisher Scoring iterations: 18
```

Problem #4, Part C: Discuss your result. Which predictor variables were important? Are there interactions?

Results: In both models, the most important predictor variable is the ‘balance’. This variable has a significance value of far less than 0.001. In the **default** model, ‘student’ is also statistically significant at 0.001.

In the **alternative** model, the two variables, ‘student’ and ‘PoorStudentHighBal’, do not have a statistically significant effect on the model. They do, however, improve the AIC score which was discussed, in detail, in my results of Exercise 3, Part D. ‘Balance’ is still very significant.

In both models, I attempted to include other interactions - multiplying ‘student’ and ‘balance’, for example - but each iteration produced a higher AIC score than the ‘base’ model and the alternative model.

Problem #4, Part D: How good is your model? Assess the performance of the logistic regression classifier. What is the error rate?

Results: After using each fitted model to predict defaults, I printed both confusion matrices. ‘Def.Predicted’ represents the predictions from the **default** model and ‘Alt.Predicted’ from the **alternative** model. From the confusion matrices, we see that the **alternative** model is superior in predicting ‘default’ from the Default dataset.

Lastly, I included the error rates from both models. As we can see, the error rate from the **alternative** model is 0.01% less than that of the **default** model.

Based on these results, my chosen model is the **alternative** model.

```
##                Observed
## Def.Predicted   No  Yes
##            No  9627   40
##            Yes  228  105
```

```
##                Observed
## Alt.Predicted   No  Yes
##            No  9628   39
##            Yes  228  105
```

```
## [1] "The error rate for the default model is:     2.68 %"
```

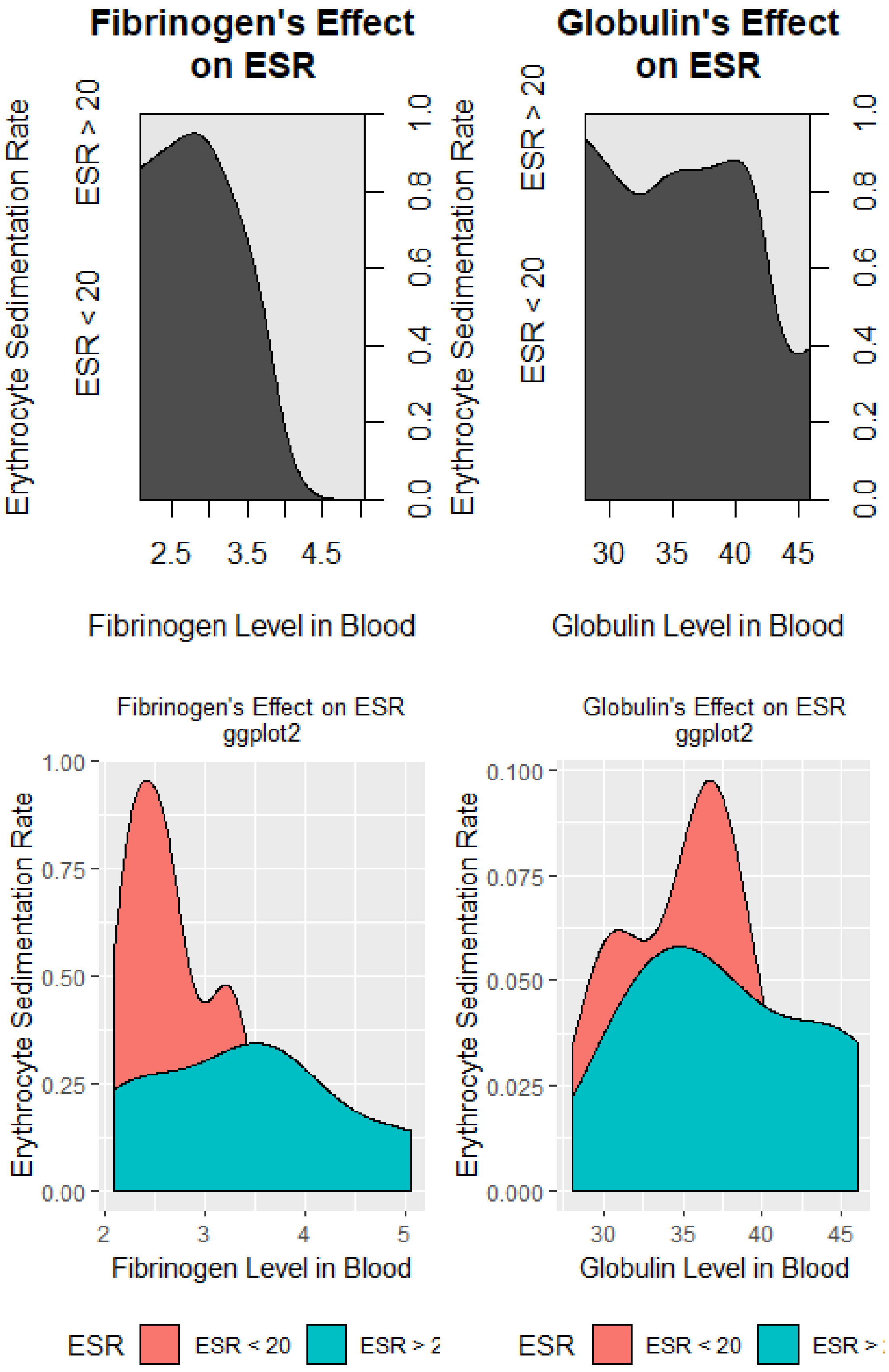
```
## [1] "The error rate for the alternative model is: 2.67 %"
```

```
## [1] "The better model is the alternative model."
```

Problem #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.

Results: Below I’ve shown the plots, as they were presented in the text. I added axis and title labels to the textbook’s version for explanatory purposes. Next, I added the analogous plots using ggplot2’s ‘qplot’ functionality.

I think the ggplot version is more informative in that it shows the difference based on the factor value of ESR.


Fibrinogen's Effect on ESR
Erythrocyte Sedimentation Rate
ESR > 20
ESR < 20
1.0
0.8
0.6
0.4
0.2
0.0
2.5
3.5
4.5
Fibrinogen Level in Blood
Globulin's Effect on ESR
Erythrocyte Sedimentation Rate
ESR > 20
ESR < 20
1.0
0.8
0.6
0.4
0.2
0.0
30
35
40
45
Globulin Level in Blood
Fibrinogen's Effect on ESR
ggplot2
Erythrocyte Sedimentation Rate
1.00
0.75
0.50
0.25
0.00
2
3
4
5
Fibrinogen Level in Blood
ESR
ESR < 20
ESR > 2
Globulin's Effect on ESR
ggplot2
Erythrocyte Sedimentation Rate
0.100
0.075
0.050
0.025
0.000
30
35
40
45
Globulin Level in Blood
ESR
ESR < 20
ESR >

Problem #5: continued

Results: Here I reproduced the summary and exponent values from the textbook. For readability, I put the Confidence Interval values in a dataframe.

There is a large confidence range, as the text mentions, due to the lack of data observations where ESR > 20. The summary of the logistic regression model shows that the treatment if ‘fibrinogen’ is statistically significant on ESR being greater than 20 at a level of 0.05 (0.0425).

```
##
## 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
```

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

```
##   Confidence Intervals  Tails
## 1             1.403209  2.5 %
## 2            54.515884 97.5 %
```

Problem #5: continued

Results: Below we've created a different model that takes both 'fibrinogen' and 'globulin' as the treatments. We see, from the summary, that globulin does not have a statistically significant impact on the ESR level.

We also can see that the model that includes the 'globulin' treatment has a p-value of 0.1716. That means that the fitted model with 'globulin' is not statistically different than the model without 'globulin' at a level of alpha = 0.05.

Next we use the anova() function to compare the previous model with this one. This function output further shows a chi square, on a single degree of freedom, that 'globulin' is not related with the ESR level.

```
##
## 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
```

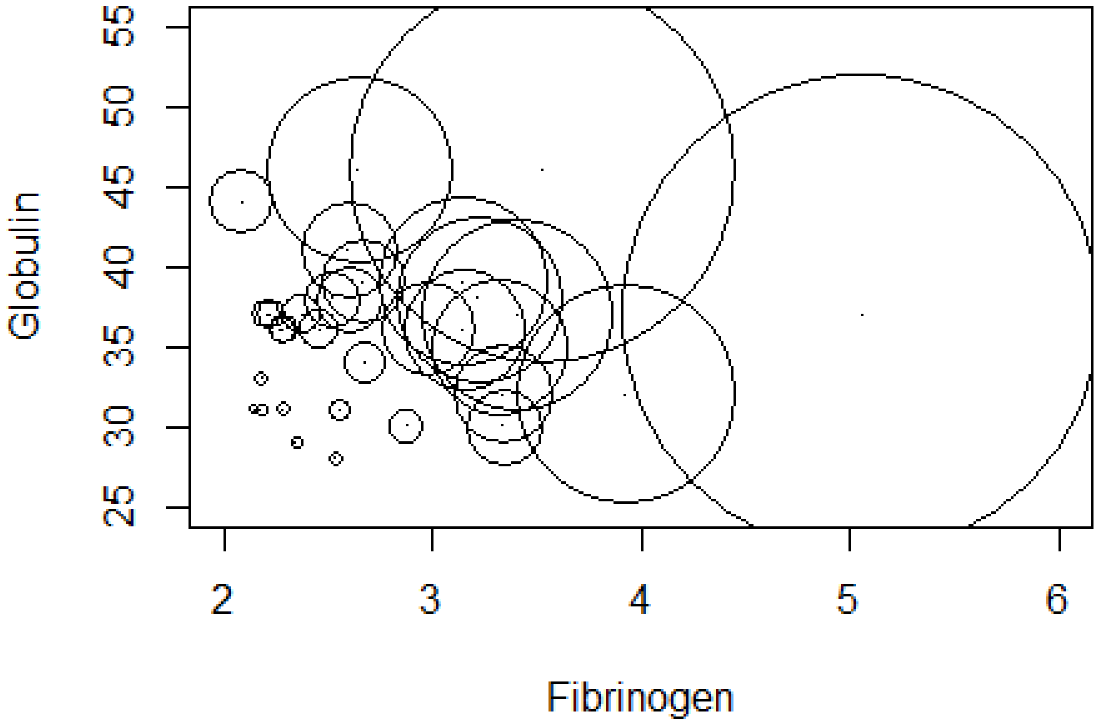
```
## Analysis of Deviance Table
##
## Model 1: ESR ~ fibrinogen
## Model 2: ESR ~ fibrinogen + globulin
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        30     24.840
## 2        29     22.971  1   1.8692   0.1716
```

Problem #5: continued

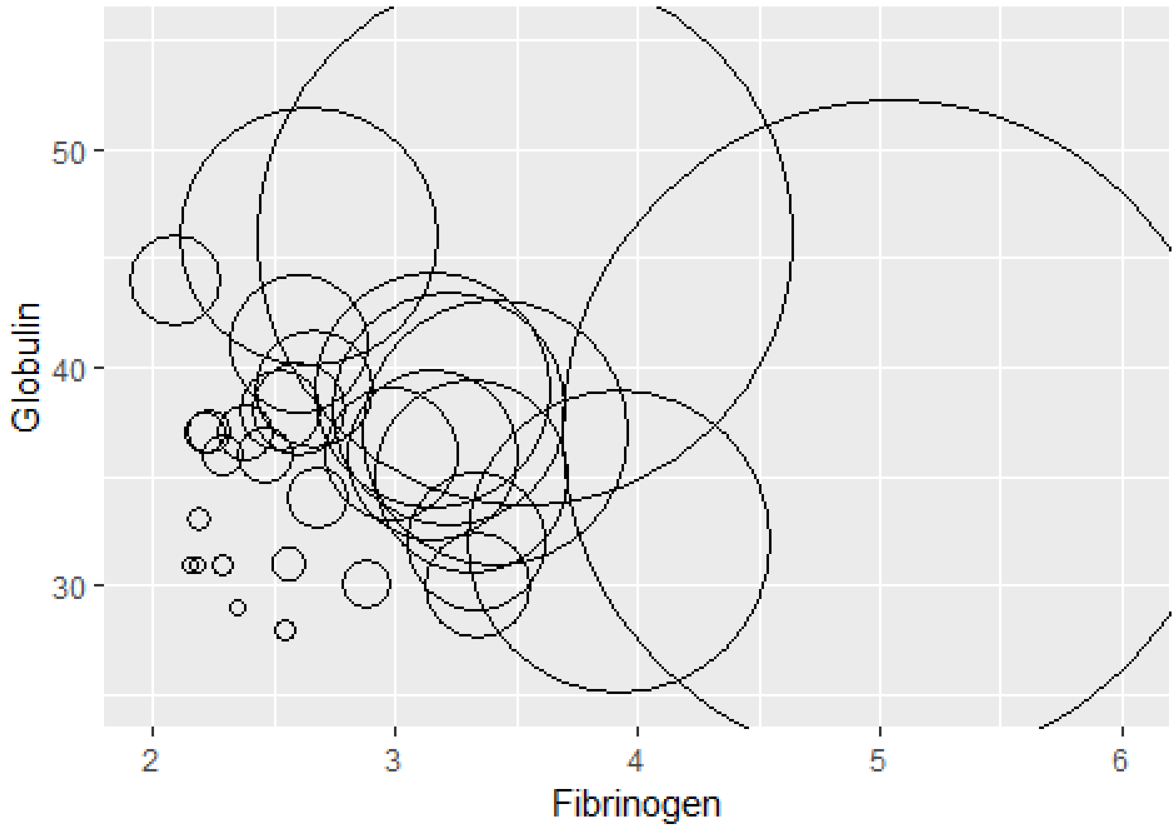
Results: The plot below shows the probability of ESR greater than 20 (larger circles) and how it is impacted by Fibrinogen and Globulin levels. We see an increasing probability of an ESR > 20 as Fibrinogen and Globulin increase.

Finally, I added a ggplot2 bubbleplot to compare the aesthetics of each type of plot. In this instance, the ggplot version is not any more informative than the base R plot. Both clearly show the increasing probability of ESR > 20 with an increase in Fibrinogen and Globulin.

Probability of ESR > 20
Based on Fibrinogen & Globulin



Probability of ESR > 20 Based on
Fibrinogen & Globulin - ggplot2


Globulin
Fibrinogen