Homework #3

Justin Robinette

September 11, 2018

No collaborators for any problem

Problem #1, Part A: Use the **bladdercancer** data from the 'HSAUR3' library to construct graphical and numerical summaries that will show the relationship between tumor size and the number of recurrent tumors. Discuss your discovery. *(Hint: mosaic plot may be the best way to assess this.)*

Results: I first plotted a stacked barplot (*Figure 1.1*) that shows the **number** of occurences of each number of recurrent tumors variable by tumor size. As we see from this illustration, at a number of 1 recurrent tumors, most are smaller than 3 cm. As we move to 2 and 3 recurrent tumors, the proportion of tumors that are less than 3 cm shrinks. By the 3 and 4 recurrent tumors, about half of the tumors were greater than 3 cm. There is an analogous base R plot included as well.

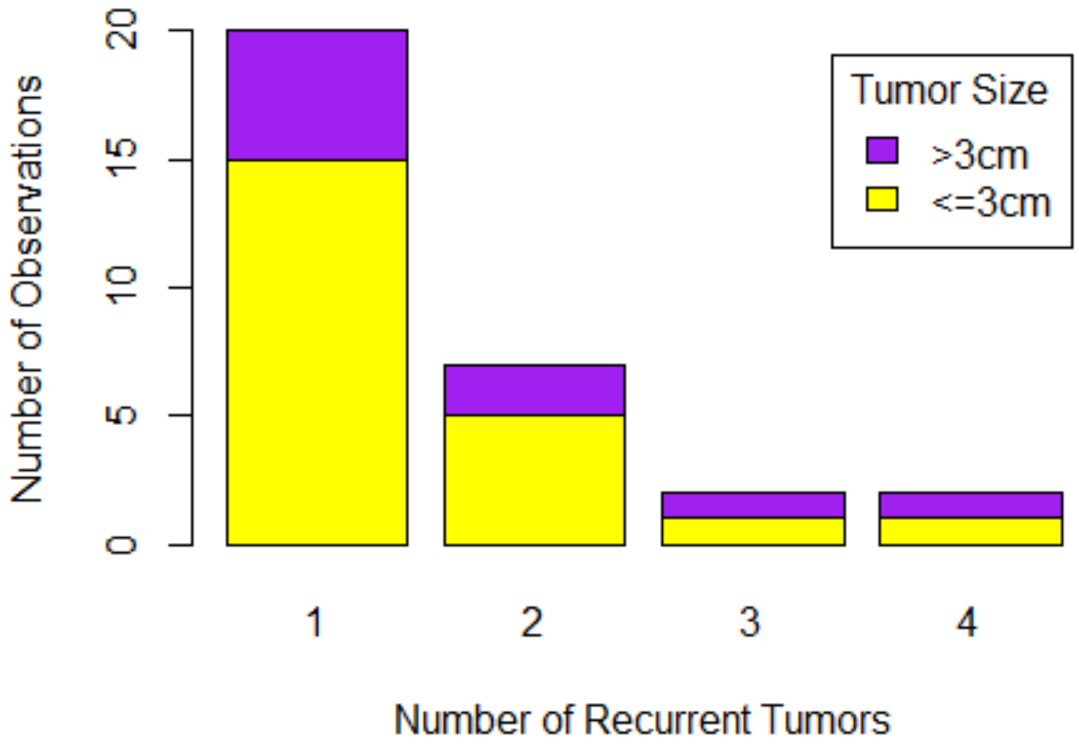
Next, I included a mosaic plot (*Figure 1.2*) that also shows how the **proportion** of tumors greater than 3 cm grows as the number of recurrent tumors increases. Again, there is a comparable base R version of this plot included for review. We again see, from these plots, how the proportion of tumors greater than 3 cm increases as the number of recurrent tumors increases.

Lastly, I've included a table depicting the information in the mosaic plot Figure 1.2 (*Figure 1.3*). This table shows a breakdown by the *Number of Recurrent Tumors*, *Tumor Size* and *Percentage of Total*. The percentage column corresponds to the percentage of observations containing the corresponding 'RecurrentTumors' based on 'TumorSize'.

Number of Recurrent Tumors
by Tumor Size

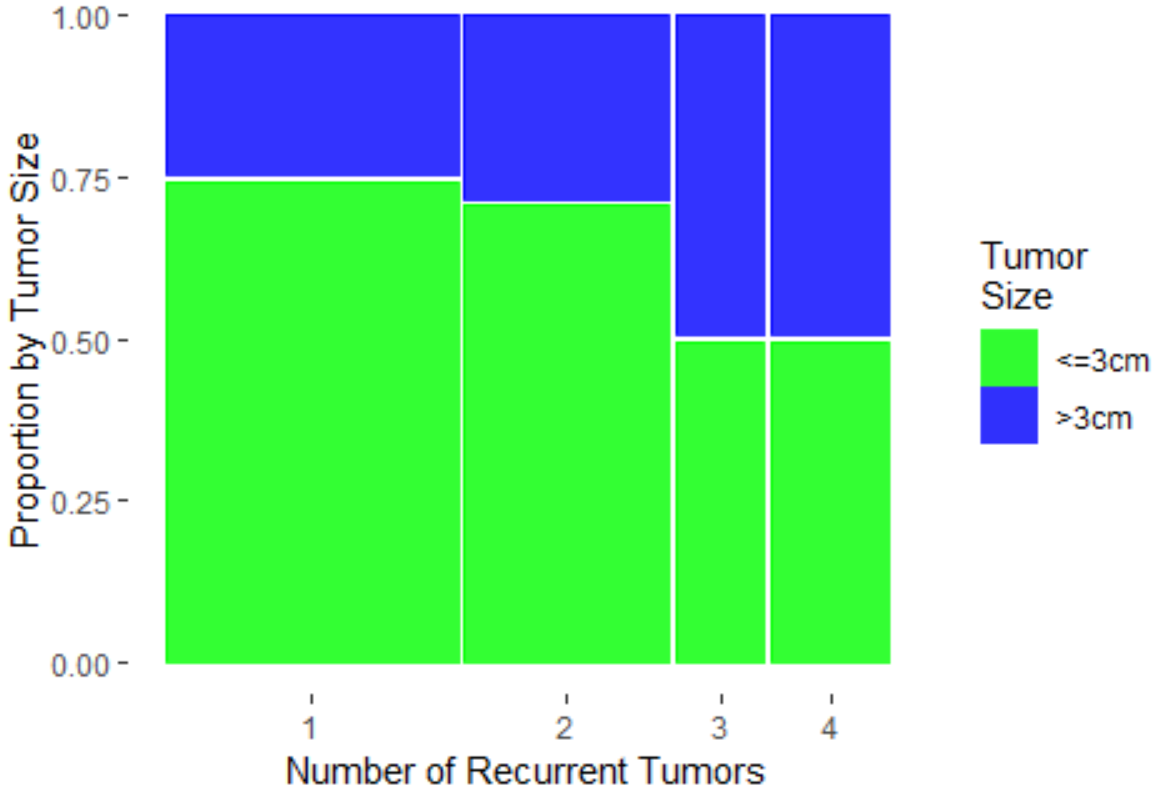
Figure 1.1

Number of Recurrent Tumors by Tumor Size - Base R



Proportion of Recurrent Tumors by Tumor Size

Figure 1.2



Proportion of Recurrent Tumors by Tumor Size - Base R

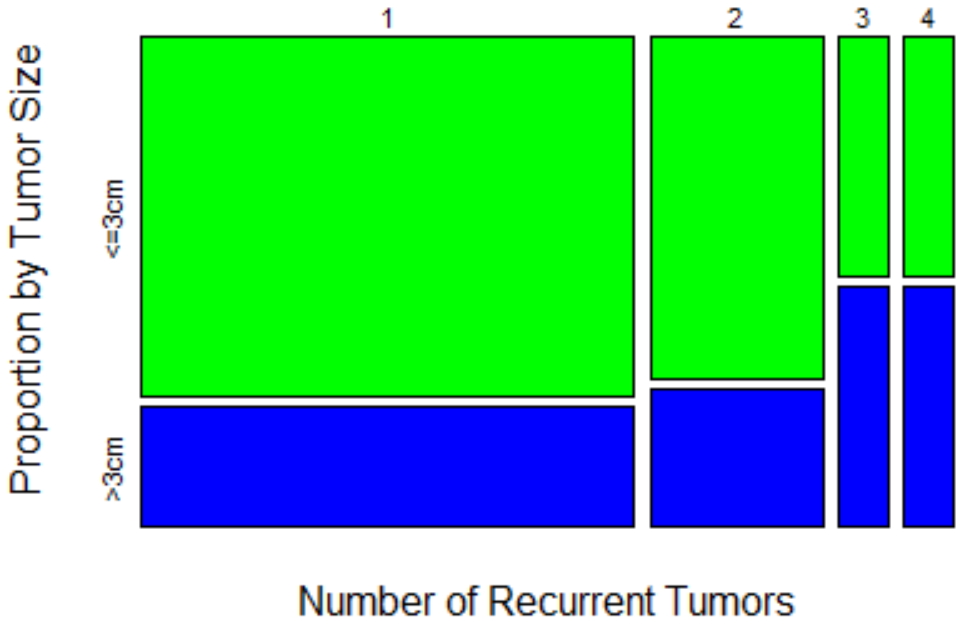


Figure 1.3

Recurrent Tumors	TumorSize	Percentage
1	<=3cm	75.00
1	>3cm	25.00
2	<=3cm	71.43
2	>3cm	28.57
3	<=3cm	50.00
3	>3cm	50.00
4	<=3cm	50.00
4	>3cm	50.00

Problem #1, Part B: Use the **bladdercancer** data from 'HSAUR3' library to build a Poisson regression that estimates the effect of size of tumor on the number of recurrent tumors. Discuss your results.

Results: First, we look at a summary of a Poisson regression that has 'number' of recurrent tumors as the response with 'time' and 'tumorsize' as the treatments. This model will be used as a comparison to a model with just 'number' and 'tumorsize' to help us visualize the effect of 'tumorsize' on the 'number'. We see that, in this model, none of the variables are statistically significant. We can see that 'Tumorsize > 3cm' has a positive relationship with the 'number'. Our AIC for this model is 88.568.

```
##
## Call:
## glm(formula = number ~ time + tumorsize, family = poisson(),
##     data = bladdercancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8183  -0.4753  -0.2923   0.3319   1.5446
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.14568    0.34766   0.419    0.675
## time           0.01478    0.01883   0.785    0.433
## tumorsize>3cm  0.20511    0.30620   0.670    0.503
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 12.800  on 30  degrees of freedom
## Residual deviance: 11.757  on 28  degrees of freedom
## AIC: 88.568
##
## Number of Fisher Scoring iterations: 4
```

Next, we look at a summary of a similar Poisson regression model that, this time, has 'number' again as the dependent variable and only 'tumorsize' as the treatment, **as the question requests**. With this model, we see an intercept p-value of 0.034, which is significant at an *alpha* = 0.05. Additionally, the AIC is lower which indicates a superior model. Despite the improvement in the model when 'time' is removed as an independent variable, we still see that 'tumorsize' does not impact 'number' in a statistically significant manner.

```
##
## Call:
## glm(formula = number ~ tumorsize, family = poisson(), data = bladdercancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6363  -0.3996  -0.3996   0.4277   1.7326
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.3747     0.1768   2.120    0.034 *
## tumorsize>3cm   0.2007     0.3062   0.655    0.512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 12.80  on 30  degrees of freedom
## Residual deviance: 12.38  on 29  degrees of freedom
## AIC: 87.191
##
## Number of Fisher Scoring iterations: 4
```

Finally, lets look at a comparison of the AIC values to confirm that the 2nd model, which has 'number' as response and 'tumorsize' as treatment, is superior to the 1st model which includes 'time' as a treatment variable. Figure 1.3 shows that *Model #2* is slightly better, due to a small variation in model AIC values, *despite 'tumorsize' not being a significant treatment of the response 'number'.*

Figure 1.3: Comparison of AIC Values by Model

AIC of Model #1	AIC of Model #2
88.56765	**87.19128**

Problem #2, Part A: The following data is the number of new AIDS cases in Belgium between the years of 1981 - 1993. Let *t* denote time. y <- c(12, 14, 33, 50, 67, 74, 123, 141, 165, 204, 253, 246, 240) t <- 1:13

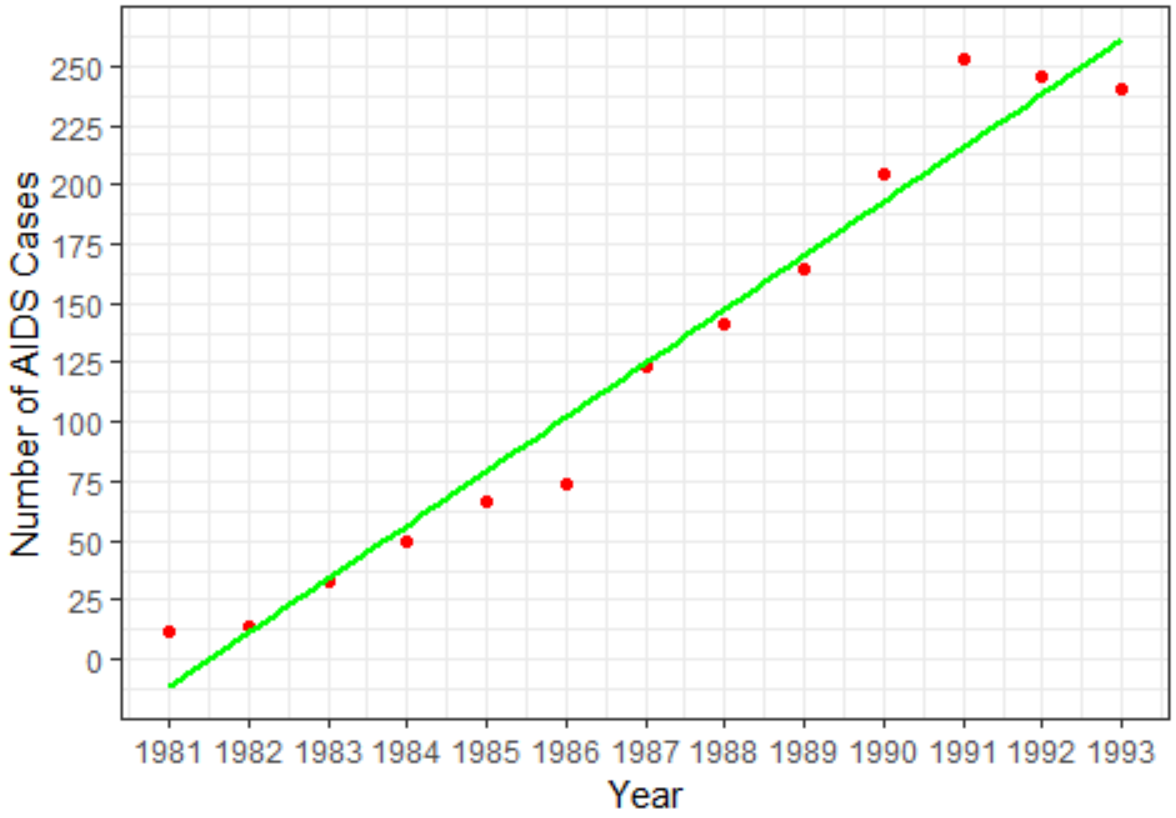
Plot the relationship between number of AIDS cases against time. Comment on the plot.

Results: First, I made the 'time' variable more specific by changing it from 1,2,…13 to specify the year (1981, 1982….1993). Next I included both a ggplot and base R plot showing the relationship between the number of AIDS cases and the year.

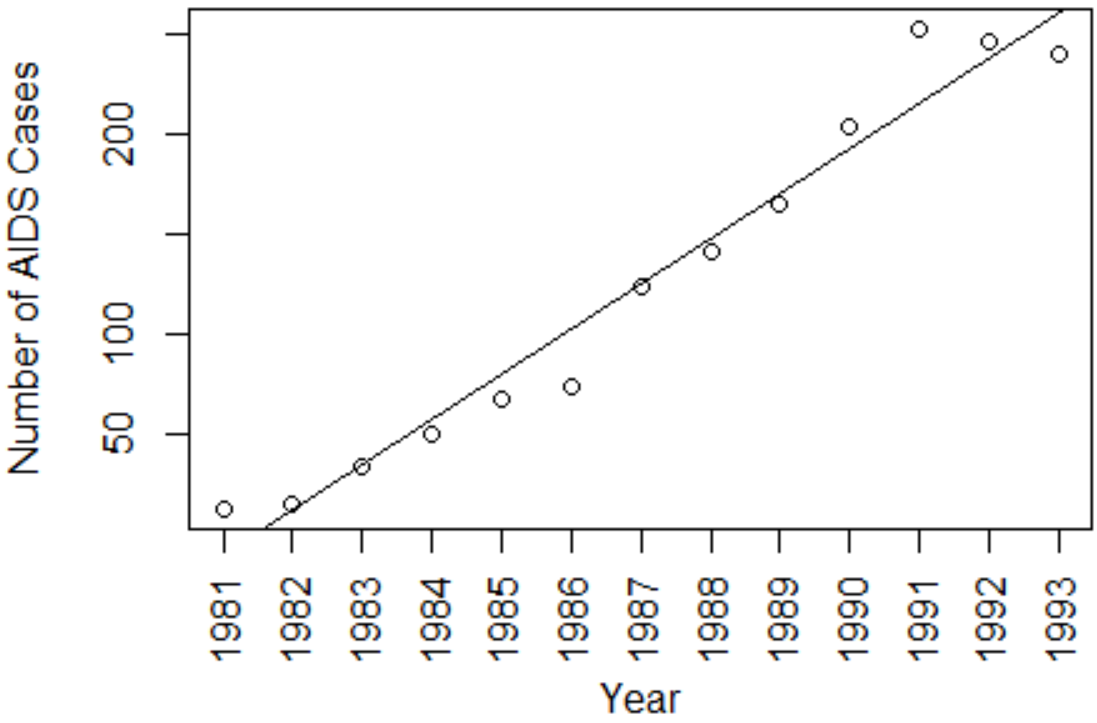
As we can see, there appears to be a strong relationship between the number of cases and the year. As the year gets bigger, the number of cases grows. I've added a regression line to the plots to help show the relationship between the two variables.

AIDS Cases in Belgium from 1981-1993

Figure 2.1



AIDS Cases in Belgium from 1981-1993
Base R



Problem #2, Part B: Fit a Poisson regression model $\boldsymbol{log(\mu_i) = \beta_0 + \beta_1 t_i}$. Comment on the model parameters and residuals (deviance) vs Fitted plot.

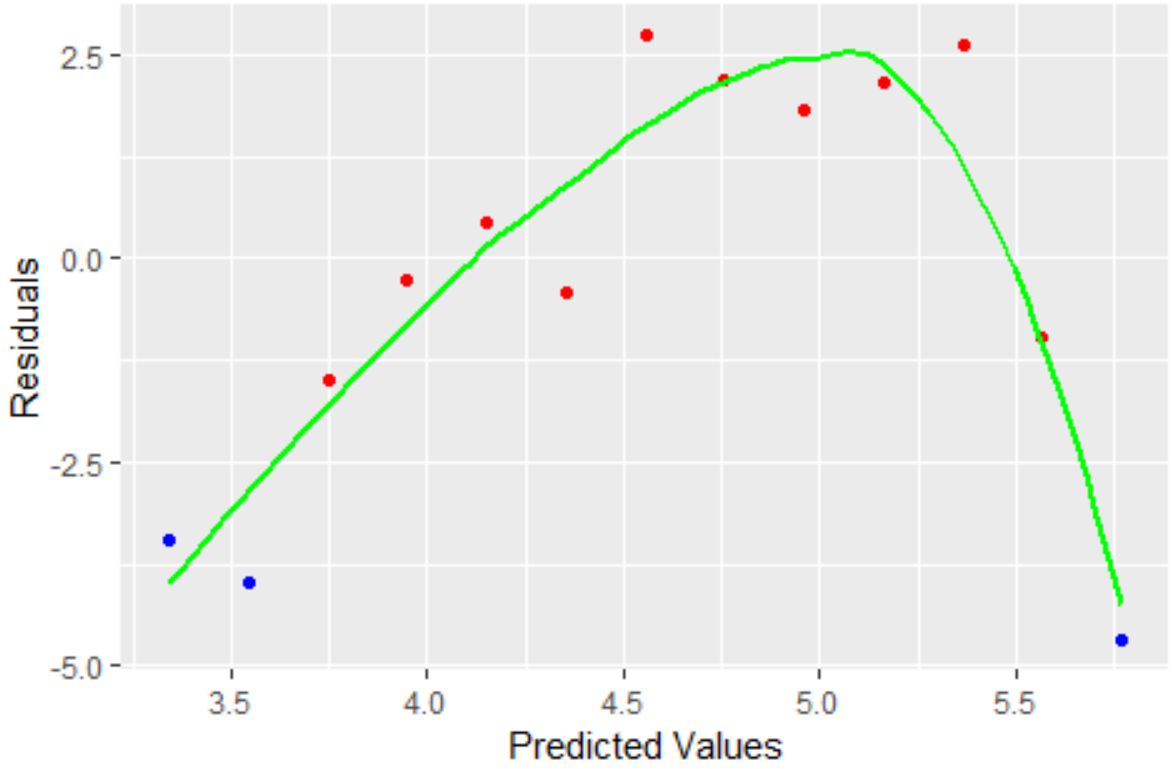
Results: Here, I fit a Poisson Regression model with 'Number' as the response variable and 'Time' as the treatment. We see, from the summary, that time has a significant positive influence on the number of cases.

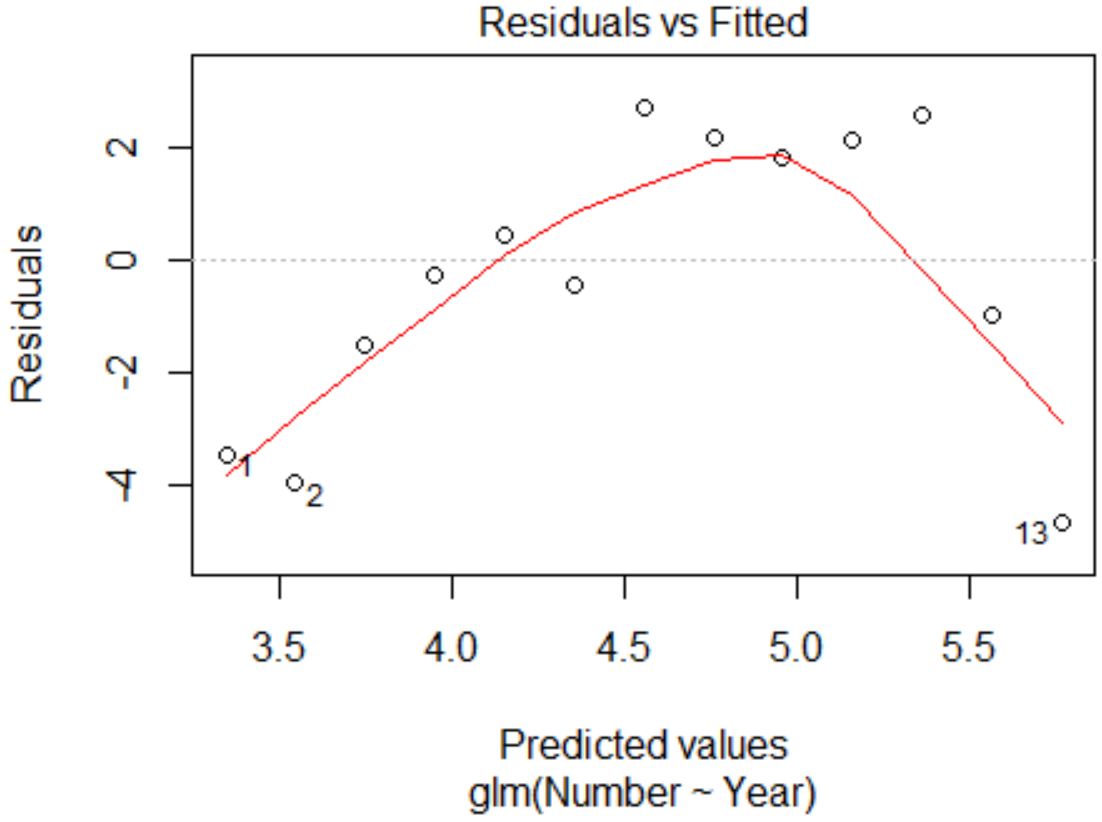
The plot below (Figure 2.2) shows that the data is overdispersed. The 3 years with the largest absolute standardized residuals correspond to 1981, 1982, and 1993 (time = 1, 2, and 13). These points are labeled with **blue** dots on Figure 2.2. An analogous *Base R* plot is included for comparison. In the *Base R* plot, the largest absolute residuals are identified by the numbered points. The numbers correspond to the years of 1981, 1982, and 1993. These are the 1st, 2nd, and 13th years in our data set.

```
##
## Call:
## glm(formula = Number ~ Year, family = poisson, data = aids.dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6784  -1.5013  -0.2636   2.1760   2.7306
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.971e+02  1.546e+01  -25.68   <2e-16 ***
## Year         2.021e-01  7.771e-03   26.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 872.206  on 12  degrees of freedom
## Residual deviance:  80.686  on 11  degrees of freedom
## AIC: 166.37
##
## Number of Fisher Scoring iterations: 4
```

Residuals vs. Fitted
Poisson Model

Figure 2.2





Problem #2, Part C: Now add a quadratic term in time $\boldsymbol{log(\mu_i) = \beta_0 + \beta_1 t_i + \beta_2 t_i^2}$ and fit the model. Comment on the model parameters and assess the residual plots.

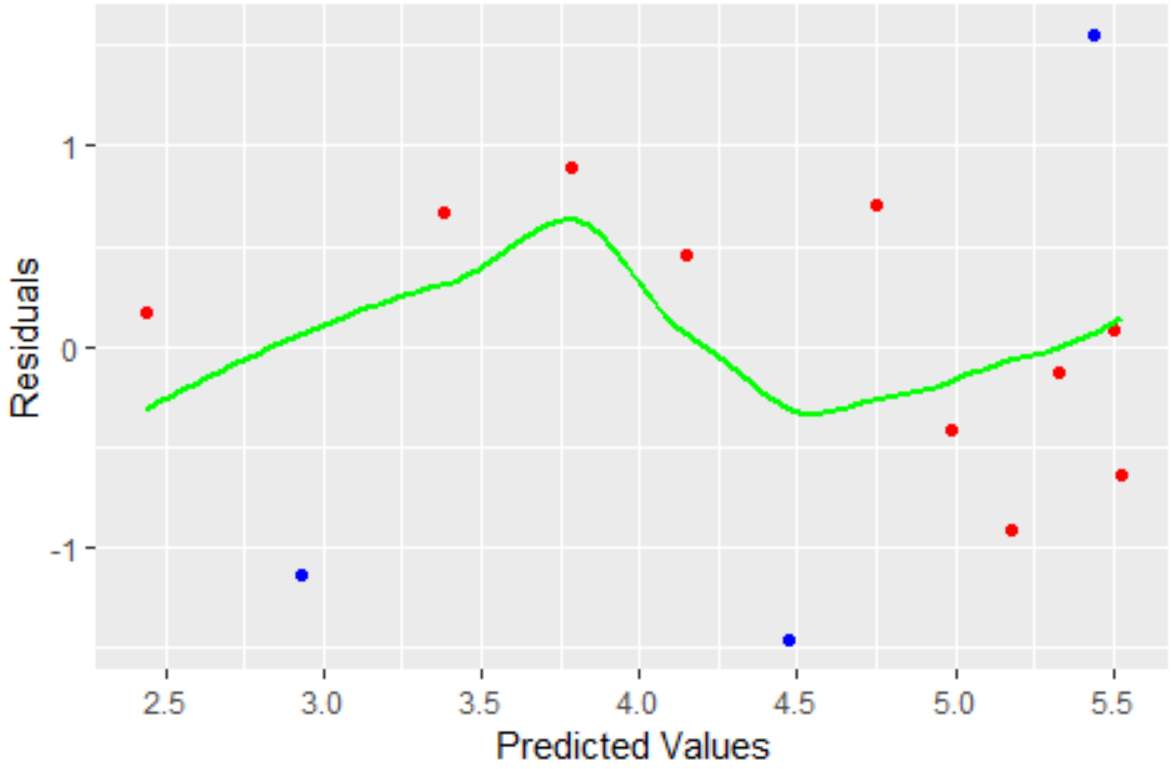
Results: Here I've added a quadratic term, to the Poisson model discussed in *Part B*, and fit the model. I've plotted the Residuals vs. Fitted again and highlighted the 3 largest absolute standardized residuals in **blue** on Figure 2.3. These correspond to the years 1982, 1986, and 1991.

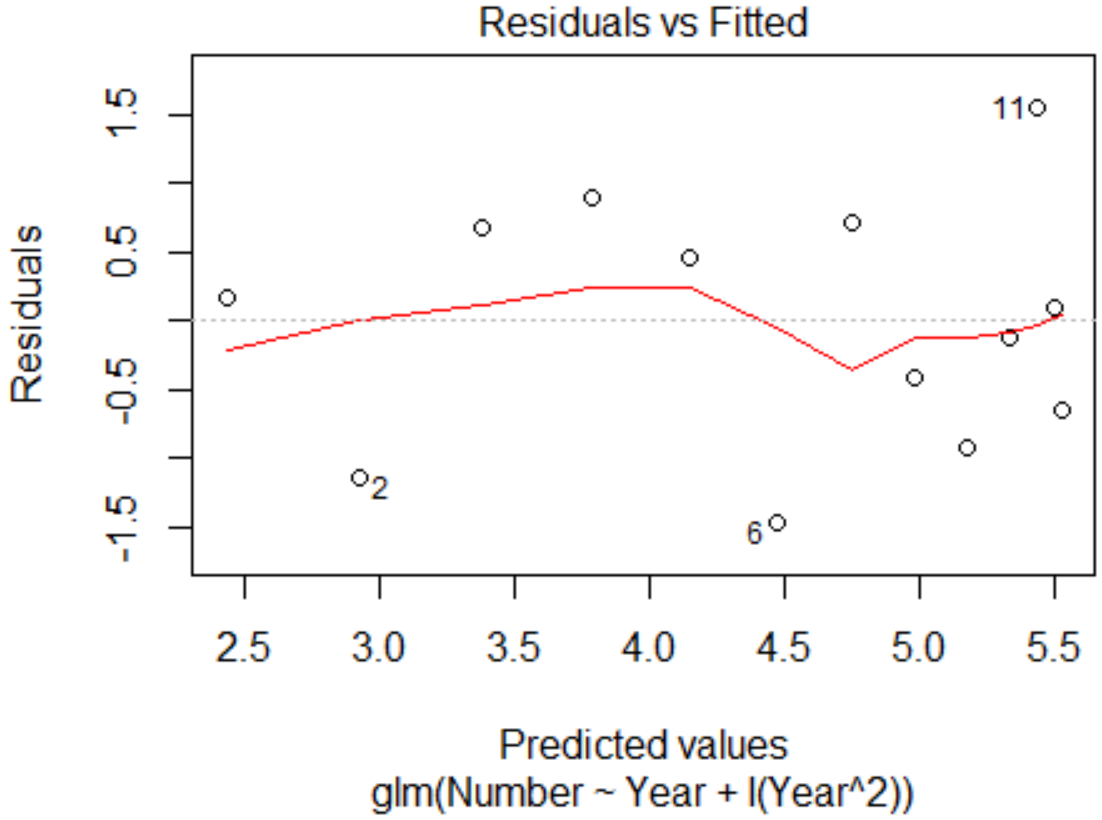
This plot shows that the Poisson model that includes the quadratic is not overdispersed as compared to the plot in Figure 2.2. I've included a Base R plot here for comparison.

```
##
## Call:
## glm(formula = Number ~ Year + I(Year^2), family = poisson(),
##     data = aids.dat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.45903  -0.64491   0.08927   0.67117   1.54596
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.478e+04  1.051e+04  -8.066 7.29e-16 ***
## Year         8.509e+01  1.057e+01   8.048 8.45e-16 ***
## I(Year^2)   -2.135e-02  2.659e-03  -8.029 9.82e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 872.2058  on 12  degrees of freedom
## Residual deviance:   9.2402  on 10  degrees of freedom
## AIC: 96.924
##
## Number of Fisher Scoring iterations: 4
```

Residuals vs. Fitted
Poisson Model with Quadratic

Figure 2.3





Problem #2, Part D: Compare the two models using AIC. Which model is better?

Results: As we can see below, the model that includes a quadratic term, ($\boldsymbol{time^2}$), is superior due to its lower AIC. With AIC, as the difference between AIC values increases, we get stronger evidence for one model over another. In this case, the Poisson Quadratic model has an AIC that is a little more than half the size of the Poisson Regression model's AIC. This indicates that we prefer the model with the quadratic term over the model without the quadratic term.

Figure 2.4: Comparison of Model AIC Values

Poisson Model AIC	Poisson Quadratic Model AIC
166.3698	**96.92358**

Problem #2, Part E: Use *anova*() function to perform $\boldsymbol{x^2}$ test for model selection. Did adding the quadratic term improve the model?

Results: Here we used the *anova*() function to perform $\boldsymbol{x^2}$ testing for selecting the superior model. In this example, our hypotheses are as follows:

$\boldsymbol{H_0}$: ***The models are equally effective in explaining the number of AIDS cases***

$\boldsymbol{H_a}$: ***The models are not equally effective in explaining the number of AIDS cases***

Based on the ANOVA table below, we can state that we have evidence to reject the *Null hypothesis*, due to the p-value of less that 0.001 and the difference in Residual Deviation.

```
## Analysis of Deviance Table
##
## Model 1: Number ~ Year
## Model 2: Number ~ Year + I(Year^2)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        11     80.686
## 2        10      9.240  1   71.446 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem #3, Part A: Load **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 4 dimensional dataset with 10,000 observations. You had developed a logistic regression model on HW #2. Now consider the following two models:

Model1 -> Default = Student + Balance

Model2 -> Default = Balance

With the whole data compare the two models (Use AIC and/or error rate)

Results: First, I obtained the AIC values from both *Model1* and *Model2*. As we can see in Figure 3.1, the AIC for *Model1* is slightly lower than the AIC for *Model2*. This indicates that *Model1* is the superior model from these two models.

To confirm, I calculated the actual error rate of the predictions compared to the Default data set. Confusion matrices were done to show the accuracy differences for *Model1* and *Model2*. Figure 3.2 summarizes this accuracy comparison.

Finally, Figure 3.3 shows a comparison of the Mean Square Error for these two models. Again, we see that *Model1* is superior to *Model2*.

For all three measures, *Model1*, which includes both 'Student' and 'Balance' as treatment variables, is slightly better at explaining the data than *Model2*, which only includes 'Balance' as a treatment variable. *This shows that the inclusion of 'Student' improves the model's predictability.*

Figure 3.1: Comparison of AIC Models

Model1 AIC	Model2 AIC
1577.682	1600.452

Figure 3.2: Comparison of Error Rates

Model1 Error Rate (%)	Model2 Error Rate (%)
2.67	2.75

Figure 3.3: Comparison of Mean Square Error

Model1 MSE	Model2 MSE
0.0213018	0.0217058

Problem #3, Part B: Use validation set approach and choose the best model. Be aware that we have few people who defaulted in the data.

Results: Figure 3.4 shows the comparison between AIC for *Model1* and *Model2* using the validation set approach. In this approach, we randomly split the original data set into 'train' and 'test' subsets. In doing so, the 'test' data does not bias the model since only the training data is used in fitting the model. 75% of the original 'Default' data set are included in the 'train' data set and the remaining 25% in the 'test' data set.

Figure 3.5 shows the comparison between Mean Square Error(MSE) for the two models.

With this approach, similar to *Part A*, *Model1* measures slightly better in both AIC and MSE. From this we deduce that *Model1*, which includes the 'balance' and 'student' variables as independent variables, is slightly superior to *Model2* which only has one independent variable - 'balance'. Again, we receive evidence that the inclusion of 'student' improves the model accuracy.

Figure 3.4: Comparison of AIC Models

Model1 AIC	Model2 AIC
1168.229	**1181.365**

Figure 3.5: Comparison of Mean Square Error (Validation Set Approach)

Model1 MSE	Model2 MSE
0.0219947	**0.0227541**

Problem #3, Part C: Use LOOCV approach and choose the best model.

Results: For this problem, I created a *for loop* that iterates through the entire 'Default' data set leaving out one observation as the 'test' data set in each iteration. The loop calculates the MSE at each iteration. Figure 3.6 shows the MSE for both models for comparison.

Again we can see that *Model1* has a lower MSE, using this approach, than that of *Model2*. From this, we conclude that *Model1* is better than *Model2* and the inclusion of the 'student' variable as a treatment improves the model.

Figure 3.6: Comparison of Mean Square Error (LOOCV)

Model1 MSE	Model2 MSE
0.0213727	**0.0321647**

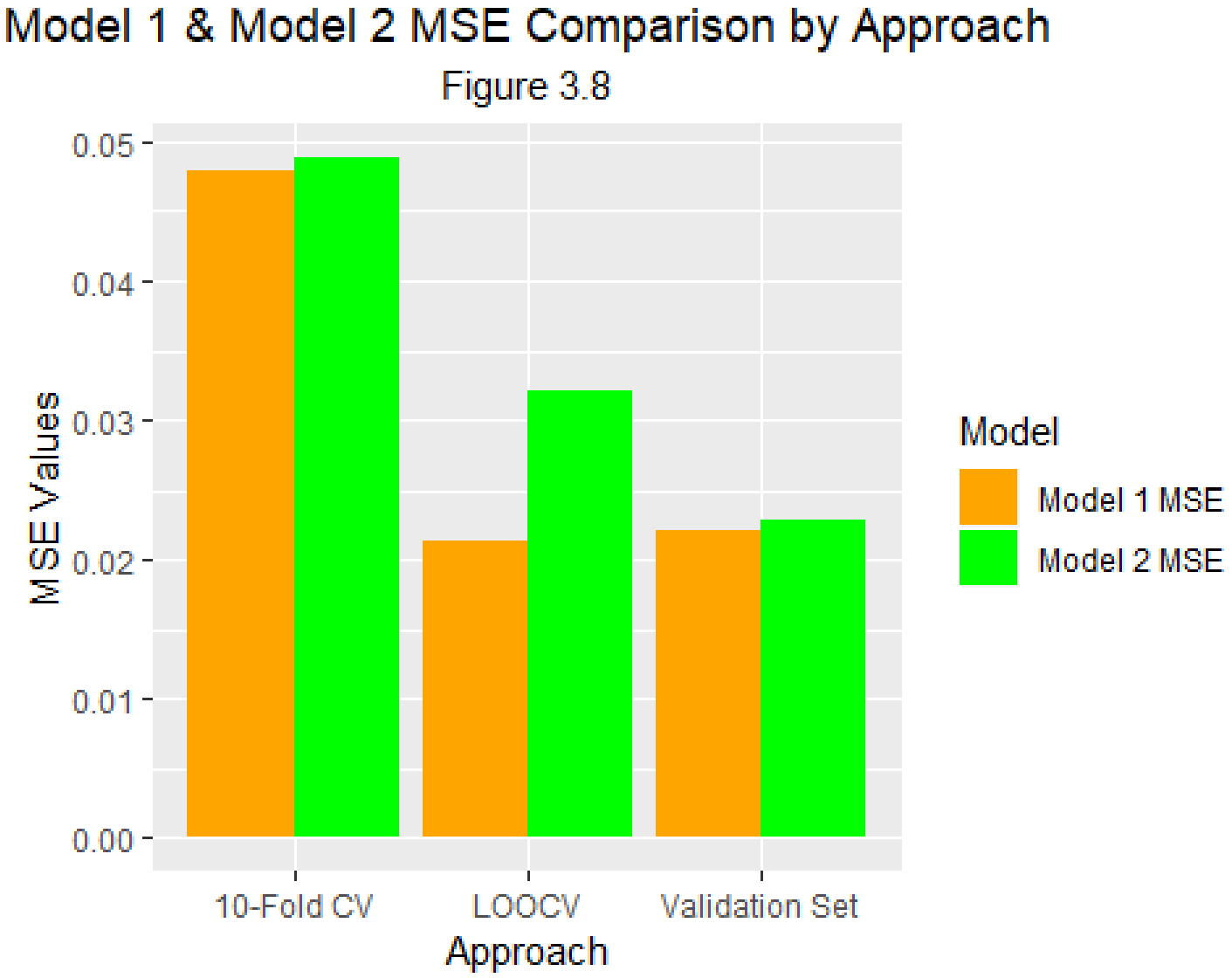
Problem #3, Part D: Use 10-fold cross-validation approach and choose the best model.

Results: Again, *Model1* emerges as the best model due to having a slightly smaller Mean Standard Error than *Model2*. Thus we can conclude that the inclusion of the treatment variable 'student' in the model improves it's accuracy compared to *Model2* which excludes this treatment variable.

Figure 3.7: Comparison of Mean Square Error (10-fold CV)

Model1 MSE	Model2 MSE
0.048	**0.0488**

Problem #3 Summary: Figure 3.8 provides a visual depiction of these MSE values from each approach. As we can see, *Model1* has a lower MSE regardless of the comparison approach taken. *There is not an analogous Base R plot for Figure 3.8 because Problem 3 does not require any plots.*



Problem #3 Summary, cont'd: Here I have the misclassification rate summaries from Part B, C, and D of Exercise 3 in Figures 3.10, 3.11, and 3.12 respectively. We can conclude that, regardless of the modeling approach, *Model1* is better at predicting 'default' than *Model2*. Therefore, the addition of the 'student' independent variable to the model does improve its performance.

Figure 3.10: Comparison of Error Rate(Validation Set)

Model1 Error Rate	Model2 Error Rate
2.8	2.92

Figure 3.11: Comparison of Error Rate(LOOCV)

Model1 Error Rate	Model2 Error Rate
2.14	3.22

Figure 3.12: Comparison of Error Rate(10-fold CV)

Model1 Error Rate	Model2 Error Rate
4.8	4.88

Problem #4: In the **ISLR** library, load the **Smarket** dataset. This contains Daily percentage returns for the S&P 500 stock index between 2001 and 2005. There are 1250 observations and 9 variables. The variable of interest is Direction which is a factor of levels Down and Up indicating whether the market had a positive or negative return on a given day. Since the goal is to predict the direction of the stock market in the future, here it would make sense to use the data from years 2001 - 2004 as training and 2005 as validation. According to this, create a training set and testing set. Perform logistic regression and assess the error rate.

Results: First I split the data into *train* and *test* data sets. I then created two logistic regression models (*model.1* and *model.2*) using the *train* data set. For **model.1**, I used the 5 'Lag' variables and the 'Volume' variable as my treatment variables. For **model.2**, I removed 'Volume' as a treatment variable.

Using the *validation set approach*, we see that **model.2** has a smaller Mean Square Error than **model.1**, the model that includes the 'Volume' treatment variable. This means that, using this method, **model.2** is better at predicting the outcomes in the *test* data set than **model.1**.

Figure 4.1: Comparison of Mean Square Error(Validation Set Approach)

Model.1 MSE	Model.2 MSE
0.2507619	0.2483559

Results: To further compare the 2 models, I have also included *Figure 4.2* which compares the AIC values of the models. Here we see that *model.2* has a smaller AIC indicating that it is superior. Because the difference is so small, I also want to check the actual error rate of each model before determining which is the better model.

Figure 4.3 shows the actual error rate for both *model.1* and *model.2*. As we can see, *model.2* was more accurate in predicting the market direction in the 'test' data set. Therefore, we can say that the model improved by removing the 'Volume' treatment variable.

Figure 4.2: Comparison of AIC

Model.1 AIC	Model.2 AIC
1395.105	**1393.341**

Figure 4.3: Comparison of Error Rate

Model.1 Error Rate(%)	Model.2 Error Rate(%)
51.98413	**41.26984**