Homework 3

Snigdha Peddi

Exercises

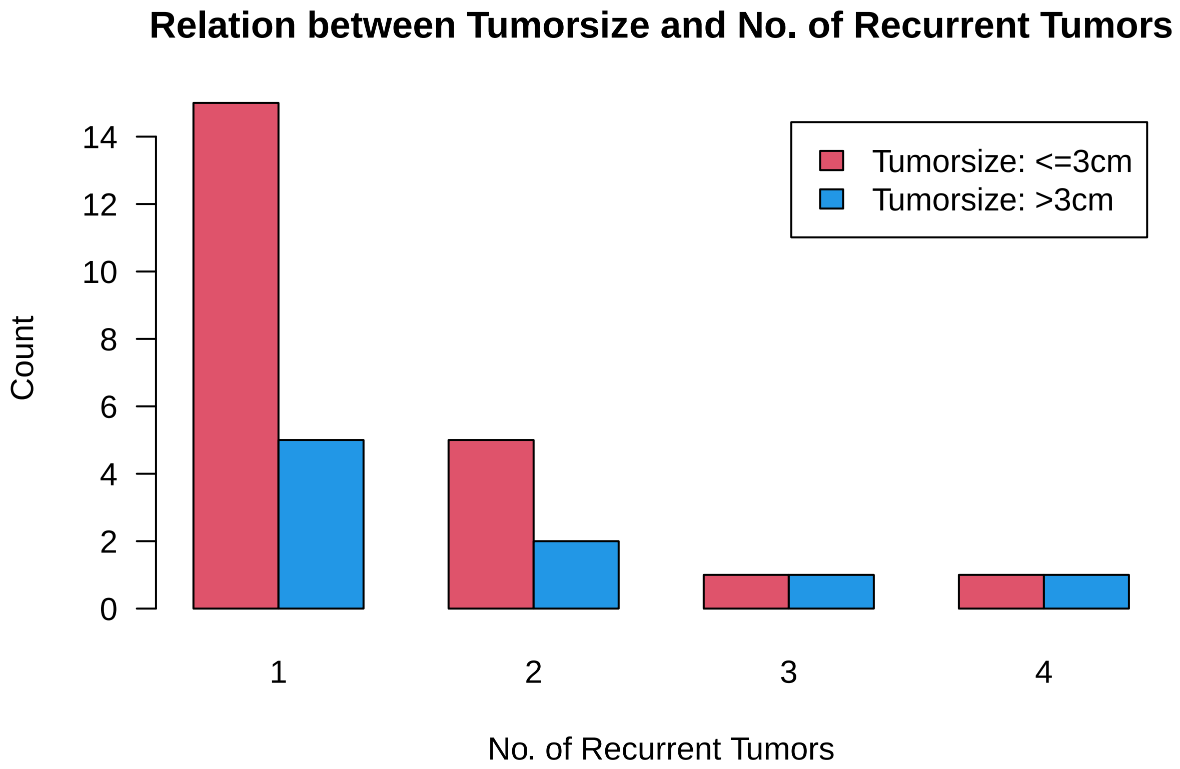
Question 1. (Ex. 7.3 pg 147 in HSAUR, modified for clarity) Use the **bladdercancer** data from the **HSAUR3** library to answer the following questions.

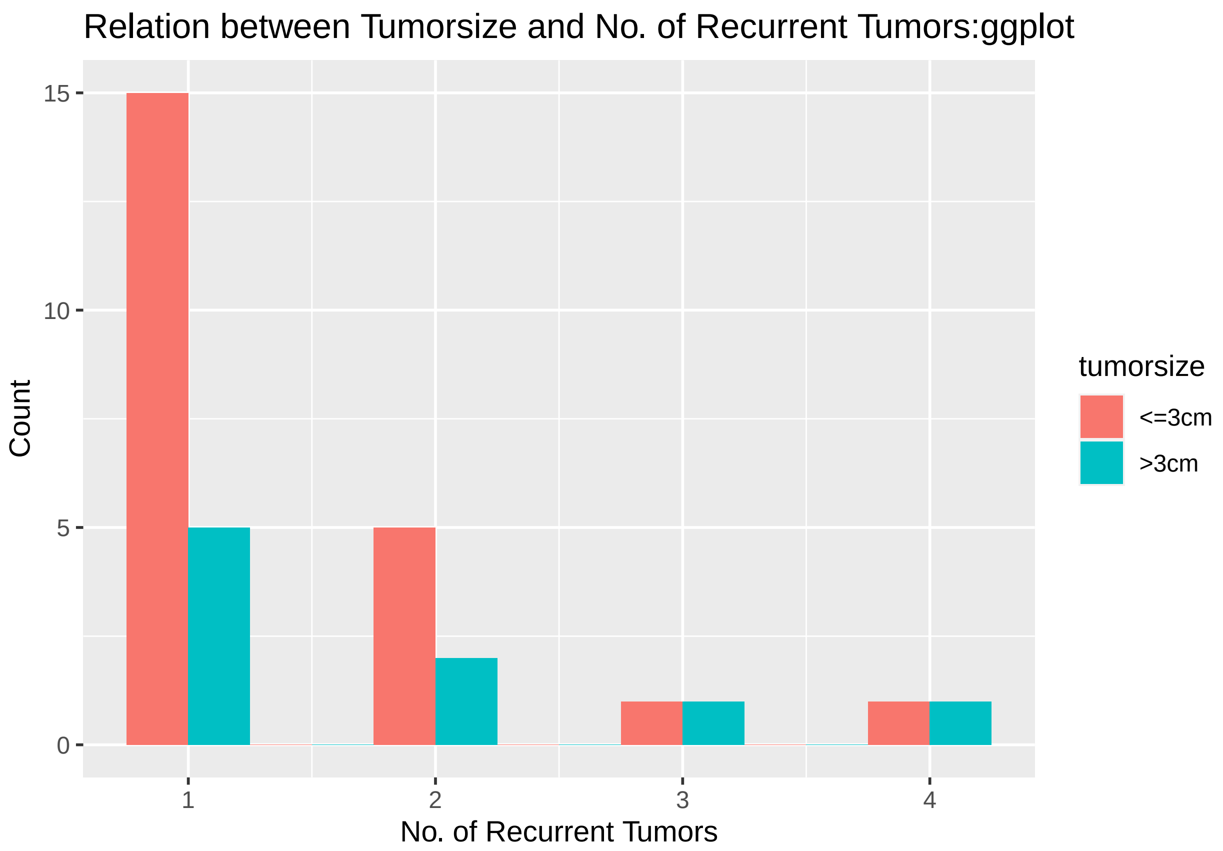
a) Construct graphical and/or numerical summaries to identify a relationship between tumor size and the number of recurrent tumors. Discuss your discovery. (For example, a mosaic plot or contingency table is a good starting point. Otherwise, there are other ways to explore this data.)

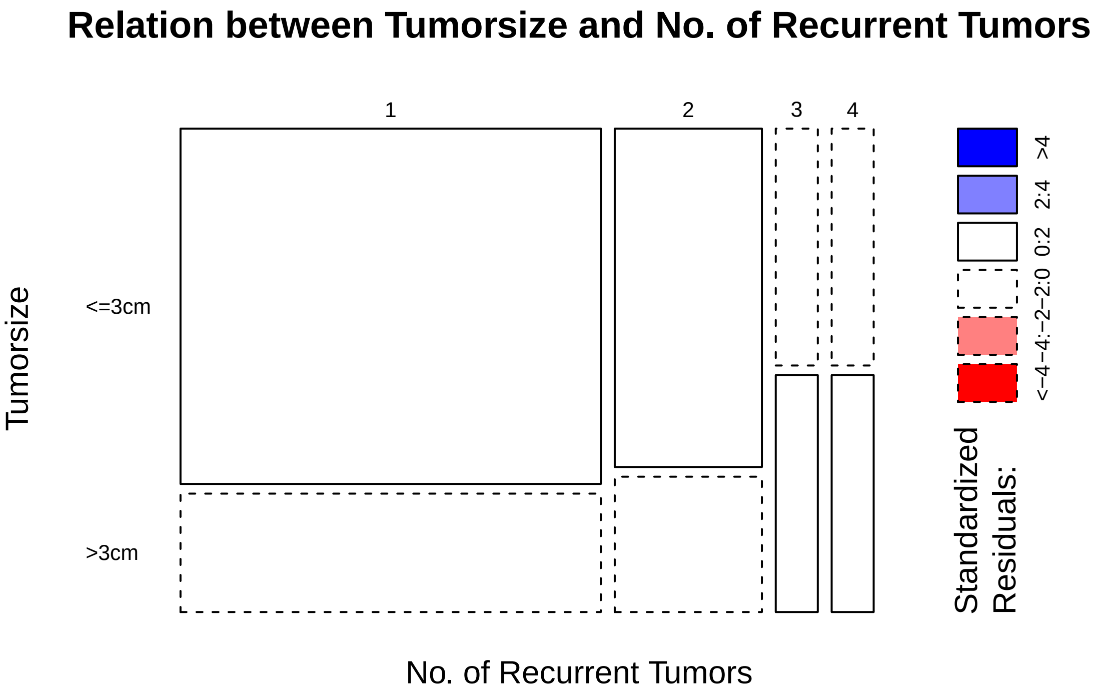
Answer 1.a:The Contingency table shows that the frequency of tumors of size <=3cm is more than the >3cm. Both the Barplot and Mosaic plot shows that the tumors of size <=3cm were more recurrent than size >3cm for 1 and 2 tumors and equal for 3 and 4 tumors. And the frequency of 1 0r 2 tumors is more than 3 or 4 tumors.There is no significant pattern describing the relation between the number of tumors and the size of the tumor.

```
## Contingency Table:

##
##          1  2  3  4
##   <=3cm 15  5  1  1
##   >3cm   5  2  1  1
```


Relation between Tumorsize and No. of Recurrent Tumors
Count
0
2
4
6
8
10
12
14
1
2
3
4
No. of Recurrent Tumors
Tumorsize: <=3cm
Tumorsize: >3cm


Relation between Tumorsize and No. of Recurrent Tumors:ggplot
Count
15
10
5
0
1
2
3
4
No. of Recurrent Tumors
tumorsize
<=3cm
>3cm



b) Assume a Poisson model describes the relationship found in part a). Build a Poisson regression that estimates the effect of tumor size on the number of recurrent tumors. Does the result of this analysis support your discovery in part a)?

Answer 1.b: A Poisson regression model is fit.

A poisson regression model between number of recurrent tumors and the tumor size indicate that the variable tubmor size is not significant with a high P value. Residual error is lower than the degrees of freedom indicating an under-dispersion. To correct the standard error a quasipoisson regression is fit.

The resultant model has a significant intercept but is still not significant with a high P value indicating that the tumor size is not correlated with the number of recurrent tumors.

```
##
## 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
```

```
##
## Call:
## glm(formula = number ~ tumorsize, family = quasipoisson(), data = bladdercancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6363  -0.3996  -0.3996   0.4277   1.7326
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.3747     0.1255   2.985  0.00571 **
## tumorsize>3cm   0.2007     0.2174   0.923  0.36368
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.5043225)
##
##     Null deviance: 12.80  on 30  degrees of freedom
## Residual deviance: 12.38  on 29  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

Another model is built with both the features time and tumor size.

The P value indicate that both the variables are not significant at 95% confidence interval. Also, the AIC is higher than the previous model with just the Tumor size variable indicating that time variable is not improving the model performance.

```
##
## 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
```

One more model is fit with both the features time and tumor size and a interaction term.

The P value of the higher order term(interaction term) indicate that there is no significant effect of the interaction term. Also, the AIC is higher than both the previous models indicating that the interaction term is not improving the model performance.

```
##
## Call:
## glm(formula = number ~ time + tumorsize + tumorsize * time, family = poisson(link = log),
##     data = bladdercancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6943  -0.5581  -0.2413   0.2932   1.4644
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.03957    0.43088   0.092    0.927
## time                 0.02138    0.02418   0.884    0.377
## tumorsize>3cm        0.46717    0.66713   0.700    0.484
## time:tumorsize>3cm -0.01676    0.03821  -0.439    0.661
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 12.800  on 30  degrees of freedom
## Residual deviance: 11.566  on 27  degrees of freedom
## AIC: 90.377
##
## Number of Fisher Scoring iterations: 4
```

The Chi square test indicates that the models with all the variable and with interaction term are not significant.

```
## Analysis of Deviance Table
##
## Model 1: number ~ time + tumorsize + tumorsize * time
## Model 2: number ~ time + tumorsize
## Model 3: number ~ tumorsize
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        27     11.566
## 2        28     11.757 -1 -0.19095   0.6621
## 3        29     12.380 -1 -0.62363   0.4297
```

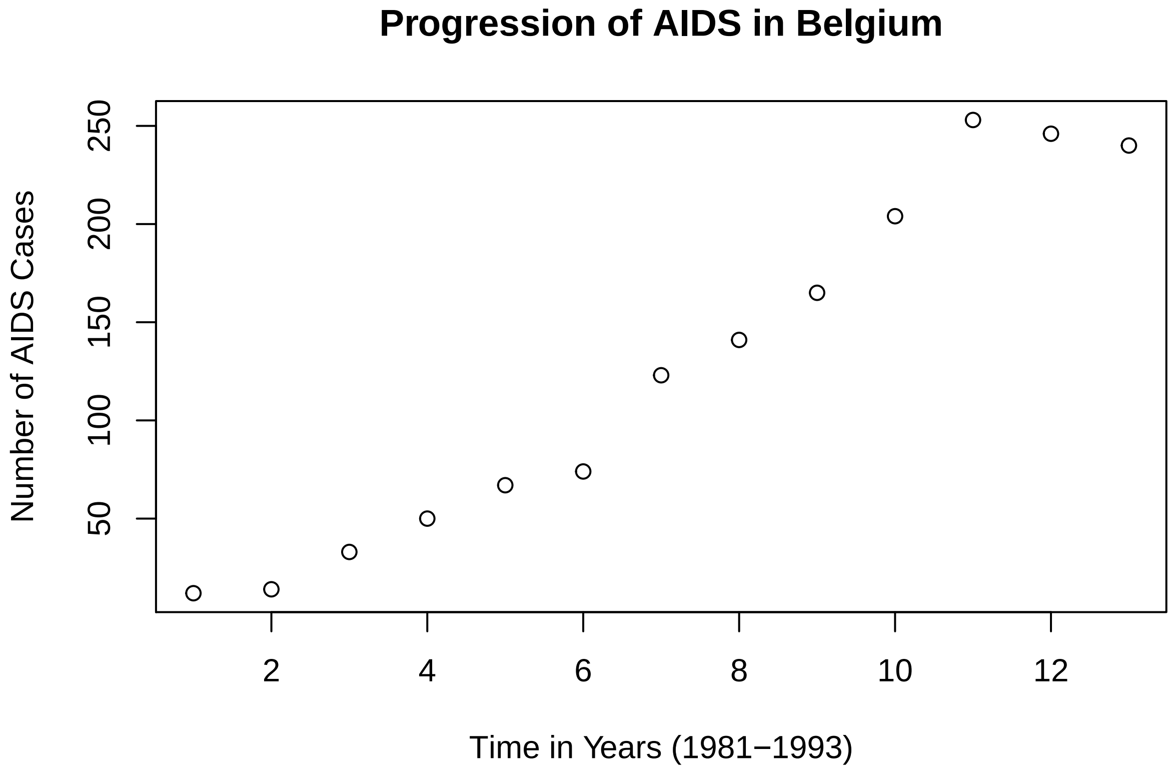
Final Answer: The P values of all the models indicate that none of the variables nor the interaction term is significant and we can reject the alternate hypothesis and there is no correlation between the tumor size and the number of recurrences.The contingency table and distribution plots also indicate that there is no correlation between these two variables.

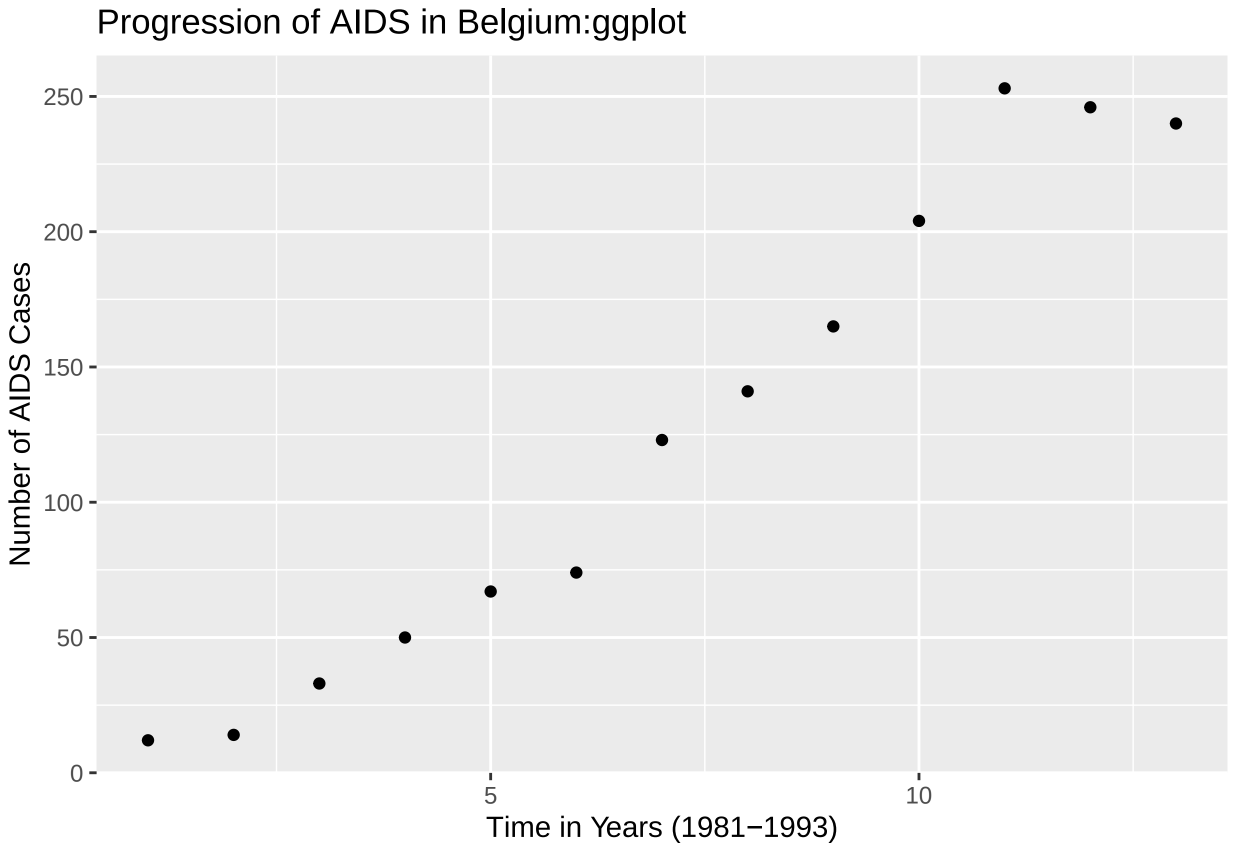
2. Let y denote the number of new AIDS cases in Belgium between the years 1981-1993. Let t denote time.

```
y = c(12, 14, 33, 50, 67, 74, 123, 141, 165, 204, 253, 246, 240)
t = c(1:13)
```

a) Plot the progression of AIDS cases over time. Describe the general nature of the progress of the disease.

Answer 2.a: A scatterplot between Number of AIDS cases over time clearly shows that there is a relation between them. As the time ncreases the number of cases increase until year 1991 and then starts declining.





b) Fit a Poisson regression model $log(\mu_i) = \beta_0 + \beta_1 t_i$. How well do the model parameters describe disease progression? Use a residuals (deviance) vs Fitted plot to determine how well the model fits the data.

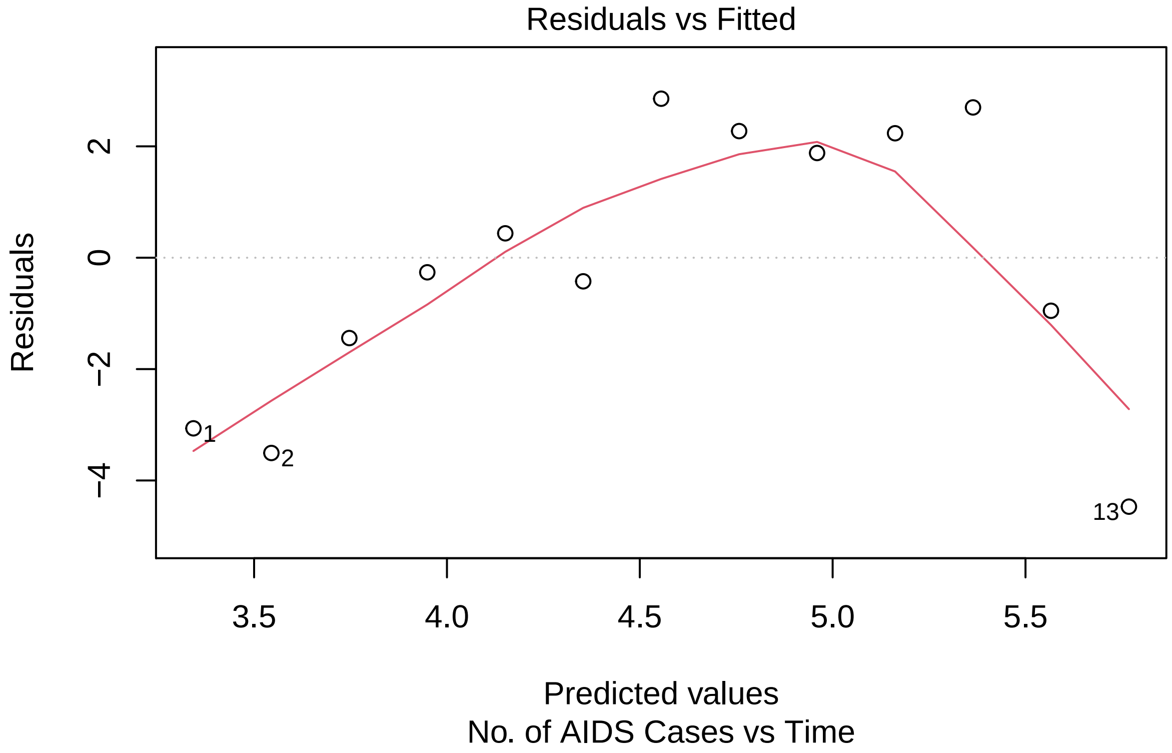
Answer 2.b: A poisson regression model is fit.

After reviewing the model parameters it is clear that the variables are significantly correlated at 95% confidence interval. However,a lower std.error and higher residual deviance than the degrees of freedom indicates an over dispersion (a variance that is higher than mean).

```
##
## Call:
## glm(formula = y ~ t, family = poisson(), data = AIDS.Belgium)
##
## 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) 3.140590   0.078247   40.14   <2e-16 ***
## t           0.202121   0.007771   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
```

The residual plot further confirms that the mean and variance are not the same and the data is over dispersed.It is clear that there is significant variance from the fitted line and there are outliers.



c) Now add a quadratic term in time (*i.e.,* $log(\mu_i) = \beta_0 + \beta_1 t_i + \beta_2 t_i^2$) and fit the model. Do the parameters describe the progression of the disease? Does this improve the model fit? Compare the residual plot to part b).

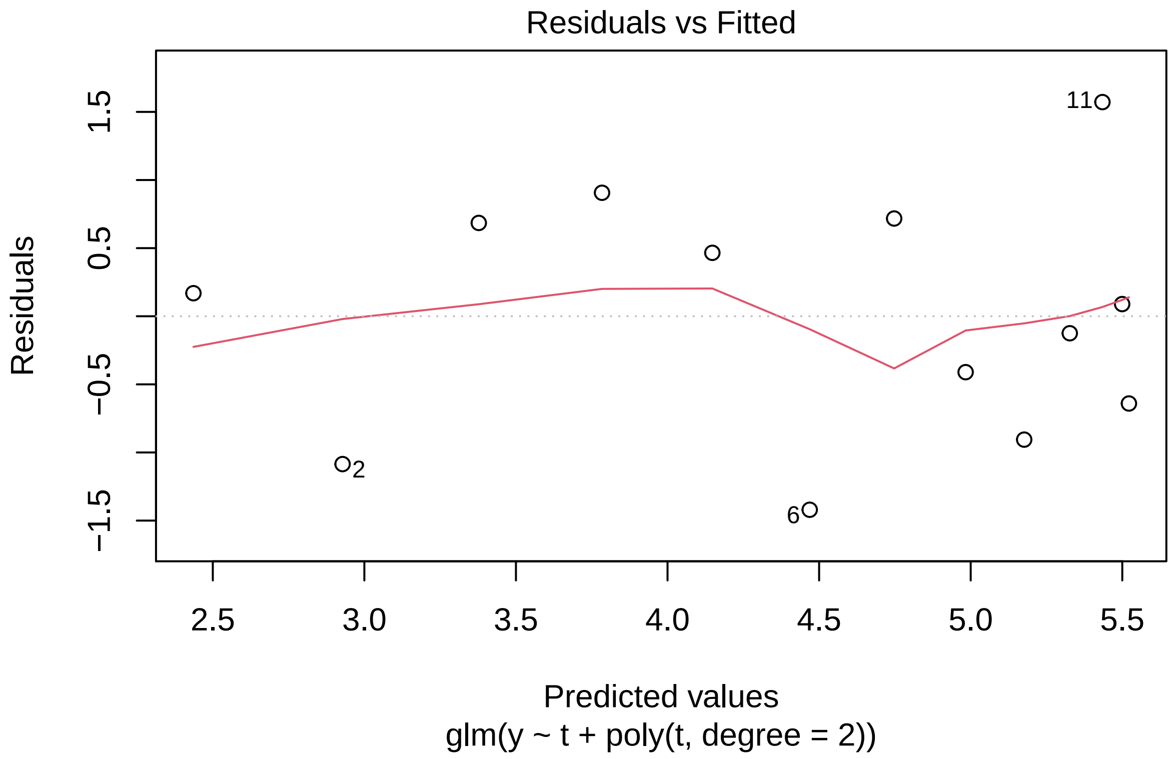
Answer 2.c: A poisson regression model with a quadratic term is fit.

With addition of a second order polynomial to the model the difference between models residual deviance and degrees of freedom is relatively small indicating no over dispersion. Also, the residual deviance is lower indicating less variance with improved standard error.The lower AIC values of 96.924(compared to 166.37 of previous model) indicate that addition of the polynomial improved the model performance.

```
##
## Call:
## glm(formula = y ~ t + poly(t, degree = 2), family = poisson(),
##     data = AIDS.Belgium)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -1.45903  -0.64491   0.08927   0.67117   1.54596
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.64858    0.11323  23.390  < 2e-16 ***
## t                     0.25716    0.01167  22.038  < 2e-16 ***
## poly(t, degree = 2)1       NA         NA      NA       NA
## poly(t, degree = 2)2 -0.95511    0.11896  -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
```

The residual plot further confirms the better performance of the model and lower residual deviance.



d) Compare the two models using AIC. Did the second model improve upon the first? Does this confirm your position from part c)?

Answer 2.d: The second models AIC is lower than the first model(96.92358 and 166.3698 respectively) indicating that the second model explains the data better and it confirms my assumptions that adding a second order polynomial to the linear model reduces the over dispersion and improves model performance.

```
## Residual deviance of linear model: 166.3698
```

```
## Residual deviance of Quadratic model: 96.92358
```

e) Compare the two models using a χ^2 test (`anova` function will do this). Did the second model improve upon the first? Does this confirm your position from part c) and/or d)?

Answer 2.e: The high deviance and low P values clearly show that adding a second degree polynomial in the second model improved the performance.The p value is statistical significant in the second model and hence I reject the null hypothesis that both models are same. This confirms my previous assumptions that quadratic model is better over linear model.

```
## Analysis of Deviance Table
##
## Model 1: y ~ t
## Model 2: y ~ t + poly(t, degree = 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
```

3. (Adapted from ISLR) Load the **Default** dataset from **ISLR** 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. You had developed a logistic regression model on HW #2. Now consider the following two models

Model 1: Default = Student + balance

Model 2: Default = Balance

Answer 3: Two models were fit. One with default vs student+balance and other with default vs balance.

```
##
## Call:
## glm(formula = default ~ student + balance, family = binomial(),
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4578  -0.1422  -0.0559  -0.0203   3.7435
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
## studentYes  -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
## balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
## ---
## 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: 1571.7  on 9997  degrees of freedom
## AIC: 1577.7
##
## Number of Fisher Scoring iterations: 8
```

```
##
## Call:
## glm(formula = default ~ balance, family = binomial(), data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
## ---
## 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: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

Compare the models using the following four model selection criteria.

a) AIC Model 1 had lower AIC than Model 2.Also, MSE of Model 1 is slightly lower than Model 2. Both the terms in Model 1 are significant indicating that the customer being a student or not a student has correlation with default status.

```
## AIC of model with Student and balance variables: 1577.682
```

```
## AIC of model with only balance variable: 1600.452
```

```
## MSE of model with Student and balance variables: 0.02130176
```

```
## MSE of model with only balance variable: 0.02170579
```

```
## Error Rate in % for Model with both variables(Student and balance):  2.67
```

```
## Error Rate in % for Model with only balance variable:  2.75
```

b) Training / Validation set approach. Be aware that we have few people who defaulted in the data.

As only few people defaulted in this data, I have first separated my data based on levels in default.Then,did a 70:30(Train:Test) split for the datasets and did a full join to get my Train and Test data. The two models were fit. There is a very little difference in the model performance with the Student variable in the Model1. The AIC and MSE of Model 1(Student+balance) are slightly better than Model 2(balance). The observed difference between the Training/Validation approach and using full data is that The AIC of both models are less when training/validation approach is used.

```
##
## Call:
## glm(formula = default ~ student + balance, family = binomial(),
##     data = Data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4535  -0.1410  -0.0559  -0.0206   3.7371
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.712492   0.439802 -24.358  < 2e-16 ***
## studentYes   -0.610353   0.174616  -3.495 0.000473 ***
## balance       0.005718   0.000276  20.719  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2043.7  on 6998  degrees of freedom
## Residual deviance: 1101.6  on 6996  degrees of freedom
## AIC: 1107.6
##
## Number of Fisher Scoring iterations: 8
```

```
## MSE of model with Student and balance variables using Training /validation approach : 0.02115002
```

```
##
## Call:
## glm(formula = default ~ balance, family = binomial(), data = Data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3010  -0.1448  -0.0585  -0.0219   3.7551
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.660597   0.432822  -24.63   <2e-16 ***
## balance       0.005535   0.000265   20.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2043.7  on 6998  degrees of freedom
## Residual deviance: 1114.4  on 6997  degrees of freedom
## AIC: 1118.4
##
## Number of Fisher Scoring iterations: 8
```

```
## MSE of model with only balance variables using Training /validation approach : 0.02166547
```

```
## Error Rate in % for Model with both variables(Student and balance):  2.7
```

```
## Error Rate in % for Model with only balance variable:  2.77
```

c) LOOCV

The below models were used for Leave One Out Cross Validation,

```
model7 <- glm(default ~ student + balance,data=Default,family=binomial())
```

```
model8 <- glm(default ~ balance,data=Default,family=binomial())
```

The MSE of the Models show that the Model with variables Student and balance is better than the other model with LOOCV approach.

```
## Error Rate in % of model with Student and balance variables using LOOCV approach : 2.67
```

```
## Error Rate in % of model with only balance variables using LOOCV approach : 2.75
```

d) 10-fold cross-validation.

The below models were used for Leave One Out Cross Validation,

```
model13 <- glm(default ~ student + balance,data=Default,family=binomial())
```

```
model14 <- glm(default ~ balance,data=Default,family=binomial())
```

The MSE of the Models show that the Model with variables Student and balance is better than the other model with 10-fold cross-validation approach.

```
## Error Rate in % of model with Student and balance variables using Kfold approach : 2.69
```

```
## Error Rate in % of model with only balance variables using Kfold approach : 2.77
```

Report validation misclassification (error) rate for both models in each of the four methods (we recommend using a table to organize your results). Select your preferred method, justify your choice, and describe the model you selected.

```
## Error Rate Comparison:
```

```
##      Method.I_AIC Method.II_AIC Method.I_Training_Validation
## [1,]         2.67          2.75                          2.7
##      Method.II_Training_Validation Method.I_LOOCV Method.II_LOOCV
## [1,]                          2.77           2.67            2.75
##      Method.I_Kfold Method.II_Kfold
## [1,]           2.69            2.77
```

Final Answer: From the Error rate comparison of the 2 models and 4 methods, it is evident that error rates from all the methods are almost similar and addition of student variable to the model reduced the error rate slightly.The performance of model with the student and balance variable is better over the other.

I prefer 10 fold cross-validation over other methods as it is simple,faster(computational time is reduced), reduced bias, variance of the resulting estimate goes up with increasing k and every data point gets to be tested exactly once and used in training k-1 times.

4. Load the **Smarket** dataset in the **ISLR** library. 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. Direction is a factor with levels Down and Up, indicating whether the market had a negative or positive return on a given day.

   Develop two competing logistic regression models (on any subset of the 8 variables) to predict the direction of the stock market. Use data from years 2001 - 2004 as training data and validate the models on the year 2005. Use your preferred method from Question #3 to select the best model. Justify your selection and summarize the model.

Answer 4: The Smarket data is split into train (with all data expect from year 2005) and test data(with data of year 2005).2 models were fit. The first model is fit with percentage return from previous 5 days,second and third order polynomial terms of percent return variables and volume. The model parameters show that none of the terms are significant at 95% confidence interval.And the Misclassification rate is also higher.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     poly(Lag1, 3) + poly(Lag2, 2) + poly(Lag3, 2) + poly(Lag4,
##     2) + poly(Lag5, 3) + I(Volume^2), family = binomial, data = train.smarket)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.465  -1.185   1.002   1.161   1.479
##
## Coefficients: (5 not defined because of singularities)
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.158231   0.184463   0.858    0.391
## Lag1           -0.050856   0.052206  -0.974    0.330
## Lag2           -0.046821   0.052636  -0.890    0.374
## Lag3            0.006364   0.052754   0.121    0.904
## Lag4            0.001355   0.052575   0.026    0.979
## Lag5           -0.008332   0.052248  -0.159    0.873
## poly(Lag1, 3)1        NA         NA      NA       NA
## poly(Lag1, 3)2  1.167931   2.321006   0.503    0.615
## poly(Lag1, 3)3  2.033027   2.060567   0.987    0.324
## poly(Lag2, 2)1        NA         NA      NA       NA
## poly(Lag2, 2)2 -0.261738   2.270270  -0.115    0.908
## poly(Lag3, 2)1        NA         NA      NA       NA
## poly(Lag3, 2)2 -0.562699   2.177221  -0.258    0.796
## poly(Lag4, 2)1        NA         NA      NA       NA
## poly(Lag4, 2)2 -2.235985   2.262459  -0.988    0.323
## poly(Lag5, 3)1        NA         NA      NA       NA
## poly(Lag5, 3)2  2.721508   2.286025   1.190    0.234
## poly(Lag5, 3)3 -1.320991   2.118791  -0.623    0.533
## I(Volume^2)    -0.064727   0.089233  -0.725    0.468
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1383.3  on 997  degrees of freedom
## Residual deviance: 1377.1  on 984  degrees of freedom
## AIC: 1405.1
##
## Number of Fisher Scoring iterations: 4
```

```
## Error Rate in % of Model with quadratic terms : 52.78
```

The second model is a linear model with an interaction term.The p values shows that there is a significant correlation between percentage return from previous day and 5 days earlier. The AIC is lower than the previous model but however the misclassification rate is still high.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag5 + Lag1 * Lag5 * Volume,
##     family = binomial, data = train.smarket)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5556  -1.1835   0.8745   1.1674   1.6526
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.16827    0.34060   0.494   0.6213
## Lag1             -0.31002    0.25136  -1.233   0.2174
## Lag5              0.27569    0.25848   1.067   0.2861
## Volume           -0.10340    0.24475  -0.422   0.6727
## Lag1:Lag5        -0.32926    0.16051  -2.051   0.0402 *
## Lag1:Volume       0.17746    0.16607   1.069   0.2853
## Lag5:Volume      -0.19841    0.17631  -1.125   0.2604
## Lag1:Lag5:Volume  0.17311    0.09891   1.750   0.0801 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1383.3  on 997  degrees of freedom
## Residual deviance: 1374.1  on 990  degrees of freedom
## AIC: 1390.1
##
## Number of Fisher Scoring iterations: 4
```

```
## Error Rate in % of Model with interaction term : 53.57
```

Kfold cross validation is performed using the whole data set with k of 10. Similar models were used. I believe that the linear model is better over model with quadratic terms as AIC of the linear model is lower than the Quadratic model.The method I have used to make this selection is Kfold cross-validation. Comparing the models build, it is clear that the the Error rate of K-fold CV is lower than Training/validation approach. And Kfold CV reduces bias and variance of the models while using every data point for testing exactly once and in training k-1 times.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     poly(Lag1, 3) + poly(Lag2, 2) + poly(Lag3, 2) + poly(Lag4,
##     2) + poly(Lag5, 3) + I(Volume^2), family = binomial, data = Smarket)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.555  -1.200   1.047   1.141   1.479
##
## Coefficients: (5 not defined because of singularities)
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.009949   0.125338  -0.079    0.937
## Lag1           -0.069697   0.050489  -1.380    0.167
## Lag2           -0.043388   0.050878  -0.853    0.394
## Lag3            0.008763   0.050966   0.172    0.863
## Lag4            0.002132   0.050818   0.042    0.967
## Lag5            0.006045   0.050498   0.120    0.905
## poly(Lag1, 3)1        NA         NA      NA       NA
## poly(Lag1, 3)2  0.537418   2.299190   0.234    0.815
## poly(Lag1, 3)3  2.604522   2.072045   1.257    0.209
## poly(Lag2, 2)1        NA         NA      NA       NA
## poly(Lag2, 2)2 -1.087557   2.263760  -0.480    0.631
## poly(Lag3, 2)1        NA         NA      NA       NA
## poly(Lag3, 2)2 -0.780208   2.183607  -0.357    0.721
## poly(Lag4, 2)1        NA         NA      NA       NA
## poly(Lag4, 2)2 -2.331578   2.280666  -1.022    0.307
## poly(Lag5, 3)1        NA         NA      NA       NA
## poly(Lag5, 3)2  2.152979   2.300857   0.936    0.349
## poly(Lag5, 3)3 -1.335082   2.113019  -0.632    0.527
## I(Volume^2)     0.036473   0.048333   0.755    0.450
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1731.2  on 1249  degrees of freedom
## Residual deviance: 1723.6  on 1236  degrees of freedom
## AIC: 1751.6
##
## Number of Fisher Scoring iterations: 4
```

```
## Error rate of the model-4c in % : 50.88
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag5 + Lag1 * Lag5 * Volume,
##     family = binomial, data = Smarket)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.643  -1.203   1.011   1.144   1.640
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.16104    0.24224  -0.665   0.5062
## Lag1             -0.27678    0.22408  -1.235   0.2168
## Lag5              0.08376    0.22943   0.365   0.7150
## Volume            0.15717    0.15930   0.987   0.3238
## Lag1:Lag5        -0.34553    0.15851  -2.180   0.0293 *
## Lag1:Volume       0.13951    0.14338   0.973   0.3305
## Lag5:Volume      -0.05411    0.15178  -0.356   0.7215
## Lag1:Lag5:Volume  0.18178    0.09682   1.877   0.0605 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1731.2  on 1249  degrees of freedom
## Residual deviance: 1721.5  on 1242  degrees of freedom
## AIC: 1737.5
##
## Number of Fisher Scoring iterations: 4
```

```
## Error rate of the model-4d in % : 46.8
```