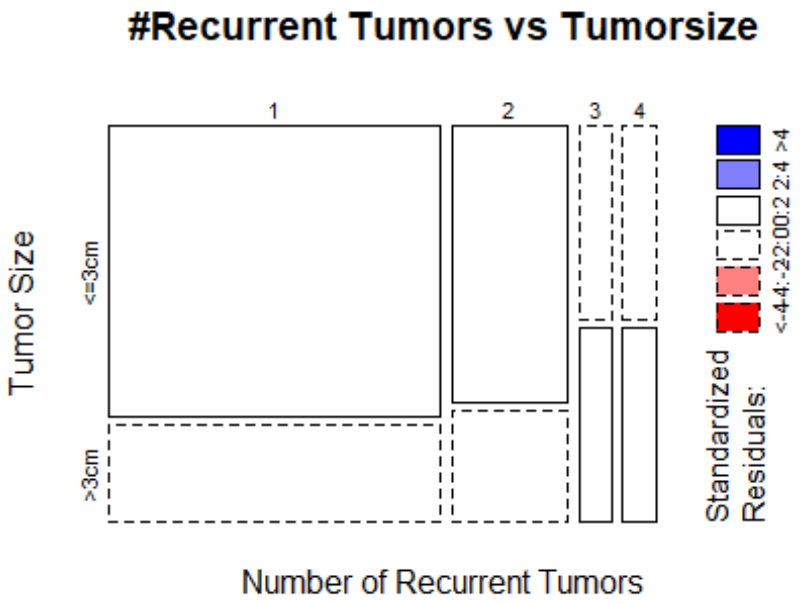
Generlized Linear Models-2

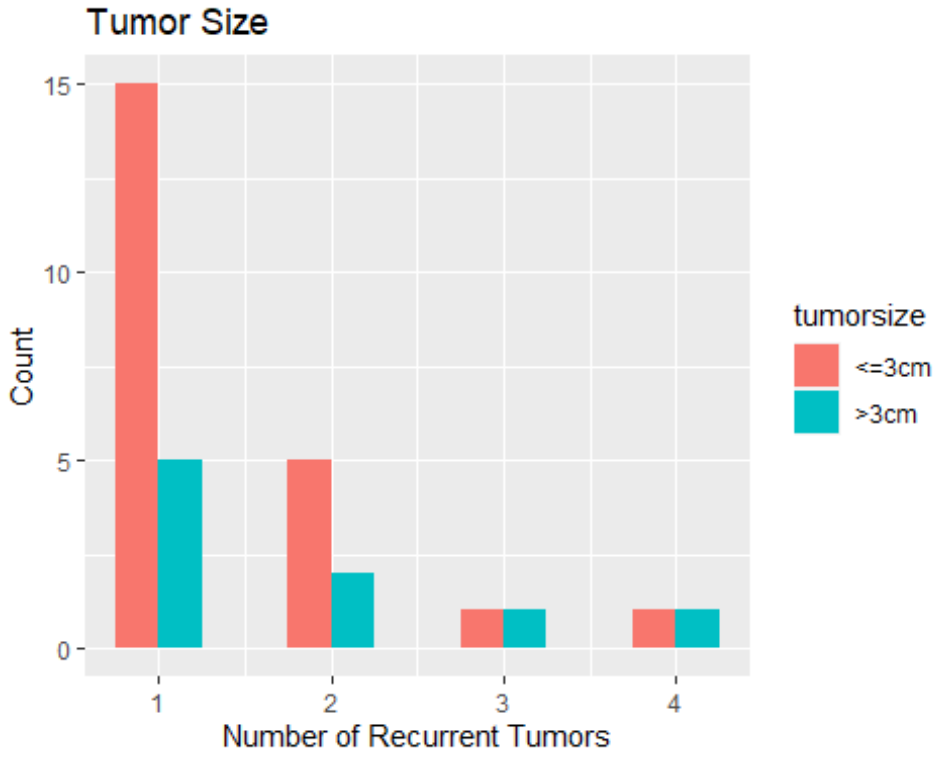
STAT 601

Exercises

1. (Ex. 7.3 pg 147 in HSAUR, modified for clarity) Use the data from the 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: By, looking at the histogram plot, tumor>3cm and tumor<= 3cm don't have normal distribution. Th histogram shows that one tumor is occurring the most and the frequency of more tumors occurring is smaller than having two tumor.

```
## # A tibble: 2 x 5
## # Groups:   tumorsize [2]
##   tumorsize `number of tumor~ `number of tumo~ `number of tumo~ `number of
tumo~
##   <fct>                 <int>            <int>            <int>
<int>
## 1 <=3cm                    15                5                1
1
## 2 >3cm                      5                2                1
1
```

The frequency table show that having one tumor occurs the most for all tumor sizes and the decreases as the number of tumors increase. The second highest most occurring is two tumors and the lowest is three and four tumors.

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)?

```
##
## 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 + time, 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
## tumorsize>3cm  0.20511    0.30620   0.670    0.503
## time           0.01478    0.01883   0.785    0.433
##
## (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
```

Answer: For the first variable, the explanatory variable intercept is significant with p-value<0.05. The tumor size variable is not significant with a p-value > 0.05. For the second model,the explanatory variable intercept is not significant with p-value >0.05. The tumor size variable is not significant with a p-value > 0.05. The extra term time is not significant with p-value > 0.05 which means it does not improve the second model. The deviance for model one is 12.38 on 29 degrees of freedom which indicates under dispersion. For model two, the deviance. is 11.757 on 28 which indicates under dispersion. The AIC for model one (87.191) is lower than model two (88.568) which means that it is better model.

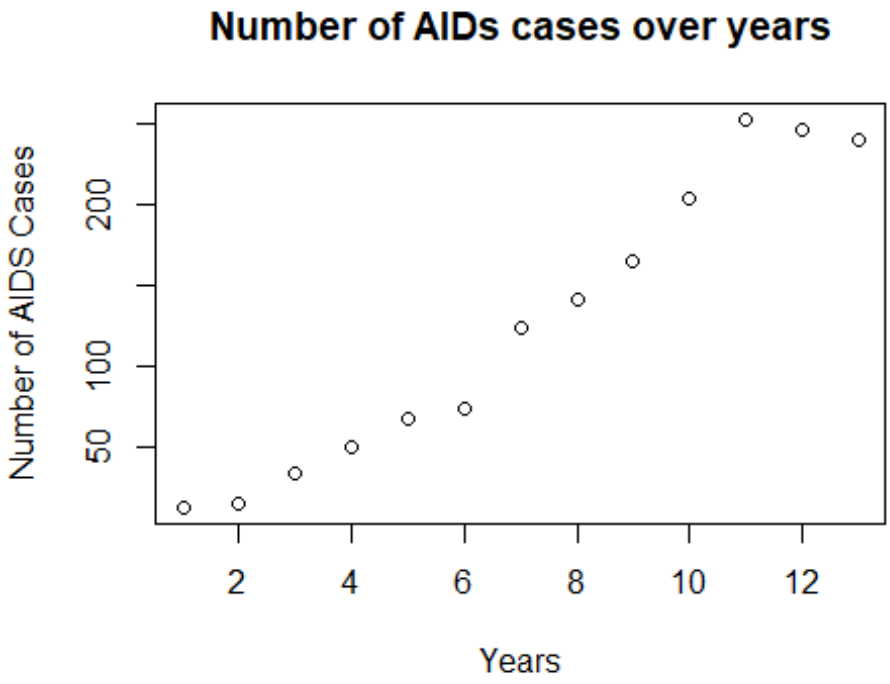
```
## Analysis of Deviance Table
##
## Model 1: number ~ tumorsize
## Model 2: number ~ tumorsize + time
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        29     12.380
## 2        28     11.757  1  0.62363   0.4297
```

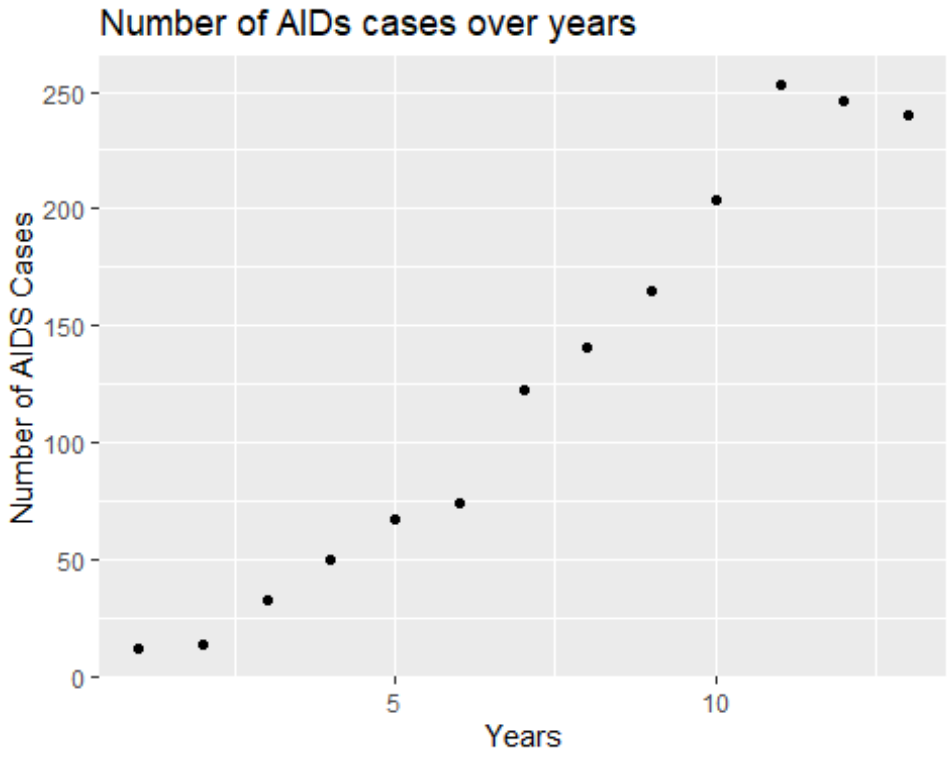
We can use X^2 test to to see if the added variable time had any effect on the model. The deviance(12.380) for the first model is far from its df(29). For the second model the deviance is (11.757) which is far from its df(28). The P-value>0.05 which implies that there is no statistical significance in the second model. Adding a variable does not improve the model. By looking at the residuals vs fitted values plot, we can see that both models have either high residuals or low residuals.Both models did not fit the data well. We will choose model two to be the better model because the chi square test accepts the null hypothesis that model one is a better model

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

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

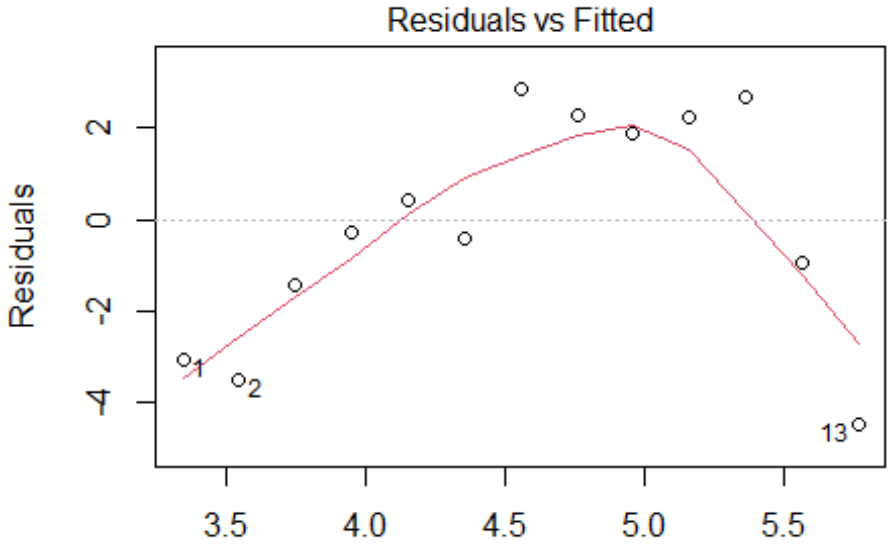
The general trend is increasing linear relationship between the number of aids cases increases and years. Approximately after year 11, there is a drop in the number of cases.

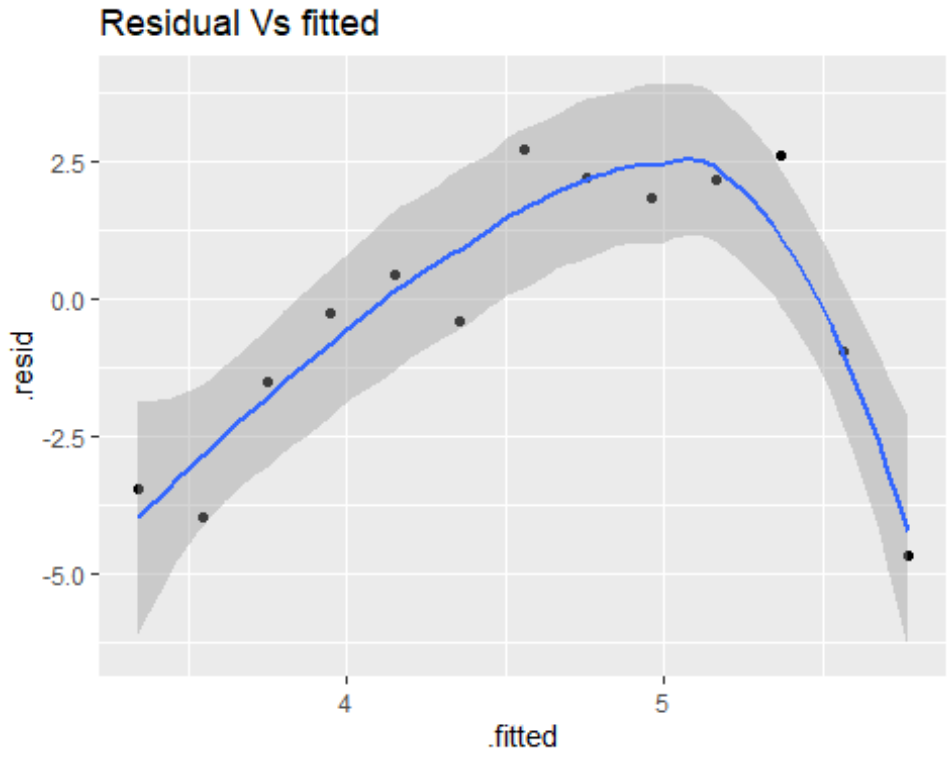




b) Fit a Poisson regression model $log(\mu_i) = \beta_0 + \beta_1 t_i$. How well do the model parameters describe disease progression? Use a residuals (deviance) vs Fitted plot to determine how well the model fits the data.

```
##
## Call:
## glm(formula = N.cases ~ years, family = poisson, data = B.data)
##
## 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 ***
## years       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
```





B. Answers:

How well do the model parameters describe disease progression?

The variable years is highly significant with a p-value< 0.05 , small standard error, and coefficient value>0, which means that as years go by, we expect the number of AIDS cases

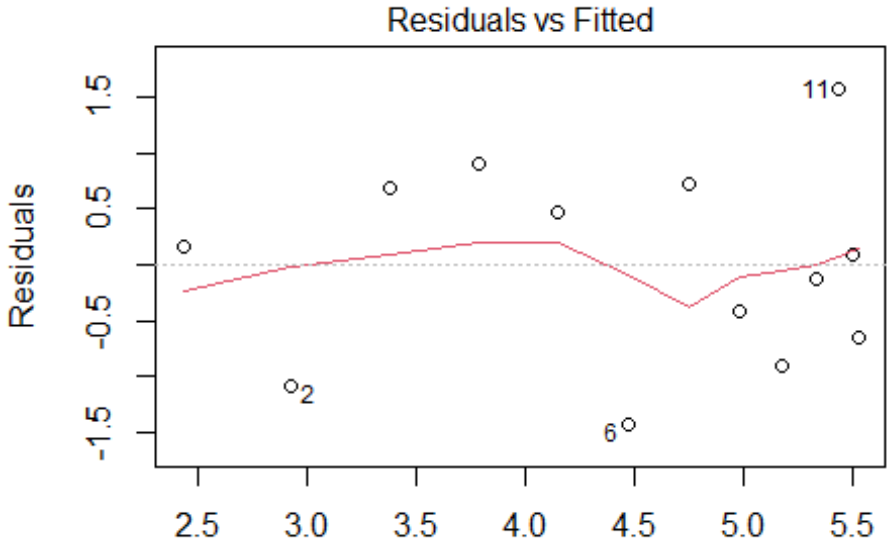
to increase.
The explanatory variable year indicates that the rate ratio is 1.22, we concluded that the number of AIDS cases increased by 22% each year from 1981 to 1993. For one unit increase in years, the number of AIDS cases will increase.

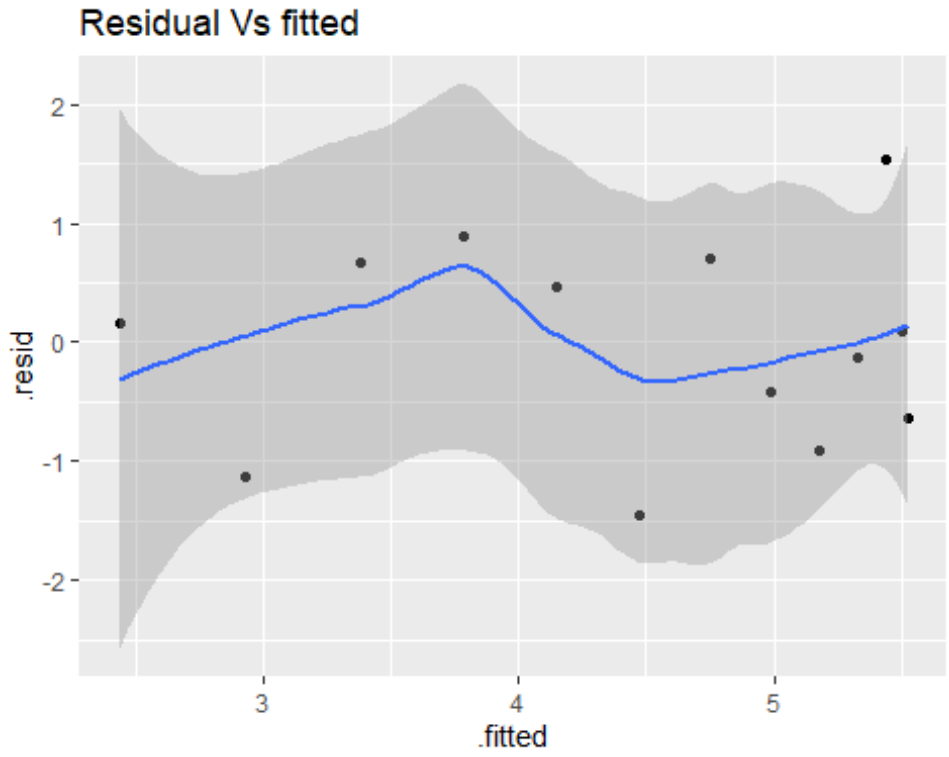
Use a residuals (deviance) vs Fitted plot to determine how well the model fits the data

For a model that fits the data well, residuals should be randomly scattered around zero for the entire range of fitted values. That would show that the model's predictions are correct on average instead of being too low or too high. For a well fitted model, the explanatory variables should explain the relationship well that only random errors remains meaning that the errors should not have a pattern. For this model, we can clearly see that the residual errors have a pattern and the fitted values are not scattered around zero at all. This means that model can be improved to fit the data well by adding more explanatory variables. Aslo, Data points 1,2, and 13 are ouliers because they are far from zero.

c) Now add a quadratic term in time ( ) 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).

```
##
## Call:
## glm(formula = N.cases ~ years + years2, family = "poisson", data = B.data)
##
## 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)  1.901459   0.186877  10.175  < 2e-16 ***
## years        0.556003   0.045780  12.145  < 2e-16 ***
## years2      -0.021346   0.002659  -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
```





Answers:

C. Do the parameters describe the progression of the disease?

The variable years is significant with a p-value< 0.05 , small standard error, and coefficient value>0, which means that as years go by, we expect the number of AIDS cases to increase.

The variable years^2 is significant with a p-value< 0.05 , small standard error, and coefficient value>0, but the coefficient is very close to zero. This means one years go by, we expect the number of AIDS cases to increase.

The parameter years shows that the rate ratio is (1.74), which means that the number of cases increases by 73% each years from 1981 to 1993. For one unit increase in years, the number of AIDS cases will increase.

```
## (Intercept)       years      years2
##   6.6956535   1.7436895   0.9788799
```

Does this improve the model fit?

By looking at the graph we can say that this model fits the data better than the first model. By adding one quadratic variable, we can see that the residuals are scattered around zero and there is randomness in the layout of the points. These two factors are a good indicator that this model improves the model fit. we still have some outliers with high and low residual values(2,6,11).

Compare the residual plot to part b

By looking at booth graphs, we can see that the quadratic model fit the data better. The first model errors follow a pattern and the quadratic model errors are random. Random errors imply that the explanatory variables explain the relationship better. Also, unlike the first model, the quadratic model residuals are centered around zero which indicates that the quadratic model predictions are are correct on average. Overall, the second model is a better fit for the data.

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

```
## [1]   2.0000 166.3698

## [1]  3.00000 96.92358
```

Model1 AIC: 166.37 Model2 AIC: 96.924 AIC indicates in sample prediction error and the quality of the model. The second model has a lower Akaike information criterion (AIC) than the first model which shows that adding a quadratic term improved the quality of the second model. AIC also strongly favors a quadratic model. Yes The lower AIC for the second model confirms my position from Part c that the second model overall is a better model for the data than the first model.

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

```
## Analysis of Deviance Table
##
## Model 1: N.cases ~ years
## Model 2: N.cases ~ years + years2
##   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
```

We can use X^2 test to to see if the quadratic variable had any effect on the model. The deviance(9.240) for the second model is closer to its df(10. For the first model the deviance is (80.686) which is far from its df(11). The one degrees of freedom for the the X^2 test indicates that the quadratic variable is statistically significant predictor of the number of AIDS cases. The P-value<0.05 which implies that there is statistical significance in the second model. Adding a quadratic variable improves the model. This confirms my position from part c & d.

3. (Adapted from ISLR) Load the dataset from library. The dataset contains four features on 10,000 customers. We want to predict which customers will default on their credit card debt based on the observed features. You had developed a logistic regression model on HW #2. Now consider the following two models

Compare the models using the following four model selection criteria.

a) AIC

```
## [1] 1577.682
```

```
## [1] 1600.452
```

Answer: Both models have high AIC values. The first model has a lower Akaike information criterion (AIC) than the second model which shows that excluding the covariate term Student reduced the quality of the second model. since the AIC for the second model is lower, we can say the second model is better than the first model.

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

```
##
## Call:
## glm(formula = default1 ~ student + balance, family = "binomial",
##     data = training.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4929  -0.1290  -0.0479  -0.0168   3.8448
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.131e+01  4.512e-01  -25.07  < 2e-16 ***
## studentYes  -6.270e-01  1.695e-01   -3.70 0.000215 ***
## balance      6.012e-03  2.802e-04   21.46  < 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: 2218.1  on 7999  degrees of freedom
```

```
## Residual deviance: 1162.0  on 7997  degrees of freedom
## AIC: 1168
##
## Number of Fisher Scoring iterations: 8

##
## Call:
## glm(formula = default1 ~ balance, family = "binomial", data =
training.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3249  -0.1324  -0.0505  -0.0180   3.8564
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.213456   0.442201  -25.36   <2e-16 ***
## balance       0.005791   0.000267   21.69   <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: 2218.1  on 7999  degrees of freedom
## Residual deviance: 1176.3  on 7998  degrees of freedom
## AIC: 1180.3
##
## Number of Fisher Scoring iterations: 8

## Model One Error 0.0305

## Model Two Error 0.031
```

Dissuasion

The first model has a lower AIC than the second model. Using the training/ validation approach resulted in lower AIC values for both models compared to step A models. All the variables for both models are significant with p-value<0.05. The Deviance for deviance values for both models are are lower than their degrees of freedom which suggests that the models are under dispersed. Again, since model one has lower AIC, We conclude that model one is better than model two.

c) LOOCV

```
## Model One

## [1] 0.02849921

## Model Two

## [1] 0.02956515
```

Answers:

Model one error is (0.02849921), and model two error is (0.02956515). Model one has a smaller error; therfore, it is more accurate. We choose model one to be the better model.

d) 10-fold cross-validation.

```
## [1] 0.01959019

## [1] 0.01980939
```

Answer: Using the 10-fold cross-validation, we can see that the first model has a lower error(0.0196)than the second model(0.0198); therefore, we can consider the first model to be better than the second model.

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.

```
##              Method   Model.1   Model.2
## 1               AIC 1168.0000 1180.3000
## 2 Train/Validation    0.0305    0.0310
## 3            Loocv    0.0280    0.0300
## 4 Cross-Validation    0.0196    0.0198
```

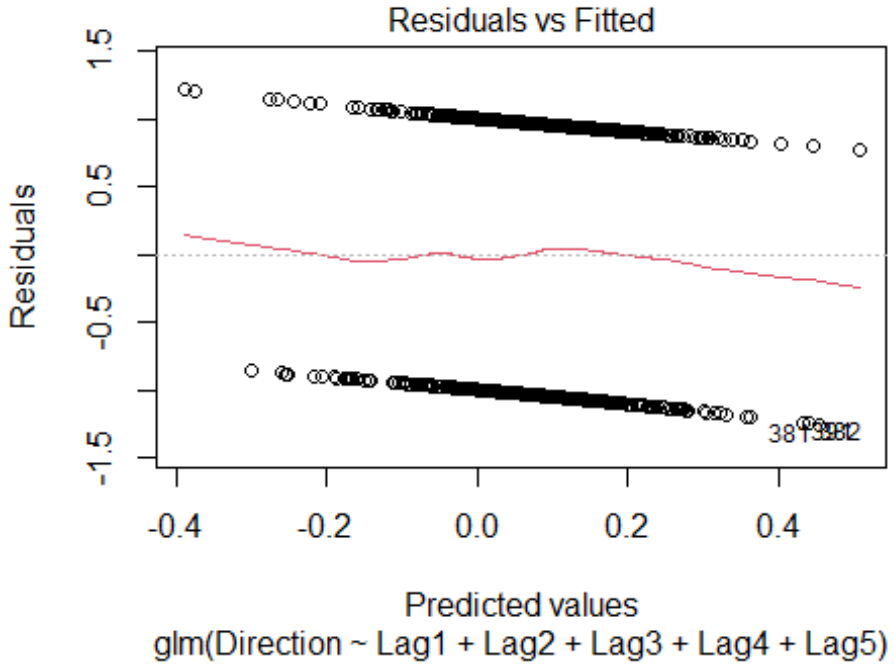
The preferred method is 10-fold cross-validation because this method helps build k different models, so we are able to make predictions on all of our data set. We can perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model which can give us more accurate estimate of out-of-sample accuracy. The first model has a lower AIC than the second model. All the variables for both models are significant with p-value<0.05. The Deviance for deviance values for both models are are lower than their degrees of freedom which suggests that the models are under dispersed. Also, for all the methods used above to split the datat, model one had a smaller error rate. We conclude that model one is better than model two.

4. Load the dataset in the 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.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5, family =
"binomial",
##     data = Smarket)
```

```
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.376  -1.204   1.071   1.145   1.347
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.074163   0.056674   1.309    0.191
## Lag1        -0.071325   0.050104  -1.424    0.155
## Lag2        -0.044136   0.050025  -0.882    0.378
## Lag3         0.009229   0.049879   0.185    0.853
## Lag4         0.007211   0.049898   0.145    0.885
## Lag5         0.009311   0.049490   0.188    0.851
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1731.2  on 1249  degrees of freedom
## Residual deviance: 1728.3  on 1244  degrees of freedom
## AIC: 1740.3
##
## Number of Fisher Scoring iterations: 3
```

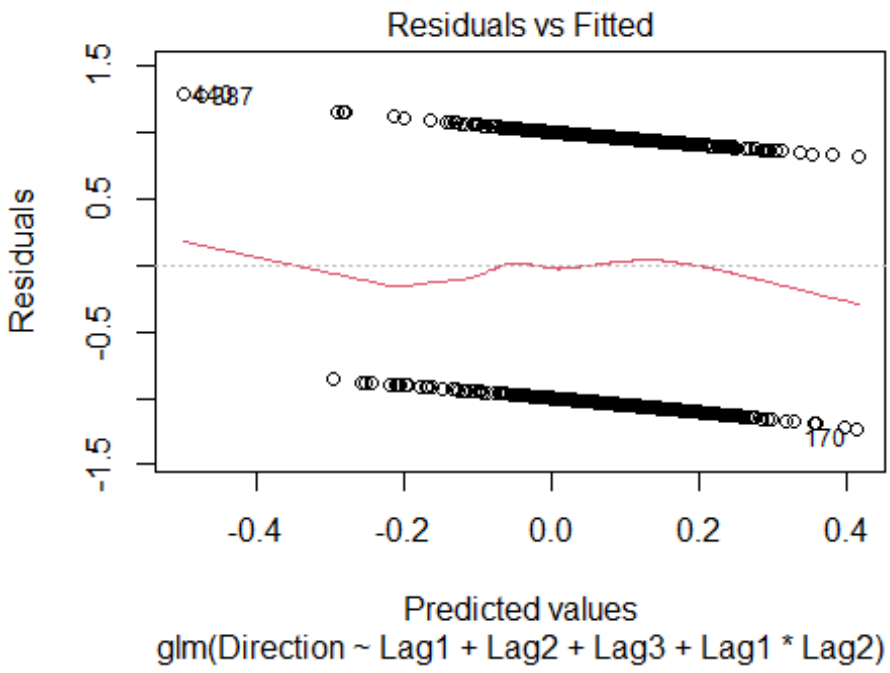



```
## ERROR 0.2524667
```

Discussion: This model uses all the variables that were the data set as explanatory variables for the model. All the explanatory variables are insignificant since they all p-values>0. By looking at the residuals polt, we can see that the errors have a pttarn and the

errors are far from zero, which indicates that the model is a bad fit. The residual deviance is 1728 with df 1244 which indicates that the model is over dispersed. cross validation was used for this model. The model has an error rate of 0.252 about 25%

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag1 * Lag2, family =
"binomial",
##     data = Smarket)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.357  -1.205   1.078   1.145   1.396
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.073930   0.056687   1.304    0.192
## Lag1        -0.072928   0.050526  -1.443    0.149
## Lag2        -0.044485   0.050029  -0.889    0.374
## Lag3         0.009019   0.049859   0.181    0.856
## Lag1:Lag2   -0.008054   0.033959  -0.237    0.813
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1731.2  on 1249  degrees of freedom
## Residual deviance: 1728.3  on 1245  degrees of freedom
## AIC: 1738.3
##
## Number of Fisher Scoring iterations: 3
```




```
## ERROR TWo 0.2506402
```

This model uses the variables Lag1 and lag2 as explanatory variables for the model. All the explanatory variables are insignificant since they all p-values > 0. By looking at the residuals plot, we can see that the errors have a pattern and the errors are far from zero, which indicates that the model is a bad fit. The residual deviance is 1728.4 with df 1246 which indicates that the model is over dispersed. cross validation was used for this model. The model has an error rate of 0.252 about 25%

```
## Analysis of Deviance Table
##
## Model 1: Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5
## Model 2: Direction ~ Lag1 + Lag2 + Lag3 + Lag1 * Lag2
##   Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
## 1      1244     1728.3
## 2      1245     1728.3 -1 0.0014366
```

We can use X^2 test to to see if the interaction variable had any effect on the model. The deviance(1245) for the second model is far from its df(1728.3). For the first model the deviance is (1244) which is far from its df(1728.3). The P-value>0.05 which implies that there is no statistical significance in the second model. Model one has an error rate of o.252 and model two has and error rate of 0.25. The AIC is lower for the second model indicating that it is better model. We conclude that the seocn model is a better model overall.