Homework 3

Amin Baabol

Instructions

Answer all questions stated in each problem. Discuss how your results address each question.

Submit your answers as a pdf, typeset (knitted) from an Rmd file. Include the Rmd file in your submission. You can typeset directly to PDF or typeset to Word then save to PDF In either case, both Rmd and PDF are required. If you are having trouble with .rmd, let us know and we will help you. If you knit to Word, check for any LaTeX commands that will not be compatible with Word.

This file can be used as a template for your submission. Please follow the instructions found under “Content/Begin Here” titled . No code should be included in your PDF submission unless explicitly requested. Use the `echo = F` flag to exclude code from the typeset document.

For any question requiring a plot or graph, answer the question first using standard R graphics (See ?graphics). Then provide a equivalent answer using `library(ggplot2)` functions and syntax. You are not required to produce duplicate plots in answers to questions that do not explicitly require graphs, but it is encouraged.

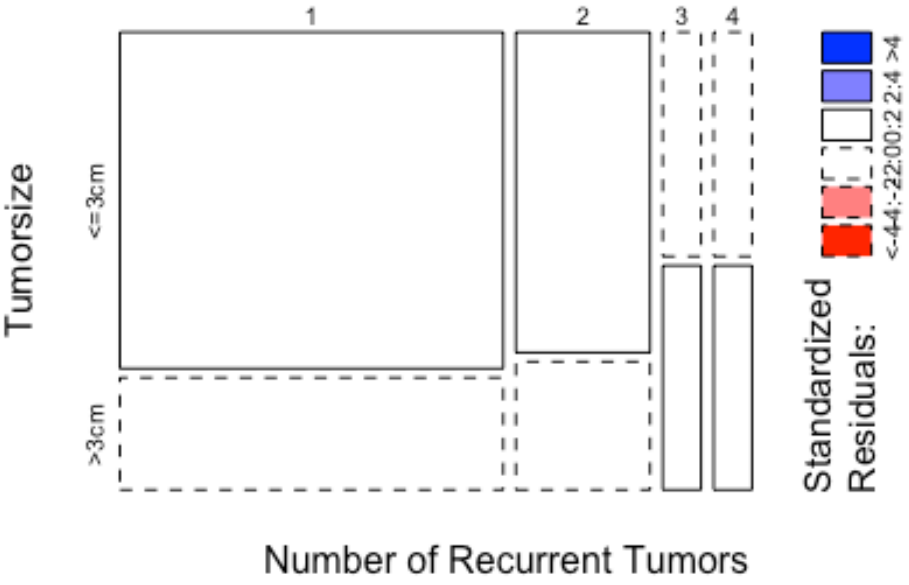
You can remove the `Instructions` section from your submission.

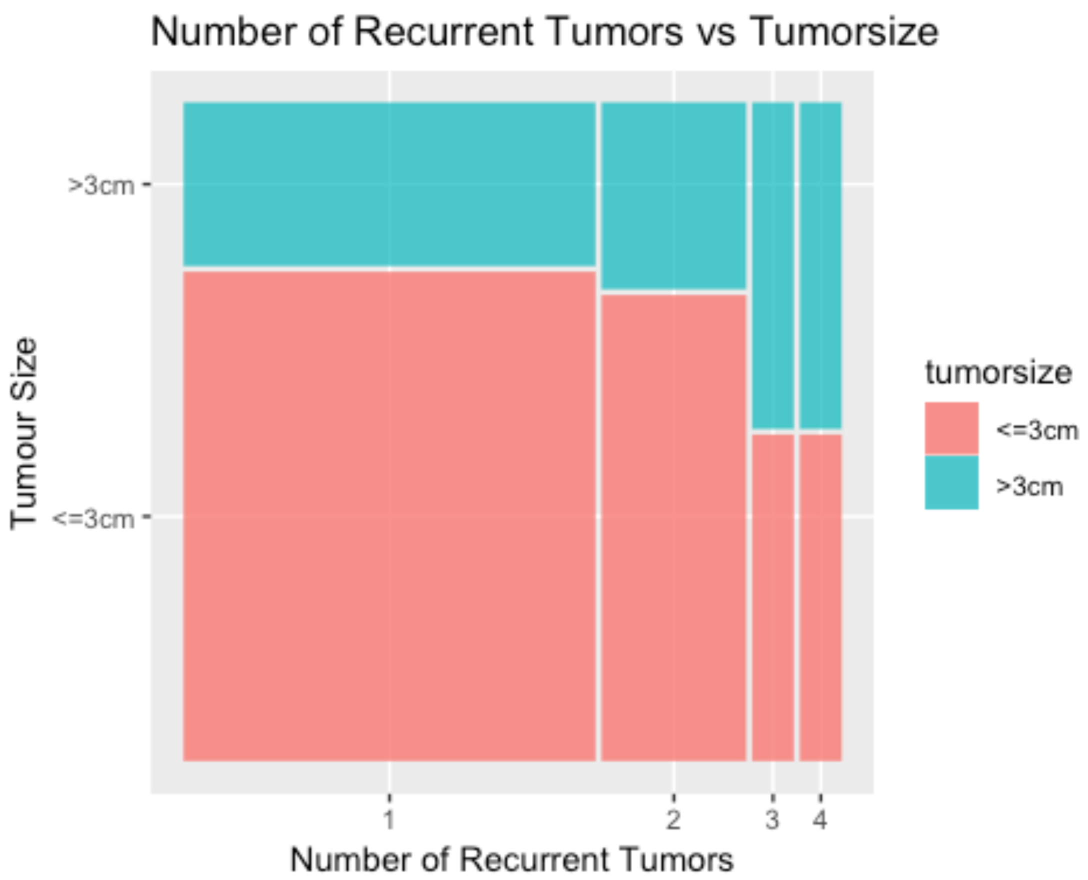
Exercises

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.)

Number of Recurrent Tumors vs Tumorsize





Discussion: Based on the visual observtion of the two mosaicplots, the observed frequency for one or two tumors greater than 3cm is lower than expected whereas the oberseved frequency for 3 or 4 tumors less than or equal to 3cm is lower than expected.

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
```

Discussion: According to the first poisson regression model, only the intercept is statistically signifcant at the 0.05 level. None of the explanatory variables have any meaningful significance despite the null deviance,residual deviance and the AIC are all being pretty low. At this point we can't draw a definite conclusion,So we'll compare the first model (model0) to a second model(model1) where we add an interaction between the explanatory variables, then we'll create a third model where the intercept is omitted because by suppressing the intercept all levels of the factor variable(tumorsize) are estimated, as opposed to using the standard constrast where the level "<=3cm" is set as the baseline in relation to the other coefficients. This will also allow us to examine if the significance of the coefficients change.

```
##
## 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
```

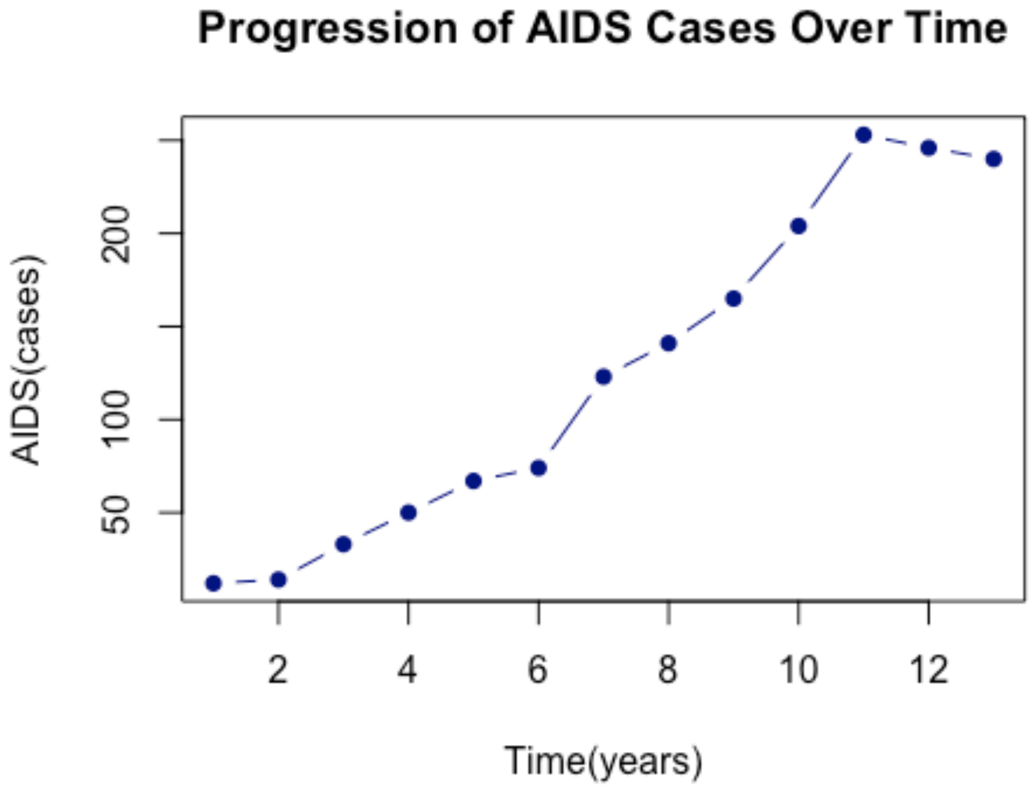
```
##
## Call:
## glm(formula = number ~ tumorsize - 1, family = poisson, data =
bladdercancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6363  -0.3996  -0.3996   0.4277   1.7326
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## tumorsize<=3cm   0.3747     0.1768   2.120   0.0340 *
## tumorsize>3cm    0.5754     0.2500   2.302   0.0214 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 20.772  on 31  degrees of freedom
## Residual deviance: 12.380  on 29  degrees of freedom
## AIC: 87.191
##
## Number of Fisher Scoring iterations: 4
```

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

Discussion: From the above analysis I accept the null hypothesis that there is nothing within the data to explain an increase in the number of tumors. Neither time nor the tumor size have any affect on increasing the number of tumors. This can be further proved through part one and interpreting the distribution of the data.

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.



Discussion: On a quick-glance it seems like there is a linear-type characterization of the relationship between number of AIDS cases and time.However, we'll have to do further analysis to draw a conclusion.

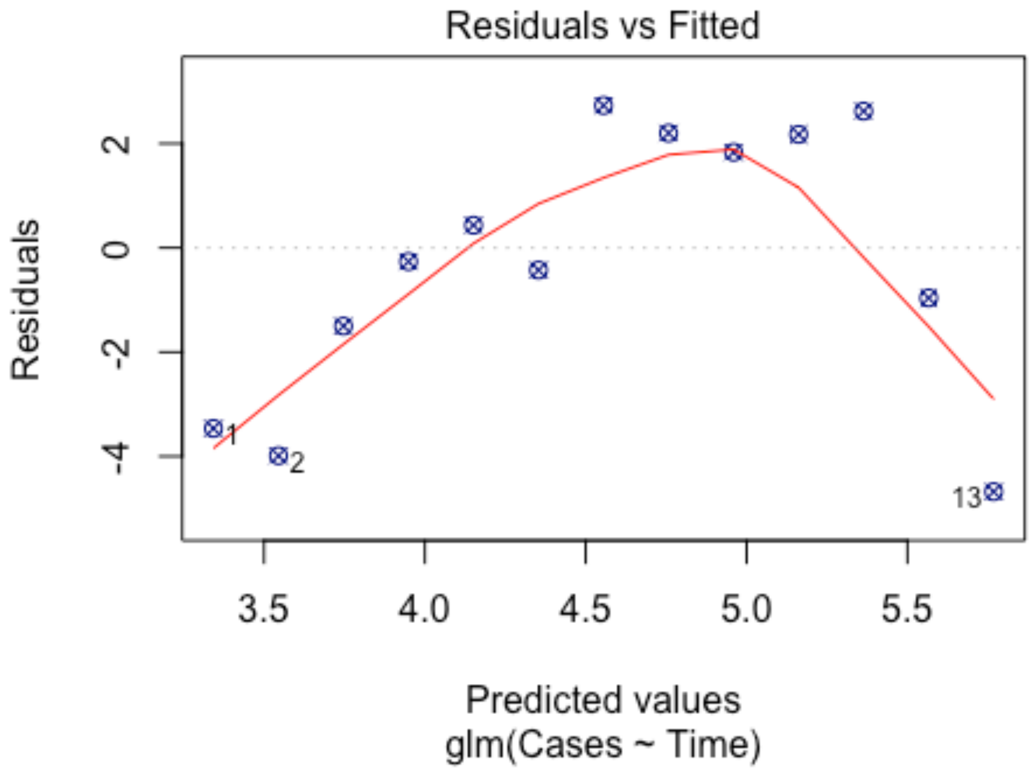
b) Fit a Poisson regression model $log(\mu_i) = \beta_0 + \beta_1 t_i$. How well do the model parameters describe disease progression? Use a residuals (deviance) vs Fitted plot to determine how well the model fits the data.

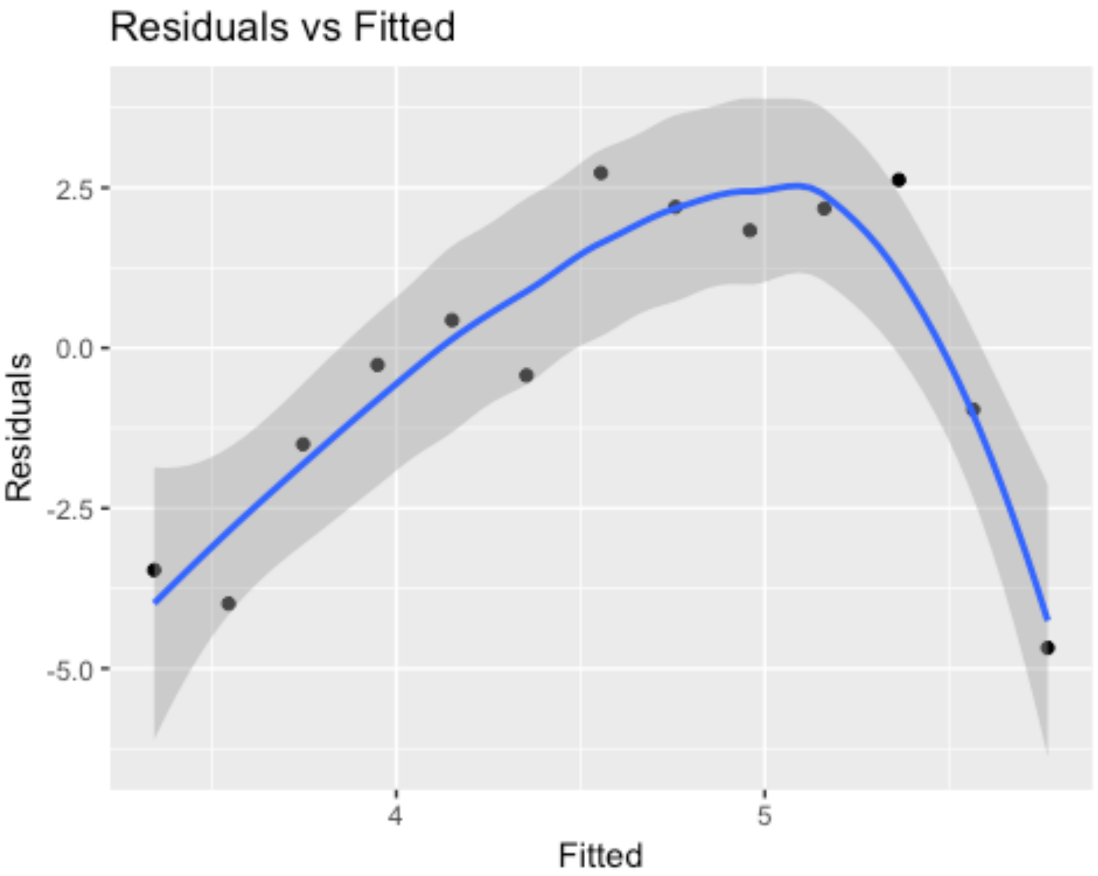
```
##
## Call:
## glm(formula = Cases ~ Time, family = poisson, data = pois.dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6784  -1.5013  -0.2636   2.1760   2.7306
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 3.140590   0.078247   40.14   <2e-16 ***
## Time        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
```

```
## (Intercept)        Time 
##   23.117491    1.223996
```

```
##                 2.5 %    97.5 %
## (Intercept) 19.789547 26.894433
## Time         1.205624  1.242922
```


Residuals vs Fitted
Residuals
2
0
-2
-4
1
2
13
3.5
4.0
4.5
5.0
5.5
Predicted values
glm(Cases ~ Time)



Discussion: The intercept and the time predictor variable are statistically significant at the 0.001 level with a confidence interval of 97.5% there is 21% increase int the AIDS cases, which means there is a strong relationship between the response and the explanatory variables. Having said that, taking a closer at the the residual deviance and the degree of freedom it is evident that this model needs further work to improve its accuracy. Poisson distribution assumes a ratio of 1, meaning the mean and the variance are equal. Therefore, we should be striving to have deviance/degree of freedom ratio closer to 1.

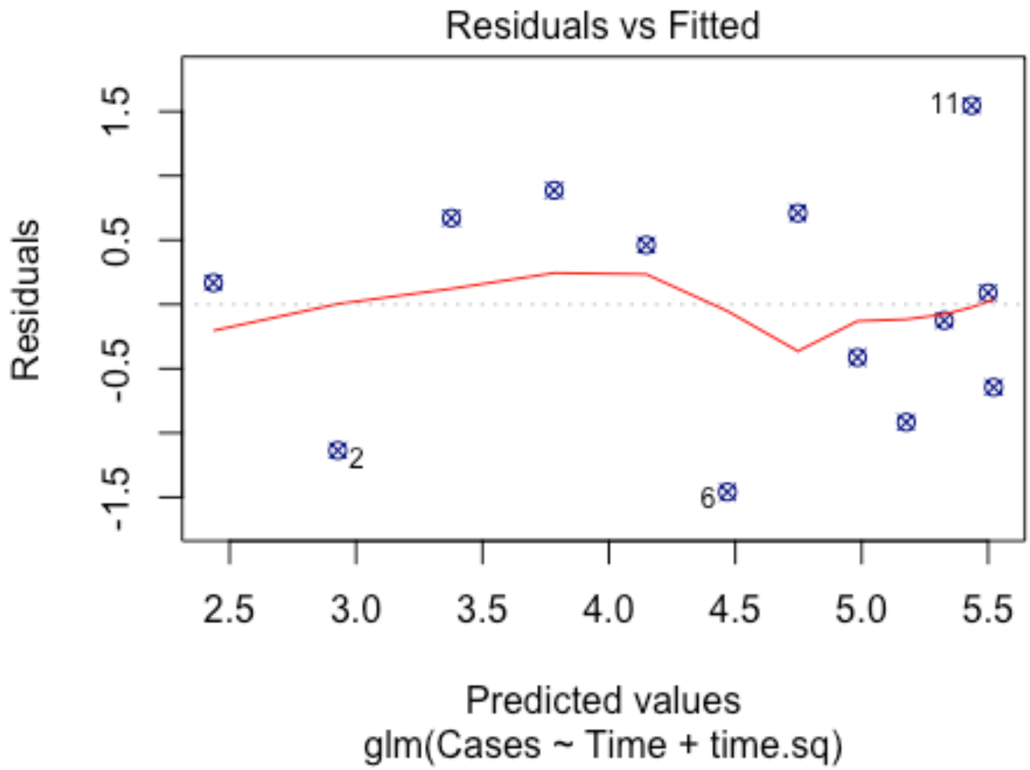
In the context of this plot the residuals and the fitted values indicated an overdispersion because the points don't center around 0. There is also an apparent quadratic shape taking place, I'm not sure if this is a coincidence but there shouldn't be recognizable patterns except random errors. It's therefore, evident we need to introduce another term to potentially reduce the deviances and the AIC, thus improving the accuracy of the model.

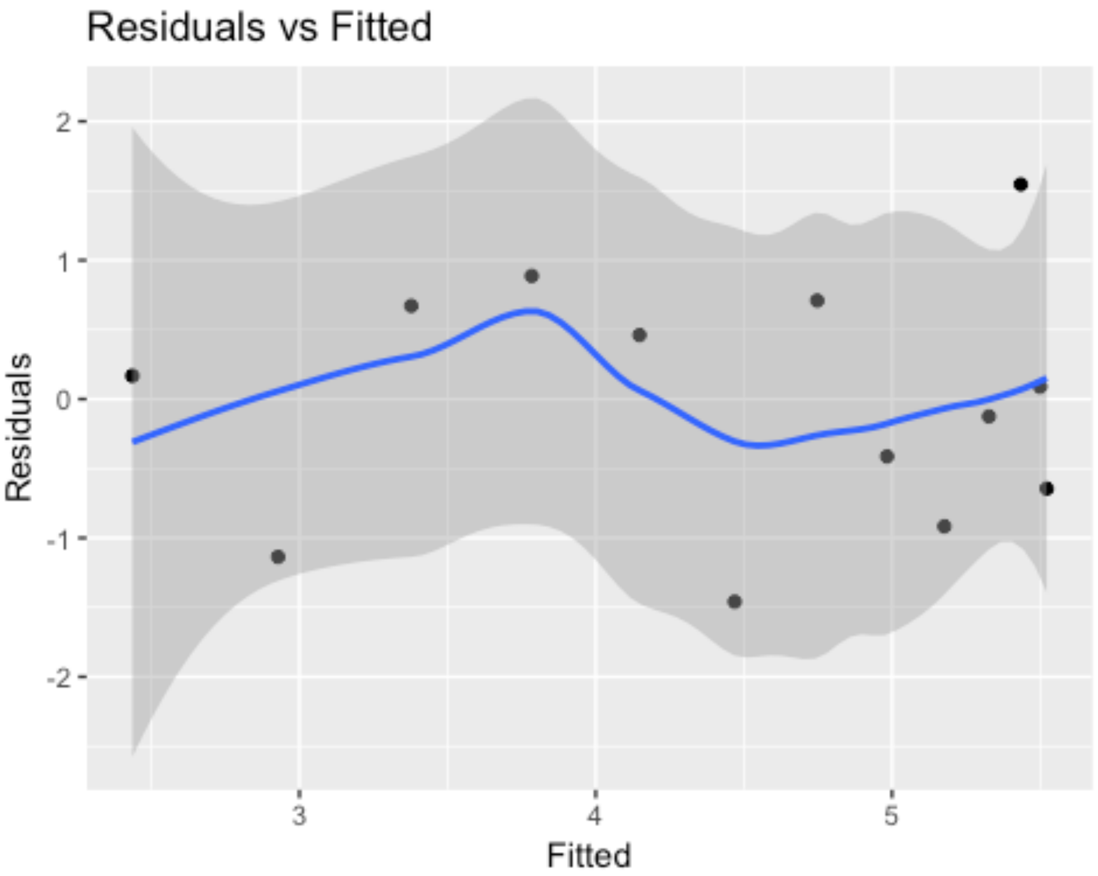
c) Now add a quadratic term in time ( ) and fit the model. Do the parameters describe the progression of the disease? Does this improve the model fit? Compare the residual plot to part b).

```
##
## Call:
## glm(formula = Cases ~ Time + time.sq, family = "poisson", data =
pois.dat2)
##
## 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 ***
## Time         0.556003   0.045780  12.145  < 2e-16 ***
## time.sq     -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
```

```
## (Intercept)        Time     time.sq
##   6.6956535   1.7436895   0.9788799
```


Residuals vs Fitted
Residuals
1.5
0.5
-0.5
-1.5
11
2
6
2.5
3.0
3.5
4.0
4.5
5.0
5.5
Predicted values
glm(Cases ~ Time + time.sq)



Discussion: The introduction of the second explanatory variable(t^2) certainly improved the model. First, the p values are extremely low such that these variables are statistically significant at the 0.001 level.The model suggests that there's a 74% jump in the AIDS cases with a 97.5% confidence interval. The accuracy of this model is supported by the residual deviance dropping from 80 to 9 and degrees of freedom from 11 to 10 which gives us a ratio of 0.9, very close to 1. This is further supported by the new residuals vs fitted plot. The values randomly dispersed around 0 with no recognizable pattern, the fitted line seems to be indicated better prediction.Hence, this is an improvement to our original model.

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

```
## [1] 166.3698

## [1] 96.92358
```

The first model has higher AIC where as the second model has significantly lower AIC, thus indicating that the sample prediction error and the quality of the second model is much better than the first model. This does confirm my position from part C after comparing the residuals vs fitted plots of both models.

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: Cases ~ Time
## Model 2: Cases ~ Time + time.sq
##   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
```

Discussion: As the chi-square test indicates, the second model(quadratic) does improve on the first model based on residual deviance reduction and the extremly low p-value for the explanatory covariates which indicates that the second model is statistically significant at the 0.001 level and the more accurate model. Hence, we accept the alternative hypothesis and reject the null hypthosis.

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

## Analysis of Deviance Table
##
## Model 1: default ~ student + balance
## Model 2: default ~ balance
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      9997     1571.7
## 2      9998     1596.5 -1   -24.77 6.459e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discussion: The AIC of the first model with the two explanatory covariate has lower AIC which suggests the first model is more accurate than the second model with the one predictor variable.Furthermore, this also suggest student may not be a good predicting covariate of who'll default on their credit card.

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

```
## [1] 0.02145588

## [1] 0.02193536
```

Discussion: The two models seem to be having the almost the error rate, at this point it's not starkly clear which model has better accuracy although the first model does have the edge with a slightly lower error rate but not much different.

c) LOOCV

```
## [1] 0.0008405851

## [1] 0.001071851
```

d) 10-fold cross-validation.

```
## [1] 0.02138929

## [1] 0.02175712
```

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.

Discussion: Splitted the data into 70/30 between the training and validation data sets gave us extremely close error rates.We performed further examination by subjecting both models to the train/validation set approach for which gave us: MSE1 = 0.02145588 MSE2 = 0.02193536d considering these two slightly different error rates, I'd recommend model 1 as the more accurate model.

We continued on by trying out two other validation methods. The loocv adjusted prediction errors for model1 and model2 respectively were 0.0008405851, 0.001071851 which again indicates that our model selection in 3b was correct.

Finally, we ended up using the 10-fold cross-validation approach setting K=10. The results were respectively 0.02138929,0.02175712.In conclusion, although the two models are similarly close in accuracy,however, model 1 with student as a covariate along with balance is the better model with lower predictions error as all the validation methods we performed indicated.

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, family = binomial(), data =
train.data)
##
```

```
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -1.345  -1.188   1.074   1.164   1.326  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.03222    0.06338   0.508    0.611
## Lag1        -0.05562    0.05171  -1.076    0.282
## Lag2        -0.04449    0.05166  -0.861    0.389
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1383.3  on 997  degrees of freedom
## Residual deviance: 1381.4  on 995  degrees of freedom
## AIC: 1387.4
## 
## Number of Fisher Scoring iterations: 3
```

```
## 
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + 
##     Lag1 * Lag2 + Lag1 * Lag3 + Lag1 * Lag4 + Lag1 * Lag5, family = 
binomial(), 
##     data = train.data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.6559  -1.1871   0.9298   1.1617   1.6054  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept)  0.032417   0.063568   0.510    0.610  
## Lag1        -0.054447   0.053976  -1.009    0.313  
## Lag2        -0.047600   0.052247  -0.911    0.362  
## Lag3         0.001429   0.052209   0.027    0.978  
## Lag4         0.003842   0.052336   0.073    0.941  
## Lag5        -0.001769   0.051752  -0.034    0.973  
## Lag1:Lag2   -0.006555   0.035474  -0.185    0.853  
## Lag1:Lag3    0.043823   0.032503   1.348    0.178  
## Lag1:Lag4   -0.016897   0.028637  -0.590    0.555  
## Lag1:Lag5   -0.062438   0.033833  -1.845    0.065 .
## ---
## 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: 1375.7  on 988  degrees of freedom
## AIC: 1395.7
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
## Analysis of Deviance Table
##
## Model 1: Direction ~ Lag1 + Lag2
## Model 2: Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Lag1 * Lag2 +
##     Lag1 * Lag3 + Lag1 * Lag4 + Lag1 * Lag5
##   Resid. Df Resid. Dev Df Deviance
## 1       995     1381.4
## 2       988     1375.7  7   5.6706
```

```
## [1] 0.2483177
```

```
## [1] 0.2487672
```

Discussion: I started out by building two models that had different number of explanatory covariates. The first model a simple model that had only three explanatory covariates consisting of Lag1,Lag2,Lag3 from the subsetted smarket dataset. The second model was more complex. It included all the Lag explanatory covariates and I threw additional interactions between the covariates. Unfornately, both models' summary isn't so promising as neither of them have any significant p-values. However, for the sake of model selection I moved on to cross validate both models and comapred their mean square errors. It turns out the simple model has an error rate of 0.24832, where as the complex model has an error rate of 0.224877. Also, the AIC for the simple model was 1387 and when I ran the compplex model the AIC jumped up to 1395 though the ratio of residuals to degrees of freedom did improved significantly from 1.99 for the simple model to 1.39 for the complex model.

Inconclusion, I'd recommend moving forward with the simple model based on its lower error rate and AIC.

Citation

"A Handbook of Statistical Analyses Using R, third Edition" by Everitt and Hothorn
R Graphics Cookbook" by Winston Chang published through O'Reilly (Basic guide to Grammar of Graphics in R)
www.stackoverflow.com

http://r-statistics.co/Linear-Regression.html