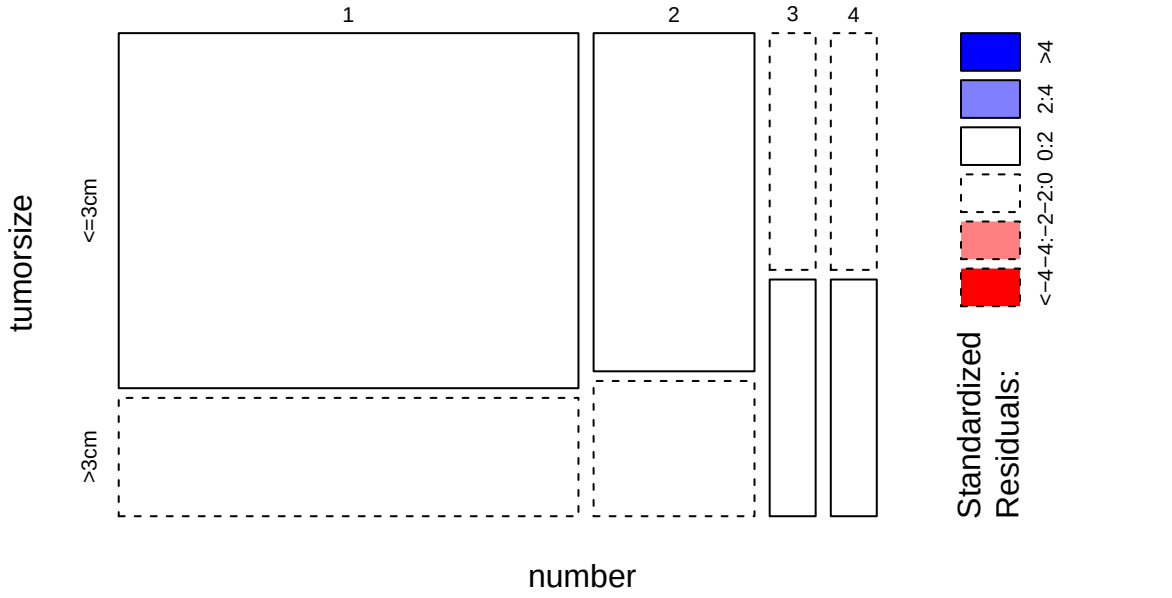
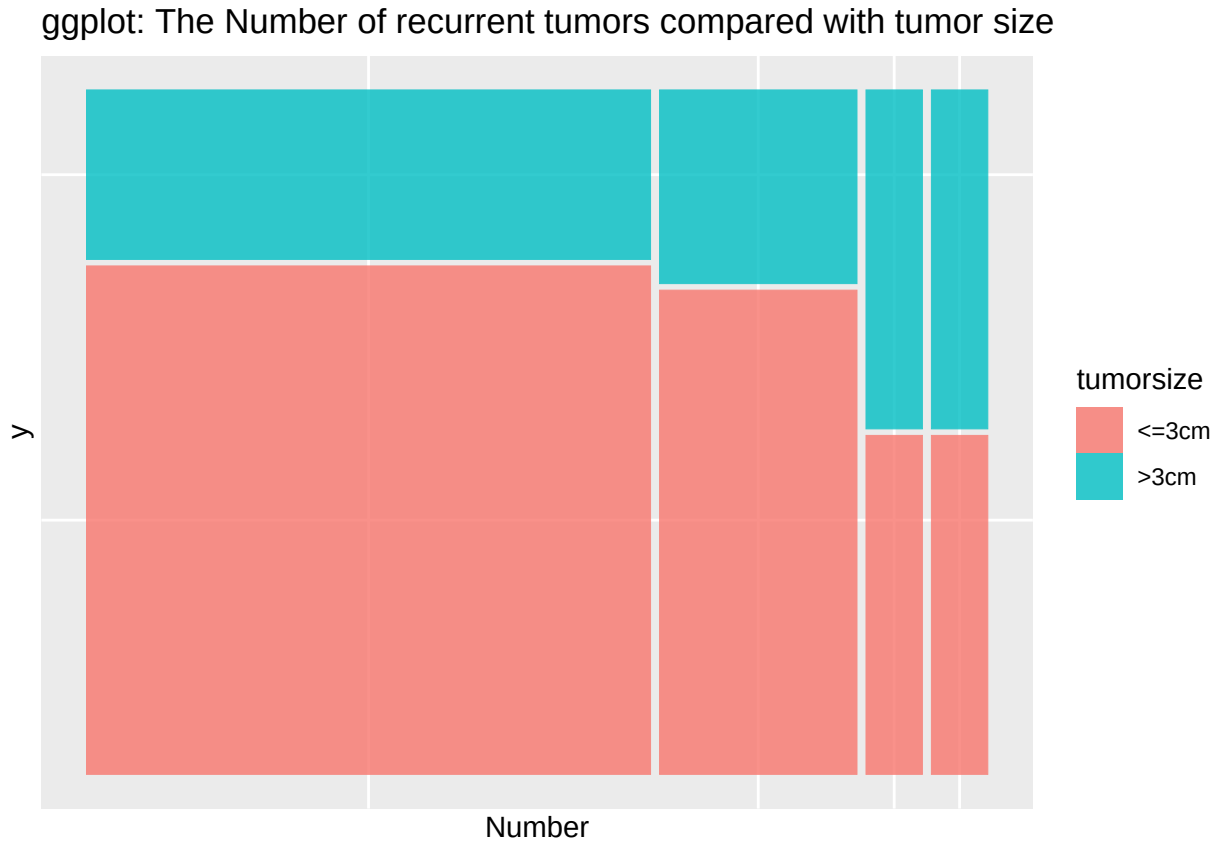
Analysis of Bladder Cancer from HSAUR3

Yamuna Dhungana

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.)





From the graphical summary, we find out that the relationship between the tumor size of the bladder cancer and the number of recurrent tumors. The graph shows that the number of less than or equal to three is more than the number of tumors greater than three. The number of tumor 1 is more comparatively. Like-wise the number of three and four tumors is less than one and two.

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
```

When we exclude the time variable from our model, the intercept adds significance. Whereas, tumor size does not add significance to our model. The value of the AIC of the model is 87.191. Dataset bladder cancer is slightly positively skewed. The median of the model is close to zero therefore we can say that the model is not biased in one direction.

```
##
## 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
```

When we see the relationship between the number of tumors, tumor size, and the time we find out that the model is insignificant. We also can see AIC is 88.568 which is greater than the previous model.

```
##
## Call:
## glm(formula = number ~ tumorsize + time + time * tumorsize, family = poisson,
##     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
## tumorsize>3cm        0.46717    0.66713   0.700    0.484
## time                 0.02138    0.02418   0.884    0.377
## tumorsize>3cm:time -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
```

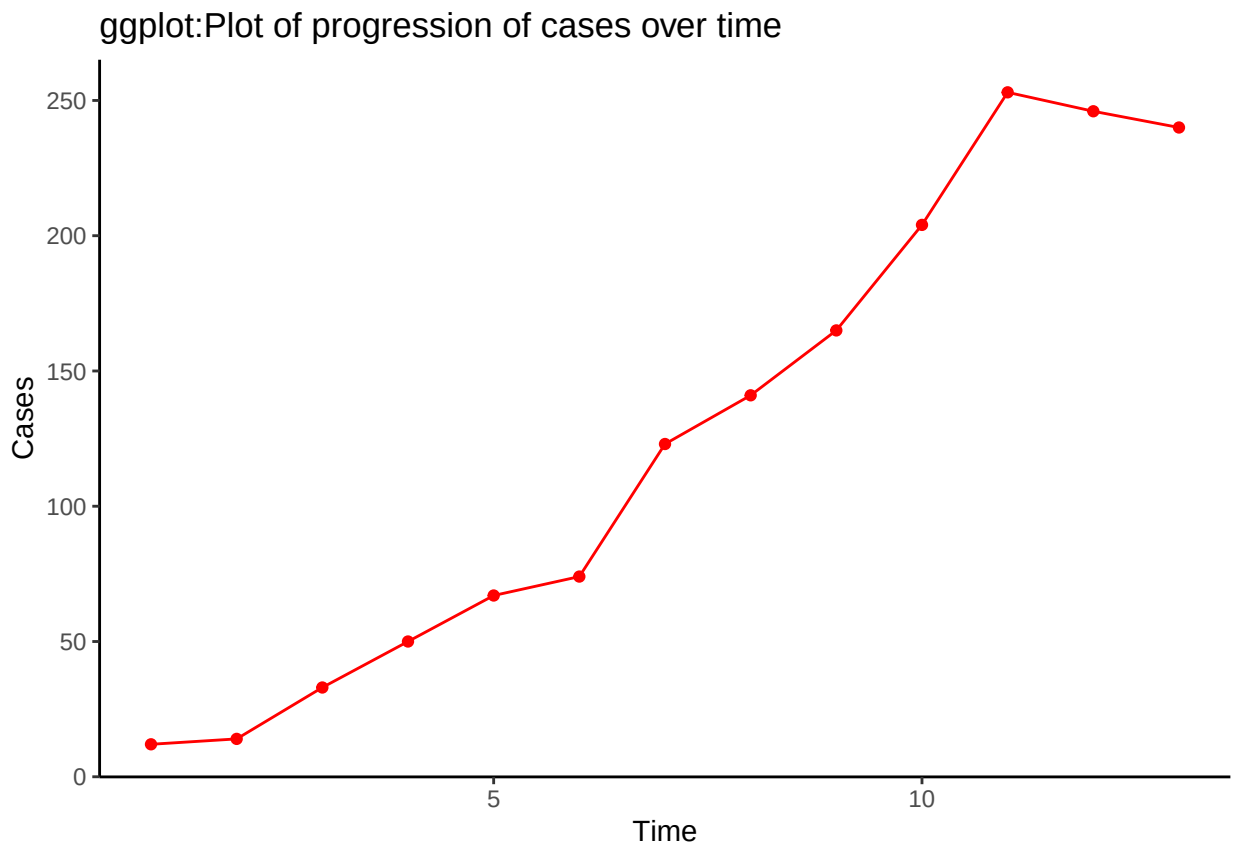
```
##
## Number of Fisher Scoring iterations: 4
```

When we see the interaction between the number of tumors, tumor size, and the time we find out that the model is insignificant. We also can see AIC is 90.377 which is greater than the previous models. Therefore, I find the first model, which is Poisson regression with a number and the tumor size more relatable with the first part of the question.

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.

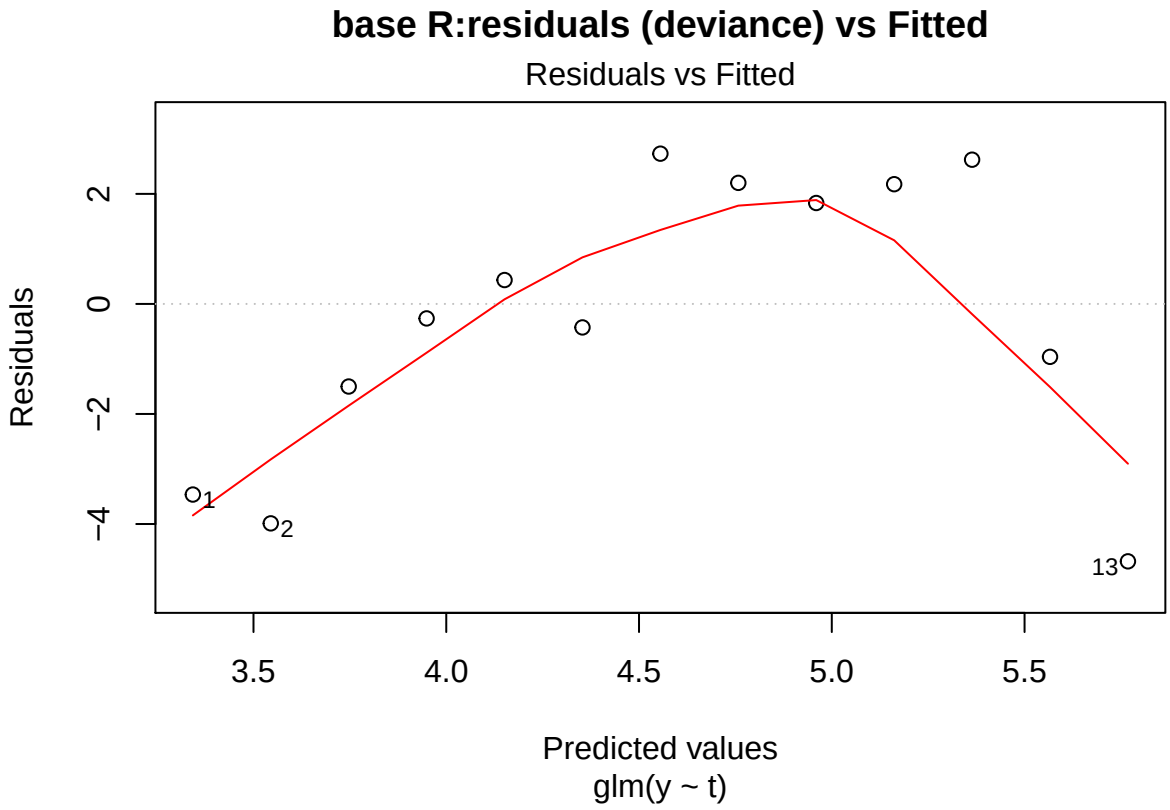


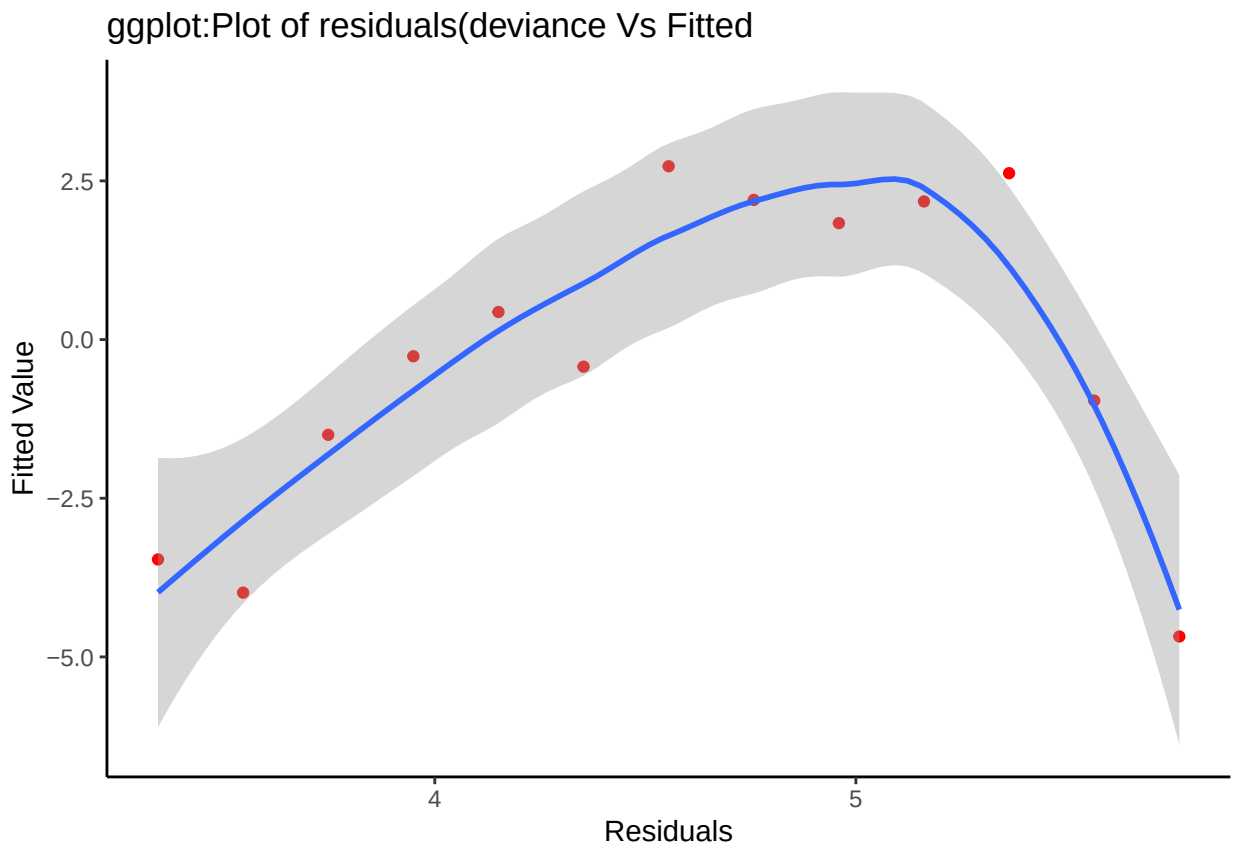
The above graphs show the relationship between the time and the progression of the AID cases in Belgium. In the first year, the number of cases was relatively stable. The graphs show a similar steep between 1982 to 1985 and 1987 and 1991. There is an increase in the number of cases. From 1986 to 1987, the cases of AIDS increased, which is approximately 75 to 175 cases. The last year that is 1991 to 1993, the cases began to decrease from the peak. The highest number of cases seen in the 13 years was 250.

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 = y ~ t, family = poisson, data = my_df)
##
## 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
```






```
## (Intercept)           t
##   23.117491    1.223996
```

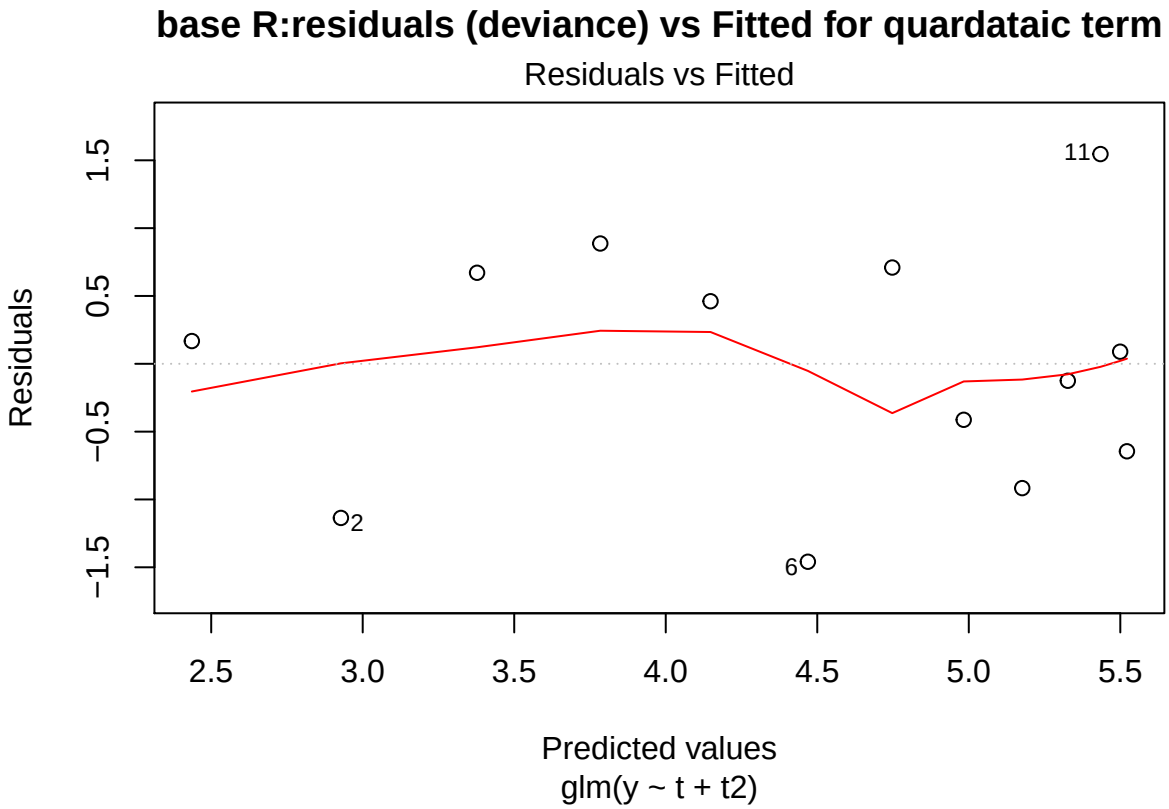
```
##                  2.5 %    97.5 %
## (Intercept) 19.789547 26.894433
## t            1.205624  1.242922
```

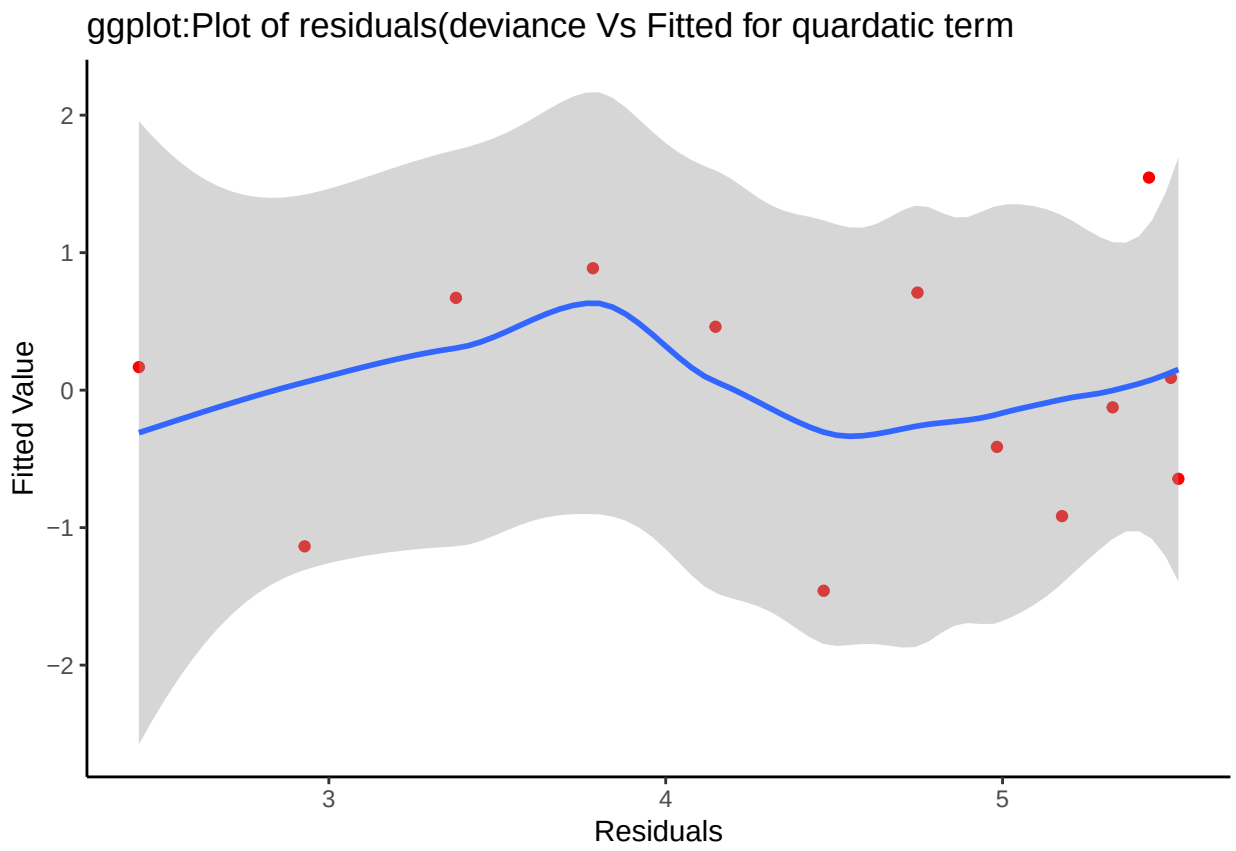
Both (b0) and (b1) are statistically significant from zero. The residual and predicted plot is normally distributed which says that our assumption is correct. We also can find out the one outliers.

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).

```
##
## Call:
## glm(formula = y ~ t + t2, family = poisson, data = my_df)
##
## 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 ***
## t            0.556003   0.045780  12.145  < 2e-16 ***
## t2          -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)           t          t2
##   6.6956535   1.7436895   0.9788799
```

```
##                  2.5 %    97.5 %
## (Intercept) 4.5976982 9.5678396
## t           1.5965138 1.9104525
## t2          0.9737254 0.9839292
```

The model proves that all the variables are statistically significant.The progression of AIDS cases decreased when we performed the quadratic model.

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

```
## AIC for the model-1
```

```
## [1] 166.3698
```

```
## AIC for the model-2
```

```
## [1] 96.92358
```

The model is always better if the value of AIC is less. Since the value of AIC is less for model-2 than model-1. Therefore, model-2 is better. Yes,the model improved from the first.

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)?

```
## Analysis of Deviance Table
##
## Model 1: y ~ t
```

```
## Model 2: y ~ t + t2
##   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
```

The p value of the model 2 is less than 0.05 which is statistically significant. Therefore model-1 improves by adding the quadratic term.

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

Compare the models using the following four model selection criteria.

a) AIC

```
## AIC of the first model is

## [1] 1577.682

## AIC of second model is

## [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
```

The AIC of the first model is smaller than the second model. Therefore the model-1 is better.

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

```
## mean square error for first model

## [1] 0.02152873

## mean square error for second model

## [1] 0.02178644
```

I split the data into a 70:30 ratio, as the test and the training data. The mean squared error for model-1 is 0.0215. Likewise, the mean squared error for the model-2 is 0.0217. Based on the MSE, we can choose model-1.

c) LOOCV

```
## [1] 0.0267

## [1] 0.0275

##  LOOCV for the first model

## [1] 0.0267
```

```
## LOOCV for the second model

## [1] 0.0275
```

LOOCV error is adjusted for bias and we still want the smallest prediction errors.From the model-1 the LOOCV is 0.02267 and the LOOCV for model-2 is 0.0275. Therefor model-2 is better.

d) 10-fold cross-validation.

```
## 10-fold cross-calidationmodel-1

## [1] 0.0268

## 10-fold cross-calidation for model-2

## [1] 0.0276
```

Using K=10 for the 10-fold cross-validation approach, for model 1, the CV error rate is 0.0266 and for model 2 is 0.0275, From most of the model we find out that the model-1 is better.

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.

Table 1: A table for errors.

	AICs	T/v set approach	LOOCV	K-fold
Model-1	1577.68	0.02	0.03	0.03
Model-2	1600.45	0.02	0.03	0.03

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.

```
## Analysis of Deviance Table
##
## Model 1: Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume
## Model 2: Direction ~ Lag1 + Lag2 + Lag1 * Lag3 + Lag1 * Lag5 + Lag3 *
##     Lag4 + Lag3 * Lag5
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       991     1381.1
## 2       988     1373.8  3   7.2708  0.06375 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 0.2507619

## [1] 0.248777

## 10-fold cross-validationmodel-1

## [1] 0.5230461

## 10-fold cross-validation for model-2

## [1] 0.488978
```

Here, I have two models. We find out that both the model is not statistically significant. We considered statistically significant. If the value of p is less than or equal to 0.05. Here the model-2 has the p-Value of 0.06, which is statistically insignificant. The error of the model-1 is through mean squared error is 0.25, and model-2 is 0.24. Likewise, the error of 10-fold cross-validation for model-1 is 0.51 and model-2 us 0.48.