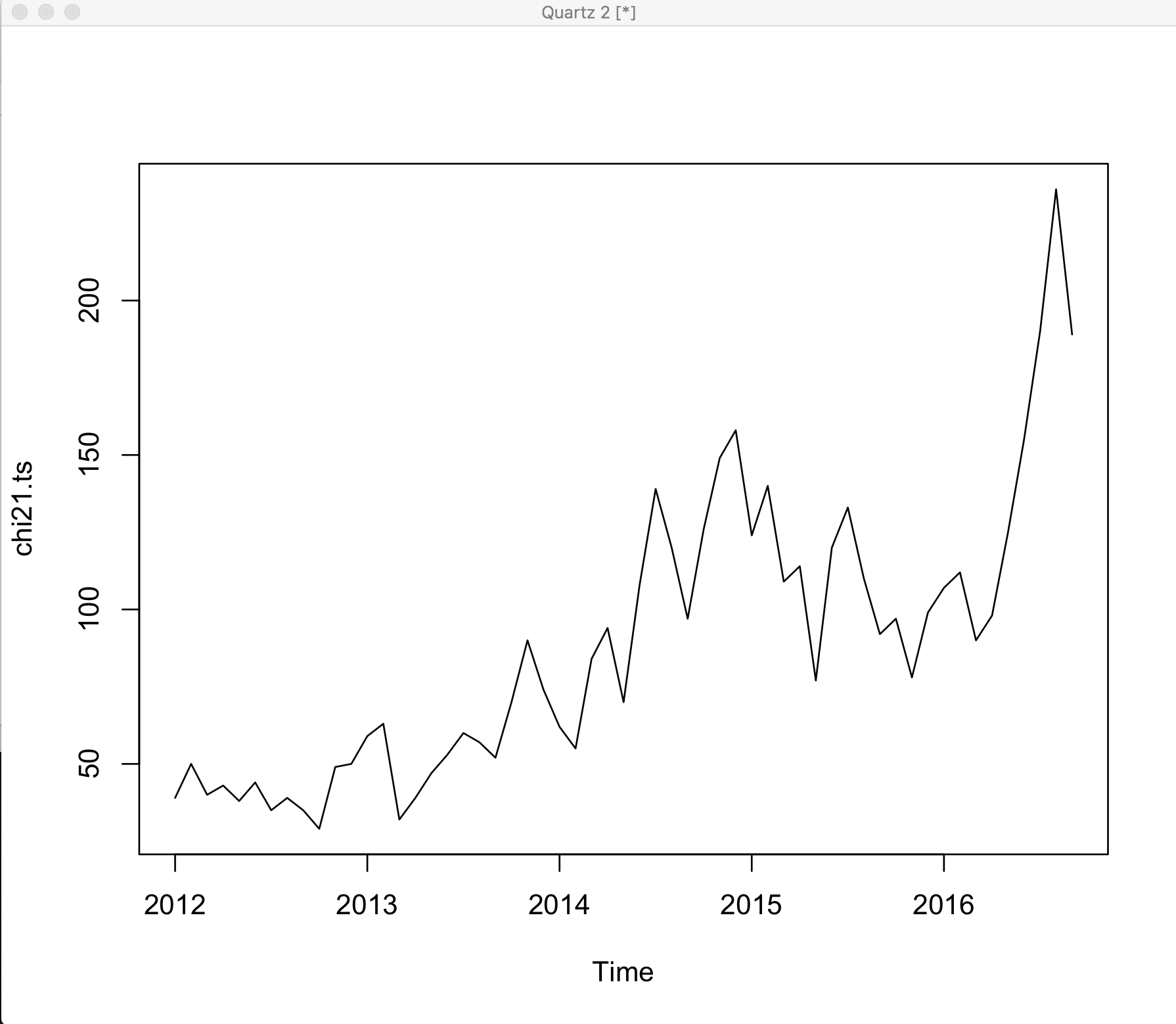
**Number of street crimes in Chicago.**

In this project, our main is to apply different techniques to model the time series data for ‘Number of street crimes in Chicago’ in order to evaluated their fit and choosing the best model which can forecast values for October and November 2016.

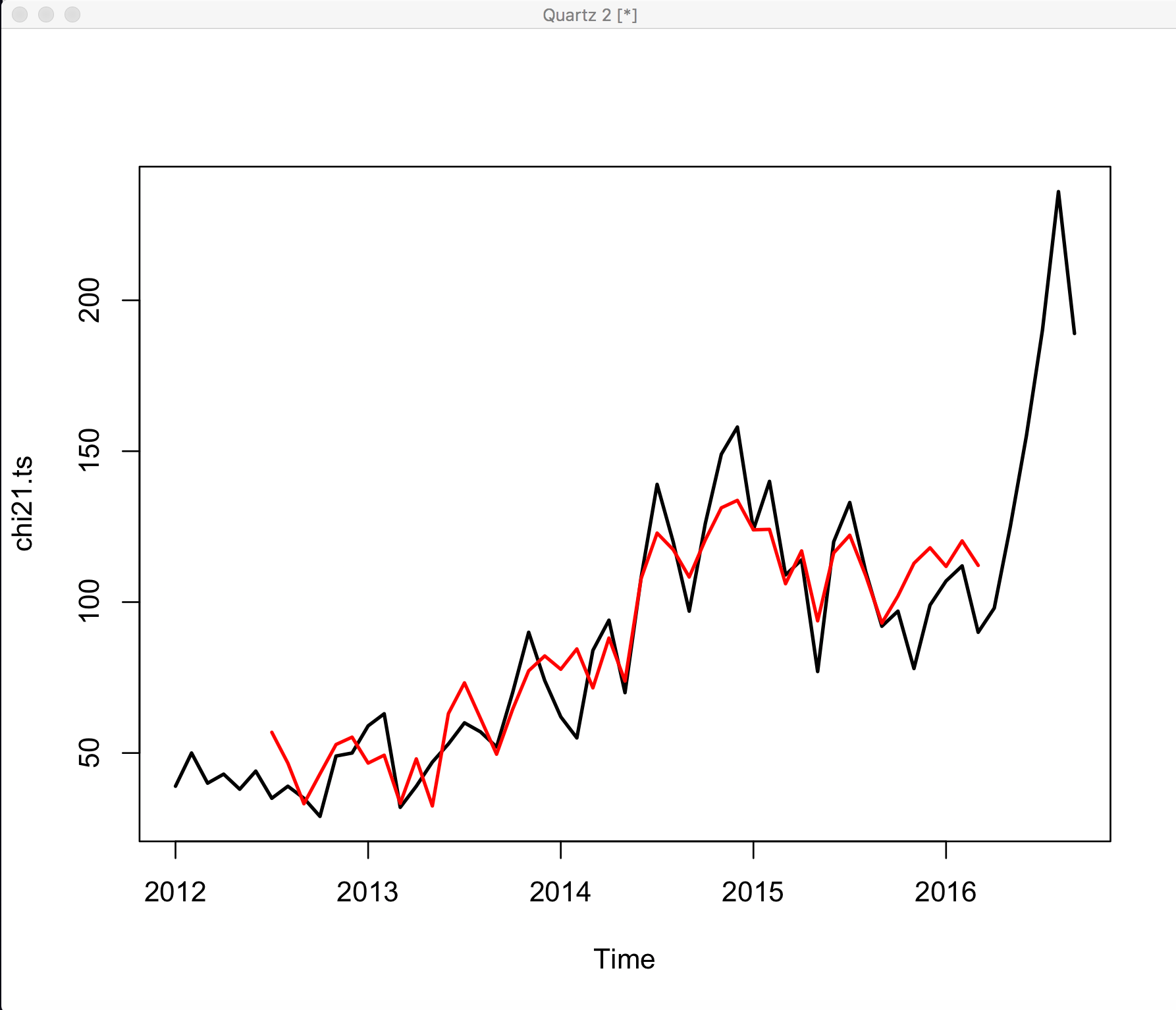
Below is the given data:



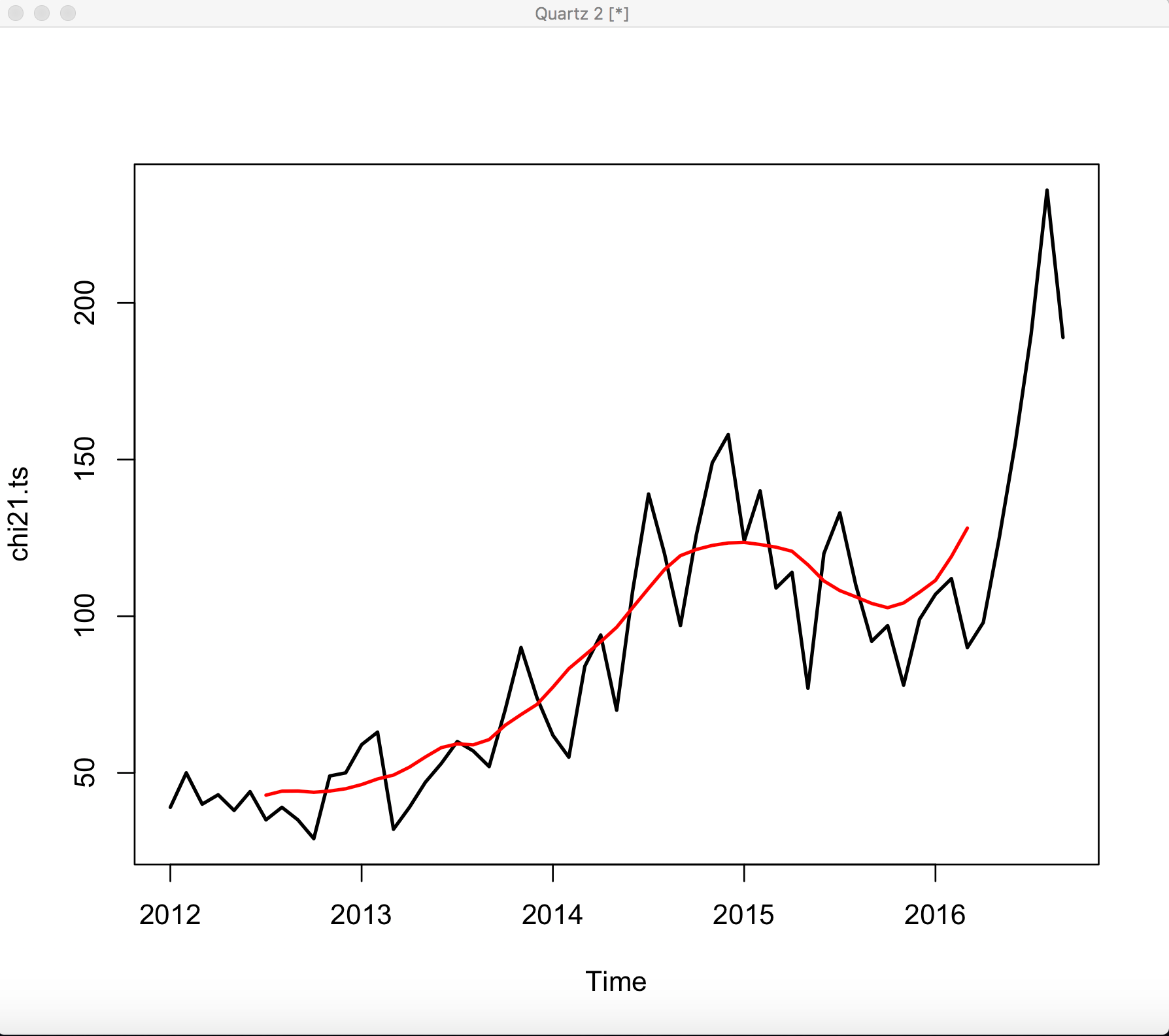
**Model 1: Using moving average decomposition model**

Starting with the moving average decomposition model which is also smoothing model. This model has two types: additive and multiplicative. We will apply both the models and then compare their results. The models applied are indicated with red line on the plotted data:

**Decomposition moving average- Additive**



**Decomposition moving average- Multiplicative**



From the above plots, we cannot infer which models fits the best. That’s why we will use the mean root square error(RMSE) to find out the best fit model.

1. Finding the best model from the above two.

By comparing the RMSE of both the models, we can conclude which model fits the data best. Below is the RMSE value for both the models:

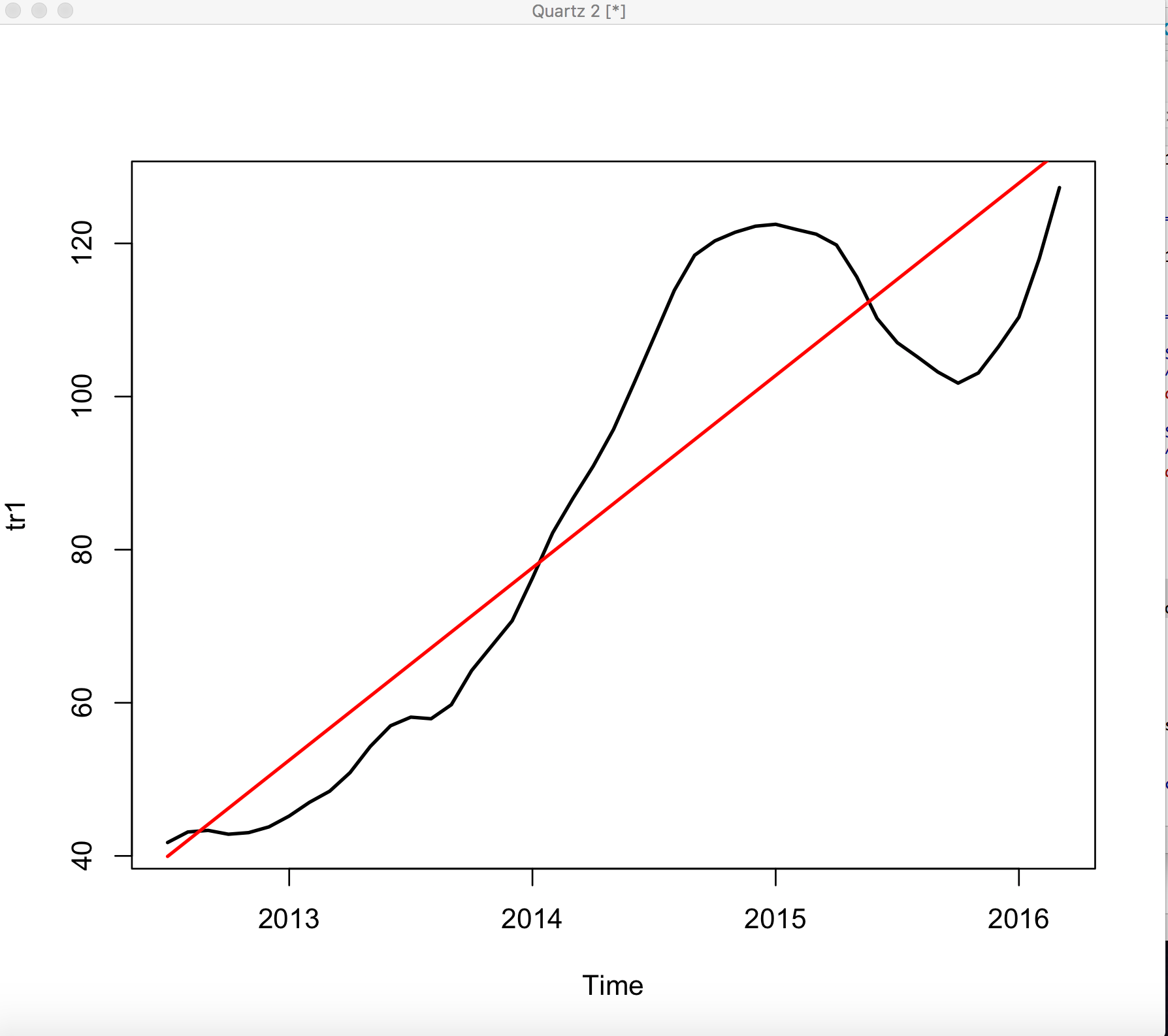
|  |  |
| --- | --- |
| RMSE additive | 12.91045 |
| RMSE multiplicative | 0.9826991 |

As we know that the model with lower RMSE is better fit for the data. In this case the multiplicative model has lower RMSE value it means it fits the data better than the additive model.

The other reason is that in our time series the season effect increases as the trend increases which can be seen in the above plotted graph.

1. Fit trend line to smoothed deseasonalized series in order to get forecasting.

**Linear Regression- Decomposition Trend**



1. Obtaining forecast for Oct and Nov 2016 using the multiplicative model.

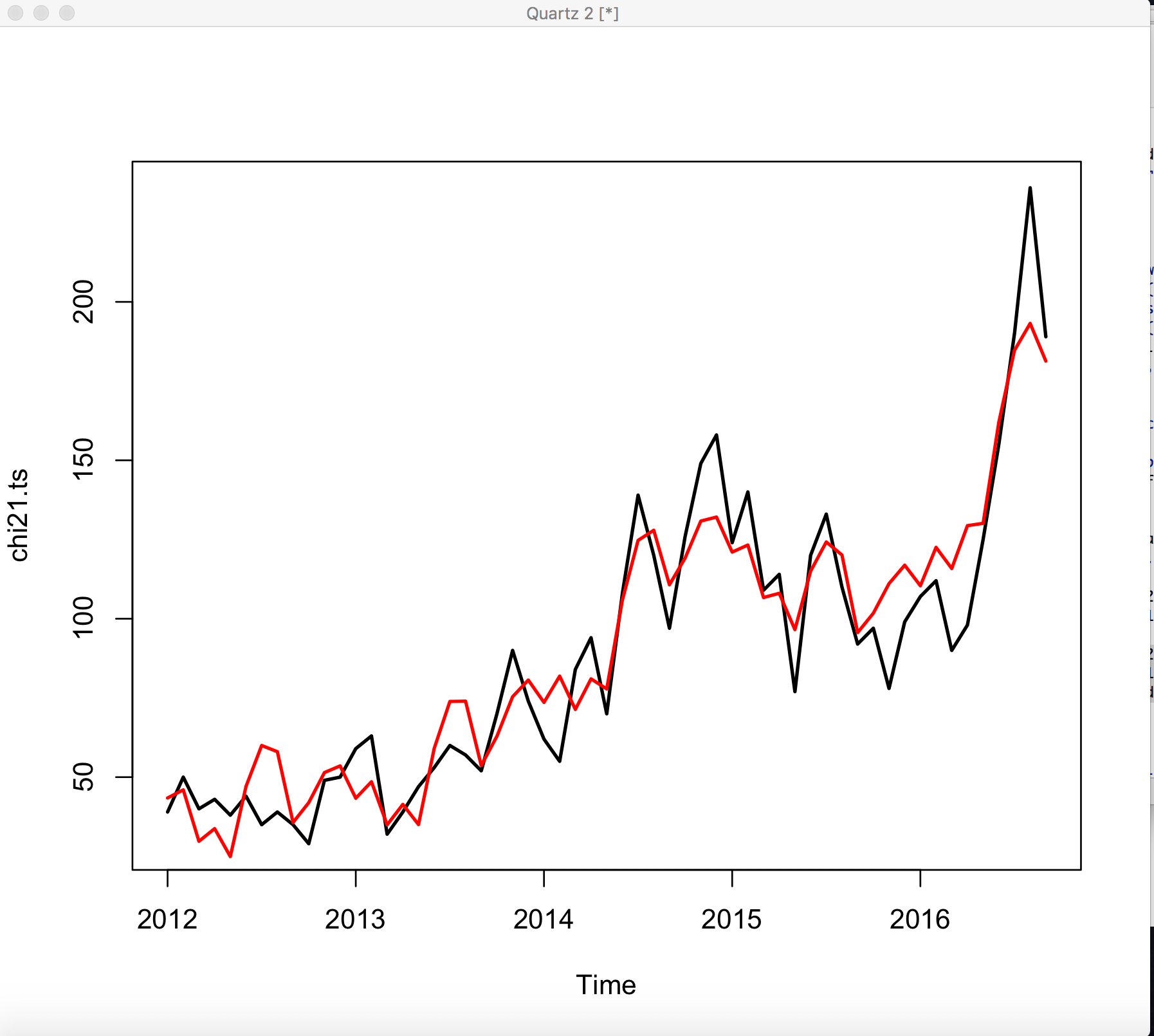
From the above liner model, we have got the forecast values for Oct and Nov 2016.

|  |  |
| --- | --- |
| Oct 2016 | 134.1644 |
| Nov 2016 | 136.2500 |

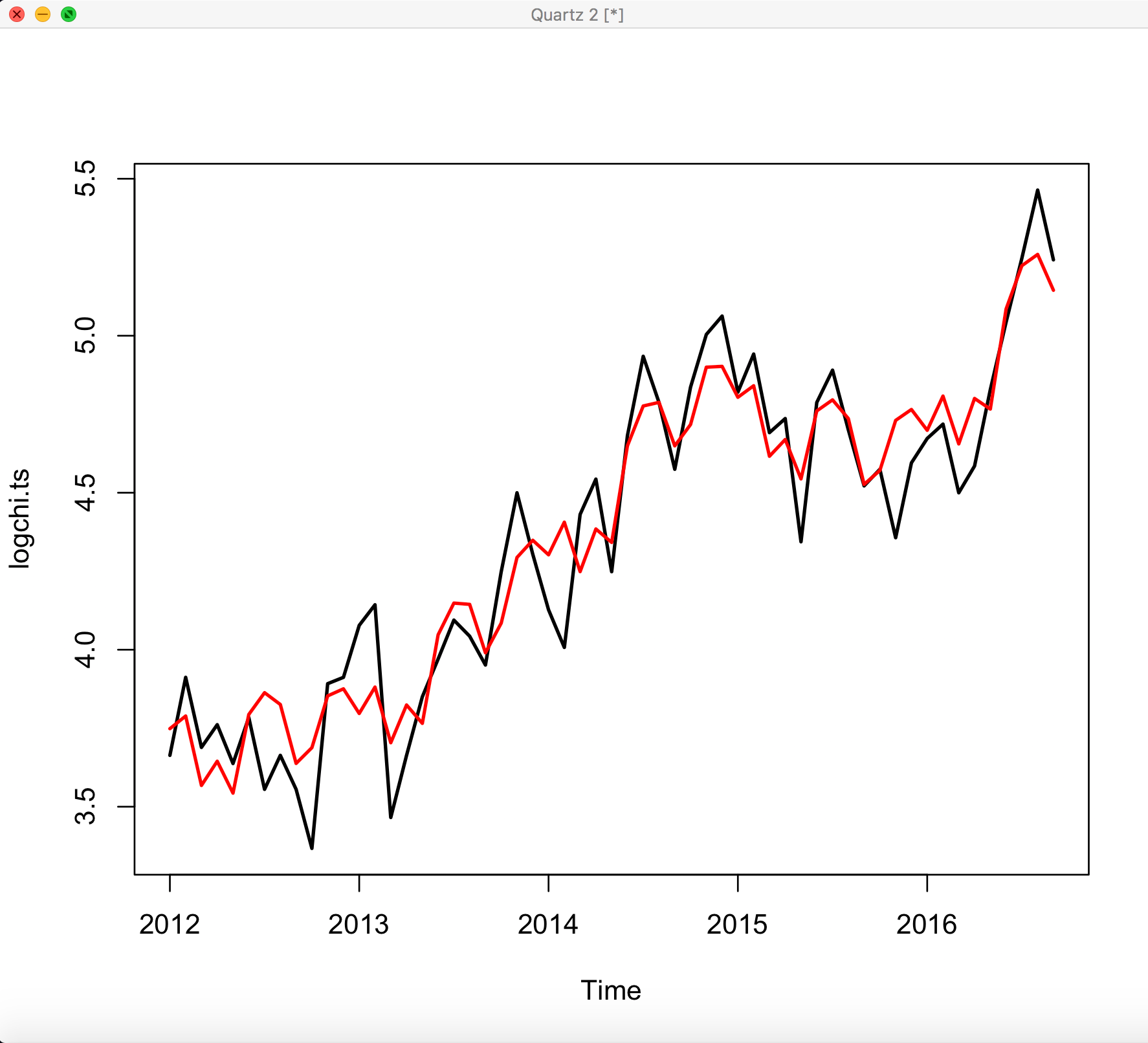
**Model 2: Using loess decomposition model**

Using another smoothing method which uses a weighted regression technique called loess, we will plot the additive and multiplicative model using red line on top of data.

**Loess Additive**



**Loess Multiplicative**



1. Finding the best model from the above two.

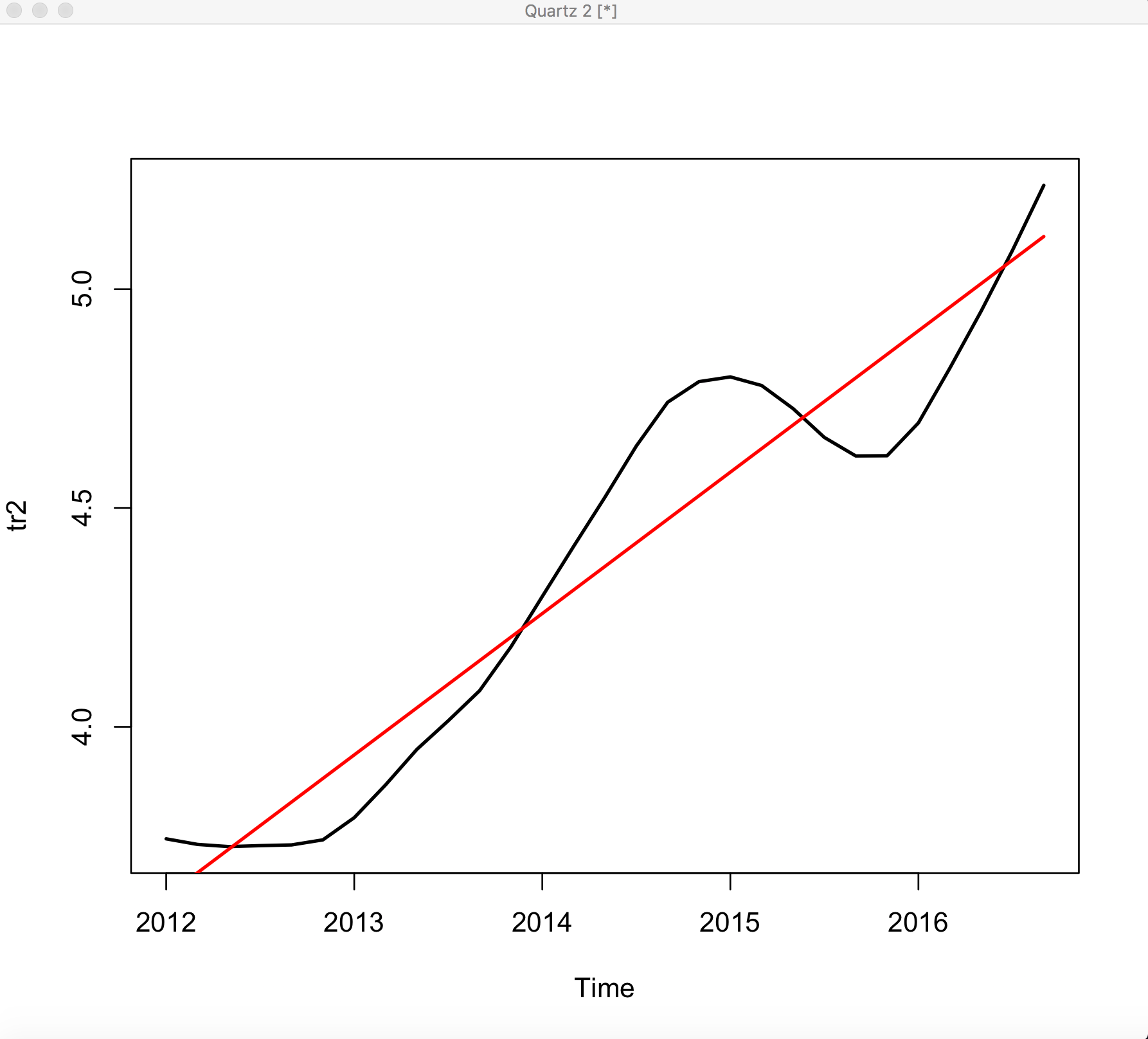
Once again, to get the best model we will calculate RMSE for both the models.

|  |  |
| --- | --- |
| RMSE additive | 14.45475 |
| RMSE multiplicative | 0.1532234 |

Again, as we know that the model with lower RMSE is better fit for the data. In this case the multiplicative model has lower RMSE value it means it fits the data better than the additive model. And the other reason is that in our time series the season effect increases as the trend increases which can be seen in the above plotted graph.

1. Fit trend line to smoothed deseasonalized series in order to get forecasting.

**Linear Regression- Loess Trend**



1. Obtaining forecast for Oct and Nov 2016 using the multiplicative model.

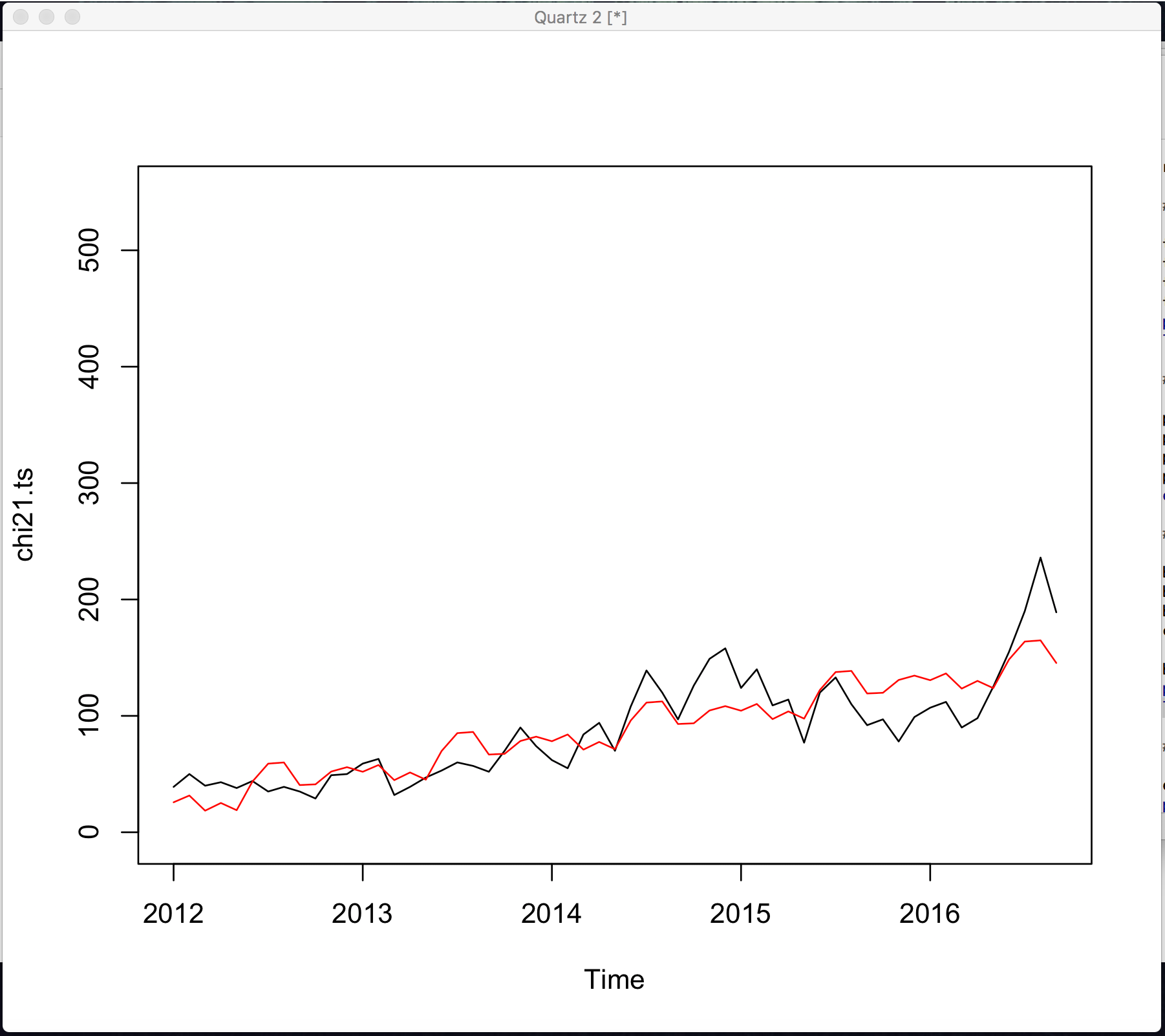
From the above liner model, we have got the forecast values for Oct and Nov 2016.

|  |  |
| --- | --- |
| Oct 2016 | 146.3015 |
| Nov 2016 | 150.2766 |

**Model 3: Using regression model**

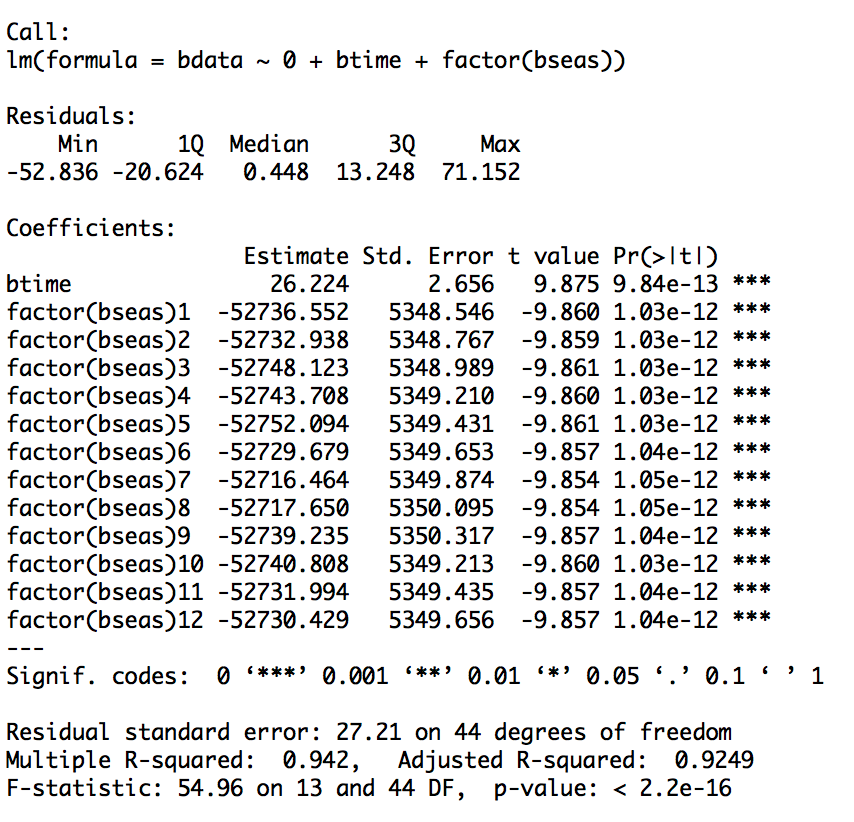
Here, we will use regression model as it also uses season and trend as input variables. As the errors are correlated in time series regression, it is not same as the linear regression. The regression is shown below using red line on the data plotted:

**Regression Model**



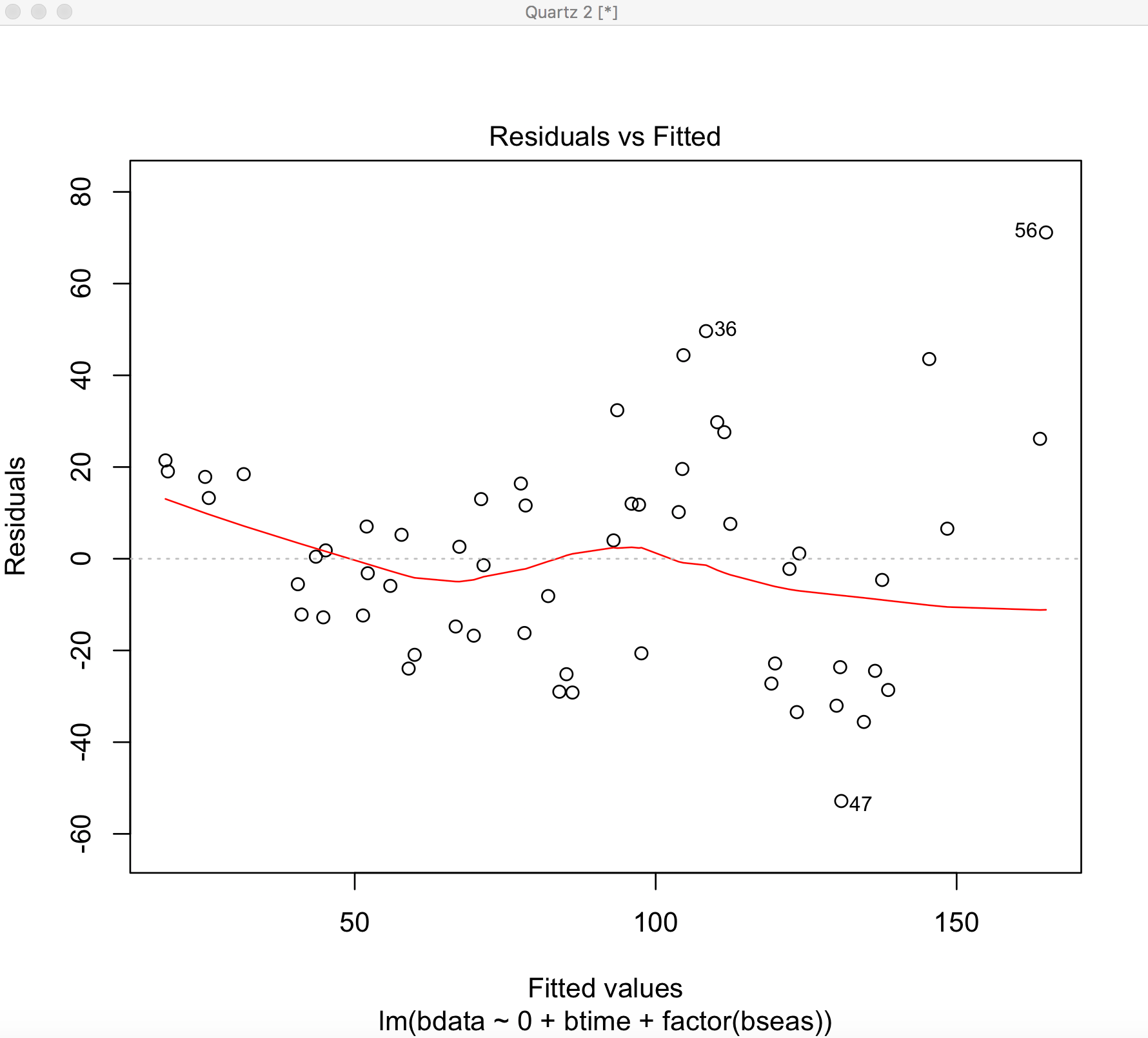
1. Determining the fit of the model:

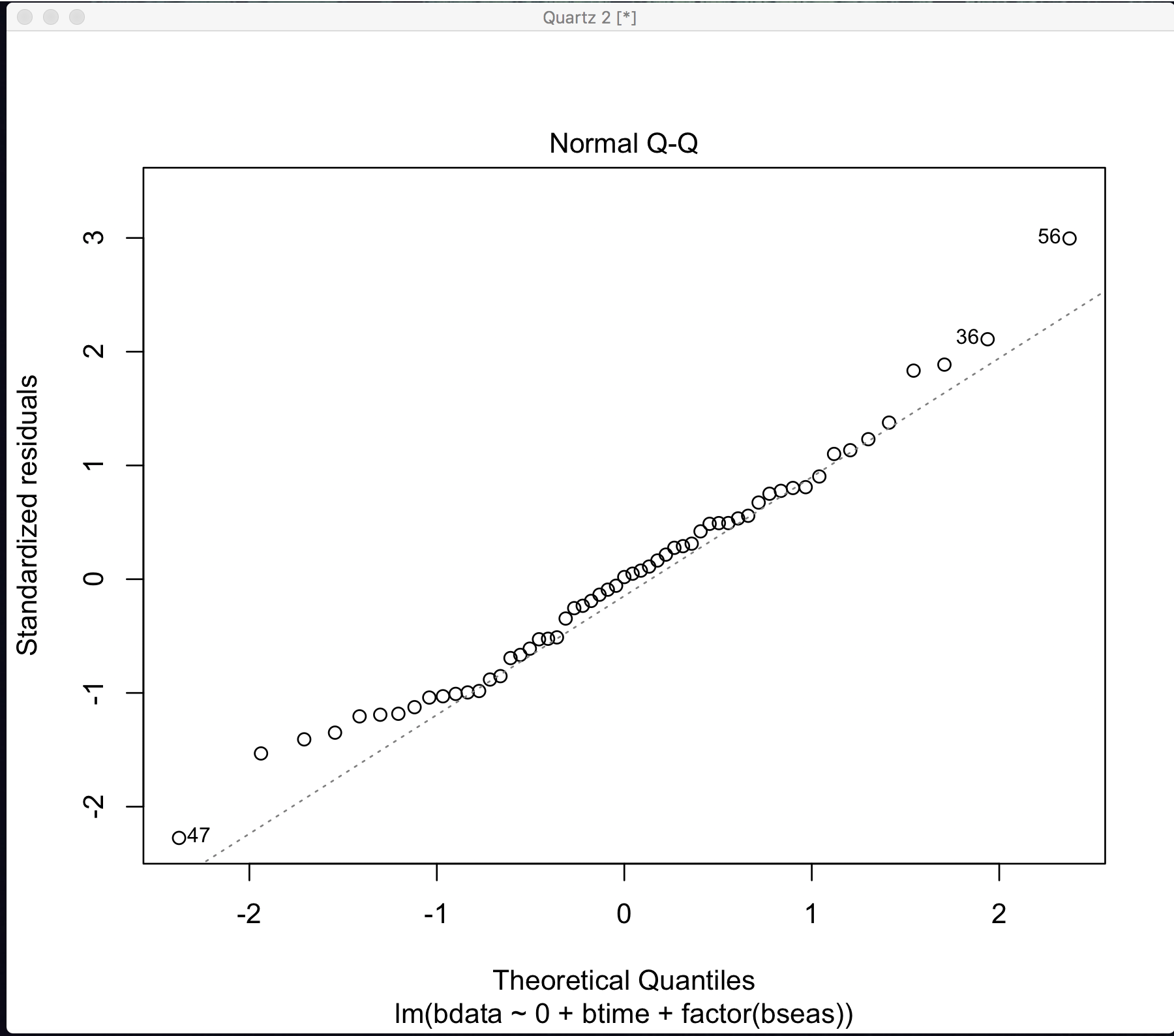
In order to validate the linear regression, we will evaluate some information. By checking the coefficients, R squared and standard error:

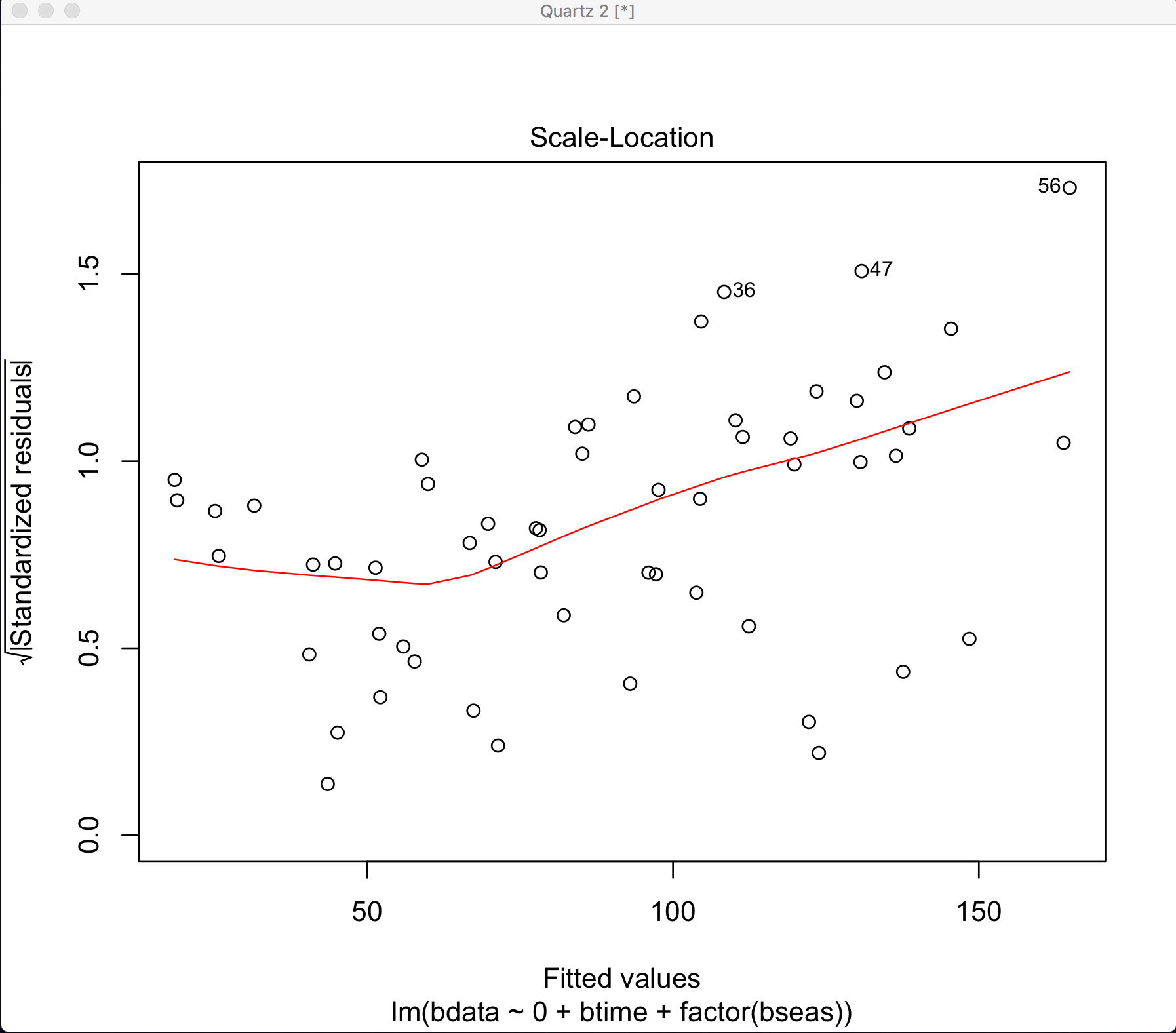


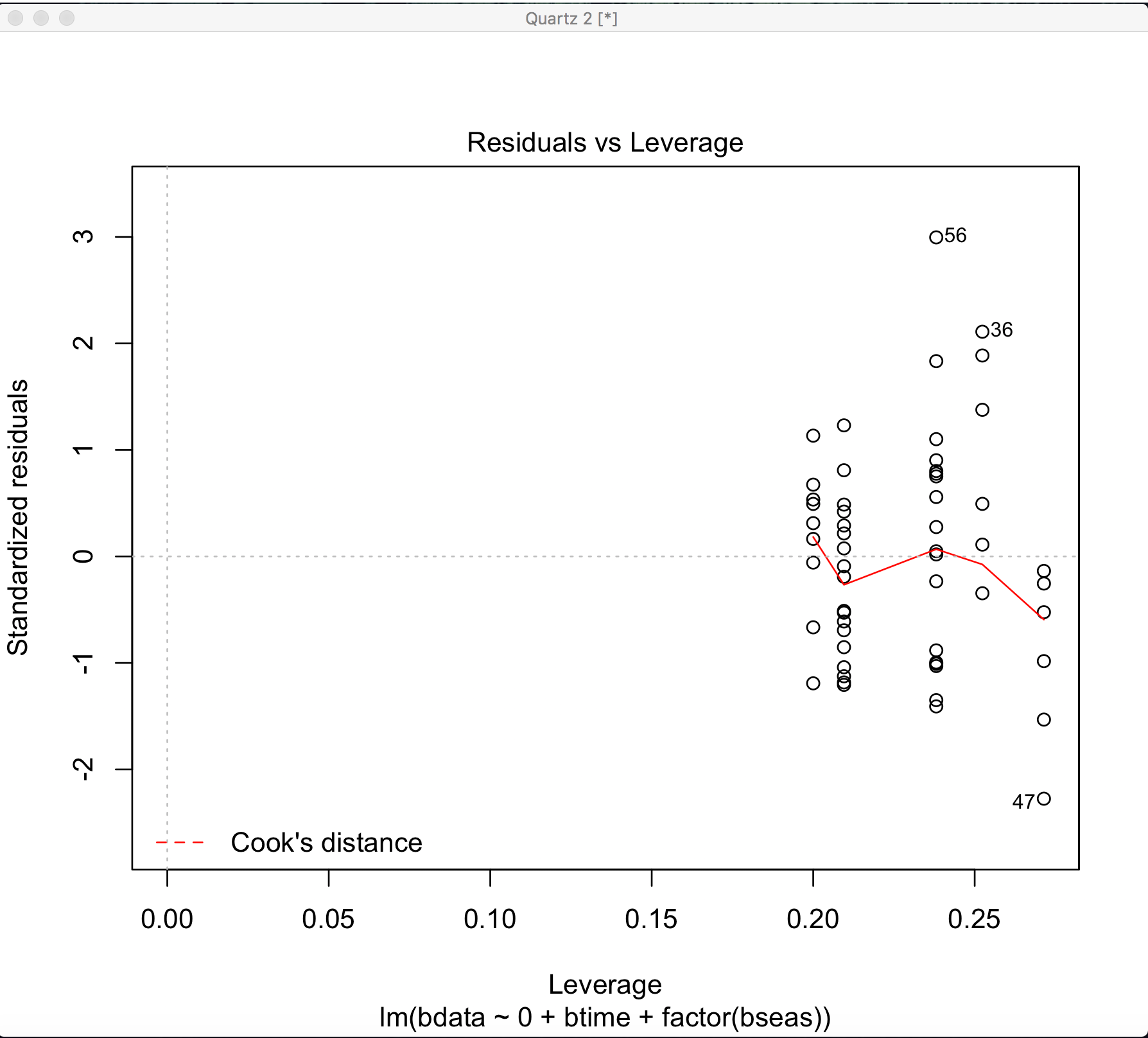
R2 is 0.942 which means 94% of variance can be explained by the above model. As the p-value is very low in can be inferred that all predictors are significant. We have to cautions while interpreting the information as some assumptions might not meet. The diagnostic plot which needs to be interpreted in order to validate:

* Equal variance or errors
* Iid errors
* Normality of errors
* Linear relationships between dependent variables & predictors





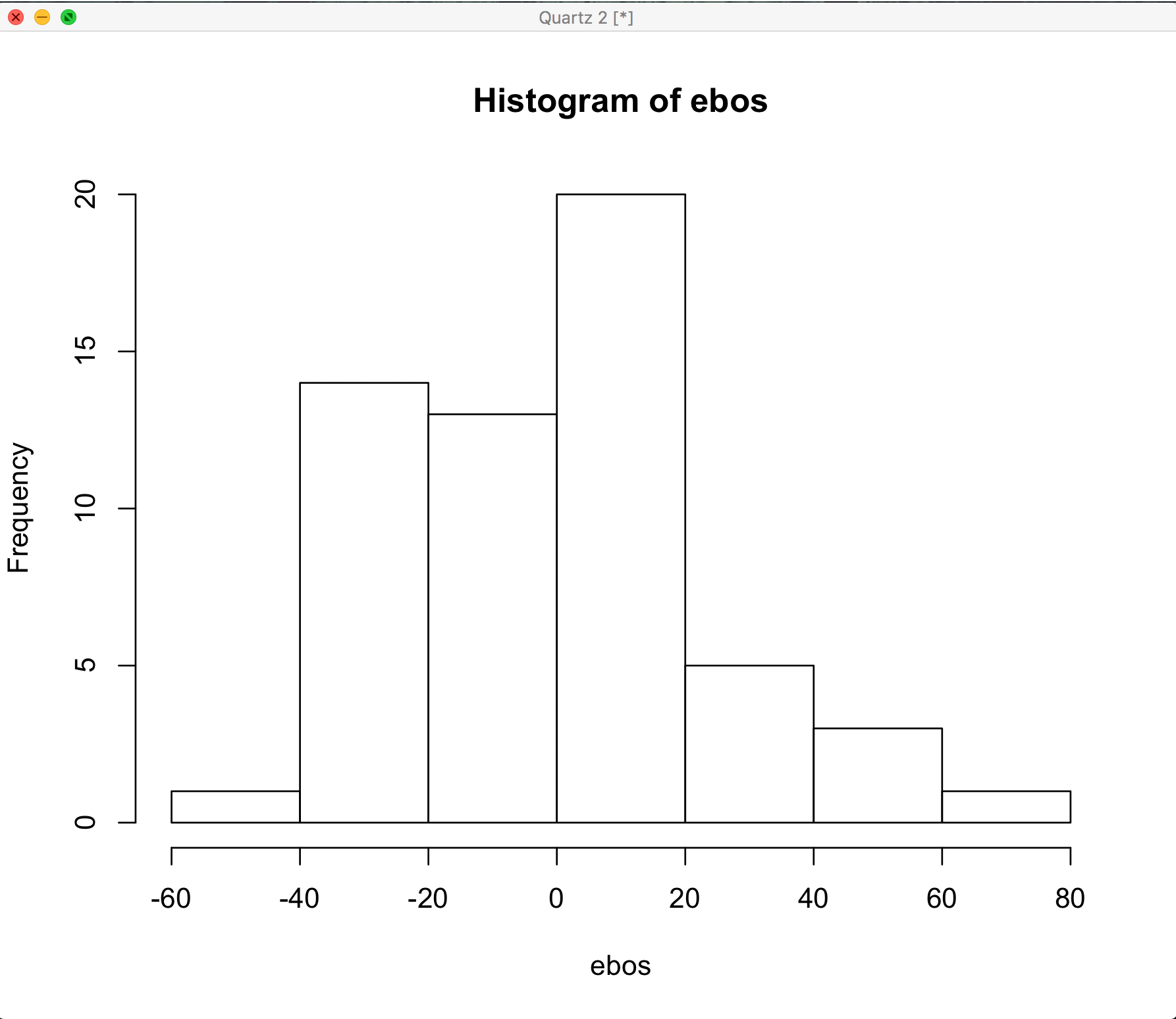


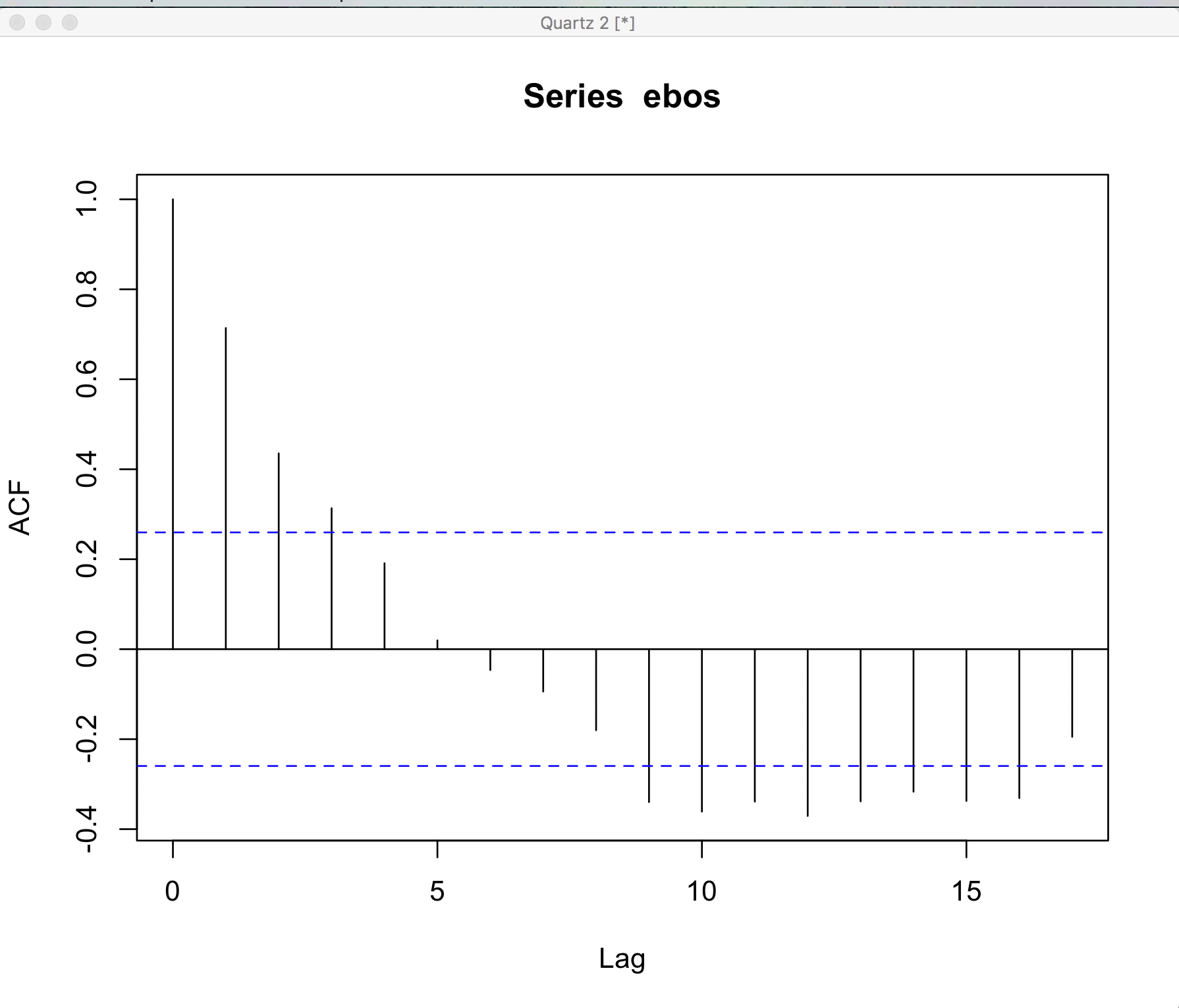


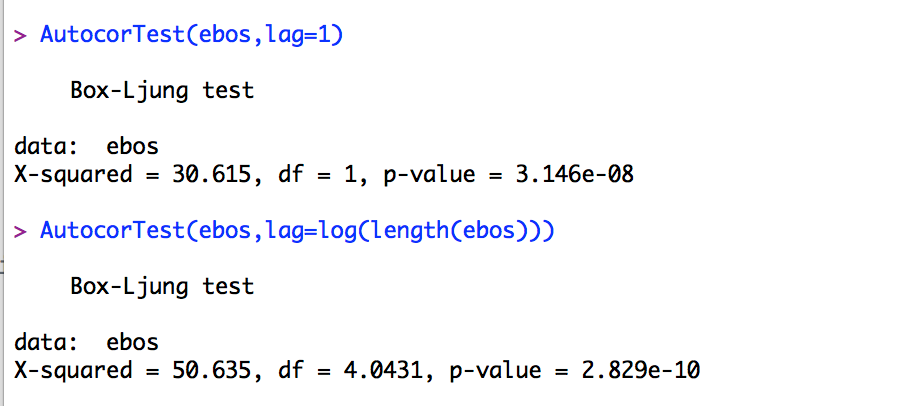
By evaluating the above plots:

* Residual vs fitted: Non-linear patterns in the residuals can be seen in this plot. In the above plot the relationship is not so clear as the red lines are not horizontal. For the better fit, polynomial model can be considered.
* Normal Q-Q: Error distribution can be inferred by this plot weather it is normal or skewed. In our plot, the line deviates from dotted line which states that the distribution is slightly skewed.
* Scale-Location: Using this plot, we can calculate the assumption of equal variance. If there is a horizontal line with equally distributed points, we can infer that the model meets the assumption. In the above plot, the assumption of equal variance is not met at all as the residuals vary as they are spread along the line.
* Residual vs Leverage: This plot is used to check for the significant outliers in the data. In the above plot, we don’t have any outliers.

As there are some assumptions that are not met, we have to check the histogram and autocorrelation of residue:





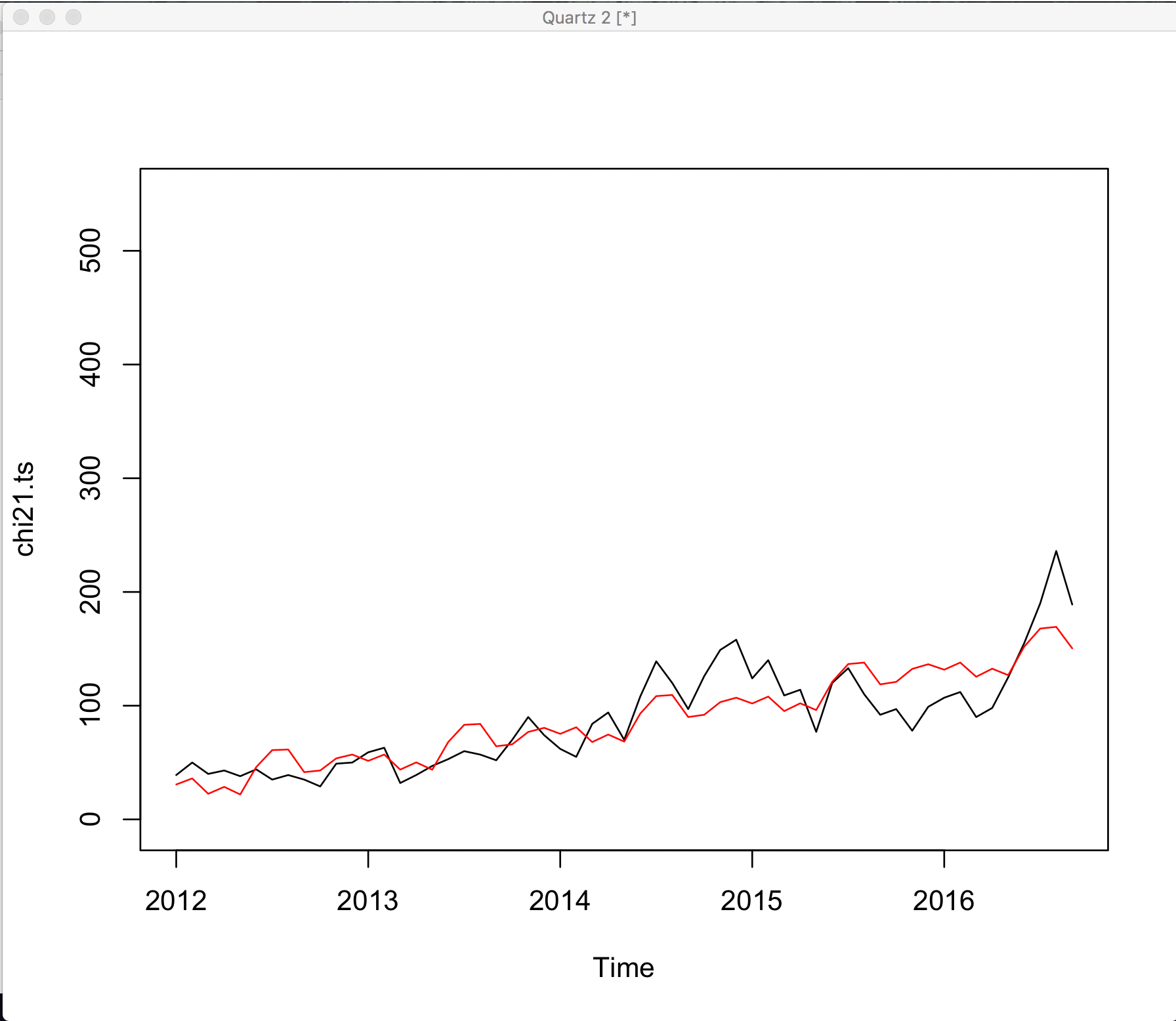


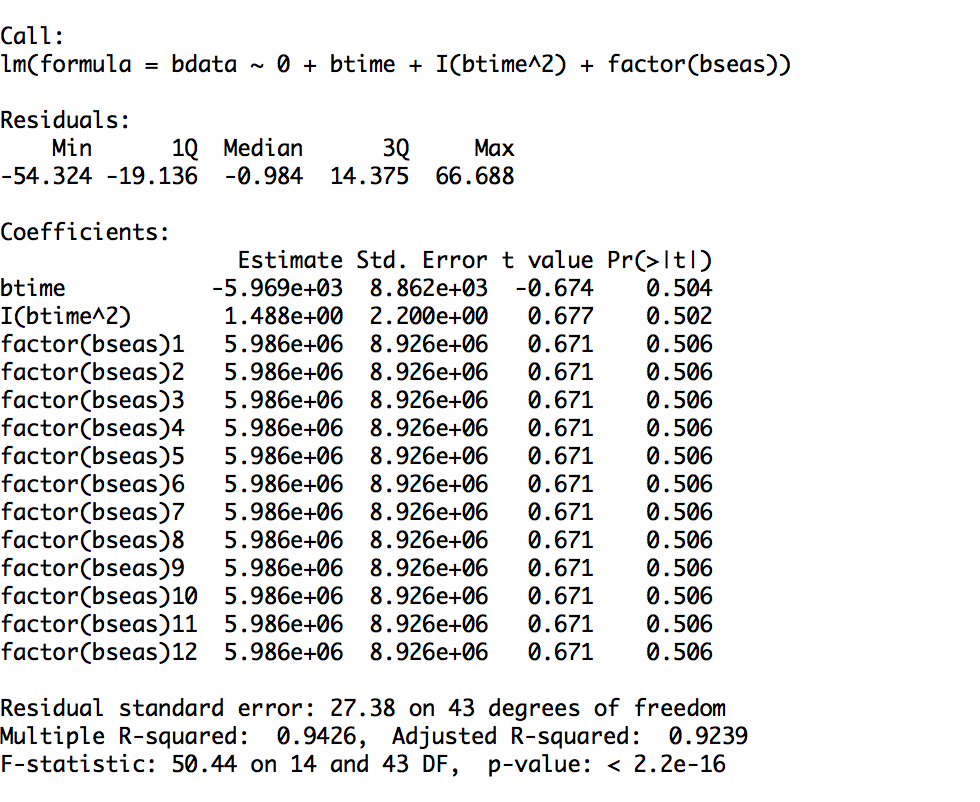
From the above histogram, we can infer that the residuals are not normally distributed. The autocorrelation output as well as tests says that there is little correlation at the start but not a lot.

1. Using quadratic model to check the fitness.

The above models stated that some assumptions are not met, there might be a case of non-linear relationships. We will use quadratic model in order to find the better fit, where the time will be squared and can be used as new variable. Below see the quadratic model:

**Regression Model**

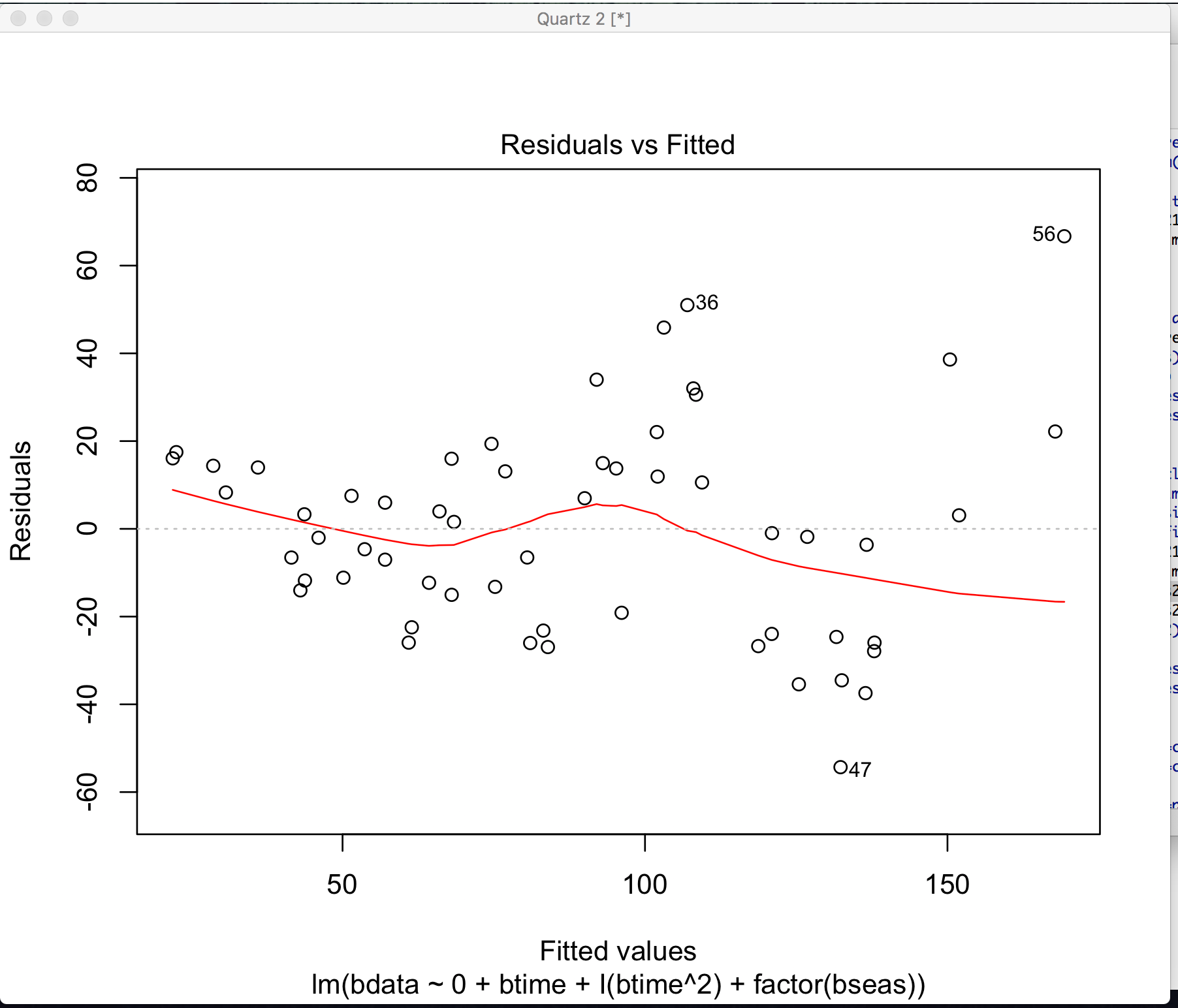


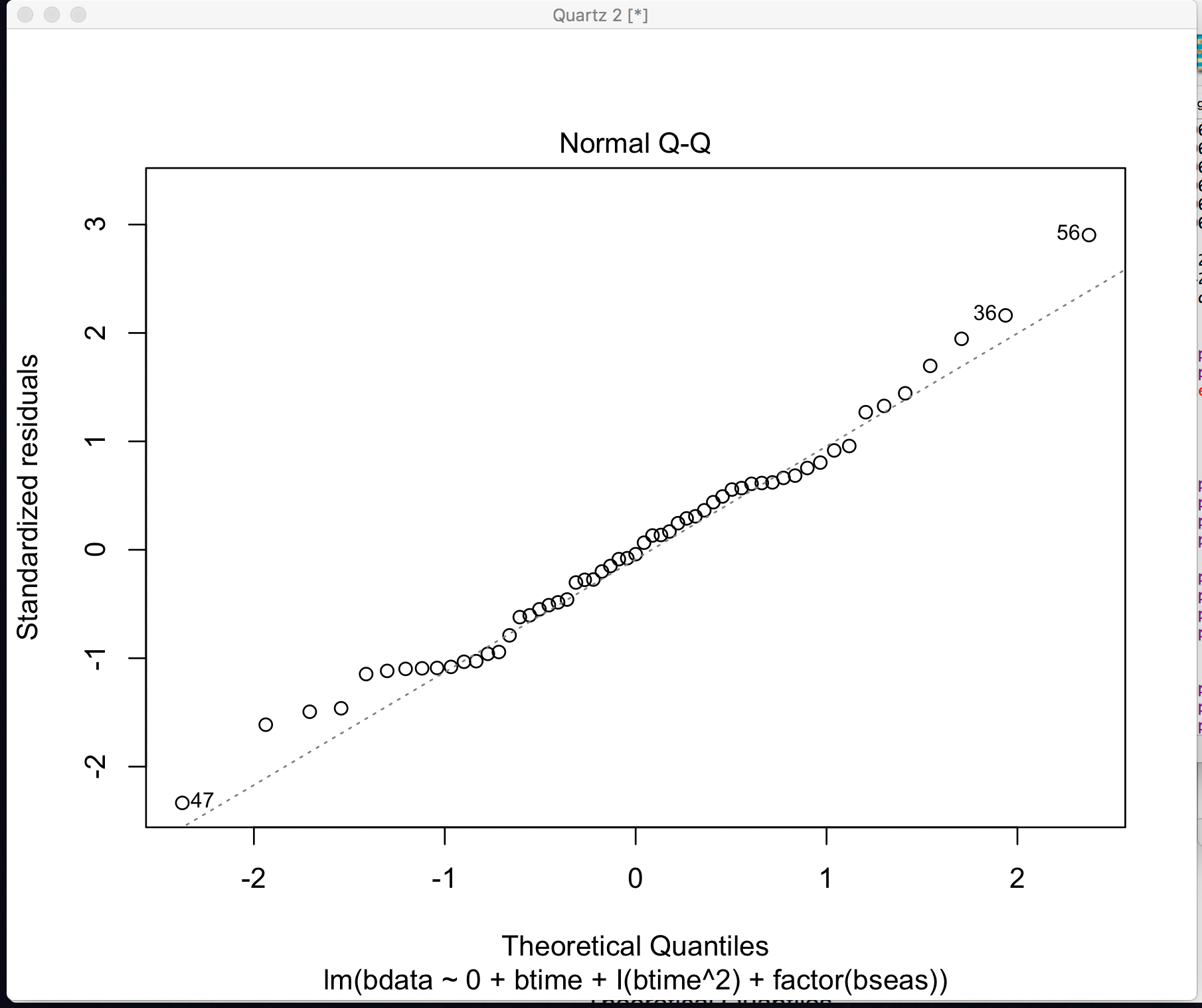


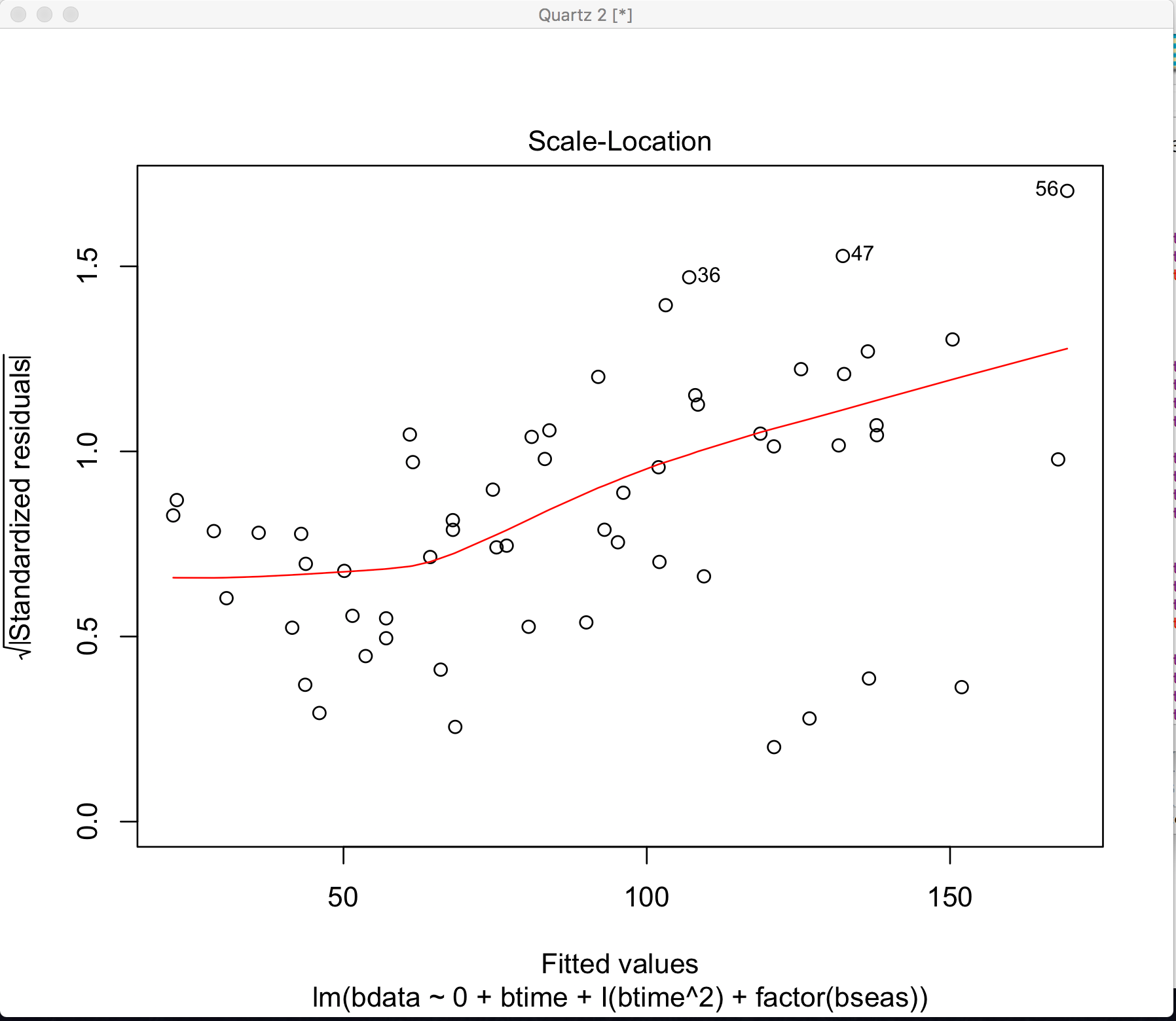
In order to compare the quadratic model with linear regression model, we have to evaluate adjusted R squared, as that the fitness of fit for the model and consider the predictors along with that:

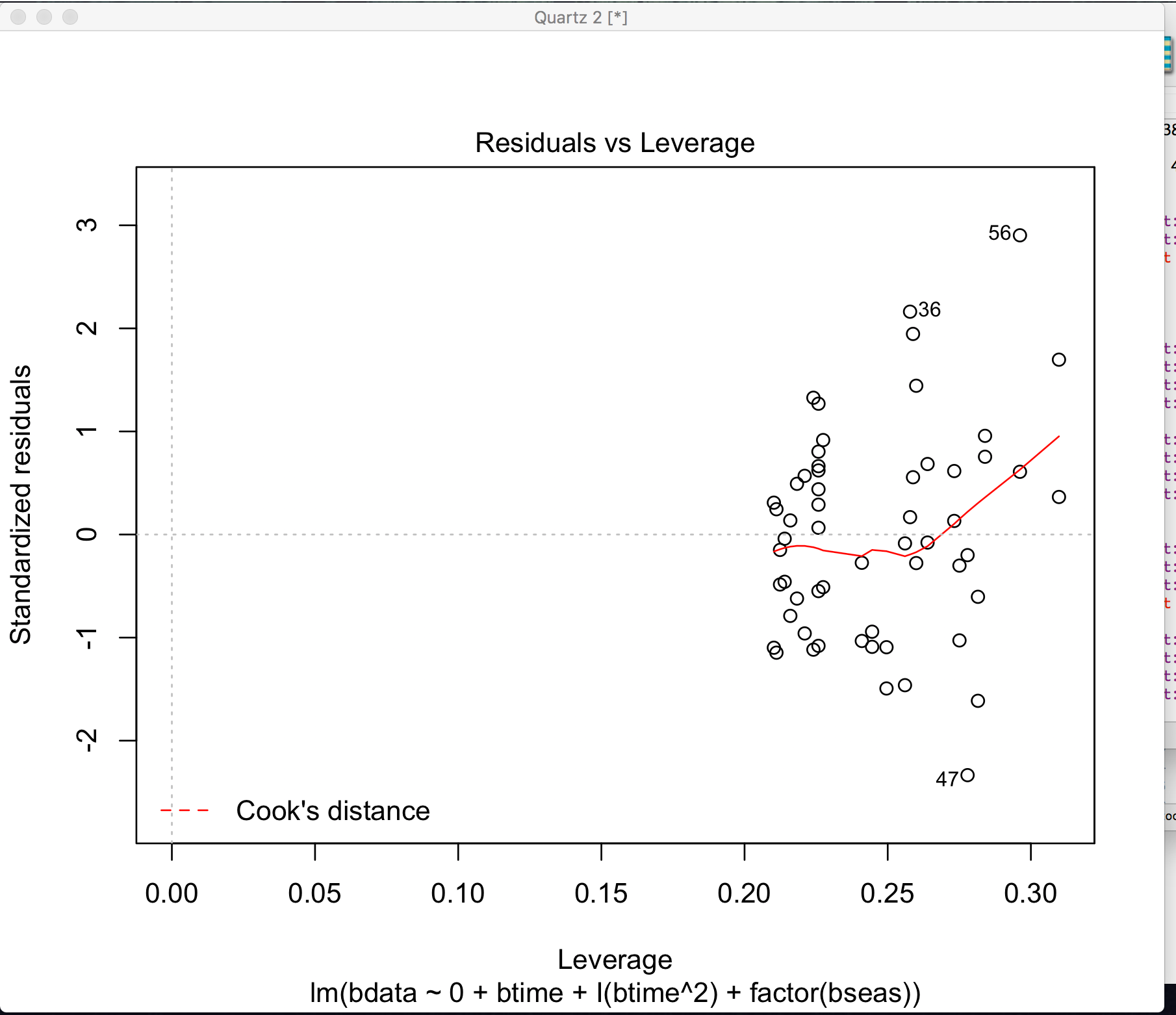
* adR2 Linear: 0.9249
* adR2 Quadratic: 0.9239

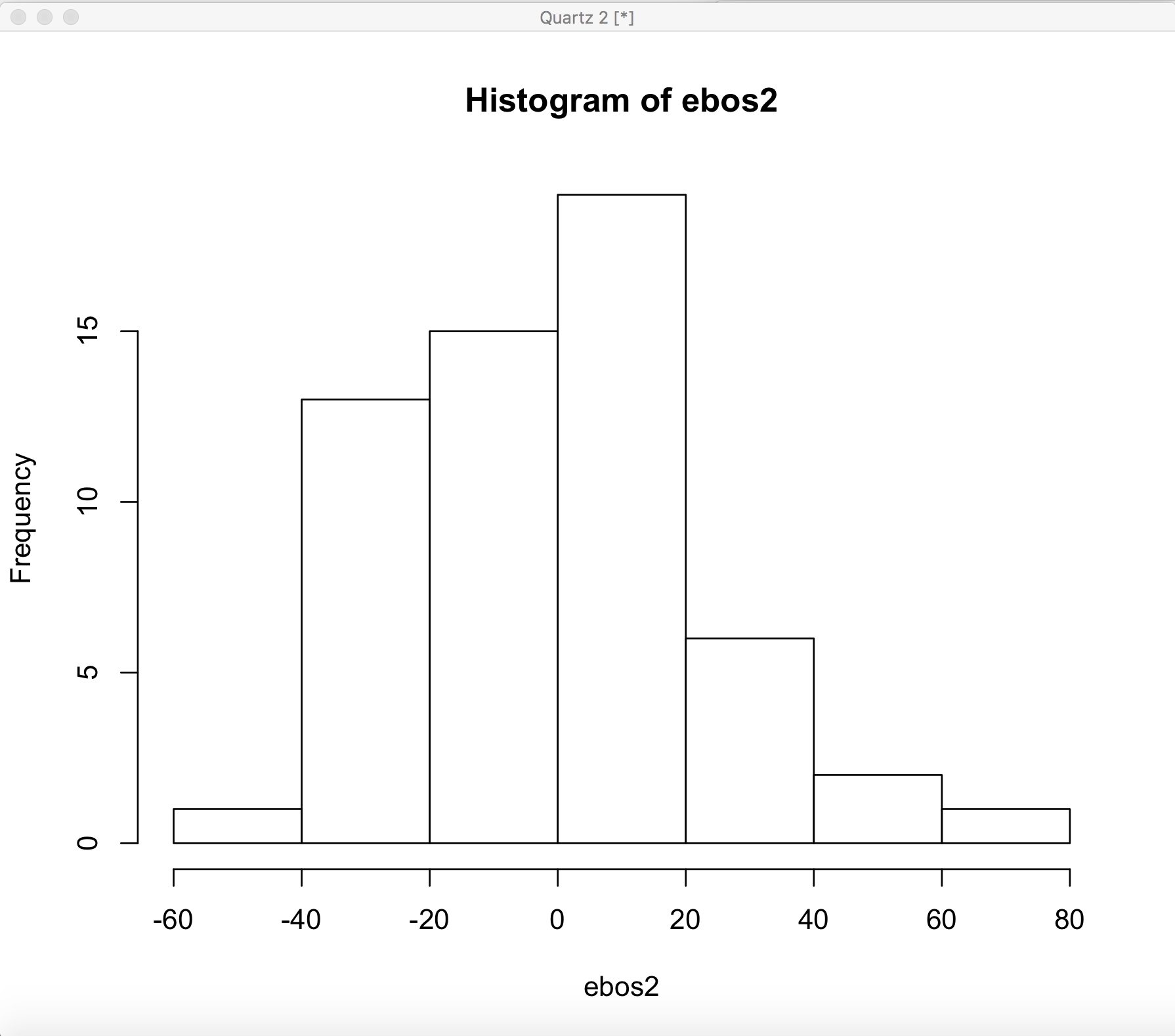
By analyzing the plot, we can conclude that the quadratic model fits better but we have to analyze the diagnostic plots in order to see the improved linear assumptions.

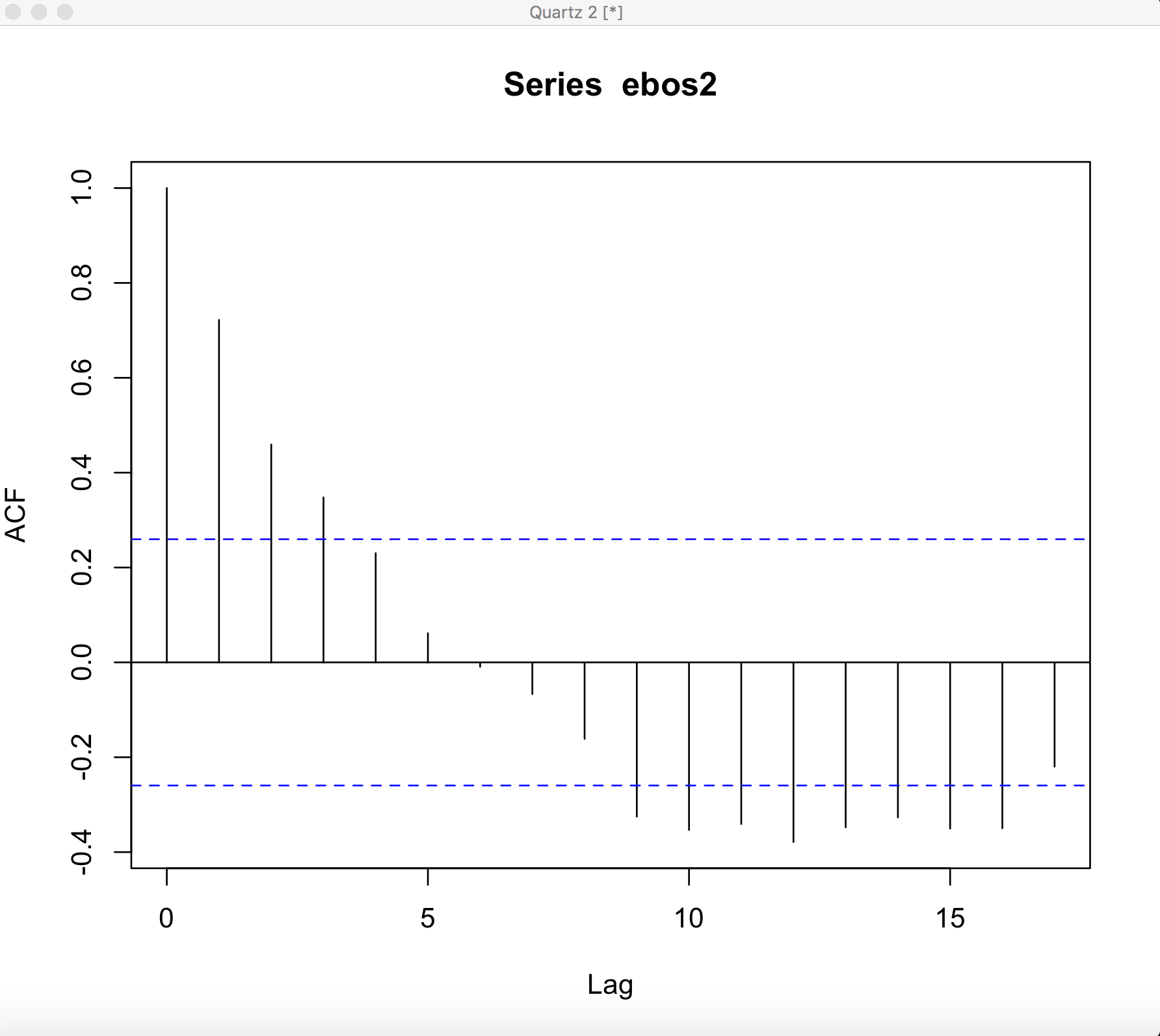


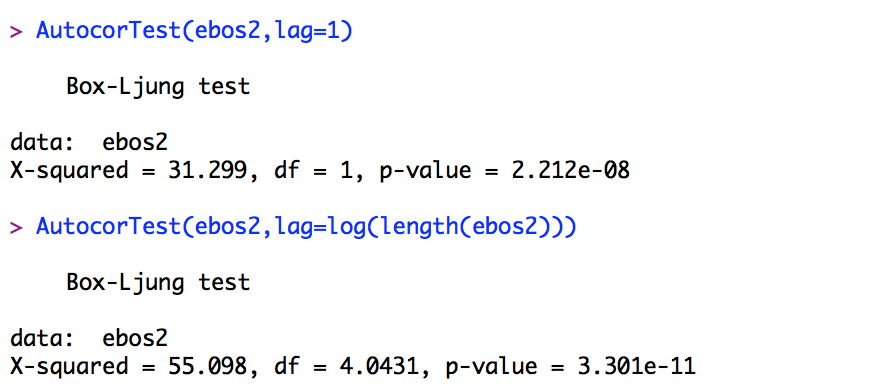












* Residual vs Fitted: The red line is closer in the earlier model than the quadratic model.
* Normal Q-Q: Here little improvements are there which can indicate that errors are normal than before.
* Scale-Location: There is an improvement in this plot as the errors are equally spread which means same error variance.
* Residual vs Leverage: We don’t have any outliers.

This model is better fit than the previous model.

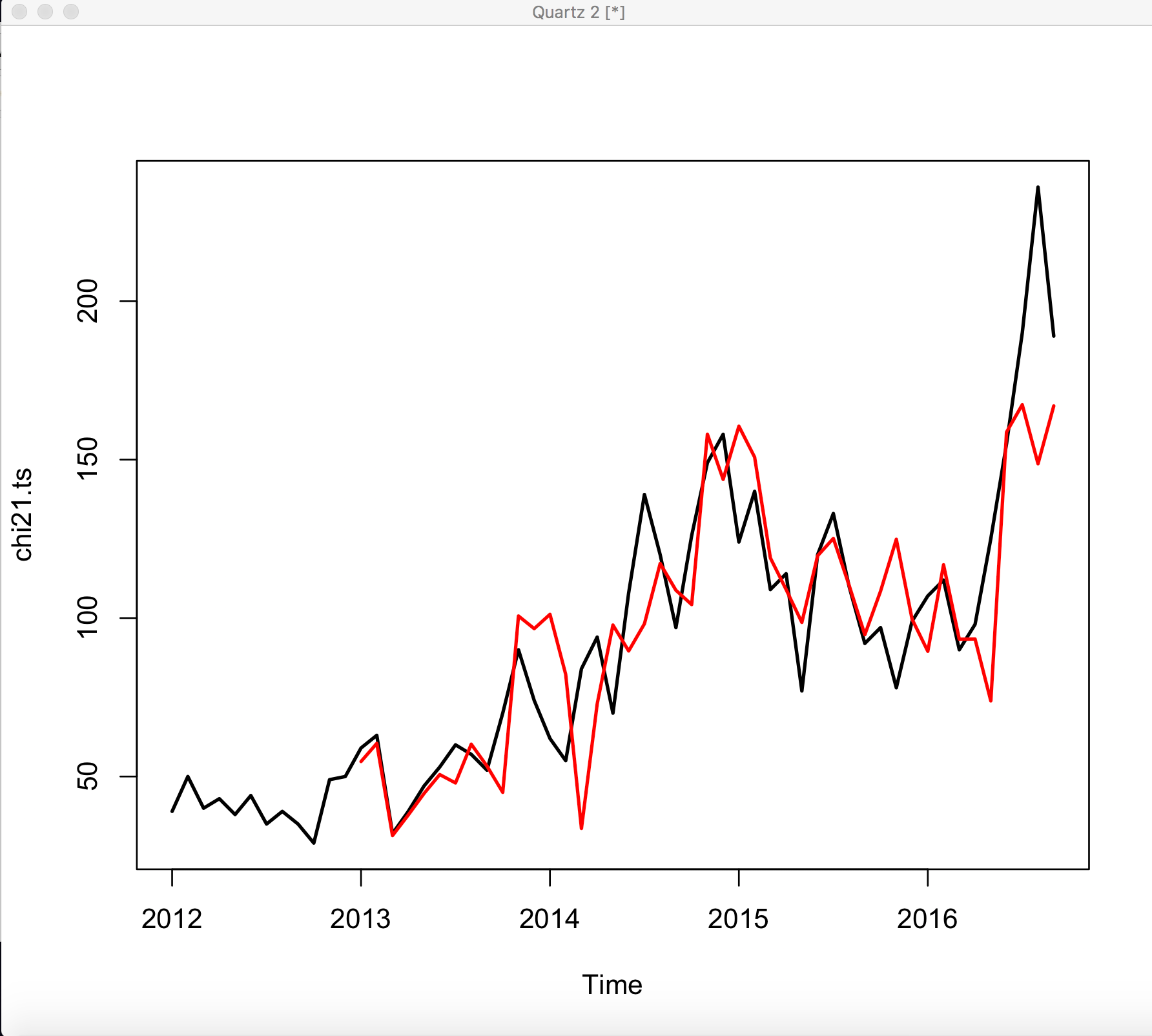
1. Using the best model to predict Oct and Nov 2016.

|  |  |
| --- | --- |
| Oct 2016 | 136.5440 |
| Nov 2016 | 148.0294 |

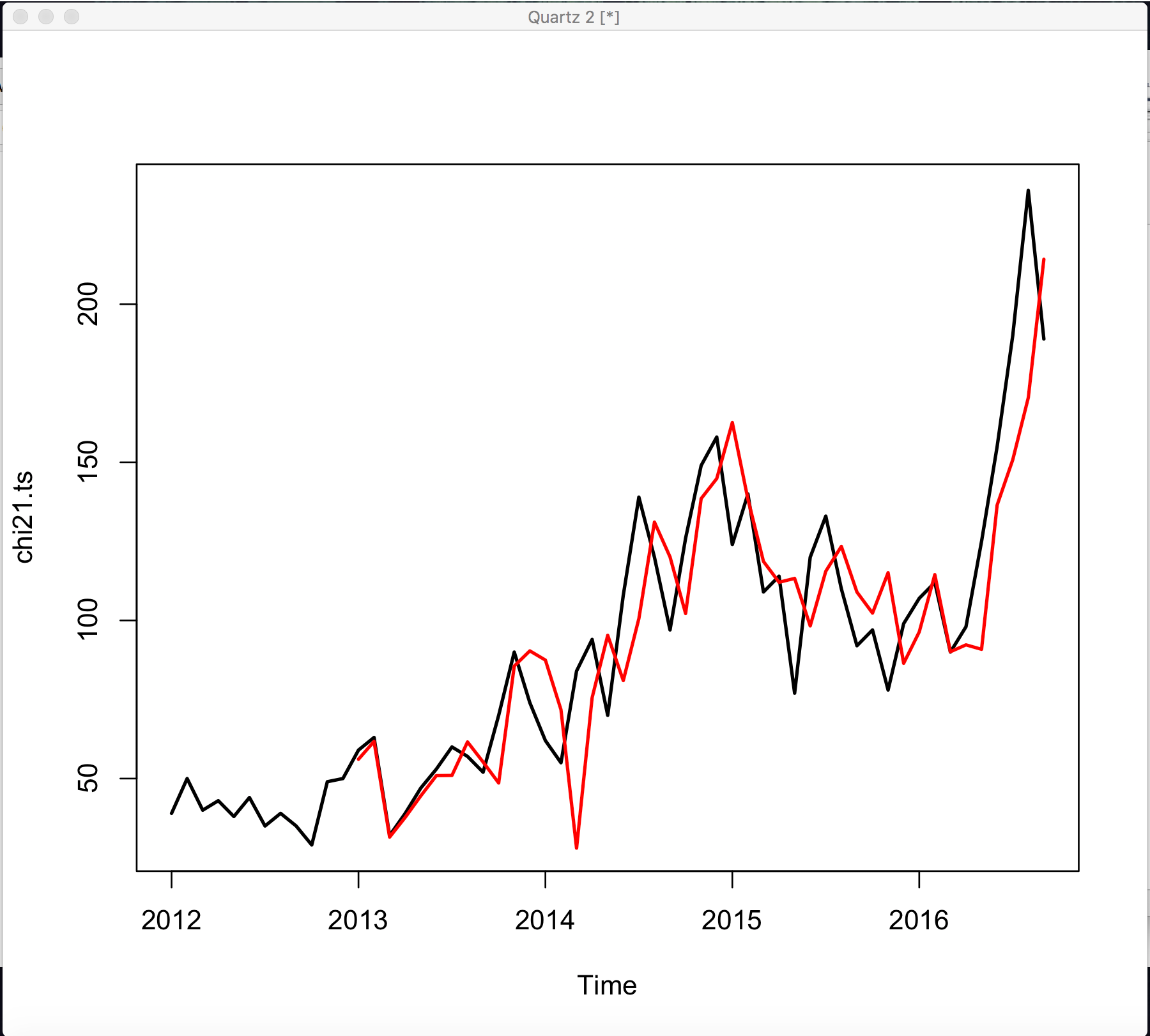
**Model 4: Using Holt-Winter Model**

The Holt-Winter model assumes that there is no systematic trend and seasonal effect. Now we will use this on both additive and multiplicative model:

**Holt-Winter Multiplicative**



**Holt-Winter Additive**



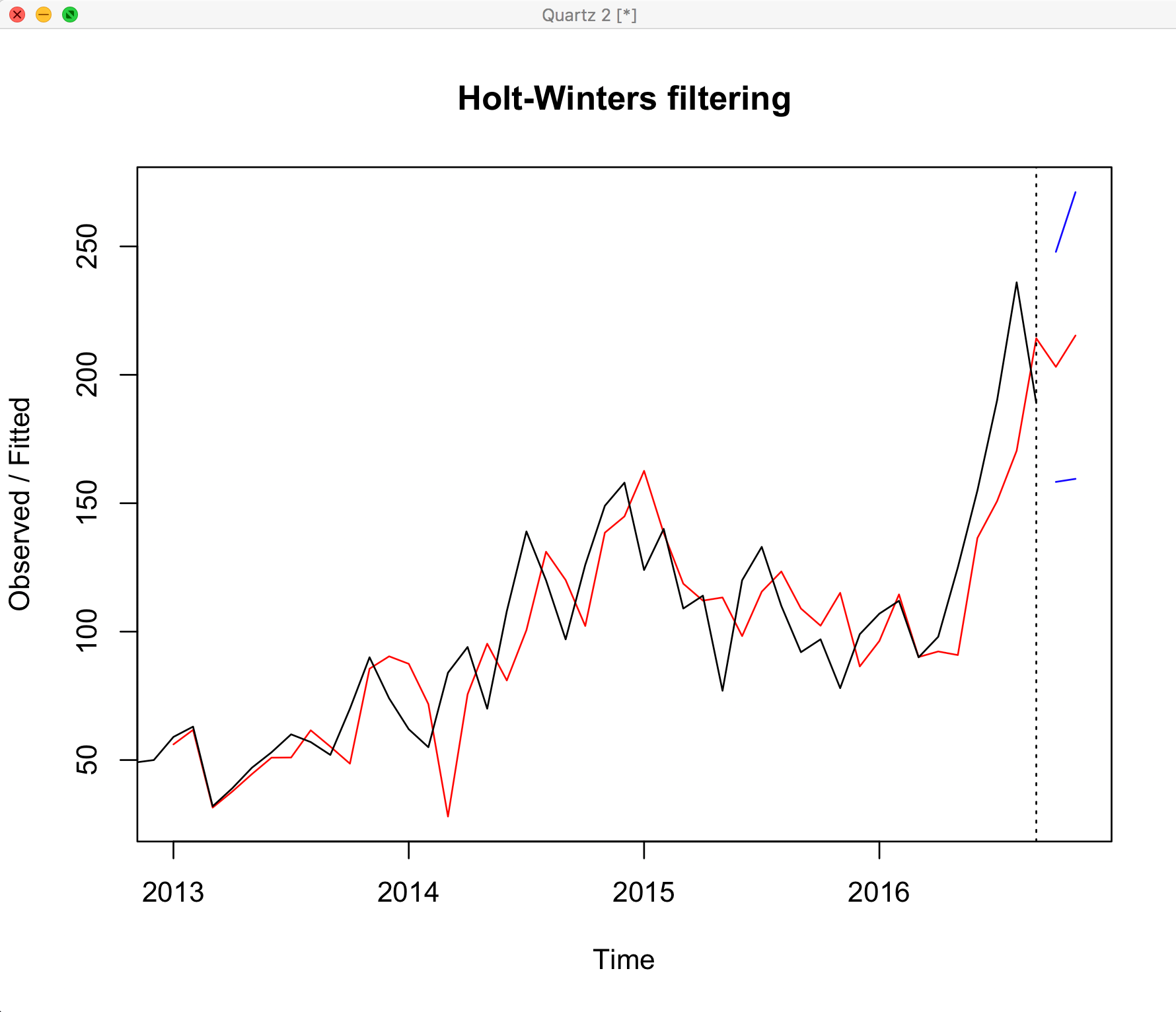
We will use RMSE to check which model fits the data better:

|  |  |
| --- | --- |
| RMSE additive | 22.8442 |
| RMSE multiplicative | 24.23891 |

In our case the additive model fits better.

We will use the additive model predict the value of Oct and Nov 2016.

|  |  |  |  |
| --- | --- | --- | --- |
|  | fit | upr | lwr |
| Oct 2016 | 203.0939 | 247.8858 | 158.302 |
| Nov 2016 | 215.3024 | 271.1188 | 159.486 |



**Comparison and analysis**

Here we will analyze all the above model and choose the best model using all the above forecasts made for Oct and Nov 2016. First, we will explain the power of the models that we have taken:

* Decomposition model: As decomposition models are not good for extrapolating we don’t prefer using this model for forecasting. This model focuses mainly on getting a good fit.
* Regression to trend decomposition model: In this model, the seasonality is not considered so it cannot make reliable predictions.
* Regression to trend loess model: Even in this model the seasonality is not considered but this model uses various transformation like log transformation to make better predictions.
* Regression model: Although the regression assumptions are not met 100% but they are close and they are good fit as higher R2 make it the best model so far.
* Holt-Winter model: This model works on systematic trends as well as seasonal effects which are not present in our case, that is the reason we will not use this model for prediction.

So by analyzing the above models and the visual evaluation we have inferred that the quadratic model is the best to make predictions.