**Title:** Data7202 A2 Report

**Name:** Xinqian Wang

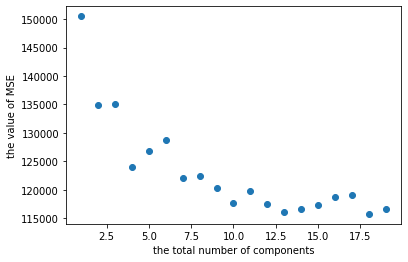
**ID:** s4565489

**Tips:** the code screenshot will be attached in the Appendix. For code script will also be attached in the folder.

* **1.**

**Question:** Apply Principal Component Regression (PCR) with all possible number of principal components. Using the 10-Fold Cross-Validation, plot the mean squared error as a function of the number of components and determine the optimal number of components.

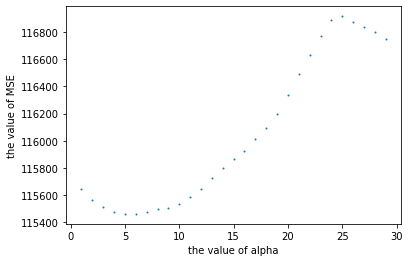
**Answer:**



The optimal number of components is 18.

**Question:** Apply the Lasso method and plot the 10-Fold Cross-Validation mean squared error as a function of λ. Determine the best λ and the corresponding mean squared error.

**Answer:**



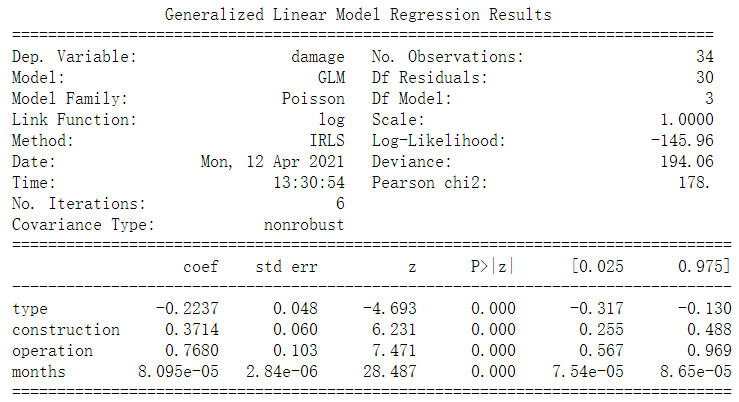
The best λ should be the alpha equals to 5.

* **2.**

**Question:** Construct a Poisson regression model and report the coefficients (for type, construction, operation, and months), and the corresponding 95% CIs.

**Answer:**

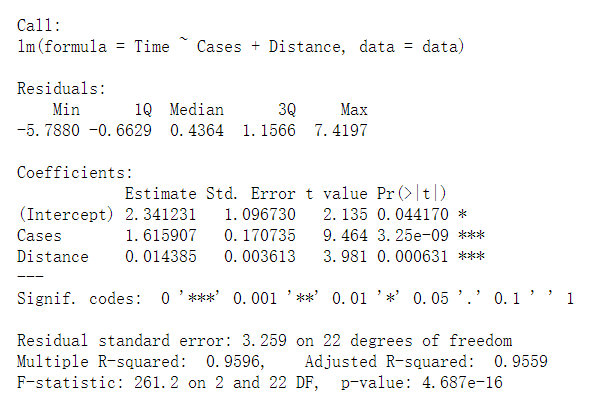
By using the code model = sm.GLM(Y,X,family=sm.families.Poisson()), we build a Poisson Regression Model. The coefficients for **type**, **construction**, **operation**, and **months** are below:



* **3.**

**Question:** Compute the multiple regression of Time on Cases and Distance. State the fitted model, the estimated residual standard deviation, and the P-values for the overall model and each of the two predictors.

**Answer:**



The = 3.259 on 22 degrees of freedom.

The P-value of overall model is 4.687e-16.

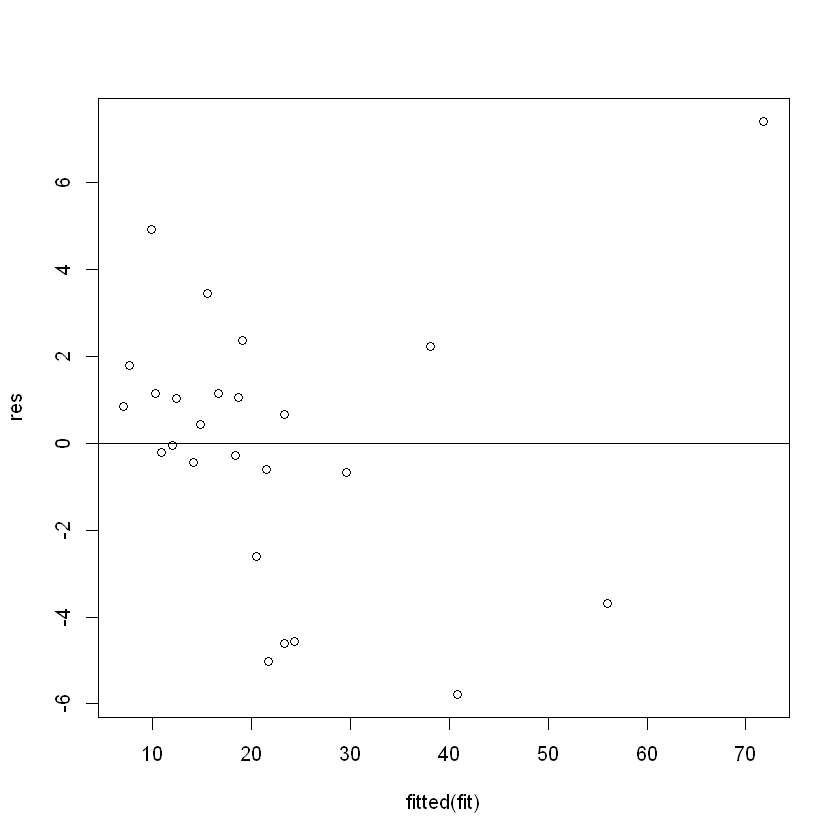
The P-value of the attribute Cases is 3.25e-09.

The P-value of the attribute Distance is 0.000631.

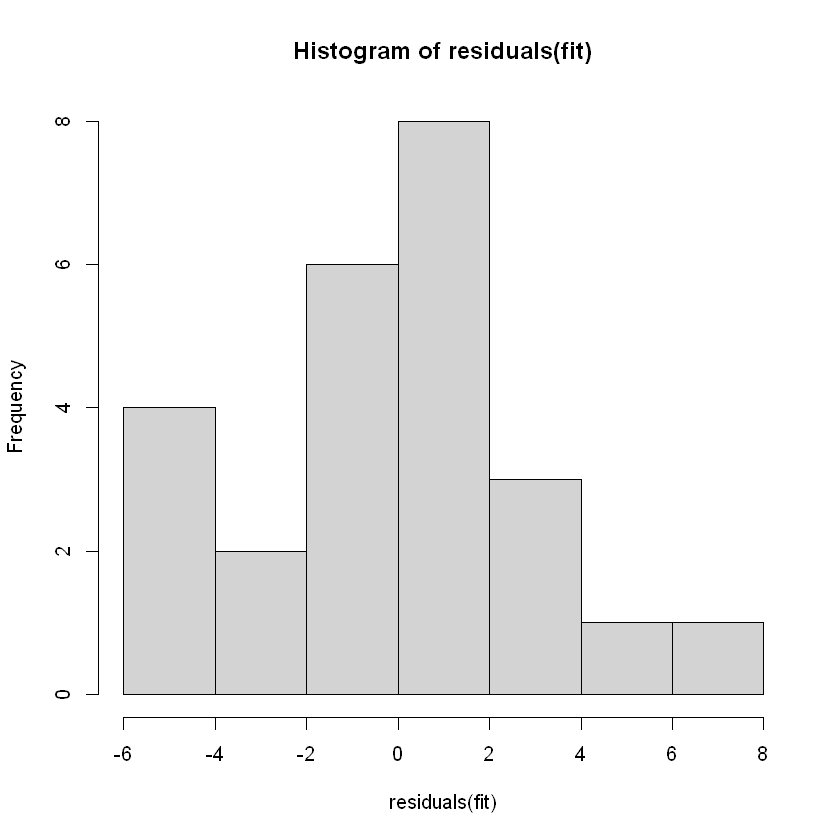
**Question:** Obtain residual plots and the histogram of the residuals. Comment on these.

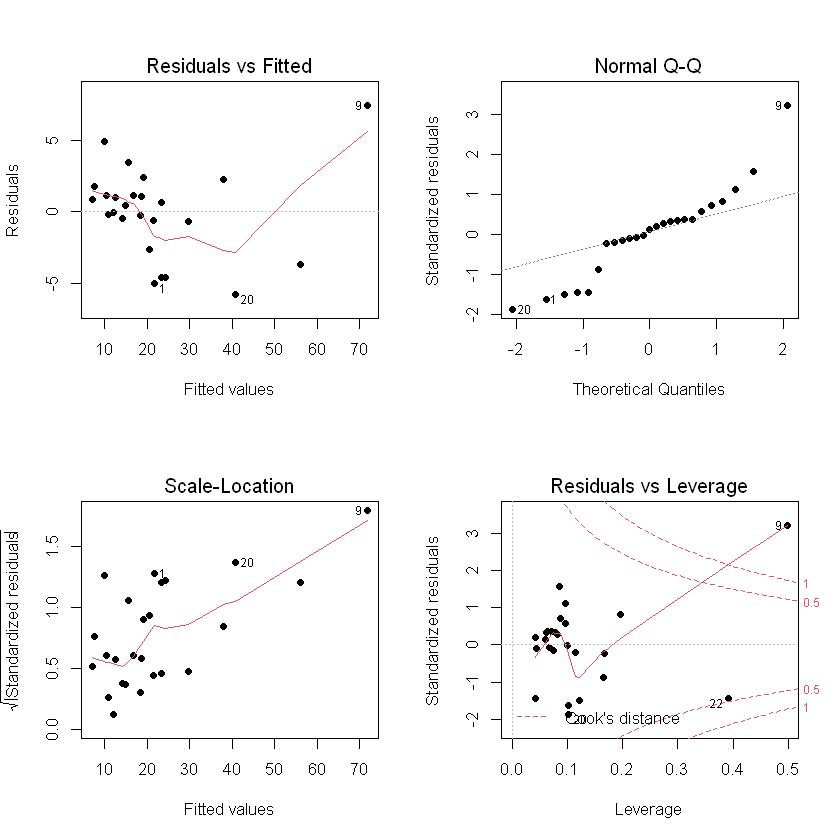
**Answer:**

Residual plot is as below:



Residual histogram is as below:





The **histogram** shows most of the residuals have the values range between -2 and 2.

The **Residuals vs Fitted plot** shows whether the residuals have non-linear patterns. The plot fails to show equally spread residuals around a horizontal line without distinct patterns, which indicates that the non-linear relationship was not explained by the model and was left out in the residuals.

The **QQ plot** is used to show if the residuals are normally distributed or not. The plot just tells that the residuals are probably not following a normal distribution.

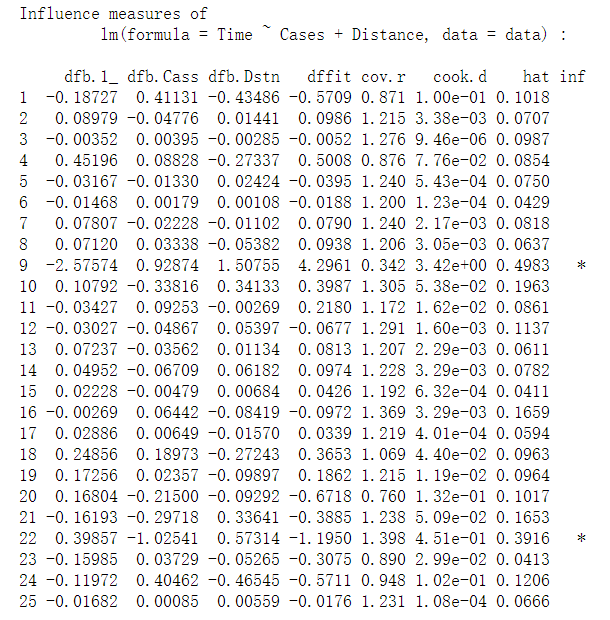
The **Spread-Location plot** shows if residuals are spread equally along the ranges of predictors. The plot shows that most of the residuals are spread in the left side of the plot whereas only a few points distributed on the right. Especially the 9th point, it’s value seems not common as other points.

This **Residuals vs Leverage plot** helps us to find influential cases if any. Not all outliers are influential in linear regression analysis (whatever outliers mean). It is usually the case that a point is judged as an outlier if its cook's distance is greater than 0.5. In this case, the point 22 and point 9 would interest us to do further research, which also means they are more influential than any other points when the regression model is decided.

**Question:** There is an observation in this data set which is extremely influential according to Cook’s distance. Which observation is it? Display a Cook’s distance plot to determine the Cook’s distance of the next most influential observation.

**Answer:**

The 9-th point is the most influential point, according to the Influence table below and the plots above. The Cook’s distance plot is placed on the former plot. The next most influential observation would be the 22-ed point.



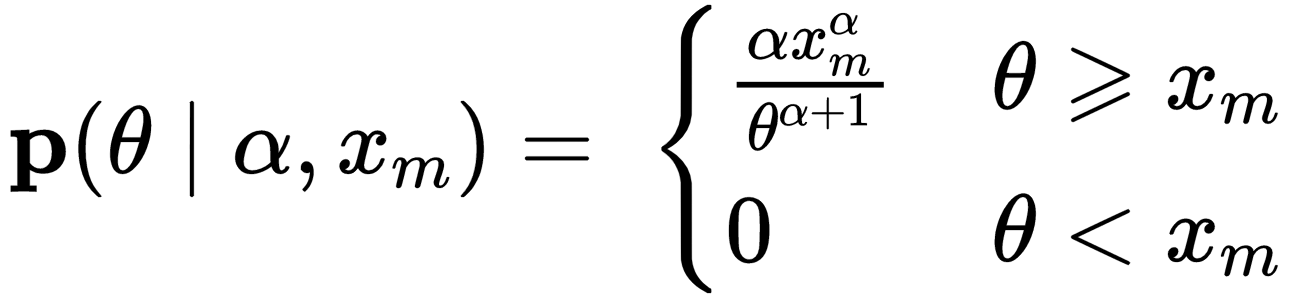
* **4.**

**Question:** Derive the posterior distribution of .

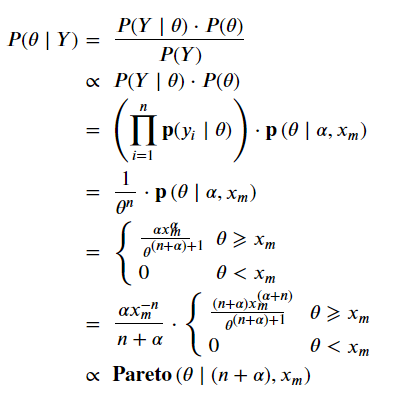
**Answer:**

Firstly, calculating the likelihood:

Secondly, calculating the prior:



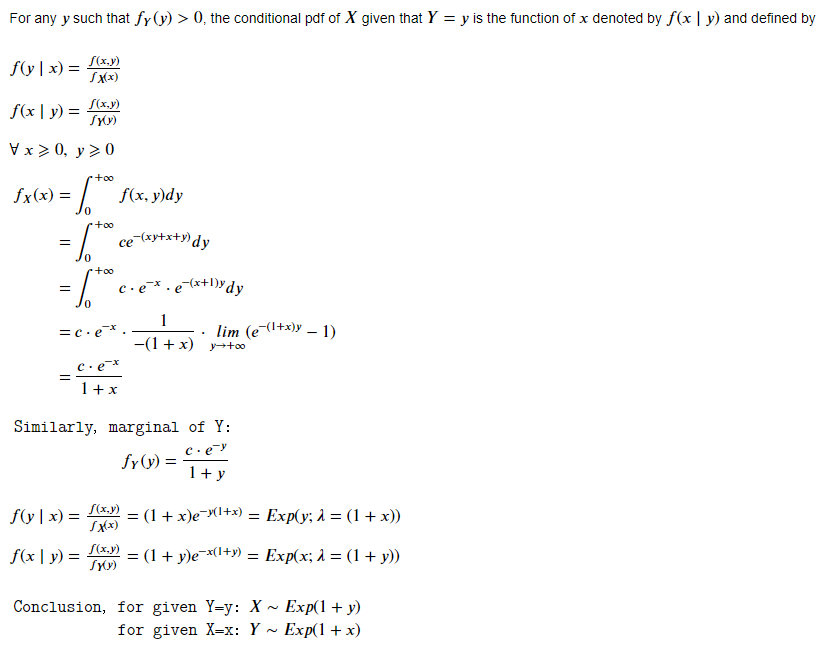
Thirdly, calculating the posterior and prove it proportional to a Pareto distribution:



* **5.**

**Question:** Find the conditional pdf of X given Y = y, and the conditional pdf of Y given X = x.

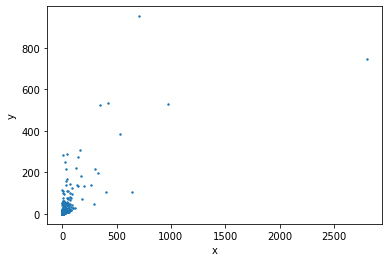
**Answer:**



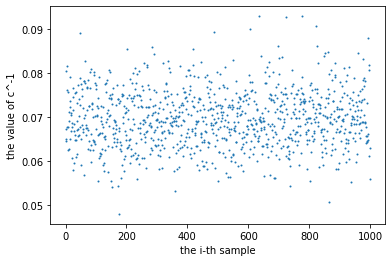
**Question:** Write working Python code that implements the Gibbs sampler and outputs 1000 points that are approximately distributed according to f.

**Answer:**

Attached in the code. The scatter plot as below:

****

I can also calculate the approximation of the constant c since it is hard to be calculated from the equation of . By calculating the sum of the 1000 samples and divided by 1000, we can get a good approximation of c. Repeating it with 1000 times, we can get the range of value . I also calculated the mean value of , which is around **0.069578**. I draw a scatter plot with the value of 1000 number of below:

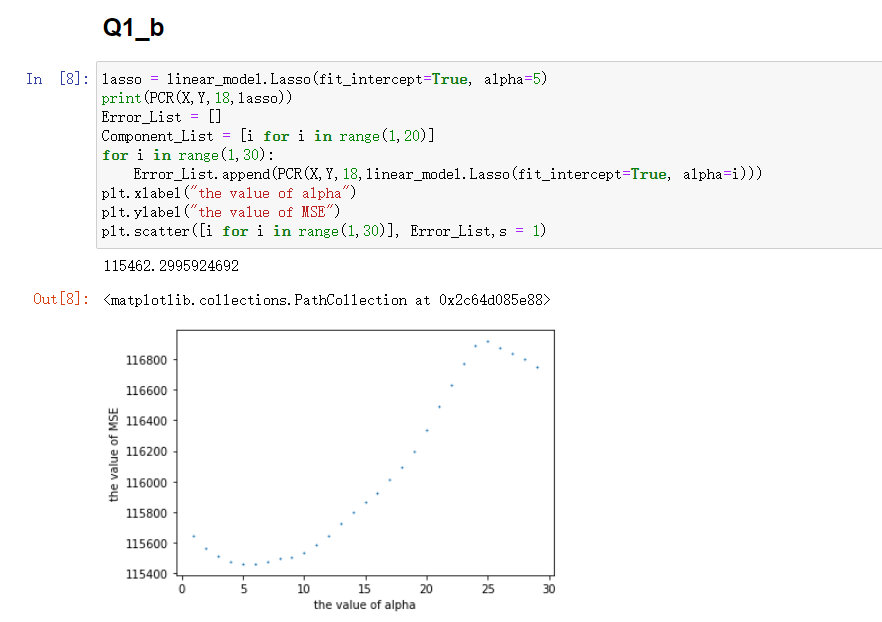


**Appendix**

* **1.**

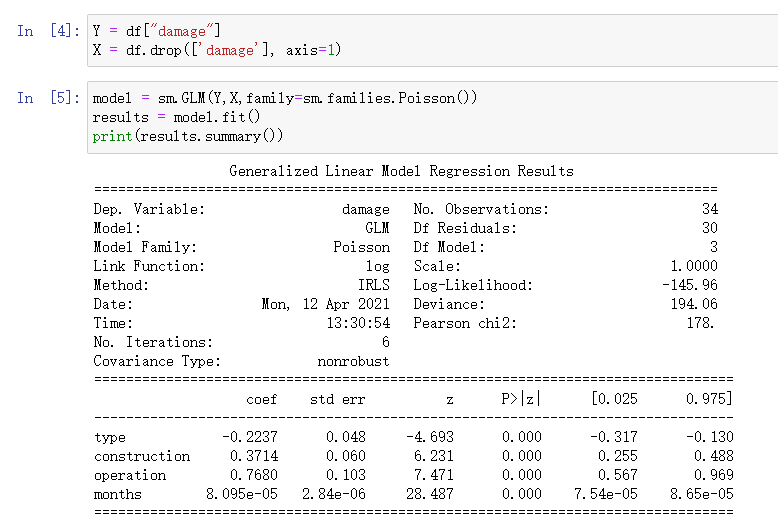






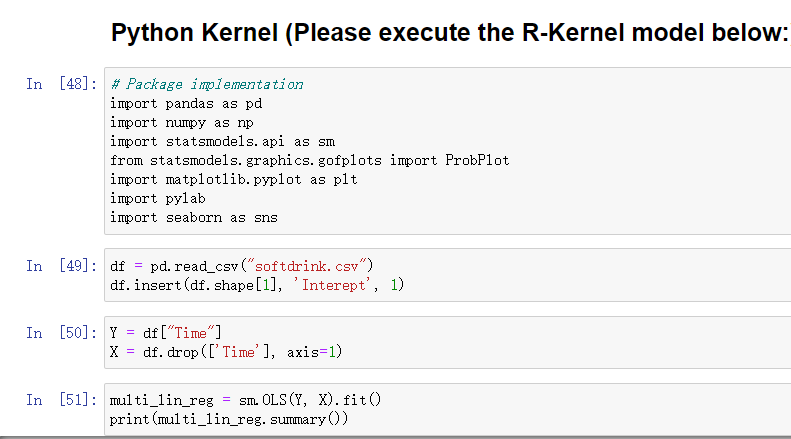
* **2.**

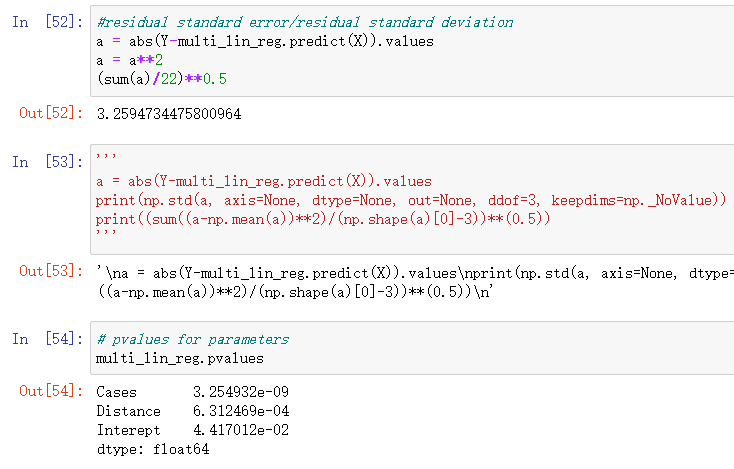


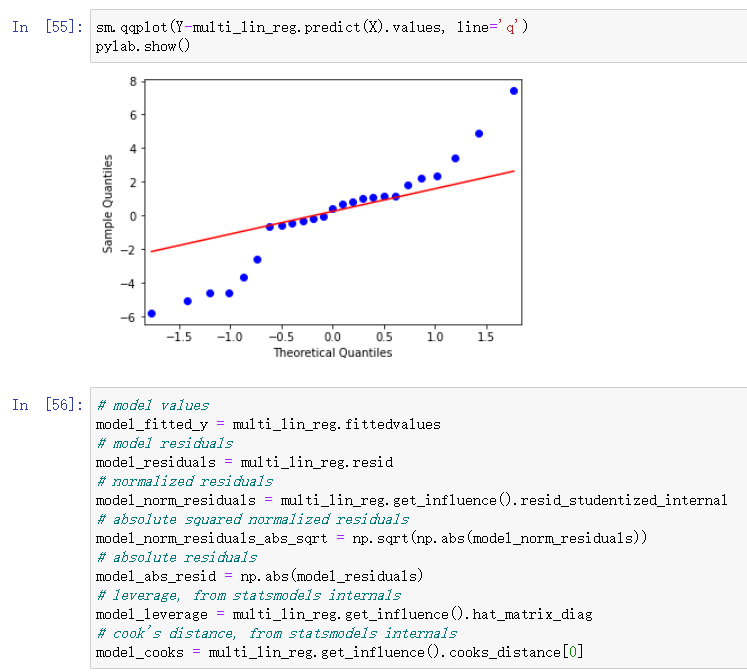


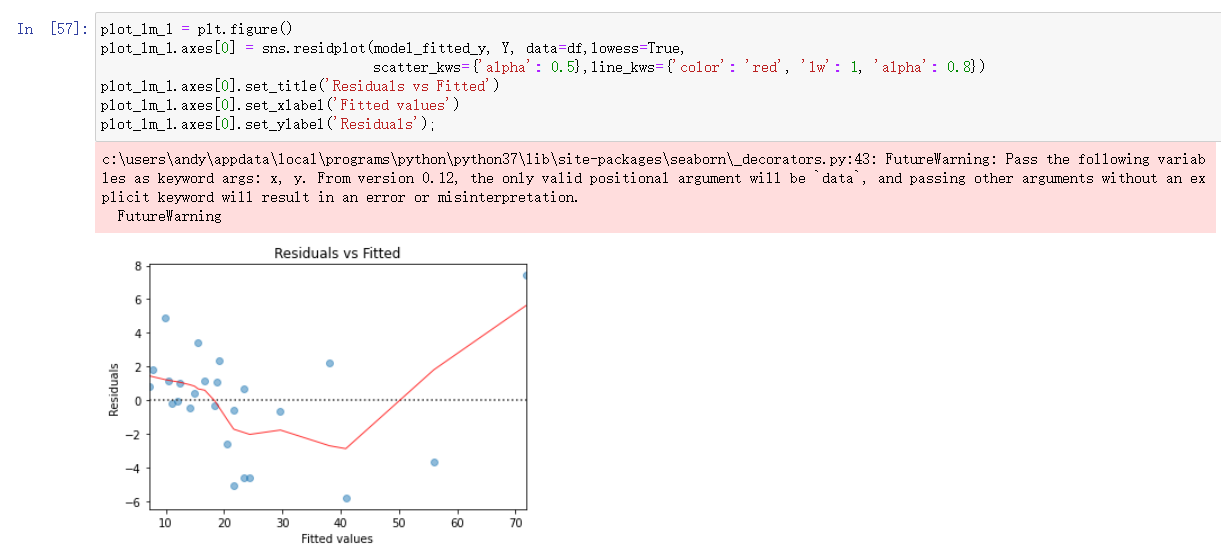
* **3.**

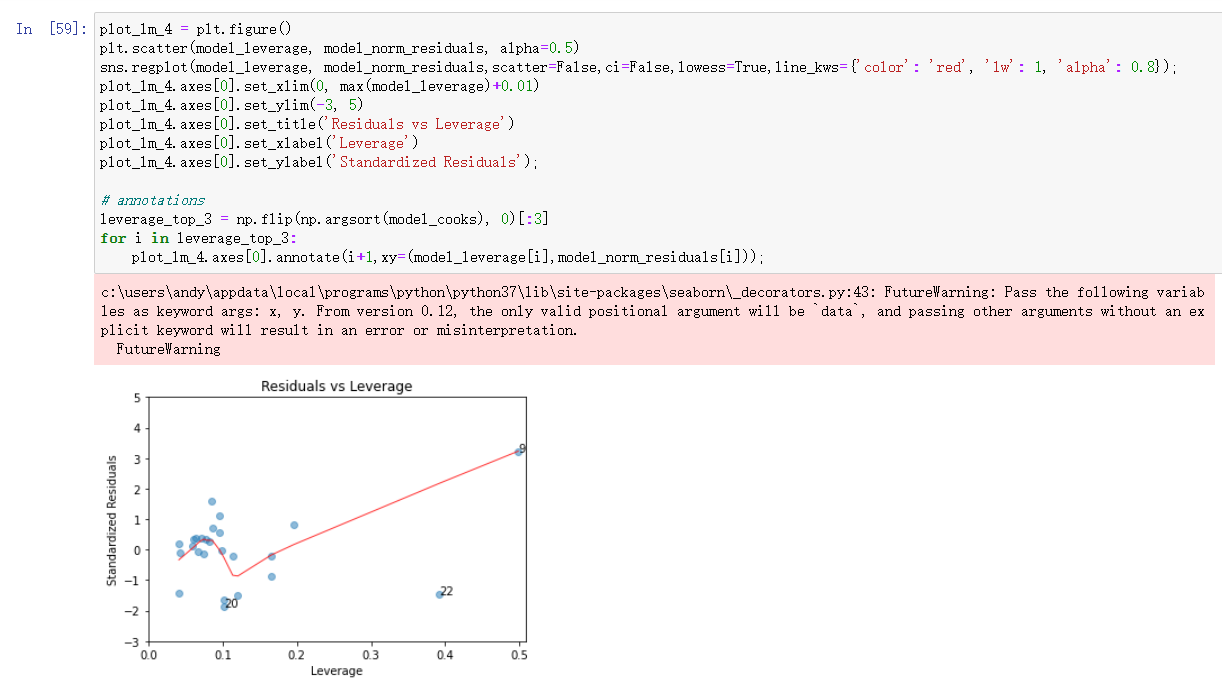
**Python Kernel**



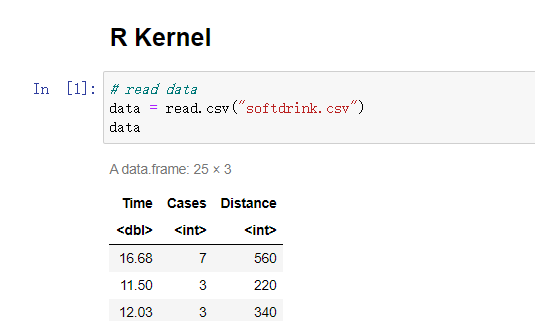


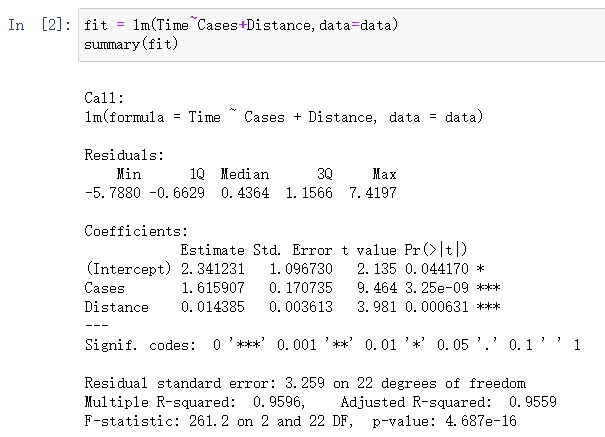


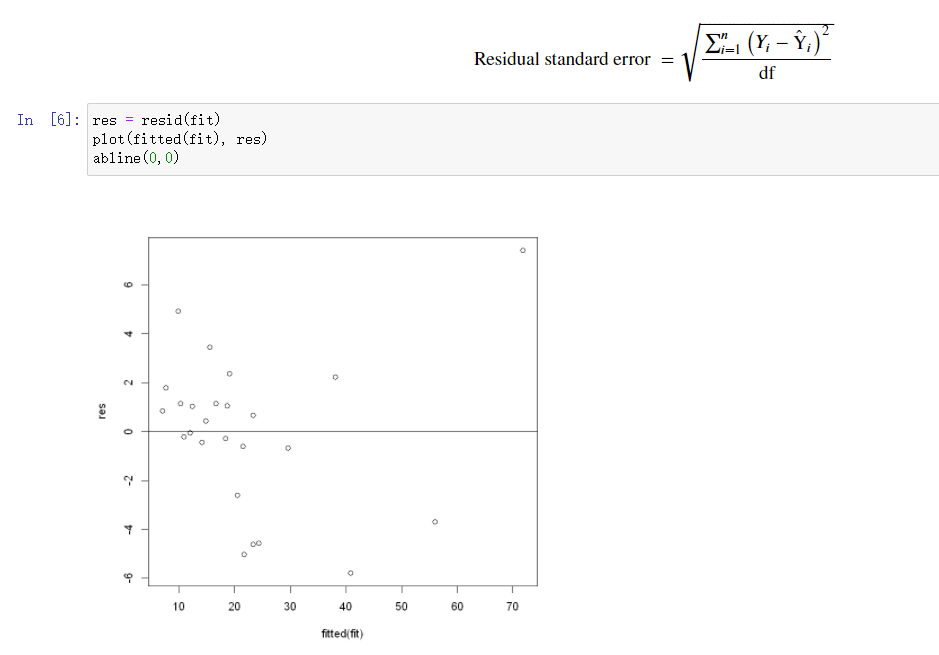


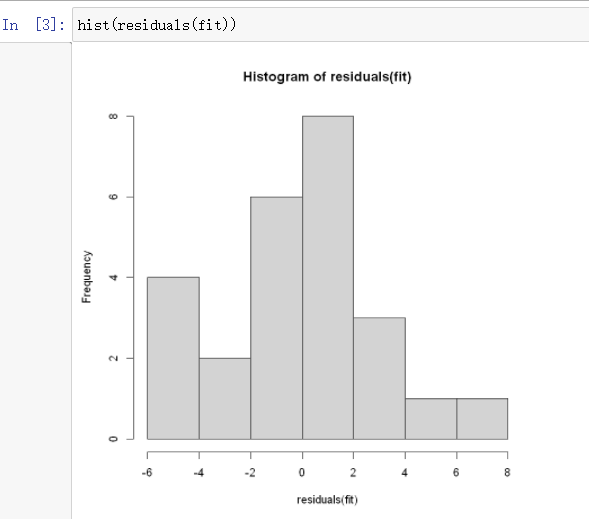


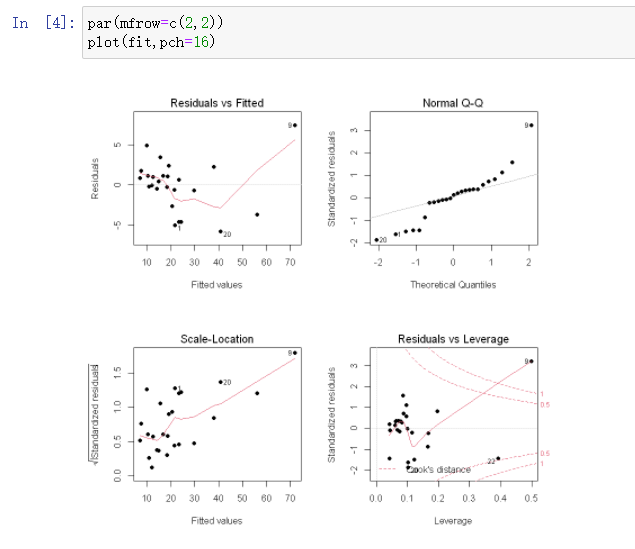
**R Kernel**

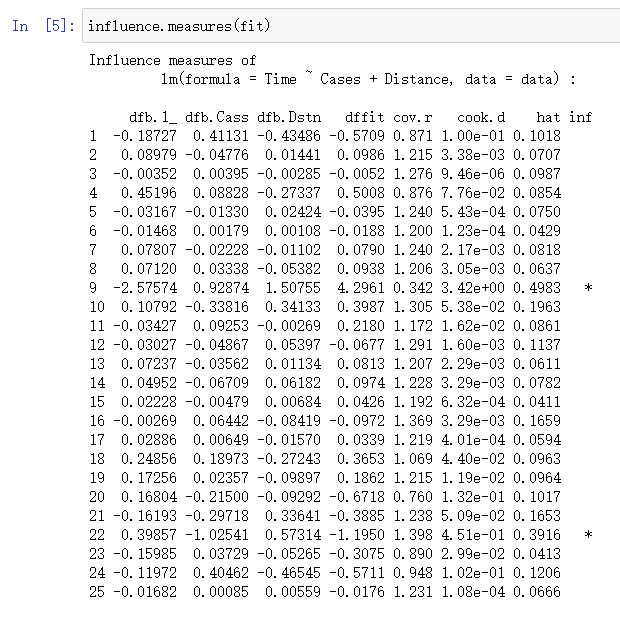








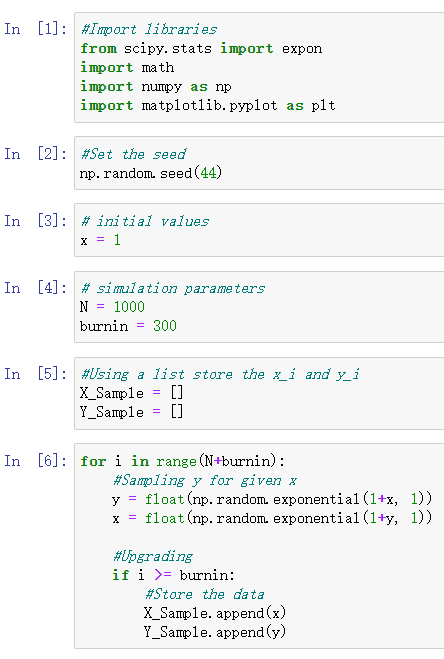


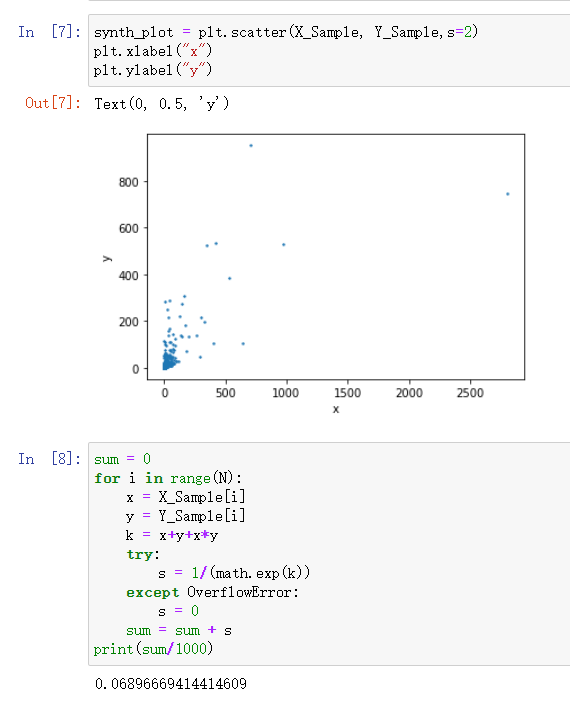


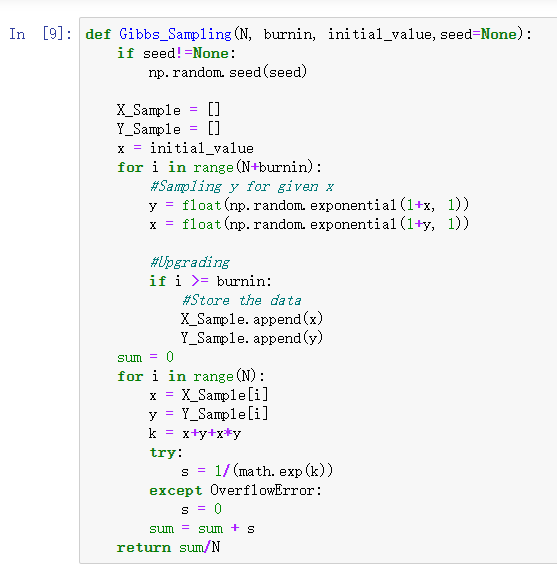
* **4.**

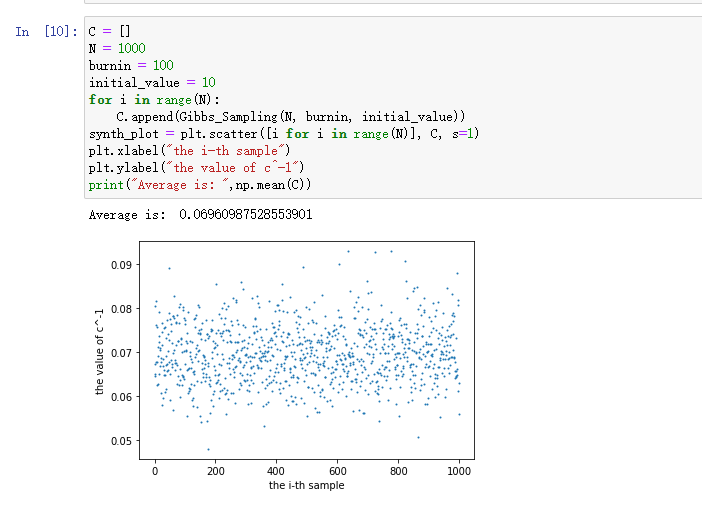
None

* **5.**









***Thank you for reading my report.***