DATA7202 : Assessment 2

Name : Peng Yu

Student ID : 4664588

Q1:

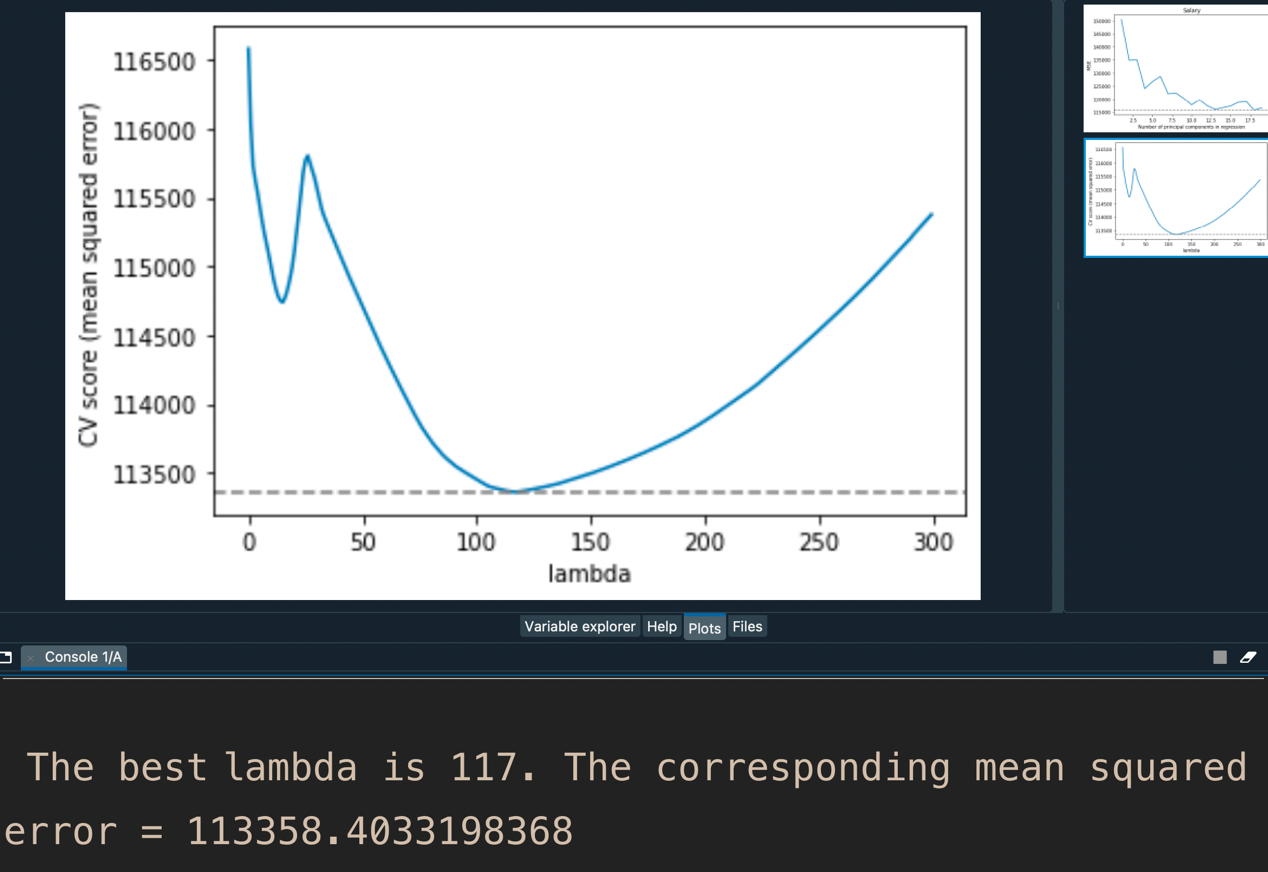
Answer (a):

Through drawing and calculation, the optimal number of components is 18, and the corresponding mean squared error is 115733.66222777616.



Answer (b):

Through drawing and calculation, the best is 117, and the corresponding mean squared error is 113358.4033198368.



Q2:

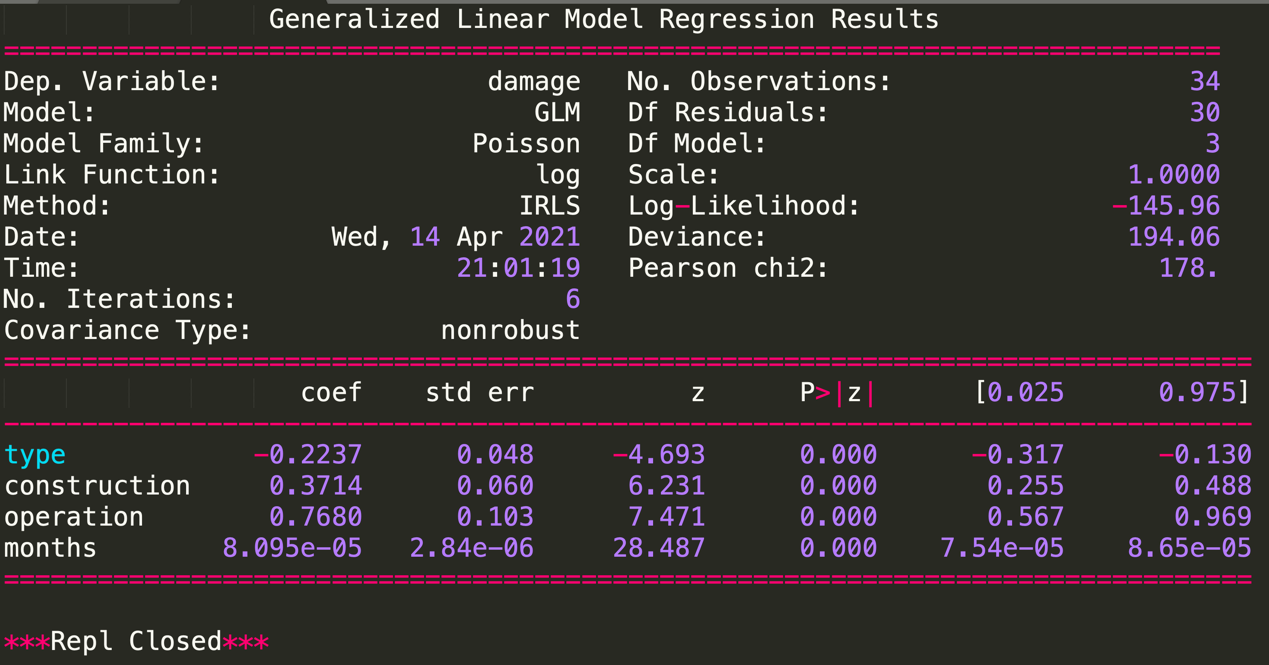
Answer:

The coefficient for type is -0.2237, and the corresponding 95% Cis is (-0.317, -0.130).

The coefficient for construction is 0.3714, and the corresponding 95% Cis is (0.255, 0.488).

The coefficient for operation is 0.7680, and the corresponding 95% Cis is (0.567, 0.969).

The coefficient for months is 8.095\*, and the corresponding 95% Cis is (7.54\*, 8.65\*).

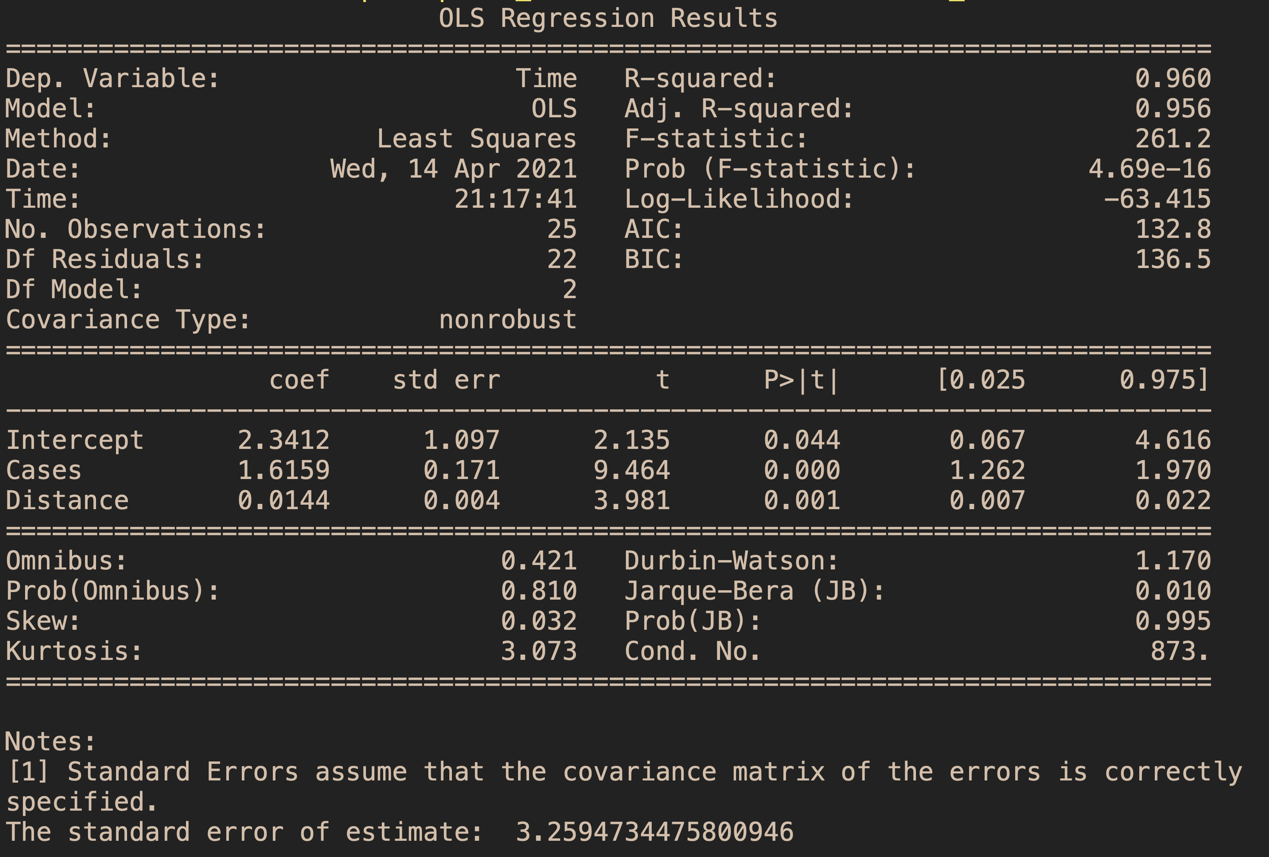


Q3:

Answer (a):

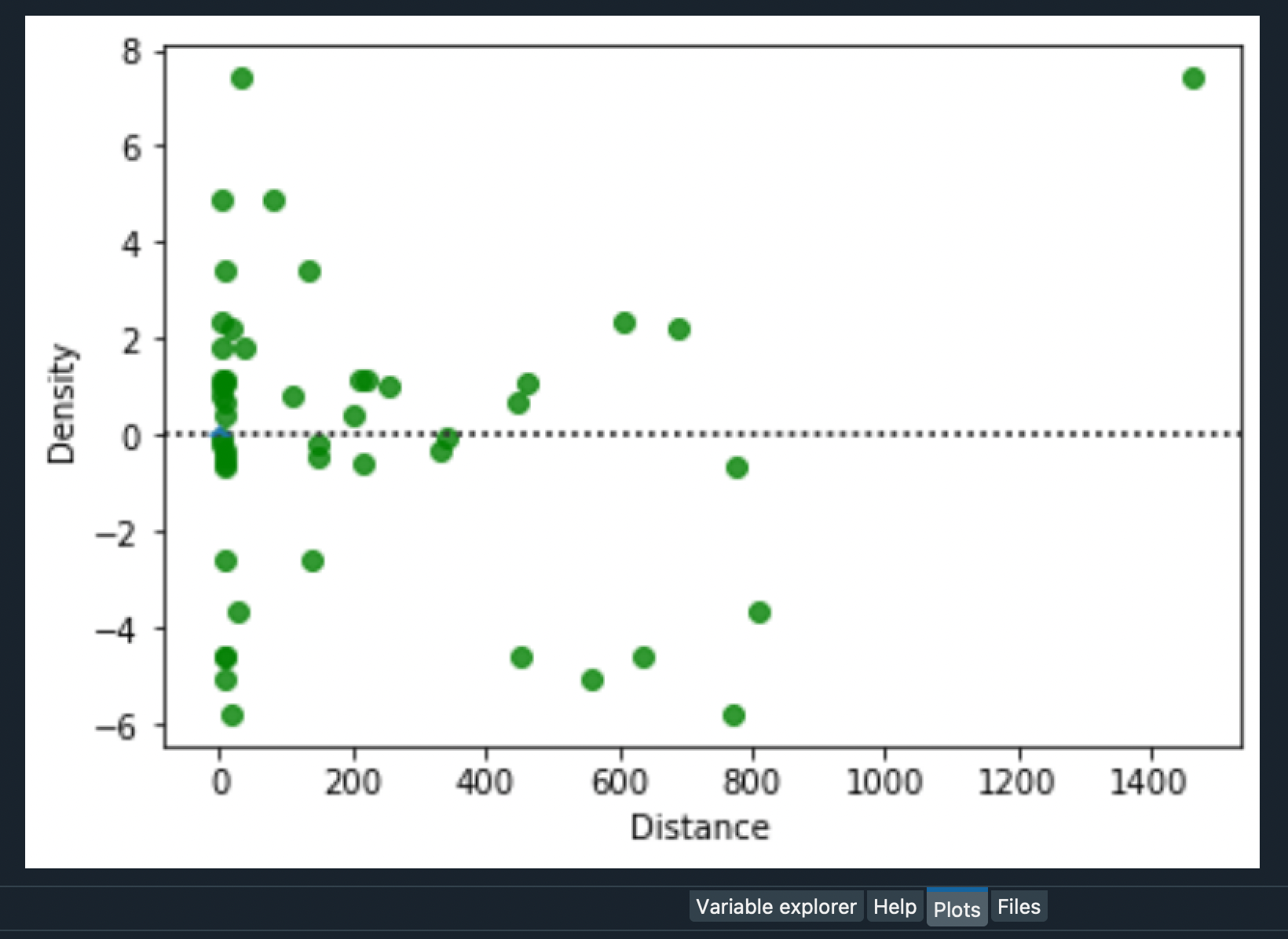
The estimated residual standard deviation is 3.2594734475800946, the p-value for Cases is approximately 0, and the p-value for Distance is 0.001.

R-squared of this model is 0.960, it is mean that the fitting degree of this model can reach 96.0%, indicating that the fitting degree is good. P-value is 4.69\*, therefore, we can think that the number of cases of product stocked and the distance walked by the route driver can better predict the delivery time. The smaller p-value corresponding to Case indicates that the model with Cases have predictive advantages over the model without Cases, same as the model with Distance have predictive advantages over the model without Distance.

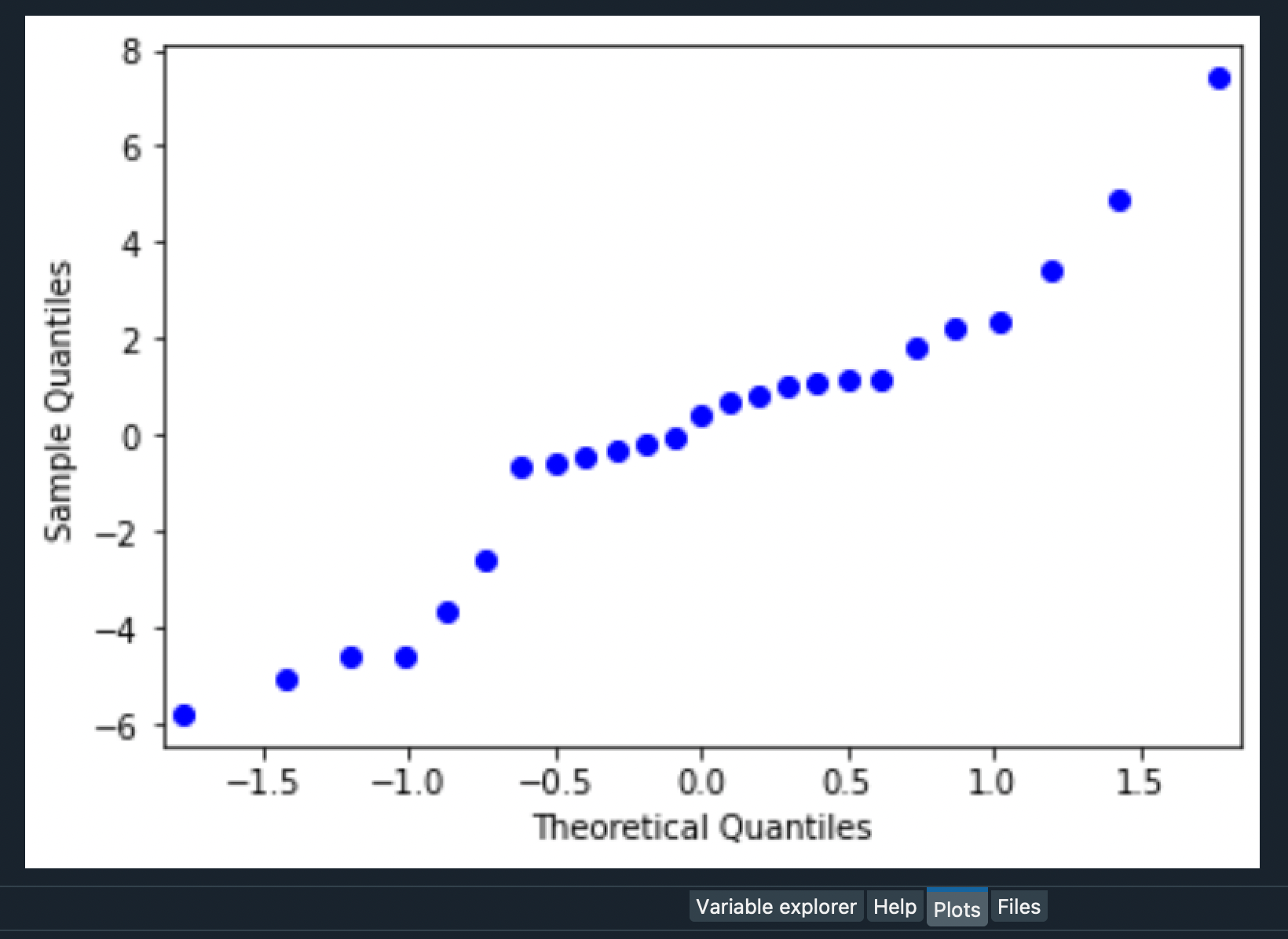


Answer (b):

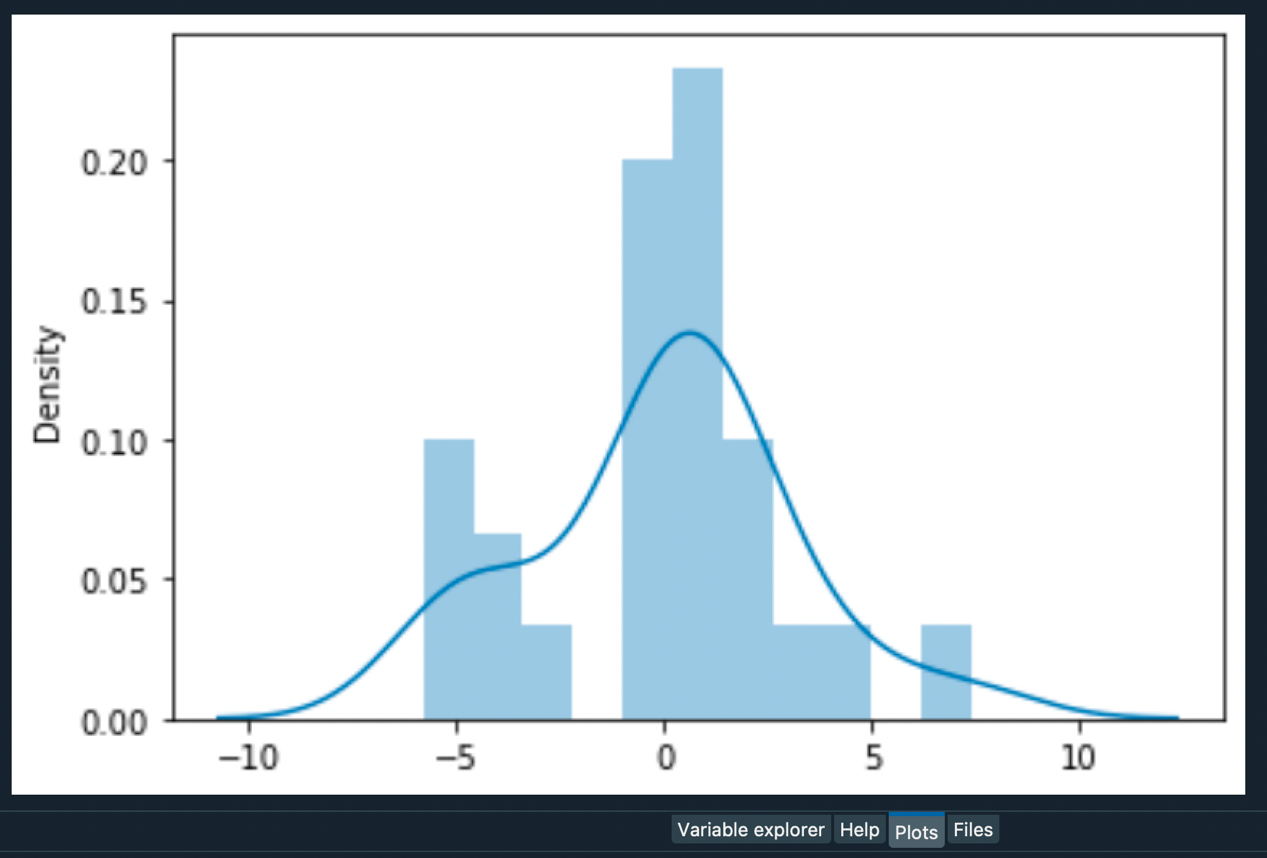
For a good model, its residuals should be points with randomness and unpredictability, forming a graph close to the normal distribution. However, in our residual graph, the residuals are not distributed very evenly on both sides of 0.



So we plotted q-qplot. If the residuals conform to the normal distribution, the points in q-qplot would form a straight line state. But our point is obviously not, so our residuals don't fit a normal distribution.

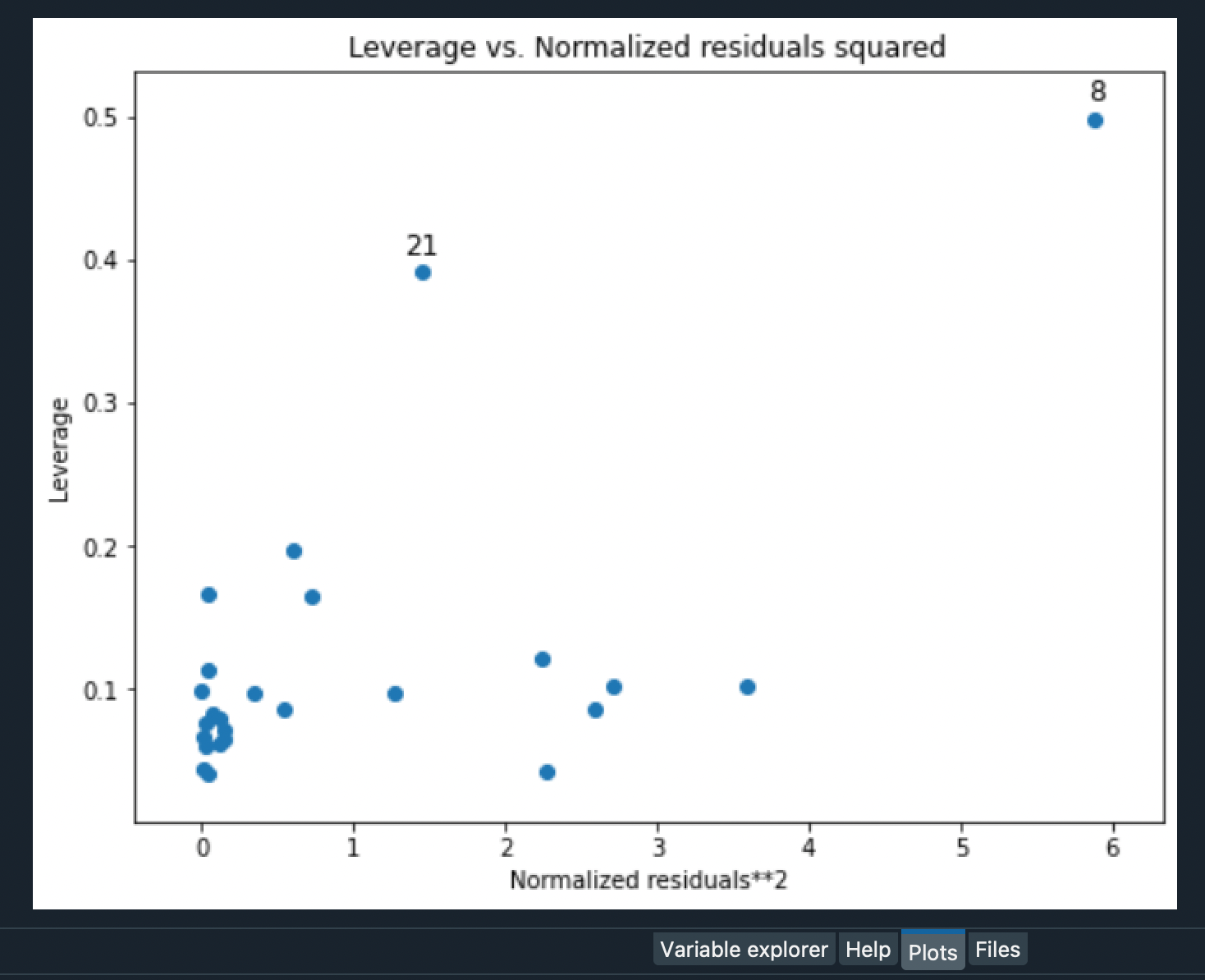


Finally, we draw the histogram of the residuals to prove our conclusion.



Answer (c):

The most influential observation is 8, and the second most influential observation is number 21.



Q4:

Answer:

(y|) =

(Y|) = (|) = =

() =

(|Y) =

=

=

=



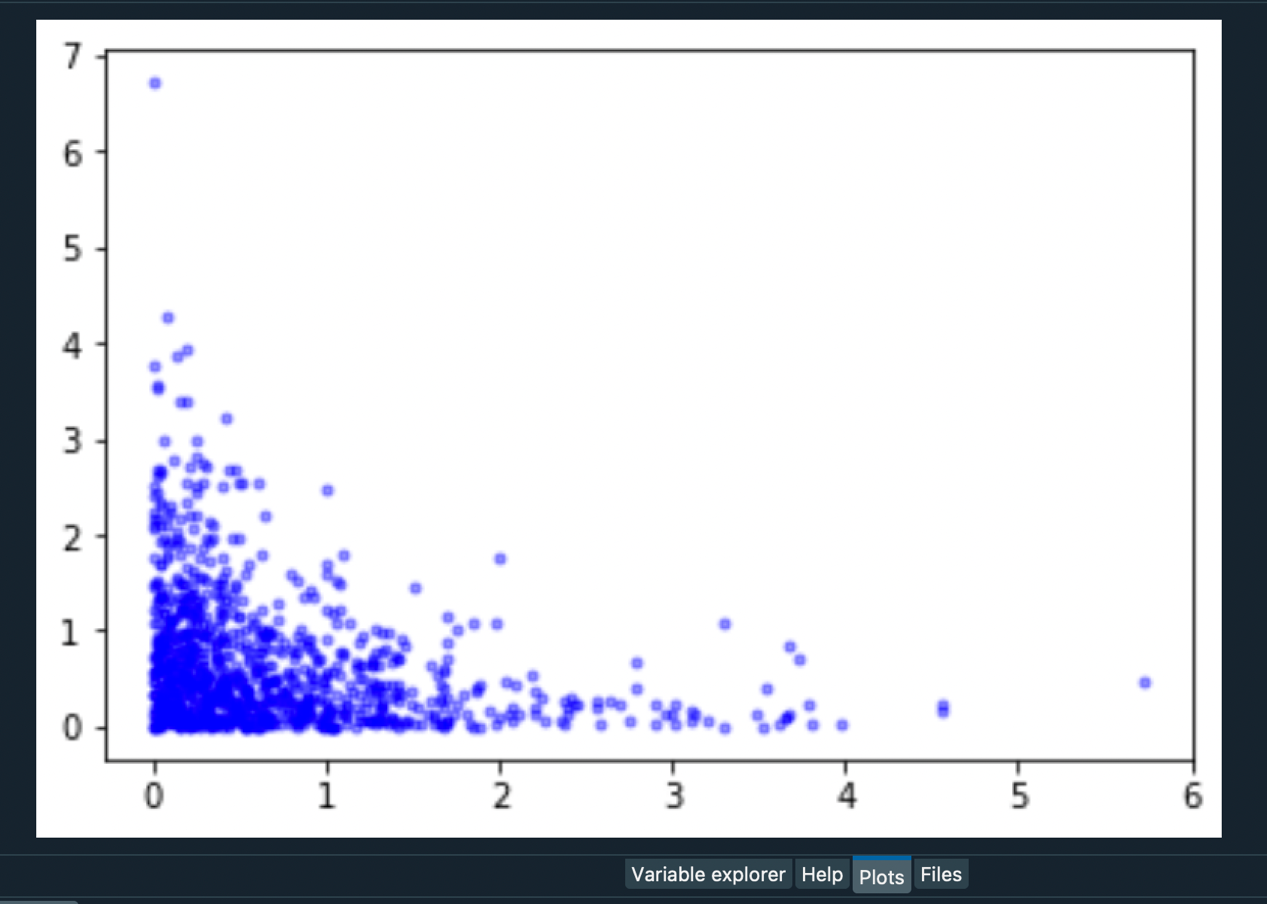
Therefore, derive the posterior distribution of is the Pareto distribution Pareto().

Q5:

Answer (a):

Answer (b):

From the above formula, it can be seen that conforms to the exponential distribution of the parameter , the same as conforms to the exponential distribution of the parameter .



**Code Appendic**

**Question 1**

**Question 2**

**Question 3**

**Question 5**