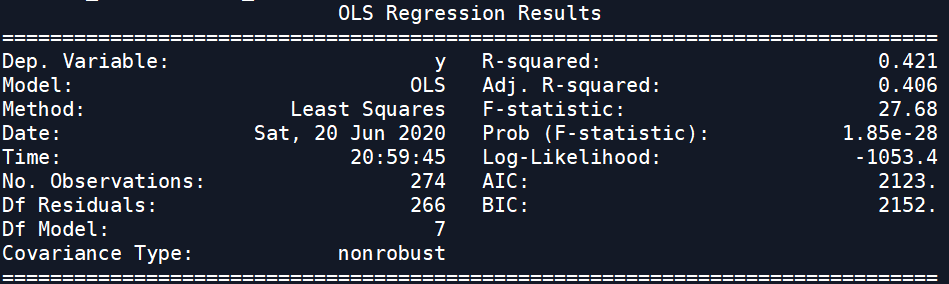
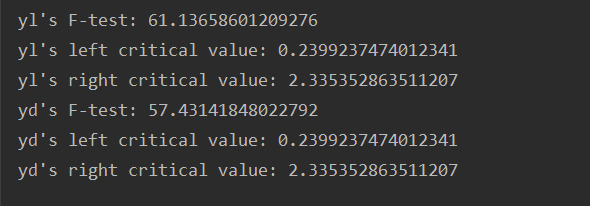
**Multiple Linear Regression**





**1) Model Test**

In ftest.py, we take a Two Tailed Test of the multiple linear regression. Outputs are attached with the document as pictures. （请将上面的两张图片放在附录中）

1. Hypothesis (Two Tailed Test)

H0: b1 = b2 = b3 =b4 = b5 = b6 = b7 = 0

H1: At least there is a non-zero coefficient.

1. Decision Rule

First, We choose f\_test as test statistics, which is in F Distribution with freedom of (7, 275-7-1).

Second, the critical values at the significance of 0.05 are:

P(f<f(left critical value)) = 0.025, P(f>f(right critical value)) = 0.025

Then, we import f of scipy.stats. The left critical value is about 0.23992, and the right critical value is about 2.33535.

1. Test Statistics

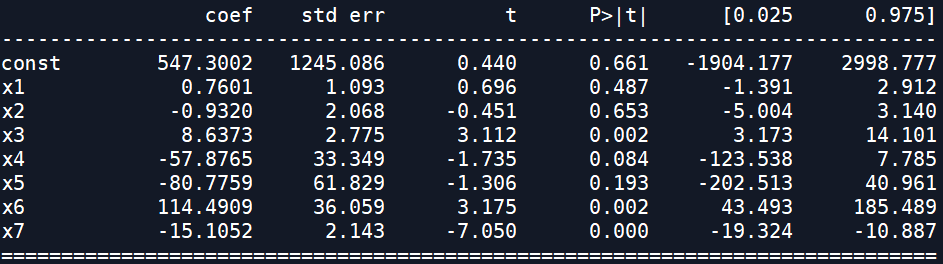
f\_test could be calculated from (SSR / p) / (SSE / (n - p - 1)). SSR is Sum of Squares due to Regression, SSE is Sum of Squares due to Error.

From ftest.py, f\_test = 57.43142.

1. Consion

From b and c, we can find that f\_test (57.43142) is much larger than right critical (2.33535) value. It implies that the multiple linear regression based on seven predictors is bad, since we do not have 0.05 significance to refuse H0.

**2) Coefficient Significance**

Similarly, we take Two Tailed Tests of b1- b7, respectively. By importing and calling summary() from statsmodels.api, we got p-values of coefficients’ test statistics. Output is shown as follows. 

From column “P>|t|”, the significance of rejecting hypothesis “coefficient=0” is:

(line) x1—0.974

(authors) x2—0.694

(shas) x3—0.004

(first time) x4—0.168

(last time) x5—0.386

(average time) x6—0.004

(total\_shas) x7—0.000

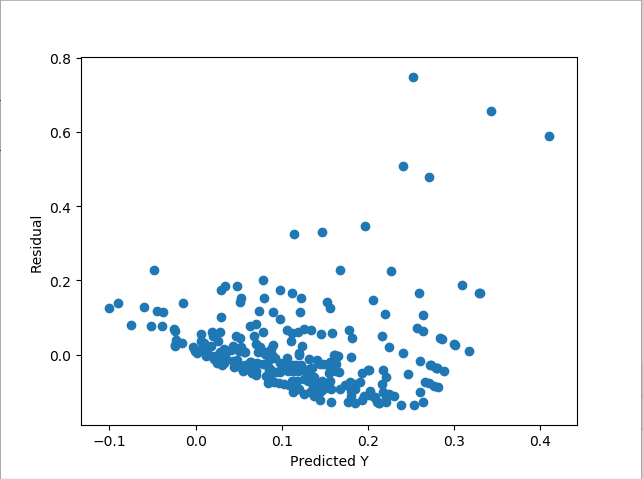
total\_shas has the minimal significance and is high relative to fixes percent. average time and shas come second in significance. On the other hand, the probability of line is closed to 1, which may imply that the file line has no relationship with fixes percent. authors also has relatively high significant in this multiple linear regression.

In conclusion, it suggests that total\_shas and average time are essential in linear regression, while lines and authors are redundant.

Besides, considering total shas represents the total commits of files from v3.0 to HEAD, shas is the number of commits among existing lines, and both have low significance, total\_shas and shas may have dependency, which needs to be explored furtherly.

**3) Residual Chart**

The residual chart is shown as follows.

Summary：

**Multiple Linear Regression**

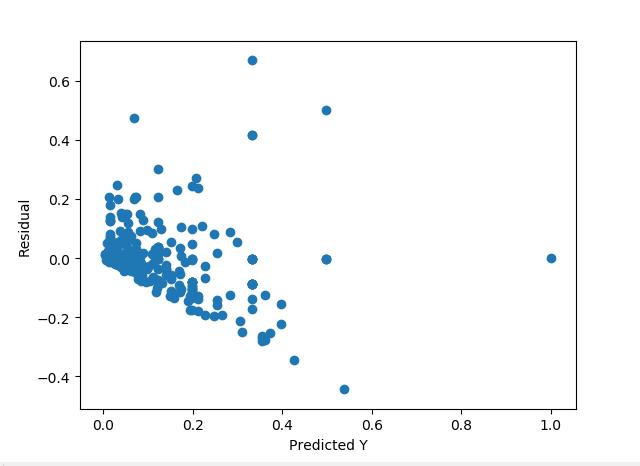
In this photo, the Residual and the Predicted Y show a non-linear trend and are not independent of each other.

Solution：

It would be better to change a new curve inertia instead of linear regression. Also, Y may have an autocorrelation.

**Decision Tree**

**1) Residual Chart**

The residual chart is shown as follows.

**Decision Tree**

Summary：

This figure shows that the relationship between the residual value and the predicted value Y is linearly abnormal, and the variance of the residual is not the same, which increases as the predicted value of y increases.

Solution:

This requires transforming Y or other solving methods to eliminate this effect.

Version used in the project:

matplotlib 3.2.1

numpy 1.18.2

pandas 1.0.3

scipy 1.4.1

setuptools 41.2.0

six 1.14.0

sklearn 0.0

subpreocess

unicodedata

re

time

math

seaborn