Machine Learning: Homework #2

Due on March 30, 2018 at 23:59

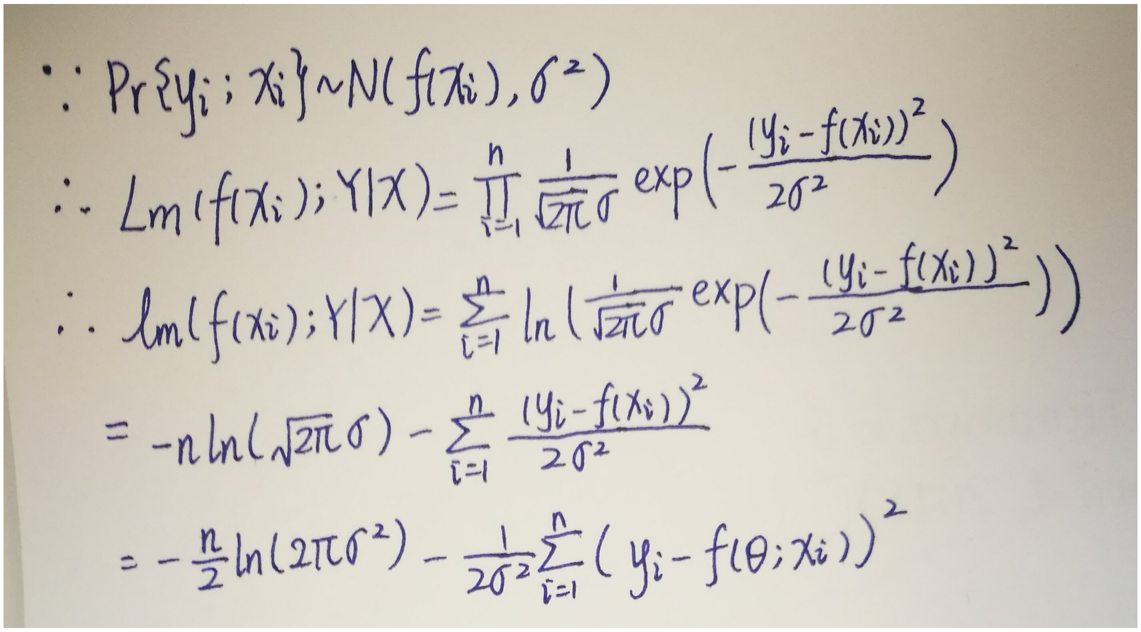
*Professor ***

Name: Zhang ye di ID: **

Problem 1

The relationship between the maximum likelihood and distance metric

Part One



$$\because \Pr\{y_i ; x_i\} \sim N(f(x_i), \sigma^2)$$

$$\therefore Lm(f(x_i); Y|X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - f(x_i))^2}{2\sigma^2}\right)$$

$$\therefore lm(f(x_i); Y|X) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - f(x_i))^2}{2\sigma^2}\right)\right)$$

$$= -n\ln(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{2\sigma^2}$$

$$= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(\theta; x_i))^2$$

Figure 1: the log likelihood function of θ

Part Two

When we maximize the log likelihood function, we minimize the loss function in MSE at the same time. So MSE is a good loss function.

Part Three

Once there exist a outlier (x', y'), when we calculate the loss function in MSE, $(y' - f(x'))^2$ will have a massive influence on the loss, what's more, in each iteration, the outliers will also have a big impact on the gradient we calculate in MSE. For example, in Homework 1 (Fig.2), the red line is the regression result considering the two outliers with MSE method, and the blue one is the result we get after drop the outliers. It is obvious that the two outliers make the linear function more closer to the x-axis.

Problem 2

Linear Regression via Gradient Descent Method

Part One

We use y to denote the university GPA, and x high school GPA. Since there is only one

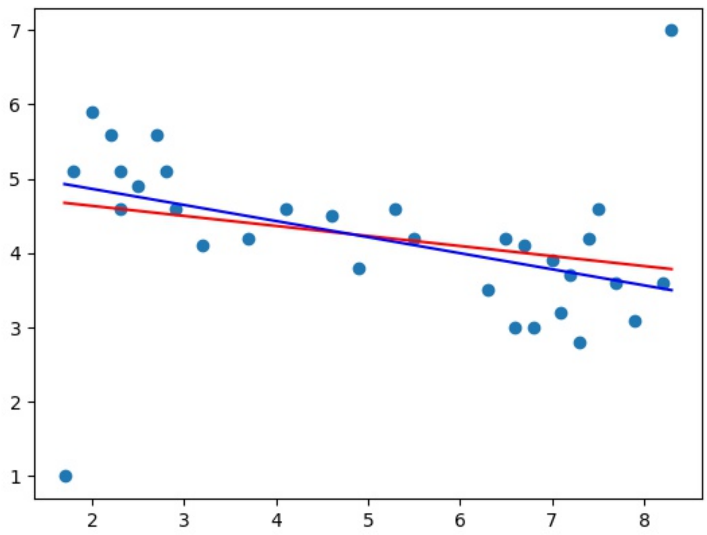


Figure 2: the impact of outliers in MSE

variable that impacts our objective result(university GPA), we formulate the question as:

$$f(x) = \theta_0 + \theta_1 x$$

f(x) is the predicted university GPA. Then we give a function like: $f(\mathbf{x}) = \theta^T \mathbf{x}$, supposed that we have m samples, and (x^j, y^j) is the j^{th} sample. $\mathbf{x} = [1\ x]^T$, $\theta = [\theta_0\ \theta_1]^T$, $f(\mathbf{x}^j) = \theta^T \mathbf{x}^j$. And the cost function $J(\theta)$ is:

$$J(\theta) = \frac{1}{2m} \Sigma_{j=1}^{m} (\theta^T \mathbf{x}^j - y^j)^2$$

Part Two

The termination criterion is: **the iteration time reach to the maximum value (iter=max-iteration=1000)**.

However, the result shows that after almost 100 iterations, the loss function can converge to a good(very small) value.

Part Three

The fitted curve and the convergence result are shown in Fig.3 and Fig4. respectively.

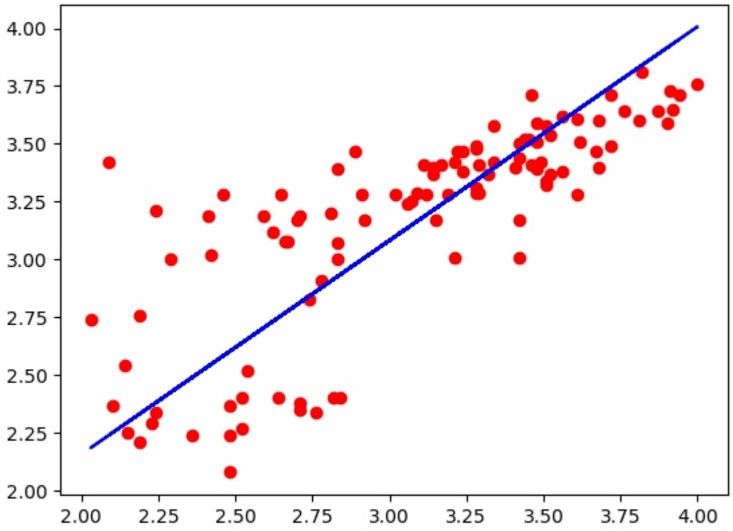


Figure 3: the fitted curve

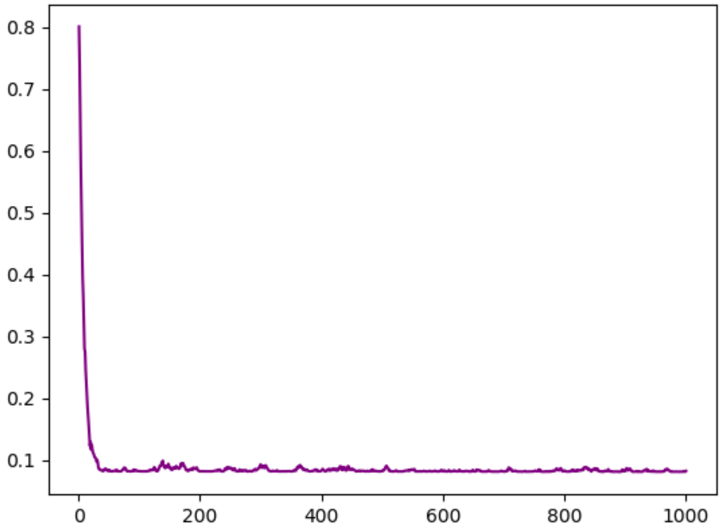
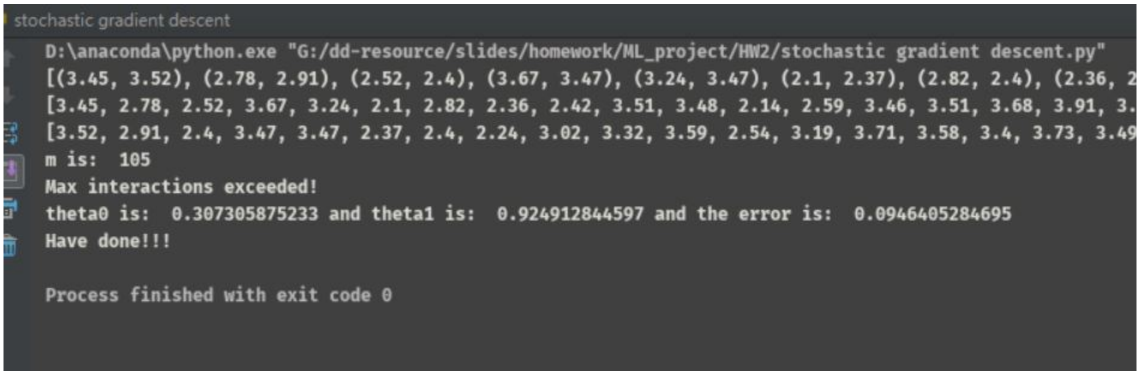


Figure 4: the convergence result

Our linear regression result is (shown in Fig.5):

$$\theta = [0.3073\ 0.9249]^T \quad J = 0.0946$$



```
stochastic gradient descent
D:\anaconda\python.exe "G:/dd-resource/slides/homework/ML_project/HW2/stochastic gradient descent.py"
[(3.45, 3.52), (2.78, 2.91), (2.52, 2.4), (3.67, 3.47), (3.24, 3.47), (2.1, 2.37), (2.82, 2.4), (2.36, 2
[3.45, 2.78, 2.52, 3.67, 3.24, 2.1, 2.82, 2.36, 2.42, 3.51, 3.48, 2.14, 2.59, 3.46, 3.51, 3.68, 3.91, 3.
[3.52, 2.91, 2.4, 3.47, 3.47, 2.37, 2.4, 2.24, 3.02, 3.32, 3.59, 2.54, 3.19, 3.71, 3.58, 3.4, 3.73, 3.49
m is:  105
Max interactions exceeded!
theta0 is:  0.307305875233 and theta1 is:  0.924912844597 and the error is:  0.0946405284695
Have done!!!

Process finished with exit code 0
```

Figure 5: the convergence result

Problem 3

Multivariable Linear Regression

Part One-1

Here we use y to denote the car price, and x for the variables that impact the car price. Since in the first question there is only one variable considered, mileage, that impacts the car price, we can formulate the linear equation $f(\theta;x)$ like what we have done in Problem 2:

$$f(\theta;x) = \theta_0 + \theta_1 x$$

Fix θ, $f(\theta;x)$ is a predict price when given a mileage value x. Assume there are m samples, and (x^j, y^j) is the j^{th} sample. $\theta = [\theta_0, \theta_1]^T$, $\mathbf{x} = [x_0, x_1]^T$, $x_0 = 1$. $f(\mathbf{x}) = \theta^T \mathbf{x}$. And the cost function $J(\theta)$ is:

$$J(\theta) = \frac{1}{2m} \Sigma_{j=1}^{m} (\theta^T \mathbf{x}^j - y^j)^2$$

The termination criterion is: iteration time get the maximum value (iter=max-iteration=10000)

Here is our regression result: the fitted curve and convergence result are shown in Fig.6 and Fig7. respectively. (Note that, in this SGD method, we use batch-SGD, and set $|batch| = 10$, what's more, when calculate the convergence, we plot the average of every 50 loss-values. It can be checked in the code.)

Our linear regression result is (shown in Fig.8):

$$\theta = [2.5046, -0.1716] \quad J = 0.4784 \quad R^2 = 0.01954 \approx 2\%$$

i.e. **Equation 1: Price** $= 25046 - 0.1716$ **mileage**

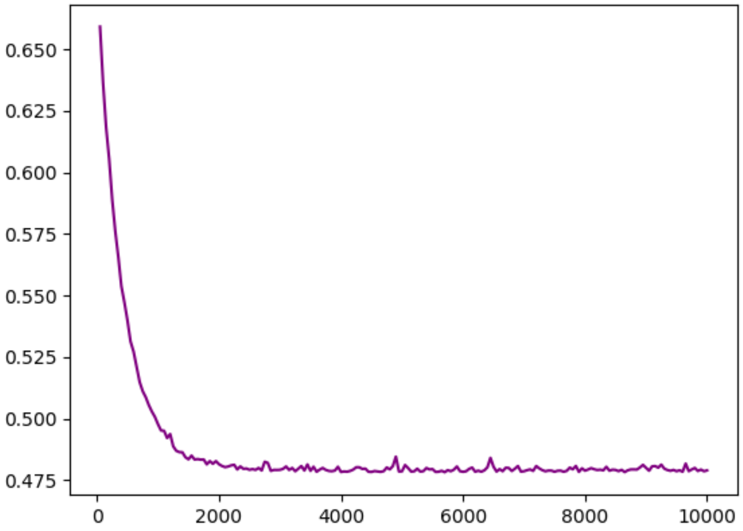


Figure 6: the fitted curve in the first linear regression equation

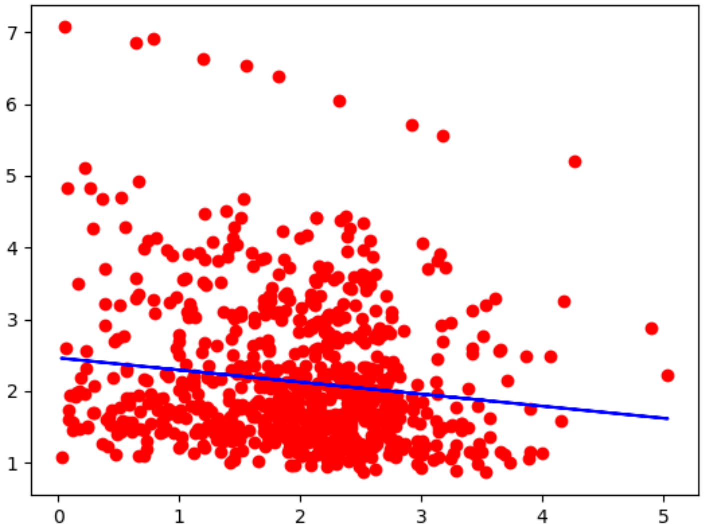
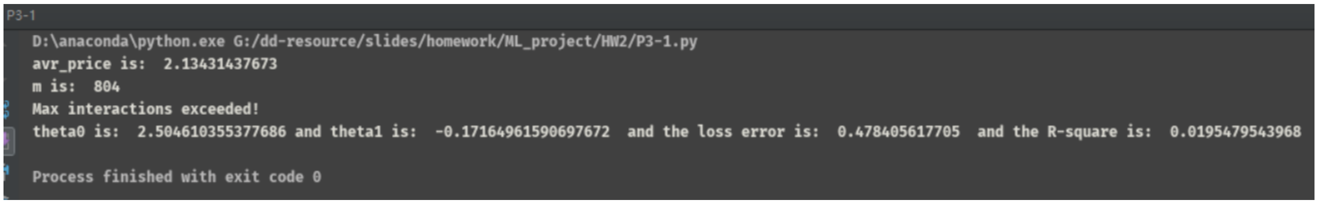


Figure 7: the convergence result in the first linear regression equation

Part One-2

Since the R^2 is 0.2, far less than 0.8, this equation is not a very good fit for the data. What's more, the histogram of the residuals in Fig.9 also shows the residuals aren't center on zero.



```
D:\anaconda\python.exe G:/dd-resource/slides/homework/ML_project/HW2/P3-1.py
avr_price is:  2.13431437673
m is:  804
Max interactions exceeded!
theta0 is:  2.504610355377686 and theta1 is:  -0.17164961590697672  and the loss error is:  0.478405617705  and the R-square is:  0.0195479543968

Process finished with exit code 0
```

Figure 8: θ and R^2 in the first linear regression equation

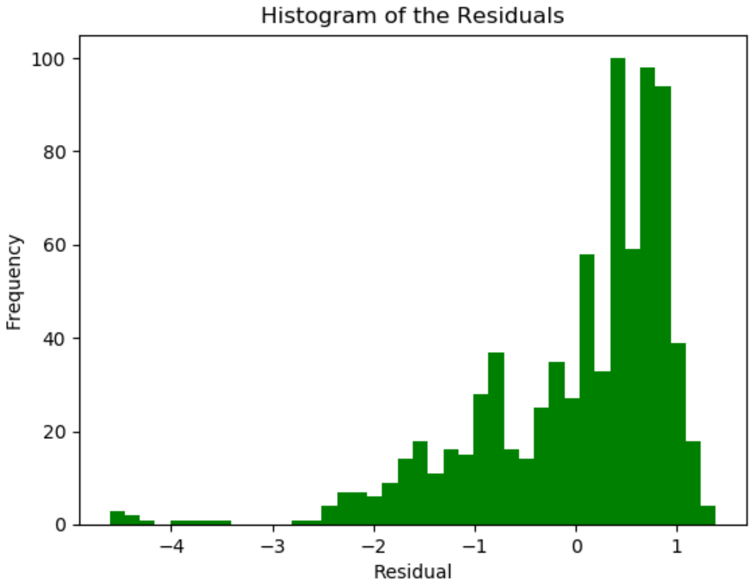


Figure 9: Histogram of the residuals in the first linear regression equation

Part One-3

Since the R^2 value is quite small, we can say the model is not a very good fit. It seems that mileage is not so important. However, mileage still have an impact on the price. When the amount of mileage gets bigger, the price will get lower. For example, in this model, the price can change a lot if two cars are identical except one has 50,000 more miles. So although the mileage has a small impact on the car price, we can not ignore it. So to deal with such a contradictory, we can introduce more variables in our regression model.

Problem 3 continued on next page...

Part Two-1

Again, we use y to denote the car price. Now we have multiple variants that affect the price. The linear regression equation should be like:

$$f(\theta;\mathbf{x}) = \theta^T\mathbf{x} \quad \mathbf{x} = [x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ x_0]^T \quad \theta = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_0]^T$$

Fix $x_0 = 1$. Here we have another 7 variables in each sample $\mathbf{x}$: $x_1 = \text{Mileage}, x_2 = \text{Cylinders}, x_3 = \text{Liters}, x_4 = \text{Doors}, x_5 = \text{Cruise}, x_6 = \text{Sound}, x_7 = \text{Leather}$. And θ_0 is a bias. $\mathbf{X}_{m\times 8} = [\mathbf{x}^1\ \mathbf{x}^2\ \cdots\ \mathbf{x}^m]^T$, $\mathbf{y} = [y_1\ y_2\ \cdots\ y_m]^T$. The loss function $J(\theta)$is :

$$J(\theta) = \frac{1}{2m}||\mathbf{X}\theta - \mathbf{y}||_2^2$$

We still use batch-SGD method to do linear regression. In our code, $|batch| = 10$, and when showing the convergence result, we plot the average of each $|base| = 50$ samples. The convergence result is shown in Fig.10

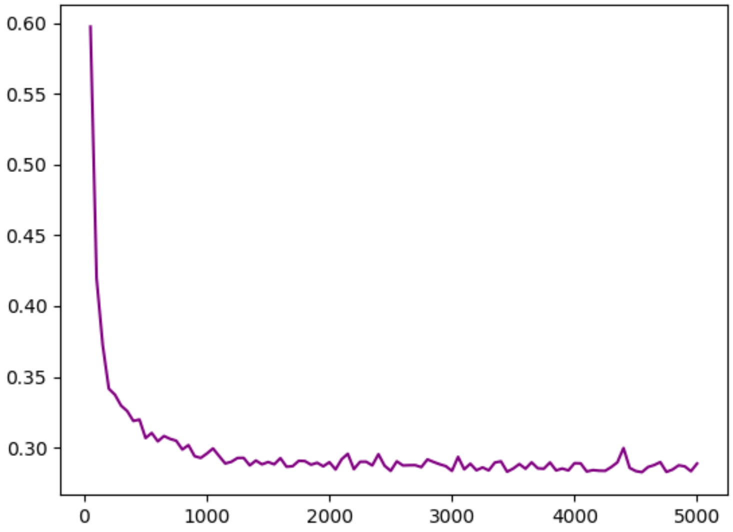
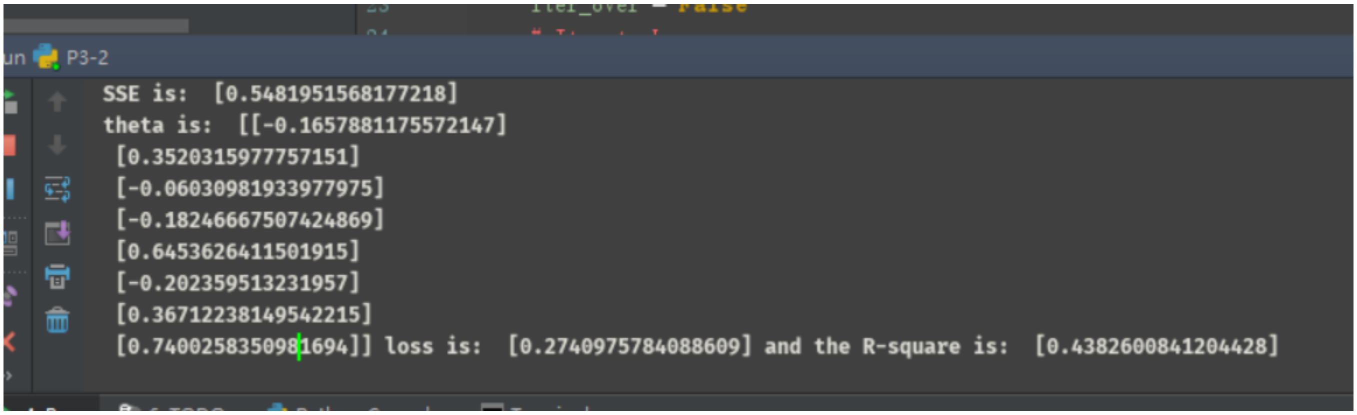


Figure 10: the convergence result in the **second** linear regression equation

The linear regression equation is as follows, shown in Fig.11 Since we processed the data **price** and **mileage** at the beginning, dividing them by 10000. So the result is:

$$\theta = [\text{-0.1658 3520 -603 -1824 6453 -2023 3671 7400}]^T \quad R^2 = 0.4382 \approx 44\%$$

i.e. **Equation 2: Price** $= 7400 - 0.1685$ **mileage** $+ 3520$ **Cylinders** $- 603$ **Liters** $- 1824$ **Doors** $+ 6453$ **Cruise** $- 2023$ **Sound** $+ 3671$ **Leather**



```
SSE is:  [0.5481951568177218]
theta is:  [[-0.1657881175572147]
 [0.3520315977757151]
 [-0.06030981933977975]
 [-0.18246667507424869]
 [0.6453626411501915]
 [-0.202359513231957]
 [0.36712238149542215]
 [0.7400258350981694]] loss is:  [0.2740975784088609] and the R-square is:  [0.4382600841204428]
```

Figure 11: θ and R^2 in the **second** linear regression equation

Part Two-2

Since the $R^2 \approx 44\%$ is get larger than before, we can say such an equation is a better fit for this data than last fit (only consider mileage). And the histogram of the residuals in Fig.12 shows the residuals more center on zero than the last fit shown in Fig.9. But it seems that it is still not a very good fit for the data.

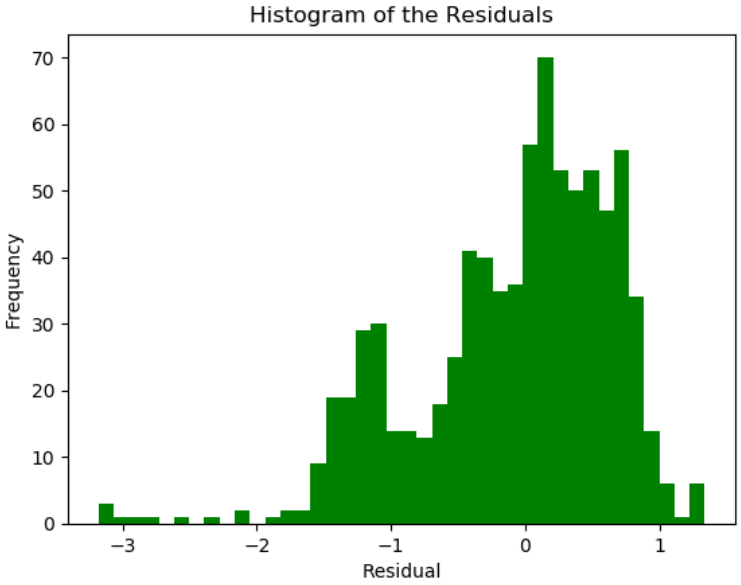


Figure 12: Histogram of the residuals in the second linear regression equation

Part Two-3

In this part, we consider all the variables list in the excel file: **Price, Mileage, Make,Model, Trim, Type, Cylinder, Liter, Doors, Cruise, Sound, Leather**.

We turn **Make, Model, Trim, Type** into discrete variables first via *pandas.dummy*. Then the new $\mathbf{X}$ is a 804×98 matrix, i.e. $\mathbf{X} \in R^{804\times 98}, \theta \in R^{98\times 1}$.

Use the same method, we get the convergence result like Fig.13. The histogram of

Problem 3 continued on next page. . .

residuals is shown in Fig.14. As we can see, if we consider all the variables that can impact the price, the most residuals are center at zero, which means it is a good fit. What's more, the $R^2 = 0.9978$ shown in Fig.15 also indicates that it is very good fit for the data.

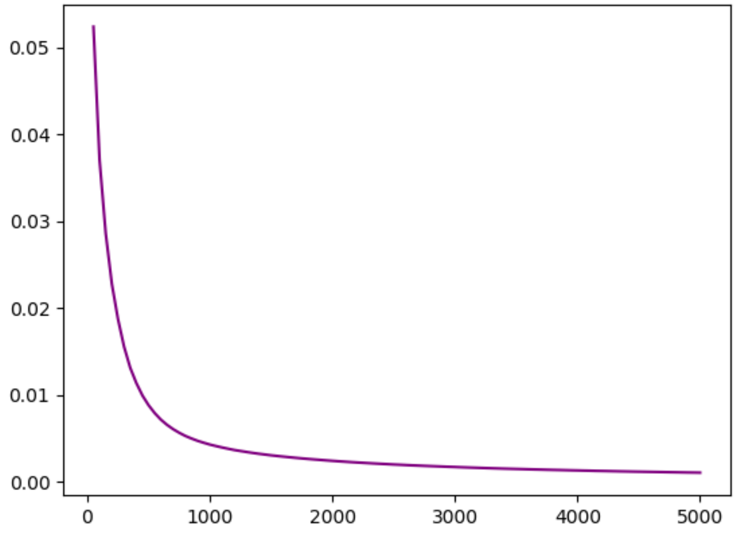


Figure 13: the convergence result in the third linear regression equation

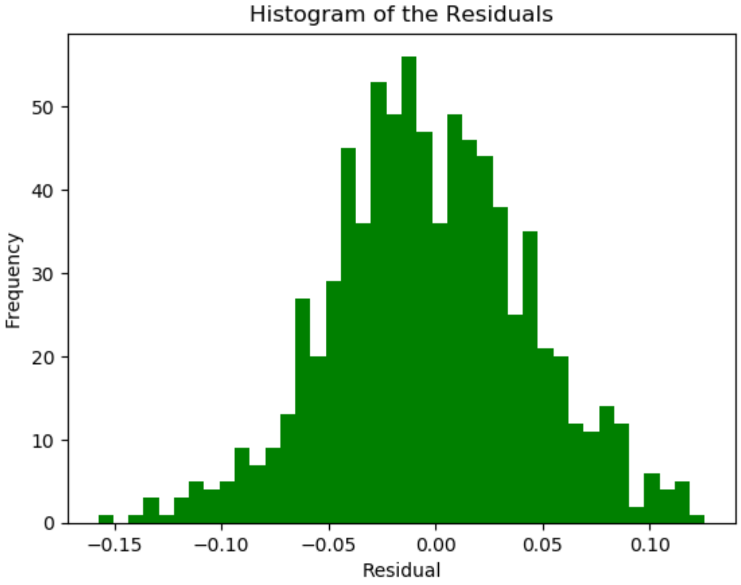
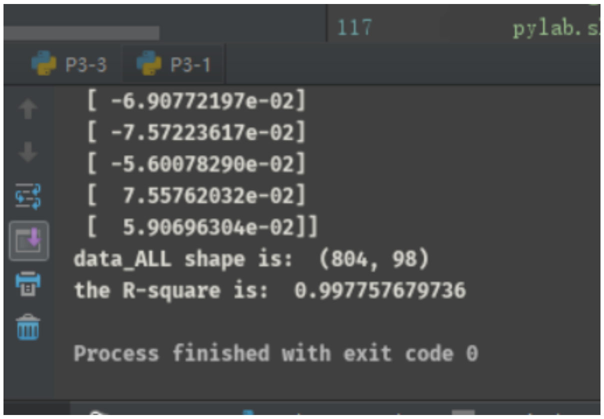


Figure 14: θ and R^2 in the third linear regression equation



```
 [ -6.90772197e-02]
 [ -7.57223617e-02]
 [ -5.60078290e-02]
 [  7.55762032e-02]
 [  5.90696304e-02]]
data_ALL shape is:  (804, 98)
the R-square is:  0.997757679736

Process finished with exit code 0
```

Figure 15: θ and R^2 in the third linear regression equation