**Title:** Data7202 A1 Report

**Name:** Xinqian Wang

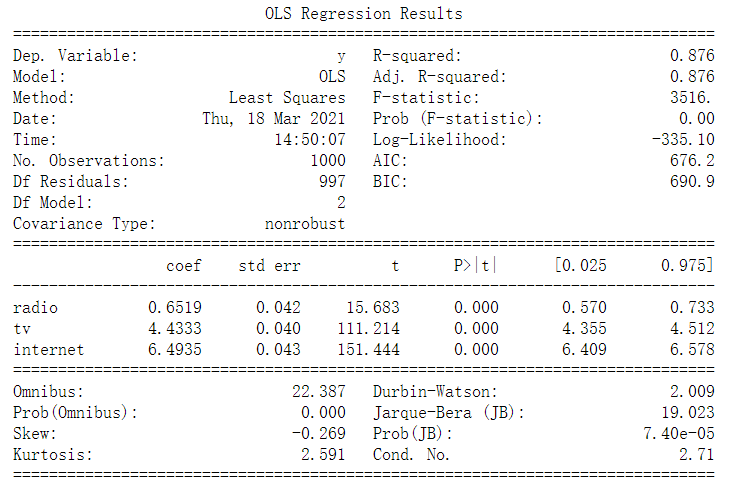
**ID:** s4565489

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

* **1.**

**Question:** For the linear regression, make an inference about the coefficients, specifically, comment about the contributions of different advertisement types to sales.

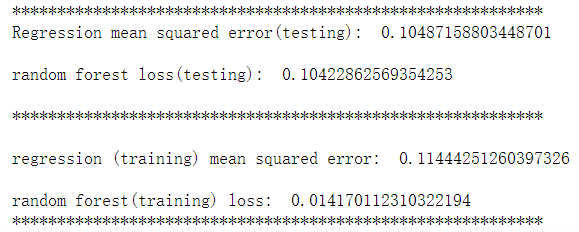
**Answer:**



From the table above we can see from the coefficient that the most relevant advertisement to the sales is the internet, the second is tv and the last is radio.

**Question:** Use the linear model and the RF (with 500 trees), to make a prediction (using the test set), and report the corresponding mean squared errors.

**Answer:**



* **2.**

**Question:** Is this a good method? Do you expect to obtain the true prediction error? Explain your answer.

**Answer:**

No, It's not a good method. The model we got based on the question is a biased model with the correlations average not closed to 0. Hence, the samples have already been saw by the predictors. The correct way to do the K fold CV should be some procedures like this below:

1. Divide the samples into K folds randomly.

2. For each fold, finding a subset of good predictors which show strong correlations with the labels by using all the samples

except in fold k.

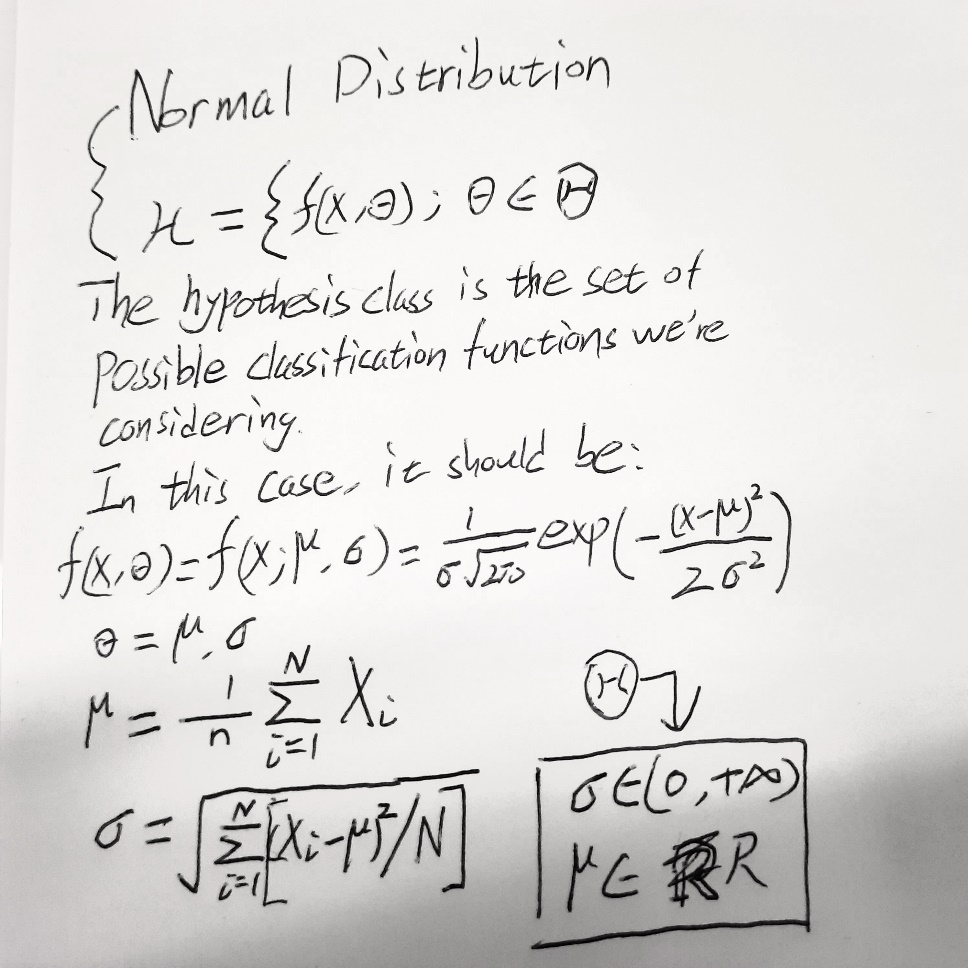
3. Using the subset of these predictors to build a multivariate classifier, using all the samples except those in fold k.

4. Using the multivariate classifier to make predictions on the samples in fold k.

* **3.**

**Question:** For this problem, determine the hypothesis class and state explicitly what is θ and Θ.

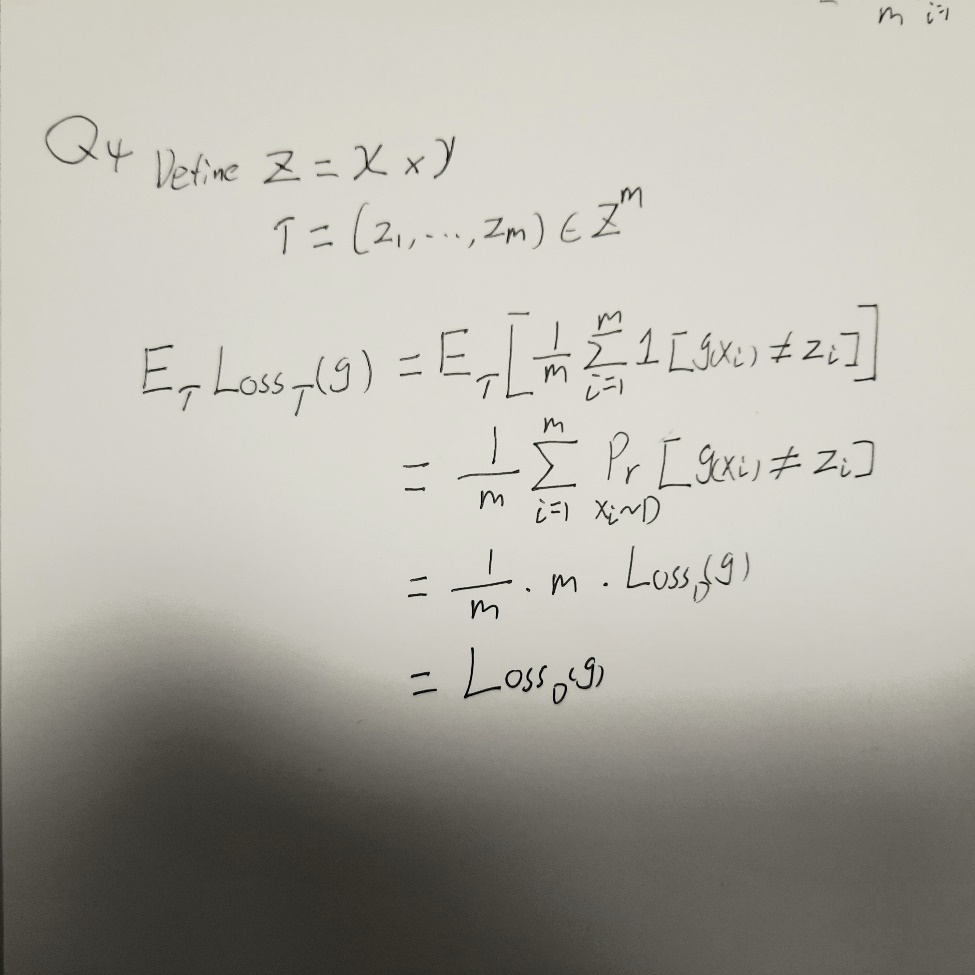
**Answer:**



* **4.**

**Question:** Show that the expected value of LossT (g) over the choice of T equals LossD(g).

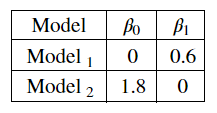
**Answer:**

****

* **5.**

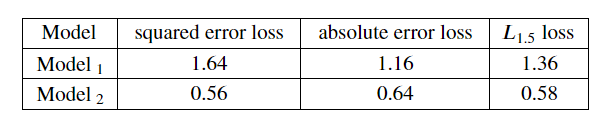
**Question:** Fit these models tot the data and write the corresponding coefficients. Namely, fill the following table:

**Answer:**



**Question:** Consider the squared error loss, the absolute error loss, and the L1.5 loss. Find the average loss for each model. Namely, fill the following table:

**Answer:**



**Question:** Draw a conclusion from the obtained results.

**Answer:**

Only doing simple linear regression seems not fit the label Y so well. We may need to involve some complexity. We can see from the loss value that model 2 would fit better than model 1.

* **6.**

**Question:** Load the data-set and replace all categorical values with numbers. (You can use the LabelEncoder object in Python).

**Answer:**

Attached in the code

**Question:** Generally, it is better to use OneHotEncoder when dealing with categorical variables. Justify the usage of LabelEncoder in (a).

**Answer:**

OneHotEncoder can handle the features with 3 or more unique values. But I think the drawbacks would be it will make the number of columns larger.

However, if there exist the features with only 2 unique values, LabelEncoder would be better.

**Question:** Fit linear regression and report 10-Fold Cross-Validation mean squared error.

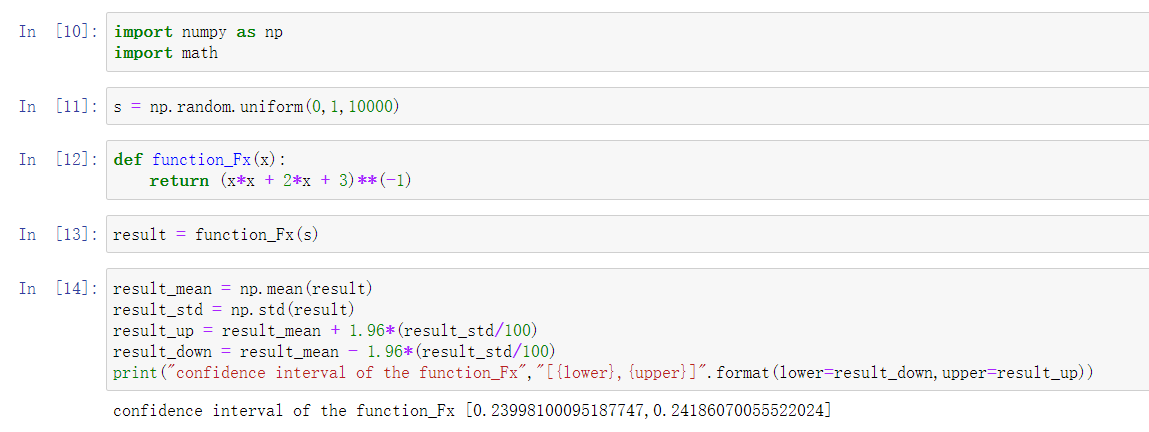
**Answer:**

Attached in the code

* **7.**

**Question:** Deliver the 95% confidence interval.

**Answer:**



**Question:** Compare the obtained estimation with the true value ` as given in (2).

**Answer:**

The result calculate by the Crude Monte Carlo Algorithm is 0.24092085075354872.

The real value calculated by the formula is 0.24030098317248838.

The gap between them is just 0.00062.

**Appendix**

* **1.**

[code](Data7202_A1Q1.ipynb)

* **2.**

None

* **3.**

[Image](DATA7202_A1Q3.jpg)

* **4.**

[Image](DATA7202Q4_A1Q4.jpg)

* **5.**

[code](DATA7202_A1Q5.ipynb)

* **6.**



[code](DATA7202_A1Q6.ipynb)

* **7.**

[code](DATA7202_A1Q7.ipynb)