PS3: EC48E: Fin. Appl. of Machine Learning

Burak Saltoğlu

Instructions

- This is a group project, each group can have up to 3 students. If so, please report the group members in the doc you submitted.
- Our python code in slides are very instructive you may need to modify it.
- Submit your work to Moodle
- Present your work clearly.
- You could use codes used in the class as sample codes.
- For any further questions, you can consult our TA
- Deadline: **by 11:59 PM on November 21, 2024**

Question 1

You are given a data file on GDP, Capital and Labor as we studied in the class. Now let us consider another specification such as $GDP = \beta_0 + \beta_1 log(L) + \beta_2 log(K) + e$. By modifying data given to you:

a. As we did in the class, first estimate the model's parameters and assume that they are the "true population parameters".

b. Find the fitted and actual data. Comment on the goodness of fit of this model.

c. First by generate pseudo random variables by using normal distribution.

d. Rerun the R codes to verify the models estimates.

   i. Now by modifying my code simulate by using various student t distributed random error.
   ii. First start the nu (degree of freedom parameter to be equal to 5)
   iii. Then increase the degree of freedom parameter to 10, 20, 25, 100.

e. Estimate the model parameters and comment on your parameter estimates and compare them with that of true parameters under those cases (normal, t(5), t(10), t(20), t(25), t(100)).

f. Find and $var(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$ and $SE(\hat{\beta})$.

g. Compare parametric SE's with your Monte Carlo average of SE's. (under these 6 random error cases).

h. Set up a confidence interval around estimated parameters of capital and employment (labor) variables with %5 and 1%. Comment on them carefully and compare your findings with the benchmark normal error.

i. Plot the average estimated residuals for the normal and t(5) cases and compare your findings.

j. Find Rsquare and R adjusted_square averages over 100,000 trials.

Question 2

Conventional Validation: Divide the total data into training and test data. Reserve the last 10 observations for testing. Compare the MSE of the training and test sample do the same for the following 4 models. Summarize and comment your results on the basis of your findings.

a. $M1 : GDP = \beta_0 + \beta_1 log(L) + \beta_2 log(K) + e$
b. $M2 : GDP = \beta_0 + \beta_1 L + \beta_2 K + e$
c. $M3 : GDP = \beta_0 + \beta_1 log(L) + \beta_2 K + e$
d. $M4 : GDP = \beta_0 + \beta_1 L + \beta_2 log(K) + e$

Question 3

For the above 4 alternative models use Leave One out Cross Validation. Make a summary table to compare MSE's of each alternative methods. Comment on your findings.

Question 4

For the above 4 models use k-fold Cross Validation assuming k=2 and k=4. Comment on your findings.

Question 5

You may now perform cross-validation on a simulated data set.

a. Generate a random data set as follows:

```
> set.seed(1)
> x=rnorm(100)
> y=x-2*x^2+rnorm(100)
```

b. Create a scatterplot of X against Y. Comment on what you find.

c. Set a random seed, and then compute the Leave One out Cross Validation (LOOCV) errors that result from fitting the following four models using least squares:

i. $Y = \beta_0 + \beta_1 X + e$
ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e$
iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + e$
iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + e$

Note you may find it helpful to use the `data.frame()` function to create a single data set containing both X and Y.

d. Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?

e. Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.

f. Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

Question 6

Bootstrapping: Consider the LSAT and GPA correlation example we studied in the lectures. Remember that the $N_{population} = 82$ and $N_{sample} = 15$.

a. Repeat the same bootstrapping experiment for $N_{sample} \in \{40, 30, 20, 10, 5\}$. What do you see about the average correlation estimates in these samples? **Hint** the population is given. Pick a sample from it by using sample() function.

b. Do you see the any difference between the distribution function of population and bootstrapped samples of the above sizes. Comment on the choice of sample size when you need to do boostrapping.

c. Calculate $99\%, 95\%, 90\%, 10\%, 5\%, 1\%$ quantiles of the correlation coefficient for the population (of $N = 82$) and the above samples. Comment on your findings.

d. Test the null hytpothesis that: $\rho_{SAT,GPA} = 0.75$ (but using $N_{sample} = 18$ only)

e. Create a 95% confindence interval on $\rho_{SAT,GPA}$ (but using $N_{sample} = 18$ only)

Question 7

Central Limit Theorem: CLT states that any sampling distribution from any distribution would converge to normal distribution. the mean of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed Formally, let $X_1, X_2, \ldots, X_n$ be i.i.d. random samples from any distribution with finite mean μ and variance σ^2.

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{1}$$

$$E(\bar{X}) = \mu \tag{2}$$

it can be represented as follows.

$$X \sim N(\mu, \sigma^2/n) \tag{3}$$

You are given a Python code (clt_48e.py). What we discussed in the class. We choose alternative distributions and we show that as we increase the sample size N=1,10,50,100,1000 we can show that X converges to μ. You will do the same for the alternative distributions below.

a. As an alternative distribution t-distribution with 3 degrees of freedom and show that CLT works as N increases (i.e use N=5,10,100,500,1000,10000 and show that $E(\bar{X}) = \mu$ and σ^2/n. Show that both the distribution of sample mean converges to Normal distribution similar to the codes output.

b. Use F distribution with your own choice of degrees of freedom parameters as another distribution.

c. Very briefly explain what have you found.

d. Very briefly discuss the implications of CLT in statistical inference with your own sentences