# Applied Data Analysis

Memory Protocol

03.08.2020

maximum 71points available in total, 24poins required to pass the exam.

duration : 100 min (90 min + bonus 10min)

1. (15 pts) Given is the linear model of the form Y = B \* beta + epsilon\_i, i є {1, 2, 3}. Epsilon~N(0,sigma^2 I)

The design matrix B is the following one:



* 1. Use Least Squares Estimation to calculate beta = [beta\_1\_cap, beta\_2\_cap] if the obtained sample is y\_cap = [1, 2, 1].
  2. epsilon\_1, epsilon\_2, epsilon\_3 are iid and ~ N(mu, sigma^2). Derive the distribution of beta.
  3. ??. If you didn’t calculate the LSE in the a), assume that beta\_1\_cap = 1, beta\_2\_cap = 2.
  4. Test the null hypothesis that beta\_2 = 0 using appropriate two-tailed F-test. matrix forms with explicit numbers are required.

1. Conditional distribution was given like P(X|Y = j) j={1,2} and it was normal EDF. The description was so long and a proper modeling was asked. Do not remember much about this.
2. X~N(mu, sigma^2 \* 2-dim. identity). Given are 4 matrices A\_0 to A\_3:

|  |  |  |  |
| --- | --- | --- | --- |
| A\_0 | A\_1 | A\_2 | A\_3 |
|  |  |  |  |

* 1. (3 pts) Are the following pairs of vectors independent? Justify your answer.
     1. X’A\_0X and X’A\_1X
     2. X’A\_0X and X’A\_2X
     3. X’A\_0X and X’A\_3X
  2. (3 pts) Derive the distributions:
     1. with mu’=(0, 0) for X’A\_0X
     2. with mu’=(-1, 1) for X’A\_1X
  3. (6 pts) With mu’=(-1, 1), are X’A\_0X and X’A\_3X independent? You can use the hint: symmetrize the A\_3 to A\_3\_\* = ½ \* (A\_3 + A\_3’) and apply the covariance formula for symmetric matrices B\_1, B\_2:

Cov(X’B\_1X, X’B\_2X) = 2 \* trace (B\_1\*B\_2) + 4 \* mu’ \* B\_1 \* B\_2 \* mu.

1. ?? pi(Y = 1| X=x) is some value (value itself was not provided). The f^Y(X | Y=j) ~ N(mu, sigma^2).
   1. Show that the regression formula would be logit = beta\_0 + beta\_1 \* x + beta\_2 \* x^2, with logit link log(mu) = log(P(Y=1)) + log(sigma\_0/sigma\_1) + mu\_0^2/(2\*sigma\_0^2) - mu\_1^2/(2\*sigma\_1^2).
   2. Specify constraints on mu and sigma such that the regression function can be simplified to logit = beta\_0 + beta\_1 \* x.
   3. Given the function from the previous task, specify the constraints such that beta\_0 would be >, =, < 0 (each of 3 options), so that logit would be continuously increasing, equal to 0 and continuously decreasing.
   4. ??
2. (10 pts) R code task. You have the dataset Task5.csv. *x1, x2, x3* are continuous features used for prediction and *cat* is the target categorical feature.
   1. Write code to train an analysis of variance model.
   2. Which function produces this output? (ANOVA table follows)
   3. The *summary()* gave this output. The p-value for *x3* is 0.101. What hypotheses it describes?
   4. Based on the previous task, formulate and test the hypotheses at significance level alpha = 0.1.
   5. r code for tukey's test.
   6. Tukey test interpretation from the given picture showing CI of each comparison, asking which ones can be rejected?
3. (8 pts) R code task. You have a Poisson GLM and the dataset Task6.csv (already loaded in the workspace). Write the code for the following:
   1. using rbinom (), getting indies of training set from a given data set.
   2. using the result of a, Sample ⅔ of the data using binomial distribution and save it as a training set
   3. Save the rest as the test set
   4. Train the corresponding Poisson GLM
   5. Use your GLM to predict data
   6. Calculate PRESS

I do not know where the below questions were.

* delta method to calculate asymptotic covariance of sqrt(p1-p2) ). 2x2 covariance matrix was given. It was similar to the Frage 6 in E Test 2.
* Gamma EDF - under given conditions, a corresponding fisher information matrix was asked.