Final Exam A (with solutions)

Advanced Statistics for Data Science

Reichman University
Spring 2022

Instructions

- You have 3 hours to complete the exam.
- The exam contains two parts. Part I contains 8 problems, each has a maximal credit of 5 points. Part II contains 3 questions, each has a maximal credit of 20 points. The maximal number of points in the exam is 100.
- For maximal grade, you should answer *all* problems correctly.
- You may bring to the exam up to two personal two-sided A4 pages containing relevant material.

Part I

For the following problems, either indicate **True** or **False** or fill-in-the-blanks to complete correct statement or answer (whichever applies).

1. (5 points) Let H be the hat matrix for a regression with n observations and p predictors. The underlying design matrix $Z \in \mathbb{R}^{n \times p}$ has full rank. The trace of $H(I - H)$ is ________________
   Explain: ____
   ____.

   Answer: 0, because $H^2 = H$.

2. (5 points) We fit a linear model using ordinary least squares regression and obtain the fitted response $\hat{\epsilon}$. It is possible that
   $$\hat{\epsilon} = \begin{pmatrix} -1 & -1 & 1 & 1 & 1 \end{pmatrix}^\top .$$
   (True/False)

Explain: ____

____.

Answer: If we have an intercept term (one of the columns of Z is a constant) then the answer is false, because we must have $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$ in this case. Otherwise, it may be possible and the answer is true as one can find a specific example.

Additional Explanation: To see why $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$ when the j-th column of Z is a constant, recall that $\hat{\epsilon}^\top Z = 0$ so in particular $\sum_{i=1}^{n} \hat{\epsilon}_i z_{ij} = 0$.

3. (5 points) The random variables X and Y are independent $\mathcal{N}(0,1)$. The distribution of $Y/|X|$ is called ______________ ______________ .

   Explain: ____

   ____.

   Answer: t distribution over 1 degree of freedom.

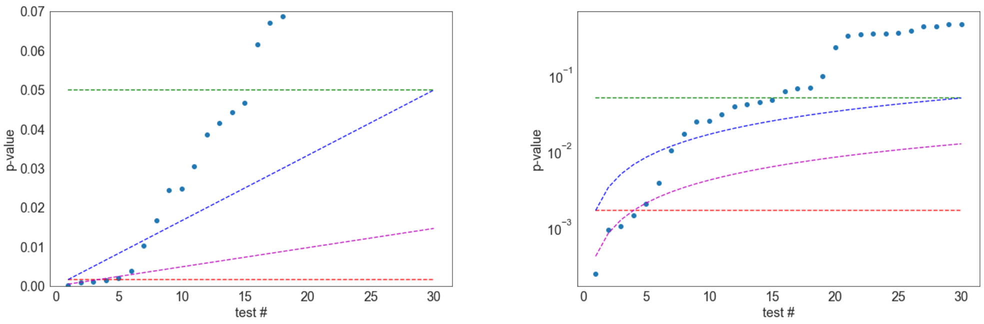
4. (5 points) Suppose we run 10 independent hypotheses tests and obtained P-values $p_{(1)} \leq \ldots \leq p_{(10)}$. If $p_{(1)} = 0.006$ and $p_{(10)} = 0.1$, it is possible that we reject 2 hypotheses after using the Binjamini-Hochberg procedure for controlling the false-discovery rate at level 0.05. **(True/-False)**

   Explain: ____

   ____.

   Answer: True. With BH we reject all P-values until and including $p_{(i^*)}$, where i^* is the largest i for which $p_{(i)} \leq \alpha \cdot i/10$. Consequently, it is possible that we reject 1, 2, or even 10 hypotheses even if $p_{(1)} > \alpha \cdot 1/10$.

5. (5 points) The figures bellow describe sorted P-values obtained from 30 individual hypothesis tests (the only difference between the figures is the scale of the y-axis, which is logarithmic on the right).



We also have the following legend:

curve number	curve description
(1)	$y = 0.05$
(2)	$y = 0.05 \cdot x/30$
(3)	$y = 0.05 \cdot x/(30 \cdot C_{30})$
(4)	$y = 0.05/30$

$(C_m = \sum_{i=1}^{m} i^{-1})$

- The tests selected by Binjamin-Hochberg's (BH) procedure for controlling the false discovery rate (FDR) at level $\alpha = 0.05$ are those whose P-values have ranks ____________ **Answer:** 1-7 .
- The tests selected after a Bonferroni correction to control the family-wise error rate at level $\alpha = 0.05$ are those whose P-values have ranks ____________ **Answer:** 1-4 .
- The tests selected by Binjamin-Hochberg's (BH) procedure for controlling the false discovery rate (FDR) at level $\alpha = 0.05$ for any type of dependency among the tests are those whose P-values have ranks ____________ **Answer:** 1-5 .

(the rank of a P-value p is said to be k is there are $k - 1$ P-values that are smaller than p)

6. (5 points) The cross-validation (CV) residuals sum-of-squares is never smaller than the residuals sum-of-squares. **(True/False)**

   Explain: ____

   ____.

   Answer: True, because $H_{ii} \leq 1$,

$$E_{CV} = \sum_{i=1}^{n} \frac{\hat{\epsilon}_i^2}{1 - H_{ii}} \leq \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

7. (5 points) We fit a linear model with $p = 5$ predictors using least squares and obtain coefficients $\hat{\beta}_j$ for $j = 1, \ldots, 5$. We conduct a t-test for each one of the coefficients to check whether they are different than zero – we obtain that only 2 out of the 5 tests are significant in the sense that the absolute value of their t statistics exceed the $1-\alpha/2$ quantile of the t distribution, where $\alpha \in (0, 1)$ is some significant level. Is it possible that all coefficients will turn out to have significant t-test P-values if we replace each test by a one-sided t-test test that rejects only when the coefficient is significantly *larger* than zero? **(True/False) Explain:** ____________________

   ____.

   Answer: True. The situation in which this may occur is when all t values are larger than zero, large enough so that $p_i^{\text{one-sided}} = \Pr(t_i \geq t_{n-p}) < \alpha$ but not too large so that $p_i^{\text{two-sided}} = \Pr(|t_i| \geq t_{n-p}) = 2p_i^{\text{one-sided}} > \alpha$.

8. (5 points) We examine a linear model with 5 predictors. Below are three tables, each potentially describing a path of a model/variable selection procedure for our model. Which of the following paths may correspond to a *backward* step-wise selection procedure?

R^2	variables included
0	$\emptyset$
.3	$\{2\}$
.5	$\{2,3\}$
.6	$\{2,3,5\}$
.62	$\{2,3,5,4\}$

R^2	variables included
.85	$\{1,2,3,4,5\}$
.81	$\{1,2,3,4\}$
.79	$\{2,3,4\}$
.78	$\{2,3\}$
.785	$\{2\}$

R^2	variables included
1	$\emptyset$
.65	$\{2\}$
.6	$\{2,3\}$
.5	$\{2,3,4\}$
.3	$\{2,3,4,5\}$

Explain: ____

____.

Answer: The correct answer is that non of the tables corresponds to a backward step-wise search, although we accepted all answers that indicated the middle table. Indeed, in backward step-wise procedures we remove variables so only the middle table makes sense. The issue with the middle table is that the value of R^2 increases in the last step which is impossible because we removed a variable. This last increase is due to an typo in the question.

Furthermore, since we used the F-test for extra sum of squares as the stopping criterion, the difference in the R^2 values in the first entry of the middle table is somewhat too large compared and would probably trigger a stop in the first iteration.

Part II

The questions below may have multiple sections. You should write your response on a separate piece of paper.

1. (20 points) We consider a balanced 2-group model:

$$y_{1j} = \mu_1 + \epsilon_{1j}, \qquad y_{2j} = \mu_2 + \epsilon_{2j}, \qquad j = 1, \ldots, n$$

(it is called *balanced* because $n_1 = n_2 = n$). The standard assumption $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, $j = 1, 2$, applies. We have the null hypothesis:

$$H_0 : \mu_1 = \mu_2 + 10$$

- Design a level-α test against H_0: Describe the test statistic and explain for what values of this statistic you decide to reject H_0 and why (you can use the quantile function of any of the distributions we have seen in class).
- Repeat the previous item for testing

$$H_0' : \mu_1 = 10\mu_2$$

Answer: We have:

$$\bar{y}_{1\cdot} \sim \mathcal{N}(\mu_1, \sigma^2/n)$$
$$\bar{y}_{2\cdot} \sim \mathcal{N}(\mu_2, \sigma^2/n)$$

Therefore, under H_0,

$$\bar{y}_{1\cdot} - \bar{y}_{2\cdot} - 10 \sim \mathcal{N}(0, \sigma^2(1/n + 1/n)),$$

so

$$\frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot} - 10}{s\sqrt{2/n}} \sim t_{2n-2}.$$

Similarly, under H_0':

$$\bar{y}_{1\cdot} - 10\bar{y}_{2\cdot} \sim \mathcal{N}(0, \sigma^2(1/n + 100/n)),$$

hence

$$\frac{\bar{y}_{1\cdot} - 10\bar{y}_{2\cdot}}{s\sqrt{101/n}} \sim t_{2n-2}.$$

2. (20 points) We observe $y_1, \ldots, y_n$. We are given some $\mu_0 \in \mathbb{R}$ and would like to test the hypothesis

$$H_0 \,:\, y_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma^2), \qquad i = 1, \ldots, n.$$

(i) Propose a test for H_0.

(ii) Express the test's P-value in terms of the quantile function of one of the distributions we have seen in class.

(iii) Suppose that in reality

$$y_i \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2), \qquad i = 1, \ldots, n.$$

Explain what factors affecting your ability to detect $\mu_1 \neq \mu_0$ and how they affect.

Answer: (i) Because we don't know σ^2, we use a one-sample t-test. Set:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

We reject H_0 for large values of t.
(ii) The P-values is

$$p = \Pr(|T| > |t|), \quad T \sim t_{n-1},$$

so we reject if $p < \alpha$. Equivalently, if $|t| > t_{n-1}^{1-\alpha/2}$.

(iii) The ability to detect $\mu_1 \neq \mu_0$ depends on the signal strength $|\mu_1 - \mu_0|/\sigma$ and the number of samples n. This is formally articulated in the power analysis of this test using the noncentral F distribution we have done in class.

3. We would like to compare the quality of two wine series based on a dataset containing scores of many participating wines in many contests. Each series is rated only once in each contest it participated. For each competing wine we record the following variables: series name, contest id, and score. The table below provides a general description of how the data may look like.

series name	contests id	score
Series1	⋮	⋮
Series2	⋮	⋮
Series2	⋮	⋮
Series1	⋮	⋮
⋮	⋮	⋮
Series2	⋮	⋮

(i) Describe a process to decide which series is better. Write out the form of the t statistic for testing this hypothesis. State the null distribution of the t statistic and give conditions under which we reject H_0. Introduce and define the notation you need. We can assume that the measurements are independent normally distributed random variables and that they all have the same variance.

(ii) Suppose that we know that both series have competed in each contest in the dataset. Would that change your process? If yes, explain the new process.

Answer: (i) We use two-sample t-test of the average score in each series. Let n_1 be the number of wines in Series1 and n_2 their number in Series2. We denote by y_{ij} the score of wine j in Series i and by $\bar{y}_{i\cdot}$ the mean over Series i, $i \in \{1, 2\}$. The t-statistic is

$$t = \frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \qquad s^2 = \frac{1}{n-2}\left[\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1\cdot})^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{2\cdot})^2\right].$$

Under the null hypothesis H_0 stating that both wines are of equal quality we have $t \sim t_{n-2}$. We reject H_0 at level 0.05 if $|t| > t_{n-2}^{0.975}$. If we had rejected H_0, we can say that Series1 is better if $t > 0$ and that Series2 is better if $t < 0$.

(ii) Since for each contest we have one score for Series1 and one score for Series2, it makes more sense to consider the difference between these scores and thus eliminate variations across contests. In this case, we take:

$$y_i = \text{score of Series1 in contest } i - \text{score of Series2 in contest } i.$$

Say that the number of contests is n. We define

$$t = \frac{\bar{y}}{s/\sqrt{n}}, \qquad s = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Under the null hypothesis H_0 that the two Series are of equal quality, we have $t \sim t_{n-1}$. We reject the null hypothesis at level $\alpha = 0.95$ if $|t| > t_{n-1}^{0.975}$. In this case, $t > 0$ indicates that Series1 is better while $t < 0$ indicates the opposite.