

Applied Statistics Qualifier Exam

This exam consists of 2 parts. You must answer 5 questions total and must answer at least 2 questions from each part. Make sure to clearly indicate which problems you are attempting. Some formulas and tables are given on the last two pages of this exam.

PART A.

QUESTION A1: Suppose that $y_i = \alpha + x_i + \epsilon_i$, $i = 1, \dots, N$ where $\epsilon_1, \dots, \epsilon_N$ are independent $\text{Normal}(0, \sigma^2)$. Assuming the data points (x_i, y_i) , $i = 1, \dots, n$, are not collinear, find the maximum likelihood estimators of α and σ^2 by differentiating the likelihood function. Justify that your solutions maximize the likelihood function.

QUESTION A2: Suppose we have the following model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \dots, N$$

where $\epsilon_1, \dots, \epsilon_N$ are independent $\text{Normal}(0, \sigma^2)$. Four univariate regression models were run and the following results were obtained:

$$\hat{y} = \frac{27 - 4x_1}{5}, \quad \hat{y} = \frac{227 - 15x_2}{40}, \quad \hat{x}_1 = \frac{21 - x_2}{8}, \quad \hat{x}_2 = \frac{29 - 2x_1}{5}.$$

Find the maximum likelihood estimates of β_0 , β_1 , and β_2 for the multiple regression model.

QUESTION A3: Suppose

$$y_{ijk} = \mu_i + \alpha_{ij} + \epsilon_{ijk}, \quad i, j, k = 1, 2$$

where $\alpha_{11} + \alpha_{12} = \alpha_{21} + \alpha_{22} = 0$ and $\epsilon_{ijk} \sim \text{independent Normal}(0, \sigma^2)$.

Given the data

		<i>i</i>	
		1	2
<i>j</i>	1	10,12	4,3
	2	12,10	8,10

test the hypothesis $H_0 : \alpha_{11} = \alpha_{12} = \alpha_{21} = \alpha_{22} = 0$ at level 0.05.

QUESTION A4: Suppose that $y_i = \alpha + \beta x_i + \epsilon_i$ for $i = 1, 2, 3$ where $\epsilon_1, \epsilon_2,$ and ϵ_3 are independent $\text{Normal}(0, \sigma^2)$.

(a) Given the data points

$$(-1, y_1), (0, y_2), \text{ and } (1, y_3)$$

show that the F -statistic for testing $H_0 : \alpha = \beta$ can be expressed in the form

$$F = C \left(\frac{\hat{\alpha} - \hat{\beta}}{\hat{\sigma}} \right)^2$$

where C is a constant and $\hat{\alpha}, \hat{\beta},$ and $\hat{\sigma}^2$ are the maximum likelihood estimates of $\alpha, \beta,$ and $\sigma^2,$ respectively. Give the value of C .

(b) Given the data points

$$(-1, 2), (0, 4), \text{ and } (1, 12)$$

find a 95% confidence ellipsoid for $(\alpha, \beta)'$. State your answer in the form

$$A_1(\hat{\alpha} - \alpha)^2 + A_2(\hat{\beta} - \beta)^2 + A_3(\hat{\alpha} - \alpha)(\hat{\beta} - \beta) \leq D,$$

giving the values of $\hat{\alpha}, \hat{\beta}, A_1, A_2, A_3,$ and D .

PART B.

QUESTION B1: Suppose $y_i \sim$ independent Normal($\beta x_i, \sigma^2$) for $i = 1, \dots, N$.

(a) For fixed $\lambda > 0$, find the bias and variance of the two estimators

$$\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}, \quad \tilde{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \lambda}.$$

Which estimator has the larger bias? Which estimator has the larger variance?

(b) For $C > 1$, what choice of λ will reduce the variance of $\hat{\beta}$ by a factor of C ? In this case, what happens to the bias of $\tilde{\beta}$?

QUESTION B2: Suppose we have a $m \times 1$ input vector $\mathbf{x} = (x_1, \dots, x_m)'$ which is either from group 0 ($y = 0$) or group 1 ($y = 1$) and we model them according to

$$p_k(\mathbf{x}; \boldsymbol{\beta}) = \Pr(y = k | \mathbf{x}) = \frac{e^{k\boldsymbol{\beta}'\mathbf{x}}}{1 + e^{\boldsymbol{\beta}'\mathbf{x}}}, \quad k = 0, 1,$$

for an unknown $m \times 1$ coefficient vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)'$.

(a) Given training data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$, show that the log-likelihood equation can be expressed as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^N \ln p_{y_i}(\mathbf{x}_i; \boldsymbol{\beta}) = \sum_{i=1}^N \left\{ y_i \boldsymbol{\beta}' \mathbf{x}_i - \ln(1 + e^{\boldsymbol{\beta}' \mathbf{x}_i}) \right\}.$$

(b) Show that the maximum likelihood estimator of $\boldsymbol{\beta}$ must satisfy

$$\mathbf{X}'(\mathbf{y} - \mathbf{p}) = \mathbf{0}$$

where \mathbf{X} is a matrix containing the rows $\mathbf{x}'_1, \dots, \mathbf{x}'_N$, \mathbf{y} is the $N \times 1$ column vector $(y_1, \dots, y_N)'$, and \mathbf{p} is the $N \times 1$ column vector $(p_1(\mathbf{x}_1; \boldsymbol{\beta}), \dots, p_1(\mathbf{x}_N; \boldsymbol{\beta}))'$.

QUESTION B3: Consider the data set with eight observations $(x_{1,i}, x_{2,i}, y_i)$:

$$(1, 2, 1), (2, 5, 0), (3, 3, 1), (4, 5, 0), (5, 1, 1), (6, 0, 0), (7, 4, 1), (8, 3, 0).$$

- (a) Using a classification tree based on both input variables $\mathbf{x}_1 = (x_{1,1}, \dots, x_{1,8})'$ and $\mathbf{x}_2 = (x_{2,1}, \dots, x_{2,8})'$, find the best first split based on misclassification rate.
- (b) After the split in (a) is made, what split should be made next based on misclassification rate? Give all splits which optimize the misclassification rate.

QUESTION B4: Consider a clustering model where each cluster is labeled by an integer 1 through $K \in \mathbb{N}$. The assignments to a cluster are characterized by an encoder $C(i) = k$ which assigns the i th observation to cluster k . We choose our loss function to be the *within-point scatter*

$$W(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(i')=k} \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2.$$

- (a) Show that, for a given cluster assignment C , the K -means algorithm minimizes the within-point scatter.
- (b) Is convergence of the K -means algorithm guaranteed? Why or why not?

FORMULAS:

Normal(μ, σ^2) density: $n(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$

If $\mathbf{y} \sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ and $\mathbf{X}'\mathbf{X}$ is invertible:

- MLE of $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- MLE of $\sigma^2 = \hat{\sigma}^2$ and $N\hat{\sigma}^2/\sigma^2 \sim \chi_{N-p}^2$
- $\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/\sigma^2 \sim \chi_p^2$

- $\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/p}{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(N-p)} \sim f_{p, N-p}$

If, in addition, $\mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$:

- (restricted) MLE of $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_0$
- $\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)\|^2/\sigma^2 \sim \chi_q^2$
-

$$\begin{aligned}
 F &= \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)\|^2/q}{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(N-p)} \\
 &= \frac{(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m})'(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K})^{-1}(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m})/q}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(N-p)} \sim f_{q, N-p}
 \end{aligned}$$

TABLES:

The tables below gives the 95th and 97.5th percentile of the χ^2 distribution with df degrees of freedom.

	95%	97.5%
1	3.841	5.024
2	5.991	7.378
3	7.815	9.348
4	9.488	11.143
5	11.071	12.833

The tables below gives the 95th and 97.5th percentile of the F distribution with $df1$ and $df2$ degrees of freedom.

	95%			
	$df1$			
	1	2	3	4
1	161.447	199.500	215.707	224.583
2	18.513	19.000	19.164	19.247
3	10.128	9.552	9.277	9.117
4	7.709	6.944	6.591	6.388

	97.5%			
	$df1$			
	1	2	3	4
1	647.789	799.500	864.163	899.583
2	38.506	39.000	39.165	39.248
3	17.443	16.044	15.439	15.101
4	12.218	10.649	9.979	9.605