Regression Methods: Mock Exam

18 November 2025

Instructions: The time allotted for the examination is 120 minutes. You may answer in either English or French. No written material may be brought into the examination.

Notation: $a_+ = \max(a, 0)$ for $a \in \mathbb{R}$; $A_{r\times s}$ means that A is an $r \times s$ matrix; $X \sim \mathcal{N}_p(\mu, \Omega)$ means that X has a p-dimensional multivariate normal distribution with mean vector $\mu_{p\times 1}$ and variance matrix $\Omega_{p\times p}$; and $X_{p\times 1} \sim (\mu, \Omega)$ means that $\mathrm{E}(X) = \mu_{p\times 1}$ and $\mathrm{var}(X) = \Omega_{p\times p}$.

Name:

First name:

SCIPER:

Exercise	Points
1	
2	
3	
4	
Total	

Exercise 1 An observation Y has probability function of the form

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\}, \quad y \in \mathcal{Y}, \theta \in \Theta, \phi > 0,$$

where $\mathcal{Y}, \Theta \subset \mathbb{R}$. Find the cumulant-generating function of Y and hence show that $\mathrm{E}(Y) = b'(\theta)$ and $\mathrm{var}(Y) = \phi b''(\theta)$.

1. Explain the terms in italics in the phrase 'the *link function* and *variance function* are key ingredients of a *generalized linear model*'.

2. Independent responses $Y_1, \dots, Y_n$ with corresponding $p \times 1$ covariate vectors $x_1, \dots, x_n$, with $p < n$, follow a logistic regression model. Unfortunately, instead of the Y_j, the noisy responses $Z_j = (1 - I_j)Y_j + I_j(1 - Y_j)$ are observed, where $\Pr(I_j = 1) = 1 - \Pr(I_j = 0) = \gamma \in (0, 1)$, and the I_j are independent, and independent of the Y_j. The probability γ is known from another study.

   Formulate a generalised linear model with responses $Z_1, \dots, Z_n$, and give its link and variance functions. What if $\gamma = 1/2$?

   How would you estimate an unknown γ?

Exercise 2 Independent continuous observations $Y_1, \dots, Y_n$ arise from a regression model

$$Y_j = x_j^{\mathrm{T}}\gamma + \sigma\varepsilon_j, \quad \varepsilon_j \stackrel{\mathrm{iid}}{\sim} F, \quad j = 1, \dots, n,$$

but only the $p \times 1$ covariate vectors $x_1, \dots, x_n$ and the values z_j of the indicator variables $Z_j = I(Y_j > 0)$ are observed.

1. Find the probability that $Z_j = 1$. Are the parameters γ and σ identifiable? Explain your reasoning.

2. Derive the likelihood based on $z_1, \dots, z_n$, and put it into generalized linear model form, identifying the response distribution and the link function in the special cases where (i) $\varepsilon_j \stackrel{\mathrm{iid}}{\sim} \mathcal{N}(0, 1)$, (ii) the ε_j have distribution function $\exp\{-\exp(-u)\}$, for $u \in \mathbb{R}$.

3. Now suppose that the observed variables are the indicators that $Y_j \in \mathcal{I}_k$, where

   $$\mathcal{I}_1 = (-\infty, \zeta_1], \quad \mathcal{I}_2 = (\zeta_1, \zeta_2], \quad \dots, \quad \mathcal{I}_K = (\zeta_{K-1}, \infty), \quad K > 2,$$

   and $\zeta_1 < \cdots < \zeta_{K-1}$ are unknown. In what circumstances would such a model be useful? Give an expression for the corresponding likelihood function. If $x_j^{\mathrm{T}}\gamma = \gamma_0 + \gamma_1 x_j$, which of the parameters are estimable? Explain.

Exercise 3 Consider a linear model $y = X\beta$ with $n \times 1$ response vector y and $n \times p$ design matrix $X = (X_0, X_1, X_2)$ of rank p. Let H_0, H_1 and H_2 respectively denote the 'hat matrices' representing orthogonal projections of $\mathbb{R}^n$ onto the subspaces $\mathcal{V}_0$, $\mathcal{V}_1$ and $\mathcal{V}_2$ of $\mathbb{R}^n$ spanned by the columns of X_0, (X_0, X_1) and X.

1. Give the general form of a 'hat matrix' and show that it is symmetric and idempotent.

2. Show that $H_0H_1 = H_1H_0 = H_0$, give the corresponding equations for H_1H_2 and H_0H_2, and show that $H_1 - H_0$ and $H_2 - H_1$ are orthogonal. What can you say about the orthogonality of $H_1 - H_0$, $H_2 - H_1$, H_0 and $(I_n - H_2)$? What does this imply about the decomposition

$$(I_n - H_0)y = (I_n - H_2)y + (H_2 - H_1)y + (H_1 - H_0)y$$

if the components of y are independent normal random variables with the same variance?

3. The table below shows clotting times (minutes) of blood plasma from eight subjects, treated by four methods.

	Subject							
Treatment	1	2	3	4	5	6	7	8
A	8.4	12.8	9.6	9.8	8.4	8.6	8.9	7.9
B	9.4	15.2	9.1	8.8	8.2	9.9	9.0	8.1
C	9.8	12.9	11.2	9.9	8.5	9.8	9.2	8.2
D	12.2	14.4	9.8	12.0	8.5	10.9	10.4	10.0

Give the form of the design matrix for a linear model in which the mean response for subject s and treatment t is

$$\eta_{ts} = \mu + \alpha_t + \beta_s, \qquad s \in \{1, \ldots, 8\}, \quad t \in \{A, B, C, D\},$$

explain how the table below is computed, and give the values of d_{S} and d_{Tot}.

	Sum of squares	df
Treatments	13.02	3
Subjects	78.99	d_{S}
Residual	13.77	21
Total	105.78	d_{Tot}

Exercise 4

1. In the usual linear model notation, show that a test of the hypothesis $\beta = \beta'$ can be based on $(\hat\beta - \beta')^{\mathrm{T}} X^{\mathrm{T}} X (\hat\beta - \beta')/\sigma^2$, and give its null distribution in a normal model. Give a suitable test statistic and its null distribution when σ^2 is unknown.

2. The angles of a triangle ABC are each measured once independently and unbiasedly with common error variance σ^2. Find unbiased estimators of the angles at A and B and of σ^2. If the errors are normally distributed, say how you would test whether the triangle is equilateral.