Comprehensive Exam: Linear Models

Fall 2026

Department of Mathematical Sciences, MTU

Directions:

- This is an in-class, closed book, closed notes exam.
- Answer all problems.
- You are only required to answer what is asked so be brief but be precise.
- ***NO CALCULATORS*** or other electronic devices are allowed. Any questions that require numerical computation can be answered with exact expressions.
- Partial credit will be given to incomplete answers so show relevant work.
- Time limit is 3 hours.
- ***PUT*** your code on the top of each page.
- ***DO NOT*** write your names on any place of your answer paper.
- ***DO NOT*** write on the back of your answer paper.
- Start each question on a ***NEW PAGE***.
- Number the pages separately for each problem.
- Write neatly and organize your answers in readable form.

Please note: Any violations of the Honor Code will be punished to the full extent allowable under the MTU Honor Code.

1. Let X be an n-dimensional random vector with mean vector μ and positive definite covariance matrix Σ. Let A be an $n \times n$ be a symmetric matrix of constants of rank r.

   a) Provide the definition of what it means for Σ to be positive definite.

   b) Show that
   $$\mathrm{E}(X^T AX) = \mathrm{tr}(A\Sigma) + \mu^T A\mu$$

   c) Suppose that A is idempotent. Give a definition of what this means, and prove that all eigenvalues of A are either 0 or 1. Further prove that
   $$r = \mathrm{rank}(A) = \mathrm{tr}(A)$$

   d) If X has the multivariate normal distribution with identity covariance matrix (but potentially non-zero mean vector), determine the distribution of the quadratic form
   $$X^T AX$$
   still under the assumption that A is idempotent.

2. Suppose that $X = (X_1, X_2, X_3)^T$ has a multivariate normal distribution with mean vector and covariance matrix
   $$\mu = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 4 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

   a) Determine the correlation matrix of X.

   b) Determine the distribution of the random vector
   $$Y = \begin{bmatrix} X_1 - X_2 \\ 3X_1 + X_3 \end{bmatrix}$$

   c) Find a nontrivial linear combination of X_1, X_2, X_3 which is independent of Y.

   d) Note that X_1 and X_3 are independent. Does this imply that they conditionally independent, given X_2? Determine the conditional distribution of $(X_1, X_3)|X_2$ and use this to answer the question.

3. Consider the simple linear *no-intercept* regression model given by
   $$Y_i = \beta x_i + \varepsilon_i, \quad i = 1, \dots, n$$
   where the ε_i's are all independent, zero-mean Gaussian random variables with common variance σ^2. Here, β and $\sigma > 0$ are unknown parameters, and $[x_1, \dots, x_n]^T$ is assumed to a vector of known constants (not all the same!).

   a) Derive the least-squares estimator of β, call it $\hat{\beta}$, and show that it is unbiased. What is the distribution of $\hat{\beta}$?

   b) Derive an unbiased estimator of σ^2, call it $\hat{\sigma}^2$. What is the distribution of $\hat{\sigma}^2$?

c) Demonstrate that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent of one another.

d) Derive a hypothesis test of $H_0: \beta = 2$ against the alternative $H_1: \beta \neq 2$. To be specific, assume that $n = 10$ and use a significance level of $\alpha = 0.1$.

e) Describe a situation where it might be preferable to fit this no-intercept model instead of a simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

4. When gasoline is pumped into the tank of a car, vapors are vented into the atmosphere. An experiment was conducted to determine whether y, the amount of vapor, can be predicted using the following four variables based on initial conditions of the tank and the dispensed gasoline:

x_1 = tank temperature (°C)
x_2 = gasoline temperature (°C)
x_3 = vapor pressure in tank (psi)
x_4 = vapor pressure of gasoline (psi)

Consider a linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ with $\varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$ for a sample size of $n = 30$. Let X be the 30×5 design matrix of this model. Here are some summary data:

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)^T = (10, 2, 8, 1, 4)^T,$$

$$SSE = 50,$$

$$(X^T X)^{-1} = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 2 & 3 \end{bmatrix}.$$

For parts (a) and (b), use the provided t-table to carry out the necessary calculations and state your final statistical conclusion. For part (c), full numerical computation is not required; providing the requested formula and notation is sufficient.

a) Find a 95% confidence interval for β_1.

b) Test $H_0: \beta_1 = \beta_2$ against $H_a: \beta_1 \neq \beta_2$ at the $\alpha = 0.05$ level.

c) Test $H_0: \beta_1 = \beta_2 = 12\beta_3 = 12\beta_4$ at the $\alpha = 0.05$ level. Provide the algebraic formula for this test statistic, its degrees of freedom, and the notation for its exact critical value.

d) A junior analyst suggests dropping x_3 and x_4 from the model entirely because their parameters look small. Based on the structure of the $(X^T X)^{-1}$ matrix, explain whether the OLS estimates for β_1 and β_2 would change if you fit a reduced model containing only x_1 and x_2.

5. Consider the true linear regression model:

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

where Y is an $n \times 1$ vector, X_1 is an $n \times p_1$ design matrix with rank p_1, X_2 is an $n \times p_2$ design matrix with rank p_2, and $p_1 + p_2 < n$. The parameter vectors β_1 and β_2 are $p_1 \times 1$ and $p_2 \times 1$, respectively. The error term ε follows a multivariate normal distribution with mean 0 and covariance matrix $\sigma^2 I_n$.

Suppose an investigator incorrectly fits the under-fitted model $Y = X_1\beta_1 + \varepsilon^*$, obtaining the ordinary least squares (OLS) estimate $\tilde{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y$ and the residual sum of squares $SSE_{\text{under}} = \tilde{\varepsilon}^T \tilde{\varepsilon}$, where $\tilde{\varepsilon} = Y - X_1\tilde{\beta}_1$.

a) Find the expected value $E[\tilde{\beta}_1]$, and specify the condition(s) under which $\tilde{\beta}_1$ is an unbiased estimator of β_1.

b) Find the expected value of the residual sum of squares, $E[SSE_{\text{under}}]$. If an investigator uses SSE_{under} to estimate the error variance σ^2 via $\hat{\sigma}^2 = SSE_{\text{under}}/(n - p_1)$, will σ^2 be systematically overestimated or underestimated? Justify your answer briefly.

c) Derive and compare the covariance matrices $\text{Cov}(\hat{\beta}_1)$ from the *true full model* and $\text{Cov}(\tilde{\beta}_1)$ from the *under-fitted model* to show that $\text{Cov}(\hat{\beta}_1) - \text{Cov}(\tilde{\beta}_1)$ is positive semi-definite. Explain why this result does not contradict the Gauss-Markov theorem.

HINTS:

i. Multiply both sides of the full model by $(I_n - H_2)$, where $H_2 = X_2(X_2^T X_2)^{-1} X_2^T$, to express $\hat{\beta}_1$ in terms of X_1, H_2, and Y.

ii. You may use without proof that for any symmetric positive-definite matrices A and B of the same dimension,

$$A - B \text{ is positive semi-definite } \Leftrightarrow A^{-1} - B^{-1} \text{ is positive semi-definite.}$$

6. An agricultural researcher is comparing the yield effects of three distinct fertilizer formulations. Each fertilizer formulation is randomly assigned to two experimental plots, yielding a total of six observations. We model this scenario as follows:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2, \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

a) Write this model in a matrix form $Y = X\beta + \varepsilon$ by explicitly presenting the parameter vector $\beta = (\mu, \tau_1, \tau_2, \tau_3)^T$ and the 6×4 design matrix X. What is the rank of X?

b) Determine whether the following individual parameters or linear combinations are *estimable*:

i. μ

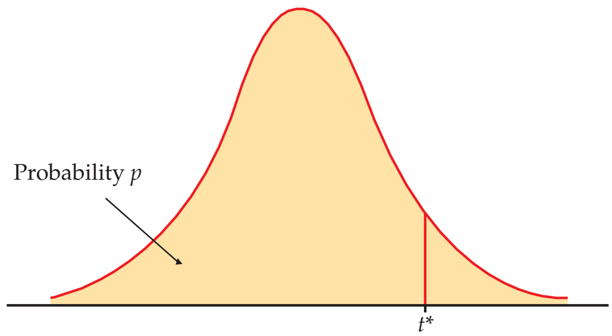
ii. $\tau_1 - \tau_2$

Answers without proper mathematical justification will not receive credit.

c) To overcome the identification problem, the researcher imposes the structural parameter constraint that the treatment effects sum to zero: $\tau_1 + \tau_2 + \tau_3 = 0$. Incorporate this constraint directly into the model equation to eliminate τ_3. Write out the newly reduced, full-rank parameter vector β^* and its corresponding 6×3 re-parameterized design matrix X^*. Show that the rank of X^* is exactly 3.

d) Using the re-parameterized, full-rank design matrix X^* from part (c), write down the formula for the OLS estimator $\widehat{\beta}^*$. What distribution does $\widehat{\beta}^*$ follow? Make sure to state all parameters of this distribution as well.

Table entry for p and C is the critical value t^* with probability p lying to its left and probability C lying between $-t^*$ and t^*.



TABLE

t distribution critical values

	Lower-tail probability p											
df	.75	.80	.85	.90	.95	.975	.98	.99	.995	.9975	.999	.9995
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											