

Problem 1

Let $\mathbf{Z} = (Z_1, Z_2, Z_3)^\top$ be a multivariate normal random vector with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$.

- (a) Find (i) $P(2Z_1 - 3Z_2 \leq 3)$, (ii) the joint distribution of $U = Z_1 + Z_2$ and $V = 3Z_1 - Z_3$ and (iii) the conditional distribution of $Z_1 + Z_3$ given $Z_1 = z_1$.

- (b) Let $\mathbf{A} = \begin{bmatrix} \frac{2}{3} & 0 & \frac{\sqrt{2}}{3} \\ 0 & 1 & 0 \\ \frac{\sqrt{2}}{3} & 0 & \frac{1}{3} \end{bmatrix}$ and consider $\mathbf{Y} = \mathbf{AZ}$. Derive the distribution of $\mathbf{Y}^\top \mathbf{Y}$.

Problem 2

Let $\mathbf{Y} = (Y_1, \dots, Y_n)^\top \sim \text{MVN}_n(\mu \mathbf{1}, \sigma^2 \mathbf{I})$, where $\mathbf{1}$ denotes an $n \times 1$ vector of ones.

- (a) Derive the expectation and variance of the sample mean and variance

$$\bar{Y} = \frac{1}{n} \mathbf{1}^\top \mathbf{Y}, \quad S^2 = \frac{1}{n-1} \mathbf{Y}^\top \left(\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \right) \mathbf{Y}.$$

- (b) Are $\bar{Y}$ and S^2 independent? Justify your answer.

Problem 3

A dataset contains information on 100 employees and their productivity scores. The goal is to study how productivity is related to

- whether the employee received training (0 = no, 1 = yes),
- the employee's number of years of experience, and
- the employee's department (Sales, HR, IT).

A multiple linear regression model is fit to the data with the productivity score as the response variable. The variable department is included in the regression model using dummy coding, with HR as the reference category. The model summary is provided below.

- (a) The regression model can in general be expressed as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. For this dataset and problem, what is the length of $\boldsymbol{\beta}$? Write down the dimension of $\mathbf{X}$ and describe briefly what numbers are stored in each column of $\mathbf{X}$.
- (b) Using the model summary, find the estimated productivity score for an employee with no training, 3 years of experience and who works in the HR department.
- (c) Give an interpretation of the estimated coefficient named `departmentIT` in the model summary.

Using the model summary, describe which test has been conducted to give the p-value 0.0759 for the coefficient `departmentSales` and discuss the result of this test.

- (d) We want to test whether department has an overall effect on productivity. Formulate this as a general linear hypothesis of the form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ vs. $\mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}$. Specify the matrix $\mathbf{C}$ and the distribution of the appropriate test statistic under the null. Sketch (roughly) the density of the test statistic under the null hypothesis, including a rejection region for significance level 0.01.

The general linear hypothesis test gives $F_{\text{obs}} = 14.299$. Include this observation in your sketch and state your conclusion.

```
Call:lm(formula = y ~ training + experience + department)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-15.7670	-3.5980	-0.1962	3.7613	13.6110

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	47.0892	1.6157	29.144	< 2e-16 ***
training	4.9441	1.1296	4.377	3.09e-05 ***
experience	1.9876	0.1116	17.804	< 2e-16 ***
departmentIT	8.0761	1.5293	5.281	8.13e-07 ***
departmentSales	2.2937	1.2780	1.795	0.0759 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.541 on 95 degrees of freedom

Multiple R-squared: 0.7962, Adjusted R-squared: 0.7876

F-statistic: 92.78 on 4 and 95 DF, p-value: < 2.2e-16

Problem 4

A 2^2 factorial experiment is conducted with two factors, A and B , each at two levels. The levels are coded as -1 (low level) and $+1$ (high level). Each combination of factor levels is replicated four times. A response Y is measured and the resulting estimated regression model (R output) is given below.

- (a) Fill in the 4 questions marks (?) found in the R output.
- (b) What is the estimated effect of changing the level of factor A (i) when B is at low level and (ii) when B is at high level?

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	9.7074	0.3474	27.943	2.73e-12	***
A	2.3457	0.3474	6.752	2.04e-05	***
B	-1.2155	?	?	0.00439	**
A:B	2.9075	0.3474	8.369	2.36e-06	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.39 on 12 degrees of freedom

Multiple R-squared: 0.9142, Adjusted R-squared: 0.8928

F-statistic: 42.63 on ? and ? DF, p-value: 1.125e-06