3. (a) Consider the following small data, where X_1 is a numerical explanatory variable and X_2 is categorical explanatory variable having class values $\{a, b, c\}$.

	X1	X2	Y
1	3	a	46.0
2	3	b	55.4
3	3	c	57.9
4	6	a	55.5
5	6	b	66.7
6	6	c	68.6
7	9	a	65.3
8	9	b	76.5
9	9	c	78.3

Consider modeling the response variable Y by the following linear model:

$$\mathcal{M}_{12}: \quad Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

where index j is related to the categories of X_2. The model $\mathcal{M}_{1|2}$ can be written in matrix form as

$$\mathcal{M}_{1|2}: \ \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \mathrm{Cov}(\mathbf{y}) = \sigma^2 \mathbf{I}.$$

Write in details what kind forms the model matrix $\mathbf{X}$ and parameter vector $\boldsymbol{\beta}$ have in case of given data is modeled by the model $\mathcal{M}_{12}$.

(2 points)

Solution:

For the variable X_2, we select the class a as baseline category. In a case of interaction effect model, we need dummy variables for the categories b and c for the variable X_2. Let us denote those dummy variables as x_{i2b}, x_{i2c}. Then the interaction effect can be written as

$$\begin{aligned}\mathcal{M}_{12}: Y_i &= \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1} + \varepsilon_i, \\ Y_i &= \beta_0 + \beta_1 x_{i1} + \alpha_2 x_{i2b} + \alpha_3 x_{i2c} + \gamma_2 x_{i1} x_{i2b} + \gamma_3 x_{i1} x_{i2c} + \varepsilon_i.\end{aligned}$$

This means that the model matrix $\mathbf{X}$ and parameter vector $\boldsymbol{\beta}$ are

$$\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \mathbf{x}_{2b} : \mathbf{x}_{2c} : \mathbf{x}_{1,2b} : \mathbf{x}_{1,2c}) = \begin{pmatrix} 1 & 3 & 0 & 0 & 0 & 0 \\ 1 & 3 & 1 & 0 & 3 & 0 \\ 1 & 3 & 0 & 1 & 0 & 3 \\ 1 & 6 & 0 & 0 & 0 & 0 \\ 1 & 6 & 1 & 0 & 6 & 0 \\ 1 & 6 & 0 & 1 & 0 & 6 \\ 1 & 9 & 0 & 0 & 0 & 0 \\ 1 & 9 & 1 & 0 & 9 & 0 \\ 1 & 9 & 0 & 1 & 0 & 9 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \alpha_2 \\ \alpha_3 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}.$$

(b) Below is the R estimation output of the lm-function related to the particular linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.8753 -1.8275 -0.0943  2.1809  7.2335

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.0026     1.5837  11.367 6.02e-14 ***
factor(x1)2  -1.7752     1.5259  -1.163   0.2517
factor(x1)3  -2.9361     1.4336  -2.048   0.0473 *
factor(x2)2   1.2917     1.3836   0.934   0.3563
factor(x2)3   0.3726     1.4915   0.250   0.8040
factor(x2)4  -3.8796     1.5543  -2.496   0.0169 *
---

Residual standard error: 3.388 on 39 degrees of freedom
Multiple R-squared:  0.3396,    Adjusted R-squared:  0.2549
F-statistic: 4.011 on 5 and 39 DF,  p-value: 0.004953
```

i. What kind of linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ the output is related to?

A. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$,

B. $Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_j + \varepsilon_i$,

C. $Y_i = \beta_0 + \beta_j + \alpha_h + \varepsilon_i$,

D. $Y_i = \beta_0 + \beta_j + \alpha_h + \gamma_{jh} + \varepsilon_i$.

ii. Calculate the maximum likelihood estimate of the expected value μ when the explanatory variables X_1, X_2 are set on the values

$$x_1 = 3,$$
$$x_2 = 3.$$

(2 points)

Solution:

The output is related to the two-way analysis of variance main effect model $Y_i = \beta_0 + \beta_j + \alpha_h + \varepsilon_i$. The maximum likelihood estimate of the expected value μ when $x_1 = 3, x_2 = 3$ is

$$\hat{\mu}_{33} = \hat{\beta}_0 + \hat{\beta}_3 + \hat{\alpha}_3 = 18.0026 + (-2.9361) + 0.3726 = 15.4391.$$

(c) Consider the linear model

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}),$$
$$\boldsymbol{\mu} = \mathbf{1}\beta_0,$$

where $\mathbf{1}$ is a vector of ones $\mathbf{1} = (1, 1, \dots, 1)'$. The sample mean

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{1}{n}\mathbf{1}'\mathbf{y}$$

is the maximum likelihood estimator for the parameter β_0, i.e., $\hat{\beta}_0 = \bar{y}$. Make yourself familiar with Theorem 1.1 in section 1.3.2 Multivariate Normal Distribution and then calculate the expected value $\mathrm{E}(\hat{\beta}_0)$ and the variance $\mathrm{Var}(\hat{\beta}_0)$.

(2 points)

Solution:

The expected value $\mathrm{E}(\hat{\beta}_0)$ is

$$\mathrm{E}(\hat{\beta}_0) = \mathrm{E}(\bar{y}) = \mathrm{E}\left(\frac{1}{n}\mathbf{1}'\mathbf{y}\right) = \frac{1}{n}\mathbf{1}' \cdot \mathrm{E}(\mathbf{y}) = \frac{1}{n}\mathbf{1}' \cdot \mathbf{1}\beta_0 = \frac{n}{n}\beta_0 = \beta_0.$$

The variance $\mathrm{Var}(\hat{\beta}_0)$ is

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{y}) &= \mathrm{Var}\left(\frac{1}{n}\mathbf{1}'\mathbf{y}\right) = \frac{1}{n}\mathbf{1}' \cdot \mathrm{Cov}(\mathbf{y}) \cdot \mathbf{1}\frac{1}{n} = \frac{1}{n}\mathbf{1}' \cdot (\sigma^2\mathbf{I}) \cdot \mathbf{1}\frac{1}{n} \\ &= \sigma^2\frac{1}{n} \cdot n \cdot \frac{1}{n} = \frac{\sigma^2}{n}.\end{aligned}$$