# Chapter 1

# Introduction to Modeling by Linear Models

## 1.1 Statistical Modeling of the Expected Value

### 1.1.1 General framework for Modeling the Expected Value

- In statistical modeling, the aim is often to model the relationship between several explanatory variables $X_1, X_2, \ldots, X_p$ and one response variable $Y$.

- The response variable $Y$ is assumed to be a random variable with the probability distribution defined by the density function $f_Y(y|\mathbf{x}, \boldsymbol{\beta})$, where $\mathbf{x} = (X_1, X_2, \ldots, X_p)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is unknown parameter vector, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$.

- In many modeling situations, it is assumed that the density function $f_Y(y|\mathbf{x}, \boldsymbol{\beta})$ depends on $\mathbf{x}$ and $\boldsymbol{\beta}$ thorough the expected value $\mathrm{E}(Y) = \mu = \mu(\mathbf{x}, \boldsymbol{\beta})$ and the variance $\mathrm{Var}(Y) = \sigma^2 = \sigma^2(\mathbf{x}, \boldsymbol{\beta})$:

$$f_Y(y|\mathbf{x}, \boldsymbol{\beta}) = f_Y\left(y|\mu(\mathbf{x}, \boldsymbol{\beta}), \sigma^2(\mathbf{x}, \boldsymbol{\beta})\right). \tag{1.1}$$

- Often the explanatory variables $\mathbf{x} = (X_1, X_2, \ldots, X_p)'$ are assumed to affect only the level of the expected value $\mu(\mathbf{x}, \boldsymbol{\beta})$ and the variance $\mathrm{Var}(Y) = \sigma^2$ is independent on $\mathbf{x}$ and $\boldsymbol{\beta}$:

$$f_Y(y|\mathbf{x}, \boldsymbol{\beta}) = f_Y\left(y|\mu(\mathbf{x}, \boldsymbol{\beta}), \sigma^2\right). \tag{1.2}$$

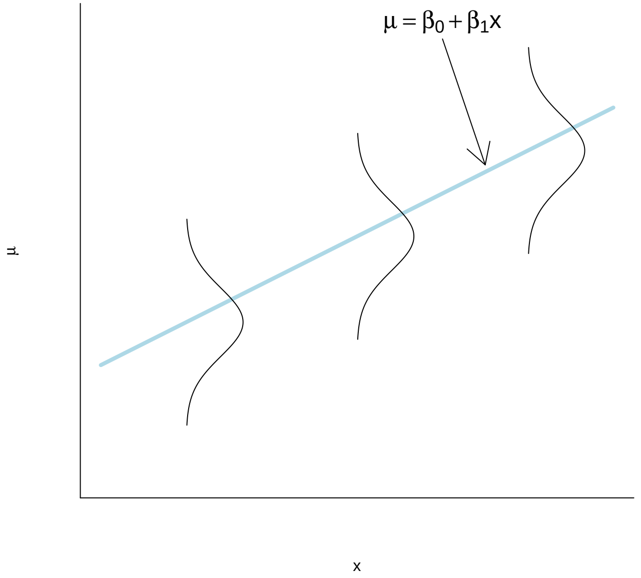
- Statistical modeling is largely about finding out a suitable function $h$, i.e., $\mu = h(\mathbf{x}, \boldsymbol{\beta})$, which would map the relationship between the explanatory variables $\mathbf{x} = (X_1, X_2, \ldots, X_p)'$ and the expected value $\mathrm{E}(Y) = \mu$.

### 1.1.2 Simple Linear Regression Model

- In **simple linear regression** model, there is only one numerical explanatory variable $X$ and the expected value (population mean) of the continuous random variable $Y$, denoted by $\mu$, is assumed to have a linear form

$$\mu = \beta_0 + \beta_1 X, \tag{1.3}$$

where $\beta_0, \beta_1$ are unknown parameters.



- The response variable $Y$ is assumed to follow the normal distribution $Y \sim N(\mu, \sigma^2)$, that is,

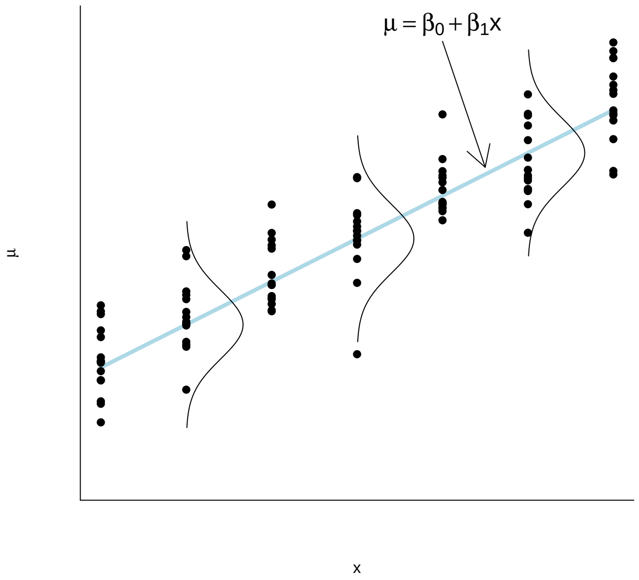
$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2), \tag{1.4}$$

where $\sigma^2$ is unknown variance parameter.

- The simple linear regression model is often presented in a form

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1.5}$$

where $\varepsilon$ denotes the random error term following the normal distribution $\varepsilon \sim N(0, \sigma^2)$.



- For each sampling unit $i$, the measured or observed value $y_i$ is assumed to be a realization from the model

$$\mathcal{M}: \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1.6}$$

  where each $\varepsilon_i$ is assumed to follow the normal distribution $\varepsilon_i \sim N(0, \sigma^2)$.

- In R, the numerical estimates $\hat{\beta}_0, \hat{\beta}_1, \tilde{\sigma}^2$ in the simple linear regression model for unknown parameters $\beta_0, \beta_1, \sigma^2$ are obtained by using lm()-function.

## Example 1.1.

Let us consider the following satsuma.txt dataset :

```
Time      3.0  3.0  3.0  3.0  3.0  3.0  3.0  3.0  3.0   6.0  ... 12.0  12.0  12.0
VitaminC 59.9 60.3 62.4 59.2 60.1 62.1 60.5 61.4 63.5  54.4  ... 38.5  39.6  41.8
```

S. Qiu and J. Wang (2015). "Effects of Storage Temperature and Time on Internal Quality of Satsuma Mandarin
by Means of E-Nose and E-Tongue Based on Two-Way MANOVA Analysis and Random Forest,"
Innovative Food Science and Emerging Technologies, Vol. 31, pp. 139-150.

The dataset is related to a study, where it was investigated how the storage time of Satsuma mandarins (3,6,9 or 12 days) affects the amount of vitamin C the mandarins are containing. In data, the sampling unit is a single mandarin, and the variable $X = \mathsf{Time}$ measures how long the mandarin was stored. The variable $Y = \mathsf{VitaminC}$ then measures the amount of vitamin C is containing in mandarin after the storage time.

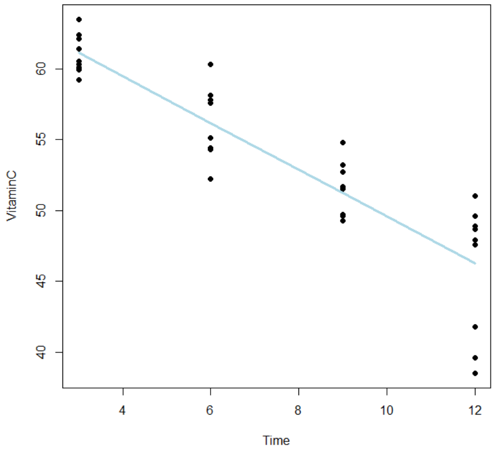
Let us model the dataset by the simple linear regression model

$$\mathcal{M}: \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2).$$

With use of $\mathsf{lm}$-function, the estimates of $\beta_0, \beta_1, \sigma^2$ are

$$\hat{\beta}_0 = 66.077778, \qquad \hat{\beta}_1 = -1.651481, \qquad \tilde{\sigma}^2 = 8.091003.$$

```
> model<-lm(VitaminC~Time)
> coef(model)
(Intercept)         Time
  66.077778    -1.651481
> sigma(model)^2
[1] 8.091003
```



### 1.1.3 One-Way Analysis of Variance

- The one-way analysis of variance examines the effect of a single categorical explanatory variable $X$ on the average values of the numerical response variable $Y$.
- Based on the value of the explanatory variable $X$, the population of observation units are divided into $k$ different subpopulations, i.e., $x_i \in \{1, 2, 3, \ldots, k\}$. The index $j$ is used to denote the subpopulation $j = 1, 2, 3, \ldots, k$.
- In the one-way analysis of variance, the following assumptions are classically made for the random variables $Y_i$:
  - The random variables $Y_i$ follow the normal distributions $Y_{ij} \sim N(\mu_j, \sigma^2)$.
  - The random variables $Y_i$ are independent of each other for $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, k$.
- The main research problem in one-way analysis of variance is to test whether the expected values $\mu_j$ of the subpopulations are the same for each subpopulation $j$. In one-way analysis of variance, considered hypotheses are

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_k, \\ H_1 &: \mu_1 \neq \mu_2 \neq \ldots \neq \mu_k. \end{aligned}$$

- The assumption $Y_i \sim N(\mu_j, \sigma^2)$ can be written as a model

$$\mathcal{M}: \quad Y_i = \beta_0 + \beta_j + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2). \tag{1.7}$$

- In this course, the first category $j = 1$ is selected to be a control or a baseline category, and hence in parametrization $\mu_j = \beta_0 + \beta_j$, the restriction $\beta_1 = 0$ holds.
- If the null hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ holds, then the random variables $Y_i$ can be considered to follow a model

$$\mathcal{M}_0: \quad Y_i = \beta_0 + \varepsilon_i.$$

- The original hypotheses $H_0$ and $H_1$ can be equivalently stated as

$$\begin{aligned} H_0 &: \text{Model } \mathcal{M}_0 \text{ is the true model,} \\ H_1 &: \text{Model } \mathcal{M} \text{ is the true model.} \end{aligned}$$

## Example 1.2.

The aim of the study was to compare the average lengths of heliconia plants in the situations of variety being Bihai, red Caribaea or yellow Caribaea. The dataset can be found in the file heliconia.txt.

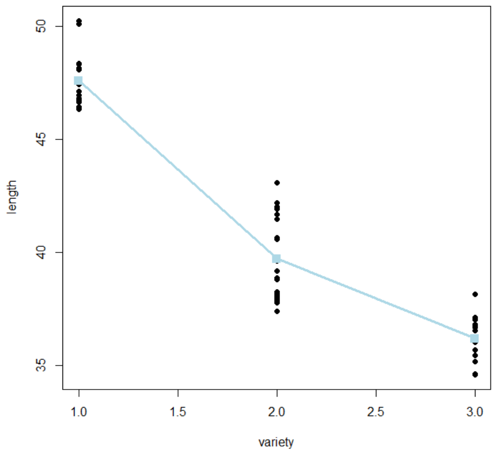
```
  variety length
1   bihai  47.12
2   bihai  46.75
.
17    red  41.90
18    red  42.01
.
40 yellow  36.78
41 yellow  37.02
.
53 yellow  34.57
54 yellow  34.63
```

The variable $X = \mathsf{variety}$ defines the subpopulations $j = \mathsf{bihai} = 1, j = \mathsf{red} = 2$ and $j = \mathsf{yellow} = 3$. The response variable is $Y = \mathsf{length}$. Let the observed values $y_i$ be realizations of the random variables $Y_i$ which are assumed to follow the normal distribution $Y_i \sim N(\mu_j, \sigma^2)$.

That is, let us test the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ by the model

$$\mathcal{M} : \quad Y_i = \beta_0 + \beta_j + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2).$$

```
> model<-lm(length~factor(variety))
> model0<-lm(length~1)
> anova(model0, model, test="F")
Analysis of Variance Table

Model 1: length ~ 1
Model 2: length ~ factor(variety)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     53 1189.44
2     51  106.57  2    1082.9 259.12 < 2.2e-16 ***
```



## 1.2 Different Types of Linear Models

### 1.2.1 Linear Regression Analysis

– In statistical modeling, the aim is often to model the relationship between several numerical explanatory variables $X_1, X_2, \ldots, X_p$ and one numerical response variable $Y$.

– In linear regression model, the relationship between explanatory variables $X_1, X_2, \ldots, X_p$ and the expected value of random variable $Y$ is assumed to have a linear form

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p, \tag{1.8}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are unknown parameters.

– The response variable $Y$ is assumed to follow the normal distribution $Y \sim N(\mu, \sigma^2)$, that is,

$$Y \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p, \sigma^2), \tag{1.9}$$

where $\sigma^2$ is unknown variance parameter.

– Linear regression model is often presented in a form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \tag{1.10}$$

where $\varepsilon$ denotes the random error term following the normal distribution $\varepsilon \sim N(0, \sigma^2)$.

– For each sampling unit $i$, the measured value $y_i$ is assumed to be realization from the model

$$\mathcal{M}: \quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \tag{1.11}$$

where each $\varepsilon_i$ is assumed to follow the normal distribution $\varepsilon_i \sim N(0, \sigma^2)$.

## Example 1.3.

Consider the dataset blackcherry.txt. This dataset provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. Note that girth is the diameter of the tree (in inches) measured at 4 ft 6 in above the ground.

```
          Girth  Height   Volume
1         21.08   21.34     0.29
2         21.84   19.81     0.29
3         22.35   19.20     0.29
4         26.67   21.95     0.46
.
30        45.72   24.38     1.44
31        52.32   26.52     2.18
```

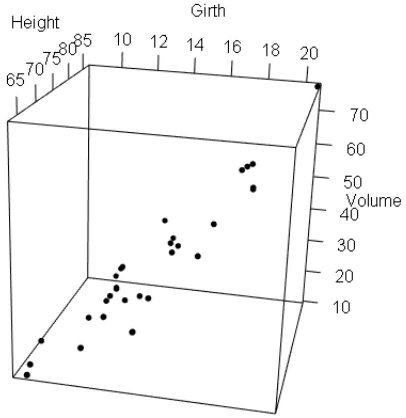
```
X1=Girth - Tree diameter in inches
X2=Height - Height in ft
Y=Volume - Volume of timber in cubic ft
```

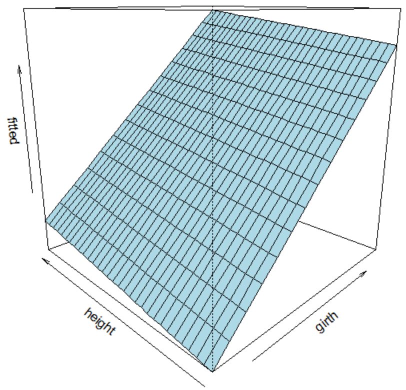
Let us model the Volume variable $Y$ with the linear regression model

$$\mathcal{M}_{1|2}: \quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

```
> model<-lm(Volume~Girth+Height)
> summary(model)

Call:
lm(formula = Volume ~ Girth + Height)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.4065 -2.6493 -0.2876  2.2003  8.4847 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *  
---

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,     Adjusted R-squared:  0.9442 
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Hence $\hat{\beta}_0 = -57.9877, \hat{\beta}_1 = 4.7082, \hat{\beta}_2 = 0.3393$ are the maximum likelihood estimates of the parameters $\beta_0, \beta_1, \beta_2$, and the unbiased estimate $\tilde{\sigma}^2$ for the the variance parameter $\sigma^2$ is $\tilde{\sigma}^2 = 3.882^2 = 15.06862$.





### 1.2.2 Linear Analysis of Covariance

- Analysis in linear statistical models is called linear analysis of covariance, if some of explanatory variables are defined on numerical scale and some of them are defined on categorical scale. In linear analysis of covariance, the response variable $Y$ is numerical variable.

- We consider here linear analysis of covariance with two explanatory variables, in such way that $X_1$ variable is considered to be a numerical variable and $X_2$ categorical variable.

- Categorical variable $X_2$ can have $k$ different values. Each value defines a class or subpopulation, and different values are usually coded in numbers $j = 1, 2, \ldots, k$.

- In this course, the first category $j = 1$ is selected to be a control or a baseline category.

- In linear analysis of covariance, linear model is called the **main effect model** if in case of sampling unit $i$ belonging to the category $x_{i2} = 1$, linear model has the form

$$x_{i2} = 1 : \quad Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i. \tag{1.12}$$

- In case of other categories, the main effect model has the forms

$$\begin{aligned} x_{i2} = 2: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_2 + \varepsilon_i, \\ x_{i2} = 3: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_3 + \varepsilon_i, \\ &\vdots \\ x_{i2} = k: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_k + \varepsilon_i. \end{aligned} \tag{1.13}$$

- The main effect model can be written as

$$\mathcal{M}_{1|2}: \quad Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_j + \varepsilon_i, \quad \alpha_1 = 0, \tag{1.14}$$

  where index $j$ denotes to the categories defined by the variable $X_2$.

- In linear analysis of covariance, linear model is said to include **interaction effect** in model, if in case of category $x_{i2} = 1$, linear model still has the form

$$x_{i2} = 1: \quad Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i \tag{1.15}$$

  but in case of other categories, the model has the forms

$$\begin{aligned} x_{i2} = 2: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_2 + \gamma_2 x_{i1} + \varepsilon_i, \\ x_{i2} = 3: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_3 + \gamma_3 x_{i1} + \varepsilon_i, \\ &\vdots \\ x_{i2} = k: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_k + \gamma_k x_{i1} + \varepsilon_i. \end{aligned} \tag{1.16}$$

– The model with interaction effect can be written as

$$\mathcal{M}_{12}: \quad Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1} + \varepsilon_i, \quad \alpha_1 = 0,\ \gamma_1 = 0, \tag{1.17}$$

where $\beta_0, \beta_1, \alpha_j, \gamma_j$ are unknown parameters, the random error term $\varepsilon_i$ is assumed to follow the normal distribution $\varepsilon_i \sim N(0, \sigma^2)$ as $\sigma^2$ being unknown variance parameter.

## Example 1.4.

The research group was interested in finding out how vitamin C affects the length of the teeth of the guinea pigs. In the study, a total of 60 guinea pigs were divided into three equal groups, with 0.5 milligrams of vitamin C per day administered to guinea pigs in the first group, 1 mg of vitamin C per day, and 2 mg of vitamin C per day. In addition, each of these three groups was divided into two so that half of the group's guinea pigs received vitamin C in the juice (OJ) and half as crystals (VC). The teeth of the guinea pig were measured prior to the beginning of the experimentation and subsequent measurements were made, see file teethpig.txt.

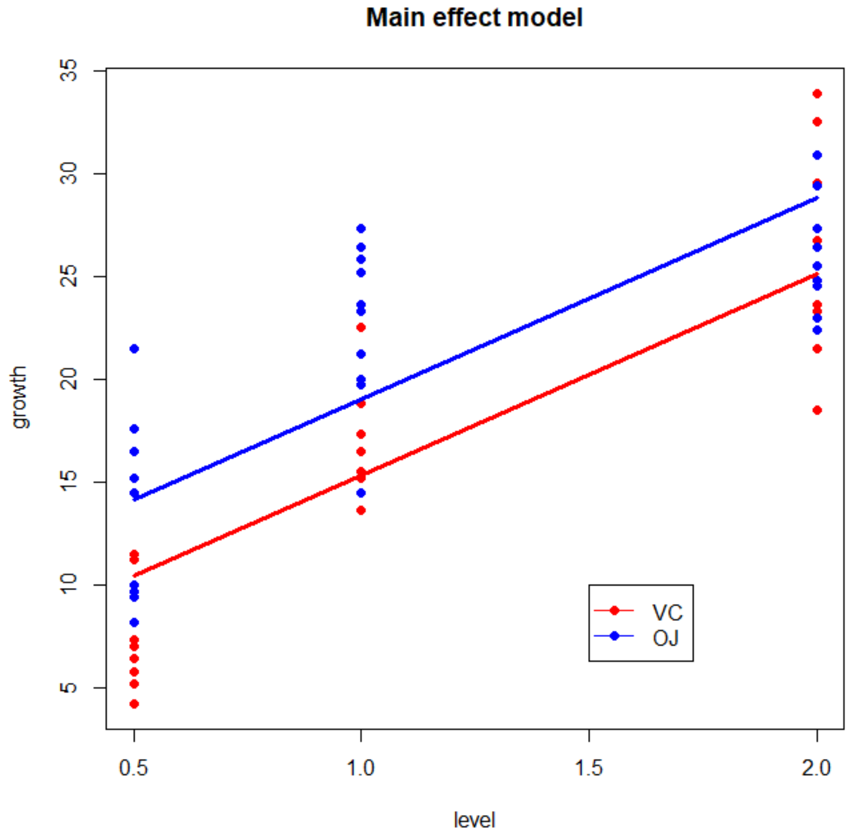
```
        growth          dose        level
1          4.2            VC          0.5
2         11.5            VC          0.5
3          7.3            VC          0.5
.
59        29.4            OJ          2.0
60        23.0            OJ          2.0
```

Let us denote the explanatory variables as $X_1 =$ level and $X_2 =$ dose. Let us model the response variable $Y =$growth by the following linear covariance models

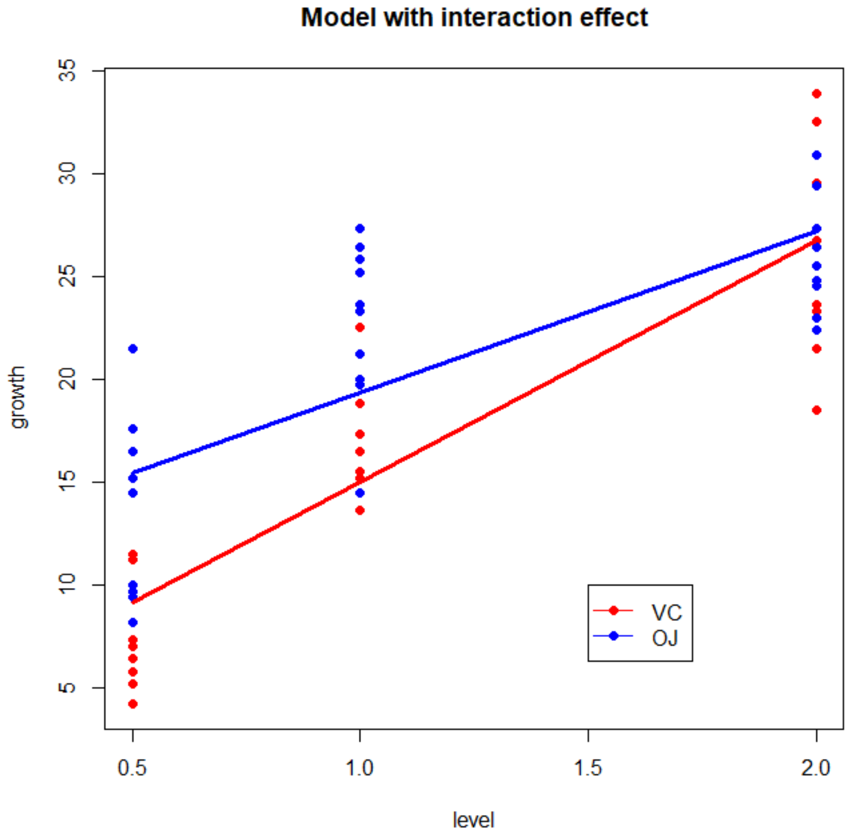
$$\begin{aligned} \mathcal{M}_{1|2} : \quad & Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_j + \varepsilon_i, \\ \mathcal{M}_{12} : \quad & Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1} + \varepsilon_i. \end{aligned}$$

```
> model.main<-lm(growth~level+factor(dose))
> summary(model.main)

Call:
lm(formula = growth ~ level + factor(dose))

Residuals:
   Min     1Q Median     3Q    Max
-6.600 -3.700  0.373  2.116  8.800

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.2725     1.2824   7.231 1.31e-09 ***
level            9.7636     0.8768  11.135 6.31e-16 ***
factor(dose)VC  -3.7000     1.0936  -3.383   0.0013 **
---

Residual standard error: 4.236 on 57 degrees of freedom
Multiple R-squared:  0.7038,    Adjusted R-squared:  0.6934
F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

```
> model.full<-lm(growth~level+factor(dose)+level:factor(dose))
> summary(model.full)

Call:
lm(formula = growth ~ level + factor(dose) + level:factor(dose))

Residuals:
    Min      1Q  Median      3Q     Max
-8.2264 -2.8462  0.0504  2.2893  7.9386

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            11.550      1.581   7.304 1.09e-09 ***
level                   7.811      1.195   6.534 2.03e-08 ***
factor(dose)VC         -8.255      2.236  -3.691 0.000507 ***
level:factor(dose)VC    3.904      1.691   2.309 0.024631 *
---

Residual standard error: 4.083 on 56 degrees of freedom
Multiple R-squared:  0.7296,    Adjusted R-squared:  0.7151
F-statistic: 50.36 on 3 and 56 DF,  p-value: 6.521e-16
```



Main effect model



Model with interaction effect

### 1.2.3 Linear Analysis of Variance

- Analysis in linear statistical models is called linear analysis of variance, if all explanatory variables are defined on categorical scale. In linear analysis of variance, the response variable $Y$ is numerical variable.
- We consider here linear analysis of variance with two explanatory variables, called two-way ANOVA, in such way that index $j = 1, 2, \ldots, k$ is related to categories defined by $X_1$ variable, and index $h = 1, 2, \ldots, l$ is is related to categories defined by $X_2$ variable.
- In this course, the first categories $j = 1$ and $h = 1$ are selected to be a joint control or a joint baseline category.
- Often in analysis of variance, the response variable $Y_i$ is assumed to follow normal distribution $Y_i \sim N(\mu_{jh}, \sigma^2)$ with expected value $\mu_{jh}$ depending on explanatory variables by linear structure.

- In two-way ANOVA situation, possible competing linear models for the response variable $Y_i$ are

$$\begin{aligned} \mathcal{M}_0 : &\quad Y_i = \beta_0 + \varepsilon_i, && (1.18a)\\ \mathcal{M}_1 : &\quad Y_i = \beta_0 + \beta_j + \varepsilon_i, && (1.18b)\\ \mathcal{M}_2 : &\quad Y_i = \beta_0 + \alpha_h + \varepsilon_i, && (1.18c)\\ \mathcal{M}_{1|2} : &\quad Y_i = \beta_0 + \beta_j + \alpha_h + \varepsilon_i, && (1.18d)\\ \mathcal{M}_{12} : &\quad Y_i = \beta_0 + \beta_j + \alpha_h + \gamma_{jh} + \varepsilon_i, && (1.18e) \end{aligned}$$

  with in each model $\varepsilon_i \sim N(0, \sigma^2)$.

- The model $\mathcal{M}_{1|2}$ is called as two-way main effect model, and the model $\mathcal{M}_{12}$ is called the two-way model with interaction.

- Two-way main effect model

$$\mathcal{M}_{1|2} : \quad Y_i = \beta_0 + \beta_j + \alpha_h + \varepsilon_i \qquad (1.19)$$

  means that for control categories $x_{i1} = 1$ and $x_{i2} = 1$ the model has a form

$$X_1 = 1, X_2 = 1 : \quad Y_i = \beta_0 + \varepsilon_i.$$

For different categories of $X_1$ when $X_2 = 1$ the model have the forms

$$\begin{aligned} X_1 = 2, X_2 = 1: &\quad Y_i = \beta_0 + \beta_2 + \varepsilon_i, \\ X_1 = 3, X_2 = 1: &\quad Y_i = \beta_0 + \beta_3 + \varepsilon_i, \\ &\vdots \\ X_1 = k, X_2 = 1: &\quad Y_i = \beta_0 + \beta_k + \varepsilon_i. \end{aligned}$$

Similarly, for different categories of $X_2$ when $X_1 = 1$ the model have the forms

$$\begin{aligned} X_1 = 1, X_2 = 2: &\quad Y_i = \beta_0 + \alpha_2 + \varepsilon_i, \\ X_1 = 1, X_2 = 3: &\quad Y_i = \beta_0 + \alpha_3 + \varepsilon_i, \\ &\vdots \\ X_1 = 1, X_2 = l: &\quad Y_i = \beta_0 + \alpha_l + \varepsilon_i. \end{aligned}$$

– Thus in two-way main effect model, the expected value $\mu_{jh}$ is assumed to follow the model

$$\mu_{jh} = \beta_0 + \beta_j + \alpha_h, \qquad \beta_1 = 0, \alpha_1 = 0. \tag{1.20}$$

- When $k = 3$ and $l = 3$, then two-way main effect model has a structure

$$
\begin{aligned}
\mu_{11} &= \beta_0, \\
\mu_{12} &= \beta_0 + \alpha_2, \\
\mu_{13} &= \beta_0 + \alpha_3, \\
\mu_{21} &= \beta_0 + \beta_2, \\
\mu_{22} &= \beta_0 + \beta_2 + \alpha_2, \\
\mu_{23} &= \beta_0 + \beta_2 + \alpha_3, \\
\mu_{31} &= \beta_0 + \beta_3, \\
\mu_{32} &= \beta_0 + \beta_3 + \alpha_2, \\
\mu_{33} &= \beta_0 + \beta_3 + \alpha_3
\end{aligned}
$$

- Two-way model with interaction

$$
\mathcal{M}_{12}: \quad Y_i = \beta_0 + \beta_j + \alpha_h + \gamma_{jh} + \varepsilon_i \tag{1.21}
$$

is actually model with following constrains on the parameters

$$
\begin{aligned}
\mathcal{M}_{12}: \quad Y_i = &\beta_0 + \beta_j + \alpha_h + \gamma_{jh} + \varepsilon_i, \\
&\beta_1 = 0, \alpha_1 = 0, \\
&\gamma_{1h} = 0 \text{ for every } h = 1, \ldots, l, \\
&\gamma_{j1} = 0 \text{ for every } j = 1, \ldots, k.
\end{aligned} \tag{1.22}
$$

- When $k = 3$ and $l = 3$, then two-way model with interaction has a structure on the expected value $\mu_{jh}$ as following

$$\begin{aligned}
\mu_{11} &= \beta_0, \\
\mu_{12} &= \beta_0 + \alpha_2, \\
\mu_{13} &= \beta_0 + \alpha_3, \\
\mu_{21} &= \beta_0 + \beta_2, \\
\mu_{22} &= \beta_0 + \beta_2 + \alpha_2 + \gamma_{22}, \\
\mu_{23} &= \beta_0 + \beta_2 + \alpha_3 + \gamma_{23}, \\
\mu_{31} &= \beta_0 + \beta_3, \\
\mu_{32} &= \beta_0 + \beta_3 + \alpha_2 + \gamma_{32}, \\
\mu_{33} &= \beta_0 + \beta_3 + \alpha_3 + \gamma_{33}
\end{aligned}$$

- The main research problem is to find out which of the following models $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_{1|2}, \mathcal{M}_{12}$ the best describes the behaviour of the random variables $Y_i$.
- The chosen model then describes what kind effect the explanatory variables $X_1$ and $X_2$ have on the expected value of $\mu_{jh}$.

- Part of analysis of variance is to consider different kind of tests between competing models. For example, in two-way ANOVA, often following hypotheses are considered

$$H_0 : \text{The model } \mathcal{M}_{1|2} \text{ is the true model,} \tag{1.23a}$$

$$H_1 : \text{The model } \mathcal{M}_{12} \text{ is the true model.} \tag{1.23b}$$

- After the appropriate model is chosen, pairwise comparison, i.e., post hoc testing, of differences $\mu_{jh} - \mu_{j_* h_*}$ are usually considered. Also testing of pairwise average differences are often considered:

$$H_0 : \left(\frac{\mu_{j1} + \mu_{j2} + \cdots + \mu_{jl}}{l}\right) - \left(\frac{\mu_{j_*1} + \mu_{j_*2} + \cdots + \mu_{j_*l}}{l}\right) = 0, \qquad j \neq j_*, \tag{1.24a}$$

$$H_0 : \left(\frac{\mu_{1h} + \mu_{2h} + \cdots + \mu_{kh}}{k}\right) - \left(\frac{\mu_{1h_*} + \mu_{2h_*} + \cdots + \mu_{kh_*}}{k}\right) = 0, \qquad h \neq h_*. \tag{1.24b}$$

## Example 1.5.

In a rat study, it was investigated how much weight gains in rats occur within 85 days when they are exposed to different amounts of recombinant bovine growth hormone (rBGH). The full dataset can be found in file ratsRBGH.txt.

```
      gender rbGH weightgain
1          1    1     274.99
2          1    1     289.67
3          1    1     346.40
4          1    1     344.32
5          1    1     364.63
6          1    2     478.62

.
.
59         2    6     164.93
60         2    6     177.85

Description: Weight gains in rats over 85 day period in 6 treatment
conditions of recombinant bovine growth hormone (rbGH):
1=Control (0 mg/kg per day)
2=Subcutaneous Injection (1.0 mg/kg per day)
3=Oral Glavage (0.1 mg/kg per day)
4=Oral Glavage (0.5 mg/kg per day)
5=Oral Glavage (5 mg/kg per day)
6=Oral Glavage (50 mg/kg per day)

Gender: 1=Male, 2=Female

Source: J.C. Juskevich and C.G. Guyer (1990). "Bovine Growth Hormone: Human
Food Safety Evaluation," Science, Vol.249,#4971,pp875-884.
```

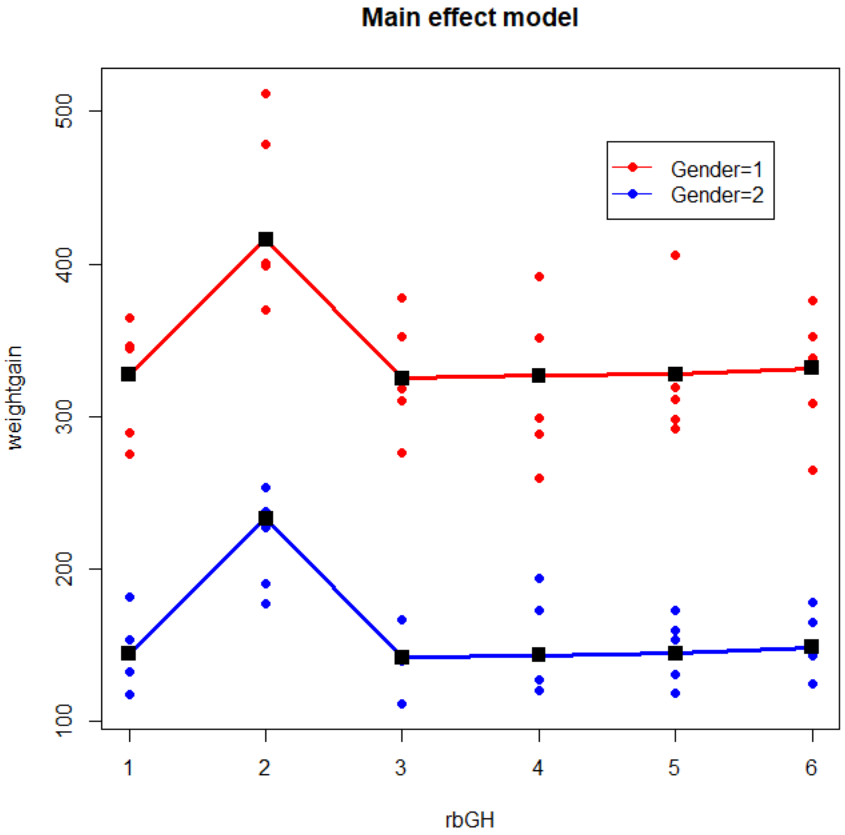
Denote explanatory variables as $X_1$=rbGH and $X_2$=gender. Let us consider modeling the response variable $Y$=weightgain by following two different models

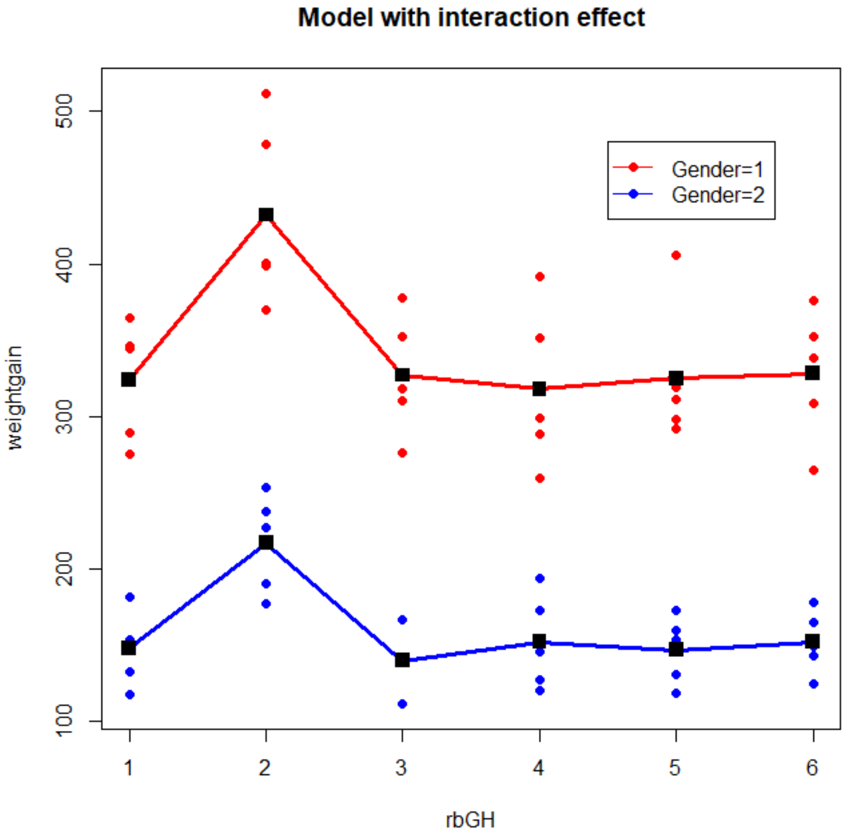
$$\begin{aligned} \mathcal{M}_{1|2}: \quad & Y_i = \beta_0 + \beta_j + \alpha_h + \varepsilon_i, \\ \mathcal{M}_{12}: \quad & Y_i = \beta_0 + \beta_j + \alpha_h + \gamma_{jh} + \varepsilon_i, \end{aligned}$$

where in each model the random error term $\varepsilon_i$ is assumed to follow normal distribution $\varepsilon_i \sim N(0, \sigma^2)$. Particularly let us test the hypotheses

$$\begin{aligned} H_0 :& \text{Model } \mathcal{M}_{1|2} \text{ is the true model,} \\ H_1 :& \text{Model } \mathcal{M}_{12} \text{ is the true model.} \end{aligned}$$

```
> model.main<-lm(weightgain~factor(rbGH)+factor(gender))
> model.full<-lm(weightgain~factor(rbGH)*factor(gender))
> anova(model.main, model.full, test="F")
Analysis of Variance Table

Model 1: weightgain ~ factor(rbGH) + factor(gender)
Model 2: weightgain ~ factor(rbGH) * factor(gender)
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1     53 73195
2     48 69564  5    3630.3 0.501  0.774
```


Main effect model
weightgain
100
200
300
400
500
rbGH
1
2
3
4
5
6
Gender=1
Gender=2


Model with interaction effect
weightgain
100
200
300
400
500
rbGH
1
2
3
4
5
6
Gender=1
Gender=2

## 1.3 Linear Models with Matrices

### 1.3.1 Matrix Notations for Linear Model

- Considered statistical model is usually assumed to hold for every unit of the population $\Omega$ in question.

- For the unit $i \in \Omega$, the linear model under normal distribution can be defined by the equations

$$Y_i \sim N(\mu_i, \sigma^2), \tag{1.26a}$$

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \mathbf{x}_i' \boldsymbol{\beta}, \tag{1.26b}$$

  where $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})'$ are observed or set values of the explanatory variables for the unit $i$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)'$ is the vector of unknown parameters and $Y_i$ is an observable random variable associated to the unit $i$.

- It is useful to define the considered statistical model to the selected sampling units $i = 1, 2, \ldots, n$. For example for the units $i = 1, 2, \ldots, n$, the linear model under normality with independence is defined by equations

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}), \tag{1.27a}$$

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{1.27b}$$

where

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \text{ is } n \times 1 \text{ observable random vector,}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \text{ is } n \times 1 \text{ unknown vector of expected values,}$$

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & 1 \end{pmatrix} \text{ is } n \times n \text{ identity matrix,}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \left(\mathbf{1} : \mathbf{x}_{(1)} : \mathbf{x}_{(2)} : \cdots : \mathbf{x}_{(p)}\right) = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ & & & \vdots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

is $n \times (p+1)$ model matrix containing the values of the explanatory variables.

Note that in these notes (and in statistics generally!) the vector $\mathbf{y}$ has double meaning. It is used for denoting the random variable $\mathbf{y} = (Y_1, Y_2, \ldots, Y_n)'$ but also as the vector of realizations of the random variables: $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$.

– A linear normal model is often expressed in a form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{1.28}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ is called as unknown vector of random errors with the expected value $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0} = (0, 0, \ldots, 0)'$ and the covariance matrix $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

## Example 1.6.

Let us consider the following small data, where $X_1$ and $X_2$ are categorical explanatory variables both having class values $\{a, b, c\}$.

```
  X1 X2     Y
1  a  a  89.0
2  a  b 104.4
3  a  c  97.9
4  b  a 105.5
5  b  b 103.7
6  b  c 104.6
7  c  a 103.3
8  c  b  93.5
9  c  c 102.3
```

When modeling the response variable $Y$ by the following linear model

$$\mathcal{M}_{1|2}: \quad Y_i = \beta_0 + \beta_j + \alpha_h + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2),$$

the model $\mathcal{M}_{1|2}$ can be written in matrix form as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. To find out detailed forms for the model matrix $\mathbf{X}$ and parameter vector $\boldsymbol{\beta}$ in case of the model $\mathcal{M}_{1|2}$, we first select the class $a$ as baseline category for both variables $X_1, X_2$. Thus we need dummy variables for the categories $b$ and $c$ for both variables $X_1, X_2$. Let us denote those dummy variables as $x_{i1b}, x_{i1c}, x_{i2b}, x_{i2c}$. Then the main effect model $\mathcal{M}_{1|2}$ can be written as

$$\begin{aligned} \mathcal{M}_{1|2}: \quad & Y_i = \beta_0 + \beta_j + \alpha_h + \varepsilon_i, \\ & Y_i = \beta_0 + \beta_2 x_{i1b} + \beta_3 x_{i1c} + \alpha_2 x_{i2b} + \alpha_3 x_{i2c} + \varepsilon_i. \end{aligned}$$

Thus

$$\mathbf{X} = (\mathbf{1} : \mathbf{x}_{1b} : \mathbf{x}_{1c} : \mathbf{x}_{2b} : \mathbf{x}_{2c}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_3 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}.$$

## Assignment 1.1.

Consider the dataset blackcherry.txt. This dataset provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. Note that girth is the diameter of the tree (in inches) measured at 4 ft 6 in above the ground.

```
          Girth  Height   Volume
1         21.08   21.34     0.29
2         21.84   19.81     0.29
3         22.35   19.20     0.29
4         26.67   21.95     0.46
.                                                                        X1=Girth - Tree diameter in inches
30        45.72   24.38     1.44                                         X2=Height - Height in ft
31        52.32   26.52     2.18                                         Y=Volume - Volume of timber in cubic ft
```

Let us model the Volume variable $Y$ with the linear regression model

$$\mathcal{M}_{1|2}: \quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

(a) Calculate the fitted value $\hat{\mu}_{31}$ for the last observation $i = 31$ in the dataset. Also, calculate the residual $e_{31} = y_{31} - \hat{\mu}_{31}$.

(b) Find the point and 95% confidence interval estimates for the expected value $\mu_{i_*}$, when $x_{i_*1} = 30$ and $x_{i_*2} = 22$.

(c) Find the best linear unbiased prediction and 80% prediction intervals for the new observation $Y_{i_*}$, when $x_{i_*1} = 30$ and $x_{i_*2} = 22$.

(d) Test at 5 % significance level hypotheses

$$\begin{aligned} H_0 &: \text{Model } \mathcal{M}_1 : \quad \mu_i = \beta_0 + \beta_1 x_{i1} \text{ holds,} \\ H_1 &: \text{Model } \mathcal{M}_{1|2} : \quad \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \text{ holds.} \end{aligned}$$

Select the appropriate test statistic to test the above hypotheses. Calculate the value of the chosen test statistic. Calculate the associated $p$-value too.

Also, test the hypotheses

$$\begin{aligned} H_0 &: \text{Model } \mathcal{M}_{1|2} : \quad \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \text{ holds,} \\ H_1 &: \text{Model } \mathcal{M}_{12} : \quad \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} \text{ holds.} \end{aligned}$$

## Assignment 1.2.

The research group was interested in finding out how vitamin C affects the length of the teeth of the guinea pigs. In the study, a total of 60 guinea pigs were divided into three equal groups, with 0.5 milligrams of vitamin C per day administered to guinea pigs in the first group, 1 mg of vitamin C per day, and 2 mg of vitamin C per day. In addition, each of these three groups was divided into two so that half of the group's guinea pigs received vitamin C in the juice (OJ) and half as crystals (VC). The teeth of the guinea pig were measured prior to the beginning of the experimentation and subsequent measurements were made, see file teethpig.txt.

```
        growth          dose         level
1          4.2            VC           0.5
2         11.5            VC           0.5
3          7.3            VC           0.5
.
59        29.4            OJ           2.0
60        23.0            OJ           2.0
```

Let us denote the explanatory variables as $X_1 = \mathsf{level}$ and $X_2 = \mathsf{dose}$. The response variable $Y = \mathsf{growth}$ was explained with the following linear covariance models

$$\begin{aligned}
\mathcal{M}_1 : &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i, \\
\mathcal{M}_{1|2} : &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_j + \varepsilon_i, \\
\mathcal{M}_{12} : &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i1} + \varepsilon_i.
\end{aligned}$$

(a) Under the model $\mathcal{M}_{12}$, calculate the best linear unbiased prediction and 80 % prediction interval for new observation $Y_{i_*}$ when $x_{i_*1} = 2$ and $x_{i_*2}$=VC.

(b) Under the model $\mathcal{M}_{1|2}$. Consider the following hypotheses

$$H_0 : \alpha_j = 0, \qquad H_1 : \alpha_j \neq 0.$$

Select the appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic.

(c) Consider the following hypotheses

$$H_0 : \text{Model } \mathcal{M}_1 \text{ is the true model}, \qquad H_1 : \text{ Model } \mathcal{M}_{12} \text{ is the true model.}$$

Select the appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic.

### 1.3.2 Multivariate Normal Distribution

– Since $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, distributional theory of multivariate normal distribution is important.

– Assume that each independent random variable $Z_1, Z_2, \ldots, Z_n$ follows the standard normal distribution $Z_i \sim N(0, 1)$. Then the random vector $\mathbf{z} = (Z_1, Z_2, \ldots, Z_n)'$ has a joint density function

$$\begin{aligned} f_{\mathbf{z}}(\mathbf{z}) = f(z_1) \cdot f(z_2) \cdot \cdots \cdot f(z_n) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{z_1^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{z_2^2}{2}} \cdot \cdots \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{z_n^2}{2}} \\ &= \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{z_1^2+z_2^2+\cdots+z_n^2}{2}} = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{\mathbf{z}'\mathbf{z}}{2}}. \end{aligned} \tag{1.29}$$

– Denote $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix.

– Consider the linear transformation $\mathbf{y} = \boldsymbol{\mu} + \mathbf{A}\mathbf{z}$, where the rank of the matrix $\mathbf{A}$ is $\text{rank}(\mathbf{A}) = n$. Then the random vector $\mathbf{y} = (Y_1, Y_2, \ldots, Y_n)'$ follows the multivariate normal distribution $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the joint density function being

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} e^{-\frac{(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2}}, \qquad \mathbf{y} \in \mathbb{R}^n, \text{ where } \mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}, \ \mathrm{Cov}(\mathbf{y}) = \boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'.$$

**Theorem 1.1.** If $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with partitioned form

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right],$$

then the followings hold:

(a) Linear transformation $\mathbf{By} + \mathbf{b}$ follows the multivariate normal distribution $\mathbf{By} + \mathbf{b} \sim N(\mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B\Sigma B}')$.

(b) The marginal distributions are $\mathbf{y}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{y}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

(c) The random vectors $\mathbf{y}_1, \mathbf{y}_2$ are independent if and only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

(d) The conditional distribution of the random vector $\mathbf{y}_2$ given $\mathbf{y}_1$ follows the multivariate normal distribution

$$\mathbf{y}_2 | \mathbf{y}_1 \sim N\left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right).$$

– If $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$, then

$$W = \mathbf{z}'\mathbf{z} = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \tag{1.30}$$

follows the $\chi^2$ distribution with $n$ degrees of freedom, $W \sim \chi^2_{(n)}$.

- If $Z \sim N(0,1)$ and $W \sim \chi^2_{(n)}$ are independent, then

$$T = \frac{Z}{\sqrt{\frac{W}{n}}} \tag{1.31}$$

  follows the Student $t$ distribution with $n$ degrees of freedom, $T \sim t_{(n)}$.

- If $W_1 \sim \chi^2_{(n_1)}$ and $W_2 \sim \chi^2_{(n_2)}$ are independent, then

$$F = \frac{\frac{W_1}{n_1}}{\frac{W_2}{n_2}} \tag{1.32}$$

  follows the $F$ distribution with $n_1$ and $n_2$ degrees of freedom, $F \sim F_{(n_1,n_2)}$.