R403: Probabilistic and Statistical Computations with R

Lecture 15: Analysis of Variance (ANOVA)

Kaloyan Ganev

2022/2023

Lecture Contents

Introduction

What is ANOVA?

- Developed by Ronald Fisher (https://en.wikipedia.org/wiki/Ronald_Fisher)
- Decrypted as “Analysis of Variance”
- One of the most frequently used statistical techniques in empirical work
- Assumes linear relationships among variables
- Not a single model but rather a collection (a class) of models
- The basic idea consists in attempting to partition variation into components corresponding to different sources of that variation
- Yet, at the same time, its purpose is not to analyse variances but to analyse **variation in means**
- In other words, it is used to test differences in means for statistical significance

Why ANOVA?

- The method is employed in scientific studies in the analysis of data (experimental, observational, or mixed)
- Allows to identify the presence or absence of statistically significant effects from the presence of experimental treatments or observational factors
- Allows to identify effects of a variable after controlling for other variables' influence
- Simple and robustly designed technique

A Quick Review of Quadratic Forms

Quadratic Forms Defined

Definition 1

A **homogeneous polynomial** is a polynomial whose non-zero terms are of one and the same degree.

Definition 2

A **quadratic form** is a homogeneous polynomial of degree 2 in n variables.

Definition 3

A **real quadratic form** is a quadratic form whose variables and coefficients are all real.

We will deal exclusively with real quadratic forms.

Examples of Quadratic Forms

- The following is a quadratic form in X_1, X_2 and X_3:

$$X_1^2 + X_2^2 + X_3^2 - 2X_1X_2$$

- The following is NOT a quadratic form in X_1 and X_2 (why?):

$$X_1^2 + X_2^2 - 2X_1 - 4X_2 + 5$$

Sample Variance as A Quadratic Form

- Let $\overline{X}$ and S^2 be respectively the sample mean and the sample variance of an arbitrary distribution
- We know that:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

where n is sample size

- This can also be written as follows:

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$$

- The sample mean equals:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Sample Variance as A Quadratic Form (2)

- Plug this expression into the right-hand side of the variance expression to get:

$$(n-1)S^2 = \sum_{i=1}^{n} \left(X_i - \frac{X_1 + X_2 + \ldots + X_n}{n} \right)^2$$

- Expand the stuff in the parentheses:

$$(n-1)S^2 = \frac{n-1}{n} \sum_{i=1}^{n} X_i^2 - \frac{1}{n} \sum_{i \neq j} X_i X_j$$

- Obviously, this is a quadratic form in $X_1, X_2, \ldots, X_n$

Sample Variance as A Quadratic Form (3)

- Note that if the sample is drawn from a $N(\mu, \sigma^2)$ distribution, then:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

- Note also that this result is independent from the value of μ
- This, for example, allows to construct confidence intervals for σ^2 when μ is unknown

The Distribution of Certain Quadratic Forms in n.i.d. Random Variables

Theorem 1

Let $Q_1, Q_2, \ldots, Q_k, Q$ be $(k+1)$ real quadratic forms in n random variables (each random variable being n.i.d. (μ, σ^2)) where:

$$Q = Q_1 + Q_2 + \ldots + Q_{k-1} + Q_k$$

If $\frac{Q}{\sigma^2} \sim \chi^2(r)$, $\frac{Q_1}{\sigma^2} \sim \chi^2(r_1), \ldots, \frac{Q_{k-1}}{\sigma^2} \sim \chi^2(r-1)$ and $Q_k \geq 0$, then:

(a) $Q_1, Q_2, \ldots, Q_k$ *are independent, from which also follows:*

(b) $\frac{Q_k}{\sigma^2} \sim \chi^2(r_k)$, *where* $r_k = r - (r_1 + \ldots + r_{k-1})$

The Distribution of Certain Quadratic Forms in n.i.d. Random Variables (2)

- Let $X \sim N(\mu, \sigma^2)$
- Draw a sample of size $n = ab$ from the above distribution
- Each drawing is independent and produces itself a random variable with mean μ and variance σ^2
- If we arrange the observations in a rows and b columns, the arrangement would look like:

$$\begin{array}{cccc} X_{11} & X_{12} & \cdots & X_{1b} \\ X_{21} & X_{22} & \cdots & X_{2b} \\ \cdots & \cdots & \cdots & \cdots \\ X_{a1} & X_{12} & \cdots & X_{ab} \end{array}$$

- With this, we could think of having a samples (by rows) or b samples (by columns)

The Distribution of Certain Quadratic Forms in n.i.d. Random Variables (3)

We define now some statistics:

- Overall (grand) mean:

$$\overline{X}_{..} = \frac{1}{ab} \sum_{i=1}^{a} \sum_{j=1}^{b} X_{ij}$$

- Row means:

$$\overline{X}_{i.} = \frac{1}{b} \sum_{j=1}^{b} X_{ij}, \quad i = 1, 2, \ldots, a$$

- Column means:

$$\overline{X}_{.b} = \frac{1}{a} \sum_{i=1}^{a} X_{ij}, \quad j = 1, 2, \ldots, b$$

(In total: $a + b + 1$)

The Distribution of Certain Quadratic Forms in n.i.d. Random Variables (4)

- We will consider an example
- We have a sample of size $n = ab$; the sample variance is S^2
- Using the sample variance formula, we can write:

$$(ab-1)S^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \overline{X}_{..})^2$$

- Add and subtract $\overline{X}_{i.}$ in the parentheses:

$$(ab-1)S^2 \quad = \quad \sum_{i=1}^{a}\sum_{j=1}^{b}[(X_{ij} - \overline{X}_{i.}) + (\overline{X}_{i.} - \overline{X}_{..})]^2$$

The Distribution of Certain Quadratic Forms in n.i.d. Random Variables (5)

- Then expand to get:

$$(ab-1)S^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij}-\overline{X}_{i\cdot})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{X}_{i\cdot}-\overline{X}_{\cdot\cdot})^2,$$

where we made use of the fact that $2\sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij}-\overline{X}_{i\cdot})(\overline{X}_{i\cdot}-\overline{X}_{\cdot\cdot}) = 0.$

- Note that there is no j index in the parentheses in the far-right-hand-side double sum; therefore:

$$(ab-1)S^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij}-\overline{X}_{i\cdot})^2 + b\sum_{i=1}^{a}(\overline{X}_{i\cdot}-\overline{X}_{\cdot\cdot})^2$$

The Distribution of Certain Quadratic Forms in n.i.d. Random Variables (6)

- All those are quadratic forms, and we can write the last expression on the previous slide more briefly as follows:

$$Q = Q_1 + Q_2$$

- First, we will apply Theorem 1 to show that Q_1 and Q_2 are independent
- We know that $Q = \dfrac{(ab-1)S^2}{\sigma^2} \sim \chi^2(ab-1)$
- Now consider Q_1. We can write it also in the following way:

$$Q_1 = \sum_{i=1}^{a} \left[(b-1) \left(\frac{1}{b-1} \sum_{j=1}^{b} (X_{ij} - \overline{X}_{i.})^2 \right) \right]$$

The Distribution of Certain Quadratic Forms in n.i.d. Random Variables (7)

- For each value of i, the expression in the big parentheses is the sample variance of a sample of size b
- The ratio:

$$\frac{\frac{1}{b-1}\sum_{j=1}^{b}(X_{ij}-\overline{X}_{i.})^2}{\sigma^2}$$

has therefore $\chi^2(b-1)$ distribution

- $\frac{Q_1}{\sigma^2}$ is practically the sum a such ratios (multiplied by $(b-1)$)
- Therefore:

$$\frac{Q_1}{\sigma^2} \sim \chi^2(a(b-1))$$

The Distribution of Certain Quadratic Forms in n.i.d. Random Variables (7)

- It is obvious that $Q_2 \geq 0$ since it is a sum of squares multiplied by a positive number b
- From Theorem 1 follows that Q_1 and Q_2 are independent and $\frac{Q_2}{\sigma^2} \sim \chi^2(a-1)$
- The ratios of those χ^2 statistics form F statistics which can be used to test some interesting statistical hypotheses
- For an exercise, try the same by replacing $X_{ij} - \overline{X}_{..}$ by $(X_{ij} - \overline{X}_{.j}) + (\overline{X}_{.j} - \overline{X}_{..})$

One-Way ANOVA

One-Way ANOVA

- In the beginning we mentioned briefly that ANOVA is a method of comparing the means of several populations
- Often those populations are assumed to be normally distributed
- However, in addition to that, ANOVA allows point and interval estimation
- In this context, inference is based on the so-called **contrasts**
- We will make an introduction to this matter in what follows

One-Way ANOVA (2)

- Consider the following independent random variables:

$$X_j \sim N(\mu_j, \sigma^2), \quad j = 1, 2, \ldots, b,$$

  where all μ_j are unknown but σ^2 is known and common to all variables (homoscedasticity)
- Let $X_{1j}, X_{2j}, \ldots, X_{aj}$, $j = 1, 2, \ldots, b$ be random samples of size a from each variable
- Assume that each observation (the data) X_{ij} in the b samples is most appropriately described by the following model:

$$X_{ij} = \mu_j + e_{ij}, \quad i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, b,$$

  where $e_{ij} \sim N(0, \sigma^2)$

One-Way ANOVA (3)

- From the practical perspective, the b samples can be thought of as groups undergoing treatments (e.g. with different medicines in clinical trials)
- In the language of ANOVA, this is a case where we have one factor (in this case medication applied) at b levels
- Therefore, the model is called a **one-way model** (the effects of a single factor only are studied)
- The mathematical formulation of the model is interpreted as follows: the outcomes X_{ij} are the result of systematic causes (μ_j) and random causes (e_{ij})

One-Way ANOVA (4)

- We are virtually not able to separate the two types of influences
- However, by using multiple samples (in our case b samples), after comparing the means we are able to tell whether a specific treatment is effective
- The comparison is based on statistical methods, i.e. deciding whether a treatment is effective is based on tests of significance
- Note that based on the above-said, ANOVA is analogical to t-tests; the latter, however, cannot be used when more than two groups of data are compared!

One-Way ANOVA (5)

- The classical ANOVA hypothesis:

$$H_0 : \quad \mu_1 = \mu_2 = \ldots = \mu_b = \mu,$$

  where μ is unspecified, tested against a general alternative H_1 ("at least one mean is different")
- Let X_j be the response from the jth treatment, and let $\mu_j = \mathsf{E}(X_j)$
- If $X_j \sim N(\mu_j, \sigma^2)$, then H_0 states that all treatments have the same effect
- A likelihood ratio test is used to test the validity of H_0
- The aim of the test is to compare the ratio of variances with the one present when means differ only due to random influences

One-Way ANOVA (6)

- Using the formula for the multivariate normal density, the first likelihood function describing the case when all means are equal is:

$$L(\omega) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{ab} \exp\left[-\frac{1}{2\sigma}\sum_{j=1}^{b}\sum_{i=1}^{a}(x_{ij}-\mu)^2\right]$$

where
$\omega = \{(\mu_1, \mu_2, \ldots, \mu_b, \sigma^2) : -\infty < \mu_1 = \ldots = \mu_b = \mu < \infty, 0 < \sigma^2 < \infty\}$

One-Way ANOVA (7)

- The values of μ and σ^2 that maximize this function are respectively:

$$\overline{x}_{..} = \frac{1}{ab} \sum_{j=1}^{b} \sum_{i=1}^{a} x_{ij}$$

$$v = \frac{1}{ab} \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \overline{x}_{..})^2$$

and the maximum of the function equals:

$$L(\widehat{\omega}) = \left[\sqrt{\frac{ab}{2\pi \sum_{i=1}^{a} (x_{ij} - \overline{x}_{..})^2}} \right]^{ab} \exp\left(-\frac{ab}{2} \right)$$

One-Way ANOVA (8)

- The second likelihood function relates to the case when means are not equal:

$$L(\Omega) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{ab} \exp\left[-\frac{1}{2\sigma}\sum_{j=1}^{b}\sum_{i=1}^{a}(x_{ij} - \mu_j)^2\right]$$

where $\Omega = \{(\mu_1, \mu_2, \ldots, \mu_b, \sigma^2) : -\infty < \mu_j < \infty, 0 < \sigma^2 < \infty\}$

One-Way ANOVA (9)

- The values of μ and σ^2 that maximize this function are respectively:

$$\overline{x}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a} x_{ij}$$

$$w = \frac{1}{ab}\sum_{j=1}^{b}\sum_{i=1}^{a}(x_{ij} - \overline{x}_{\cdot j})^2$$

and the maximum of the function equals:

$$L(\widehat{\Omega}) = \left[\sqrt{\frac{ab}{2\pi\sum_{i=1}^{a}(x_{ij} - \overline{x}_{\cdot j})^2}}\right]^{ab} \exp\left(-\frac{ab}{2}\right)$$

One-Way ANOVA (10)

- Take the likelihood ratio:

$$\Lambda = \frac{L(\widehat{\omega})}{L(\widehat{\Omega})} = \left[\sqrt{\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij} - \overline{x}_{.j})^2}{\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij} - \overline{x}_{..})^2}} \right]^{ab}$$

- Recall from the quadratic forms example that $Q = \sum_{j=1}^{b}\sum_{i=1}^{a}(X_{ij} - \overline{X}_{..})^2$. From this follows that the statistic:

$$V = \frac{1}{ab}\sum_{j=1}^{b}\sum_{i=1}^{a}(X_{ij} - \overline{X}_{..})^2 = \frac{Q}{ab}$$

One-Way ANOVA (11)

- If (and when) you do the exercise on quadratic forms, you get to the following expressions:

$$Q_3 = \sum_{j=1}^{b} \sum_{i=1}^{a} (X_{ij} - \overline{X}_{\cdot j})^2$$

$$Q_4 = a \sum_{j=1}^{b} (\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot})^2$$

- By the theorem we also find out that Q_3 and Q_4 are independent, and:

$$\frac{Q_3}{\sigma^2} \sim \chi^2(b(a-1)), \quad \frac{Q_4}{\sigma^2} \sim \chi^2(b-1)$$

One-Way ANOVA (12)

- Thus Λ can be viewed as:

$$\Lambda = \left(\sqrt{\frac{Q_3}{Q}}\right)^{ab}$$

or:

$$\Lambda^{2/ab} = \frac{Q_3}{Q}$$

- Because $Q = Q_3 + Q_4$,

$$\Lambda^{2/ab} = \frac{Q_3}{Q_3 + Q_4} = \frac{1}{1 + Q_4/Q_3}$$

One-Way ANOVA (13)

- The significance level that corresponds to the test of H_0 is:

$$\alpha = P_{H_0}\left[\frac{1}{1+Q_4/Q_3} \leq \lambda_0^{2/ab}\right] = P_{H_0}\left[Q_4/Q_3 \geq \lambda_0^{-2/ab} - 1\right]$$

- Multiply both sides of the inequality by $\dfrac{b(a-1)}{b-1}$ (a positive number) and denote $c = \dfrac{b(a-1)}{b-1}(\lambda_0^{-2/ab} - 1)$ to get:

$$\alpha = P_{H_0}\left[\frac{Q_4/(b-1)}{Q_3/[b(a-1)]} \geq c\right]$$

One-Way ANOVA (14)

- But at the same time we know that:

$$F = \frac{Q_4/[\sigma^2(b-1)]}{Q_3/[\sigma^2 b(a-1)]} = \frac{Q_4/[(b-1)]}{Q_3/[b(a-1)]} \sim F_{b-1,b(a-1)}$$

- Therefore, H_0 can be tested with an F-statistic
- The critical point is chosen from the F-table – the one that corresponds to the specified significance level and the respective degrees of freedom
- Note: Testing means equality does not require that sample size to be equal across the b samples

ANOVA in R

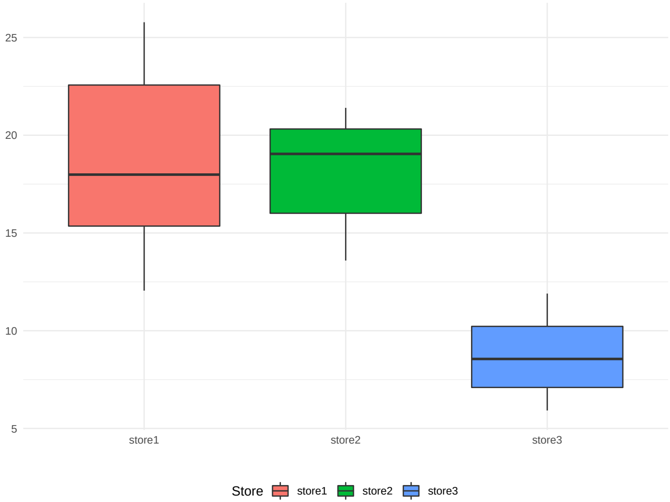
An Example of One-Way ANOVA in R

- Start with the following dataset: `three_stores.csv`[1]
- Explanation of data: three stores, for each store the values of six purchases are randomly sampled
- Question: Do the three stores have the same average amount per purchase?
- Load the data in R and calculate averages:

```
avgbuy <- read.csv("three_stores.csv")
attach(avgbuy)
colMeans(avgbuy) # Means per store
mean(unlist(avgbuy)) # Grand mean for all stores
boxplot(avgbuy, col = c("red","green","blue"))
```

[1] Data taken from Hanke and Reitsch (1991), p. 405.

An Example of One-Way ANOVA in R (2)



An Example of One-Way ANOVA in R (3)

- Null hypothesis: the three populations (stores) have equal means
- Looks like the first two means are approximately equal but the third is not; the question is whether this inequality is due to chance only or not
- Before running ANOVA, the data should be stacked:

```
avgbuys <- stack(avgbuy)
names(avgbuys)
```

- With the above names, now run the ANOVA:

```
myanova <- aov(formula = values ~ ind, data = avgbuys)
myanova
summary(myanova)
```

Two-Way ANOVA

Two-Way ANOVA

- In one-way ANOVA we dealt with one factor at b levels
- Now, let there be two factors, A and B, respectively having a and b levels
- In such a setting, we have **two-way (two-factor) ANOVA**
- Let $X_{ij} \sim n.i.d(\mu_{ij}, \sigma^2)$, $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$ the values of the responses when A is at level i and B is at level j
- The total number of levels pairs is $n = ab$ (each pair is actually a treatment)
- The mean μ_{ij} is interpreted as the mean response to a treatment

Two-Way ANOVA (2)

- Denote:

$$\overline{\mu} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}, \quad \overline{\mu}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij}, \quad \overline{\mu}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ij}$$

  where $i = 1, 2, \ldots, a,\ j = 1, 2, \ldots, b$

- These are respectively the grand mean, the row means, and the column means (if we arrange factor levels in a table)
- Consider the following additive model:

$$\mu_{ij} = \overline{\mu} + (\overline{\mu}_{i\cdot} - \overline{\mu}) + (\overline{\mu}_{\cdot j} - \overline{\mu})$$

- It is interpreted as follows: the mean μ_{ij} is the result of the additive main effect[2] of level i of factor A and the additive main effect of level j of factor B

[2] Besides main effects, there could also be interaction effects which are non-additive.

Two-Way ANOVA (3)

- For simplicity, denote:

$$\mu = \overline{\mu}, \quad \alpha_i = \overline{\mu}_{i.} - \overline{\mu}, \quad \beta_j = \overline{\mu}_{.j} - \overline{\mu}$$

- Then the model becomes:

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

where $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$

- In empirical data this relationship is not exactly fulfilled
- If we add random disturbances, we get:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

- This is referred to as the **two-way ANOVA model**

Two-Way ANOVA Setup: A Numerical Example

- Taken from Kutner et. al. (2005), p. 817
- Consider the following table of mean learning times (in minutes):

Factor A: Gender	**Factor B: Age**			**Row average**
	Young	**Middle**	**Old**	
Male	9	11	16	12
Female	9	11	16	12
Column average	9	11	16	12

- In this case we have $a = 2$ and $b = 3$

Two-Way ANOVA Setup: A Numerical Example (2)

- The main gender effects (in minutes) are:

$$\alpha_1 = 12 - 12 = 0$$
$$\alpha_2 = 12 - 12 = 0$$

- From this it turns out that gender has no effect on average learning times
- The main age effects (in minutes) are:

$$\beta_1 = 9 - 12 = -3$$
$$\beta_2 = 11 - 12 = -1$$
$$\beta_3 = 16 - 12 = 4$$

- This shows that mean learning time increases with age, i.e. age has an effect
- (This model could however be reduced to a one-way ANOVA)

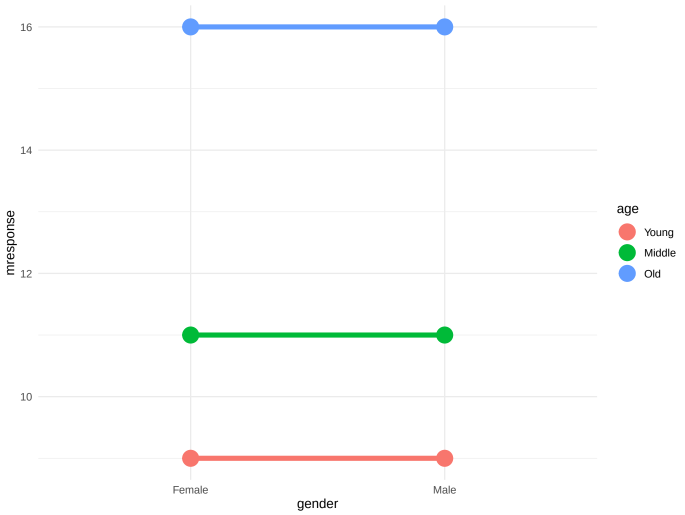
Two-Way ANOVA Setup: A Numerical Example (3)

- It is easy to see that all mean responses can be obtained using the model equation
- For example:

$$\mu_{21} = 12 + 0 - 3 = 9$$
$$\mu_{13} = 12 + 0 + 4 = 16$$

- We also note that for additive models, when we plot μ_{ij} against j, we get parallel lines
- Such plots are called **treatment mean plots**, **mean profile plots**, or **interaction plots**

Two-Way ANOVA Setup: A Numerical Example (4)



Two-Way ANOVA Setup: A Second Example

- Same as the previous, changed values of μ_{ij}

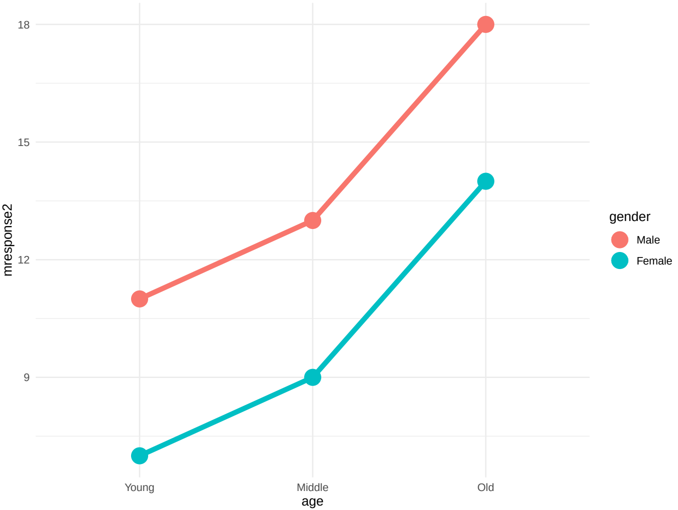
Factor A: Gender	**Factor B: Age** Young	Middle	Old	**Row average**
Male	11	13	18	14
Female	7	9	14	10
Column average	9	11	16	12

- The main gender effects (in minutes) are:

$$\alpha_1 = 14 - 12 = 2$$
$$\alpha_2 = 10 - 12 = -2$$

- This time gender has an effect on average learning times

Two-Way ANOVA Setup: A Second Example (2)



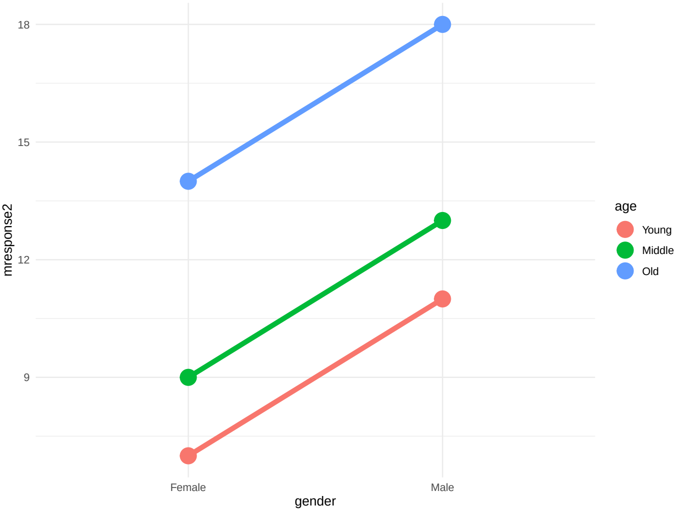
Two-Way ANOVA Setup: A Second Example (3)

- The main age effects (in minutes) are:

$$\beta_1 = 9 - 12 = -3$$
$$\beta_2 = 11 - 12 = -1$$
$$\beta_3 = 16 - 12 = 4$$

- In this case also mean learning time increases with age, i.e. age has an effect
- The graph differs, however

Two-Way ANOVA Setup: A Second Example (4)



Two-Way ANOVA: The Main Effects Hypotheses

- Having in mind the two examples and the knowledge we have on one-way ANOVA, we can now state the relevant hypotheses:

Main Effects Hypotheses

$$H_{0,A} : \alpha_1 = \ldots = \alpha_a = 0$$
$$H_{1,A} : \text{at least one } \alpha \neq 0$$

$$H_{0,B} : \beta_1 = \ldots = \beta_b = 0$$
$$H_{1,B} : \text{at least one } \beta \neq 0$$

Additional Readings

Additional Readings

- Hogg, R., J. McKean and A. Craig (2013): *Introduction to Mathematical Statistics*, Pearson, 7th edn.
- Kutner, M., C. Nachtsheim, J. Neter and W. Li (2005): *Applied Linear Statistical Models*, McGraw-Hill Irwin, 5th edn.
- Tabachnick, B. and L. Fidell (2013): *Using Multivariate Statistics*, Pearson, 6th edn.