Linear models:

One-way and two-way ANOVAs

Linear models

- Response = deterministic part + stochastic part
  - Stochastic = random
  - Deterministic = systematic
- Linear models are so called because the expected response can be treated as the results of explanatory variables whose effects are additive

General linear models

So called “general linear models” are defined by one feature:

- The assumed error around the deterministic portion of the model is normally distributed

$$y_i = \text{systematic part} + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

- Note: The assumption of normality applies to the variation around the expected value (residuals) not to the data as a whole.

T-test

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

T-test

T-test example from last time

- What is the effect of region (North or South) on mass?

$$mass_i = \alpha + \beta * region_i + \varepsilon_i$$
$$\varepsilon_i \sim Norm(0, \sigma^2)$$

- Snake mass is composed of three components:
  1. a constant (alpha),
  2. the product of another constant (beta) with the value of the indicator for region in which snake was caught
  3. The error (ε_i) that is specific to snake.
- Another way to write this model is:

$$mass_i \sim Normal\ (\alpha + \beta * region_i, \sigma^2)$$

T-test

Design matrix – what does the variable region look like?

lm(mass ~ region)
model.matrix(~region)

(Intercept)	region2
1	0
1	0
1	0
1	0
1	1
1	1

Translates into a set of equations:

$$6 = \alpha * 1 + \beta * 0 + \varepsilon_1$$
$$8 = \alpha * 1 + \beta * 0 + \varepsilon_2$$
$$5 = \alpha * 1 + \beta * 0 + \varepsilon_3$$
$$7 = \alpha * 1 + \beta * 0 + \varepsilon_4$$
$$9 = \alpha * 1 + \beta * 1 + \varepsilon_5$$
$$9 = \alpha * 1 + \beta * 1 + \varepsilon_6$$

Or in matrix notation:

$$\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 9 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

This is an ***effects parameterization*** of the t-test

Linear models

What if I change this model with two categorical variables:

$$mass_i = \alpha + \beta * region_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

To this (where j=1,2,3 so that there were 3 regions instead of two):

$$mass_i = \alpha + \beta_{j(i)} * region_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

Linear models

What is this model now?

ANOVA!!

To this (where j=1,2,3 so that there were 3 regions instead of two):

$$mass_i = \alpha + \beta_{j(i)} * region_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

ANOVA

- **AN**alysis **O**f **VA**riance
- Used to analyze the differences in means among groups (and can also be used to look at differences in variances)
- In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, generalizing the *t*-test to more than two groups.
- Why not due multiple t-tests? Increased chance of a type I error (rejected the null hypothesis when it's true)

One way ANOVA

- What is the effect of region (North, Central, South) on mass?

$$mass_i = \alpha + \beta_{j(i)} * region_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

- Snake mass is composed of three components:
  1. A constant (alpha),
  2. The product of another constant (beta) with the value of the indicator for region in which snake was caught
  3. The error (ε_i) that is specific to snake.
- Another way to write this model is:

$$mass_i \sim Normal\left(\alpha + \beta_{j(i)} * region_i, \sigma^2\right)$$

What is the interpretation of the parameters α and $\beta_{j(i)}$?

One way ANOVA

- How can we write the same model with a means parameterization?

$$mass_i = \beta_{j(i)} * region_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

- Another way to write this model is:

$$mass_i \sim Normal\left(\beta_{j(i)} * region_i, \sigma^2\right)$$

What is the interpretation of the parameters $\beta_{j(i)}$? How many $\beta_{j(i)}$ parameters are there in this model?

One way ANOVA

Suppose we have a single explanatory variable with three levels (region = North, Central, or South) on a continuous response (mass). *Question: Is mass different for snakes in different regions?*

Example with
10 data points:

Individual	Location	Mass
1	North	6
2	North	8
3	North	5
4	North	7
5	Central	5
6	Central	4
7	Central	6
8	South	8
9	South	8
10	South	9

One way ANOVA

Translates into a set of equations:

$$
\begin{aligned}
6 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_1 \\
8 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_2 \\
5 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_3 \\
7 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_4 \\
5 &= \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \varepsilon_5 \\
4 &= \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \varepsilon_6 \\
6 &= \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \varepsilon_7 \\
8 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \varepsilon_8 \\
8 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \varepsilon_9 \\
9 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \varepsilon_{10}
\end{aligned}
$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

Or in matrix notation:

$$
\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 5 \\ 4 \\ 6 \\ 8 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \end{pmatrix}
$$

One way ANOVA

Translates into a set of equations:

$$
\begin{aligned}
6 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_1 \\
8 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_2 \\
5 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_3 \\
7 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_4 \\
5 &= \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \varepsilon_5 \\
4 &= \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \varepsilon_6 \\
6 &= \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \varepsilon_7 \\
8 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \varepsilon_8 \\
8 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \varepsilon_9 \\
9 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \varepsilon_{10}
\end{aligned}
$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

Or in matrix notation:

$$
\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 5 \\ 4 \\ 6 \\ 8 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \end{pmatrix}
$$

What is kind of parameterization is this?

effects parameterization

One way ANOVA

Translates into a set of equations:

$$
\begin{aligned}
6 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_1 \\
8 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_2 \\
5 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_3 \\
7 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \varepsilon_4 \\
5 &= \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \varepsilon_5 \\
4 &= \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \varepsilon_6 \\
6 &= \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \varepsilon_7 \\
8 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \varepsilon_8 \\
8 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \varepsilon_9 \\
9 &= \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \varepsilon_{10}
\end{aligned}
$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

Or in matrix notation:

$$
\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 5 \\ 4 \\ 6 \\ 8 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \end{pmatrix}
$$

How many columns would there be in the design matrix if there were five regions?

Five including alpha; ß = levels - 1 = 5-1=4

ANOVA

- One way:

  An ANOVA hypothesis that tests the difference in population/group means based on one characteristic or factor

- Two way ANOVA:

  An ANOVA hypothesis that tests comparisons between population/group means based on two or more characteristics or factors

  - These factors can be additive (main effects) or multiplicative (interaction effects)

Two way ANOVA

- What is the effect of region (North, Central, South) and habitat type (forest or grassland) on mass?

$$mass_i = \alpha + \beta_{j(i)} * region_i + \gamma_{k(i)} * habitat_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

- Snake mass is composed of four components:
  1. A constant (alpha),
  2. The product of a constant (beta) with the value of the indicator for region in which snake was caught
  3. The product of a constant (gamma) with the value of the indicator for the habitat type in which snake was caught
  4. The error (ε_i) that is specific to snake.

Assumption: the effects of region and habitat are independent

Two way ANOVA

- What is the effect of region (North, Central, South) and habitat type (forest or grassland) on mass?

$$mass_i = \alpha + \beta_{j(i)} * region_i + \gamma_{k(i)} * habitat_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

- What kind of parameterization is this?
- How many columns are in the design matrix?
  - One for the intercept term (α) which is the mean in forests in the north region
  - Two for the $\beta_{j(i)}$ terms for the effects of region (central, south)
  - One for $\gamma_{k(i)}$ indicating the grassland habitat

Two way ANOVA

What if the relationship between snake mass and habitat type changed by region?

- In this case, we would say the factors are *interacting*

Two way ANOVA

- What is the effect of region (North, Central, South) and habitat type (forest or grassland) on mass in the case when region influences habitat type?

$$mass_i = \alpha + \beta_{j(i)} * reg_i + \gamma_{k(i)} * hab_i + \delta_{k(i)} * reg_i * hab_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

- Snake mass is composed of five components:
  1. A constant (alpha)
  2. The product of a constant (beta) with the value of the indicator for region in which snake was caught
  3. The product of a constant (gamma) with the value of the indicator for the habitat type in which snake was caught
  4. The product of a constant (delta) with the value of the indicator for the region and an indicator for habitat type
  5. The error (ε_i) that is specific to snake

Assumption: the effects of region and habitat are interactive

Two way ANOVA

- What is the effect of region (North, Central, South) and habitat type (forest or grassland) on mass in the case when region influences habitat type?

$$mass_i = \alpha + \beta_{j(i)} * reg_i + \gamma_{k(i)} * hab_i + \delta_{k(i)} * reg_i * hab_i + \varepsilon_i$$
$$\varepsilon_i \sim Norm(0, \sigma^2)$$

- How many parameters are there to estimate?

  This one is easy: 6, the number of combinations in the model (2*3). Essentially, each region/habitat combination gets its own estimate. (Plus, of course, the variance term.)

$$mass_i = \alpha_{jk(i)} * reg_i * hab_i + \varepsilon_i$$
$$\varepsilon_i \sim Norm(0, \sigma^2)$$

Lab

One-way ANOVA:

-Pull up the incomplete R script and we will work through it together

Two-way ANOVA:

-The complete script is available on Github. Work through it in small groups to make sure you understand what's going on