Design and analysis of experiments
Lecture 8

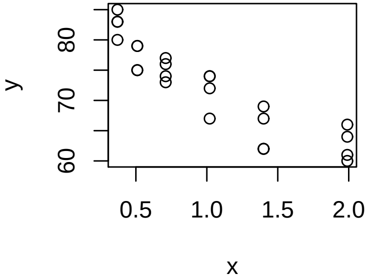
Jakob G. Rasmussen
Department of Mathematics
Aalborg University
Denmark

Second homework

- Homework number 2 is now on the homepage.
- Same rules as last time.
- At the end of todays lecture you have all the tools you need for solving this.
- The handin date is: Monday October 21th.

Lack of fit test

- If an explanatory variable is continuous but have only been measured at some discrete levels (with replicates) we can either use ANOVA or linear regression.
- This can be used to check the fit of the regression model.
- Notice that the regression only has two parameters, while the ANOVA has one for each group - the ANOVA is more flexible and contains the regression model as a special case!



Hypothesis

- We use the ANOVA as base model, and try to check if we can simplify it to the regression model:

$$
\begin{aligned}
H_0 &: y_{ij} = \alpha + \beta x_i + \epsilon_{ij} \qquad \text{(regression)} \\
H_1 &: y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{(ANOVA)}
\end{aligned}
$$

- Or equivalently (note: i is the group number)

$$
\begin{aligned}
H_0 &: \mu_i = \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \\
H_1 &: \mu_i = \bar{y}_i
\end{aligned}
$$

- We can split the variation in the data into two terms, pure error and lack of fit:

$$
y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)
$$

Test statistic

- Written as sums of square:

$$\sum_{i,j}(y_{ij} - \hat{y}_i)^2 = \sum_{i,j}(y_{ij} - \bar{y}_i)^2 + \sum_i n_i(\bar{y}_i - \hat{y}_i)^2$$

$$\begin{aligned} SS &: \quad SS_E = SS_{PE} + SS_{LoF} \\ d.f. &: \quad (n-2) = (n-k) + (k-2) \end{aligned}$$

- The usual F test statistic:

$$F_0 = \frac{MS_{LoF}}{MS_{PE}} = \frac{SS_{LoF}/(k-2)}{SS_{PE}/(n-k)} \sim F_{k-2,n-k}$$

- We reject H_0 if $F_0 > F_{k-2,n-k;\alpha}$
- Rejection means that the regression does not fit well compared to what can be achieved by the ANOVA, which suggests the model can be improved, fx by using polynomial regression.

R

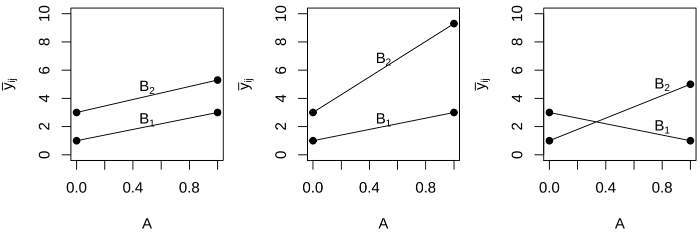
- R-demo, part 1
- Exercise 1

ANOVA with multiple factors

- We have already seen ANOVA with 2, 3 and even 4 factors, when we looked at design with blocks or latin square designs, but now we will have a thorough look at general ANOVA with more than one factor.
- The most important concept which we have not looked at in ANOVA with multiple factors is interaction.

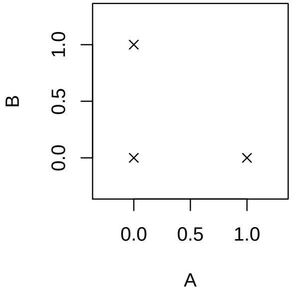
Interaction

- Example:
  - We may want to test whether the late-night driving increases the risk of dying.
  - And maybe we also want to check the increased mortality resulting from drinking.
  - These are two factors, but the combined risk is probably much higher than the sum of the risks, since drinking-and-driving is a particularly bad idea.
  - That is, the two factors interact.
- The interaction plot shows interaction:

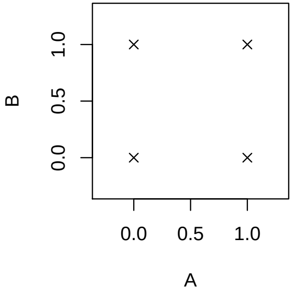


Two designs

- Change only one factor at a time
- This assumes that there is no interaction
- Typically a bad idea



- All combinations
- Interaction is accounted for
- A much better design



Two-way ANOVA with fixed factors

- Means model:

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

- Model without interaction (additive model):

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

- Model with interaction (two-way ANOVA):

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

- Note that:
  - $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, n$ (balanced case)
  - $\epsilon_{ijk} \sim N(0, \sigma^2)$ independent

Overspecification

- The model is overspecified - μ_{ij} has ab parameters, $\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ has $1 + a + b + ab$!
- Method 1 (reference cell $(1, 1)$):
  - $\alpha_1 = 0$
  - $\beta_1 = 0$
  - $(\alpha\beta)_{1j} = 0$ for all j
  - $(\alpha\beta)_{i1} = 0$ for all i
- Method 2 (zero means):
  - $\sum \alpha_i = 0$
  - $\sum \beta_j = 0$
  - $\sum_i (\alpha\beta)_{ij} = 0$ for all j
  - $\sum_j (\alpha\beta)_{ij} = 0$ for all i

Hypotheses

- Hypothesis on the interaction term:

$$H_0^{AB} : (\alpha\beta)_{ij} = 0 \text{ for all } i, j \text{ (i.e. additive model)}$$
$$H_1^{AB} : (\alpha\beta)_{ij} \text{ not all } 0$$

- Hypothesis on factor A:

$$H_0^{A} : \alpha_i = 0 \text{ for all } i$$
$$H_1^{A} : \alpha_i \text{ not all } 0$$

- Hypothesis on factor B:

$$H_0^{B} : \beta_j = 0 \text{ for all } j$$
$$H_1^{B} : \beta_j \text{ not all } 0$$

- Hierarchical principle: Test higher order terms first, i.e. interaction should be tested before main effects.

Test statistics

- We use sums of square as usual for making test statistics - we skip the details:
  - Total: $SS_T = \sum_{ijk}(y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2, \quad \nu_T = abn - 1$
  - Error: $SS_E = \sum_{ijk}(y_{ijk} - \bar{y}_{ij\bullet})^2, \quad \nu_E = ab(n-1)$
  - Interaction: $SS_{AB} = \sum_{ijk}(\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2, \quad \nu_{AB} = (a-1)(b-1)$
  - Factor A: $SS_A = \sum_{ijk}(\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2, \quad \nu_A = a - 1$
  - Factor B: $SS_B = \sum_{ijk}(\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2, \quad \nu_B = b - 1$
- As usual SS_T is the sum of all the other terms, implying that the variation in the data has been separated into various sources (same holds for ν_T).
- All SS are χ^2 distributed under the appropriate null-hypotheses, giving the following test statistics:
  - Interaction: $F_0^{AB} = \frac{MS_{AB}}{MS_E} = \frac{SS_{AB}/\nu_{AB}}{SS_E/\nu_E} \sim F_{\nu_{AB},\nu_E}$
  - Factor A: $F_0^A = \frac{MS_A}{MS_E} = \frac{SS_A/\nu_A}{SS_E/\nu_E} \sim F_{\nu_A,\nu_E}$
  - Factor B: $F_0^B = \frac{MS_B}{MS_E} = \frac{SS_B/\nu_B}{SS_E/\nu_E} \sim F_{\nu_B,\nu_E}$

The unbalanced case

- When the number n_{ij} of observations in each combination of groups are not the same, we are in the unbalanced case - here the formulas do not work, and we have to be careful when using R.
- The proportional case only changes the formulas slightly, and calculations in R will work directly:

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}}$$

- Example of numbers of observations for proportional data:

4	4	2
2	2	1
2	2	1

The unbalanced and unproportional case

- Two unproportional cases:

4	4	4
4	3	4
4	4	4

4	4	4
4	5	4
4	4	4

- Missing observations:
  - This can occur if an observation is missing.
  - This can be fixed by imputing the missing data, fx by the mean of the other measurements in the same group.
  - But why is the data missing: random (ok) or reason (problematic).
- Too many observations:
  - We do extra measurements of fx one or more new treatments.
  - We can fix this by removing the extra measurements (choose at random).
  - But obviously this is stupid if we have made the extra measurements on purpose.
- The above methods are a bit outdated, and with modern computers it is no problem to handle the more difficult formulas in the unbalanced, unproportional case.

R

- R-demo, part 2
- Exercise 2

Analysis of covariance

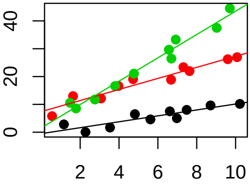
- Analysis of covariance (ANCOVA) is the combination of ANOVA and regression - we have both continuous and categorical explanatory variables (but still only a continuous response variable).
- The formulas in Chapter 15.3 are very cumbersome, so we will skip them completely and instead focus on understanding the models.

ANCOVA - a simple example

- A model with one continuous variable x and one categorical variable A:

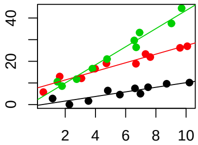
$$y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2), i = 1, \ldots, k, j = 1, \ldots, n_i$$

- Notice different slopes and intercepts depending on the level i of the factor.
- Different intercepts can be thought of as the main effect in A, while different slopes can be thought of as interaction between A and x.

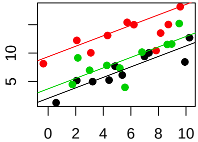


ANCOVA in R

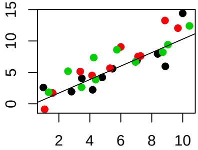
- Different slopes and intercepts:
  `lm(y~A*x)`
  $y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}$



- Different intercepts:
  `lm(y~A+x)`
  $y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij}$



- Same slopes and intercepts:
  `lm(y~x)`
  $y_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij}$



R

- R-demo, part 3
- Exercise 3