Linear normal models

Anton Rask Lundborg
arl@math.ku.dk

Copenhagen Causality Lab
Department of Mathematical Sciences

December 13, 2023

Lecture outline

- Origin of regression analysis.
- Model validation: Can the conclusions be trusted?
- Taxonomy of models with continuous (normal) response.
  - Usage and interpretation of interactions (effect modifications).
  - Validation, hypotheses, estimates.
  - ANOVA (ANalysis Of VAriance).
  - ANCOVA (ANalysis of COVAriance).
- Where is the effect?
  - Usage and interpretation of em-means (Estimated Marginal Means).
  - Multiple testing: FWER vs. FDR.
- Transformation of variables.

Origin of regression analysis

Charles Darwin, "On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life", published November 1859.

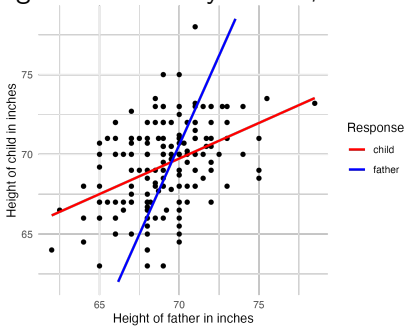
Probabilistic paradox: If there is variation in the offspring, that is, offspring inherit from properties from their parents but with added (independent) random variation, then we would expect the variation to grow with rate $\sqrt{\text{number of generations}}$?
— The variation is observed to be roughly constant across generations.

Sir Francis Galton (half-cousin to Charles Darwin) investigated the relation between the heights of children and their parents, and resolved the probabilistic paradox by a recognizing a phenomenon, which he called

Regression toward the mean.

Data example 1: Simple linear regression

Heights collected by Galton, cosine of angle between lines is correlation:



- Tall fathers get tall sons and tall sons have tall fathers.
- – but on average not as extreme as themselves.
- Instead, extremes tend to regress toward the mean in the other "generation".

Regression line for **Y=height of son** against **X=height of father**:

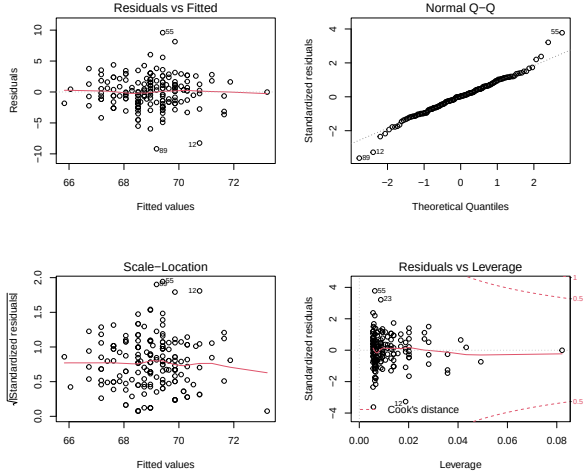
$$Y - \underbrace{69.00}_{\mu_Y} \approx \underbrace{0.4475}_{\rho_{YX} \cdot \frac{\sigma_Y}{\sigma_X}} \cdot (X - \underbrace{69.10}_{\mu_X})$$

Validation of model $Y_i = \alpha + \beta \cdot X_i + \epsilon_i$

- Model assumption: error terms ϵ_i are independent $\mathcal{N}(0, \sigma^2)$.
- Error terms ϵ_i are predicted by the residuals $r_i = Y_i - (\underbrace{\hat{\alpha} + \hat{\beta} \cdot X_i}_{\text{predicted value}})$.
- Validation of assumptions on the error terms:
  - Normal distribution: normal quantile plot of r_i.
  - Mean equals zero: Residual plot of r_i against predicted values.
  - Homogeneous variance: Based on standardized residuals $s_i = \frac{r_i}{\sqrt{\widehat{\text{var}}(r_i)}}$.

    Default method in R plots $\sqrt{|s_i|}$ against the predicted values.
  - Independence: Postulated by design (or add random effect, see Day 5).
- Observations having both large standardized residuals (potential outliers) and large degree of self-estimation (measured by the so-called leverage) may be critical.
  - These measures are combined in the so-called Cook's distance.

Validation plots: `plot(lm(son ~ father, data = height))`



You could also use `gof::cumres()` and, if possible, a Lack-of-Fit test.

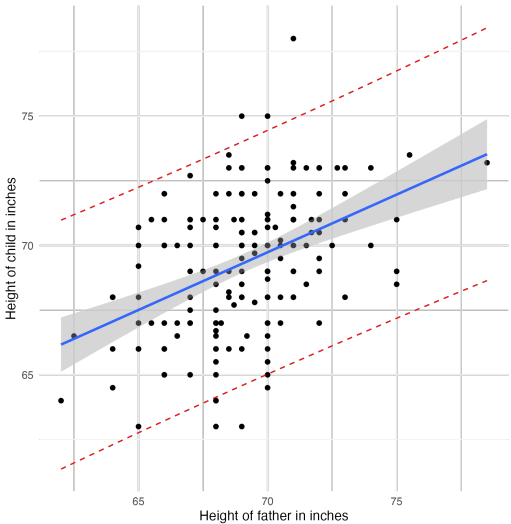
R code for the height example (I) – see `Galton.R`

```
# Data from HistData-package: select one random son per father
library(HistData)
data("GaltonFamilies")
height <- GaltonFamilies %>%
  filter(gender == "male") %>%
  group_by(family) %>%
  slice_sample(n = 1)

# Fit the linear regression
m1 <- lm(childHeight ~ father, data = height)

# Model validation
par(mfrow = c(2, 2))
plot(m1)
par(mfrow = c(1, 1))

# Is there an effect?
drop1(m1, test = "F")

# What is the effect?
cbind(coef(m1), confint(m1))
```

R code for the height example (II) – see `Galton.R`

```
x <- seq(min(height$father), max(height$father), length.out = 100)
pred_height <- cbind(
  data.frame(father = x),
  predict(m1,
    interval = "prediction",
    newdata = data.frame(father = x)
  )
)
ggplot(height, aes(x = father, y = childHeight)) +
  geom_point() +
  geom_smooth(method = "lm") +
  geom_line(aes(x = father, y = lwr),
    data = pred_height,
    col = "red", lty = 2
  ) +
  geom_line(aes(x = father, y = upr),
    data = pred_height,
    col = "red", lty = 2
  ) +
  scale_x_continuous(name = "Height of father in inches") +
  scale_y_continuous(name = "Height of child in inches") +
  coord_equal()
```

Observations and model plot



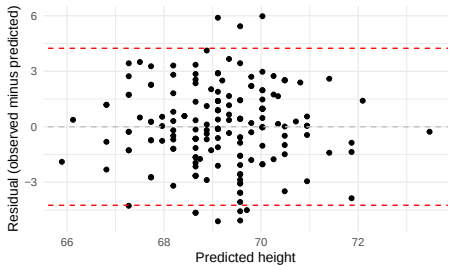
- The confidence interval (shaded area) gives the uncertainty on the estimate for the population mean.
- The prediction interval (area between red lines) gives the height prediction for an individual son knowing the height of his father.

(Root) mean squared error

A useful measure of the predictiveness of a model is given by its mean-squared error (MSE) or the square root of this quantity: RMSE. In R an estimate of the RMSE is usually reported as the "Residual Standard Error".

The RMSE is an estimate of the standard deviation σ of the residuals ϵ_i: $\hat{\sigma} = \text{RMSE}$.

When the errors are normally distributed with the same standard deviation σ, we therefore expect most model deviations to lie in an interval $(-1.96\hat{\sigma}, 1.96\hat{\sigma})$:

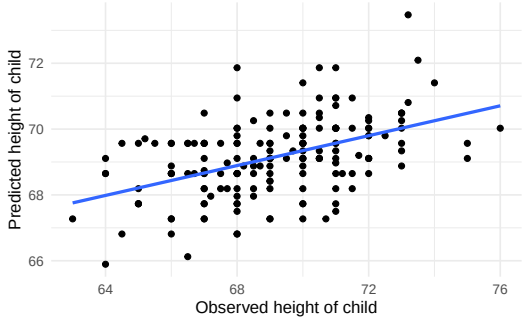


Explained variance (R^2)

Proportion of variability explained by the model (in Danish "forklaringsgrad"), available in output from `summary(m1)`:

$$R^2 = \text{correlation}(\text{observed}, \text{predicted})^2.$$

R^2 can also be found as the slope from linearly regressing the observed values on the predicted values (see plot).

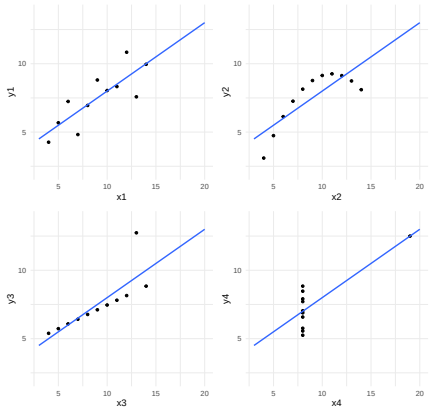


R^2 should not be used as an indicator for model quality

While we prefer large R^2 in a valid model, this is not a guarantee that a model is valid!

A famous example is Anscombe's quartet consisting of 4 datasets with the same R^2.

Similarly, a perfectly valid model may have small R^2 if the signal is weak.



Taxonomy – Continuous response with i.i.d. normally distributed errors

Name	**Explanatory variables**	**Standard interactions**
t-test	1 nominal with 2 levels	—
simple regression	1 numerical	—
multiple regression	2 or more numerical	none
1-way ANOVA	1 categorical	—
2-way ANOVA	2 categorical	factor_1:factor_2 (†)
N-way ANOVA	N categorical	all up to some degree (†)
ANCOVA	categorical + numerical	factor:covariate
Linear normal model	several categorical and numerical	as few as "possible"

(†) But only main effects of "block" factors.

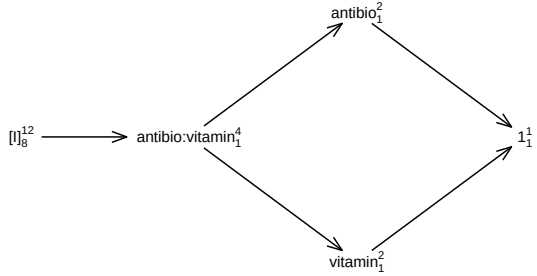
- Categorical explanatory variables are sometimes called factors. These may either be nominal or ordinal.
- Numerical explanatory variables are sometimes called covariates.

Data example 2: Two-way ANOVA (ANalysis Of VAriance)

Growth of rats ($N = 12$)

Variable	Type	Range	Usage
`antibio`	ordinal	$0 < 40$	fixed
`vitamin`	ordinal	$0 < 5$	fixed
`growth`	numerical	[1.000 ; 1.560]	response

	antibio	vitamin	growth
1	0	0	1.30
2	0	0	1.19
3	0	0	1.08
4	0	5	1.26
5	0	5	1.21
6	0	5	1.19
7	40	0	1.05
8	40	0	1.00
9	40	0	1.05
10	40	5	1.52
11	40	5	1.56
12	40	5	1.55

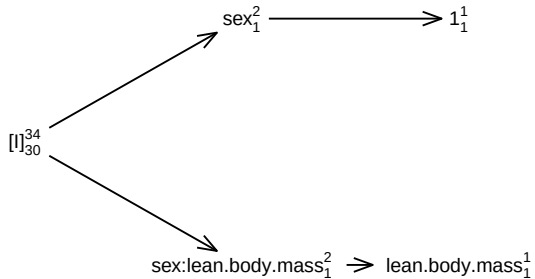


Data example 3: ANCOVA (Analysis of COVAriance)

Physical strength of men and women ($N = 34$)

Variable	Type	Range	Usage
`sex`	nominal	men, women	fixed
`lean.body.mass`	numerical	[28.00 ; 59.46]	fixed
`strength`	numerical	[56.42 ; 176.80]	response

```
      sex lean.body.mass  strength
1   woman        35.44753  95.79802
2   woman        30.82749  82.90753
3   woman        40.70181 111.84551
4   woman        31.99461  56.41715
5   woman        38.69096 105.31021
6   woman        44.04707 114.53475
...
30    man        59.46444 164.82921
31    man        50.19311 155.09919
32    man        48.38956 123.47673
33    man        54.71320 176.80209
34    man        58.17427 165.67084
```

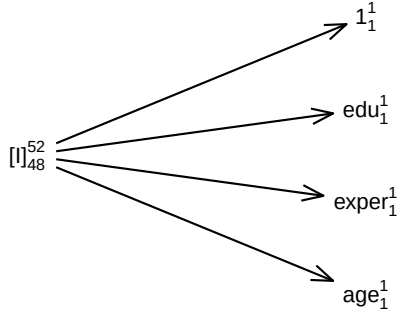


Data example 4: Multilinear regression

Worker wages in the USA in 1985 ($N = 52$)

Variable	Type	Range	Usage
edu	numerical	[2 ; 18]	fixed
exper	numerical	[0 ; 55]	fixed
age	numerical	[18 ; 68]	fixed
wage	numerical	[1.8 ; 26.3]	response

	edu	exper	age	wage
1	16	6	28	12.05
2	17	3	26	6.00
3	14	10	31	12.00
4	15	33	57	10.61
5	16	0	18	10.00
...				
48	9	34	50	5.75
49	15	11	32	7.67
50	16	6	28	11.79
51	12	33	54	6.10
52	17	25	49	23.25



Checkpoint

- Questions?
- After the break we discuss interactions and their interpretation

Time for a break!

Interactions and their interpretation

Interactions and their interpretation

'factor' and 'covariate' denote categorical and numerical variables, respectively.

- Interactions between factors consist of all combinations.
  - May be understood as factors themselves.
  - E.g., the interaction between `status` (levels: healthy, ill) and `sex` (levels: male, female) has 4 levels:

    (healthy male, healthy female, ill male, ill female)

- Interactions between factors and covariates allow slopes against the covariate to depend on the level of the factor.
  - E.g., an interaction between `sex` (levels: men, women) and `lean.body.mass` (in kg) allows the slope against the lean body mass to depend on the gender.

- Interactions between covariates is the same as multiplying their numerical values. Since the interpretation often becomes fuzzy, this is not used much, with the exception of polynomial regression.
  - E.g., a quadratic regression of `son` against `father` height:

$$\texttt{son}_i = \alpha + \beta \cdot \texttt{father}_i + \gamma \cdot \left(\texttt{father}_i\right)^2 + \texttt{error}_i$$

Data example 2: Two-way ANOVA

Growth of rats ($N = 12$)

- Two doses of antibiotics. Two doses of vitamin:

antibio	vitamin	
	0	5
0	1.30, 1.19, 1.08	1.26, 1.21, 1.19
40	1.05, 1.00, 1.05	1.52, 1.56, 1.55

- Statistical model with parameters $\alpha_{0,0}, \alpha_{0,5}, \alpha_{40,0}, \alpha_{40,5}$ and σ:

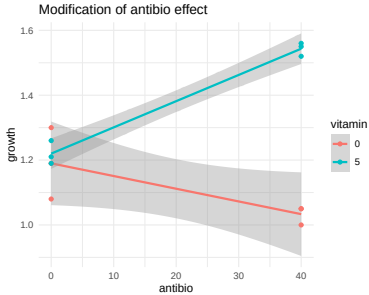
$$\text{growth}_i = \alpha(\text{antibio}_i, \text{vitamin}_i) + \epsilon_i, \qquad \epsilon_i\text{'s independent } \mathcal{N}(0, \sigma^2)$$

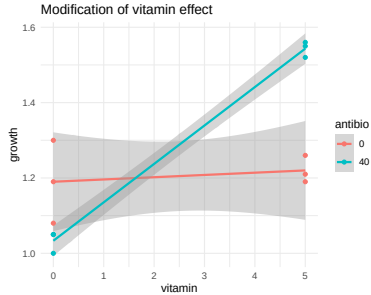
- May also be viewed as 1-way ANOVA:

treat = antibio:vitamin			
$(0,0)$	$(0,5)$	$(40,0)$	$(40,5)$
1.30, 1.19, 1.08	1.26, 1.21, 1.19	1.05, 1.00, 1.05	1.52, 1.56, 1.55

Interaction plot

Interactions may also be understood as and denoted "effect modifications"





R code for left panel (for right panel: `antibio` $\leftrightarrows$ `vitamin`):

```
ggplot(rats, aes(x = antibio, y = growth, col = factor(vitamin))) +
geom_smooth(method = "lm") +
geom_point() +
scale_color_discrete(name = "vitamin") +
ggtitle("Modification of antibio effect")
```

Is there an effect?

Is there an effect?

Model selection: (sometimes not necessary for simple setups)

- Backward-forward model selection (starting from the "largest" model) using the Akaike Information Criterion automatized by `step(.., direction="both")`.
- Selecting among all models using either Akaike Information Criterion or Bayes Information Criterion automatized by `MuMIn::dredge()`.

Hypothesis testing:

- May be done via `drop1(.., test="F")`, where the option `test="F"` ask for F-tests. However, occasionally it is necessary to use the `anova()` function as exemplified on slide 26.
- Preferably all tests should be of scientific interest.
- Beware of potential multicollinearity and non-orthogonal designs.
  - Weak forms of multicollinearity or non-orthogonality are rarely problematic.
  - Other issues like mediation and confounding briefly discussed on Day 6.

Which hypotheses are testable?

- In N-way ANOVAs an effect $\text{fac}_1 : \cdots : \text{fac}_k$ is only testable if it does not appear in higher order interactions. Since interactions can be understood as effects by themselves this rule may also be stated as:

Hierarchical principle

A main effect is only testable if it does not appear in an interaction.

`drop1()` in R obeys to the hierarchical principle.

- In R interactions are denoted by ":". The "*" is short syntax for

$$\texttt{antibio} * \texttt{vitamin} = \texttt{antibio} + \texttt{vitamin} + \texttt{antibio:vitamin}$$

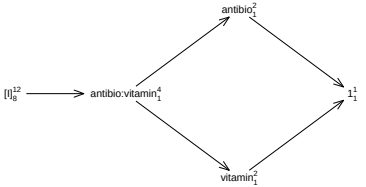
- The standard hypothesis on an interaction is that there is no interaction, but still lower order effects. This explains why the lower order terms are also included in the model, e.g.

$$\text{growth} \sim \text{antibio} + \text{vitamin} + \text{antibio:vitamin}$$

Design Diagrams – visualizing design structure

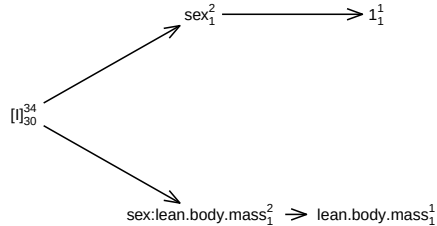
- Two special variables are always adjoined:
  - I identifies each observation $\Longrightarrow$ $[I]$ is the error term.
  - 1 collapses all observations $\Longrightarrow$ 1 is the intercept.
- $[\cdot]$ means that the variable has random effect.
- Superscripts = #levels, Subscripts = degrees of freedom.
- $A \to B$ means that "model" provided by variable B is part of the "model" provided by variable A.
  - For factors, we also say that A is nested in B.
- Model reduction: A systematic effect is "absorbed" by the nested random effect.
- Technicality: Minima between effects must be included. This refines the hierarchical principle.

```
plot(DD(~ antibio * vitamin, data = rats))
```



Data example 3: Possible hypotheses in an ANCOVA

Variable	Range	Usage
`sex`	men, women	fixed
`lean.body.mass`	[28.00 ; 59.46]	fixed
`strength`	[56.42 ; 176.80]	response



- Main effect of `sex` and interaction term `sex:lean.body.mass` are both testable.
  - `drop1()` misbehaves as it does not provide test on `sex`!
- The null hypothesis for the test on `sex` is that males and females with lean body mass = 0 kg on average have the same strength.
  - Question: Would that be biologically meaningful?
  - Question: How to test hypothesis that strength of men and women are equal for lean body mass = 40 kg, still allowing for different slopes?
  - We can subtract 40 from lean body mass before we fit the model!

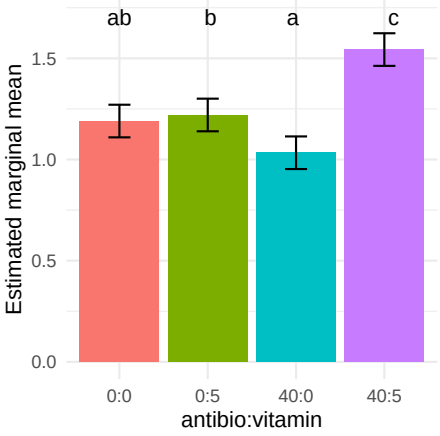
Checkpoint

- Questions?
- After the break we discuss the question: Where is the effect? This also leads to a short discussion of the multiple testing problem.

Time for a break!

Where is the effect?

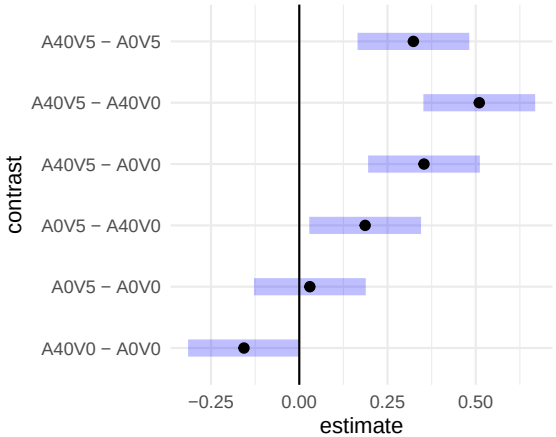
Data example 2: Growth of rats – Where is the effect? (I)



- There are $3 + 2 + 1 = 6$ pairwise comparisons between 4 treatments $\Longrightarrow$ there is a multiple comparisons problem.
- Here, this is solved by the Tukey method.
- Treatments not sharing a letter are significantly different.
- Confidence interval interpreted separately for each treatment.
- One potential visualization can be seen in plot.

```
multcomp::cld(emmeans(m1, ~ antibio * vitamin), Letters = letters)
```

Data example 2: Growth of rats – Where is the effect? (II)

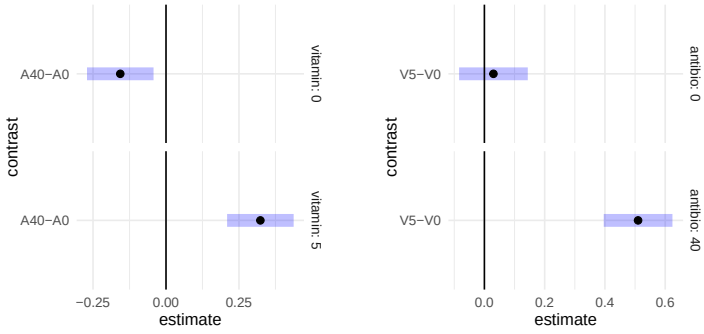


- There are 3+2+1=6 pairwise comparisons between 4 treatments $\Longrightarrow$ there is a multiple comparisons problem.
- Here this is solved by the Tukey method.
- Confidence intervals are enlarged to have simultaneous coverage.
- Alternative visualization can be seen in plot.

```
plot(pairs(emmeans(m1, ~ antibio * vitamin), reverse = TRUE),
     int.adjust = "tukey") + geom_vline(xintercept = 0)
```

Data example 2: Growth of rats – Where is the effect? (III)

We can also compare contrasts by fixing values of one factor (see two different examples below)



- In this formulation the multiple comparisons problem is less severe
  - In what sense?
  - Why is it problematic to consider both of the above plots?
- See `growth.R` for code to produce plots.

What is the effect?

Estimated marginal means – useful tool to answer: What is the effect?

- em-means = predicted values across levels of a factor, where the other explanatory variables in the model are set to neutral values.
- Possible neutral values for factors:
  - Levels are weighted equally.
  - Levels are weighted according to sample distribution.
  - For an ordinal factor we may use e.g. the first level.
- For covariates the sample mean is typically used as the neutral value.
- Pairwise comparisons of em-means in a post-hoc analysis identify differences between the levels of a factor in the model.
  - Should be corrected for multiple testing!
- In SAS there is a `BYLEVEL` option, which means that sample properties are taken within the strata corresponding to the factor under consideration.
  - This can also be done in R, see R script "strength.R".

Data example 3: Balanced vs. Marginal em-means (I) – see `strength.R`

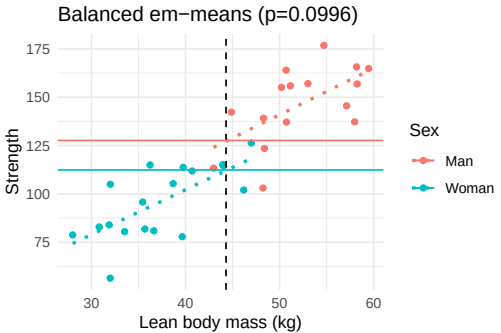
Are men physically stronger than women? Welch t-test $= 7.49$, df $= 30.83$, $p < 0.0001$. What does this difference really mean?

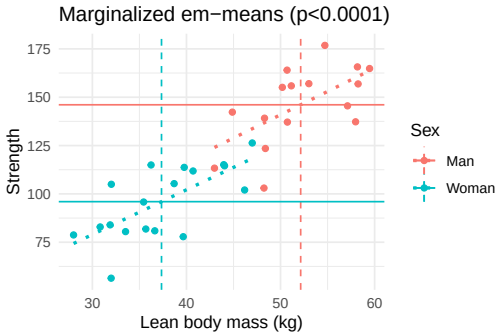
Variable	Type	Range	Usage
`sex`	nominal	men, women	fixed
`lean.body.mass`	numerical	[28.00 ; 59.46]	fixed
`strength`	numerical	[56.42 ; 176.80]	response

- We can compute the difference between the sexes for a fixed common value of lean body mass (the mean). These are balanced em-means (the default): t=1.697, df=31, $p = 0.0996$. Conclusion: Men and women do not differ in strength for the same lean body mass.
- We can compute the difference between the sexes for a grouped value of lean body mass (mean within the sexes). These are marginalized em-means: t=9.715, df=31, $p < 0.0001$. Conclusion: The average man and the average woman differ in strength.

Data example 3: Balanced vs. Marginal em-means (II) – see `strength.R`

We can visualize this; full lines indicate the computed em-mean, dashed lines indicate the value that is used for grouping and the dotted lines are the linear regression fits:





Some useful R functions for em-means

`emmeans()`	Constructs the `emmeans`-object (of class `emmGrid`). Option `type="response"` asks for back-transformations.
`confint()`	Computes confidence intervals.
`pairs()`	Makes pairwise comparisons (post hoc analysis).
`test()`	Performs tests. Option `joint=TRUE` makes the overall F-test.
`multcomp::cld()`	Constructs compact letter display. Requires `multcomp` and `multcompView` packages to be installed (but not necessarily loaded).
`plot()`	Visualizes the marginal means.

Checkpoint

- Questions
- After the break we discuss the question: Can the conclusions be trusted? That is, we recap how to do model validation. In this context we also discuss transformation of variables.
- We end with a digression: More on the multiple testing problem.

Time for a break!

Can the conclusions be trusted?

Can the conclusions be trusted?

In principle model validation is done as for regression analyses, although ANOVA's (only categorical explanatory variables) have special properties.

- **Normal residuals:** Normal Quantile Plot on standardized residuals.
- **Residuals have zero mean:** If we use a full factorial design, i.e. all interactions, then this assumption is automatically satisfied!
  - For non-saturated models we often have a Lack-of-Fit test.
  - We may also use the residual plot. Since the predicted values are only at "few" positions it can be better to use a Box-plot than a scatter plot.
- **Residuals have homogeneous variance:** Length of standardized residuals vs. predicted values.
  - It is also possible to apply various Goodness-of-Fit tests, e.g. Levene and Bartlett. This is most often done for 1-way ANOVA's.
- **Independence of residuals:** Postulated by design. See also Day 5.

Transforming data – useful when normality assumption fails

- **Standard transformations (for $y > 0$):**
  - `log` transform: $y \mapsto \log(y)$.
  - Square root transformation: $y \mapsto \sqrt{y}$.
  - The inverse transformation: $y \mapsto \frac{1}{y}$ (changes the order of the observations).
  - Box-Cox transformation with index λ:

$$y \mapsto \begin{cases} \frac{y^\lambda - 1}{\lambda \cdot \text{GeoMean}^\lambda} & \text{for } \lambda \neq 0 \\ \log(y) & \text{for } \lambda = 0 \end{cases}$$

    Note the order of the observations is changed when $\lambda < 0$.
    Some particular cases:

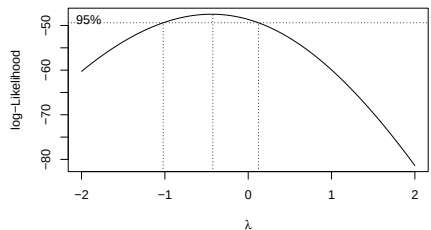
$\lambda = -1$	$\lambda = 0$	$\lambda = 0.33$	$\lambda = 0.5$	$\lambda = 1$
Inverse	`log`	*cubic root*	*square root*	*no transformation*

- **Arcus sine transformation (for $y \in [0, 1]$):**
  - $y \mapsto \arcsin(\sqrt{y})$.
  - May be appropriate when y measures the proportion of successes out of a number of trials.

Data example 4: Wages of female workers – see `wage.R`

```
m1 <- lm(wage ~ edu + exper + age, data wage_subset)
```

```
# R code for Box-Cox
# transformation

library(MASS)
bc <- boxcox(m1)
bc$x[which.max(bc$y)]
```



The last line gives $\hat{\lambda} = -0.4242424$, which we approximate as $\hat{\lambda} = -0.4$. This gives:

$$\text{wage}_i^{-0.4} = \alpha \cdot \text{edu}_i + \beta \cdot \text{exper}_i + \gamma \cdot \text{age}_i + \epsilon_i$$

$$\text{wage}_i = \left(\alpha \cdot \text{edu}_i + \beta \cdot \text{exper}_i + \gamma \cdot \text{age}_i + \epsilon_i\right)^{-2.5}$$

This model is valid, but the interpretation is awkward! Better to use log(wage), possibly also with log transformation of explanatory variables.

More on the multiple testing problem

The multiple testing problem

	Null hypothesis is		
	true	false	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

- Suppose you perform $m_0 = 100$ independent valid hypothesis tests of true null hypotheses ("no effect"). Then the probability that you find $V \geq 1$ rejections of true null hypotheses when rejecting at 0.05 is

$$1 - (1 - 0.05)^{100} = 0.9941.$$

  Thus, the family-wise Type I error (FWER) is 99.4%, which is far from the 5% significance level. We should correct for this misbehavior.
- The false discovery rate (FDR) is defined as V/R. An alternative is to control this quantity.

Correcting for multiple comparisons – using several p-values simultaneously

- To ensure that the FWER ≤ 0.05 we must reduce the significance level. Another way of formulating this is to adjust the p-values.

- The adjusted p-value (adjusted for FWER) of an effect is below α means:

  If the significance level is such that the FWER is (at most) α,
  then this effect is significant.

  For this to make sense we, of course, also need to specify which family of tests we are considering.

- Similar interpretation when adjusting for FDR. In the literature the FDR-adjusted p-values are sometimes referred to as q-values.

Some methods for correcting for multiple comparisons

R functions: `p.adjust()`. R packages: `emmeans`, `multcomp`, `hommel`, `TMTI`

- FWER methods that can be applied to any set of p-values:
  - **Bonferonni:** Makes no assumptions, and may be done "by hand". However, it is notoriously conservative (too few new discoveries).
  - **Holm (1979):** Makes no assumptions, and may be much less conservative. Always preferred over the Bonferroni correction.
  - **Hochberg (1988)** and **Hommel (1988):** May be even less conservative, but are only valid for non-negatively associated p-values.

- FDR methods that can be applied to any set of p-values:
  - **Benjamini & Hochberg (1995):** Valid if the so-called PRDS condition holds. In particular, if the p-values are independent.
  - **Benjamini & Yekutieli (2001):** Makes no assumptions.

- FWER methods using joint distribution of p-values:
  - **Tukey:** Often used together with Tukey grouping, cf. slide 29.
  - R-package multcomp developed by **Hothorn, Bretz, Westfall (2008)**. Preferably accessed via emmeans package by **Lenth et al. (2018)**.

Beyond FWER and FDR

Example from Goeman & Solari (2011): Multiple Testing for Explorative Research:

A PhD student in organic chemistry has synthesized 5 anti-cancer candidate drugs labelled A to E. Initial mice studies of the drug effects have been performed to evaluate the marginal null hypotheses:

$$H_i: \text{Drug } i \text{ has no effect.}$$

As a result the following p-values were obtained:

Drug	p-value
A	0.055
B	0.068
C	0.075
D	0.090
E	0.210

Question: Should the PhD student be disappointed?

Avoid Type II errors: Ask precise questions!

Q1: Are any of the marginal null hypotheses false?

Q2: How many of the marginal null hypotheses are false?

Q3: Which of the marginal null hypotheses are false?

Avoid Type II errors: Ask precise questions!

Q1: Are any of the marginal null hypotheses false?

Q2: How many of the marginal null hypotheses are false?

Q3: Which of the marginal null hypotheses are false?

Q1 answered by testing the global null hypothesis.

Global null hypothesis: None of the 5 drugs have an effect.

Too Many Too Improbable (TMTI) test, one test from the large zoo of combination tests, rejects this: $p = 0.0014$

Strong evidence that at least 1 of the drugs have an effect!

Avoid Type II errors: Ask precise questions!

Q1: Are any of the marginal null hypotheses false?

Q2: How many of the marginal null hypotheses are false?

Q3: Which of the marginal null hypotheses are false?

Q2 answered by confidence set for the number of false null hypotheses (i.e. number of effective drugs).

Top-down procedure using the TMTI-test gives the following 95pct confidence sets:

- $\{3, 4, 5\}$ among all 5 drugs
- $\{3, 4\}$ among the 4 drugs with lowest p-values

This gives ground for making new confirmatory experiments!

Avoid Type II errors: Ask precise questions!

Q1: Are any of the marginal null hypotheses false?

Q2: How many of the marginal null hypotheses are false?

Q3: Which of the marginal null hypotheses are false?

Q3 answered by Family-Wise-Error-Rate (FWER) adjusted p-values for the marginal null hypotheses.

Close testing procedure using TMTI-test gives

$$\text{adjusted } p \geq 0.079$$

Thus, none of the drugs can be pinpointed as having an effect at the usual 5% significance level.

Summary (I): Validation of linear normal models

- **Do residuals have mean=0?** Plot of residuals vs predicted values.
- **Do residuals have same variance?** Plot of standardized residuals vs predicted values.
  - R plots $\sqrt{|s_i|}$ against $\hat{y}_i$ but the interpretation is the same.
  - Often occurs that variance increases with predicted values ('trumpet shape' in residual plots). Sometimes solved by a log-transformation.
- **Are residuals normal distributed?** Normal quantile plot of standardized residuals.
  - Banana shape indicates need for log-transformation.
- **Are residuals independent?** Not validated formally. Use knowledge about design of experiment. See also course Day 5.
- **Are there any outliers?** Plot of standardized residuals vs leverages. Critical lines in terms of Cook's distances (D=0.5, D=1.0).
  - Generally it is not advisable to remove observations!
  - For robustness of the results excluding some observations may be tried.

Summary (II): Building statistical models

- Both the response and the explanatory variables may be transformed.
- Models need to be statistically valid if *p*-values and confidence intervals are to be trusted.
- We also like models with simple/sensible interpretations.
- If focus is on prediction only, then it is less/not important that the model is valid.
- If you have several explanatory variables, then there is the possibility of multi collinearity. This phenomenon can mess up interpretations and hypothesis tests. See Exercise 4.1 for an example of this.