Linear normal models

(Regression analysis & Analysis of Variance)

Bo Markussen
bomar@math.ku.dk

Department of Mathematical Sciences

December 9, 2020

Lecture outline

- Origin of regression analysis.
- Model validation: Can the conclusions be trusted?
- Taxonomy of models with continuous (normal) response.
  - Usage and interpretations 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.

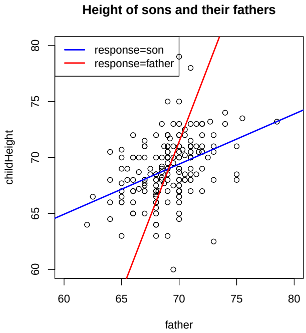
Probabilistic paradox: If there is variation in the off-spring, how can it be that the variation doesn't grow with rate $\sqrt{\text{number of generations}}$?
— The variation is most often more of less constant across generations.

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

Regression toward the mean.

Data example 1: Simple linear regression

Heights of sons and fathers collected by Galton



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

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

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

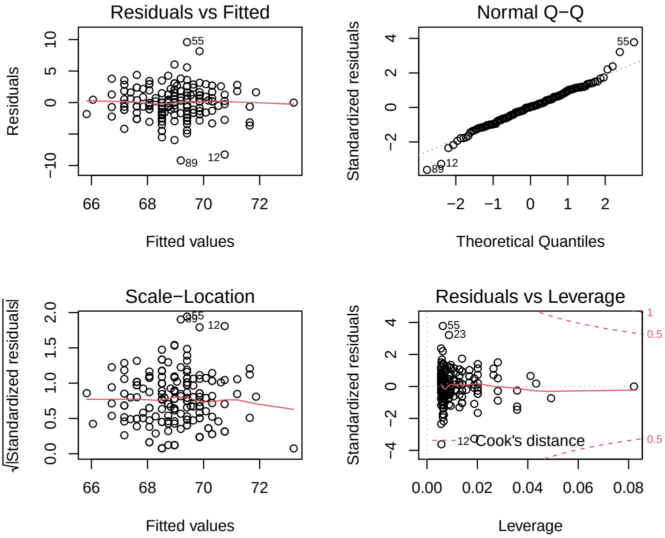
Validation of model $Y_i = \alpha + \beta * X_i + \epsilon_i$

- Model assumption: error terms ϵ_i are independent $\mathcal{N}(0, \sigma^2)$.
- Error term ϵ_i is predicted by the residual $r_i = Y_i - (\underbrace{\hat{\alpha} + \hat{\beta} * X_i}_{\text{predicted value}})$.
- 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, but see also Day 5.
- Observations having both large standardized residuals (potential outliers) as well as 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))`

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



R code for the height example (I)

See script “Galton.R” for more fancy R code

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

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

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

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

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

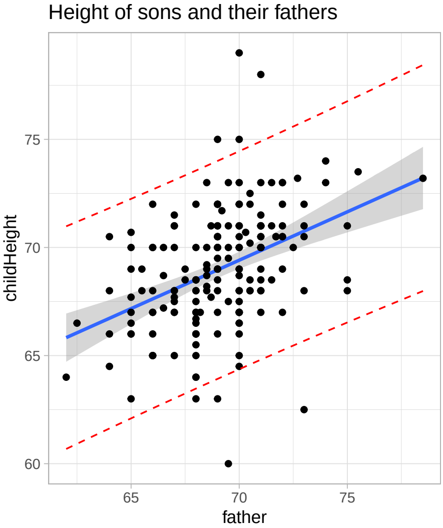
R code for the height example (II)

```
# Plot of observations together with model predictions
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) +
 coord_equal()
```

- The confidence interval gives the uncertainty on the estimate for the population mean.
- The prediction interval gives the height prediction for an individual son knowing the height of his father.

Observations and Model plot

Data points with confidence and prediction intervals



Actual by Predicted Plot

Graphical visualization of the R-squared value

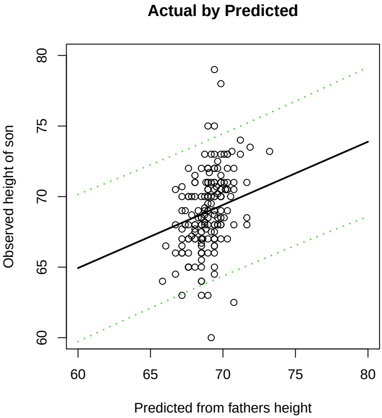
- Proportion of variability explained by the model (in Danish "forklaringsgrad"), available i output from `summary(m1)`:

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

- The closer R^2 is to 1 the better, but R^2 does not quantify whether the model is good or bad!

- RMSE (Root Mean Square Error) of the error terms ϵ_i, called "Residual Standard Error" in R, estimates the variation around the systematic effect in the population:

$$\hat{\sigma} = \text{RMSE}$$



Questions?

- And then a break.
- After the break we discuss the taxonomy of experimental designs, with special emphasis on linear normal models.

Taxonomy

Continuous response with i.id. normally distributed errors

Name	**Explanatory variables**	**Standard interactions**
T-test	1 nominal with 2 levels	—
simple regression	1 continuous	—
multiple regression	2 or more continuous	none
1-way ANOVA	1 categorical	—
2-way ANOVA	2 categorical	$\text{factor}_1\text{:factor}_2$ (†)
N-way ANOVA	N categorical	all up to some degree (†)
ANCOVA	categorical + continuous	factor:covariate
Linear normal model	several categorical and continuous	as few as "possible"

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

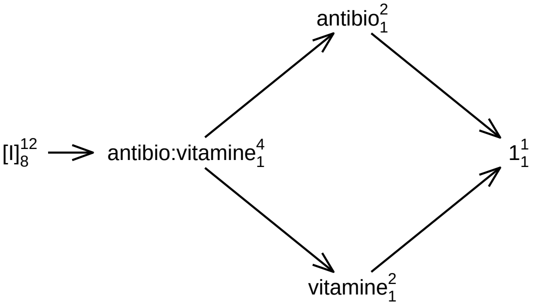
- Categorical explanatory variables are sometimes called factors. These may either be nominal or ordinal.
- Continuous explanatory variables are sometimes called covariates.

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	continuous	[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
```

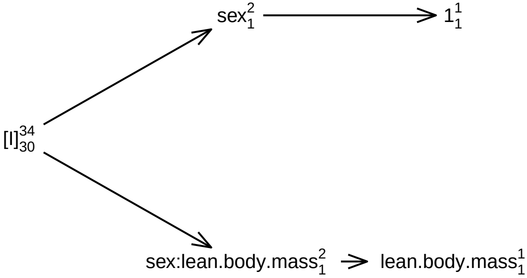


Data example 3: ANCOVA (Analysis of COVAriance)

Physical strength of men and women (N=34)

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

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

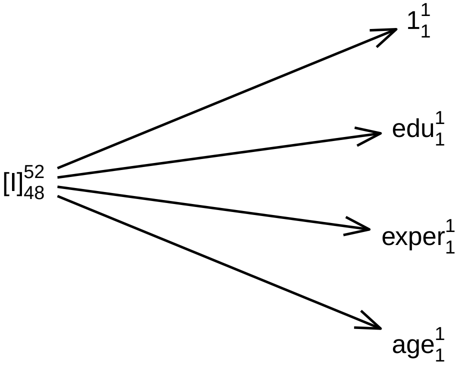


Data example 4: Multilinear regression

Worker wages in USA in 1985 (N=532)

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

```
      edu exper age wage
  1     8    21  36  5.1
  2     9    42  59  5.0
  3    12     1  19  6.7
  4    12     4  22  4.0
  5    12    17  36  7.5
  6    13     9  28 13.1
...
528    18     5  29 11.4
529    12    33  54  6.1
530    17    25  49 23.2
531    12    13  31 19.9
532    16    33  56 15.4
```



Interactions and interpretations

'factor' and 'covariate' is short for categorical and continuous 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:

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

Data example 2: Two-way ANOVA

Growth of rats (N=12)

- Two doses of antibiotics. Two doses of vitamin:

	vitamin	
antibio	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:

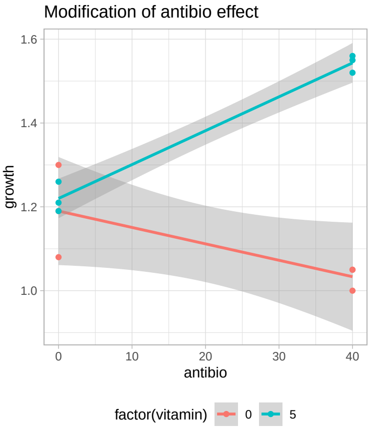
$$\text{growth}_i = \alpha(\text{antibio}_i, \text{vitamin}_i) + \epsilon_i, \quad \epsilon_i\text{'s independent } \mathcal{N}(0, \sigma^2)$$

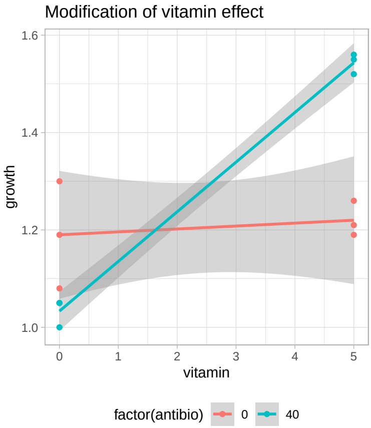
- May also be viewed as 1-way ANOVA:

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 called **“effect modifications”**





R code for left panel (for right panel: `antibio` ⇆ `vitamin`):

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

Is there an effect?

Model selection: (not necessarily needed!)

- 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 22.
- Preferably all tests should be of scientific interest.
- Beware of potential multi collinearity and non-balanced designs.
  - Non-orthogonality usually doesn't pose practicable problems.
  - Other issues like mediation and confounding briefly discussed on Day 6.

Which hypotheses are testable?

- In N-way ANOVA's an effect $\text{fac}_1\text{:}\ldots\text{:fac}_k$ 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

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

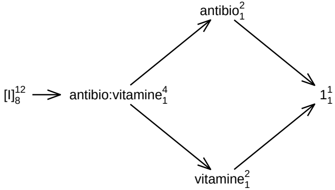
- The standard hypothesis on an interaction is that there is no interaction, but still lower order effects. This explains why the lower order terms also are included in the model, e.g.

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

Design Diagrams

- 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))
```



$[I]^{12}_{8}$

antibio:vitamine$^{4}_{1}$

antibio$^{2}_{1}$

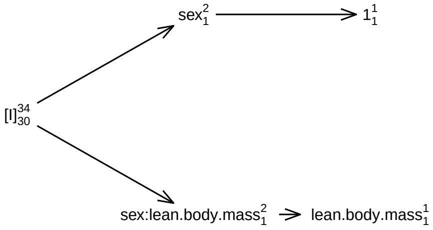
vitamine$^{2}_{1}$

1^{1}_{1}

Data example 3: Possible hypothesis in an ANCOVA

lm(strength∼sex*lean.body.mass,data=strength)

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



$[I]^{34}_{30}$

sex^{2}_{1}

1^{1}_{1}

$\text{sex:lean.body.mass}^{2}_{1}$

$\text{lean.body.mass}^{1}_{1}$

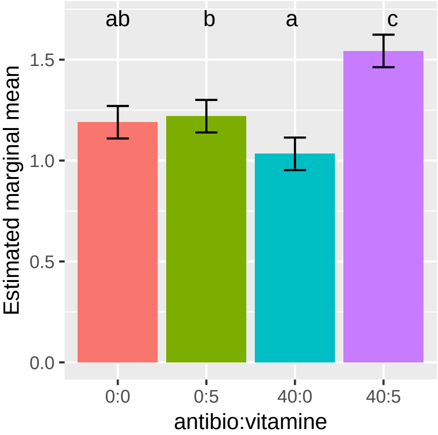
- Main effect of **sex** and interaction term **sex:lean.body.mass** are both testable.
  - `drop1()` misbehaves as it doesn't provide test on **sex**!
- The null hypothesis for the test on **sex** is that males and females with lean body mass = 0 kg on average have the same strength. Right?
  - Would that be biologically meaningful?
  - How to test hypothesis that strength of men and women are equal for lean body mass = 40 kg, still allowing for different slopes?

Questions?

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

Where is the effect?

May be visualized in different ways, here's one of them



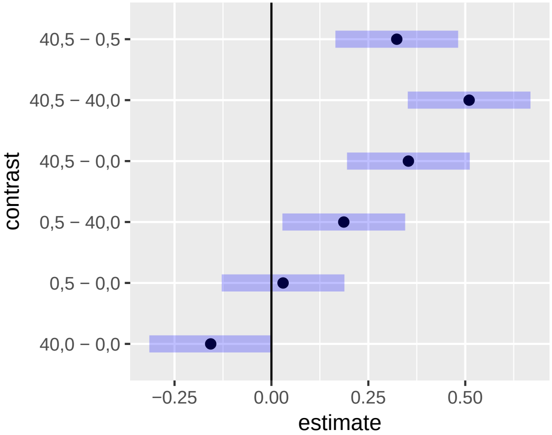
- Data example 2: Growth of rats.
- Treatments not sharing a letter are significantly different.
- There are $3+2+1=6$ pairwise comparisons between 4 treatments $\Longrightarrow$ there is a multiple comparisons problem.
- Here this is solved by the Tukey method.
- Confidence interval interpreted separately for each treatment.

```
CLD(emmeans(m1,~antibio*vitamin),Letters=letters)
```

For `ggplot()`-code see R script '`growth.R`'.

Where is the effect?

May be visualized in different ways, here's another possibility

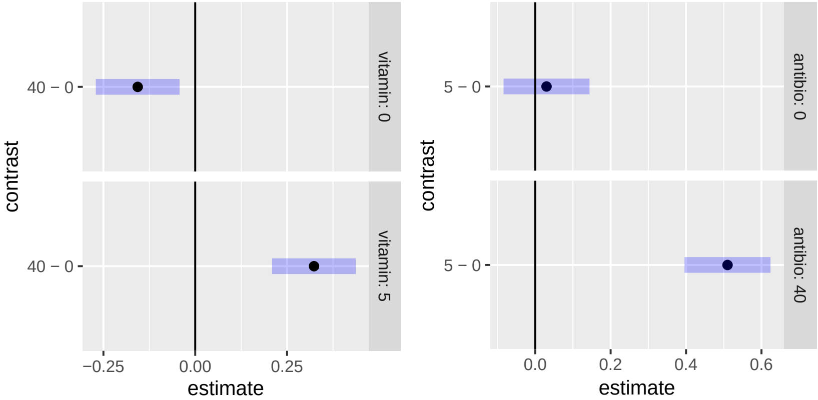


- Data example 2: Growth of rats.
- There are $3+2+1=6$ pairwise comparisons between 4 treatments $\Longrightarrow$ there is a multiple comparisons problem.
- Here this is solved by the Tukey method.
- Confident intervals are enlarged to have simultaneous coverage.

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

Where is the effect?

May be phrased in many ways, here's a third possibility



- In this formulation there is not a multiple comparisons problem!
  - Why / in what sense?

The multiple comparisons 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 $V \geq 1$ significances on the 5% level 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 misbehaviour.

- The false discovery rate (FDR) is defined as V/R. An alternative is to control this quantity.

Correcting for multiple comparisons

When you want to use more 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.

- That the adjusted p-value (for FWER) for 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 sometimes are referred to as q-values.

Some methods for correcting for multiple comparisons

R functions: p.adjust(), emmeans()

- FWER methods working on any collection of p-values:
  - **Bonferonni:** Makes no assumptions, and may be done "by hand". However, it is notoriously conservative ($\sim$ too few new discoveries).
  - **Holm (1979):** Makes no assumptions, and may be much less conservative. Hence always preferred to the Bonferroni correction.
  - **Hochberg (1988)** and **Hommel (1988):** May be even less conservative, but assume non-negatively associated p-values.

- FDR methods working on any collection 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 24.
  - R-package multcomp developed by **Hothorn, Bretz, Westfall (2008)**. Preferably accessed via emmeans package by **Lenth et. al. (2018)**.

Estimated marginal means

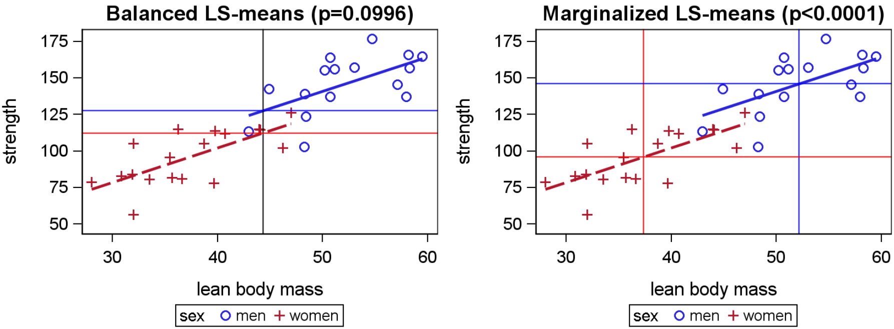
An often used answer to the question: What is the effect?

- 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 e.g. use 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

Are men physically stronger than women? Welch T-test $= 7.49$, df $= 30.83$, $p < 0.0001$

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



- Balanced em-means: t=1.697, df=31, p=0.0996
- Marginalized em-means: t=9.715, df=31, $p < 0.0001$

Questions?

- And then a break!
- 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.

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 only are 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

Often a solution 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: $y \mapsto \frac{1}{y}$.
    This transformation changes the order of the observations.
  - Box-Cox transformation with index λ:

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

    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*

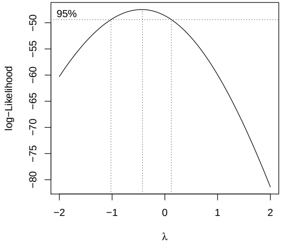
- **Arcus sine transformation (for $y \in [0, 1]$):**
  - $y \mapsto \arcsin(\sqrt{y})$.
  - May be appropriate when y measures the *proportion of successes out of a number of trials*.

Data example 4: Wages of female workers

Assume we have fitted: `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 might approximate as $\hat{\lambda} = -0.4$. This gives the model:

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

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

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

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.
    - Actually R plots $\sqrt{|s_i|}$ against $\hat{y}_i$. But the interpretation is the same.
    - Often seen is that variance increases with predicted values. May often be solved by log-transformation.
    - This is also seen as "trumpet shape" in the residual plot.
- **Are residuals normal distributed?** Normal quantile plot of standardized residuals.
    - Banana shape indicates need for log-transformation.
- **Are residuals independent?** Not validated formally. Use instead 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!
    - But 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.