Faculty of Life Sciences

# Model validation and prediction

**Ib Skovgaard and Claus Ekstrøm**
E-mail: ekstrom@life.ku.dk



## Program

- Review of statistical models/examples
- Model validation
- Prediction





## Exercise 7.1: age and percent body fat

Linear regression

| Age | 23 | 28 | 38 | 44 | 50 | 53 | 57 | 59 | 60 |
|---|---|---|---|---|---|---|---|---|---|
| Fat % | 19.2 | 16.6 | 32.5 | 29.1 | 32.8 | 42.0 | 32.0 | 34.6 | 40.5 |

Statistical model:

$$\text{fatpct}_i = \alpha + \beta \cdot \text{age}_i + e_i, \quad e_1, \dots, e_n \sim N(0, \sigma^2) \text{ independent}$$

R: `model1 <- lm(fatpct~age)`





## Exercise 6.7: weight gain in chicken

One-way ANOVA

| Feed type | Weight gain | | | | |
|---|---|---|---|---|---|
| 1 | 55 | 49 | 42 | 21 | 52 |
| 2 | 61 | 112 | 30 | 89 | 63 |
| 3 | 42 | 97 | 81 | 95 | 92 |
| 4 | 169 | 137 | 169 | 85 | 154 |

Statistical model:

$$\text{gain}_i = \alpha_{\text{feed}(i)} + e_i, \quad e_1, \dots, e_n \sim N(0, \sigma^2) \text{ independent}$$

R: `model2 <- lm(gain~factor(feed))`
Note: `factor(feed)`!

# Exercise 6.1: Gestation times for horses

A single sample

Gestation times for 13 horses:

339 339 339 340 341 340 343 348 341 346 342 339 337

Statistical model:

$$\text{gest}_i \sim N(\mu, \sigma^2) \text{ independent}$$

The model may also be written

$$\text{gest}_i = \mu + e_i, \quad e_1, \dots, e_n \sim N(0, \sigma^2) \text{ independent}$$

R: `model3 <- lm(gest~1)`





# All of the models

$$\text{fatpct}_i = \alpha + \beta \cdot \text{age}_i + e_i, \quad e_1, \dots, e_n \sim N(0, \sigma^2) \text{ independent}$$

$$\text{gain}_i = \alpha_{\text{feed}(i)} + e_i, \quad e_1, \dots, e_n \sim N(0, \sigma^2) \text{ independent}$$

$$\text{gest}_i = \mu + e_i, \quad e_1, \dots, e_n \sim N(0, \sigma^2) \text{ independent}$$

Types of variables:

- Response variable, $y$: fatpct, gain, gest
- Explanatory variable: age (quantitative), feed (factor/categorical)

Assumptions:

- All $e_i$ (or $y_i$) are normally distributed
- The mean of $y_i$ may depend on an explanatory variable
- All $e_i$ (or $y_i$) have the same standard deviation
- $e_1, \dots, e_n$ (or $y_1, \dots, y_n$) are independent





# Summary 1: Statistical models and inference

The models for linear regression, one-way ANOVA and a single sample are actually much alike!

Therefore the statistical inference is also alike in the three types of model ($p$ is the number of parameters in the mean):

- mean value parameters are estimated by LS
- the standard deviation $\sigma$ is estimated the same way
- Confidence intervals: $\text{estimate} \pm t_{0.975,n-p} \cdot \text{SE}(\text{estimate})$
- Tests of hypotheses are $t$-tests or $F$-tests

The models can be extended to include more explanatory variables — quantitative variables and/or factors. Linear normal models:

$$y_i = \text{mean}_i + e_i \quad e_1, \dots, e_n \sim N(0, \sigma^2) \text{ independent}$$





# Residuals

Expected value or fitted value or predicted value, $\hat{y}_i$:

- $\hat{y}_i = \widehat{\text{fatpct}}_i = \hat{\alpha} + \hat{\beta} \cdot x_i$
- $\hat{y}_i = \widehat{\text{gain}}_i = \hat{\alpha}_{g(i)}$
- $\hat{y}_i = \widehat{\text{gest}}_i = \hat{\mu}$

Residuals:

$$r_i = y_i - \hat{y}_i = \text{observed} - \text{fitted}$$

The residuals are our best guess of the $e$'s! Hence

- $\hat{\sigma} = s = \sqrt{\frac{1}{n-p} \sum_{i=1}^n r_i^2}$ where $p$ is the number of parameters in the model for the mean (2, $k$, 1)
- residuals are used for model validation!

The residuals may be standardized to have standard deviation 1: $\tilde{r}_i = r_i / \text{SE}(r_i)$.

# Residuals in R

```
> model1 <- lm(fatpct~age)       ## Linear regression
> fit1 <- fitted(model1)         ## Fitted values
> res1 <- residuals(model1)      ## Raw residuals
> stdres1 <- rstandard(model1)   ## Standard. residuals

> model2 <- lm(gain~factor(feed))
> fit2 <- fitted(model2)
> res2 <- residuals(model2)
> stdres2 <- rstandard(model2)
```





# Model validation: why?

Model validation aims to check if the model assumptions are reasonable for our data.

Why do we need model validation?

- If the assumptions are ok, then the 95%-CI contains the population value with 95% probability, and the $p$-values are correct.

  We may trust our results!
- If the assumptions are not ok, then we do not know if the results are trustworthy!

The assumptions about $e_1, \ldots, e_n$ are checked through the standardized residuals $\tilde{r}_1, \ldots, \tilde{r}_n$.





# Model validation: how?

Assumptions:

1. $e_i$ is normally distributed
2. $e_i$ have mean 0 — for any values of the explanatory variables
3. $e_i$ have the same standard deviation — for any values of the explanatory variables
4. $e_1, \ldots, e_n$ are independent

How?

- Independence is rather a matter of experimental design
- Check the first three assumptions about $e_1, \ldots, e_n$ using the standardized residuals $\tilde{r}_1, \ldots, \tilde{r}_n$

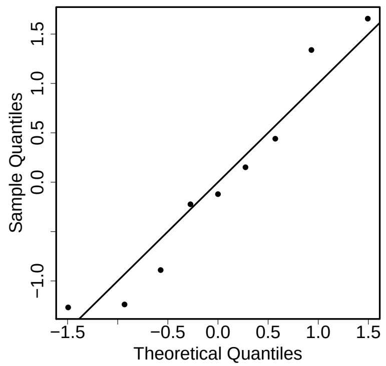
**Thorvald Nicolai Thiele, 1838–1910**

Man skal tegne før man kan regne





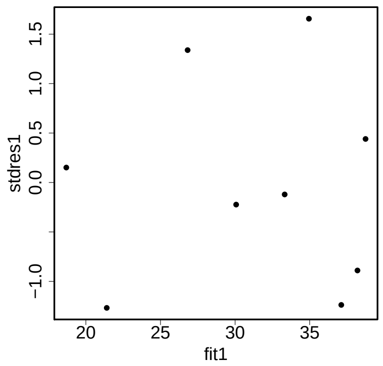
# Age and body fat: Assumption 1




Assumption 1. $e_i$ is normally distributed:

- QQ-plot of $\tilde{r}_1, \ldots, \tilde{r}_n$
- Compare with a straight line, with intercept 0 and slope 1

# Age and body fat: Assumptions 2 and 3




Assumptions 2. and 3. $e_i$ have mean 0 and same standard deviation

- Residual plot, $\tilde{r}_i$ against $\hat{y}_i$
- No systematic pattern in the vertical variation
- Are there outliers, that is, extreme observations?

  ($\tilde{r}_1, \ldots, \tilde{r}_n$ have standard deviation 1)

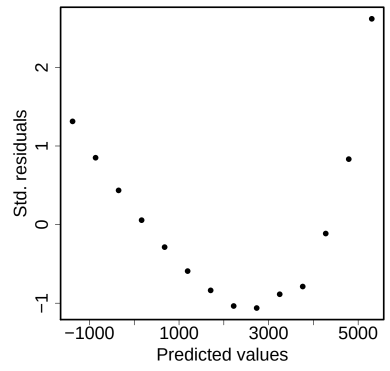
If you have installed `isdals` you can use `residualplot(model)`



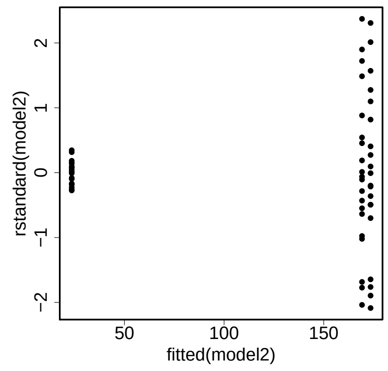


# Residual plot for the other two data sets

Duckweed (Ex. 2.4 and 6.2)




Pillbugs (case 2)

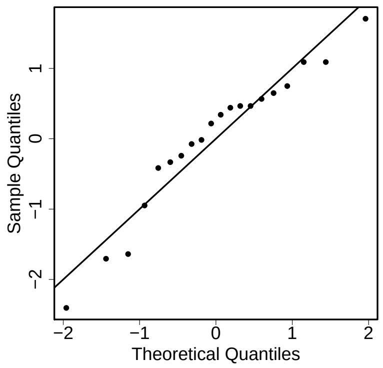




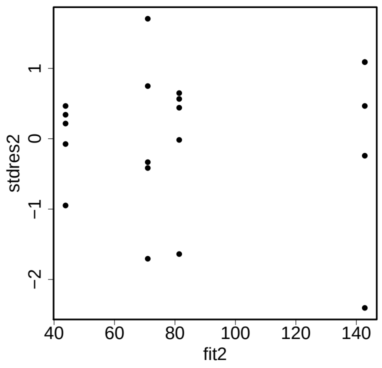



# Residual analysis for the chicken data

QQ-plot




Residualplot








# Assumptions 4: independence

Usually not to be verified from the data (residuals), but is rather a matter of the experimental design.

Subsets of the observations may not “share information”.

If an observation is larger than expected, is it likely to be mirrored by particular other observations?

Examples of dependent data:

- Data from the same fields, the same persons, the same plants, etc.
- Data from siblings, litters, ...

Some time dependence may be useful — but then dependence should be part of the model.

## Summary 2: Model validation

It is very important to validate the model, otherwise we cannot trust confidence intervals, $p$-values, etc.

Model validation is primarily graphical, comprising the residual plot and the QQ-plot for standardized residuals. The residual plot, in particular, is essential!

- In the residual plot the vertical variation should be random. Should not be systematically different "from left to right".
- Very large standardized residuals correspond to extreme observations or outliers. They should be examined further.
- In the QQ-plot the points should be scattered at random around a straight line.
- Is it reasonable to assume independence?





## Age and body fat: prediction

Person, aged 52. Expected percent body fat is $\alpha + \beta \cdot 52$, the estimate of which is

$$\hat{y} = \hat{\alpha} + \hat{\beta} \cdot 52 = 6.2254 + 0.5419 \cdot 52 = 34.4043$$

with the standard error (page 93–94)

$$\mathrm{SE}(\hat{y}_0) = s\sqrt{\frac{1}{n} + \frac{(52-\bar{x})^2}{\mathrm{SS}_x}} = 4.61 \cdot \sqrt{0.1374} = 1.709$$

95%-confidence interval:

$$34.4043 \pm 2.365 \cdot 1.709 = (30.36, 38.45)$$

A 52 year old person has 28 percent body fat. Why can we not use the confidence interval to decide whether this is unusual?





## Age and body fat: prediction

The confidence interval narrows down the expected value — not a single new observation.

The confidence interval takes into account only the estimation error — not the individual variations.

95%-prediction interval:

$$\hat{y} \pm t_{0.975,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\mathrm{SS}_x}}$$

Person, aged 52:

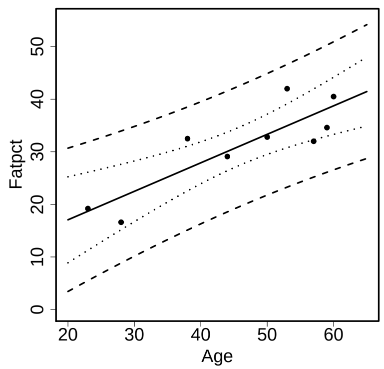
$$\hat{y} \pm 2.365 \cdot 4.61 \cdot \sqrt{1 + 0.1374} = (22.78, 46.03)$$

Is the fat percentage 28 unusual for a 52 year old person?





## Confidence interval vs. prediction interval




CI vs. PI:

- Interpretation: expected value or new observation?
- PI always wider than CI
- CI may become very narrow if $n$ is large, PI stays about the same

Prediction in a one-way ANOVA and in a single sample: see Section 7.2.3!

## Summary 3: prediction

Prediction is about "predicting" new observations.

- A 95%-prediction interval contains with probability 95% a new observation for a given value of an explanatory variable.
- A prediction interval is always wider than the corresponding confidence interval because it also takes variation between observations into account
- Are not reduced in size by increasing $n$.





## Summary 1–3

The models for linear regression, one-way ANOVA and a single sample are "same soup".

- Same assumptions — except the specification of the mean
- Two types of explanatory variables: quantitative and factors
- Statistical inference "the same": LS-estimation, confidence intervals, test, prediction, model validation
- More explanatory variables may be used — still the same type of model and methods for statistical inference





## Lecture summary: main points

- Multiple comparisons — why is a problem, and what can we do about it?
- Model validation
  - Analysis of standardized residuals — what should we look for?
- Prediction. Meaning of confidence- and prediction intervals. Computation.