SS2: Foundations in Statistics

SS2.1: Regression Analysis

Professor Caroline Brophy

School of Computer Science and Statistics
Trinity College Dublin

LegumeLegacy Event 2, Dublin, Ireland

November 2023

LegumeLegacy funding

This project has received funding from the European Union's Horizon 2021 doctoral network programme under the Marie Skłodowska-Curie grant agreement No. 101072579.

Summary of SS2: Foundations in Statistics II

- SS2.1: Regression analysis (this session)
  - Simple linear regression and One-way ANOVA models
  - Two-way ANOVA models
  - Multiple regression models
- SS2.2: Challenges with data and regression analysis
- SS2.3: The design and implementation of experiments and the management of data

What can you expect to learn from taking the SS2 module?

By the end of SS2, you should be able to:

- Fit and interpret a range of regression and ANOVA models.
- Translate and communicate the outcomes from a statistical model.
- Identify the limitations of standard regression and ANOVA modelling techniques, and situations where model assumptions may need to be altered.
- Design, implement and collect data from experiments in ways that are statistically sound.

Any difference in expected outcomes for Statistics PhD students and biological sciences PhD students? Yes!

What can you expect to learn from taking the SS2 module?

By the end of SS2, Statistics PhD students should additionally be able to:

- Fit and interpret a range of regression and ANOVA models and advise non-statisticians on same.
- Translate and communicate the outcomes from a statistical model to a non-statistics audience.
- Identify and communicate to non-statisticians the limitations of standard regression and ANOVA modelling techniques, and situations where model assumptions may need to be altered.
- Advise on how to design, implement and collect data from experiments in ways that are statistically sound.

SS2.1: Regression Analysis

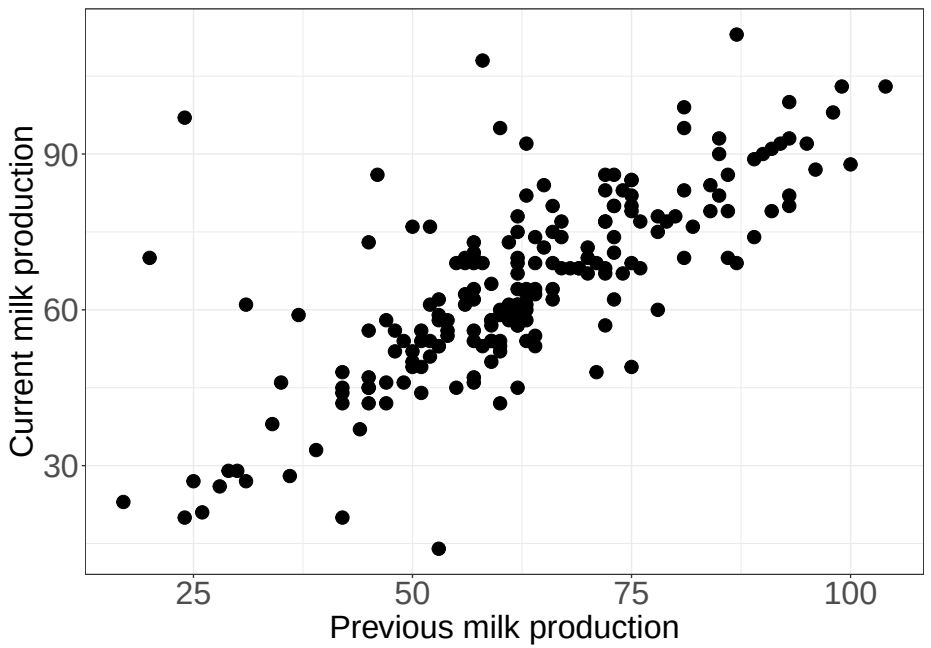
SS2.1.1: Simple linear regression

Example - milk production data

The Dairy Herd Improvement Cooperative in Update New York collects and analyses data on Milk production. It is of interest to predict current milk production from a set of measured potential predictor variables. The response variable is current milk production in pounds. Samples are taken once a month during milking.

Reference for dataset: 'Regression Analysis by Example' by Chatterjee and Hadi.

Can current milk production can be predicted from the previous month's production?



Simple linear regression model and assumptions

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where $i = 1, ..., n$.

β_0: is the intercept parameter, the expected mean of Y when $x = 0$.

β_1: is the slope parameter, the change in the expected mean of Y for a one unit increase in x.

We assume that

1. $E[\epsilon_i] = 0$.
2. $\text{Var}(\epsilon_i) = \sigma^2$ (and does not depend on i).
3. ϵ_i are independent.
4. $\epsilon_i \sim N(0, \sigma^2)$.

The model can be written in different ways, what do the different ways mean?

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$E[Y] = \beta_0 + \beta_1 x$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Fitting a simple linear regression model

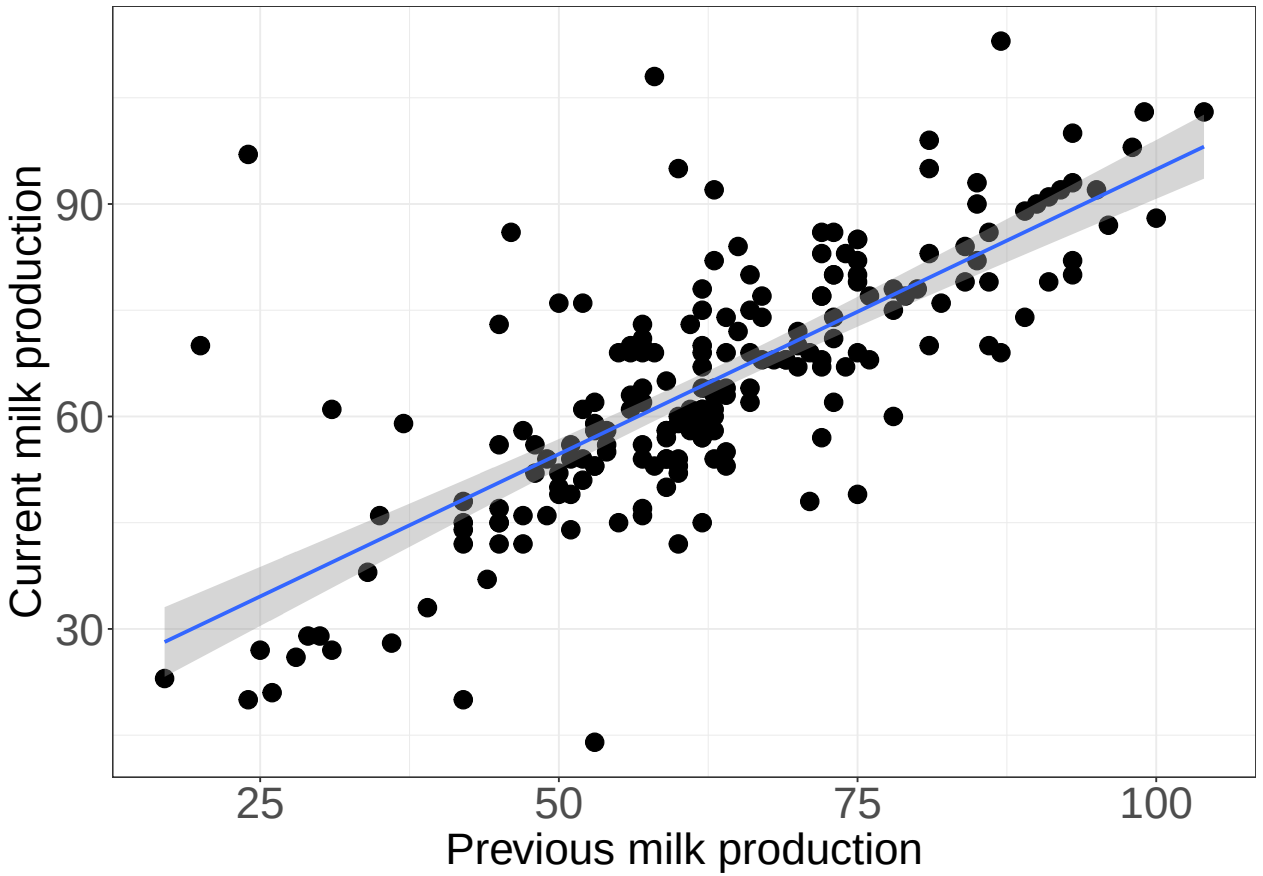
```
##
## Call:
## lm(formula = current ~ previous, data = milk)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.099  -6.403  -1.962   5.058  63.215
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.49063    3.32778   4.354 2.14e-05 ***
## previous     0.80393    0.05135  15.656  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.13 on 197 degrees of freedom
## Multiple R-squared:  0.5544, Adjusted R-squared:  0.5521
## F-statistic: 245.1 on 1 and 197 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_0 = 14.49$ is the estimated average current month's milk production (pounds) when the previous milk production was 0.

$\hat{\beta}_1 = 0.80$ is the estimated increase in average current milk production in pounds for a 1 pound increase in the previous month's milk production.

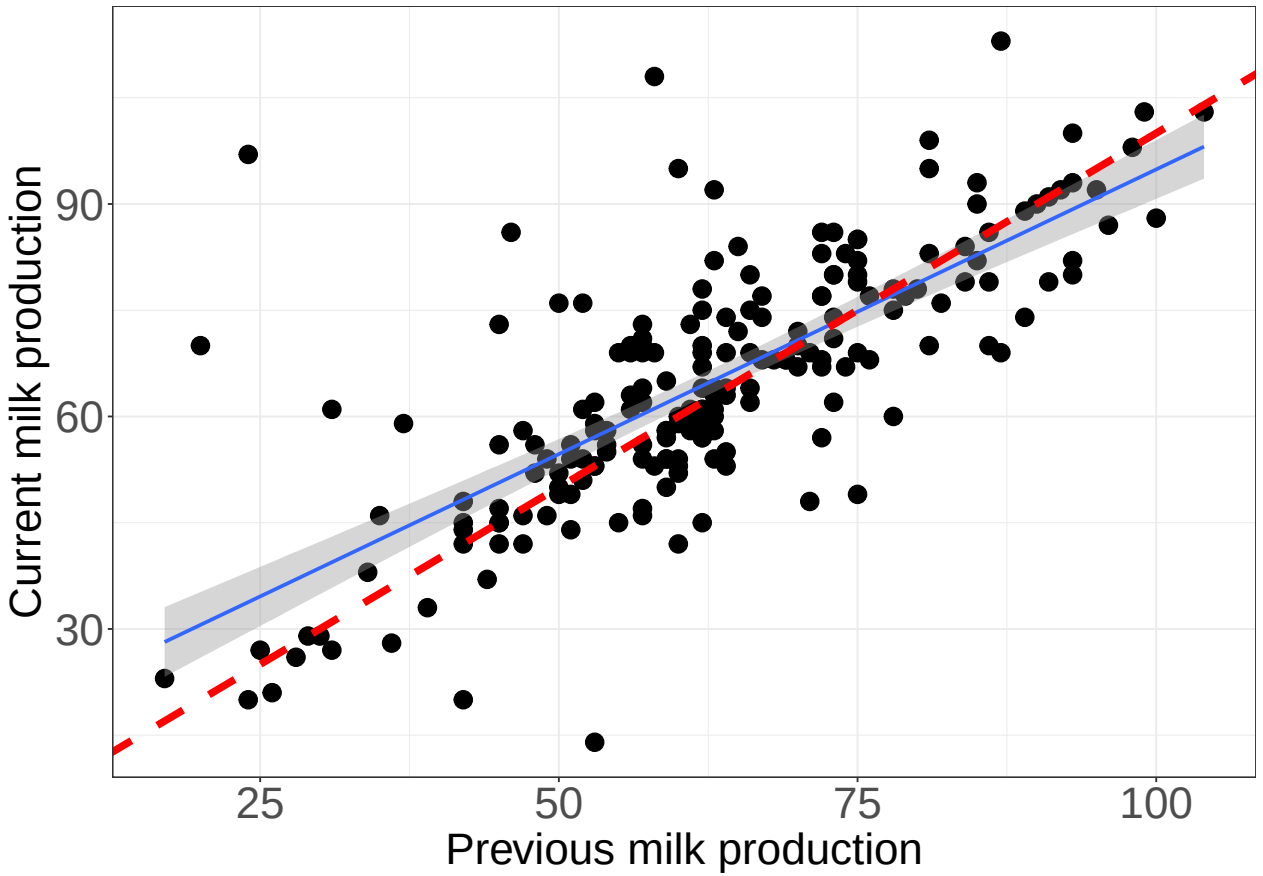
How practical / useful are the interpretations of the intercept and slope for this example?

Visualising the fitted simple linear regression model



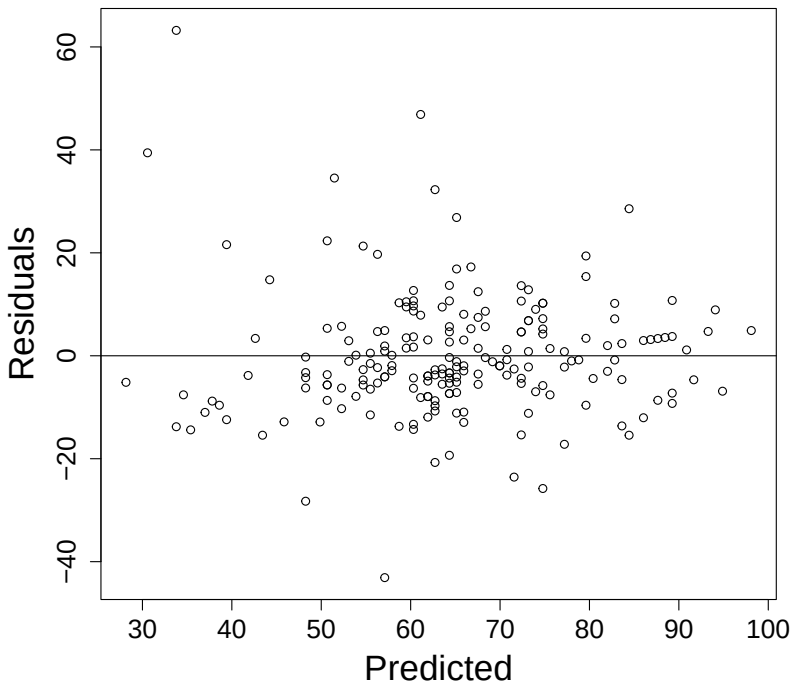
Quantifying uncertainty?

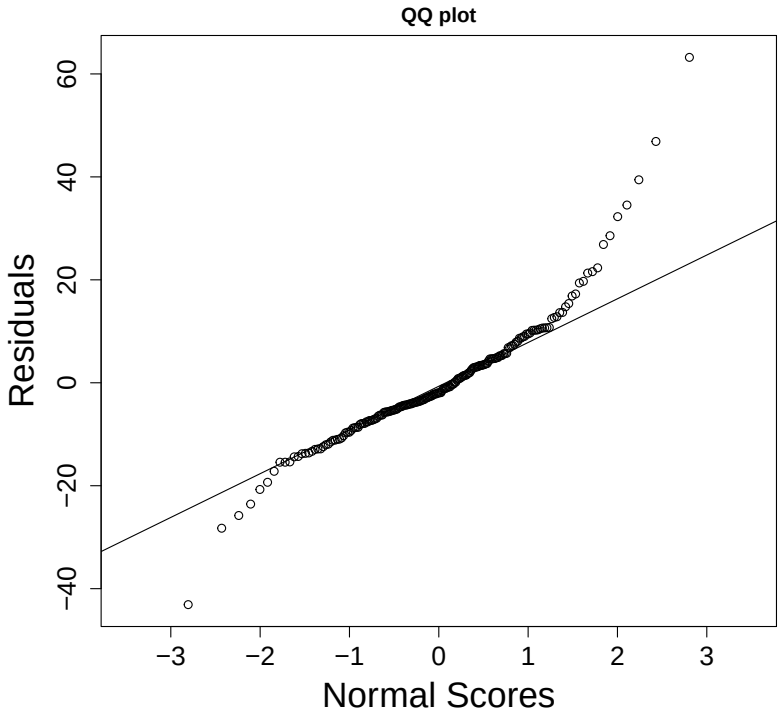
Visualising the fitted simple linear regression model



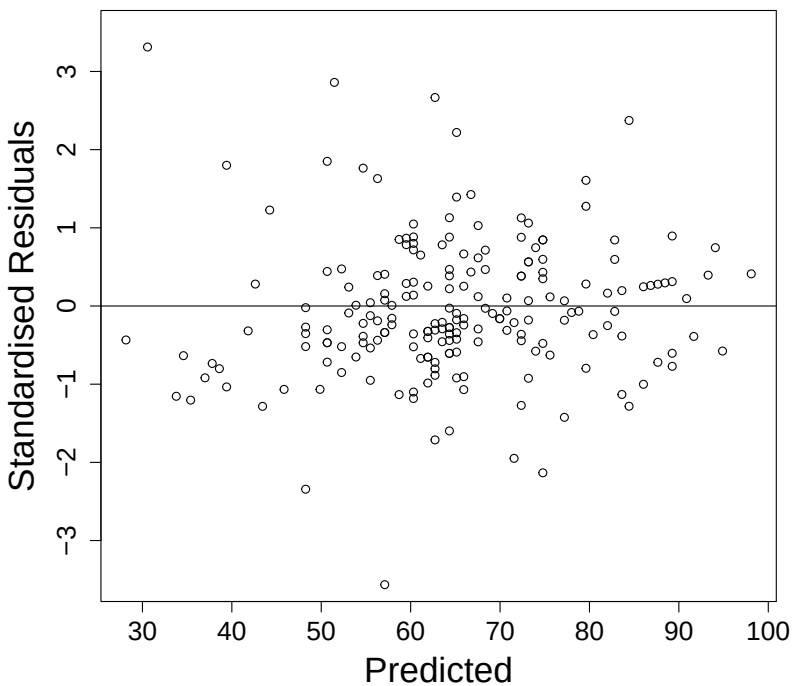
Quantifying uncertainty? Tests about the relationship?

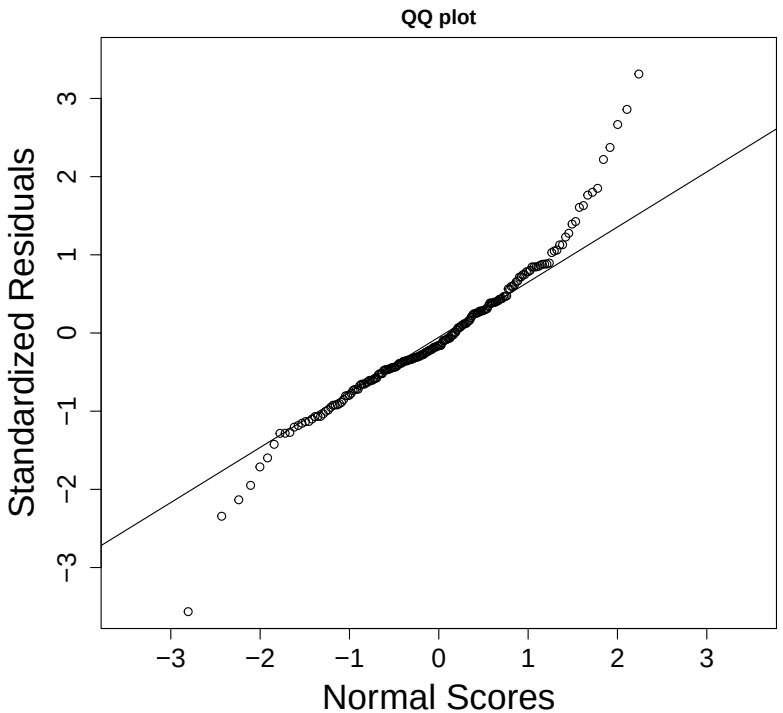
Simple linear regression diagnostics





Simple linear regression diagnostics





Simple linear regression in matrix notation

The regular form of the SLR model is:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{where} \qquad \epsilon_i \sim \text{IID } \mathrm{N}(0, \sigma^2)$$

where IID stands for independent and identically distributed.

This can be expressed as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where $\mathbf{Y}$ is the $n \times 1$ response vector, $\mathbf{X}$ is the $n \times 2$ design matrix, $\boldsymbol{\beta}$ is the 2×1 parameter vector, and $\boldsymbol{\epsilon}$ is the $n \times 1$ error vector, and:

$$\mathrm{E}(\boldsymbol{\epsilon}) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \mathrm{Var}(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

SS2.1.2: Multiple regression models

Model specification and assumptions

A multiple regression model takes the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i \qquad \text{for } i = 1 \text{ to } n$$

Model assumptions are similar to those as stated for the simple linear regression model: $\epsilon_i \sim$ I.I.D. $N(0, \sigma^2)$.

The model can be expressed using matrix notation as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

with dimensions $(n \times 1), (n \times p), (p \times 1)$ and $(n \times 1)$, where $p = k + 1$ is the number of model parameters.

β is a vector of unknown parameters to be estimated from observed data. β_j is the change in the mean value of Y per unit change in x_j, assuming all other independent variables are held constant. Consequently, the β_j parameter estimates depend on which x's are included in the model.

Milk production multiple regression example

```
##
## Call:
## lm(formula = current ~ previous + fat + protein, data = milk)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.419  -7.049  -0.657   5.697  57.750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.77476    7.28785   4.086 6.42e-05 ***
## previous     0.76054    0.05298  14.356  < 2e-16 ***
## fat          1.27922    1.05451   1.213  0.22656
## protein     -5.44158    1.75126  -3.107  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.9 on 195 degrees of freedom
## Multiple R-squared:  0.5754, Adjusted R-squared:  0.5689
## F-statistic:  88.1 on 3 and 195 DF,  p-value: < 2.2e-16
```

The interpretation of any parameter estimate assumes that other predictors in the model are held constant.

For example, it is estimated that the average milk production in pounds in the current month increases by 0.76 for every pound increase in the previous month's milk production, holding fat (x_2, percent of fat in milk) and protein (x_3, percent of protein in milk) constant.

ANOVA tables and Sums of squares

An ANOVA table is a way to partition the variation in the response according to sources.

```
## Analysis of Variance Table
##
## Response: current
##            Df Sum Sq Mean Sq  F value  Pr(>F)
## previous    1  36078   36078 254.6289 < 2e-16 ***
## fat         1      1       1   0.0048 0.94506
## protein     1   1368    1368   9.6550 0.00217 **
## Residuals 195  27629     142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This ANOVA table (generated using the anova() function) presents 'Type I' sums of squares. This means that the variation due to each predictor is 'given' the predictors that come before it in the model. Switching the order gives:

```
## Analysis of Variance Table
##
## Response: current
##            Df Sum Sq Mean Sq  F value  Pr(>F)
## fat         1    892     892   6.2987 0.01290 *
## previous    1  35186   35186 248.3350 < 2e-16 ***
## protein     1   1368    1368   9.6550 0.00217 **
## Residuals 195  27629     142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note on sums of squares

- There are three (well actually four. . .) types of sums of squares (SS) that can be generated that are usually denoted SS Type I, II and III.
- The SS values presented on the previous slide are Type I SS and often referred to as 'sequential' SS since the order matters. The others are sums of squares for predictors in the presence of other main effects, or in the presence of main effects and interactions.
- It can matter which you use! It also matters whether your model includes interactions or not (Type II and III are the same when there are no interactions) and whether or not your data are balanced.
- When a predictor is involved in a higher order interaction, it generally does not make sense to test the significance of the main effect.

This paper is a good read on this topic: Hector, von Felten and Schmid (2010) Analysis of variance with unbalanced data: an update for ecology & evolution. Journal of Animal Ecology, 79, 308–316.

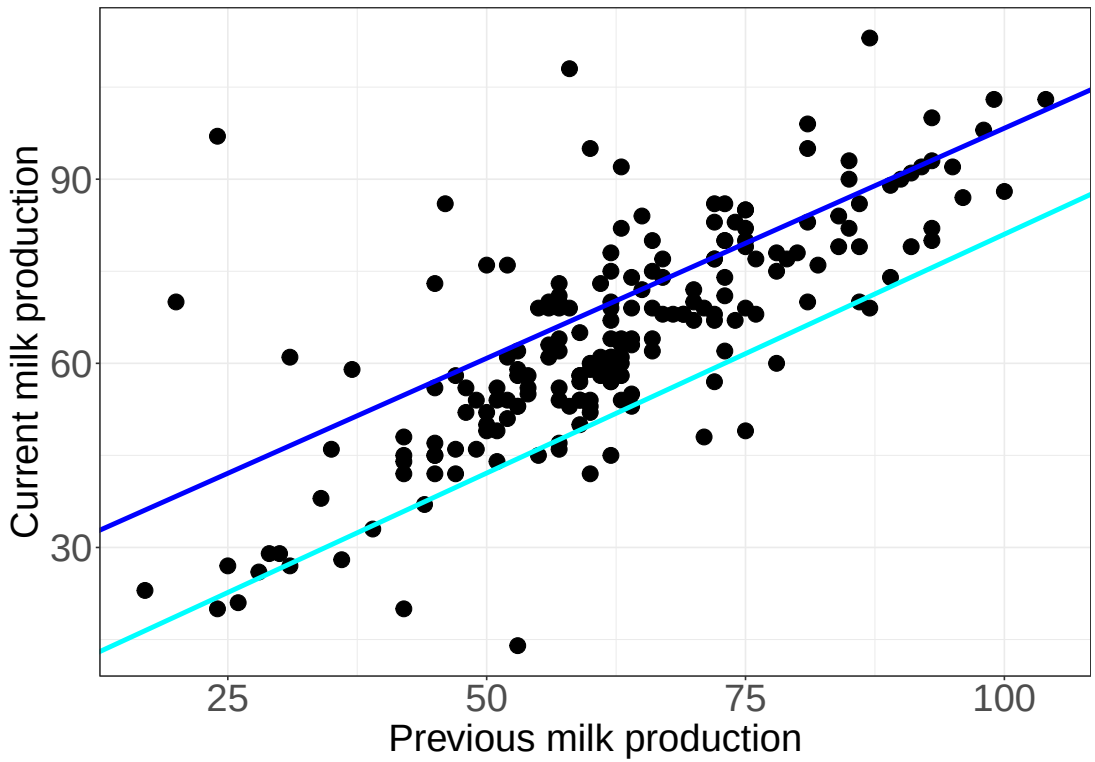
Interactions

We can include interactions between predictors in multiple regression models:

```
##
## Call:
## lm(formula = current ~ previous + protein + previous:protein,
##     data = milk)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.557  -6.687  -1.037   5.175  60.529
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      33.409283  19.826107   1.685   0.0936 .
## previous          0.735351   0.350711   2.097   0.0373 *
## protein          -5.034096   6.116428  -0.823   0.4115
## previous:protein  0.007157   0.111588   0.064   0.9489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.95 on 195 degrees of freedom
## Multiple R-squared:  0.5722, Adjusted R-squared:  0.5657
## F-statistic: 86.95 on 3 and 195 DF,  p-value: < 2.2e-16
```

Visualising predictions from multiple regression models

The lines show predictions for current milk production versus previous milk production for protein = 2 (blue) and protein = 6 (turquoise):



For protein = 2: $\hat{\beta}_0 = 33.409 - 5.034 \times 2$ and $\hat{\beta}_1 = 0.735 + 0.007 \times 2$.

For protein = 6: $\hat{\beta}_0 = 33.409 - 5.034 \times 6$ and $\hat{\beta}_1 = 0.735 + 0.007 \times 6$.

Additional considerations

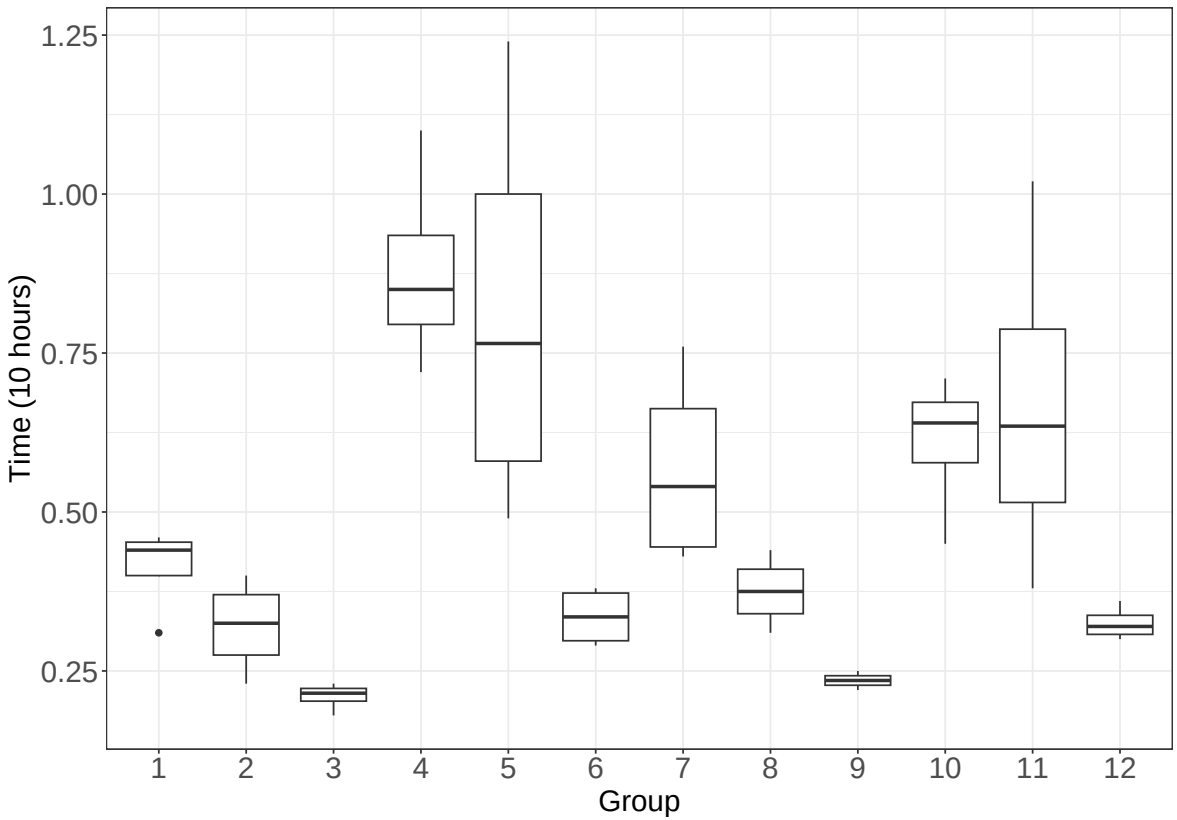
- Diagnostics for multiple regression
- Estimation methods
- Multicollinearity
- Methods for model selection

SS2.1.3: One-way ANOVA

Example - animal survival times data

An randomised experiment was carried out where animals were assigned to various manipulations and their time to survival (in units of 10 hours) was recorded. There were 12 groups to which a total of 48 animals were assigned at random, giving four animals per group.

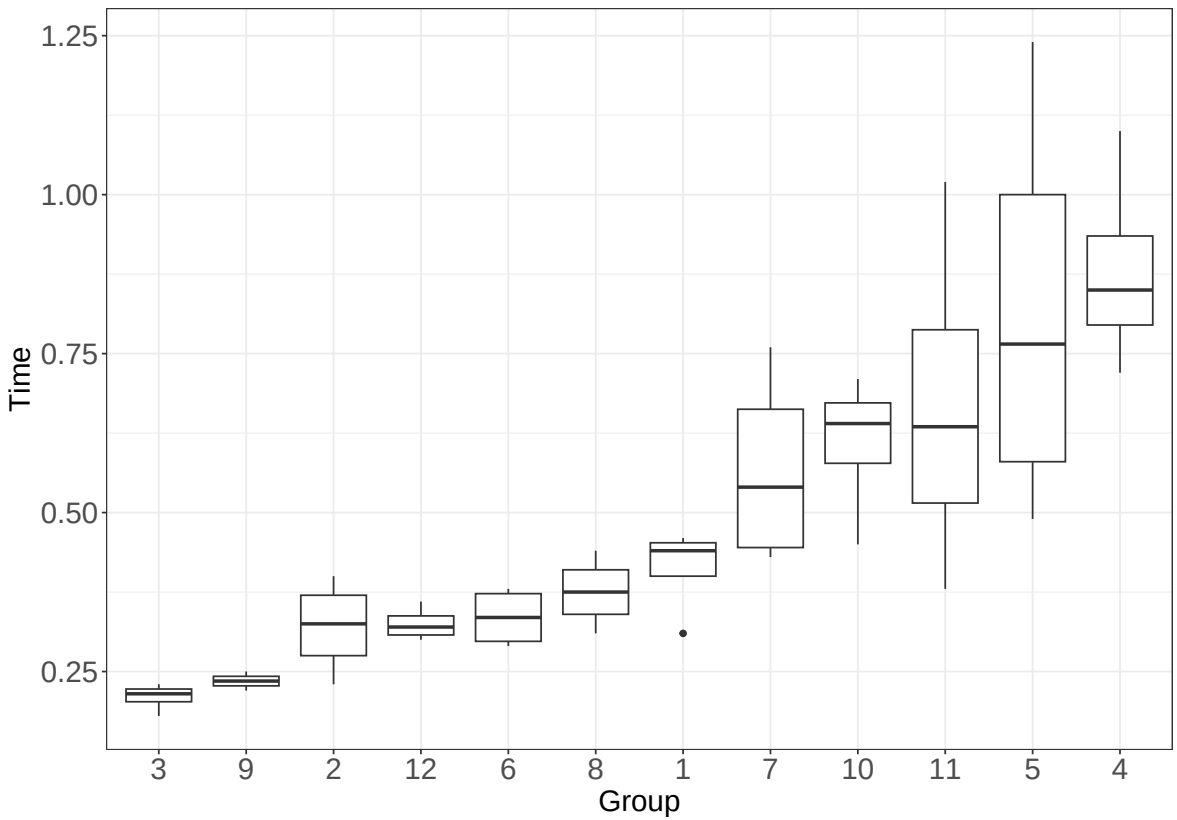
Here is a look at the data:



Example - animal survival times data

An randomised experiment was carried out where animals were assigned to various manipulations and their time to survival (in units of 10 hours) was recorded. There were 12 groups to which a total of 48 animals were assigned at random, giving four animals per group.

Here is a better (!) look at the data:



One-way ANOVA model

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Where $i = 1, ..., k$ for each level of the factor, $j = 1, ..., n_k$ is the replication at each level of the factor.

μ: is a constant.

α_i: is the effect of level i of the factor.

We assume that

1. $E[\epsilon_{ij}] = 0$.
2. $\text{Var}(\epsilon_{ij}) = \sigma^2$ (and does not depend on i or j).
3. ϵ_{ij} are independent.
4. $\epsilon_{ij} \sim N(0, \sigma^2)$.

The model can be written in different ways, what do the different ways mean?

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$$E[Y_{ij}] = \mu + \alpha_i$$

$$\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i$$

The difference between continuous and categorical predictors

The simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

The one-way ANOVA model:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- How many parameters are in each model?
- What is the difference in how the intercepts are interpreted?
- How many degrees of freedom are needed to estimate each model?

Alternative specification of the one-way ANOVA model:

$$y_{ij} = \mu + \alpha_1 x_1 + \alpha_2 x_2 + ...\alpha_{12} x_{12} + \epsilon_{ij}$$

where x_1 to x_{12} are indicator 'dummy' variables for each level of the grouping factor.

One-way ANOVA in matrix notation

The regular form of the one-way ANOVA model is

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \text{where} \qquad \epsilon_{ij} \sim \text{IID N}(0, \sigma^2)$$

In our example, $i = 1, ..., 12$, and each $n_k = 4$, so the model can also be expressed as:

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{47} \\ y_{48} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & 0 & \dots & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{12} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{47} \\ \epsilon_{48} \end{bmatrix}$$

where $\mathbf{Y}$ is the 48×1 response vector, $\mathbf{X}$ is an 48×13 design matrix, β is a 13×1 parameter vector, and ϵ is the 48×1 error vector, and:

$$\text{E}(\boldsymbol{\epsilon}) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \text{Var}(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

Fitted one-way ANOVA model: interpretation of parameter estimates?

```
##
## Call:
## lm(formula = time ~ groupF, data = survival)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32500 -0.04875  0.00500  0.04313  0.42500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.41250    0.07457   5.532 2.94e-06 ***
## groupF2     -0.09250    0.10546  -0.877 0.386230
## groupF3     -0.20250    0.10546  -1.920 0.062781 .
## groupF4      0.46750    0.10546   4.433 8.37e-05 ***
## groupF5      0.40250    0.10546   3.817 0.000513 ***
## groupF6     -0.07750    0.10546  -0.735 0.467163
## groupF7      0.15500    0.10546   1.470 0.150304
## groupF8     -0.03750    0.10546  -0.356 0.724219
## groupF9     -0.17750    0.10546  -1.683 0.101000
## groupF10     0.19750    0.10546   1.873 0.069235 .
## groupF11     0.25500    0.10546   2.418 0.020791 *
## groupF12    -0.08750    0.10546  -0.830 0.412164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1491 on 36 degrees of freedom
## Multiple R-squared:  0.7335, Adjusted R-squared:  0.6521
## F-statistic:  9.01 on 11 and 36 DF,  p-value: 1.986e-07
```

What constraint on parameters has been used and why is it needed?

The design and parameter matrices

$$\begin{pmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & \vdots & & & & & & \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
\mu \\
\alpha_1 \\
\alpha_2 \\
\alpha_3 \\
\alpha_4 \\
\alpha_5 \\
\alpha_6 \\
\alpha_7 \\
\alpha_8 \\
\alpha_9 \\
\alpha_{10} \\
\alpha_{11} \\
\alpha_{12}
\end{pmatrix}$$

How would a constraint on the parameters be enforced via the design matrix?

ANOVA table

```
## Analysis of Variance Table
##
## Response: time
##           Df  Sum Sq  Mean Sq F value    Pr(>F)
## groupF    11 2.20436 0.200396  9.0097 1.986e-07 ***
## Residuals 36 0.80072 0.022242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The hypothesis being tested here is:

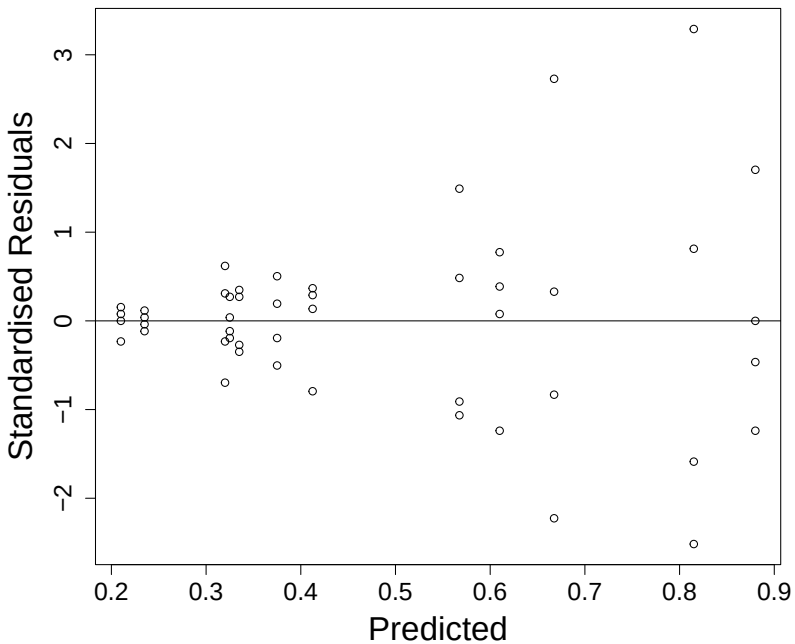
$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_{12} = 0,$$

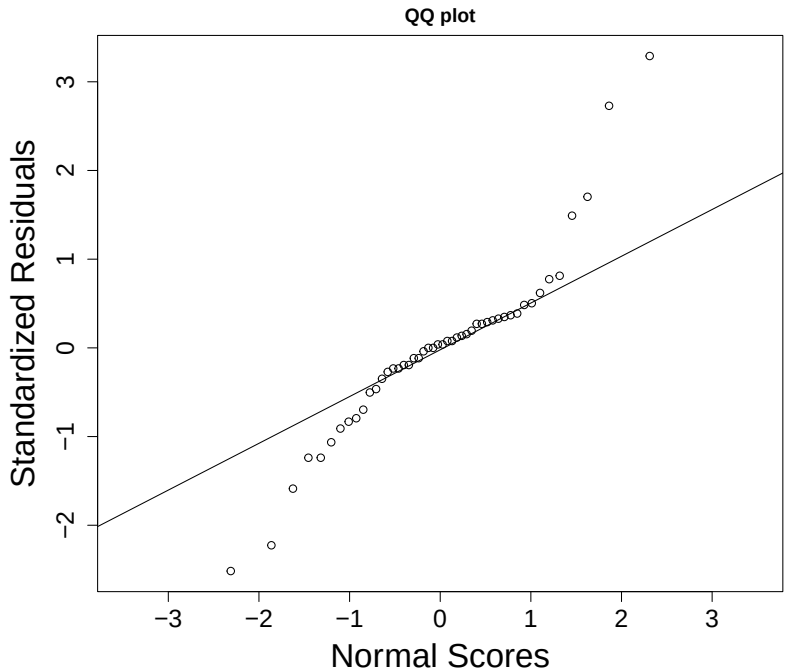
versus

$$H_A: \text{At least one } \alpha_i \text{ differs from } 0.$$

What is the interpretation from the ANOVA table and what might the next steps be?

Diagnostics



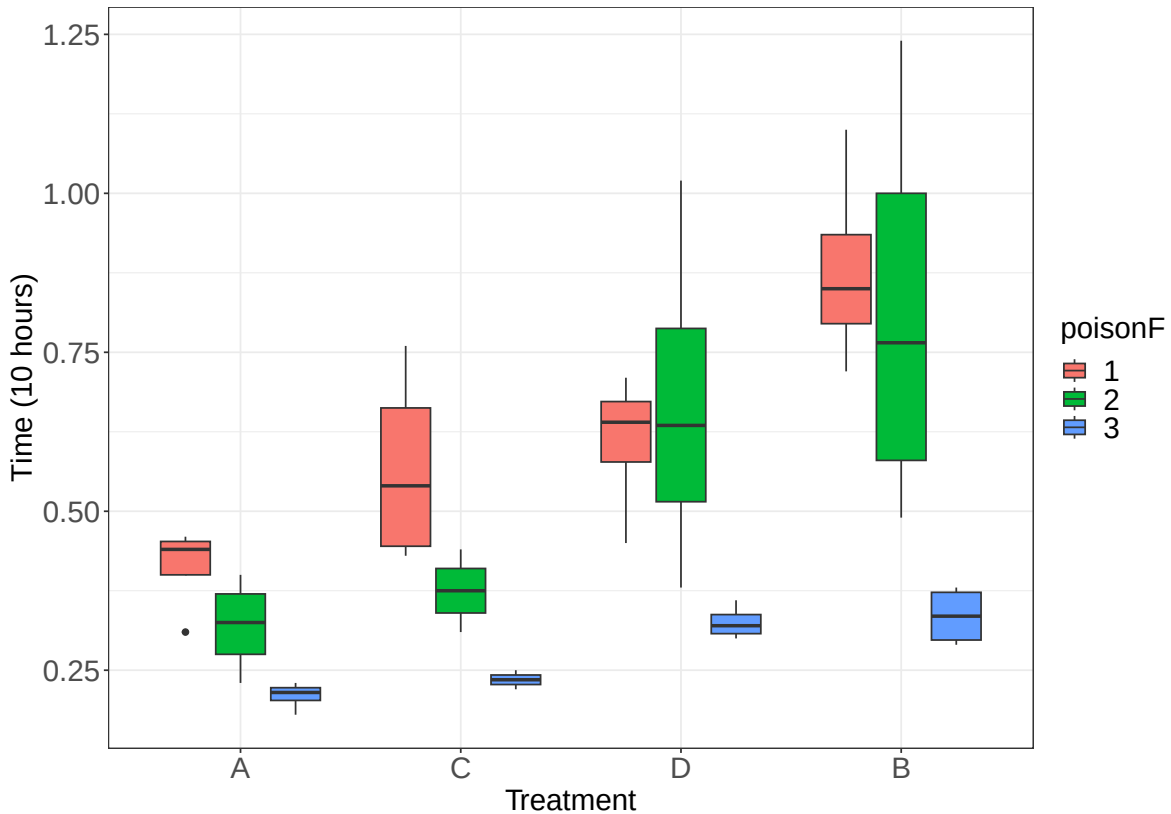


SS2.1.4: Two-way ANOVA

Example

The data used in the one-way ANOVA model example was actually a two-factor factorial design. The data was taken from Box and Cox (1964).

Each of the 48 animals was assigned to one of the three poisons (factor with levels: 1, 2, 3) and to one of the four treatments (factor with levels: A, B, C, D).



Model specification and assumptions

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

Where, for our example:

μ is a constant

α_i is the effect of level $i = 1, 2, 3$ of poison.

β_j is the effect of level $j = 1, 2, 3, 4$ for the A, B, C, D levels respectively of treatment.

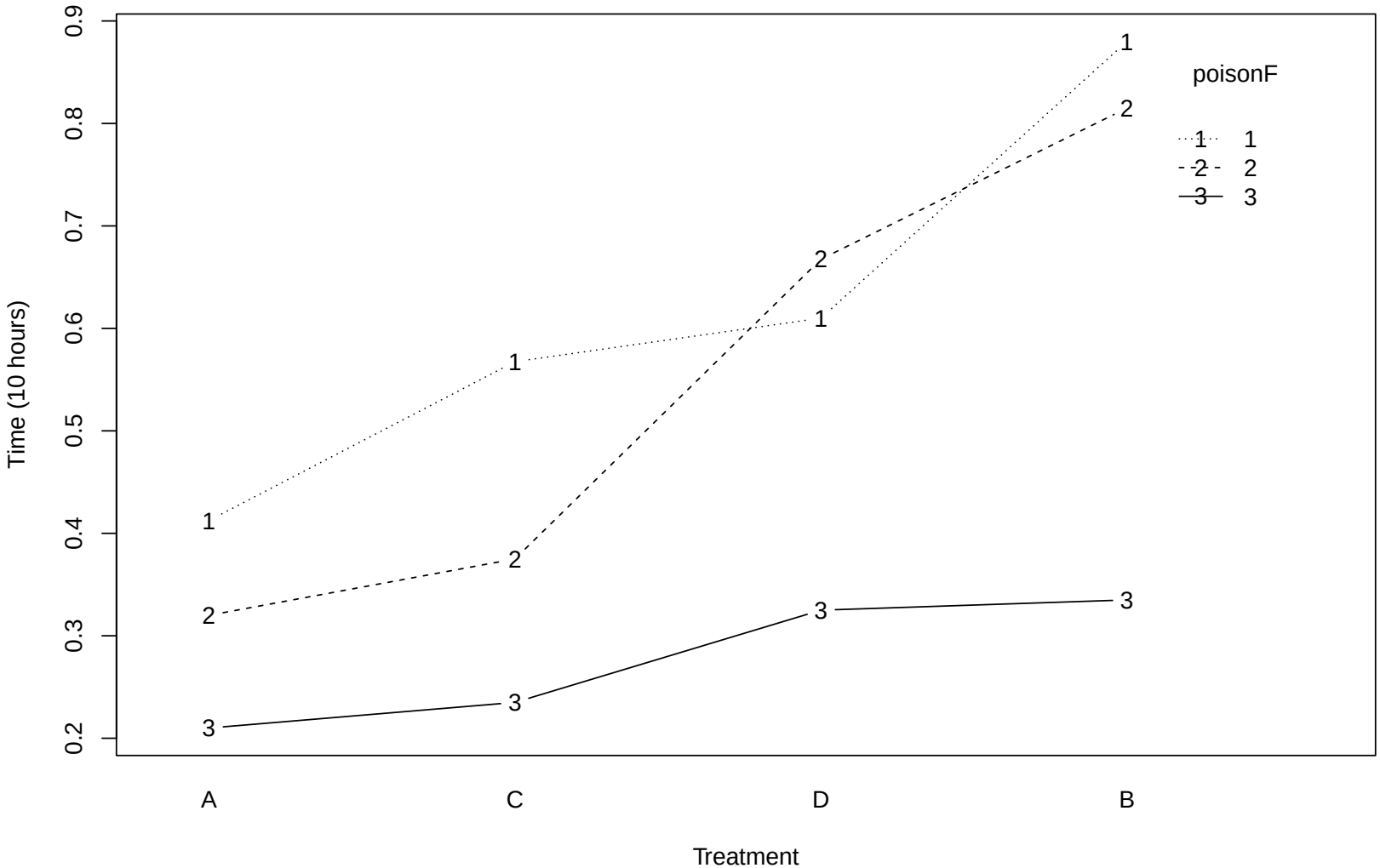
γ_{ij} is the interaction effect of level i of poison and level j of treatment.

And $k = 1, ...4$ indicates the replicates within each i, j, combination (it was a full two-factor factorial balanced design).

ANOVA table

```
## Analysis of Variance Table
##
## Response: time
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## poisonF            2 1.03301 0.51651 23.2217 3.331e-07 ***
## treatment          3 0.92121 0.30707 13.8056 3.777e-06 ***
## poisonF:treatment  6 0.25014 0.04169  1.8743    0.1123
## Residuals         36 0.80072 0.02224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Examine the two-way ANOVA model with interaction



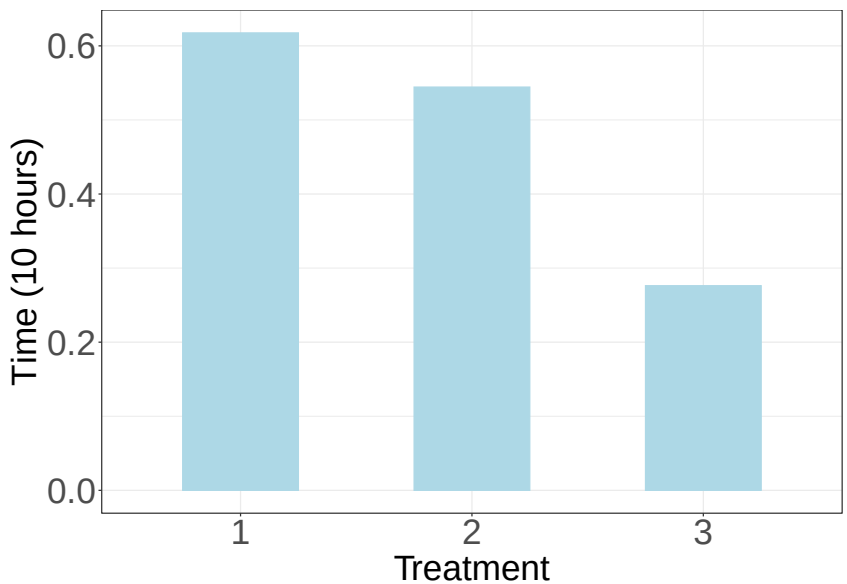
ANOVA table without the interaction

```
## Analysis of Variance Table
##
## Response: time
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## poisonF    2 1.03301 0.51651  20.643 5.704e-07 ***
## treatment  3 0.92121 0.30707  12.273 6.697e-06 ***
## Residuals 42 1.05086 0.02502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: time
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## treatment  3 0.92121 0.30707  12.273 6.697e-06 ***
## poisonF    2 1.03301 0.51651  20.643 5.704e-07 ***
## Residuals 42 1.05086 0.02502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

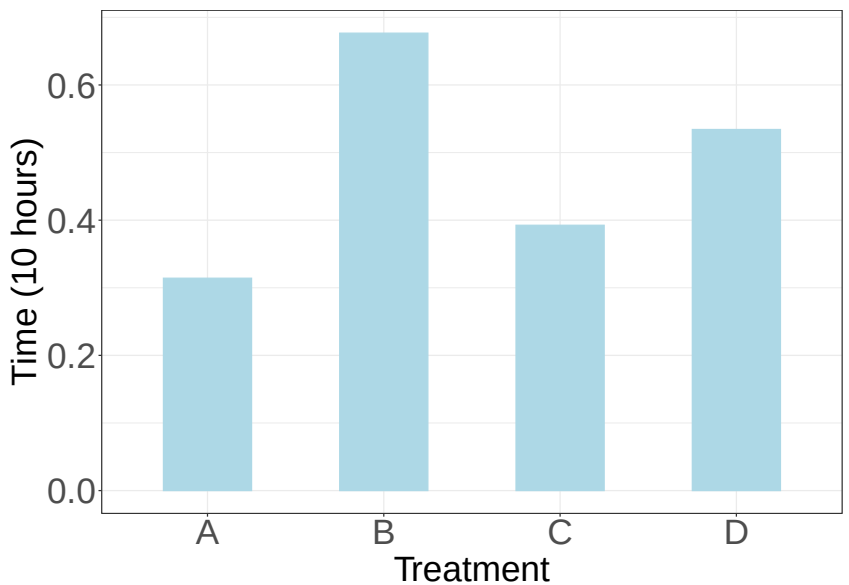
Comparisons across poison levels

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = time ~ poisonF + treatment, data = survival)
##
## $poisonF
##          diff        lwr        upr     p adj
## 2-1 -0.073125 -0.2089936  0.0627436 0.3989657
## 3-1 -0.341250 -0.4771186 -0.2053814 0.0000008
## 3-2 -0.268125 -0.4039936 -0.1322564 0.0000606
```

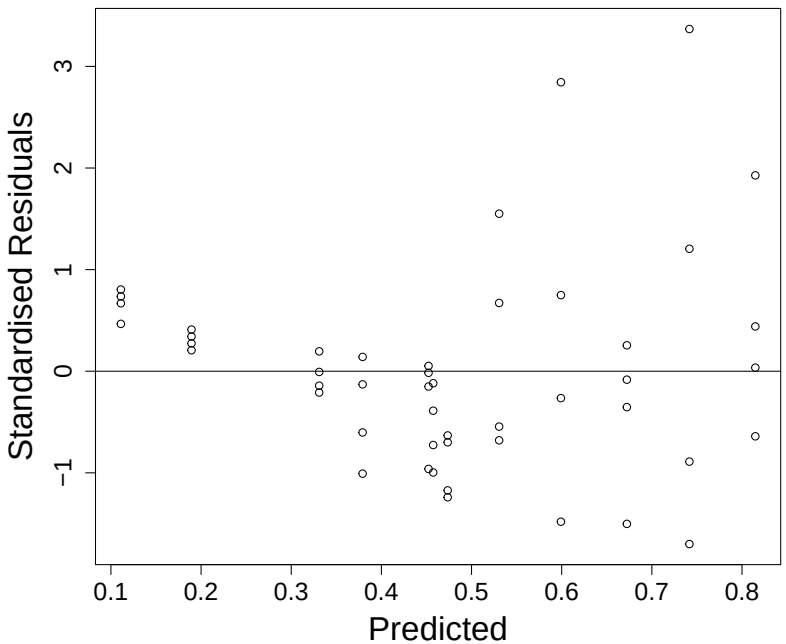


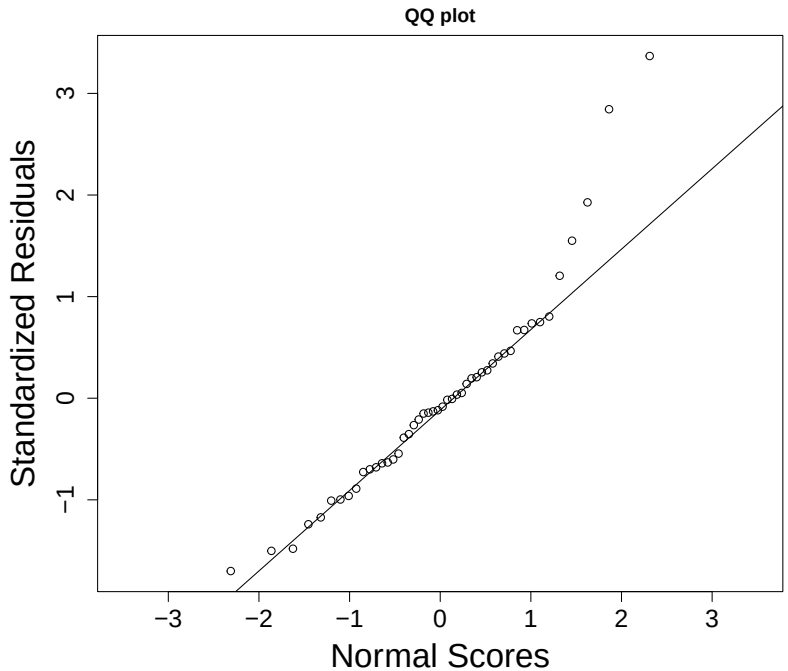
Comparisons across treatment levels

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = time ~ poisonF + treatment, data = survival)
##
## $treatment
##             diff         lwr         upr     p adj
## B-A  0.36250000  0.18976135  0.53523865 0.0000083
## C-A  0.07833333 -0.09440532  0.25107198 0.6221729
## D-A  0.22000000  0.04726135  0.39273865 0.0076661
## C-B -0.28416667 -0.45690532 -0.11142802 0.0004090
## D-B -0.14250000 -0.31523865  0.03023865 0.1380432
## D-C  0.14166667 -0.03107198  0.31440532 0.1416151
```



Diagnostics





SS2.1.5: Continuous and categorical predictors

Continuous and categorical predictors

So far, we have explored regression models with only continuous predictors, and only categorical predictors.

Multiple regression models can include both continuous and categorical predictors and interactions between continuous and categorical predictors are also possible. This type of analysis is sometimes referred to as 'Analysis of Covariance' or ANCOVA.

We will explore this topic further in the lab session.

SS2.1.6: Recommended reading

Some suggested reading material

Books

OpenIntro Statistics, by Diez, Çetinkaya-Rundel and Barr.

Applied Linear Regression, by Sandford and Weisberg.

Regression Analysis by Example, by Chatterjee and Hadi.

The Statistical Sleuth, by Ramsey and Schafer.

Statistics for experimenters, by Box, Hunter and Hunter.

Papers

Hector, von Felten and Schmid (2010) Analysis of variance with unbalanced data: an update for ecology & evolution. Journal of Animal Ecology, 79, 308–316.

Zuur,Ieno and Elphick (2009) A protocol for data exploration to avoid common statistical problems. Methods in Ecology & Evolution, 1, 3–14.

Box and Cox (1964) An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26, 211–52.