University of Zurich[UZH]

MAKERERE UNIVERSITY

Data Analysis with R:
Day 6 - Preliminary - Slides

Sonja Hartnack, Terence Odoch & Muriel Buri

October 2018

Exercise 15

Introduction: ANOVA as a special case of a linear model

Linear Modelling

So far we have considered how to determine whether statistically significant differences exist between different ”feed” groups (factors).
The explanatory variable has been categorical. We have not yet considered continuous explanatory variables.
ANOVA and linear regression are both a special cases of **linear models**.

Simple linear regression (1/3)

A **simple linear regression** fits a straight line through a set of data points. This straight line is fitted in a way that makes the sum of the squared residuals (the vertical distances between each data point and the fitted line) as small as possible.

Simple linear regression (2/3)

In a simple linear regression model for n data points $(x_i, y_i), i = 1, ...n$ the following equation is used:

$$y_i = \alpha + \beta x_i + \epsilon,$$

where

- α: intercept or constant
- β: slope, regression coefficient (effect size)
- ϵ: error, residuals

Simple linear regression (3/3)

The goal is to find the equation of the straight line

$$f(x) = y = \alpha + \beta x,$$

which would provide a ”best” fit for the data points (least square approach).

Assumptions of a linear regression

A **simple linear regression** is based on several assumptions which should be checked carefully.

- linearity of the relationship between the explanatory (independent) and the outcome (dependent) variable
- normality of the residuals
- independence
- constant variance (homoscedasticity)

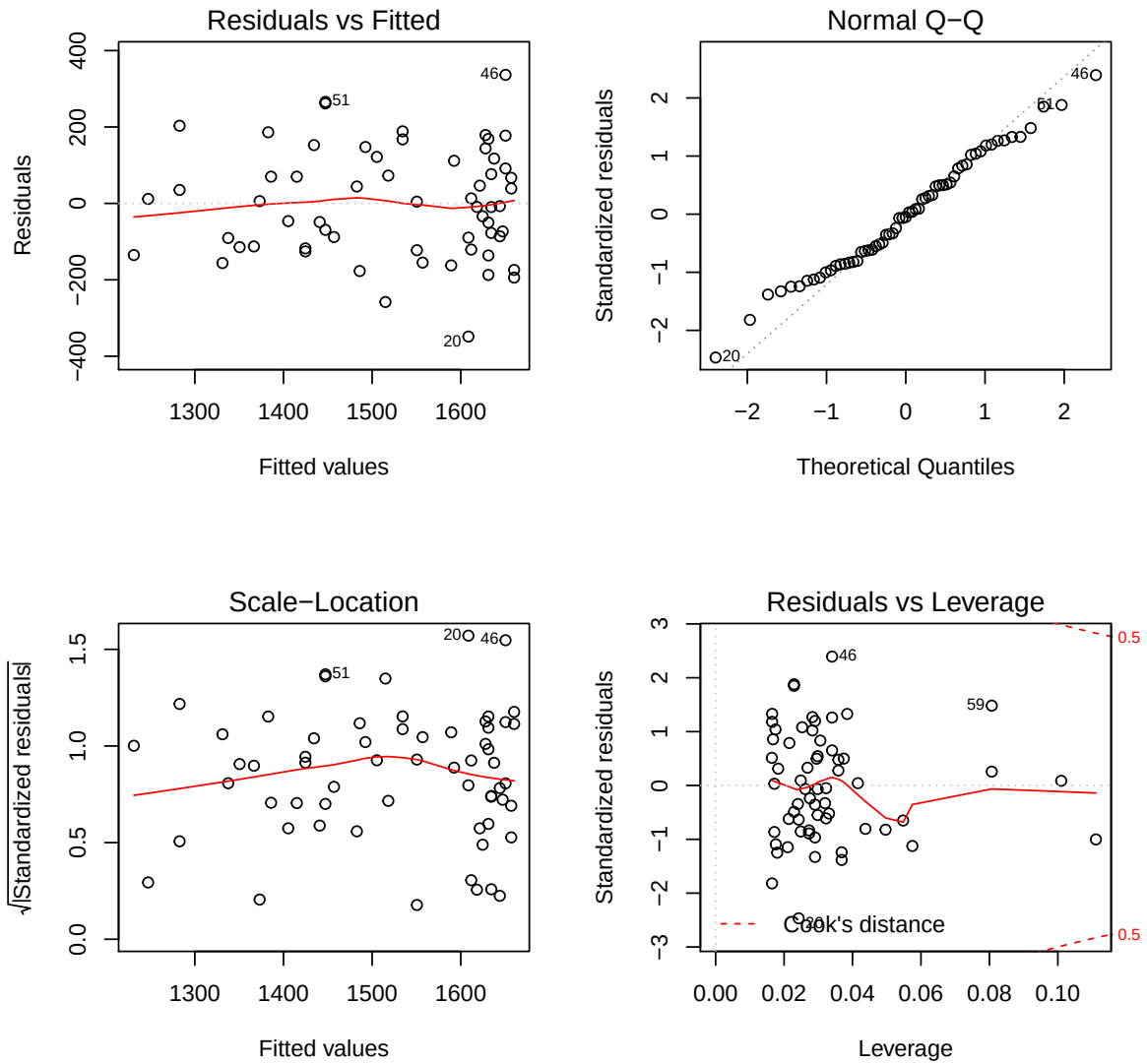
Linear regression model in R

```
data(water)
# mod.hard <- lm(mortality ~ hardness, data = water)
mod.hard <- lm(water$mortality ~ water$hardness)
summary(mod.hard)


##
## Call:
## lm(formula = water$mortality ~ water$hardness)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -348.61 -114.52   -7.09  111.52  336.45
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1676.3556    29.2981  57.217  < 2e-16 ***
## water$hardness   -3.2261     0.4847  -6.656 1.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143 on 59 degrees of freedom
## Multiple R-squared:  0.4288,Adjusted R-squared:  0.4191
## F-statistic:  44.3 on 1 and 59 DF,  p-value: 1.033e-08
```

Checking linear model assumptions

```
par(mfrow=c(2,2))
plot(mod.hard)
```



Linear model vs. t.test(...) in R

```
mod.loc <- lm(mortality ~ location, data = water)
coef(mod.loc)

##   (Intercept) locationSouth
##     1633.6000     -256.7923

t.test(water$mortality ~ water$location)

##
##  Welch Two Sample t-test
##
## data:  water$mortality by water$location
## t = 7.1427, df = 53.29, p-value = 2.584e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  184.6919 328.8928
## sample estimates:
## mean in group North mean in group South
##           1633.600            1376.808
```

The multiple linear model

An extension of the simple linear model (1/2)

If several explanatory variables are of interest, instead of performing multiple simple regressions, a **multiple linear regression** or **multivariable** approach is appropriate.

- the same assumptions as for simple linear regressions should be checked
- collinearity might be an issue (see vif(...) function from package `usdm`)
- model comparison (AIC) (...discussed later)

The multiple linear model

An extension of the simple linear model (2/2)

Simple linear regression model:

$$y_i = \alpha + \beta x_i + \epsilon$$

Multiple linear regression model:

$$y_i = \alpha + \beta_1 x1_i + \beta_2 x2_i + ... + \epsilon, \quad (1)$$
$$y_i = \alpha + \beta_1 x1_i * \beta_2 x2_i + ... + \epsilon,$$

where

- α: intercept or constant
- β_1: slope1, regression coefficient (effect size)
- β_2: slope2, regression coefficient (effect size)
- $\beta_1 * \beta_2$: interaction effect between x_1 and x_2
- ϵ: error, residuals

The multiple linear model

Interpretation of model coefficients

- α: intercept or constant
- β_1: slope1, regression coefficient (effect size)
- β_2: slope2, regression coefficient (effect size)
- $\beta_1 * \beta_2$: interaction effect between x_1 and x_2
- ϵ: error, residuals

- $\rightarrow \beta_1$ describes the number of units of a change in the outcome variable y as x_1 changes by one unit, x_2 being held constant.
- $\rightarrow \beta_2$ describes the number of units of a change in the outcome variable y as x_2 changes by one unit, x_1 being held constant.

The multiple linear model in R

```
mod.hard.loc <- lm(mortality ~ hardness + location, data = water)
summary(mod.hard.loc)
```

```
##
## Call:
## lm(formula = mortality ~ hardness + location, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -222.959  -77.281    7.143   90.751  307.836
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1695.4371    25.3285  66.938  < 2e-16 ***
## hardness        -2.0341     0.4829  -4.212 8.93e-05 ***
## locationSouth -176.7108    36.8913  -4.790 1.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.1 on 58 degrees of freedom
## Multiple R-squared:  0.5907,Adjusted R-squared:  0.5766
## F-statistic: 41.86 on 2 and 58 DF,  p-value: 5.601e-12
```

Model selection

- Multiple R^2: the larger, the better
- Akaike criterion (AIC): the smaller, the better

Nested models

Two models are nested if one of them is a particular case of the other one: the simpler model can be obtained by setting some coefficients of the more complex model to particular values.

```
mod.hard <- lm(mortality ~ hardness, data = water)                # mod 1
mod.loc <- lm(mortality ~ location, data = water)                 # mod 2
mod.hard.loc <- lm(mortality ~ hardness + location, data = water) # mod 3
```

Among the 3 above models ...

- which ones are nested?
- which ones are not nested?

Multiple R^2

R^2 is a measure of fit quality:

$$R^2 = \frac{\text{SSRegression}}{\text{SSTotal}}$$
$$R^2 = 1 - \frac{\text{SSError}}{\text{SSTotal}}$$

CAREFUL:

- The SSError always decreases as more predictors are added to the model.
- R^2 always increases and can be artificially large.

Akaike criterion (AIC)

- Model fit (R^2) always improves with model complexity. We would like to strike a good balance between **model fit** and **model simplicity**.
- AIC combines a measure of model fit with a measure of model complexity: The smaller, the better.

For a given data set and a given model:

$$AIC = -2 \cdot \log(L) + 2p$$

L stands for the likelihood. p stands for the number of parameters in the models (penalizes complex models).

Model selection strategy for AIC

- Consider a number of candidate models.
  (They need not be nested.)
- Calculate their AIC.
- Choose the model(s) with the smallest AIC.

$\rightarrow$ **CAREFUL**: The absolute value of AIC is meaningless. The relative AIC values, between models, is meaningful.

Model selection with AIC in R

```
mod1 <- lm(mortality ~ hardness, data = water)
mod2 <- lm(mortality ~ location, data = water)
mod3 <- lm(mortality ~ hardness + location, data = water)
AIC(mod1, mod2, mod3)
```

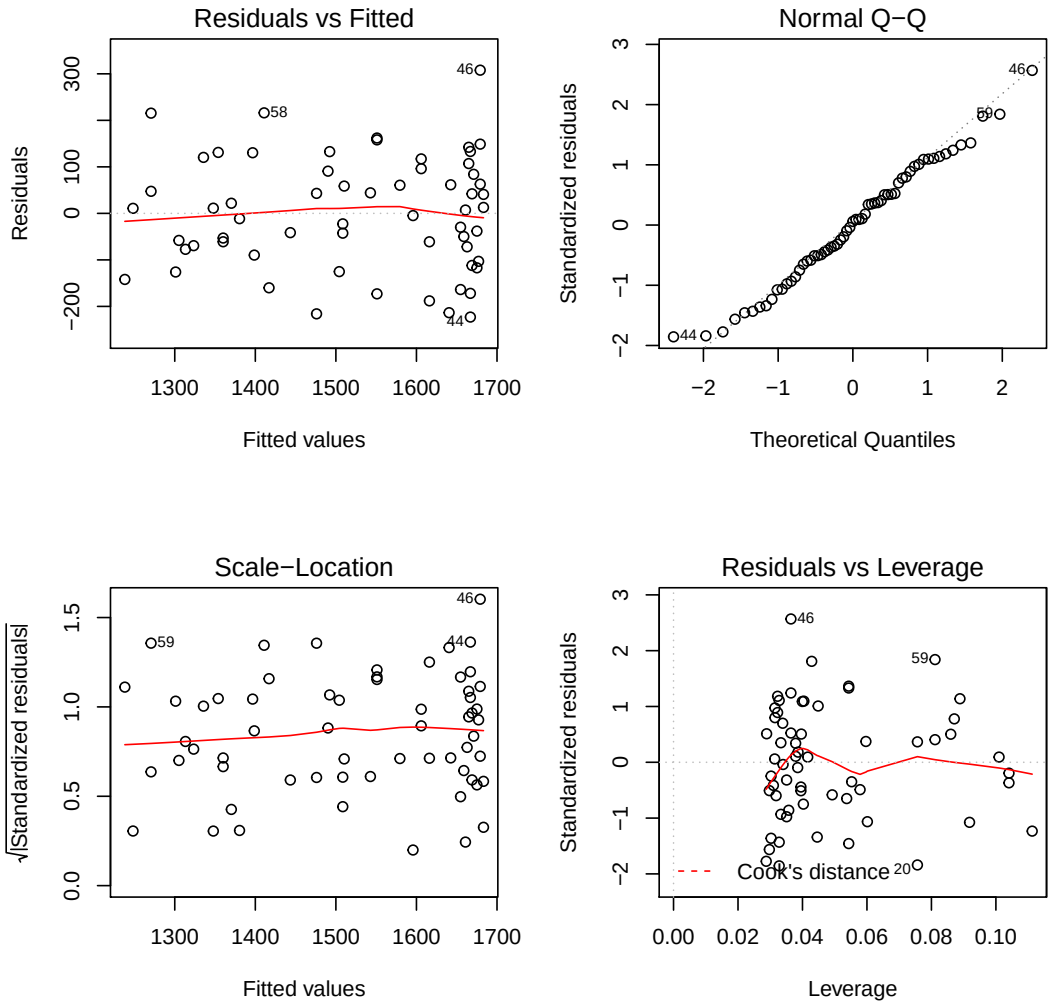
```
##      df      AIC
## mod1  3 782.5692
## mod2  3 778.5186
## mod3  4 764.2366
```

```
round(AIC(mod1, mod2, mod3), 2)
```

```
##      df    AIC
## mod1  3 782.57
## mod2  3 778.52
## mod3  4 764.24
```

`mod.hard.loc` is the best!

```
mod3 <- lm(mortality ~ hardness + location, data = water)
par(mfrow=c(2,2))
plot(mod3)
```



Exercise 16

Two-way Interactions in R (1/3)

There are three different interactions:

- **Interaction between two categorical variables**
- Interaction between one continuous and one categorical variables
- Interaction between two continuous variables

Two-way Interactions in R (2/3)

Model specification in R: Be aware, an interaction is never tested without its corresponding main effects included in the model.

```r
# Interaction between two categorical variables
# mod.dose.supp.int <- lm(len ~ dose.fac + supp + dose.fac:supp,
# data = ToothGrowth)
mod.dose.supp.int <- lm(len ~ dose.fac * supp, data = ToothGrowth)
# summary(mod.dose.supp.int)
```

$$
\begin{aligned}
y \sim \beta_{\text{baseline((dose==low) \& (supp==OJ))}} + \beta_{\text{dose==med}} + \\
\beta_{\text{dose==high}} + \beta_{\text{supp==VC}} \\
+ \beta_{\text{(dose==med) \& (supp==VC)}} \\
+ \beta_{\text{(dose==high) \& (supp==VC)}}
\end{aligned}
$$

Two-way Interactions in R (3/3)

```
# Interaction between two categorical variables
# mod.dose.supp.int <- lm(len ~ dose.fac + supp + dose.fac:supp,
# data = ToothGrowth)
mod.dose.supp.int <- lm(len ~ dose.fac * supp, data = ToothGrowth)
summary(mod.dose.supp.int)
```

```
##
## Call:
## lm(formula = len ~ dose.fac * supp, data = ToothGrowth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -8.20  -2.72  -0.27   2.65   8.27
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           13.230      1.148  11.521 3.60e-16 ***
## dose.facmed            9.470      1.624   5.831 3.18e-07 ***
## dose.fachigh          12.830      1.624   7.900 1.43e-10 ***
## suppVC                -5.250      1.624  -3.233  0.00209 **
## dose.facmed:suppVC    -0.680      2.297  -0.296  0.76831
## dose.fachigh:suppVC    5.330      2.297   2.321  0.02411 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.631 on 54 degrees of freedom
## Multiple R-squared:  0.7937,Adjusted R-squared:  0.7746
## F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

Two-way Interactions in R: Interpretation of coefficients (1/2)

```
coef(mod.dose.supp.int)
```

```
##         (Intercept)         dose.facmed        dose.fachigh
##               13.23                9.47               12.83
##              suppVC  dose.facmed:suppVC dose.fachigh:suppVC
##               -5.25               -0.68                5.33
```

$$y \sim \beta_{\text{baseline((dose==low) \& (supp==OJ))}} + \beta_{\text{dose==med}} + \beta_{\text{dose==high}} + \beta_{\text{supp==VC}} + \beta_{\text{(dose==med) \& (supp==VC)}} + \beta_{\text{(dose==high) \& (supp==VC)}}$$

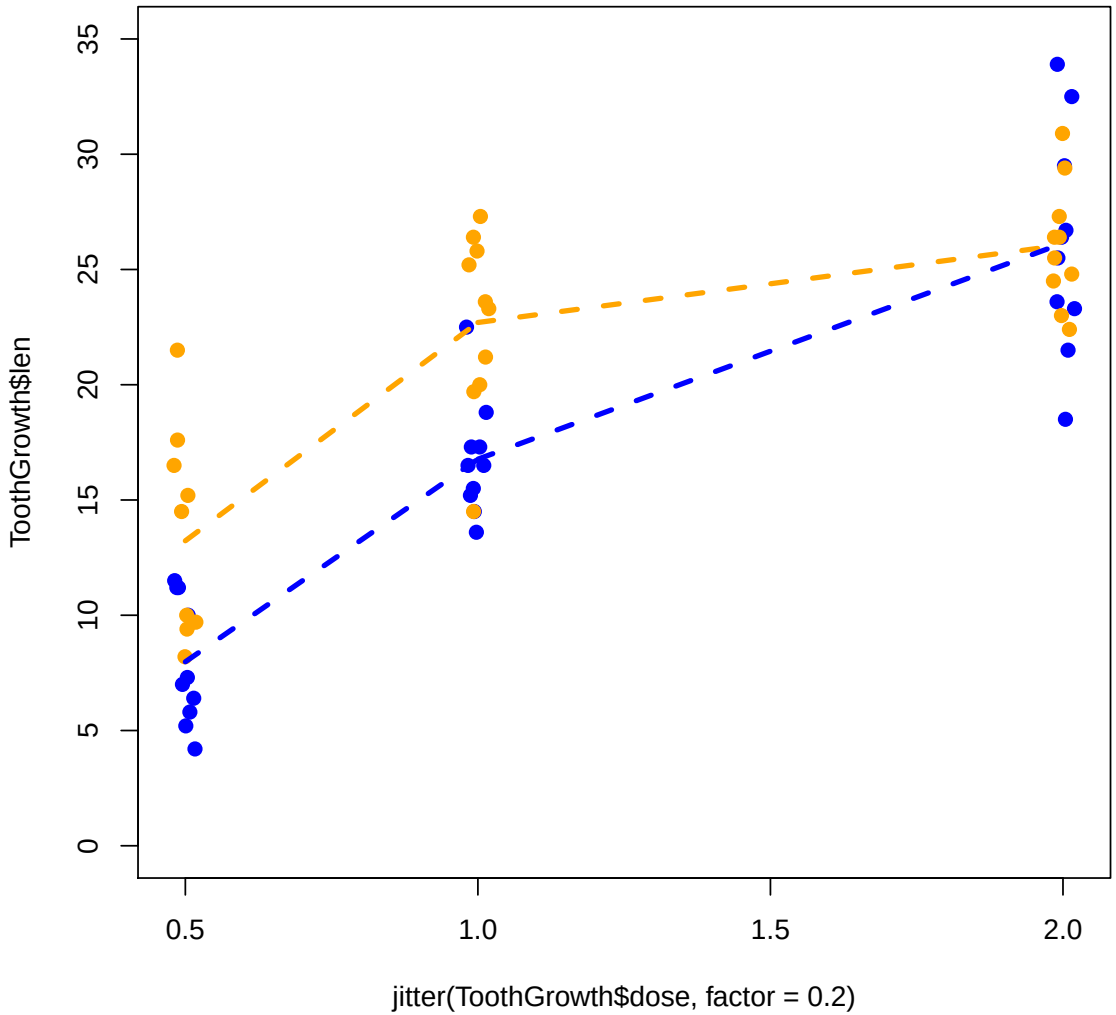
Two-way Interactions in R: Interpretation of coefficients (2/2)

```
coef(mod.dose.supp.int)
```

```
##         (Intercept)         dose.facmed        dose.fachigh
##               13.23                9.47               12.83
##               suppVC  dose.facmed:suppVC dose.fachigh:suppVC
##               -5.25               -0.68                5.33
```

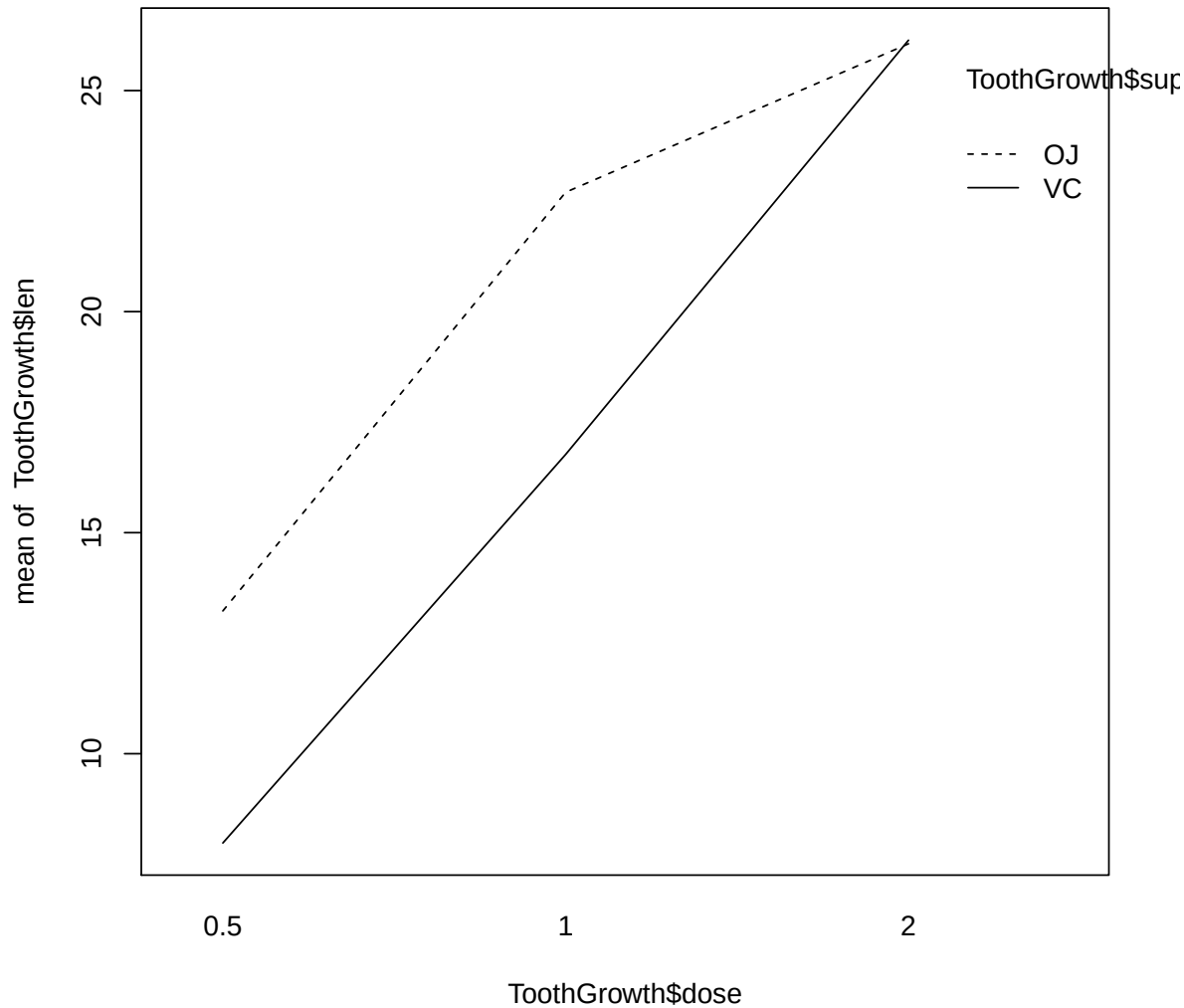
- $\rightarrow$ `dose.facmed:suppVC`
  change in the slope between the `low` and `med` `dose.fac` group under the supplement type (`supp`) `VC` in comparison to the intercept. In other words, by changing the dose from `low` to `med` within the supplement group `VC`, the slope **de**creases by approx. -0.68.
- $\rightarrow$ `dose.fachigh:suppVC`
  change in the slope between the `low` and `high` `dose.fac` group under the supplement type (`supp`) `VC` in comparison to the intercept. In other words, by changing the dose from `low` to `high` within the supplement group `VC`, the slope **in**creases by approx. $+5.33$.

Interactions



Interactions

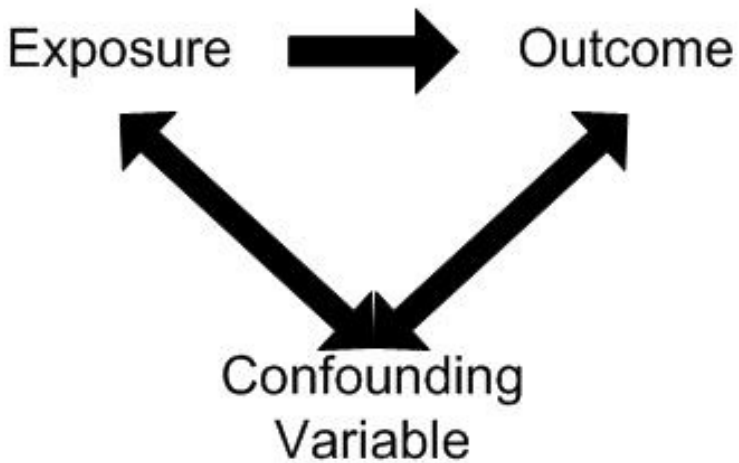
```
interaction.plot(ToothGrowth$dose, ToothGrowth$supp, ToothGrowth$len,
                 fixed = TRUE)
```



Exercise 17

Confounding

- **Confounding variable** - is associated with the exposure and the outcome
- **Confounding variable** - is not part of the causal path between exposure and the outcome



Some things to remember

- Not adjusting / controlling for a confounding variable may lead to biased results. It is good practice to present as well the crude (not adjusted ORs) and the adjusted ones. Adjustment is typically done if difference $>$ 10%.

- Check also for interaction (= effect modification) e. g. using logistic regression.

- Do not adjust for a variable C if it is a common effect of E and D (collider) or if it is in the causal pathway of E and D.

Exercise 18

Other topics

So far, we have covered some of the basic **tools** necessary for helping you to start to analyse your own study data.

Many things have been mentioned others have not been included - some are worth mentioning as they are particularly common in some areas of veterinary research.

Normally distributed response variables

So far, we have only considered analyses where the response variable is a **continuous** measurement.

We now have a brief look at two other common types of **response** variables.

The methods we use are largely similar to what we have looked at previously.

Data types in R

- numeric & integers

```
ToothGrowth$len[1:6]

## [1]  4.2 11.5  7.3  5.8  6.4 10.0

bacteria$week[1:6]

## [1]  0  2  4 11  0  2
```

- unordered factor with 2 levels

```
lung$sex[1:6]

## [1] female male   female female male   female
## Levels: female male
```

- (un/ordered) factor (with more than 2 levels)

```
chickwts$feed[1:6]

## [1] horsebean horsebean horsebean horsebean horsebean horsebean
## Levels: casein horsebean linseed meatmeal soybean sunflower
```

Generalised Linear Modelling

- random component: $\mathbf{E}(\mathbf{Y}) = \mu$
- systematic component: covariates: $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$ produce a linear predictor $\eta = \sum_1^p \mathbf{x}_j \beta_j$
- link between random and systematic components: $\mu = \eta$, e.g.
  - **Logistic Regression:**
    $\eta_i = g(\mu_i)$, $g(.)$ is logit function for binomial data
  - **Poisson Regression:**
    $\eta_i = g(\mu_i)$, $g(.)$ is exponential function for poisson data

Logistic Regression: unordered factor with 2 levels as response

E. g. a logistic regression with one covariate (`sex`) fitting a model to $\mathbb{P}(Y = \textit{TRUE}) = \pi$, e. g. presence of `respiratory symptoms` (experienced by the child over the previous 12 months) , gives

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \mathbf{x}_1$$

```
logreg.mod <- glm(respsymptoms ~ sex,
                  family = binomial, # description of the link function
                  data = lung)
summary(logreg.mod)
```

Logistic Regression: unordered factor with 2 levels as response

```
logreg.mod <- glm(respsymptoms ~ sex,
                  family = binomial, # description of the link function
                  data = lung)
summary(logreg.mod)
```

```
##
## Call:
## glm(formula = respsymptoms ~ sex, family = binomial, data = lung)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7440  -0.7440  -0.6915  -0.6915   1.7597
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1429     0.1276  -8.957   <2e-16 ***
## sexmale      -0.1663     0.1901  -0.875    0.382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 682.85  on 635  degrees of freedom
## Residual deviance: 682.08  on 634  degrees of freedom
## AIC: 686.08
##
## Number of Fisher Scoring iterations: 4
```

Logistic Regression: unordered factor with 2 levels as response

Interpretation of model coefficients

```
summary(logreg.mod)
```

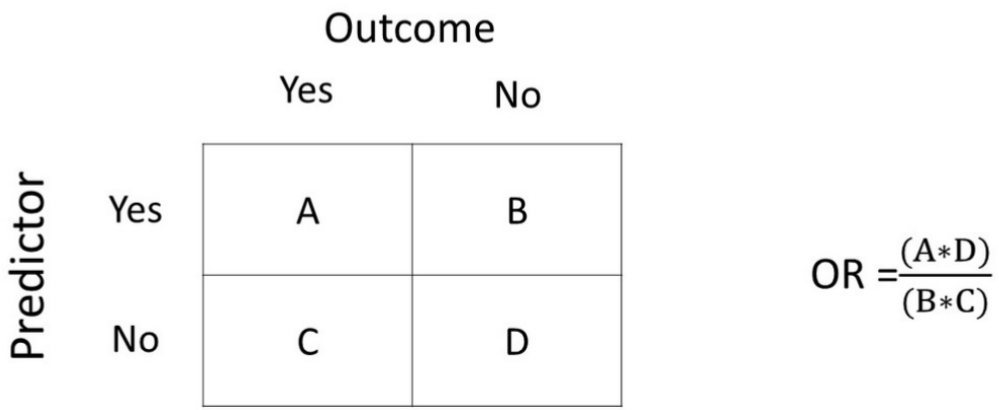
```
##
## Call:
## glm(formula = respsymptoms ~ sex, family = binomial, data = lung)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7440  -0.7440  -0.6915  -0.6915   1.7597
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1429     0.1276  -8.957   <2e-16 ***
## sexmale      -0.1663     0.1901  -0.875    0.382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 682.85  on 635  degrees of freedom
## Residual deviance: 682.08  on 634  degrees of freedom
## AIC: 686.08
##
## Number of Fisher Scoring iterations: 4
```

```
cbind(logodds = coef(logreg.mod), odds = exp(coef(logreg.mod)))
```

```
##                logodds      odds
## (Intercept) -1.1428851 0.3188976
## sexmale     -0.1662919 0.8467990
```

Odds ratios versus log odds ratios

- The logistic regression model in R outputs the log odds ratios.
- $\exp(\text{log odds ratio}) = \text{odds ratio}$
- An **odds ratio (OR)** is a measure of association between an exposure (here: `sex`) and an outcome (here: `respsymptoms`). The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.



Logistic Regression: unordered factor with 2 levels as response

Interpretation of model coefficients (cont'd)

```
cbind(logodds = coef(logreg.mod), odds = exp(coef(logreg.mod)))
```

```
##                logodds      odds
## (Intercept) -1.1428851 0.3188976
## sexmale     -0.1662919 0.8467990
```

The coefficient β_{sex} is the estimated amount by which the log odds of `respsymptoms` would in-/decrease if `sex` is male. The log odds of `respsymptoms` when `sex` is female is just above in the first row ($\beta_{\text{(Intercept)}}$, reference level of `sex`).

Logistic Regression: probability of success as response

- A researcher is examining beetle mortality after 5 hours of exposure to carbon disulphide, at various levels of concentration of the gas.

- Beetles were exposed to gaseous carbon disulphide at various concentrations (in mg/L) for five hours (Bliss, 1935) and the number of beetles killed were noted.

```
# install.packages("investr")
library("investr")
data(beetle)
colnames(beetle) <- c("Dose", "Num.Beetles", "Num.Killed")
str(beetle)

## 'data.frame': 8 obs. of  3 variables:
##  $ Dose       : num  1.69 1.72 1.76 1.78 1.81 ...
##  $ Num.Beetles: num  59 60 62 56 63 59 62 60
##  $ Num.Killed : num  6 13 18 28 52 53 61 60

beetle$Num.Surv <- beetle$Num.Beetles - beetle$Num.Killed
```

Logistic Regression: probability of success as response

```
logreg.mod <- glm(cbind(Num.Killed, Num.Surv) ~ Dose,
                  family = binomial, # description of the link function
                  data = beetle)
summary(logreg.mod)
```

```
##
## Call:
## glm(formula = cbind(Num.Killed, Num.Surv) ~ Dose, family = binomial,
##     data = beetle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5941  -0.3944   0.8329   1.2592   1.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.717      5.181  -11.72   <2e-16 ***
## Dose          34.270      2.912   11.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.202  on 7  degrees of freedom
## Residual deviance:  11.232  on 6  degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4
```

Logistic Regression: probability of success as response

Interpretation of model coefficients

```
summary(logreg.mod)
```

```
##
## Call:
## glm(formula = cbind(Num.Killed, Num.Surv) ~ Dose, family = binomial,
##     data = beetle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5941  -0.3944   0.8329   1.2592   1.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.717      5.181  -11.72   <2e-16 ***
## Dose          34.270      2.912   11.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.202  on 7  degrees of freedom
## Residual deviance:  11.232  on 6  degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4
```

```
cbind(logodds = coef(logreg.mod), odds = exp(coef(logreg.mod)))
```

```
##                logodds         odds
## (Intercept) -60.71745 4.273114e-27
## Dose         34.27033 7.645631e+14
```

Logistic Regression: probability of success as response

Interpretation of model coefficients (cont'd)

```
cbind(logodds = coef(logreg.mod), odds = exp(coef(logreg.mod)))
```

```
##                logodds         odds
## (Intercept) -60.71745 4.273114e-27
## Dose         34.27033 7.645631e+14
```

The coefficient β_{Dose} is the estimated amount by which the log odds of `being killed` would in-/decrease if `Dose` would increase by one unit. The log odds of `being killed` when `Dose` is 0 is just above in the first row $(\beta_{(\text{Intercept})})$.

Some things to remember

- Using generalised linear models (`glm`) is very similar to the previous models we have seen for normal data (`lm`) - things like residuals need to be checked for randomness, although this is more difficult with counts. Also the residuals being normally distributed is not relevant to these models.

- **As always explore the data first visually and with tables before doing any formal analyses.**