CANCER RESEARCH UK | Cambridge Institute

Together we are beating cancer

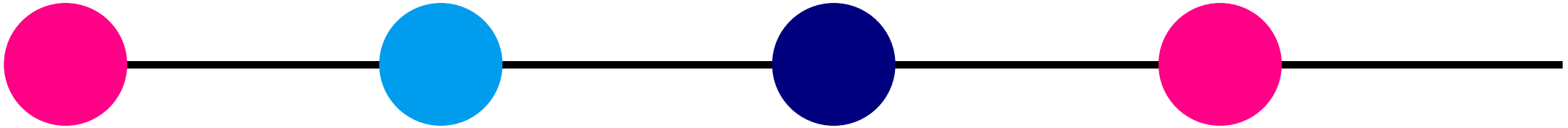
Luca Porcu & Chandra Chilamakuri (Bioinformatics core)

21st February 2025

Linear regression models

Fixed-effects models

Process flow



Stage 1

Effects and regression models

Stage 2

Analysis of variance (ANOVA)

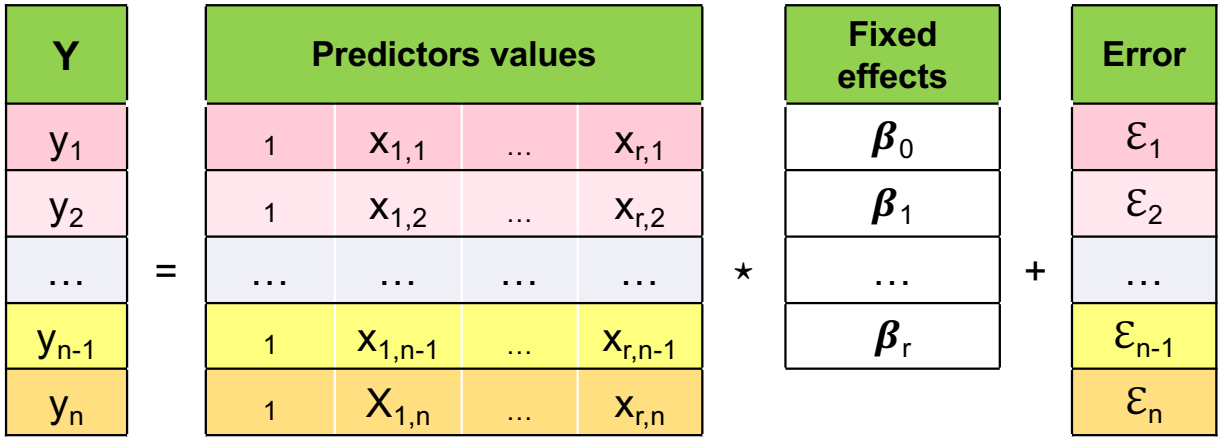
Stage 3

Simple regression model

Stage 4

Multivariable regression model

CANCER RESEARCH UK | Cambridge Institute



Multivariable regression model

Definition and classification

12.00 -12.20 am

Together we are beating cancer

Multivariable linear regression models

1. The unit k (e.g. mouse), k = 1,…,N
2. $\boldsymbol{\beta_0}$: intercept
3. $\boldsymbol{\beta_i}$: effect of predictor i, i = 1,…,r
4. $x_{i,k}$: predictor value of the unit k, i = 1,…,r; k=1,…N
5. ε_k: the *random* part of the model (i.e. error term of the model). It is a blanket characterization of the uniqueness of the k_{th} unit

Equation of the statistical model:

$$Y = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot \mathbf{x_{1,k}} + \ldots + \boldsymbol{\beta_r} \cdot \mathbf{x_{r,k}} + \varepsilon_k$$

Using language of matrices:

Y		Predictors values					Fixed effects		Error
y_1		1	$x_{1,1}$	…	$x_{r,1}$		$\boldsymbol{\beta}_0$		ε_1
y_2		1	$x_{1,2}$	…	$x_{r,2}$		$\boldsymbol{\beta}_1$		ε_2
…	=	…	…	…	…	⋆	…	+	…
y_{n-1}		1	$x_{1,n-1}$	…	$x_{r,n-1}$		$\boldsymbol{\beta}_r$		ε_{n-1}
y_n		1	$X_{1,n}$	…	$x_{r,n}$				ε_n

Assumptions of multivariable linear regression models are the following:

- The effect of each factor is additive on $\boldsymbol{\mu}$ (i.e. population mean) parameter
- ε_k is assumed to be independent of one another and normally distributed with mean = 0 and common standard deviation = $\boldsymbol{\sigma}$

Hypothesis testing in R: single predictor


```
> head(dSet)
```

IDmouse	Sex	Age (months)	Weight (grams)	Tumour Volume (mm^3)
Key1	F	8.9	93.1	160.8
Key2	F	9.3	95.1	132.8
Key3	F	11.0	83.8	128.1
Key4	F	5.0	82.2	151.9
Key5	M	2.9	83.7	150.5
Key6	M	5.5	114.2	154.0

```
> fittedModel = lm(tumourVolume ~ sex + age + weight, data=dSet)
> summary(fittedModel)
```

Output

Coefficients:

	Estimate	Std. Error	t value	**Pr(>|t|)**
(Intercept)	133.6318	22.5550	5.925	4.88e-08 ***
sex M	5.3824	5.5175	0.976	0.332
age	0.2296	0.8733	0.263	0.793
weight	0.1285	0.2403	0.535	0.594

Hypothesis testing in R: combined predictors

```
> library(multcomp)
> fittedModel = lm(tumourVolume ~ sex + age + weight, data=dSet)
```

```
> mComb = matrix(0, nrow=2, ncol=4)
> mComb[1,1] = 1; mComb[1,3] = -1; mComb[2,4] = 1
> tumVol.glht = glht(fittedModel, linfct = mComb)
> summary(tumVol.ghlt, test = adjusted("none"))
```

Output

```
Linear Hypotheses:
                    Estimate   Std. Error   t value   Pr(>|t|)
1 == 0              133.4021   22.7385      5.867     6.31e-08 ***
2 == 0              0.1285     0.2403       0.535     0.594
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- none method)
```

Hypotheses to test:

1) $\beta_{INTERCEPT} - \beta_{age} = 0$
2) $\beta_{weight} = 0$

Hypothesis testing in R: combined hypotheses

```
> fittedModel1 = lm(tumourVolume ~ sex + age + weight, data=dSet)
> fittedModel2 = lm(tumourVolume ~ sex, data=dSet)
> anova(fittedModel2, fittedModel1)
```

Output

Analysis of Variance Table

Model 1: tumourVolume ~ sex
Model 2: tumourVolume ~ sex + age + weight

	Res.Df	RSS	Df	Sum of Sq	F	**Pr(>F)**
1	98	67337				
2	96	67053	2	284.04	0.2033	0.8164

Hypothesis to test: $\beta_{age} = \beta_{weight} = 0$

Hypothesis testing in R: combined hypo. & pred.


```
> library(multcomp)
> fittedModel = lm(tumourVolume ~ sex + age + weight, data=dSet)
```

```
> mComb = matrix(0, nrow=2, ncol=4)
> mComb[1,1] = 1; mComb[1,3] = -1; mComb[2,4] = 1
> tumVol.glht = glht(fittedModel, linfct = mComb)
> summary(tumVol.ghlt, test = Ftest())
```

Output

```
Linear Hypotheses:
                Estimate
1 == 0          133.4021
2 == 0          0.1285
Global Test:
                F       DF1 DF2     Pr(>F)
1               126.4   2  96       1.285e-27
```

Hypothesis to test: $\beta_{INTERCEPT} - \beta_{age} = 0$ **and** $\beta_{weight} = 0$

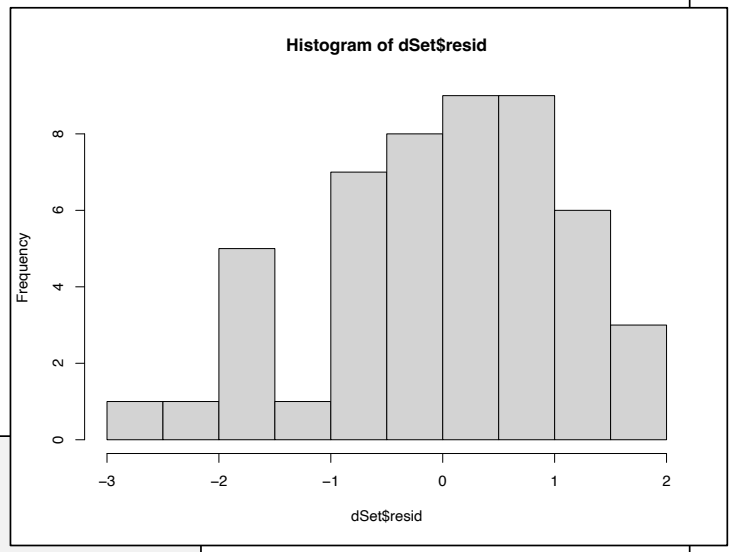
Diagnostics: residuals

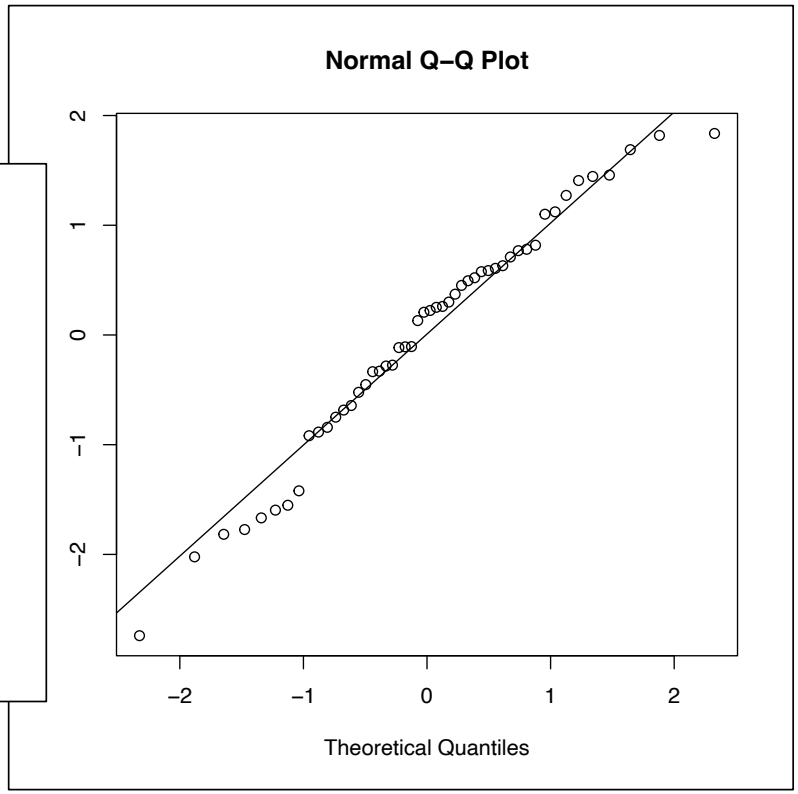
Shapiro-Wilk normality test

data: dSet$resid
W = 0.97324, p-value = 0.3119

Bartlett test of homogeneity of variances

data: resid by Predictor
Bartlett's K-squared = 1.5374, df = 1, p-value = 0.215





Assumptions of normality and homoscedasticity **must be satisfied** by residuals, overall and by each single level (e.g. residuals at female level) or combined levels (e.g. residuals at female level and weight below 90 grams)

Development of a reference model, tools

Residuals behaviour

Please, refer to slide n.9

Development of a reference model, tools

Sums of squared residuals (RSS)

The sum of the squared differences between observed and predicted values

```
> fittedModel = lm(tumourVolume ~ sex + age + weight, data=dSet)
> RSS = sum(resid(fittedModel)^2)
```

R-squared index

$$R^2 = 1 - \frac{RSS}{TSS}$$

Adjusted R-squared index

$$Adjusted\ R^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - p - 1}\right)$$

Higher values are better for both R^2 and adjusted R^2. Adjusted R^2 includes a penalty for the number of predictors introduced in the model so tends to favor more simple models with fewer predictors.

TSS = Total sum of squares (the sum of the squared differences between observed values and the mean of the observed values)
n = number of observations (data points)
p = number of predictors

Development of a reference model, tools

Information criteria: AIC and BIC indices

```
> fittedModel = lm(tumourVolume ~ sex + age + weight, data=dSet)
> AIC = AIC(fittedModel); BIC = BIC(fittedModel)
```

AIC index

$$2 \cdot K - 2 \cdot (\text{log-likelihood})$$

BIC index

$$K \cdot \log_e(n) - 2 \cdot (\text{log-likelihood})$$

Lower values are better for both AIC and BIC. AIC favors more complex models, while BIC includes a penalty for the number of parameters estimated so tends to favor more simple models with fewer parameters.

K = number of parameters
log-likelihood = maximised value of the log-likelihood function of the model
n = number of observations (data points)

Development of a reference model, tools

ANOVA and likelihood ratio tests for nested models

ANOVA test: please, refer to slide n.7

Likelihood ratio test:

```
> library(lmtest)
> fittedModel1 = lm(tumourVolume ~ sex + age + weight, data=dSet)
> fittedModel2 = lm(tumourVolume ~ sex, data=dSet)
> lrtest(fittedModel2, fittedModel1)
```

Output

```
Likelihood ratio test

Model 1: tumourVolume ~ sex
Model 2: tumourVolume ~ sex + age + weight
  #Df  LogLik  Df  Chisq  Pr(>Chisq)
1   3 -467.51
2   5 -467.30   2 0.4227      0.8095
```

Hypothesis to test: $\beta_{age} = \beta_{weight} = 0$

http://bioinformatics-core-shared-training.github.io/IntroductionToStats/practical.html