Linear Regression

APPLIED STATISTICS - FCUP

Rita Gaio
argaio@fc.up.pt

2020

Linear Regression Model

Goal: modelling the relationship between a **continuous** random variable Y and a set of *explanatory* (observed; not random) variables $X_1, \ldots, X_p$ of any type.

Possible purposes are:

- evaluation of the effect of $X_1, \ldots, X_p$ on Y
- forecasting - prediction of Y values from known values of $X_1, \ldots, X_p$

Examples:

- explain systolic blood pressure in the adult portuguese population through body mass index, age, sex, hypertension disease (yes/no) and other comorbilities (yes/no).
- understand how gross domestic product (GDP) is impacted by changes in unemployment and inflation
- forecast sales for a company when it is known that the company's sales go up and down depending on changes in GDP, and some well identified political, social and legal factors.

Linear Regression Model

Data: $\{(y_i; x_{1i}, x_{2i}, ..., x_{pi})\}_{i=1,...,n}$ experimental units: $i = 1, \ldots, n$

Assumptions:

- $y_1, \ldots, y_n$ are independent random variables
- $y_1|x_i, \ldots, y_n|x_i$ are homocedastic (have the same variance), for any $x_i = (x_{1i}, \ldots, x_{ni})$

Model:

$$y_i|x_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + u_i, \qquad u_i \sim N(0, \sigma^2),\ i = 1, \ldots, n$$

or, equivalently,

$$y_i|x_i \sim N(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}, \sigma^2), \quad i = 1, \ldots, n$$

Parameters to be estimated: $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$

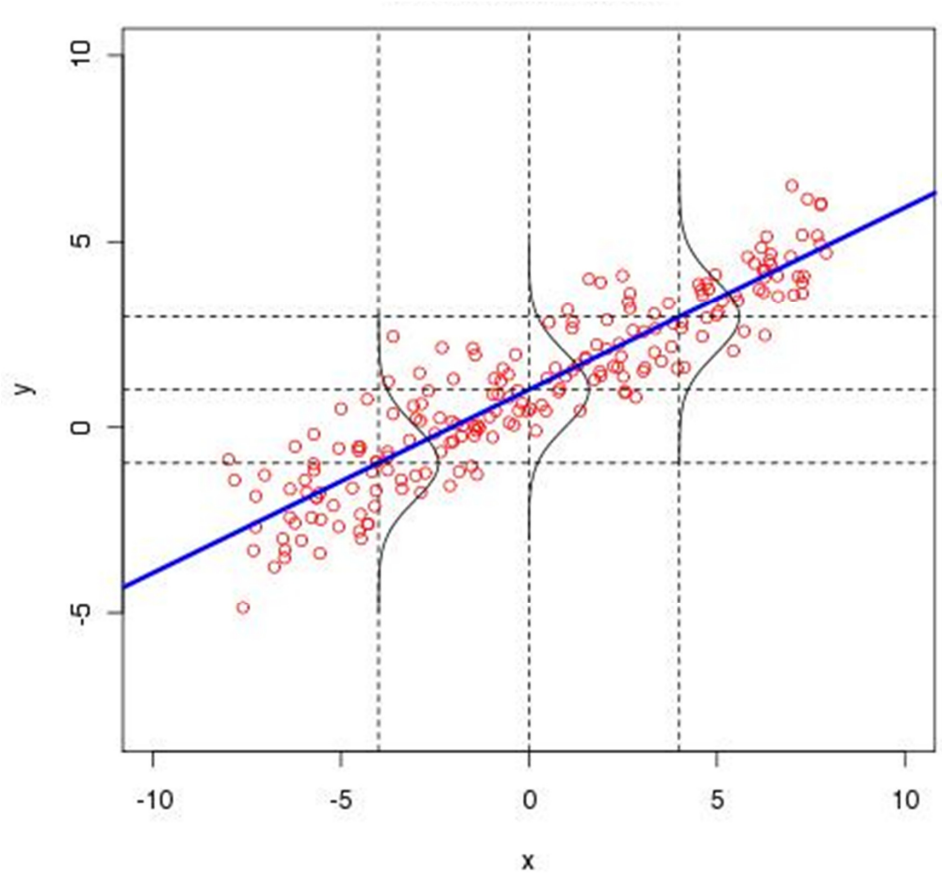
Neither the β_i's, nor σ^2 or the u_i's can ever be known.

Linear Regression Model

Simple linear regression model (a single predictor):

$$y_i|x_i = \beta_0 + \beta_1 x_i + u_i, \qquad u_i \sim N(0, \sigma^2),\ i = 1, \ldots, n$$

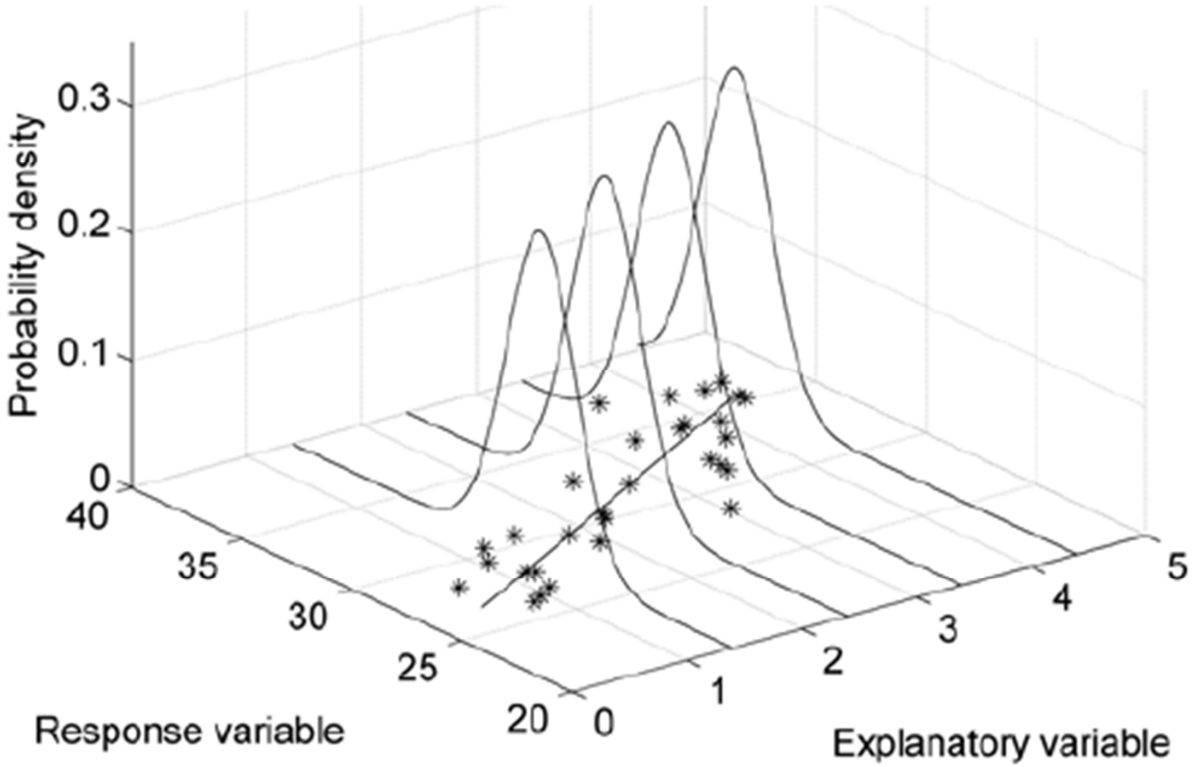
Multiple regression model: more than one predictor



Linear Regression Model

$$y_i|x_i = \beta_0 + \beta_1 x_i + u_i, \qquad u_i \sim N(0, \sigma^2),\ i = 1, \ldots, n$$

The **linear predictor** needs to be a linear function of the β_j's.



Linear Regression Model - terminology

$$y_i|x_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + u_i, \qquad u_i \sim N(0, \sigma^2),\ i = 1, \ldots, n$$

- y_i: response or dependent variable
- $X_1, \ldots, X_p$: explanatory or independent variables, or predictors.
- u_i: error
- $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$: regression (population) parameters (unknown)

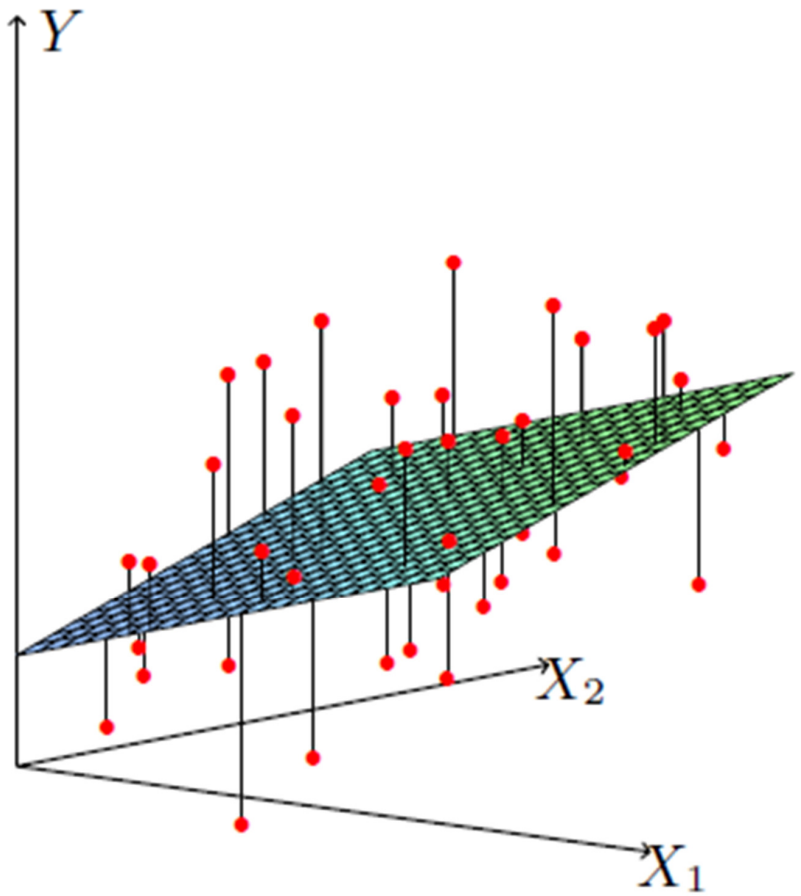
Once the model is fitted:

- $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \cdots + \widehat{\beta}_p x_{pi}$: fitted-value for observation i
- $\widehat{u}_i = y_i - \widehat{y}_i$: residual for observation i
- $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$: regression coefficients (estimates of the unknown regression parameters)

Remark: Here, lower-case letters can represent random variables. We follow matrix notation (lower-case letters for vectors; upper-case letters for matrices).

Linear Regression Model

Fitting of a linear regression model with 2 explanatory variables:



Linear Regression Model

Using matrix notation, the previous model

$$y_i|x_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + u_i, \qquad u_i \sim N(0, \sigma^2), i = 1, \ldots, n$$

can be written as

$$\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ \vdots \\ u_n \end{pmatrix}$$

i.e., $(y \in \mathbb{R}^n,\ X \in \mathbb{R}^{n \times (p+1)},\ \beta \in \mathbb{R}^{p+1},\ u \in \mathbb{R}^n)$

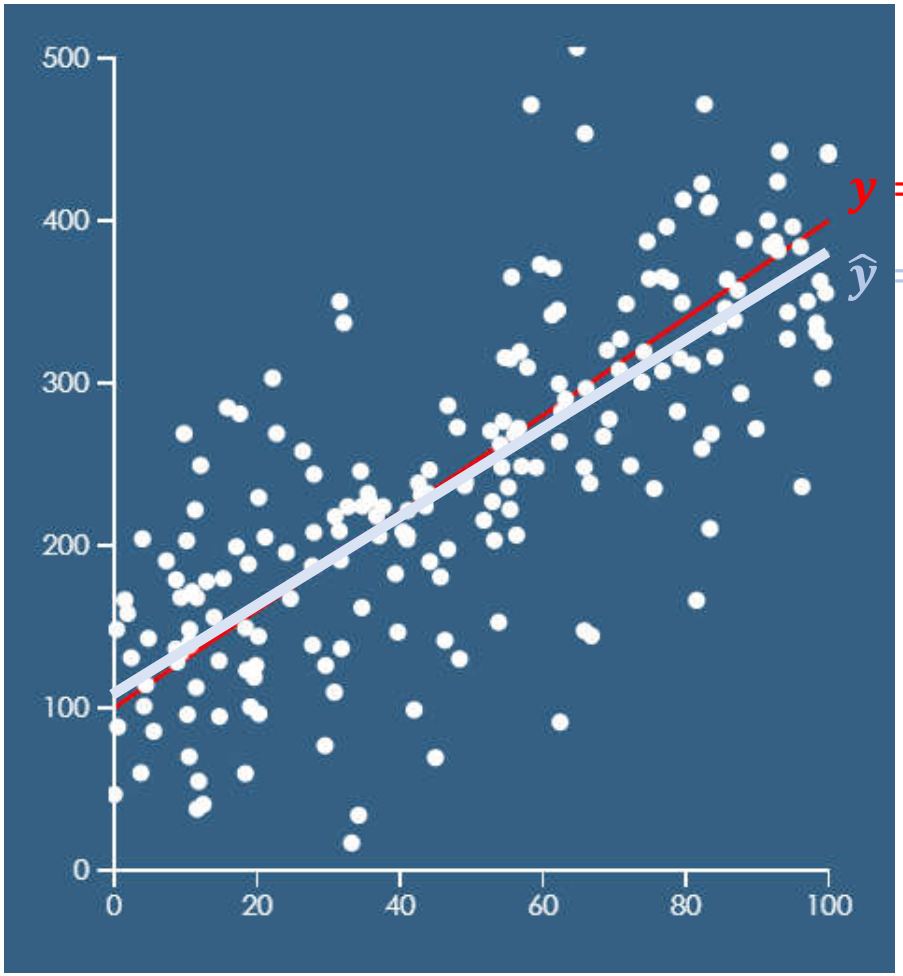
$$y = X\beta + u$$

with the assumption

$$u \sim MVN(0, \sigma^2 Id).$$

The matrix X is called the **design matrix**.

Interpretation of regression parameters - simple regression



$$y = \beta_0 + \beta_1 x$$

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

Interpretation of regression parameters - simple regression

$$y_i = \beta_0 + \beta_1 x_i + u_i, \qquad u_i \sim N(0, \sigma^2)$$

- β_0: constant term
  Expected response whenever x takes on the value 0.

- β_1: true effect of x on y
  How much the response is expected to increase ($\beta_1 > 0$) or decrease ($\beta_1 < 0$) for every unit increase in X.

BUT β_0 and β_1 will never be known!

Interpretation of regression coefficients - simple regression

$$\boxed{\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i}$$

- $\widehat{\beta}_0$: intercept or constant term
  Predicted response whenever x takes on the value 0.
  Whenever nonsense, the explanatory variable can be centered, using its sample mean $(x_i - \overline{x})$.

- $\widehat{\beta}_1$: effect of x on y
  How much the predicted response increases ($\widehat{\beta}_1 > 0$) or decreases ($\widehat{\beta}_1 < 0$) for every unit increase in x, in the data at hand.

Interpretation of regression coefficients

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \cdots + \widehat{\beta}_p x_{pi}$$

- $\widehat{\beta}_0$: predicted response whenever all predictors take on the value 0, given the data at hand
- $\widehat{\beta}_j$, $j = 1, \ldots, p$: marginal/adjusted effect of variable x_j on y
  As the model describes the regression of y on $(x_1, \ldots, x_p)$ **jointly**, $\widehat{\beta}_j$ accounts for the contribution of the other predictors, that is, it is **adjusted** or **controlled for** those predictors.

  $\widehat{\beta}_j$: difference in the predicted value of y for each one-unit difference in x_j, *holding the other explanatory variables constant*.[1]
  The latter part means the marginal effect is obtained *after removing the linear effect of the other predictors from both x_j and y*.

[1] the predictors may be inherently related, and holding some of them constant while varying the others may not be possible

Interpretation of regression coefficients - example

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i}$$

$\widehat{\beta}_1$ represents the contribution of x_1 to the predicted response **after both variables have been linearly adjusted for** x_2

(1) y is regressed on x_2
The residuals, $\widehat{u}_{y,x_2}$, correspond to the part of y that is not linearly related to x_2

(2) x_1 is regressed on x_2
The residuals, $\widehat{u}_{x_1,x_2}$, correspond to the part of x_1 that is not linearly related to x_2

(3) u_{y,x_2} is regressed on u_{x_1,x_2}

The effect of u_{x_1,x_2} on u_{y,x_2} is the (marginal) regression coefficient $\widehat{\beta}_1$.

Interpretation of regression coefficients - example

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i}$$

Interpretação de $\widehat{\beta}_1$:

Nos dados em causa, um aumento de 1 unidade em x_1 corresponde a uma alteração de $\widehat{\beta}_1$ na resposta prevista, ...

- ... mantendo x_2 constante
- ... depois de y e x_1 terem sido ajustados para x_2
- ... depois de remover o efeito linear de x_2 sobre x_1 e y:
  - $y \sim x_2 \longrightarrow u_{y,x_2}$
  - $x_1 \sim x_2 \longrightarrow u_{x_1,x_2}$
  - $u_{y,x_2} \sim u_{x_1,x_2} \longrightarrow \widehat{\beta}_1$.

Interpretation of regression coefficients

$$y_i|x_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + u_i, \qquad u_i \sim N(0, \sigma^2)$$

Size and sign of regression coefficients:

- size of $\widehat{\beta}_j$: size of the adjusted effect that X_j has on Y
- sign of $\widehat{\beta}_j$: direction of the adjusted effect
  - $\widehat{\beta}_j > 0$: positive association between X_j and Y
  - $\widehat{\beta}_j < 0$: negative association between X_j and Y

Note: the above association is not necessarily causal.

Example

```
library(foreign)
base1 <- read.spss("Cystfibr.sav", to.data.frame=TRUE)
```

```
## re-encoding from CP1252
```

```
mod1 <- lm(tlc ~ age+bmp, data=base1)
```

Example

```
summary(mod1)
```

```
##
## Call:
## lm(formula = tlc ~ age + bmp, data = base1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.0796 -10.6015   0.8405   7.7994  30.5473
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.9913    20.5443   7.642 1.25e-07 ***
## age          -1.2971     0.6638  -1.954   0.0635 .
## bmp          -0.3093     0.2797  -1.106   0.2808
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Example

- the intercept has no direct interpretation as age and bmp cannot be 0
- $\widehat{\beta}_{age} \approx -1.3$:
  In the data at hand, any 1-year increase in age predicts a decrease of 1.3 in total lung capacity, after adjustment for body mass percentage

  Question: what if the interest is on a 10-year increase?
- $\widehat{\beta}_{bmp} \approx -0.3$:
  In the data at hand, any 1-unit increase in bmp predicts a decrease of 0.3 in total lung capacity, after adjustment for age (or for individuals with the same age)

About the data:

- total lung capacity: volume of air present in the chest after full inspiration (measured in liters in the metric system)
- functional residual capacity: volume of air left in the lungs at the end of a quiet expiration
- residual volume: amount remaining at the end of a maximal expiration

Standardized regression coefficients

Standardized regression coefficients: obtained from a regression analysis where the response and explanatory variables have been standardized (mean 0 and variance 1)

$$y_i^*|x_i = \beta_0^* + \beta_1^* x_{1i}^* + \cdots + \beta_p^* x_{pi}^* + u_i, \qquad u_i \sim N(0, \sigma^2)$$

$\widehat{\beta}_j^*$, $j = 1, \ldots, p$: adjusted effect of variable X_j^* on y^*

How much the predicted response increases ($\widehat{\beta}_j > 0$) or decreases ($\widehat{\beta}_j < 0$), in terms of standard deviations, when X_j increases by 1 standard deviation, *holding the other explanatory variables constant / adjusting for the other explanatory variables*.

Standardized regression coefficients are useful for comparing effects

Standardized regression coefficients

Instructions in **R**:

```
library(QuantPsyc)
lm.beta(model)
```

where *model* is an object of class *lm*, obtained from an *lm* instruction.

Standardized regression coefficients

```
library(QuantPsyc)
lm.beta(mod1)
```

```
##         age         bmp
## -0.3867119 -0.2188176
```

Age is seen to have a greater impact on tlc than bmp.

Moreover, in the data at hand:

- any increase of 1 standard deviation in age predicts a decrease of approx. 0.4 standard deviations in tlc, adjusting for bmp
- any increase of 1 standard deviation in bmp predicts a decrease of approx. 0.2 standard deviations in tlc, adjusting for age.

Parameter Estimation

$$y_i|x_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + u_i, \qquad u_i \sim N(0, \sigma^2)$$

$$y = X\beta + u, \qquad u \sim MVN(0, \sigma^2 Id)$$

parameters to be estimated: $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ and σ^2

Possible methodologies for finding β:

- **least squares**: minimizes the **residual sum of squares** function

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= (y - \widehat{y})^t (y - \widehat{y}) \\ &= (y - X\beta)^t (y - X\beta) \end{aligned}$$

- **maximum likelihood**: maximizes the (log)**likelihood function**

$$L\left(\beta, \sigma^2 | (y_i, x_i)_i\right) = \prod_{i=1}^{n} f(y_i|x_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y_i - X_{i.}\beta)^2}{2\sigma^2}\right)$$

Parameter Estimation - Least Squares

$$RSS(\beta) = (y - X\beta)^t (y - X\beta)$$

Using differential matrix calculus (rules next page),

$$\begin{aligned} \frac{\partial}{\partial \beta} RSS(\beta) &= 2(y - X\beta)^t \frac{\partial}{\partial \beta}(y - X\beta) \\ &= -2(y^t - \beta^t X^t)X \\ &= -2(y^t X - \beta^t X^t X). \end{aligned}$$

Solving $\frac{\partial}{\partial \beta} RSS(\beta) = 0$ gives the $p+1$ **normal equations**:

$$X^t X \widehat{\beta} - X^t y = 0.$$

If $X^t X$ is invertible, ie, columns $X_1, \ldots, X_p$ are not linearly dependent,

$$\boxed{\widehat{\beta}_{OLS} = (X^t X)^{-1} X^t y}$$

- $\widehat{\beta}$ is a global minimum since RSS has no upper bound.
- when $X^t X$ is singular, the equations can be solved using *generalized inverses*.

Differential Matrix Calculus

For any $x \in \mathbb{R}^n$, $y \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ (not depending on x):

- $\frac{\partial}{\partial z}(Ax) = A\frac{\partial x}{\partial z}$, for any vector $z \in \mathbb{R}^n$
- $\frac{\partial}{\partial x}(y^t Ax) = y^t A$
- $\frac{\partial}{\partial x}(x^t Ax) = x^t(A + A^t)$.
  In particular, if A is symmetric then

$$\frac{\partial}{\partial x}(x^t Ax) = 2x^t A.$$

- $\frac{\partial}{\partial z}(y^t x) = x^t \frac{\partial y}{\partial z} + y^t \frac{\partial x}{\partial z}$.

Parameter Estimation - Maximum Likelihood

$$L\left(\beta, \sigma^2 | (y_i, x_i)_i\right) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y_i - X_{i.}\beta)^2}{2\sigma^2}\right)$$

Since *log* is an increasing function, maximizing L is equivalent to maximizing $\ell = log(L)$.

$$\begin{aligned} \ell\left(\beta, \sigma^2 | (y_i, x_i)_i\right) &= \log\left(L\left(\beta, \sigma^2 | (y_i, x_i)_i\right)\right) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_{i.}\beta)^2 \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} RSS(\beta) \end{aligned}$$

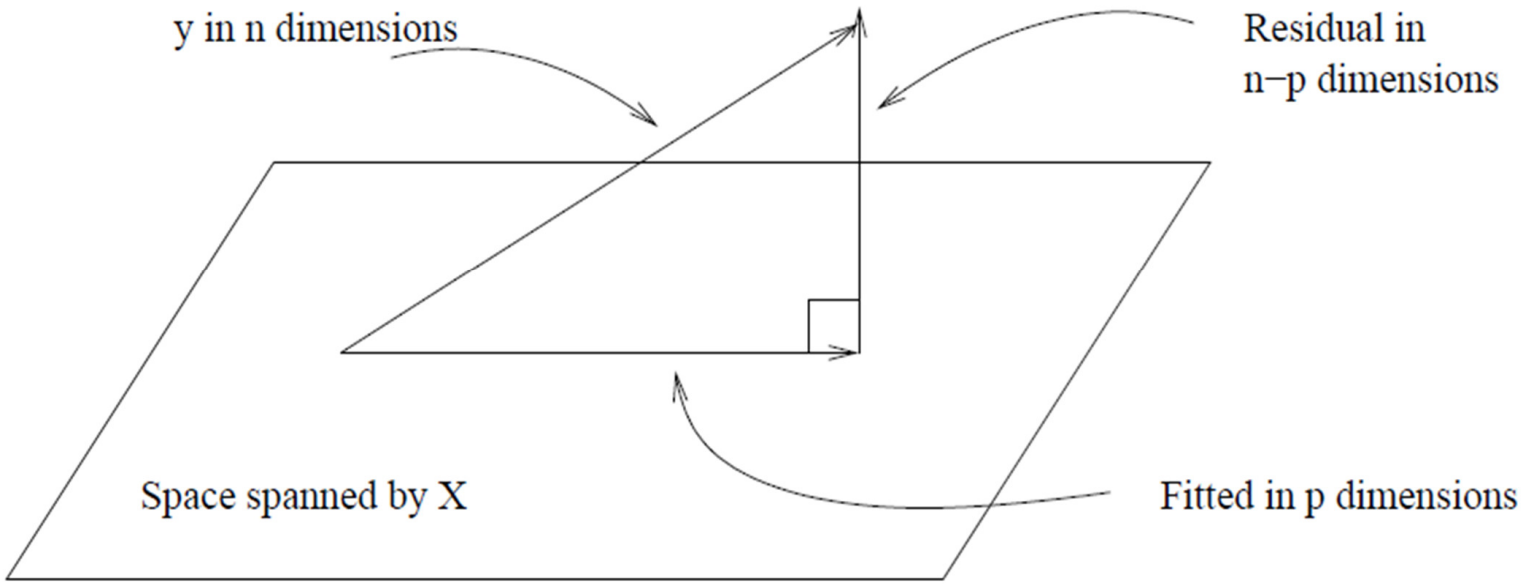
hence maximizing L is equivalent to minimizing RSS.

Result: estimating β by the ordinary least squares method provides the same result as estimating β by maximum likelihood while keeping σ^2 fixed.

$$\widehat{\beta}_{OLS} = \widehat{\beta}_{ML}.$$

Geometric interpretation of linear regression

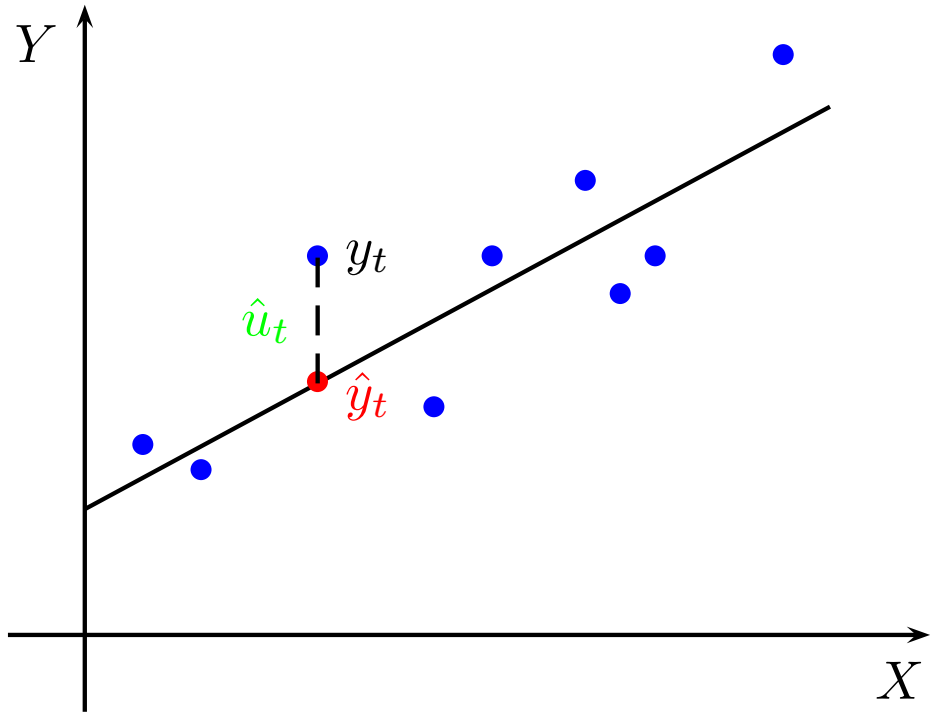
Geometrically, $\widehat{y}$ corresponds to the orthogonal projection of $y \in \mathbb{R}^n$ on the subspace of dimension $p+1$ generated by the columns of X



Residuals

- $u = y - X\beta$ model **errors**
- $\widehat{u} = y - X\widehat{\beta}$ **residuals**

Note that $\widehat{u}$ may be numeric or a random variable, depending on the context.



Residuals

Exercise: show that, in a linear regression model:

(a) $E(\widehat{u}) = 0$

(b) $X^t\widehat{u} = 0$ - the residuals are orthogonal to the design matrix

(c) the residuals are orthogonal to the fitted values

(d) if the model has a nonzero intercept, then $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$, i.e., the sum of the observed responses is equal to the sum of the fitted values. This is actually one reason for considering squares in the RSS function.

Gauss-Markov Theorem

Least squares gives good estimates and predictions if certain conditions are met.

Gauss-Markov Theorem: In a linear regression model with zero centered, non-correlated and homocedastic errors, the (ordinary) least squares estimator $\widehat{\beta}_{OLS}$ is BLUE - Best Linear Unbiased Estimator.

Remark: In Gauss-Markov theorem, the errors do not need to be gaussian.

The theorem means that:

(1) $E(\widehat{\beta}_{OLS}) = \beta$ (i.e. $\widehat{\beta}$ is centered or unbiased)

(2) $\widehat{\beta}_{OLS}$ has the lowest sampling variance within the class of linear unbiased estimators, i.e., $Cov(\tilde{\beta}) - Cov(\widehat{\beta}_{OLS})$ is a positive semi-definite matrix for any linear unbiased estimator $\tilde{\beta}$ of β.

Gauss-Markov Theorem

Exercise: Prove that the assumptions in the Gauss-Markov Theorem are equivalent to

$$\begin{aligned} E(u_i) &= 0 \\ E(u_i^2) &= \sigma^2 \\ E(u_i u_j) &= 0, \quad i \neq j \end{aligned}$$

which can also be written as

$$\begin{aligned} E(u) &= 0 \\ E(uu^t) &= \sigma^2 Id \end{aligned}$$

where $u = (u_1, \ldots, u_n)^t$.

Properties of $\widehat{\beta}_{OLS}$

It can be worked out from the formula of $\widehat{\beta}$ that:

(a) $Cov(\widehat{\beta})^2 = \sigma^2 (X^t X)^{-1}$

(b) $se(\widehat{\beta}_i) = \sigma \sqrt{(X^t X)^{-1}_{ii}}$.
Here $se()$ denotes the standard deviation. It is a common terminology in linear regression.

[2] variance-covariance matrix; the result uses the fact that $Cov(Ay) = A\, Cov(y)\, A^t$, for any $y \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$

Estimation of σ^2

$$y_i|x_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + u_i, \qquad u_i \sim N(0, \sigma^2)$$

$$y = X\beta + u, \qquad u \sim MVN(0, \sigma^2 Id)$$

Up to now, we assumed σ^2 known. In general, it is unknown.

An estimator for σ^2 may be obtained by **maximum likelihood**, using $\widehat{\beta}_{ML}$ $(= \widehat{\beta}_{OLS})$ from before. It can be shown (exercise) that

$$\frac{\partial}{\partial \sigma^2} \ell(\sigma^2 \,|\, \widehat{\beta}, (y_i, x_i)_i) = \frac{1}{2\sigma^2} \left(-n + \frac{RSS(\widehat{\beta})}{\sigma^2} \right)$$

that is equal to 0 for

$$\widehat{\sigma}^2_{ML} = \frac{RSS(\widehat{\beta})}{n}.$$

This is the maximum likelihood estimator of σ^2.

Estimation of σ^2

However $\widehat{\sigma}^2_{ML}$ is a <u>biased</u> estimator:

$$E(\widehat{\sigma}^2_{ML}) \neq \sigma^2.$$

An <u>unbiased</u> estimator for σ^2, correcting the degrees of freedom to the fact that $\widehat{\beta} \in \mathbb{R}^{p+1}$ has been estimated, is

$$\overline{\sigma}^2 = \frac{RSS(\widehat{\beta})}{n-(p+1)}.$$

Similarly to what is done for the sample variance, it can be shown that

$$(n-(p+1))\frac{\overline{\sigma}^2}{\sigma^2} \sim \chi^2(n-(p+1)).$$

Question: How can we use this distribution together with the formula for $\widehat{\beta}_{OLS}$ in order to apply confidence intervals and hypothesis tests theory on β?

Hypothesis Tests on β_j - Wald Test

Motivation: assume that we have the price of several houses together with some of their caracteristics: number of rooms, lot size, floor area, garage (yes/no) and storm windows (yes/no).

Some questions of interest:

(a) Is the selling price affected by the number of rooms in a house?

(b) Suppose the realtor says that adding a garage will add 5000 eur to the selling price of the house. Can this be true?

(c) Do lot size and floor area affect the price equally?

(d) Does either lot size or floor area have any effect on prices?

(e) Can it be true that storm windows add 6000 eur and agarage adds 4000 eur to the price of a house?

Hypothesis Tests on β_j - Wald Test

Up to now, we have essentially dealt with parameters estimation and the model errors have been assumed to be zero-centered, non-correlated and homocedastic.

In addition, we now suppose that the errors are gaussian:

$$u_i \sim N(0, \sigma^2), i = 1, \ldots, n.$$

The formulae

$$\widehat{\beta} = (X^t X)^{-1} X^t y \quad \text{and} \quad Cov(\widehat{\beta}) = \sigma^2 (X^t X)^{-1}$$

together with the fact that linear combinations of normal distributions are also normal, give

$$\widehat{\beta} \sim N(\beta, \sigma^2 (X^t X)^{-1}).$$

This formula assumes σ^2 is known...

Hypothesis Tests on β_j - Wald Test

Whenever σ^2 is unknown, $\overline{\sigma}^2$ is used and the distribution is asymptotic:

$$\widehat{\beta} \overset{a}{\sim} N(\beta, \overline{\sigma}^2 (X^t X)^{-1}).$$

Remark: If the errors are not normal but the *sample size is sufficiently large to compensate deviations from normality* (the Central Limit Theorem is more sensitive to extreme distributions in small samples), the above convergence still holds.

The formula $\widehat{\beta} = (X^t X)^{-1} X^t y$ shows that each coefficient is a weighted average of the Y values with weights that depend in a complicated way on the predictors X. That is, we can write each coefficient as

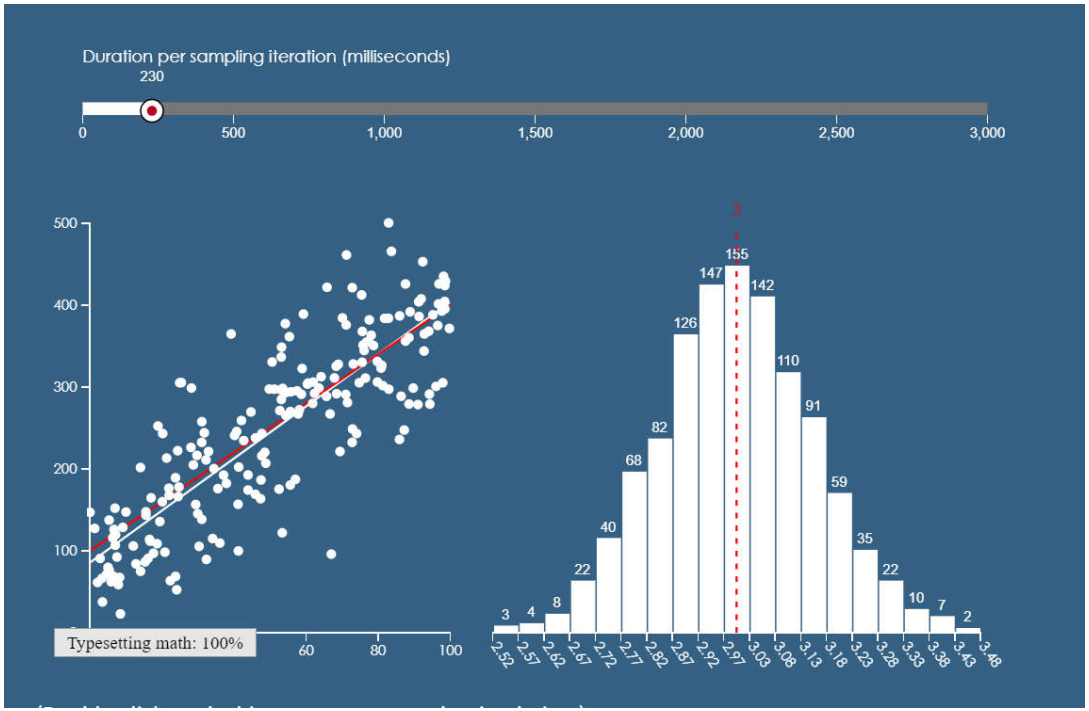
$$\widehat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} \omega_i Y_i$$

This is an average of variables $\omega_i Y_i$ that have different distributions, therefore the Central Limit Theorem still applies.

Hypothesis Tests on β_j - Wald Test

Simulation Study: https://www.econometrics-with-r.org/4-5-tsdotoe.html

The interactive simulation continuously generates random samples (x_i, y_i) of 200 observations where $E(Y|X) = 100 + 3X$, estimates a simple regression model, stores the estimate of the slope β_1, and visualizes the distribution of the r.v. $\widehat{\beta}_1$, using a histogram.



Hypothesis Tests on β_j - Wald Test

Join the following facts:

(a) if U and V are independent random variables, $U \sim N(0,1)$ and $V \sim \chi^2(n-1)$, then

$$T = U/\sqrt{V/(n-1)} \sim t(n-1).$$

(b) $\widehat{\beta}|\sigma^2 \sim MVN(\beta, Cov(\widehat{\beta}))$; in particular,

$$\frac{\widehat{\beta}_j - \beta_i}{se(\widehat{\beta}_j)} \sim N(0,1) \qquad \text{with } se(\widehat{\beta}_j) = \sigma\sqrt{(X^tX)^{-1}_{jj}}$$

(c) if σ^2 is unknown and n is *large* or the errors are normally distributed, $\widehat{\beta} \overset{a}{\sim} N(\beta, Cov(\widehat{\beta}))$; hence

$$\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \overset{a}{\sim} N(0,1) \qquad \text{with } se(\widehat{\beta}_j) = \overline{\sigma}\sqrt{(X^tX)^{-1}_{jj}}$$

Hypothesis Tests on β_j - Wald Test

(d) $(n-(p+1))\frac{\overline{\sigma}^2}{\sigma^2} \sim \chi^2(n-(p+1))$.

(e) each $\widehat{\beta}_j$ is independent from $\overline{\sigma}^2$.

We obtain

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \overset{a}{\sim} t(n-(p+1)), \qquad j = 0, 1, \ldots, p$$

where $se(\widehat{\beta}_j) = \overline{\sigma}\sqrt{(X^tX)^{-1}_{jj}}$ is the standard deviation of $\widehat{\beta}_j$.

Hypothesis Tests on β_j - Wald Test

It is common practice to test, for each $j = 0, \ldots, p$,

$$H_0 : \beta_j = 0.$$

H_0 means that, in the population:

- $j = 0$: the intercept is null
- $j = 1, \ldots, p$: variable X_j has no adjusted effect on y

Requirements: $u_i \sim N(0, \sigma^2 Id)$ (n is *sufficiently large* may replace normality...)

Test Statistic: $T_j = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \overset{a}{\sim} t(n - (p+1))$

Decision: to reject H_0 at an α level if $|t_j| \geq t_{1-\alpha/2}(n - (p+1))$

Hypothesis Tests on β_j - Wald Test

Note:

- the removal of one predictor from a regression model may influence the statistical significance of the remaining explanatory variables
- it is advisable to remove each non-significant explanatory variable at a time
- if there is a reason for it, non-significant explanatory variable may remain in the model.

Hypothesis Tests on β_j - Wald Test

```
mod2 <- lm(tlc ~ age, data=base1)
```

Hypothesis Tests on β_j - Wald Test

```
summary(mod2)
```

```
##
## Call:
## lm(formula = tlc ~ age, data = base1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.8841  -7.0328  -0.4786  10.0957  29.6902
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 136.7959     9.4514  14.474 4.82e-13 ***
## age          -1.5743     0.6175  -2.549   0.0179 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.31 on 23 degrees of freedom
## Multiple R-squared:  0.2203, Adjusted R-squared:  0.1864
## F-statistic: 6.499 on 1 and 23 DF,  p-value: 0.01793
```

Exercise

To decide whether a company is discriminating against women, the following data were collected from the company's records:

- Salary: annual salary in thousands of dollars
- Qualification: index of employee qualification
- Sex: 1, if the employee is a man; 0, if the employee is a woman.

Two linear models were fit to the data and the regression outputs are shown in the next page. Suppose that the usual regression assumptions hold.

(a) Are men paid more than equally qualified women?

(b) Are men less qualified than equally paid women?

(c) Do you detect any inconsistency in the above results? Explain.

(d) Which model would you advocate if you were the company's defense lawyer? Explain.

Example

Model 1: Response variable is Salary

Variable	Coefficient	s.e.	t-test	p-value
Constant	20009.5	0.8244	2427 1	<0.0001
Qualification	0.935253	0.0500	18.7	<0.0001
Sex	0.224337	0.4681	0.479	0.6329

Model 2: Response variable is Qualification

Variable	Coefficient	s.e.	t-test	p-value
Constant	-16744.4	896.4	- 18.7	<0.0001
Sex	0.850979	0.4349	1.96	0.0532
Salary	0.836991	0.0448	18.7	<0.0001

Confidence Interval for β_j

From

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \overset{a}{\sim} t(n-(p+1)), \qquad j = 0, 1, \ldots, p$$

a $100(1-\alpha)\%$ confidence interval for β_j can be deducted. We obtain

$$\widehat{\beta}_j \pm se(\widehat{\beta}_j) t_{1-\alpha/2}(n-(p+1))$$

Instructions in **R**:

```
confint(model, parm, level=0.95,...)
```

where

- model: object corresponding to the estimated model
- parm: set of regression parameters for which confidence intervals are required (by default, confidence intervals are computed for all parameters)
- level: confidence level.

Confidence Interval for β_j

```
confint(mod2)
```

```
##                   2.5 %      97.5 %
## (Intercept) 117.244152 156.3476686
## age          -2.851787  -0.2968197
```

```
confint(mod2,2)
```

```
##         2.5 %     97.5 %
## age -2.851787 -0.2968197
```

Confidence Interval for the Mean Response

Predictions should only be considered for values of the explanatory variables that are similar to those observed in the sample.

Need to distinguish between:

- predictions of the future mean response - **confidence interval for the mean response**

  "What would a house with characteristics x sell for, on average?"
- predictions of future observations - **prediction intervals**

  "What would a house with characteristics x sell for?"

Let $x_0 = (x_{01}, \ldots, x_{0p})^t$ be a vector of predictor values[3]. Then:

- the mean of the response is estimated by $\widehat{y}_0 = x_0^t \widehat{\beta}$
- $Var(\widehat{y}_0) = x_0^t (X^t X)^{-1} x_0 \sigma^2$ (exercise)

hence the confidence interval for the mean response for given x_0 is

$$\widehat{y}_0 \pm t_{1-\alpha/2}(n-(p+1))\overline{\sigma}\sqrt{x_0^t (X^t X)^{-1} x_0}.$$

[3] the convention is that vectors are columns

Prediction Interval

Prediction interval: interval estimate of the response y for a given value of the explanatory variables

Let $x_0 = (x_{01}, \ldots, x_{0p})^t$ be a new vector of values for explanatory variables.

$$Var(\widehat{y}_0 + u) = (1 + x_0^t (X^t X)^{-1} x_0)\sigma^2$$

hence the prediction interval for the response for given x_0 is

$$\widehat{y}_0 \pm t_{1-\alpha/2}(n - (p+1))\overline{\sigma}\sqrt{1 + x_0^t (X^t X)^{-1} x_0}.$$

Prediction Interval

Instructions in **R**:

```
predict(model, newdata, interval="predict", level, ...)
```

where

- model: object of class "lm" representing the regression model
- newdata: data frame in which to look for variables with which to predict
- level: confidence level

The instruction produces a vector of predictions, or matrix of predictions and bounds with column names 'fit', 'lwr', and 'upr' if interval is set.

Note: Decision makers should use more than just a single prediction to make rational choices.

Prediction Interval

```
# single predicted value
predict(mod2, newdata=data.frame(age=c(35)))
```

```
##        1
## 81.6953
```

```
# prediction interval
predict(mod2, newdata=data.frame(age=c(35)), interval="p")
```

```
##        fit      lwr      upr
## 1 81.6953 40.1058 123.2848
```

The 95% prediction interval of the total lung capacity for the age of 35 y.o. is between 40.1 and 123.3 minutes.

Comparison with the Null Model

Null Model, M_0:

$$y = \beta_0 + u, \quad u \sim MVN(0, \sigma^2 Id)$$

ie, for every $j = 1, \ldots, p$,

$$y_i = \beta_0 + u_i, \quad u_i \sim N(0, \sigma^2).$$

- The null model has no explanatory variables
- In simple linear regression, the regression line is horizontal (ie, X has no effect on y)
- The null model is the poorest model we can fit
- As expected, $\widehat{\beta}_0 = \overline{y}$.

Comparison with the Null Model

The following are equivalent H_0's:

$H_0 : \beta_1 = \cdots = \beta_p = 0$

H_0: goodness of fit (current model - M) = goodness of fit (null model - M_0)

Test statistic:

$$F = \frac{(TSS - RSS(M))/p}{RSS(M)/(n-(p+1))} \sim F(p, n-(p+1))$$

where

- $TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$ is the **total sum of squares**
- $RSS(M) = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \widehat{u}^t u$, as before.

Decision: to reject H_0 at an α level whenever $f > F_{1-\alpha}(p, n-(p+1))$.

Comparison with the Null Model

Remarks:

1. reject $H_0 \implies M$ has a better goodness-of-fit than M_0
2. do not reject H_0:
   - consider transforming the data
   - $(X_1, \ldots, X_p)$ may have real effect on y but data are not sufficient to prove it
   - it does not make sense to test each β_j individually
3. if any of the t-tests for the individual regression coefficient prove significant, then the F-test will usually be significant
4. puzzling case: none of the t-values for testing the regression coefficients are significant, but the F-test is significant.
   This should be looked at carefully for it may indicate highly correlated explanatory variables - multicollinearity.

ANOVA table

For any linear regression model, TSS has the following decomposition (exercise):

$$\begin{array}{rcl} \text{total variation of } Y & = & \text{variation due to regression} + \text{residual variation} \\ \sum_{i=1}^{n}(y_i - \overline{y})^2 & = & \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \\ \text{TSS} & = & \text{RegSS} + \text{RSS}. \end{array}$$

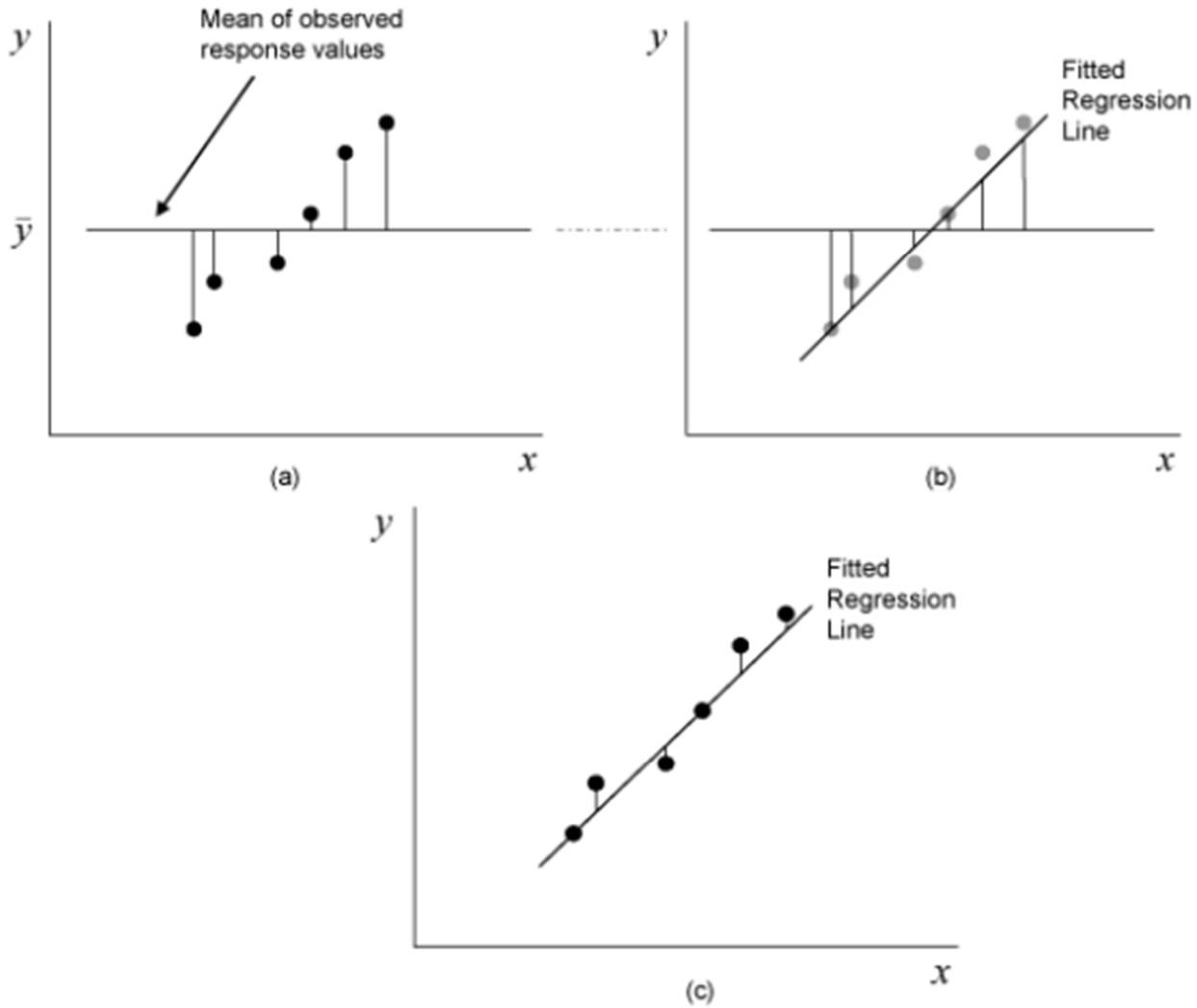
- TSS: **T**otal **S**um of **S**quares

  Represents the response total variability.
- RegSS: **Reg**ression **S**um of **S**quares

  Represents the response variability that is explained by the model
- RSS: **R**esidual **S**um of **S**quares

  Represents the response variability that is not explained by the model

ANOVA Table


Mean of observed response values
y
$\bar{y}$
x
(a)
y
Fitted Regression Line
x
(b)
y
Fitted Regression Line
x
(c)

ANOVA Table

The following is denoted by **ANOVA table** and is returned by almost all statistical analysis softwares (or a similar version of it):

Model	Sum of Squares	df	Mean Squares	F value	p-value
Regression	RegSS$=\sum(\widehat{y_i}-\overline{y})^2$	p	$\frac{RegSS}{p}$	$\frac{RegSS/p}{RSS/(n-(p+1))}$	$P(F>f)$
Residual	RSS$=\sum(y_i-\widehat{y_i})^2$	$n-(p+1)$	$\frac{RSS}{n-(p+1)}$		
Total	TSS$=\sum(y_i-\overline{y})^2$	$n-1$			

In a model with a good fit, with a significant F-test, the total variation is essentially due to the regression, and the residual variation is *small*.

ANOVA Table

In **R**, the ANOVA table has a different look:

```
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: tlc
##           Df Sum Sq Mean Sq F value  Pr(>F)
## age        1 1522.4 1522.35  6.5619 0.01779 *
## bmp        1  283.6  283.64  1.2226 0.28080
## Residuals 22 5104.0  232.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

- RSS corresponds to the 3rd entry of the last row (5104.0)
- TSS corresponds to the sum of the entries in the "Sum Sq" column.

ANOVA Table

In **R**, the test statistic and corresponding p-value from the F-test are given at the end of the *summary* instruction:

```
summary(mod1)
```

```
##
## Call:
## lm(formula = tlc ~ age + bmp, data = base1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.0796 -10.6015   0.8405   7.7994  30.5473
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.9913    20.5443   7.642 1.25e-07 ***
## age          -1.2971     0.6638  -1.954   0.0635 .
## bmp          -0.3093     0.2797  -1.106   0.2808
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.23 on 22 degrees of freedom
## Multiple R-squared:  0.2614, Adjusted R-squared:  0.1942
## F-statistic: 3.892 on 2 and 22 DF,  p-value: 0.03571
```

Coefficient of Determination: R^2

The **coefficient of determination**, R^2 is

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

It is the proportion of the total variability in the response variable that can be accounted for by the set of predictor variables.

Properties:

1. $0 \leq R^2 \leq 1$
2. $R^2 = Corr(y, \widehat{y}) = r^2_{y,\widehat{y}}$
3. simple linear regression: $R^2 = r^2_{y,x}$
4. multiple linear regression: if the predictors are pairwise uncorrelated, $R^2 = r^2_{y,x_1} + \cdots + r^2_{y,x_p}$

Coefficient of Determination: R^2

Remarks:

1. the definition of R^2 assumes the existence of a constant term in the model - this is because TSS has to be seen as RSS(null model)
2. in the absence of any linear relationship between y and the predictor variables, R^2 will be near zero
3. if the model fits the data well then R^2 is close to 1 but the inverse is not necessarily true
4. As a measure of fit, R^2 has to be used carefully
5. beware of high R^2's reported from models without an intercept
6. it is often felt that small sample sizes tend to unduly inflate R^2.

Adjusted R^2

R^2 might not be adequate to compare regression models: adding new predictors (even if non-significant) will always keep or increase the R^2 value. **Why?**

Adjusted R^2: adjusts R^2 for the sample size, dividing RSS and TSS by their respective degrees of freedom

$$R^2_a = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}.$$

Properties:

1. $R^2_a \leq R^2 \leq 1$
2. $R^2_a = 1 - \frac{n-1}{n-(p+1)}(1-R^2)$.

Adjusted R^2

Remarks:

1. R_a^2 can be used to compare models having different numbers of predictors
2. R_a^2 cannot be interpreted as the proportion of total variation in y accounted for by the predictors.
3. R_a^2 can take negative values
4. if the model fits the data well then R_a^2 is close to 1
5. R_a^2 can be critically used as a measure of fit.

Exercise

The table in the next page shows the regression output of a multiple regression model relating the beginning salaries in dollars of employees in a given company to the following predictor variables:

- Sex: 1 = man; 0 = woman
- Education: years of schooling at the time of hire
- Experience: number of months of previous work experience
- Months: number of months with the company.

In items (a)-(b) below, specify the null and alternative hypotheses, the test used, and your conclusion using a 5% level of significance.

Exercise

(a) Conduct the F-test for the overall fit of the regression.

(b) Is there a positive linear relationship between Salary and Experience, after accounting for the effect of Sex, Education, and Months?

(c) Compute R^2 and R_a^2.

(d) Which interval would you present to forecast the salary for a man with 12 years of education, 10 months of experience, and 15 months with the company?

(e) Which interval would you present to forecast the salary, on average, for women with 12 years of education, 10 months of experience, and 15 months with the company?

Exercise

ANOVA table

Source	Sum of Squares	d.f.	Mean Square	F-test
Regression	23665352	4	5916338	22.98
Residuals	22657938	88	257477	

Coefficients Table

Variable	Coefficient	s.e.	t-test	p-value
Constant	3526.4	327.7	10.76	0.000
Sex	722.5	117.8	6.13	0.000
Education	90.02	24.69	3.65	0.000
Experience	1.2690	0.5877	2.16	0.034
Months	23.406	5.201	4.50	0.000
n = 93	d.f.=88	$\widehat{\sigma}^2$ =507.4		

Comparison of Nested Models

Nested Models: Model M_1 is nested in Model M_2 if the parameters in Model M_1 are a subset of the parameters in Model M_2

Examples:

$$M_1: \quad \text{salary} = \beta_0 + \beta_1 Experience + \beta_2 Management$$
$$M_2: \quad \text{salary} = \beta_0 + \beta_1 Experience + \beta_2 Age$$
$$M_3: \quad \text{salary} = \beta_0 + \beta_1 Experience + \beta_2 Management + \beta_3 Age$$

M_1 is nested in M_3; M_2 is nested in M_3

M_1 is not nested in M_2

Comparison of Nested Models

- **Requirements**: model M_1 nested in model M_2
- H_0: M_1 and M_2 have the same goodness-of-fit
  H_1: M_2 has a better goodness-of-fit than M_1
- **Test Statistic**:

$$\frac{(RSS_1 - RSS_2)/(p_2 - p_1)}{RSS_2/(n - (p_2 + 1))} \sim F(p_2 - p_1, n - (p_2 + 1))$$

- **Decision**: reject H_0 at an α level if $f > F_{1-\alpha}(p_2 - p_1, n - (p_2 + 1))$.
  - reject $H_0 \implies$ choose M_2
  - do not reject $H_0 \implies$ choose M_1.

Instructions in **R**:

```
anova(m1, m2)
```

where $m1$ and $m2$ are the models whose goodness of fit is being compared.

Comparison with the Null Model - Special Case

Since, for the null model, $\widehat{y} = \widehat{\beta}_0 = \overline{y}$,

$$\begin{aligned} RSS(\widehat{\beta}_0) &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \overline{y})^2 \\ &= TSS \end{aligned}$$

hence the test statistic is

$$F = \frac{(TSS - RSS(M))/p}{RSS/(n-(p+1))} \sim F(p, n-(p+1))$$

which is the F-test we know from before, with $H_0 : \beta_1 = \cdots = \beta_p = 0$.

Algorithms for Selection of Variables

If the full model has p predictor variables, there are $2^p - 1$ sub-models - impossible to do exhaustive search for best submodel.

If p is *large*, we can proceed with stepwise algorithms:

- forward selection
- backward selection
- combination of the previous two.

Forward selection:(example)

1. Start with the null model M_0
2. Include the predictor variable to the current model which reduces the residual sum of squares most.
3. Continue step 2. until all predictor variables have been chosen or until a large number of predictor variables has been selected. This produces a sequence of sub-models $M_0 \subseteq M_1 \subseteq M_2 \subseteq \ldots$
4. Choose the model in the sequence $M_0 \subseteq M_1 \subseteq M_2 \subseteq \ldots$ which has the smallest AIC / BIC / ... (define a criterion)

Backward selection (example)

1. Start with the full model M (with all predictors)
2. Exclude the predictor variable from the current model which increases the residual sum of squares the least. (or other criterion)
3. Continue step 2. until all predictor variables have been deleted (or a large number of predictor variables). This produces a sequence of sub-models $M \supseteq M_1 \supseteq M_2 \supseteq \ldots$
4. Choose the model in the sequence $M \supseteq M_1 \supseteq M_2 \supseteq \ldots$ which has smallest AIC / BIC / ... (define a criterion)

Drawbacks of Variables' Selection Methods

- variables' selection algorithms can be used as an exploratory tool; the final model should combine knowledge domain with a statistically-based model choice
- the steps in the algorithm inflate the final significance level
- some algorithms are sensitive to the multicollinearity problem and to the presence of outliers.

Comparison of Non-Nested Models

M_1, M_2 non-nested models

There are no hypothesis tests for comparison of the models' goodness-of-fit.

Instead, **information criteria** can be used:

- **AIC**: Akaike Information Criterion

$$AIC = -2LL + 2p$$

- **BIC**: Bayesian Information Criterion

$$BIC = -2LL + p\,log(n)$$

where LL is the value of the model's log-likelihood function, p is the number of regression parameters and n is the number of observations.

The lower the information criterion, the better.

Comparison of Non-Nested Models

```
extractAIC(object, ..., k)
```

where

- *object* is the model at hand
- k is the penalization to be used: $k = 2$ gives AIC; $k = log(n)$ gives BIC.

Comparison of non-nested linear models can also use R_a^2.

Categorical Explanatory Variables

In a regression context, qualitative variables are often named **factors** and are represented by **indicators** or **dummy variables**.

Dummy variables:

- take on only two values, usually 0 and 1
- the two values mean that the observation belongs to one of two possible categories

A categorical variable with K categories is represented in the linear predictor by $K - 1$ dummies.

The category that has no dummy is called the **reference category.**

Categorical Explanatory Variables - Example

Example: Education, coded as 0 - completion of high school, 1 - completion of a bachelor degree, 2 - completion of an advanced degree, needs two dummies.

- choose one category of Education, say class 0, as reference class;
  e.g. in Epidemiology, it is usually the class least associated with the response
- define the two dummies; for subject i,

$$D_{1i} = \begin{cases} 1, & i\text{th subject has completed B.D.} \\ 0, & \text{otherwise} \end{cases}$$

$$D_{2i} = \begin{cases} 1, & i\text{th subject has completed A.D.} \\ 0, & \text{otherwise} \end{cases}$$

D_1 and D_2 uniquely represent the three education groups

Categorical Explanatory Variables - Example

In the regression model, variable **'Education' will be represented by** D_1 **and** D_2

We do not write (why not?)

$$y_i = \beta_0 + \beta_1 \text{Education} + u_i$$

but instead we should consider

$$y_i = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + u_i$$

That is, the regression model having Education as a single predictor has not one but two explanatory variables.

Categorical Explanatory Variables - Example

Z: **body mass index**

0: < 18.5, Underweight; 1: $18.5 - 24.9$, Normal ref; 2: 25.0 and above, Overweight/Obese

Z is represented by 2 dummies:

Z_1: 1-Underweight; 0-Otherwise
Z_2: 1-Overweight/Obese; 0-Otherwise

W: **age group**

0: < 40 y.o. - ref; 1: $41 - 64$ y.o.; 2: $65 - 79$ y.o.; 3: > 80 y.o.

W is represented by 3 dummies:

W_1: 1- age $41 - 64$ y.o.; 0-Otherwise
W_2: 1- age $65 - 80$ y.o.; 0-Otherwise
W_3: 1- over 80 y.o.; 0-Otherwise

Categorical Explanatory Variables - Example

The regression model for a response variable y including body mass index and age group as predictors is

$$y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 W_{1i} + \beta_4 W_{2i} + \beta_5 W_{3i} + u_i, \quad u_i \sim N(0, \sigma^2)$$

The predicted response...

- for a 57 y.o. individual with BMI=26 is $\widehat{\beta}_0 + \widehat{\beta}_2 + \widehat{\beta}_3$
- for a 37 y.o. individual with BMI=18 is $\widehat{\beta}_0 + \widehat{\beta}_1$
- for an 82 y.o. individual with BMI=24 is $\widehat{\beta}_0 + \widehat{\beta}_5$
- for a 20 y.o. individual with BMI=19 is $\widehat{\beta}_0$.

Example (from Chatterjee et al)

The Salary Survey data set was developed from a salary survey of computer professionals in a large corporation (data on next slide).

The response variable is salary (S) and the predictors are: experience (X), measured in years; education (E), coded as 1 for completion of a high school (H.S.) diploma, 2 for completion of a bachelor degree (B.S.), and 3 for the completion of an advanced degree (A.D.); and management (M), which is coded as 1 for a person with management responsibility and 0 otherwise.

Among the predictors, the only factor is education. We fix 'completion of an advanced degree' as its reference class and define the dummy E_1 (resp. E_2) for completion of a high school (resp. bachelor) degree.

Regression model:

$$S = \beta_0 + \beta_1 X + \beta_2 D_1 + \beta_3 D_2 + \beta_4 M + u$$

Example (from Chatterjee et al)

Row	S	X	E	M	Row	S	X	E	M
1	13876	1	1	1	24	22884	6	2	1
2	11608	1	3	0	25	16978	7	1	1
3	18701	1	3	1	26	14803	8	2	0
4	11283	1	2	0	27	17404	8	1	1
5	11767	1	3	0	28	22184	8	3	1
6	20872	2	2	1	29	13548	8	1	0
7	11772	2	2	0	30	14467	10	1	0
8	10535	2	1	0	31	15942	10	2	0
9	12195	2	3	0	32	23174	10	3	1
10	12313	3	2	0	33	23780	10	2	1
11	14975	3	1	1	34	25410	11	2	1
12	21371	3	2	1	35	14861	11	1	0
13	19800	3	3	1	36	16882	12	2	0
14	11417	4	1	0	37	24170	12	3	1
15	20263	4	3	1	38	15990	13	1	0
16	13231	4	3	0	39	26330	13	2	1
17	12884	4	2	0	40	17949	14	2	0
18	13245	5	2	0	41	25685	15	3	1
19	13677	5	3	0	42	27837	16	2	1
20	15965	5	1	1	43	18838	16	2	0
21	12336	6	1	0	44	17483	16	1	0
22	21352	6	3	1	45	19207	17	2	0
23	13839	6	2	0	46	19346	20	1	0

Example

Variable	Coefficient	s.e.	t-test	p-value
Constant	11031.800	383.2	28.80	< 0.0001
X	546.184	30.5	17.90	< 0.0001
E_1	-2996.210	411.8	-7.28	< 0.0001
E_2	147.825	387.7	0.38	0.7049
M	6883.530	313.9	21.90	< 0.0001
$n = 46$	$R^2 = 0.957$	$R^2_a = 0.953$	$\hat{\sigma} = 1027$	d.f.= 41

- each additional year of experience is estimated to be worth an annual salary increment of 546 dollars
- 6883.5 is the average incremental value in annual salary associated with a management position
- an A.D. is worth 2996 dollars more than a H.S. diploma, a B.S. is worth 148 dollars more than an A.D. (this differential is not statistically significant), and a B.S. is worth 3144 dollars more than a H.S. diploma.

Interactions

Consider a regression model with 2 explanatory variables, x_1 and x_2, and x_1 is a factor (binary, without loss of generality).

The interaction variable between x_1 and x_2 is defined as the product $x_1 x_2$, and the regression model including it is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + u_i, \quad u_i \sim N(0, \sigma^2)$$

There is an interaction between x_1 and x_2 whenever the effect of x_2 on y depends on the levels of x_1. In fact:

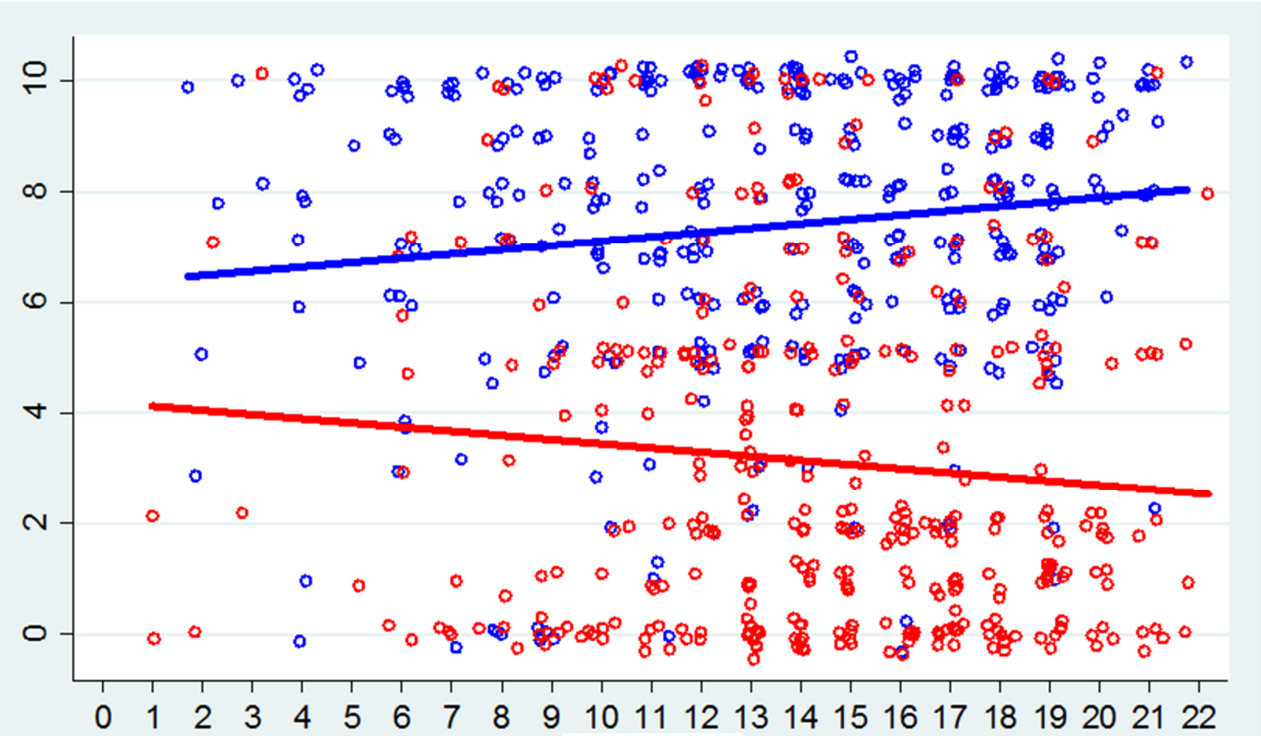
- $x_1 = 0$: the model is
$$y = \beta_0 + \beta_2 x_2 + u_i$$
- $x_1 = 1$; the model is
$$y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_2 + u_i,$$

Example

Regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + u_i, \quad u_i \sim N(0, \sigma^2)$$

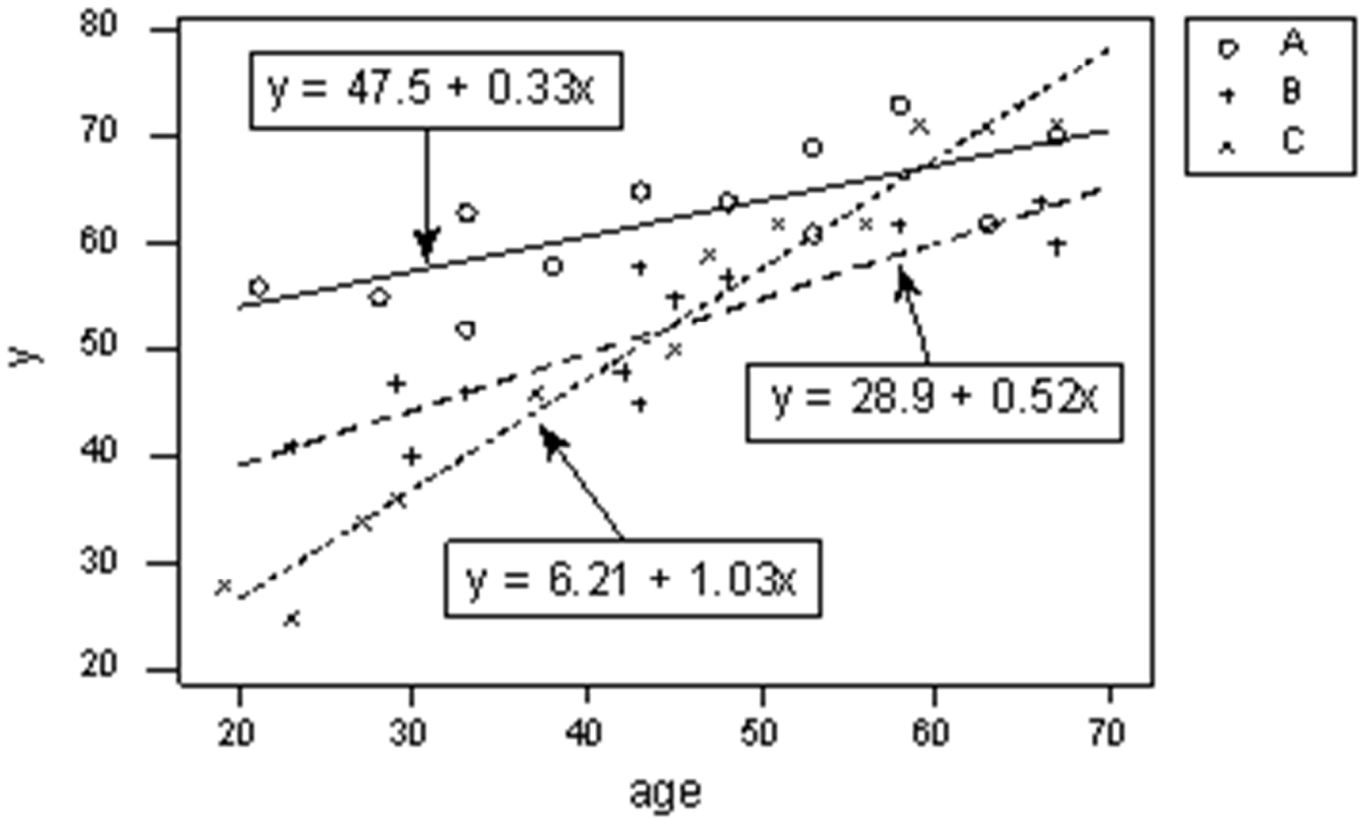
In the picture, x_1 has 2 levels (red and blue) and x_2 corresponds to the x-axis. The response variable is on the y-axis.



Example

In the picture, the response y varies with age, in the x-axis, and a 3-level BMI-body mass index.

There is an interaction between age and the 3-level factor BMI.



Exercise

Recover the previous regression model for prediction of the salary of computer professionals from experience (measured in years), education (3 levels, ref=A.D.) and management responsibilities (0-no; 1-yes).

A model with an interaction between education and management provides the following estimates. Interpret each effect from the model.

Variable	Coefficient	s.e.	t-test	p-value
Constant	11203.40	79.07	141.7	< 0.0001
X	496.99	5.57	89.3	< 0.0001
E_1	-1730.75	105.30	-16.4	< 0.0001
E_2	-349.08	97.57	-3.6	0.0009
M	7047.41	102.60	68.7	< 0.0001
$E_1 \cdot M$	-3066.04	149.30	-20.5	< 0.0001
$E_2 \cdot M$	1836.49	131.20	14.0	< 0.0001
$n = 46$	$R^2 = 0.999$	$R^2_a = 0.999$	$\hat{\sigma} = 173.8$	d.f.= 39

Outliers

Outlier: observation for which the residual is large in magnitude compared to other observations in the data set.

Can correspond to:

- errors in the data collection, or
- genuine observations.

Empirical Rule for Outliers' Identification: look at

- observations with standardized residuals higher, in absolute value, than 3.3 (0.999-quantile for $N(0,1)$)
- residuals' dispersion pattern

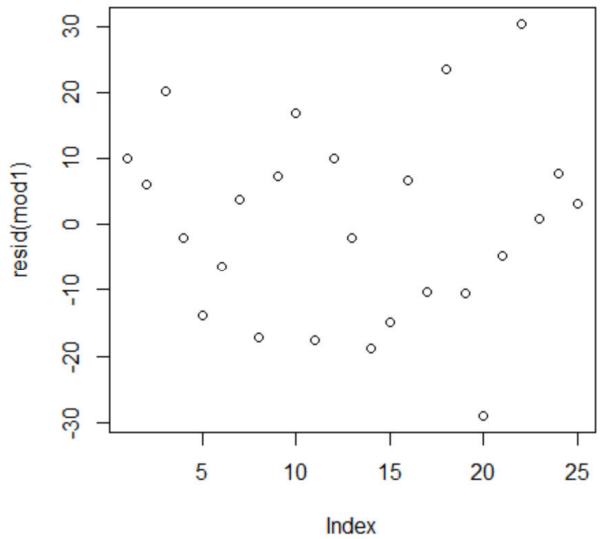
The identified observations should be carefully looked at. If necessary, remove them, fit the model again and compare the results.

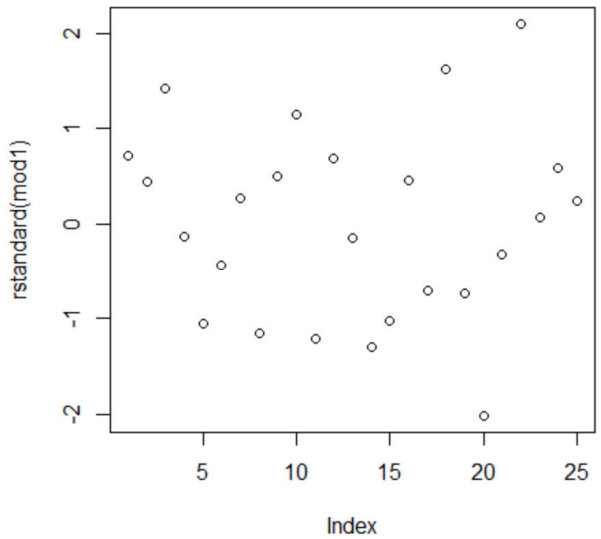
Outliers

Instructions in **R**:

```
residuals(model)  # crude residuals
resid(model)
model$resid
```

```
rstandard(model)  # standardized residuals
```



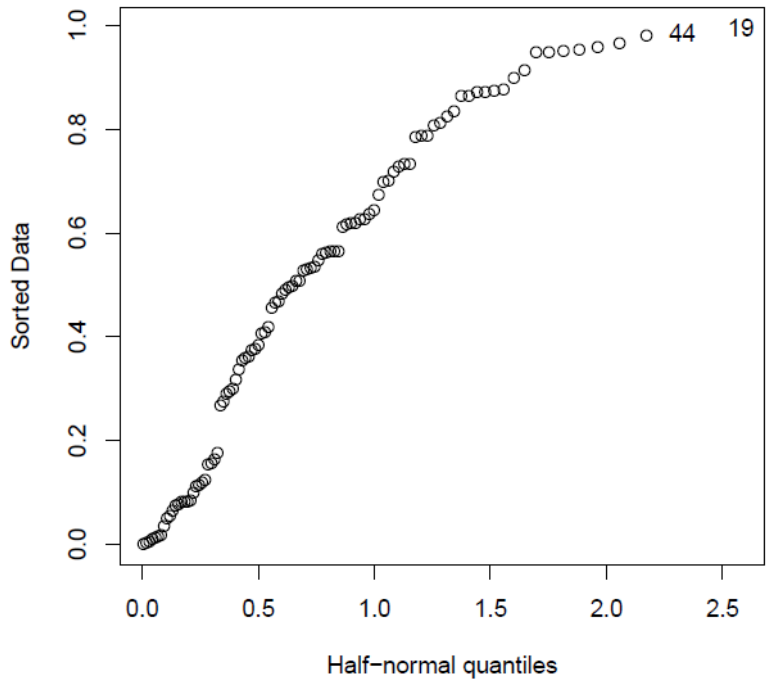


Outliers

Outliers may be identified by the **half-normal plot**, a variation of the QQ-plot for normality that plots the sorted data against the quantiles $\left(\frac{n+i}{2n+1}\right)_{i=1,\ldots,n}$ of the standard normal $N(0,1)$ distribution.

Instructions in **R**:

```
library(faraway)
halfnorm(residuals)
```



Checking Model Assumptions

residuals $\widehat{u}_i = y_i - \widehat{y}_i$:

- approximation of the unobservable error term u_i
- can help checking whether the linear model is appropriate.

Tukey-Anscombe Plot: plots the residuals $\widehat{u}_i$ (on the y-axis) versus the fitted values $\widehat{y}_i$ (on the x-axis).

As $\widehat{Corr}(\widehat{u}_i, \widehat{y}_i) = 0$, indeed $\widehat{u}_i \perp \widehat{y}_i$, the points should randomly fluctuate around the horizontal line through zero

Checking Model Assumptions

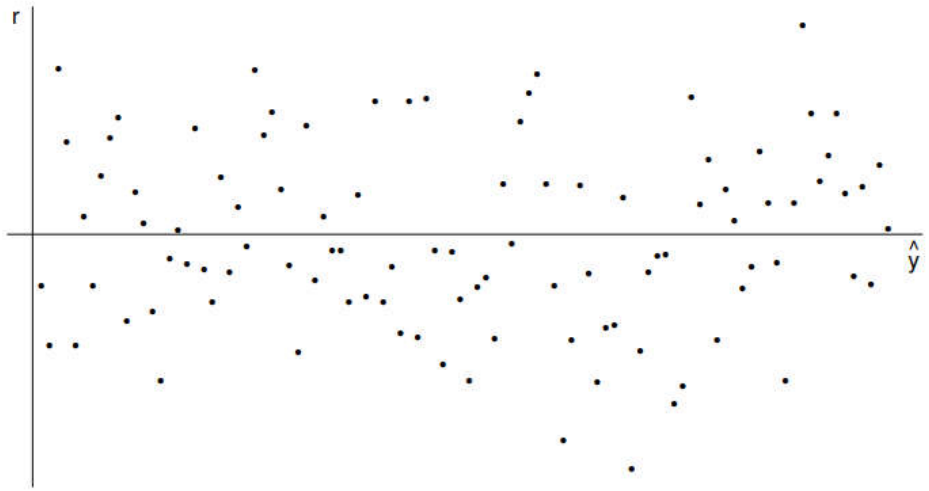


Figure 1.4: Ideal Tukey-Anscombe plot: no violations of model assumptions.

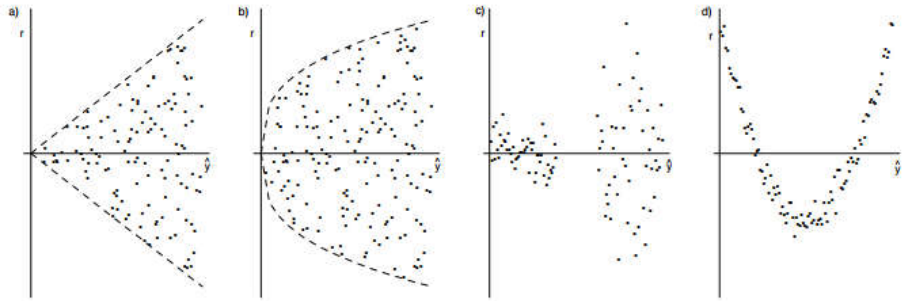
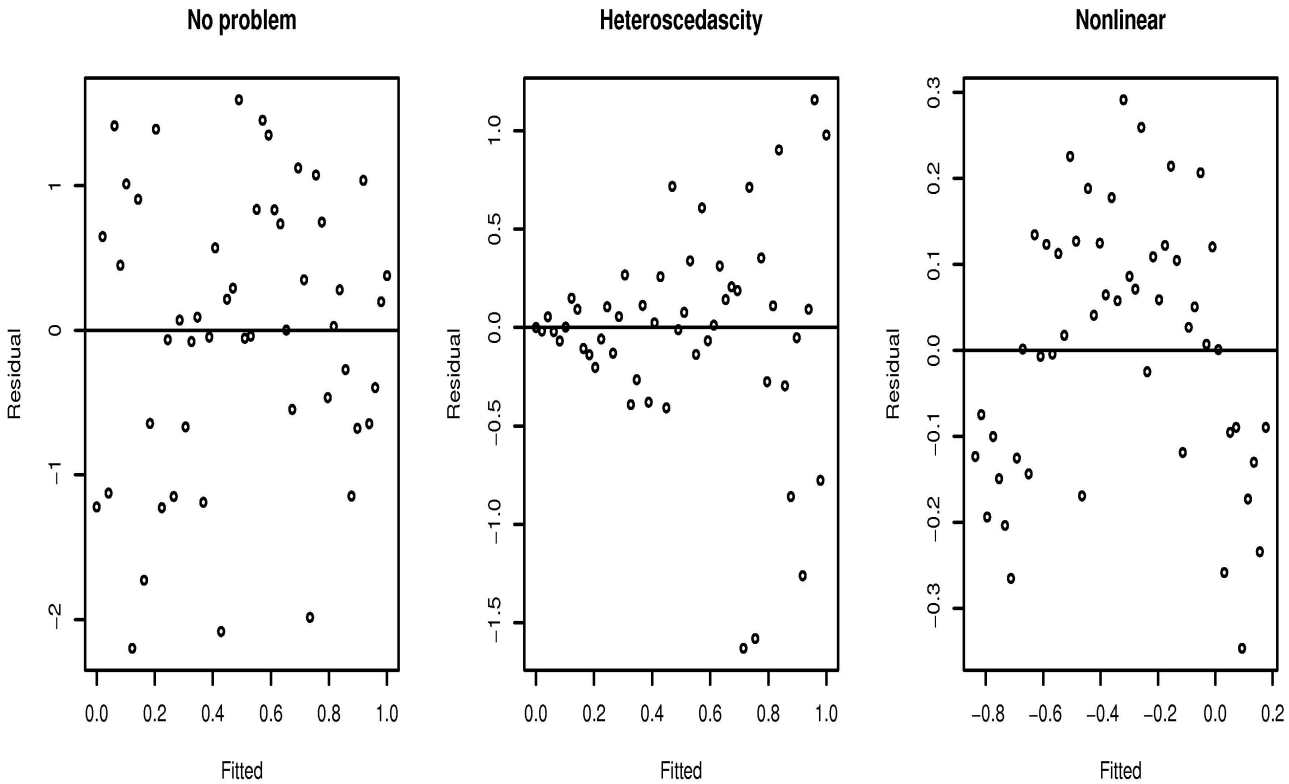


Figure 1.5: a) linear increase of standard deviation, b) nonlinear increase of standard deviation, c) 2 groups with different variances, d) missing quadratic term in the model.

Checking Model Assumptions



Checking Model Assumptions

Normality of the residuals: necessary for confidence intervals and hypothesis tests on the regression parameters or else n has to be *sufficiently large*

Plot the histogram, boxplot and QQ-plot of the standardized residuals.

If normality fails, response transformations or redefinition of the model may be considered.

Checking Model Assumptions

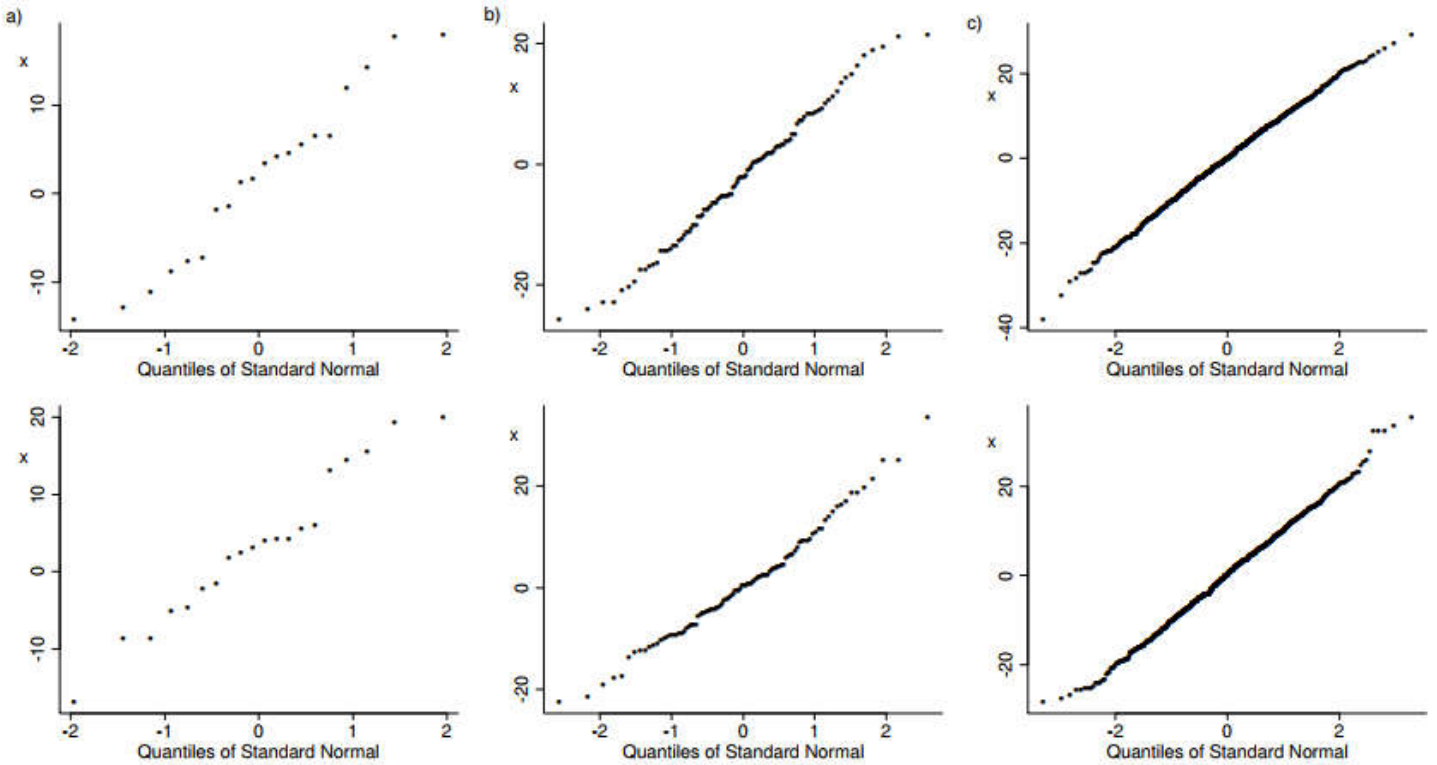


Figure 1.6: QQ-plots for i.i.d. normally distributed random variables. Two plots for each sample size n equal to a) 20, b) 100 and c) 1000.

Checking Model Assumptions

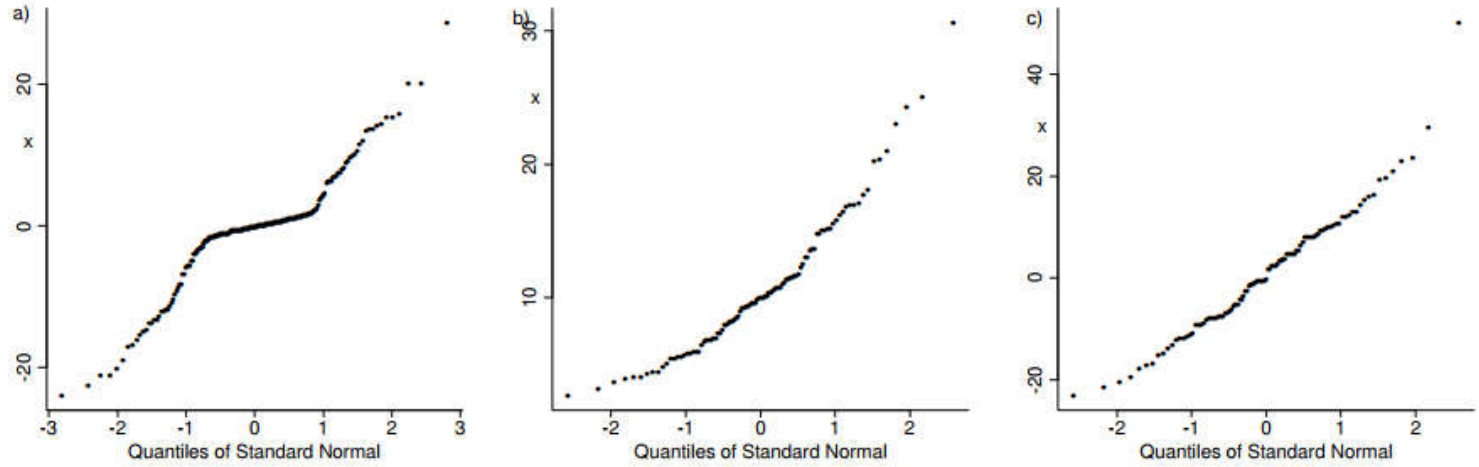
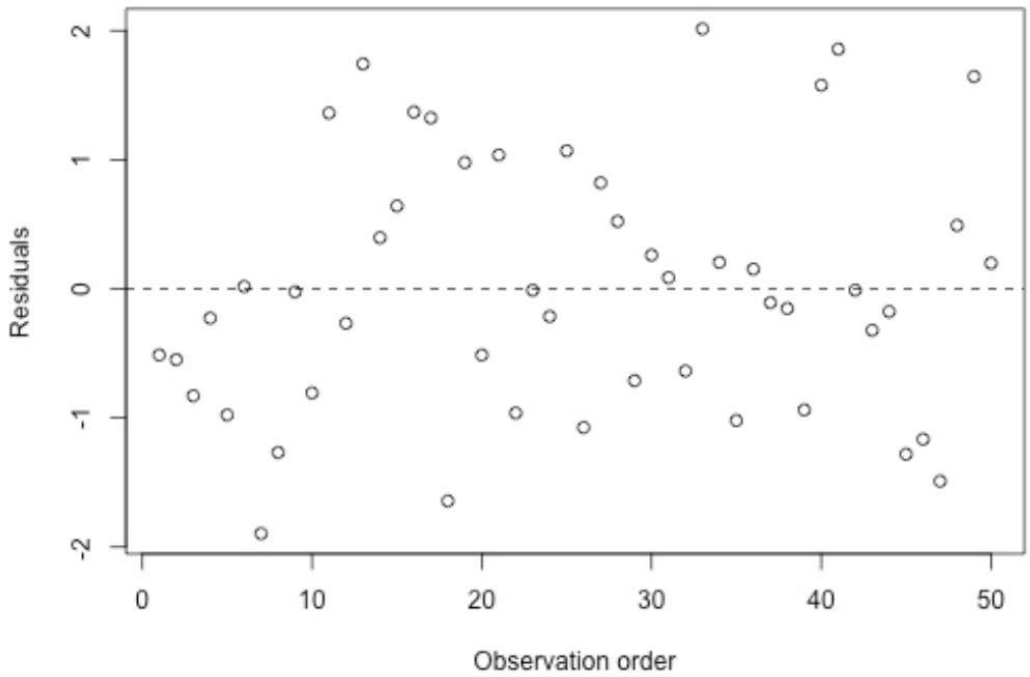


Figure 1.7: QQ-plots for a) long-tailed distribution, b) skewed distribution, c) dataset with outlier.

Checking Model Assumptions

Plot of the **residuals versus the observations numbers/indexes** (or, if available, the time of recording of each observation):

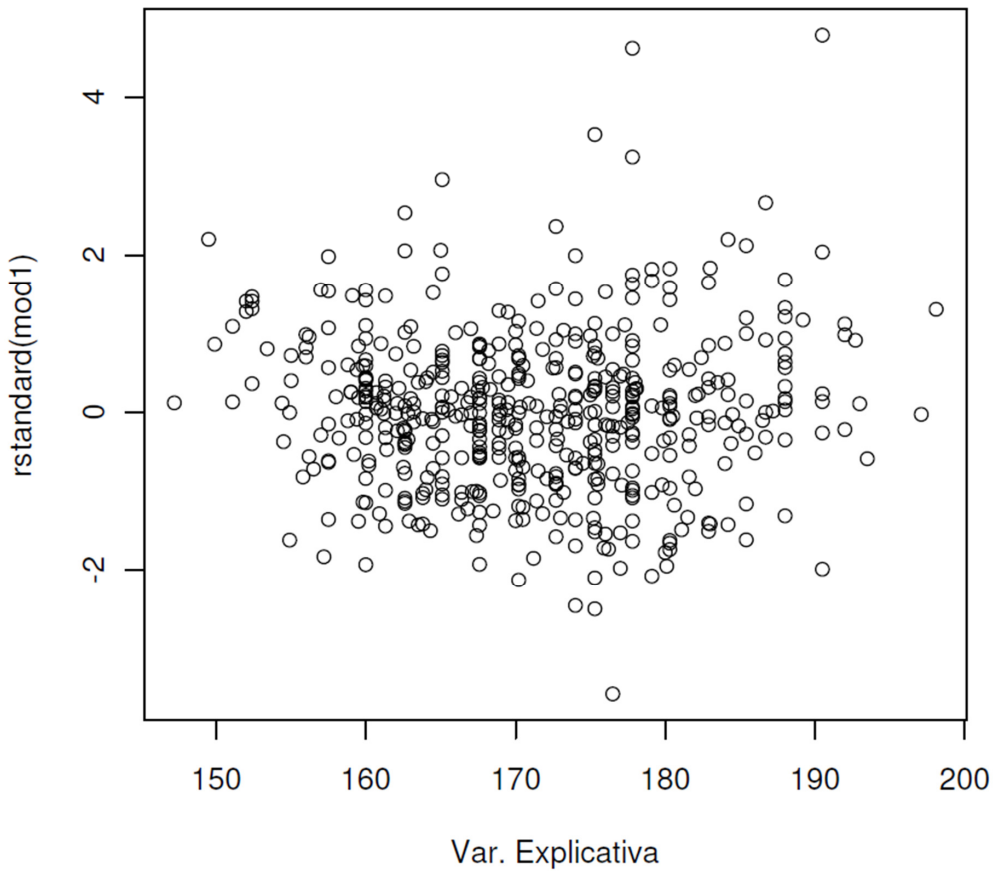
- independent observations $\Longrightarrow$ residuals vary randomly around the zero line
- neighboring (w.r.t. x-axis) residuals look similar $\Longrightarrow$ independence assumption for the errors seems violated.



Checking Model Assumptions

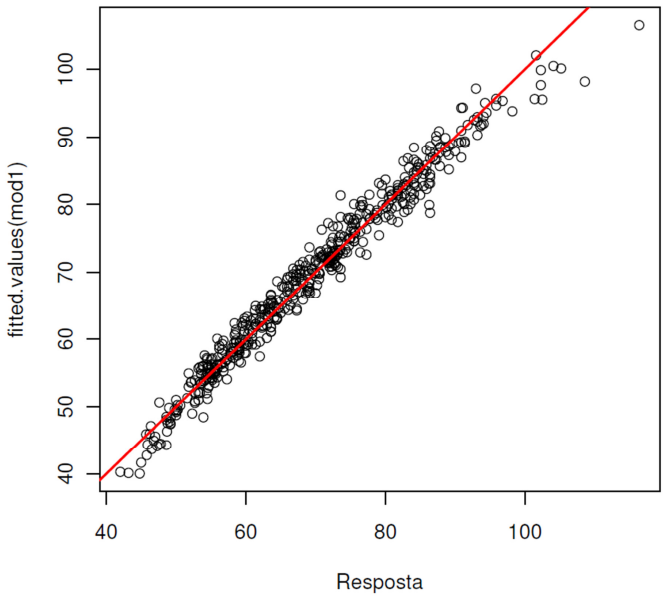
Standardized residuals versus explanatory variables: as residuals are orthogonal to the design matrix, no pattern should be observed.

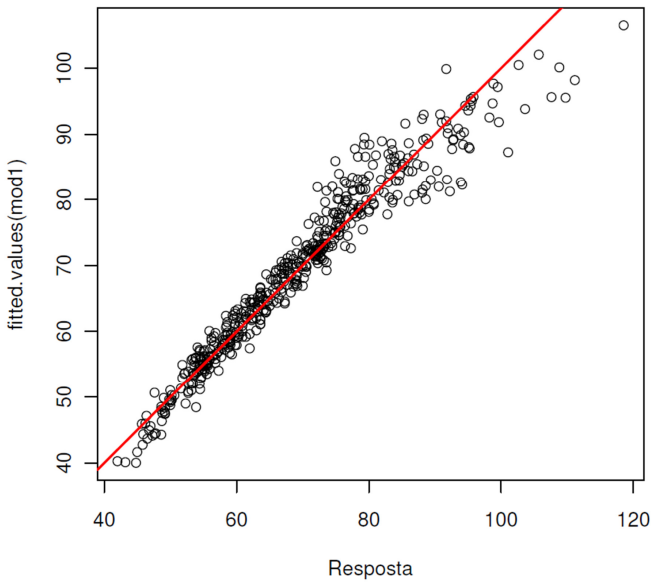
The existence of an association may suggest a transformation in the predictor.



Checking Model Assumptions

Observed response versus fitted values: supplementary plot to the Tukey-Anscombe plot.





General Linear Model*

The **general linear model** is a regression model with heterocesdastic and/or correlated errors:

$$\boxed{y = X\beta + u, \qquad u \sim MVN(0, V)}$$

where V is any symmetric and positive definite matrix.

In this situation, β can be estimated by the generalized least squares method - **GLS**, by minimizing

$$RSS(\beta) = (y - X\beta)^t V^{-1}(y - X\beta).$$

We still have

$$\widehat{\beta}_{ML} = \widehat{\beta}_{GLS}.$$

Exercise: write down the likelihood function and observe that maximizing it is equivalent to minimizing RSS given above.

General Linear Model*

The GLS estimator of β is (exercise)

$$\widehat{\beta}_{GLS} = (X^t V^{-1} X)^{-1} X^t V^{-1} y.$$

Aitken's Theorem: $\widehat{\beta}_{GLS}$ is BLUE.

For uncorrelated errors (and possibly heterocedastic), V is simply a diagonal matrix and the generalized least squared method is called a **weighted least squares method**.

Bibliographic References

- Julian J. Faraway, *Linear Models with R*. Chapman & Hall/CRC texts in statistical science series (more practical)
- Chatterjee *et al.*, *Regression Analysis by Example*. Wiley Series in Probability and Statistics
- Ashish Sen and Muni Srivastava. *Regression Analysis: Theory, Methods, and Applications*. Springer Verlag. (more mathematical)
- Julian J. Faraway, *Practical Regression and Anova using R*. Chapman & Hall/CRC texts in statistical science series
- Frank E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Series in Statistics
- John Fox, *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications
- Weisberg, *Applied Linear Regression*.
- Draper and Smith, *Applied Regression Analysis*. Wiley Series in Probability and Statistics.