Discovering associations

Biomedical Data Science

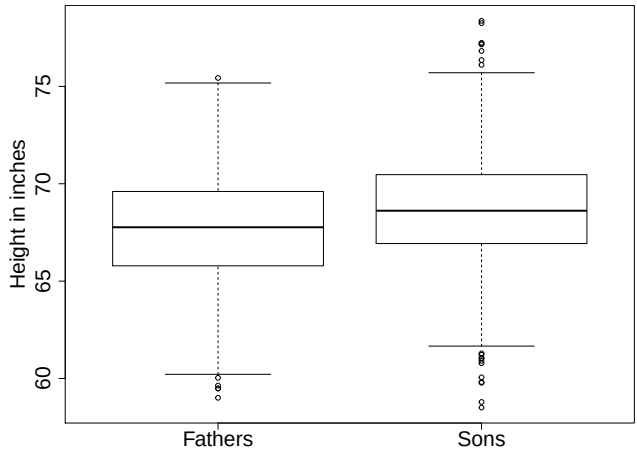
Marco Colombo

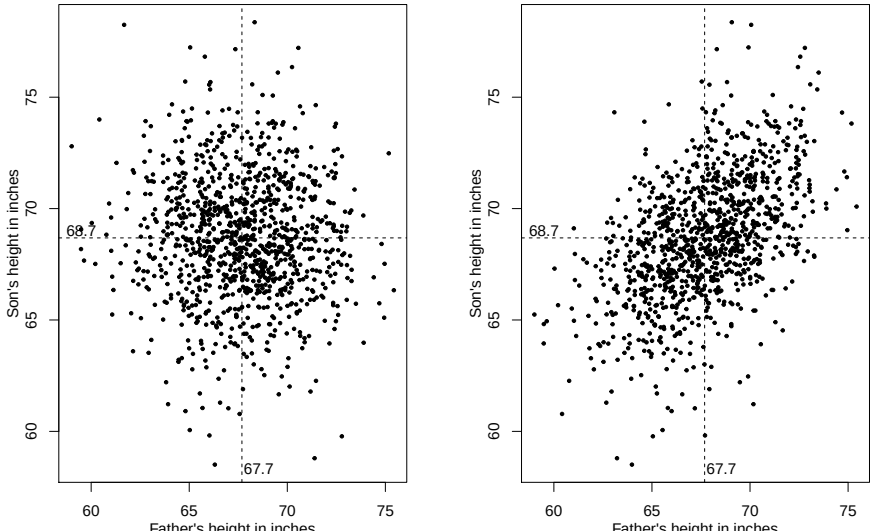
Lecture 2, 2017/2018

Father–son data

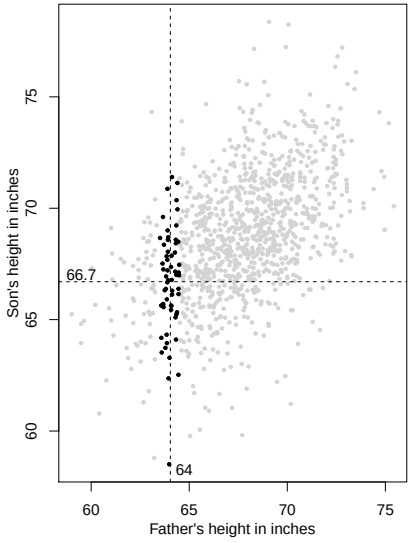
Let's start with a famous example: the study of the relationship between father and son heights. The dataset comprises heights of 1078 fathers and their adult sons in 1900 England.

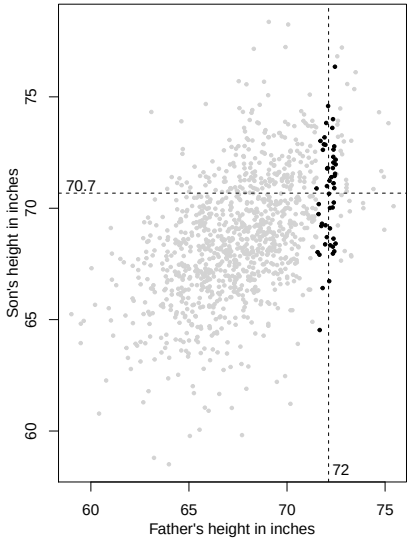
	fheight	sheight
1	65.04851	59.77827
2	63.25094	63.21404
3	64.95532	63.34242
4	65.75250	62.79238
5	61.13723	64.28113
6	63.02254	64.24221
7	65.37053	64.08231
8	64.72398	63.99574
9	66.06509	64.61338
10	66.96738	63.97944





Both plots have the same marginal distribution. Real data (not shuffled) are on the right.





By knowing the father's height, we can improve our guess. The variables are *correlated*.

Covariance

The *covariance* between two variables measures the strength of the linear dependence with each other:

$$\begin{aligned} cov(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= \frac{1}{n-1} \sum_i (x_i - \mu_X)(y_i - \mu_Y) \end{aligned}$$

From the sign of the covariance, we can infer wheter there is a positive or a negative relationship between the two variables.

For the father–son data, $cov(fheight, sheight) = 3.87$.

This quantity is hard to interpret as its measurement unit corresponds to the unit of X times the one of Y (in our example, inches^2).

Correlation coefficient

The *Pearson correlation coefficient* scales the covariance by the product of the standard deviations:

$$\rho = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

The correlation coefficient is a dimensionless value in the $[-1, 1]$ interval, so it's easier to interpret.

For the father–son data, $p = 0.501$.

Note that the correlation coefficient is a measure of the *linear relationship* of variables: it is a misleading measure if the two variables are not related linearly.

Bounds on the correlation coefficient

It's easy to prove that $\rho \in [-1, 1]$ from the formula for the angle between two vectors:

$$a^T b = \|a\|\|b\| \cos\theta_{ab}.$$

Given that variances and covariances are translation-invariant (they are measure of dispersion rather than location):

$$var(X) = var(X + k),$$

we can assume without loss of generality that we subtracted the mean from X and Y, so they both have mean 0 (that is, the variables are *centred*):

$$\cos\theta_{XY} = \frac{X^T Y}{\|X\|\|Y\|} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} = \rho$$

Regression toward the mean

It was Francis Galton who in 1886 first formally described the association between heights in fathers and sons, which would not have been obvious without considering the two variables jointly.

He used the term *regression* to describe the process according to which in the extreme ranges of fathers' heights, the sons' heights are less extreme (regression toward the mean).

The amount of reduction in deviation from the average is exactly the correlation coefficient ρ.

Statistical inference

Inference is the use of probability to learn population characteristics from data.

The purpose of a *model* is to allow us to use past observations (data) to describe in a *parsimonious* and *generalizable* way a certain phenomenon, and from that to make predictions.

- Parsimonious model: the model uses as few predictors as possible
- Generalizable model: the model describes correctly data that was not used in fitting

In order to do this we need a way of choosing what value to assign to the parameters of the model: this process is called *parameter estimation*.

Linear models

One of the most effective type of models is the *linear model*.

In a linear model we assume that the relationship between the outcome (y) and the predictors (X) is linear.

These are examples of linear models:

- $y = \beta_0 + \beta_1 X_1$
- $y = \beta_0 + \beta_1 X_1^2$
- $y = \beta_0 + \beta_1 X_1 X_2$

Whatever transformation is applied to the predictors, these models are linear in the parameters ($\beta_0, \beta 1$).

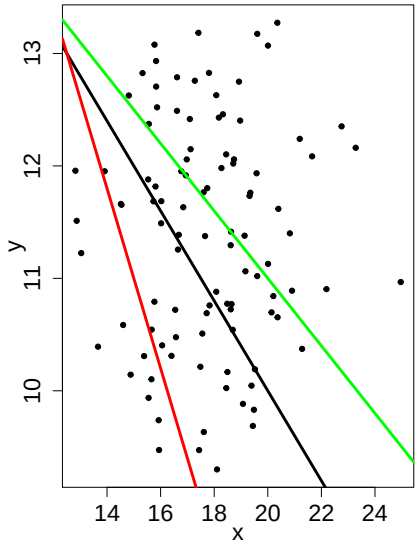
Linear regression

Linear regression is one of the simplest types of linear model.

Given n data points x_i, y_i, $i = 1, \ldots, n$, we want to find the equation of the line that best fits the data.

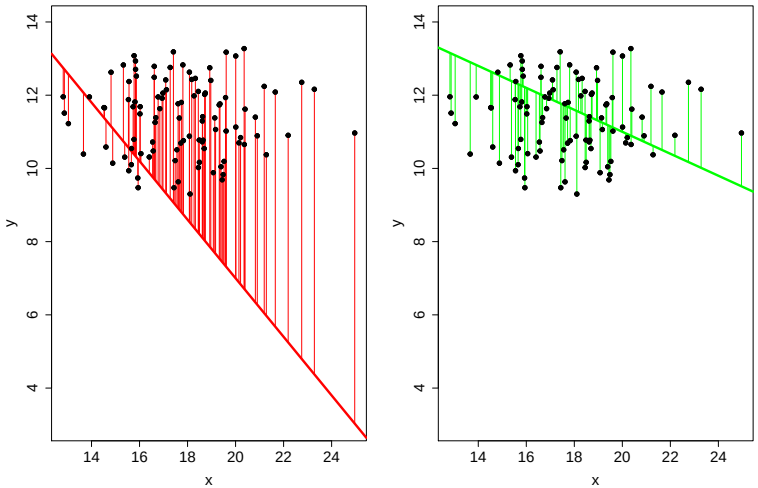
What could be a criterion to define the best fit?

In general we want to minimize the errors made when summarizing the data points by a line.



Residuals

The errors made by summarizing the data by a line are called *residuals*, and measure the distance between each data point and the regression line.



Least squares approach

Following to the *least squares approach*, we find the equation of the regression line

$$y = \beta_0 + \beta_1 X_1 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

which minimizes the sum of squared residuals (RSS):

$$\mathrm{RSS} = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

This can be framed in terms of an optimization problem:

$$\min Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Multiple linear regression

The simple linear regression model can be extended to incorporate p predictors (covariates):

$$\begin{aligned} y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \\ &= X\beta + \varepsilon \end{aligned}$$

where

$$X = [1\ X_1\ X_2\ \ldots\ X_p] \in R^{n,p+1}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \in R^{p+1}$$

X is called the *design matrix*, while β is the vector of *regression coefficients* we want estimate.

Solving the least squares problem

The general formulation of the optimization problem is the following:

$$\min Q(\beta) = (y - X\beta)^T (y - X\beta).$$

This is a convex quadratic equation. To solve it, we need to find the points for which its gradient $\nabla Q = 0$:

$$\nabla Q = -2X^T(y - X\beta) = 0 \qquad \Leftrightarrow \qquad X^T X\beta = X^T y$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Interpreting the regression results

$\hat{\beta}$ represents the vector of fitted (estimated) regression coefficients.

```
(Intercept)     fheight
     33.900       0.514
```

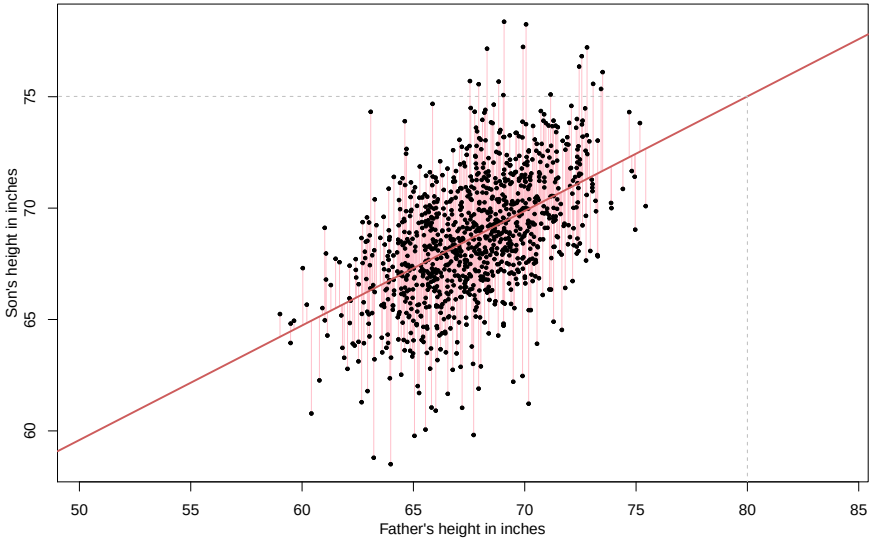
Therefore, our linear regression model reads as

$$\texttt{sheight} = 33.9 + 0.514\,\texttt{fheight}$$

$\hat{\beta}_0$ is the intercept term, and corresponds to the value of y when all $X_i = 0$.

$\hat{\beta}_j$ represents the amount by which y increases on average if we increase X_j by one unit while keeping constant all other X_i, $i \neq j$.

The regression line



Assumptions of linear regression

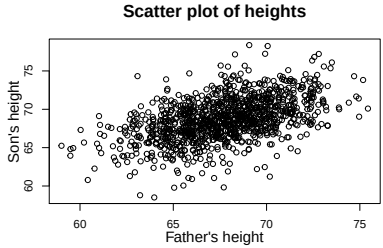
Some assumption must hold for a linear regression model to be valid.

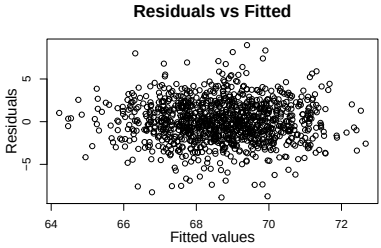
- The true underlying relationship between X and y is linear
- The variance of the error is constant (homoscedasticity)
- The residuals must be independent and normally distributed

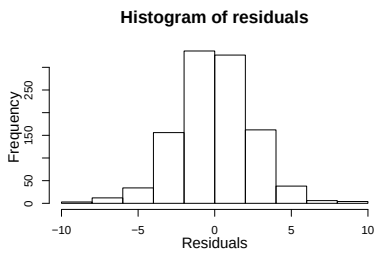
Note that there is no assumption about the distributions of X and y!

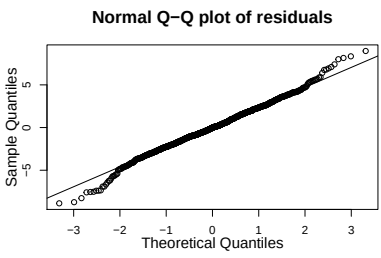
Other conditions for the usefulness of a model are that the data used in fitting the model are representative of the population of interest, and that variables can be measured without error.

Checking the model assumptions

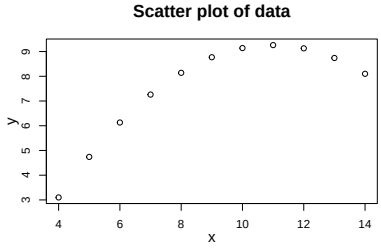


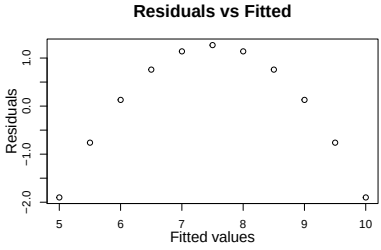


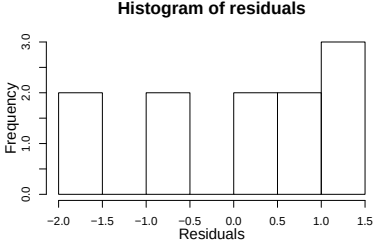


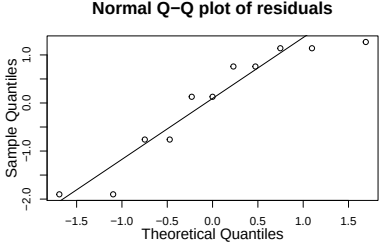


Example: violation of assumptions









Coding categorical variables

Categorical variables need to be transformed in order to be usable within a design matrix.

For a categorical variable with k levels, we create $k - 1$ dummy variables (one level is used as a reference level), and code each of them by assigning a 1 to the samples that fall into that level, and 0 otherwise.

Example: Albumin-to-creatinine ratio (ACR) quantifies the amount of protein excreted in urine, and is often categorized in three group (Normo-, Micro- or Macroalbuminuria).

sex	ACR	(Intercept)	sexM	ACRMacro	ACRMicro
F	Micro	1	0	0	1
M	Macro	1	1	1	0
F	Normo	1	0	0	0
F	Macro	1	0	1	0
M	Normo	1	1	0	0

Linear algebra considerations

Consider

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Note that this formulation requires the computation the inverse matrix $(X^T X)^{-1}$. Under what conditions does this exist?

- $(X^T X)$ has to be invertible (non-singular)
- This occurs if the columns of X are linearly independent:

$$X_i = \theta_0 + \theta_1 X_j, \quad i \neq j$$

- In an n-dimensional space, there can be at most n linearly independent vectors

Therefore, the linear regression problem cannot be solved if $p > n$.

Collinearity

Even for $n > p$, a regression problem may not be solved if two variables are linearly dependent. When this occurs, it's often because of misspecification of the model (especially when using categorical variables).

However, even when two variables are not linearly dependent but are highly correlated, the condition number of matrix $(X^T X)$ can be very high.

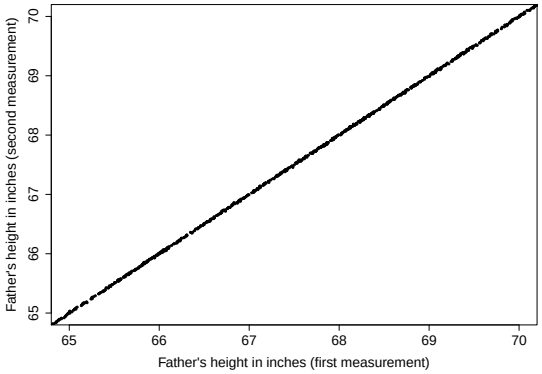
Ill-conditioning causes several problems:

- a numerical algorithm may struggle to find an inverse
- the inverse may contains very large coefficients
- some of the fitted coefficients $\hat{\beta}$ can be very large
- small changes in the data cause very large changes in $\hat{\beta}$
- standard errors in the related variables are large

Collinearity example

Suppose we had a second measurement taken of fathers' heights and we included this in the model: this is likely to be a noisy version of the first one.

	fheight	fheight.2nd
1	65.0	65.0
2	63.3	63.2
3	65.0	64.9
4	65.8	65.7
5	61.1	61.1
6	63.0	63.0



	Estimate	Std. Error
(Intercept)	33.9000	1.83
fheight	0.4480	7.52
fheight.2nd	0.0658	7.51

Hypothesis testing I

To evaluate the strength of an association we need a formal statistical approach that allows us to test how confident we are that a regression coefficient is not 0.

```
(Intercept)     fheight
     33.900       0.514
```

This can be done formally by hyphothesis testing:

- Null hypothesis: the regression coefficient $\hat{\beta}_i = 0$
- Alternative hypothesis: $\hat{\beta}_i \neq 0$

Hypothesis testing II

From the vector of residuals $\varepsilon = y - \hat{y}$ and the design matrix X we can compute the standard error for each $\hat{\beta}_i$:

$$SE(\hat{\beta}_i) = \sqrt{\frac{\varepsilon^T \varepsilon}{n-p}(X^T X)^{-1}_{ii}}$$

This can be used to compute the Wald test statistic:

$$w = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

The Wald test statistic follows the t distribution with $n - p - 1$ degrees of freedom: from this we can derive a *p-value*, which is the probability of getting a more extreme result that what was observed.

Hypothesis testing III

Choose some small value α (conventionally $\alpha = 0.05$) and reject the null hypothesis if the *p*-value is less than α.

We call α the *significance level* of the test.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.900	1.830	18.5	1.60e-66
fheight	0.514	0.027	19.0	1.12e-69

The *p*-value allows us to decide whether we can reject the null hypothesis, and if so claim that X_i has an association with the outcome after adjusting for all other covariates in the model.

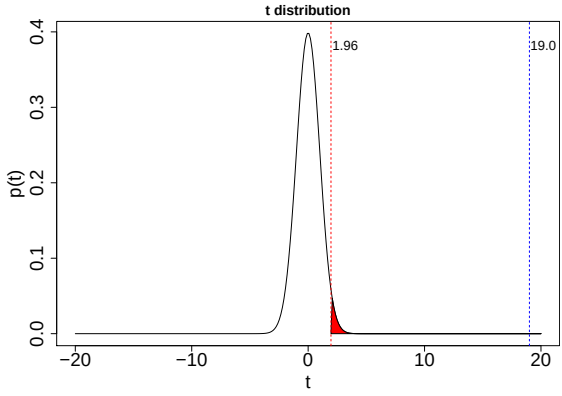
p-values

A *p*-value is a statement about the data *given* the null hyphothesis.

That is: assuming that covariate *i* has no effect on the outcome *y* (given all other covariates in the model), we can define the distribution of the test statistic.

The *p*-value is the probability of seeing a test statistic at least as big as the one that our data implies.

A *p*-value doesn't provide any information on the probability of the null hypothesis given the data!



Reporting associations

Often a *p*-value alone is used in declaring the existence of an association.

- If the *p*-value $< \alpha$ then we can claim that the association is statistically significant (at the significance level α).
- If the *p*-value $> \alpha$ the null hypothesis cannot be rejected, but this does not imply that the null hypothesis is true.

What should be reported are the effect size with a confidence interval or some statistic which gives the reader a sense of the change in a meaningful scale.

If effect sizes are small, an individual's risk for disease might not change by an amount that is practically significant, independently of how small a *p*-value may be.

Errors in hypothesis testing

There are two different sorts of errors that can be made when testing hypotheses.

- *Type I error*: rejecting the null hypothesis when it's true (false positive): In this case we claim that there is an association, when actually there is none. This is controlled by the significance level α.

- *Type II error*: not rejecting the null hypothesis when it's false (false negative): in this case we do not report an association when it's present. We measure this by β.

There is a trade-off between Type I and Type II errors: usually, by trying to decrease one type of error, the chances of committing the other type of error increase.

Power

Power is the probability of rejecting the null hypothesis when it's false: this probability is $1 - \beta$, where β is the Type II error rate.

Power depends on the standard error of the estimates, which in turn depends on the sample size and on the population standard deviations.

The only way to reduce Type I and Type II errors is by increasing sample sizes. However this is not always possible, it's costly, and in some cases (such as in certain types of drug trials) may even be unethical.

Type II error control plays a major role in planning study design and data collection procedures before seeing the data and in deciding if an investigation has good chances of succeeding.

Power considerations

The Type I error rate α can be characterized by assuming that the null hypothesis is true (ie: no difference). However, the same cannot be done for the Type II error rate β, as the alternative hypothesis does not specify a particular value for the difference.

So in power calculations (through which we can study the Type II error rate) we need to assume a particular effect size. The questions we raise are the following:

- what is the smallest difference that can be reliably distinguish from 0 given a sample size of N?
- how big does N have to be in order to detect that the absolute value of the difference is greater than 0?

Setting a lower α decreases the power of the test for a given effect size because the null hypothesis will be more difficult to reject.

Assessing how well a model fits the data

There are different approaches to evaluate the fit of the model.

- Residual standard error (RSE): estimates the standard deviation of the errors made by using the regression model (residuals), as measured by the residual sum of squares RSS

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}} = \sqrt{\frac{1}{n-p-1}\varepsilon^T\varepsilon} = \sqrt{\frac{1}{n-p-1}\sum_i (y_i - \hat{y}_i)^2}$$

- R^2 statistic: can be interpreted as the percentage of the variation in the outcome that is explained by the model

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Model complexity

Model complexity must be taken into consideration when building a model.

For an equivalent measure of fit, the most parsimonious model (least complex) should be preferred.

R^2 does not take into account model complexity: it always increases with the addition of covariates in the model.

The *adjusted* R^2 provides a better summary of the fit of the model that accounts for the number of predictors.

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Overfitting

A regression model can be made as accurate as desired: for a sample of size n, we can guarantee that by using n linearly independent predictors we can fit the data perfectly.

Would such a model be of any use?

Overfitting happens when the model fitted on the data describes it way too well, and ends up explaining the random error or noise in the data.

In such a situation, it is unlikely for the model to fit another sample of the same population: the model does not generalize, that is the model is not able to respond to new situations.

The best way of understanding if the model is overfitting is by testing the performance of the model on a separate dataset: if the performance is inferior to what we expected, then overfitting has probably occurred.

Further reading (optional)

For good presentation of linear regression, refer to the book:

- *Introduction to Statistical Learning*, Chapter 3, in particular sections 3.1-3.3. Examples of using R to fit linear regression models are in section 3.6.

A nice and easy presentation of the issues related to hypothesis testing, statistical significance and power are in the following paper:

- R.L. Lieber, *Significance and statistical power in hypothesis testing*, Journal of Orthopaedic Research (1990)