Applied Statistics and Data Analysis

4. Linear regression with a single predictor

Paolo Vidoni

Department of Economics and Statistics
University of Udine
via Tomadini 30/a - Udine

paolo.vidoni@uniud.it

Based mainly on Chapter 5 of the course textbook *Regression with a single predictor*

Table of contents

Summary

- Introduction to linear models
- Fitting a line to data
- Confidence and predictions intervals
- ANOVA models
- Regression diagnostics
- Logarithmic and other transformations
- The matrix form

Introduction to linear models

- An important objective in scientific research concerns the study of the relation (useful for both interpretation and prediction purposes) among a **response variable** and some **explanatory variables** (**regressors**, **predictors** or **covariates**).
- The focus here is on the straight line model, namely the **simple linear regression model** which is based on the linear relation of *one response variable* and a *single predictor variable*.
- Data for which these models may be appropriate can be displayed as a scatterplot. By convention, the x-variable, plotted on the horizontal axis, has the role of explanatory variable, whereas y-variable, plotted on the vertical axis, has the role of response or outcome variable.
- Although the interest is on the linear model, the use of transformations makes it possible to accommodate specific forms of non-linear relationship within this framework.
- Many of the issues that arise for these simple regression models are fundamental to any study of regression methods.

Table of contents

Data about two variables

- If data about the response Y and the regressor X are available, it is appropriate to start from a data visualization by means of a scatterplot, perhaps supplemented by the analysis of the correlation.
- If there are many observations, it is often useful to compare the fitted line with a fitted smooth curve. If they differ substantially, then straight line regression could be inappropriate.
- Fitting a straight line is often quite natural, and this corresponds to assume a **simple linear regression model** defined as

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

  where, given the x_i, the error term is **normally distributed**, namely $\varepsilon_i \sim N(0, \sigma^2)$, and errors of different units are **independent**.
- This amounts to say that, **given** x_i (which is taken as fixed in regression models), the i-th response is $Y_i \sim N(\alpha + \beta\, x_i, \sigma^2)$, independent from the other responses.

- The normality assumption and the independence assumption are important for obtaining confidence intervals and perform hypothesis testing on the model parameters, and on β in particular (β is the key parameter, connecting the two variables together).
- Two types of predictor variable may be considered:
  - *metric predictor variables*, that is measurements of some quantity that may help to predict the value of the response (i.e. if the response is the blood pressure of patients, then age or fat mass are potential metric predictors);
  - *factor predictor variables*, that is labels that serve to categorize the response measurements into groups, as in the ANOVA model (i.e. if the response is the blood pressure of patients, then a factor may be the drug treatment).
- Here, metric predictor variables are mainly considered and the case of factor predictor variables is briefly discussed, with the aim of introducing ANOVA models and comparing regression models and ANOVA models when the levels of the factor are quantitative.

Estimation and testing

- A popular procedure for estimating the regression parameters α, β is the **least squares method**, giving

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}, \quad \widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  These estimators are unbiased and consistent; as a consequence of the normality assumption, they correspond to the MLE's.

- Given the standard errors $\mathrm{SE}(\widehat{\alpha})$, $\mathrm{SE}(\widehat{\beta})$, the test statistics for the nullity of the coefficients (that is, $H_0 : \alpha = 0 \,\mathrm{vs}\, H_1 : \alpha \neq 0$ and $H_0 : \beta = 0 \,\mathrm{vs}\, H_1 : \beta \neq 0$) are, under H_0,

$$\frac{\widehat{\alpha}}{\mathrm{SE}(\widehat{\alpha})} \sim t(n-2), \quad \frac{\widehat{\beta}}{\mathrm{SE}(\widehat{\beta})} \sim t(n-2)$$

- Testing the null hypothesis that $\beta = 0$ is particularly relevant; a small p-value leads to the rejection of H_0 and it is consistent with an evident linear trend.

- Using the parameter estimates $\widehat{\alpha}$, $\widehat{\beta}$, it easy to estimate the means $\mu_i = \alpha + \beta x_i$ of the response r.v.'s Y_i, which correspond to the **fitted values** (prediction of points on the fitted line)

$$\widehat{y}_i = \widehat{\mu}_i = \widehat{\alpha} + \widehat{\beta}\, x_i,\ i = 1, \ldots, n$$

- The observed **residuals** are given by

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{\alpha} - \widehat{\beta}\, x_i,\ i = 1, \ldots, n$$

- An estimate for σ is the **residual standard error**

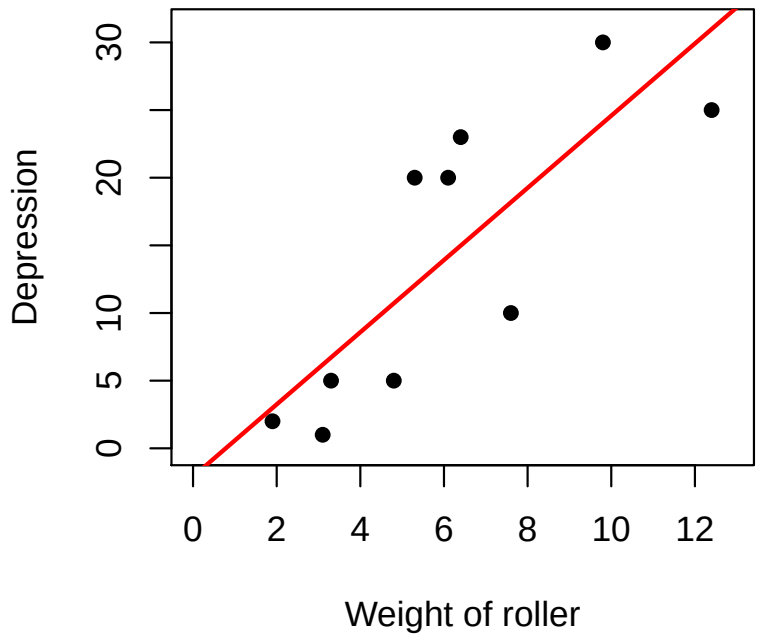
$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}\, x_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}{n-2}}$$

with $n-2$ being the **degrees of freedom**.

Example: roller data

Experiment where different `weights` (t) of roller were rolled over different part of a lawn, and the `depression` (mm) measured.

The scatterplot of the data, with the fitted regression line is given below



The intercept of the fitted line is $\widehat{\alpha} = -2.09$ ($\mathrm{SE}(\widehat{\alpha}) = 4.75$), the estimated slope is $\widehat{\beta} = 2.67$ ($\mathrm{SE}(\widehat{\beta}) = 0.70$). The standard deviation σ of the noise term is estimated by the residual standard error 6.735, with 8 d.f.

The p-value for the slope (testing $\beta = 0$) is 0.005, consistent with the evident linear trend.

The p-value for the intercept (testing $\alpha = 0$) is 0.67, i.e. the difference from zero may well be random sampling error (it would be reasonable to fit a model that lacks an intercept term).

Fitted values (red points on the fitted line) and observed residuals (blue segments) for the roller data

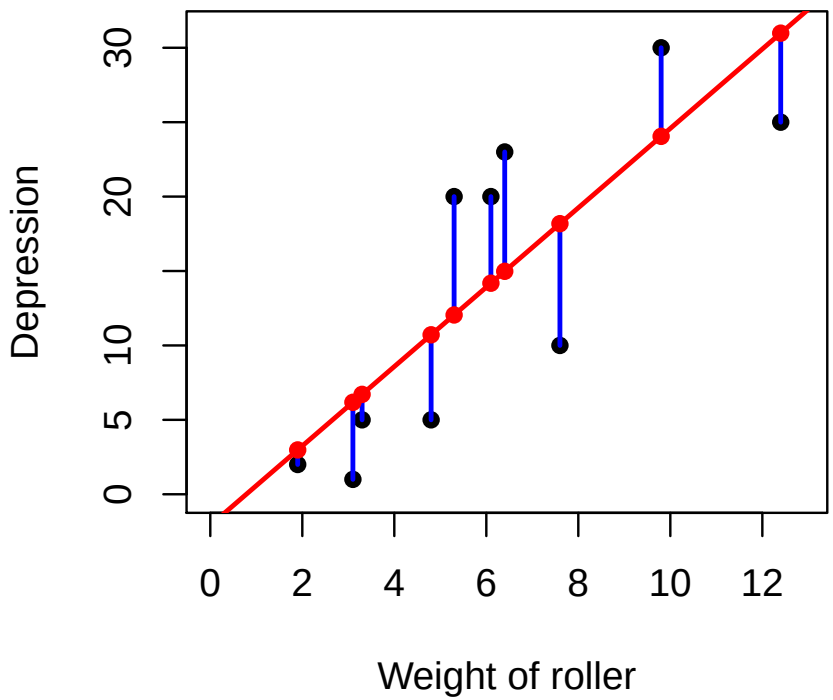


Table of contents

Confidence intervals

- Confidence intervals may be calculated for the model parameters and for the regression line at some given value x_0, namely $\mu_0 = \alpha + \beta\, x_0$.
- A 95% confidence interval for β has the form

$$\left[\widehat{\beta} \pm t_{n-2;0.025}\, \mathrm{SE}(\widehat{\beta})\right]$$

and analogously for α.

- A 95% confidence interval for μ_0 has a similar form and it is given by

$$[\widehat{\mu}_0 \pm t_{n-2;0.025}\, \mathrm{SE}(\widehat{\mu}_0)]$$

with $\widehat{\mu}_0 = \widehat{\alpha} + \widehat{\beta}\, x_0$ and

$$\mathrm{SE}(\widehat{\mu}_0) = \sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where σ can be estimated by $\widehat{\sigma}$. Note that $\mathrm{SE}(\widehat{\mu}_0)$ is affected by how far x_0 is from the covariates sample mean $\bar{x}$.

Prediction intervals

- A **prediction interval** for the response r.v. $Y_0 = \mu_0 + \varepsilon_0$ at a new predictor value x_0 can also be obtained. In this case the interval provides a set of values for a r.v. and it incorporates also the variability due to the random term ε_0.
- The best **point predictor** $\widehat{Y}_0$ for Y_0 is again $\widehat{\mu}_0$, namely

$$\widehat{Y}_0 = \widehat{\mu}_0 + \widehat{\varepsilon}_0 = \widehat{\mu}_0 + 0 = \widehat{\mu}_0$$

since the best prediction of the random term ε_0 is $E(\varepsilon_0) = 0$.

- The **prediction error** is $Y_0 - \widehat{Y}_0 = Y_0 - \widehat{\mu}_0$ and then

$$E(Y_0 - \widehat{\mu}_0) = 0, \quad V(Y_0 - \widehat{\mu}_0) = \sigma^2 + \mathrm{SE}(\widehat{\mu}_0)^2$$

The square root of $\sigma^2 + \mathrm{SE}(\widehat{\mu}_0)^2$, describing the variability of the point predictor, defines the **standard error of prediction**, whose estimate is denoted as $\mathrm{SE}(\widehat{\mathrm{Y}}_0)$, which is greater than $\mathrm{SE}(\widehat{\mu}_0)$.

The 95% prediction interval for Y_0 (wider than that for μ_0) is

$$[\widehat{Y}_0 \pm t_{n-2;0.025}\, \mathrm{SE}(\widehat{\mathrm{Y}}_0)]$$

Example: roller data

For the roller data, the 95% confidence interval for β is

$$[2.67 \pm 2.31 \times 0.70] = [1.05, 4.28]$$

It does not contain 0; as seen before, the null hypothesis $H_0 : \beta = 0$ is rejected with a 5% significance level.

Figure below shows the 95% pointwise confidence bounds for (points μ_0 on) the fitted line (dashed lines) and the 95% pointwise prediction bounds for new data Y_0 with different values for x_0 (**dashed lines**).

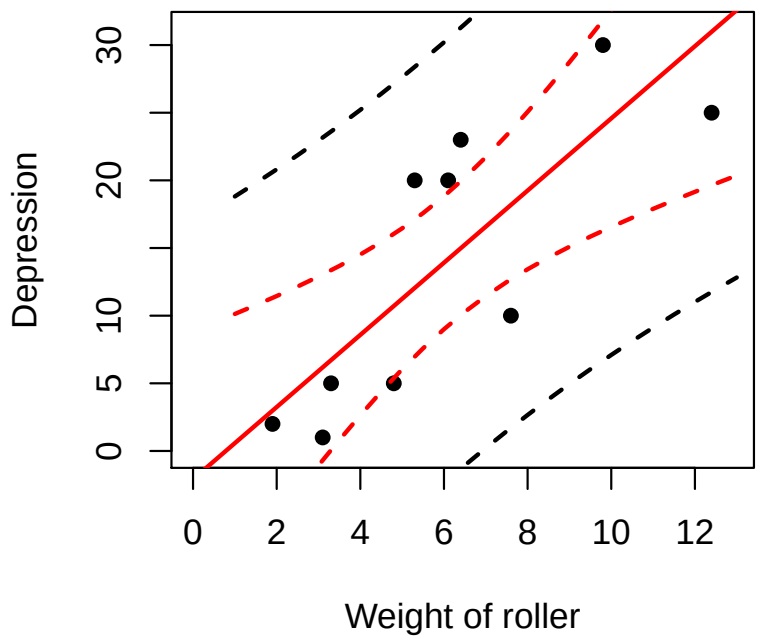


Table of contents

One-way ANOVA

- **One-way analysis of variance** (**ANOVA**) is a set of techniques to compare the means of several groups, generalizing two-sample comparisons.
- It can be interpreted as the study on how the **mean level** of a continuous response variable depends on the **level of a factor**, which may be viewed as a categorical-type regressor.
- It can be extended to more than one factor, obtaining **two-way** or **multi-way ANOVA**.
- It is widely used in data from designed experiments, but it has a role also with observational data; indeed, it is an inferential method sometimes employed as an EDA tool.
- The underlying statistical model is a special case of a linear regression model.

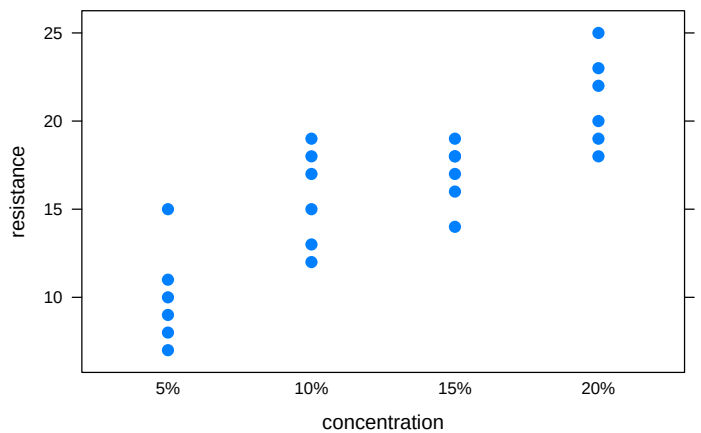
Example: paper resistance

The table below reports the data on an experiment to study the relation between `paper resistance` and `wood fibre concentration` in pulp.

There are 4 different levels of concentrations, and 6 trials are made at each level (the data are **balanced**, as appropriate for ANOVA)

	observation							
concentration	1	2	3	4	5	6	Total	Mean
5%	7	8	15	11	9	10	60	10.00
10%	12	17	13	18	19	15	94	15.67
15%	14	18	19	17	16	18	102	17.00
20%	19	25	22	23	18	20	127	21.17

A graphical representation of the data is given below



- The stripplots display *within-group* variability and give an indication of differences among the group means (of the response variable `resistance`).
- There is a single explanatory variable (regressor), namely the `concentration`, with one level for each of the different treatments that were applied.
- Variances seem similar for the four different treatments (the levels of `concentration`).
- A simple-minded approach is to calculate the means for each of the four treatments, and then examine all pairwise comparisons.
- The use of an analysis of variance (really, as noted below, the fitting of a linear model) enables an overall analysis.

Statistical model for one-way ANOVA

- The setting is that there are a levels of a factor of interest (*treatment*), which identify a different groups of observations.
- The statistical model for one-way ANOVA is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where $i = 1, \ldots, a$, $j = 1, \ldots, n$, and

  - i identifies the treatment level;
  - j identifies the observation;
  - μ is the **general mean**;
  - τ_i is the **treatment effect**, namely the deviation from the general mean for the i-th group;
  - ε_{ij} is a random error.

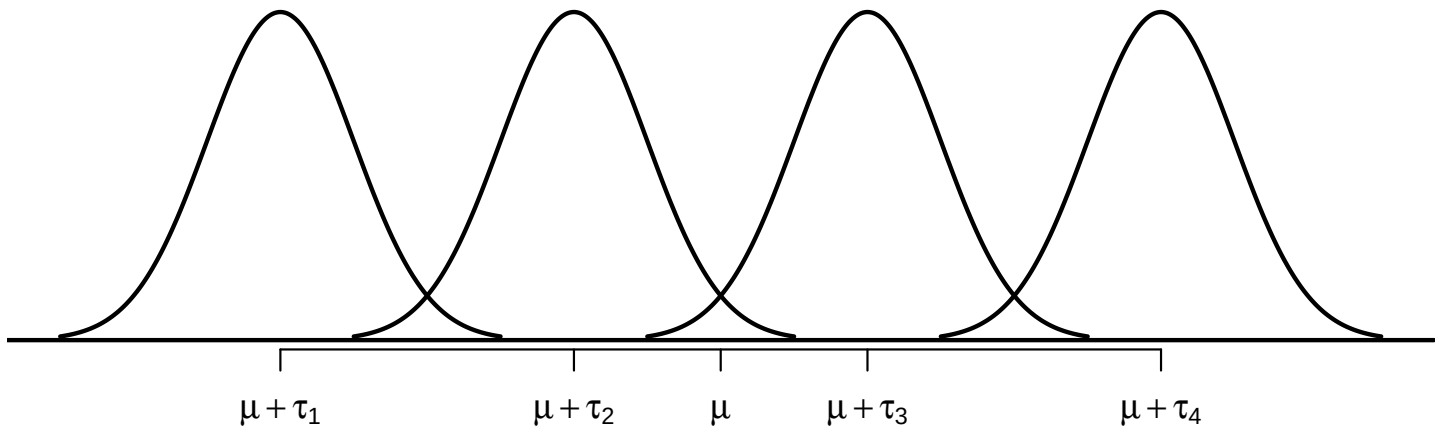
- The random errors are assumed i.i.d. normal distributed, that is

$$\varepsilon_{ij} \sim N(0, \sigma^2) \qquad \text{i.i.d.}$$

  Namely, the Y_{ij} are independent r.v.'s with

$$Y_{ij} \sim N(\mu + \tau_i, \sigma^2)$$

- For example, if there are $a = 4$ levels, the general mean is $\mu = 0$ and the treatment effects are $\tau_1 < \tau_2 < 0$ and $\tau_4 > \tau_3 > 0$, the probability distributions describing the $a = 4$ groups are such that



Hypothesis testing in one-way ANOVA

- The null hypothesis states that **all the means are equal**

$$H_0: \quad \tau_1 = \tau_2 = \cdots = \tau_a = 0$$
$$H_1: \quad \tau_i \neq 0 \text{ for at least one group}$$

  If H_0 is true, the data are a random sample from a $N(\mu, \sigma^2)$ distribution, with **no effect** of the factor on the mean response.

- The test is performed by comparing two estimates of the variance σ^2

$$\widehat{\sigma}_0^2 = \frac{\sum_{i=1}^{a} n\,(\bar{y}_i - \bar{y})^2}{a-1} \qquad \widehat{\sigma}^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2}{a\,n - a}$$

  where $\bar{y}_i$ are the **group means**, and $\bar{y}$ the **grand mean**.

- The numerator of $\widehat{\sigma}_0^2$ and $\widehat{\sigma}^2$ are referred to as **between-group sum of squares** and **residuals sum of squares**, and their denominators as treatment d.f. and residual d.f.

- While $E(\widehat{\sigma}^2) = \sigma^2$, for $\widehat{\sigma}_0^2$ it holds $E(\widehat{\sigma}_0^2) = \sigma^2$ *only when* H_0 is true: from their comparison we can glean some information on the plausibility of H_0.
- Indeed, the F test for one-way ANOVA is just the ratio $F = \widehat{\sigma}_0^2/\widehat{\sigma}^2$ of the two estimators of σ^2 and under H_0

$$F = \frac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2} \sim F(a-1, an-a)$$

- The results of the analysis can be trusted as long as the statistical model is reasonable; model checking can be made following the guidelines for linear regression models.
- The nonparametric equivalent of the F test is the **Kruskal-Wallis test** based on ranks, which extends the Mann-Whitney U test when there are more than two independent groups.

  The normal distribution for the residuals is not required and, when the group distributions have the same shape and scale, the comparison concerns the group medians.

Post-hoc analysis

- If the p-value is small, the ANOVA test is significant, so that at least one mean must differ from the others.
- In this case, it is recommendable to investigate the results by means of a **post-hoc analysis**, such as the **Least Significant Difference** (**LSD**) or the **Tukey's Honest Significant Difference** (**HSD**).
- LSD and HSD perform testing on which effects are significantly different at a given level (5% or 1% are typically used); both these two procedures provide a fixed quantity as output, and effects which differ by less than it are deemed as not statistically different.
- LSD does not take into account that there are $a\,(a-1)/2$ possible comparisons rather than a single test, hence it is often *liberal*: the reported output is too short and tends to differentiate too much the various group effects.

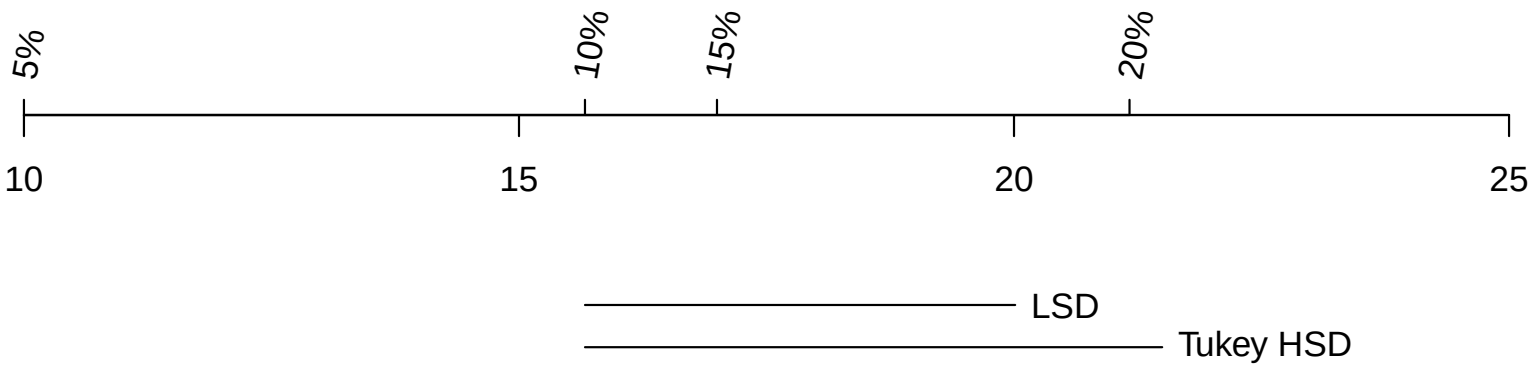
  On the contrary, HSD focuses on the *maximum difference*, hence it is *conservative*: the reported output is too long and blurs the comparisons; considering both methods is therefore a sensible approach.

Example: paper resistance

For the data on `paper resistance` at $a = 4$ different level of `wood fibre concentration`, it easy to obtain $\widehat{\sigma}_0^2 = 127.6$, with $a - 1 = 3$ d.f., and $\widehat{\sigma}^2 = 6.51$, with $a\,n - a = 20$ d.f.

The observed value of the F statistic is $F = 19.605$, that gives a p-value $p = 3.6 \cdot 10^{-6}$, leading to a substantial evidence against H_0.

Post-hoc analysis here is useful, and it is provided by considering the LDS and HDS results summarized in the following figure



The effects for 10% and 15% concentration are surely equivalent, while, for example, the effects for 5% and 15% are different.

Regression vs qualitative ANOVA

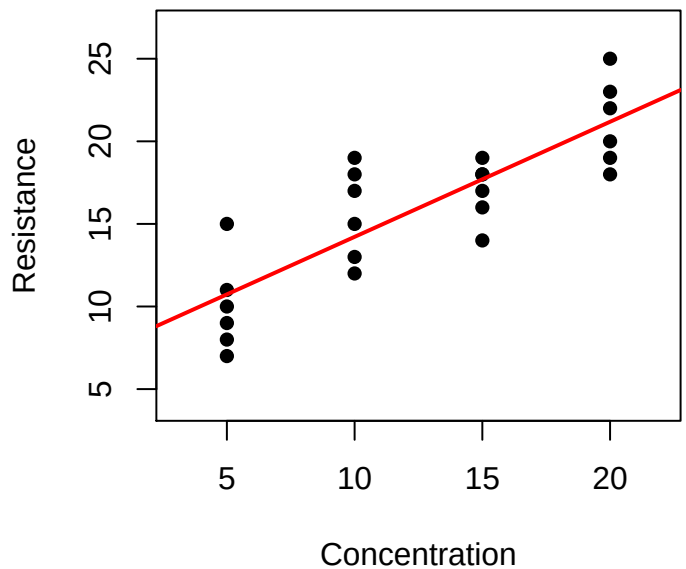
- The aim is to compare regression models and ANOVA models, when the levels of the factor are quantitative. In this context, the factor in the one-way ANOVA may be considered as the predictor in a simple linear regression model.
- In the ANOVA framework the statistical tests for *qualitative* differences between treatment effects ignore the fact that the levels are *quantitative*.
- A test for linear trend is more powerful than an analysis of variance test, that treats the levels as qualitatively different levels. The p-values in the first case will on average be smaller.
- Fitting a line or a curve, where this is possible, rather than fitting an analysis of variance model that has a separate parameter for each separate level of the explanatory variable, takes proper advantage of structure in the data.

  This allows interpolation between successive levels of the explanatory variable and it enables a convenient description for the pattern of the response variable.

Example: paper resistance

Data on `paper resistance` and `wood fibre concentration`: 4 levels of concentrations (5%, 10%, 15%, 20%) and 6 trials at each level.

One-way ANOVA uses a different mean for each concentration, but do not use the information about the actual amount. This can be done by fitting a linear model, which passes all the diagnostic checks



The p-value for the slope β is $2.43 \cdot 10^{-7}$, whereas that one found by the ANOVA analysis is $3.6 \cdot 10^{-6}$. Both values suggest a strong relation, but the p-value for the linear model is about 10 times smaller.

Table of contents

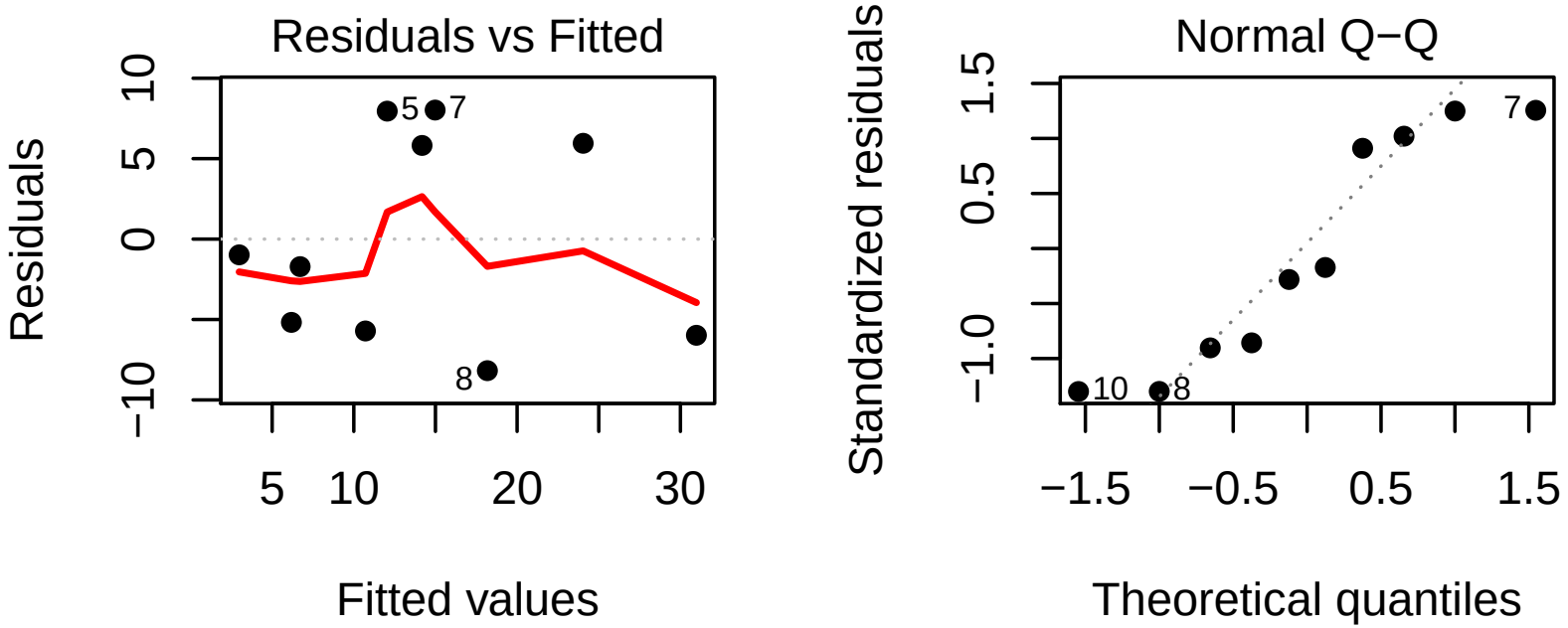
Checking the residuals

- A crucial part of any model fitting concerns judging if the model is appropriate for the data, by **checking residuals and outliers**. With small data sets, departures from assumptions will be hard to detect.
- As it is not possible to observe the error term, the observed residuals are used instead. In particular, various **residual plots** are routinely considered:
  - plot of the observed residuals $\widehat{\varepsilon}_i$ against the fitted values $\widehat{y}_i$, for checking lack of systematic patterns (e.g. correlation and clustering);
  - plot of the square root of absolute values of the residuals $\sqrt{|\widehat{\varepsilon}_i|}$ against the fitted values $\widehat{y}_i$, for checking if variance is constant;
  - plot of $\widehat{\varepsilon}_i$ against x_i, useful for detecting nonlinearity effects of x_i (for a single predictor is equivalent to the plot of $\widehat{\varepsilon}_i$ *vs* $\widehat{y}_i$);
  - normal probability plot for checking the normality assumption.
- It is possible to show that, no matter whether the model is a good one or not, the residuals invariably have sample mean equal to 0, and variance that depends (to some extent) on x_i.

Example: roller data

For the roller data, the *left panel* gives some mild suggestion of clustering, but the sample size is too small to be sure about it.

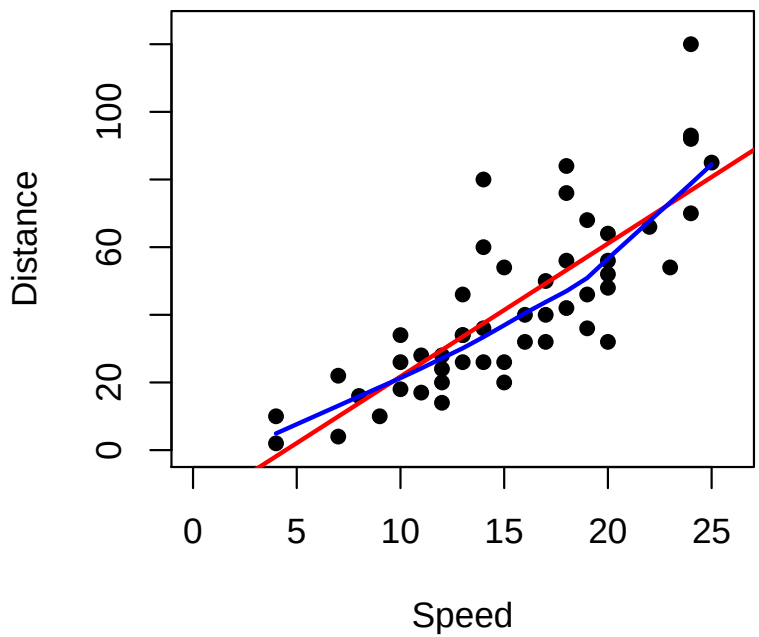
It is not easy to interpret the normal probability plot (*right panel*), due to the small data set and the lack of a reference standard. It could be useful to compare this plot against a number of independent plots from normal simulated data with the same number of observations.



Example: cars

The data set gives the speed of cars (mph) and the `distance` taken to stop (ft). The observations, recorded in the 1920s, regard $n = 50$ cars.

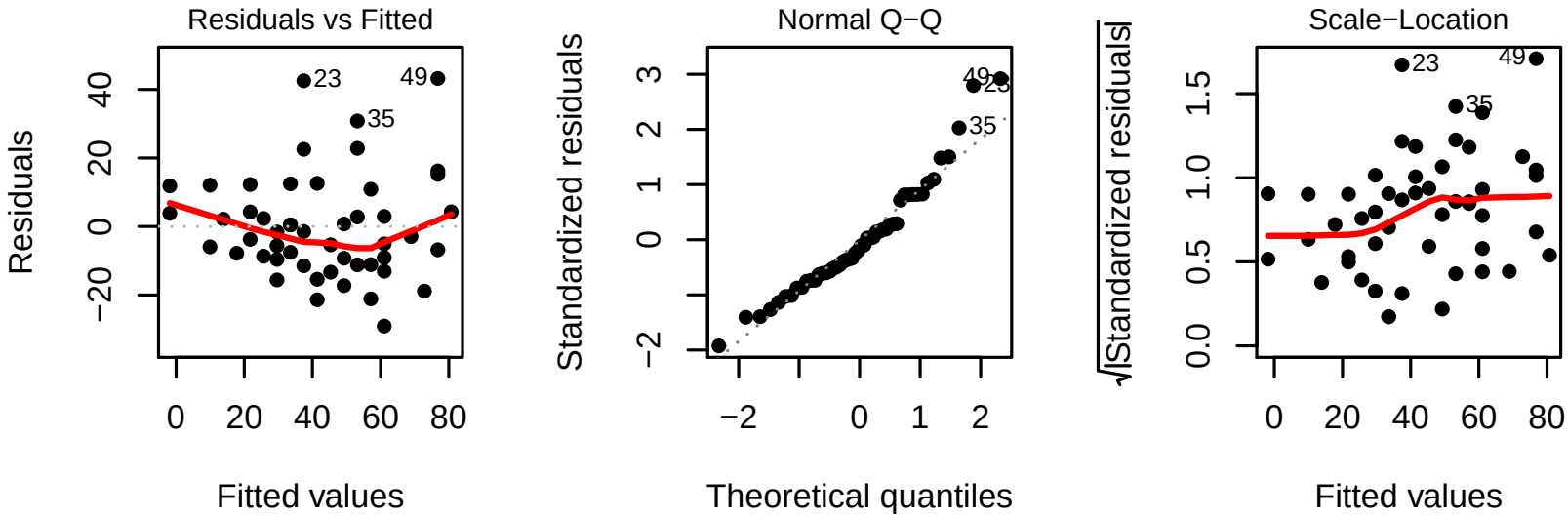
The scatterplot of the data, with the fitted regression line and the fitted smooth curve, is given below



The intercept of the fitted line is $\widehat{\alpha} = -17.58$ ($\mathrm{SE}(\widehat{\alpha}) = 6.76$), the estimated slope is $\widehat{\beta} = 3.93$ ($\mathrm{SE}(\widehat{\beta}) = 0.42$). The p-value for the slope is closed to 0, whereas the p-value for the intercept is 0.012.

The smooth curve gives a better indication of the pattern in the data than the straight line.

The following three diagnostic plots confirms that the simple linear regression model does not gives an adequate description for the cars data



The curvature in the first plot (correlation for the residuals) and in the third plot (non constant variance) is apparent.

The residuals are from the straight line model. Different conclusions could be obtained from the equivalent plots based on residuals from the smooth curve.

The ANOVA results and the R^2

- Also for linear regression models it is possible to produce the ANOVA results, which decompose the **total variability** of the response variable into two parts:
  - a part accounted for by the linear model;
  - a residual part, that for a good model should be small.
- In this context, the F test gives a statement on model adequacy which is analogous to that obtained from test statistic for the slope β.
- The total variance of the response variable Y can be evaluated by considering the **total sum of squares** (about the mean)

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathrm{SSM} + \mathrm{SSE}$$

which can be expressed as the composition of the **sum of squares accounted for by the linear model** $\mathrm{SSM} = \sum (\widehat{y}_i - \bar{y})^2$ and the **residual sum of squares** $\mathrm{SSE} = \sum (\widehat{y}_i - y_i)^2 = \sum \widehat{\varepsilon}_i^2$.

- The contribution of X in explaining the variability of Y can be summarized with the quantity R^2 (**coefficient of determination**)

$$R^2 = \frac{\text{SSM}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

  that indicates the proportion of the total variability of Y which is accounted for by the linear function of the predictor X.

- The R^2 statistic corresponds to the square of the Pearson correlation coefficient $\widehat{\rho}_{XY}$ and its values ranges from 0 in case of no contribution of X (horizontal regression line), to 1 in case of perfect regression (all observations on the regression line).
- A less optimistic measure (in general preferable to R^2) is the

$$\text{adjusted } R^2 = 1 - \frac{\text{SSE}/\text{d.f. SSE}}{\text{SST}/\text{d.f. SST}}$$

  taking into account the degrees of freedom of SSE and SST.
- Neither statistic gives any direct indication of how well the regression equation will predict when applied to a new data set.

Example: roller data

For the roller data, the total sum of square is $\mathrm{SST} = 657.97 + 362.93 = 1020.90$ and the introduction of `weight` cuts it down to 362.93.

Then the coefficient of determination is

$$R^2 = \frac{1020.90 - 362.93}{1020.90} = 1 - \frac{362.93}{1020.90} \doteq 0.64$$

that shows that about 64% of the total variability is accounted for by the linear function of `weight`.

Such result is achieved by passing from the model that just uses only the sample mean to the linear model, which uses two coefficients: the number of degrees of freedom available for estimating the noise variance and the total variance are $10 - 2 = 8$ and $10 - 1 = 9$, respectively.

The adjusted R^2 takes this into account and corresponds to

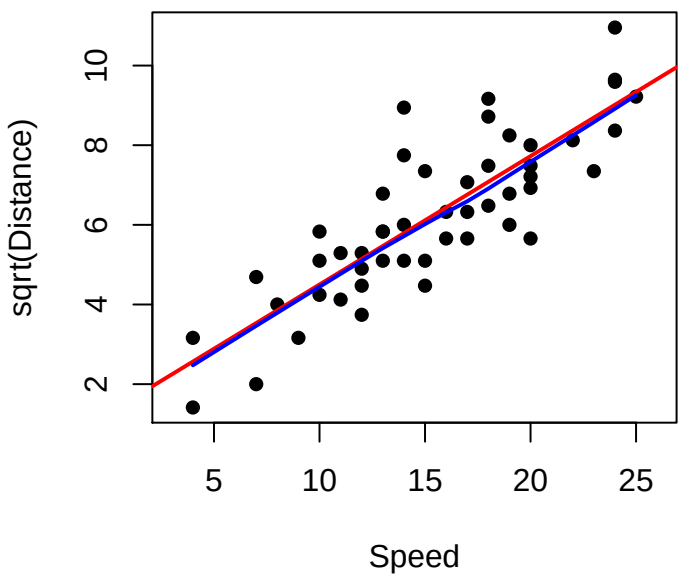
$$\text{adjusted } R^2 = 1 - \frac{362.93/8}{1020.90/9} \doteq 0.60$$

Example: cars

For the cars data, $R^2 = 0.651$ (adjusted $R^2 = 0.644$) shows that about 65% of the total variability of `distance` is described by the linear function of `speed`.

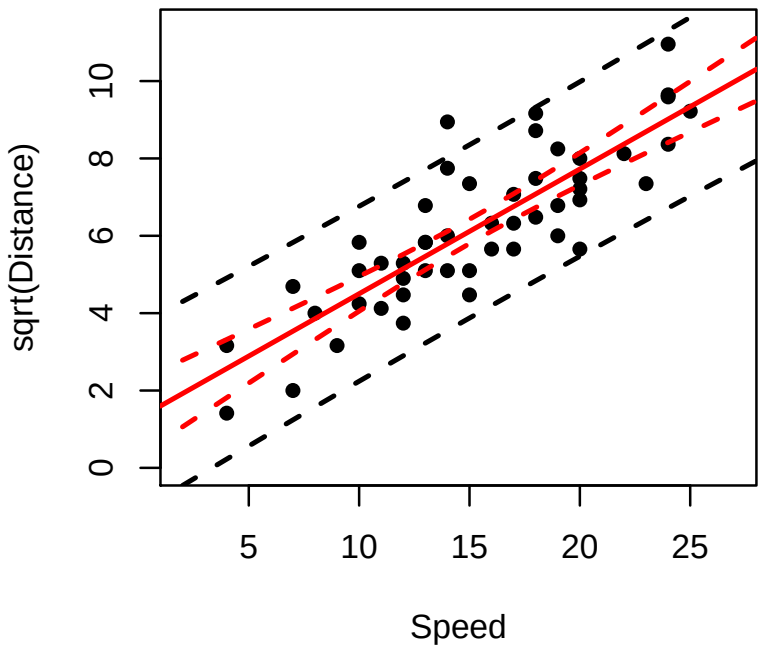
In order to account for the non-linearity, that appears in the scatterplot, a linear model with $\sqrt{\texttt{distance}}$ as response in considered.

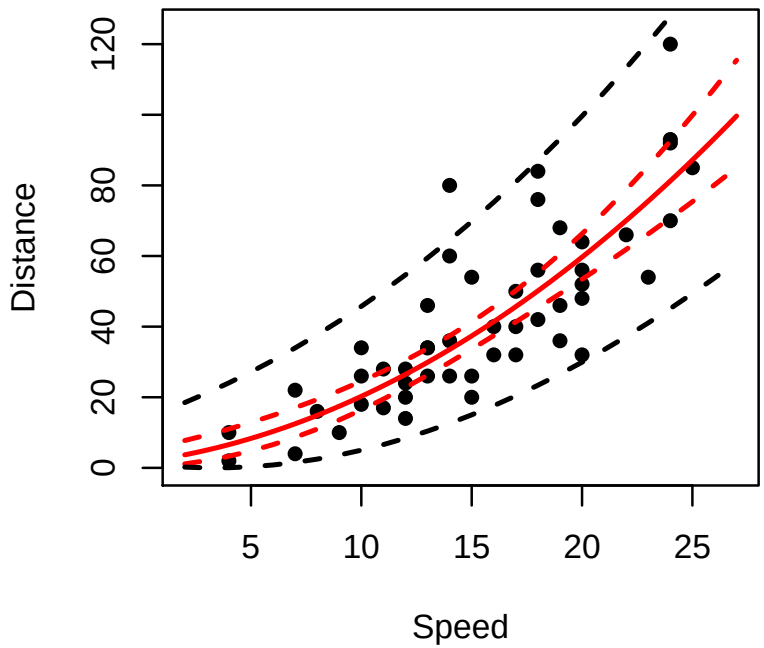
In this case, the coefficient of determination increases to $R^2 = 0.709$ (adjusted $R^2 = 0.702$) and the improved fit is confirmed by the following scatterplot, with the fitted regression line and the fitted smooth curve



For the linear model taking $\sqrt{\texttt{distance}}$ as response, the 95% confidence bounds for the mean response (dashed lines) and the 95% prediction bounds (**dashed lines**) are computed for a range of speed values.

The result are reported both on the transformed and on the original scale; the latter choice is safer.





Outliers, leverage and influence

- **Outliers** are points that lie away from the bulk of the data.
- Outliers are important because (i) they may carry very useful information, as they may corresponds to the best (or the worst) conditions and (ii) they may have an undue influence on the conclusions.
- Often, but not always, they correspond to points with a large residual. To this end, it is useful to introduce the concepts of **leverage** and **influence**.
- The leverage of a data point describes the (potential) impact on the fitted line of moving the point on the y-coordinate. **Points with high leverage** are at the extreme end of range of x-values; they *may* exert a greater pull on the regression line than points towards the center of the range.
- **Influential points** have a strong influence on the model results, and if omitted they would change the fitted line; they are points with a large residual, a large leverage or both.

Identification and treatment of outliers

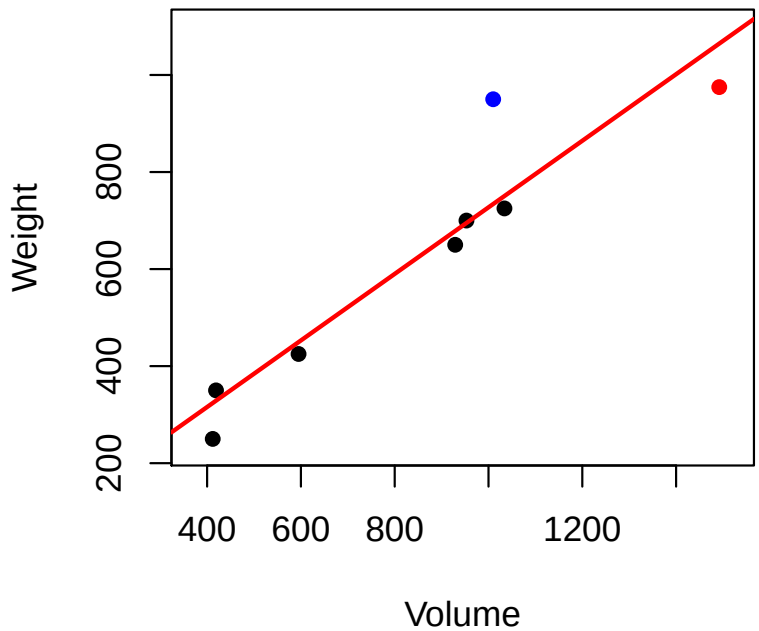
- Points with high leverage are flagged in the diagnostic plots and they usually correspond to points with large or small x-values.
- Influential points can be detected by means of the **Cook's distance**; it measures the change in the fitted line if the point were omitted.

  Points with Cook's distances greater than one, or substantially larger than for other points, require investigation.

  Looking at residuals may not reveal influential points, since an outlier with high leverage tend to attract the fitted line and therefore it may have a small residual.
- If some outliers are identified, the first thing to do is to check them carefully, as they may just be a recording error. If an outlier seems a genuine data value, a good practice is to perform the analysis both with and without the outlier, to assess its impact on the conclusions.
- Another possibility is to use **robust methods**, which fit a linear regression model reducing the influence of outliers.

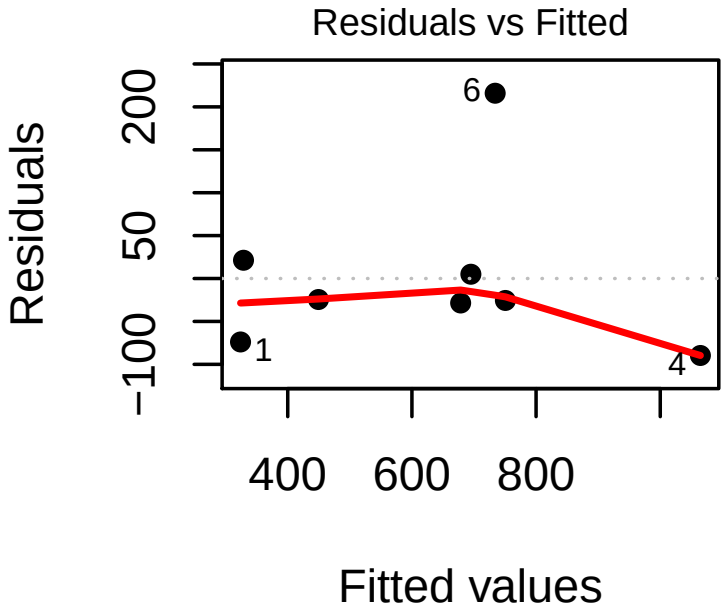
Example: books

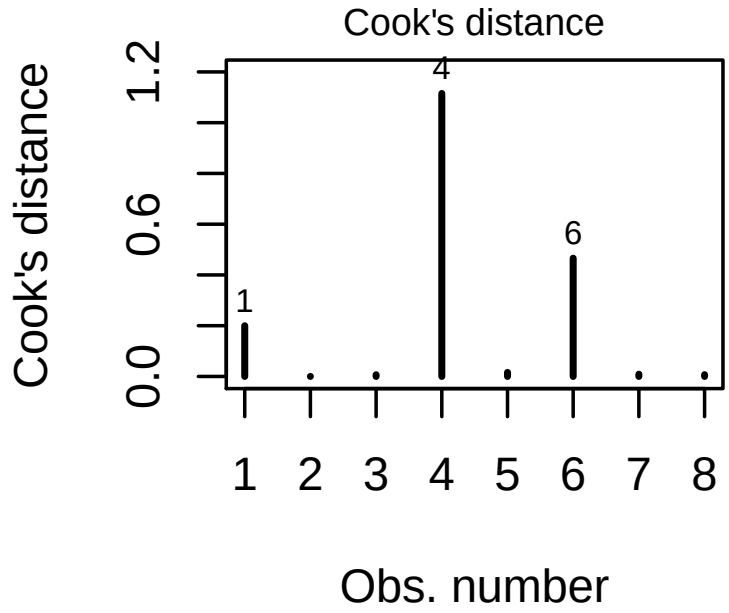
The data set gives the `volume` (cm^3) and the `weight` (gr) of $n = 8$ paperback books. The scatterplot of the data, with the fitted regression line, is given below



The intercept of the fitted line is $\widehat{\alpha} = 41.37$ ($\text{SE}(\widehat{\alpha}) = 97.56$), the estimated slope is $\widehat{\beta} = 0.686$ ($\text{SE}(\widehat{\beta}) = 0.11$). The p-value for the slope is 0.0006, whereas the p-value for the intercept is 0.686.

Indeed, $R^2 = 0.875$ and $\text{adjusted } R^2 = 0.854$.





Observations 4 (red point in the scatterplot) and 6 (blue point in the scatterplot) are influential points; observations 4 is also a leverage point.

Although observation 6 has the largest residual, its Cook's distance is relatively small.

Observation 4 has the largest Cook's distance. In part, because this point is a high leverage point. Since its y-value is lower than would be predicted by the line, it pulls the line downward.

Points 4 and 6 are both candidates for omission, however with only eight observations, it would not make sense to omit any of them.

Assessing the predictive accuracy of a model

- An important definition in Statistics (and in Machine Learning) is that of **training data**: the data set used to estimate a given model.
- If the training data is used also to evaluate the **predictive accuracy** of a model, this gives an *optimistic assessment* because the same data are used twice.

  An attempt to correct for this fact motivates the use of the adjusted R^2 in place of the R^2 for simple linear regression models.
- The ideal approach is to assess the performance of the model on a new data set. Then it is a good practice to evaluate the predictive accuracy of a given model by splitting the data into two separate groups:
  - the **training set**, used to estimate the model;
  - the **test set**, used to assess its predictive performance.

  This is a very general idea that can be applied to nearly every statistical method.

Cross-validation

- When this approach is not practical, mainly due to limitations of the available data, **cross-validation** can be used instead.

  This consists in splitting the data in k sets (*folds*), which are used in turn as a test set, with the remaining folds giving the training data.

  The predictive assessments for each of the k folds are then combined together into a single measure. Values of k between 3 and 10 are typically used.

- Several measures of performances may be defined.

  For linear regression models, the estimate of σ^2 (or of the sum of squared residuals or of the mean square error) is a good choice, as it corresponds to the residual variability left unexplained by the regression line.

- A particular case of the k-fold procedure is the **leave-one-out cross-validation**, where k is set equal to the sample dimension n, so that each single observation is used in turn as a test set.

Table of contents

Choosing the right scale

- Diagnostic plots often point out that something is wrong with the model assumptions.
- In many cases, it is just a matter of choosing the right scale for the response variable. There two notable special cases:
  - **logarithmic scale**, useful for size measurements of biological organisms, but also for many socio-economic variables that are highly skewed, such as income or satisfaction; also a good option if the ratio between the largest and the smallest response value is large, namely greater than 10 or even 100;
  - **square root scale**, useful for count data, it often stabilizes the variance and it is actually a special case of power transformations y^p, with $p = 1/2$; also $p = 1/3$, the **cube root scale**, is at times useful.
- Though the usual course is to transform the y-values, in some instances it may make sense to transform the x-values instead, or both.

The Box-Cox transformation

- Choosing the right scale for the response is often a matter of trial and error, and it might be convenient to be able to do it in a semi-automatic way.
- The **Box-Cox transformation** provides exactly that for models with positive responses. It corresponds to a *general power transformation*, depending on a real parameter λ

$$y(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}$$

- The specification of a confidence interval for λ may give a quite useful indication. In particular, this would quickly suggest whether the log or the power transformations are likely to work well.

Example: cars

For the cars data, the following plot displays a sort of log-likelihood function for λ and the 95% confidence interval for λ can be read off.

The value $\lambda = 0.5$, corresponding to the square root transformation used previously, is among the most supported values.

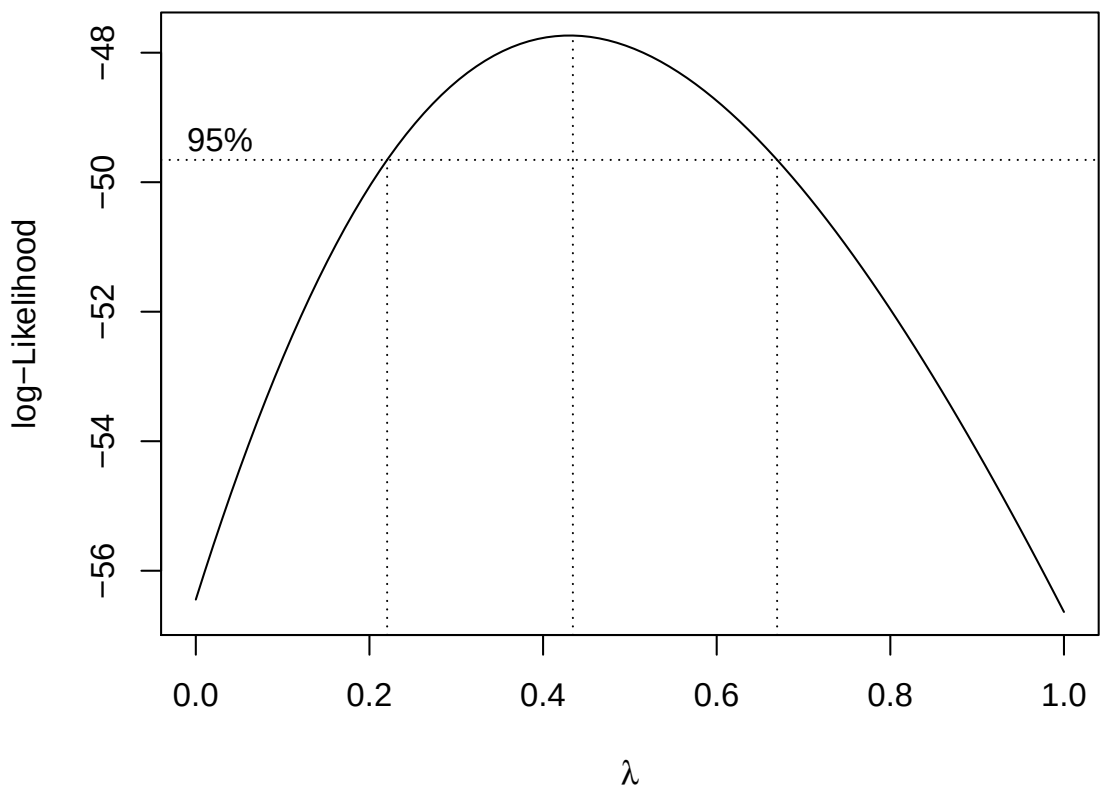


Table of contents

The matrix form of simple linear regression

- It can be useful to rewrite the simple linear regression model in the following compact matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

  - $\mathbf{y} = (y_1, \ldots, y_n)^T$ is column vector collecting all the response values;
  - $\mathbf{X}$ is the $n \times 2$ **model matrix** (or **design matrix**) defined as

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

  - $\boldsymbol{\beta} = (\alpha, \beta)^T$ is the column vector with the model coefficients;
  - $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the column vector collecting all error terms.
- It is not difficult to verify that this matrix form is exactly equivalent to the definition of the model given before.