Applied Statistics and Data Analysis

5. Towards multiple linear regression and logistic regression

Paolo Vidoni

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

paolo.vidoni@uniud.it

Based mainly on Chapters 6 and 8 of the course textbook

Table of contents

Summary

- Introduction to multiple regression
- Multiple linear regression: assumptions and inference
- Multiple linear regression: diagnostics
- Model assessment and model selection
- Covariates: selection and multicollinearity
- Factors as explanatory variables
- Discrete responses
- Logistic regression

Introduction to multiple regression

- In straight line regression, a response variable Y is regressed on a single explanatory variable. Multiple linear regression generalizes this methodology to allow **multiple explanatory (predictor) variables**, denoted as **covariates**.
- Multiple linear regression model is one of the most fundamental statistical model.
- It is not always the right model, as it is based on some assumptions that are not always reasonable. However, to some extent, a large number of statistical models are an extension of it.
- Even if it is a rather simple model, and just a rough representation of reality in many cases, it may be extremely useful for both interpretation and prediction purposes.

Table of contents

Model assumptions

- If data about the response Y and the p regressors $X_1, \ldots, X_p$ are available, a **multiple linear regression model** is defined as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

  where, given the covariates x_{ij}, $j = 1, \ldots, p$, the error term is **normally distributed**, namely $\varepsilon_i \sim N(0, \sigma^2)$, and errors of different units are **independent**.

- Thus, **given** x_{ij}, $j = 1, \ldots, p$, (which are taken as fixed), the i-th response Y_i is normally distributed, independent from the other responses, with constant variance σ^2 and mean defined as a linear combination of the covariates, namely

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

- In general, the assumption on $E(Y_i)$ is likely to be false, however, it can be a good approximation or a reasonable starting point for subsequent analysis.

- The multiple linear regression model can be expressed in matrix form as

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

  - $\mathbf{y} = (y_1, \ldots, y_n)^T$ is column vector collecting all the observed response values;
  - $\mathbf{X}$ is the $n \times (p+1)$ **model matrix**, with rank $p+1$ and $n > p+1$, defined as

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

  - $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the column vector with the model coefficients;
  - $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the column vector collecting the error terms.
- Then, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$, with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{I}_n$ the identity matrix.

- As for simple linear regression, the predictor variables, which form the basic ingredients for $\mathbf{X}$, can be *metric* (*numeric*) variables or *factor* variables.
- The focus here is on metric regressors, even though the case with factor regressors will be briefly discussed.
- To understand the structure of $\mathbf{X}$, the following example is considered: the response y is described by two numeric regressors, u and v, and by a factor g with labels dividing the observations into three groups.

  Factors may be included in the model by means of **dummy variables** (regressors which assume only two values, 1 and 0): in this case, there are three dummy indicators showing whether the corresponding observation belongs to the group, or not.

  With regard to numeric variable, they could enter the model also non-linearly: the model is linear in the parameters and error term, but not necessarily in the predictors.

A model matrix, accounting for non-linearities concerning numeric regressors and for a factor regressor with three levels, might be

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & u_1 & v_1 & v_1^2 & u_1 v_1 \\ 1 & 0 & 0 & u_2 & v_2 & v_2^2 & u_2 v_2 \\ 1 & 0 & 0 & u_3 & v_3 & v_3^2 & u_3 v_3 \\ 0 & 1 & 0 & u_4 & v_4 & v_4^2 & u_4 v_4 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & u_n & v_n & v_n^2 & u_n v_n \end{pmatrix}$$

Then, the first three observations are in the first group, the fourth is in the second group and the last is in the third group.

When working with factor regressors some care is required to ensure that the model matrix $\mathbf{X}$ has full rank, as required in the assumptions. Otherwise, there will be a lack of identifiability: the model parameters can not uniquely be determined from data.

Notice that, in this case, the first column in $\mathbf{X}$, which specifies a common intercept parameter, is not present.

Inference

- Point estimates of the linear model parameters $\boldsymbol{\beta}$ can be obtained by the method of least squares, giving

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

  The estimator is (minimum variance, linear) unbiased and consistent; for the normality assumption, it corresponds to the MLE.

- Since $\widehat{\boldsymbol{\beta}}$ is just a linear transformation of a normal random vector,

$$\widehat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\beta}))$$

  where the variance matrix is $\mathbf{V}(\boldsymbol{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

- Using $\widehat{\boldsymbol{\beta}}$, it easy to estimate the mean vector $\boldsymbol{\mu} = \mathbf{X}\,\boldsymbol{\beta}$ of the response random vector Y, which correspond to the **fitted values**

$$\widehat{\mathbf{y}} = (\widehat{y}_1, \ldots, \widehat{y}_n)^T = \widehat{\boldsymbol{\mu}} = \mathbf{X}\,\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$$

  with $\mathbf{H} = \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ the **hat matrix** (such that $\mathrm{tr}(\mathbf{H}) = p+1$ and $\mathbf{H}\mathbf{H} = \mathbf{H}$).

- A result, useful for testing hypotheses about individual β_j, as well as for finding confidence intervals for β_j, is

$$\frac{\widehat{\beta}_j - \beta_j}{\mathrm{SE}(\widehat{\beta}_j)} \sim t(n-p-1)$$

  where the (estimated) standard error $\mathrm{SE}(\widehat{\beta}_j)$ is given by the square root of the j-th diagonal element of matrix $\mathbf{V}(\boldsymbol{\beta})$.

- Interpreting the inferential results on the regression coefficient β_j is not as straightforward as it might appear. As a matter of fact, the p-value is used to test whether the coefficient could be zero, given that the other coefficients remain in the model (i.e. are non-zero).
- Since the estimators for the various coefficients are usually not independent, dropping one term (setting it to zero), will change the estimates of the other coefficients and hence their p-values.
- For this reason, if the aim is to refine the model by dropping some coefficients, the strategy is to drop only one term at a time (starting from those with the highest p-values) and to refit after each drop.

- It is also of interest to obtain distributional results for testing, for example, the simultaneous equality to zero of several model parameters. Such tests correspond to ***F* tests** with appropriate degrees of freedom.
- A relevant F test is that one focusing on the global significance of all the regression coefficients, namely on

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0, \text{ for at least one } j$$

- Furthermore, the observed **residuals** are given by $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{y}} = (\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n)^T$, with $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$, $i = 1, \ldots, n$.
- An estimate for σ is the **residual standard error**

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{n-p-1}} = \sqrt{\frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{n-p-1}}$$

with $n-p-1$ being the **degrees of freedom** and $\sum_{i=1}^n \widehat{\varepsilon}_i^2$ the **residual sum of squares**.

Confidence and predictions intervals

- Confidence intervals may be calculated for the model parameters and for the regression function at some given value for the regressors $\mathbf{x}_0 = (1, x_{01}, x_{02}, \ldots, x_{0p})^T$, that is for the expected response at $\mathbf{x}_0$

$$\mu_0 = \mathbf{x}_0^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \ldots + \beta_p x_{0p}$$

- A 95% confidence interval for β_j has the form

$$\left[\widehat{\beta}_j \pm t_{n-p-1;0.025}\, \mathrm{SE}(\widehat{\beta}_j)\right]$$

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

$$[\widehat{\mu}_0 \pm t_{n-p-1;0.025}\, \mathrm{SE}(\widehat{\mu}_0)]$$

with $\widehat{\mu}_0 = \mathbf{x}_0 \widehat{\boldsymbol{\beta}}$ and

$$\mathrm{SE}(\widehat{\mu}_0) = \sigma \sqrt{\mathbf{x}_0^T\, (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}$$

where σ can be estimated by $\widehat{\sigma}$.

- A **prediction interval** for the response r.v. $Y_0 = \mu_0 + \varepsilon_0$, at some new predictor values $\mathbf{x}_0$, 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** for Y_0 is again $\widehat{Y}_0 = \widehat{\mu}_0$ and the **prediction error** is $Y_0 - \widehat{Y}_0 = Y_0 - \widehat{\mu}_0$; 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

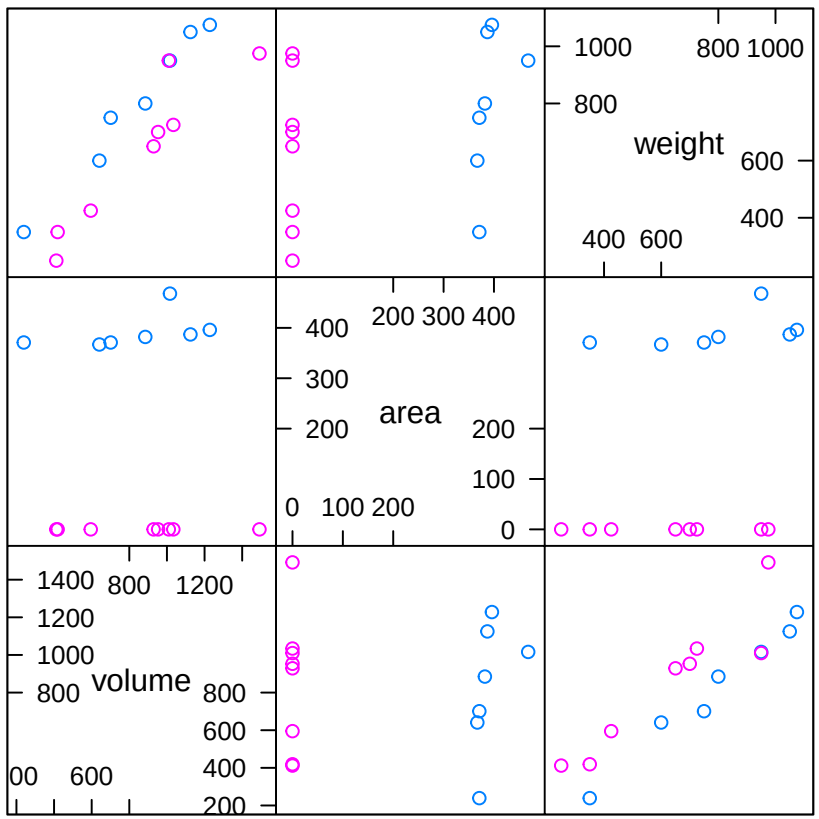
$$[\widehat{Y}_0 \pm t_{n-p-1;0.025}\, \mathrm{SE}(\widehat{\mathrm{Y}}_0)]$$

- In this framework, also more general confidence (prediction) regions or simultaneous confidence (prediction) intervals, involving more than one parameter (future observation), can be specified.

Example: book weight

Data about a sample of $n = 15$ books; the variables are `book volume` (cm^3), `hard board cover area` (cm^2), `book weight` (gr) and `cover`, a factor with levels hardback and paperback.

The scatterplot matrix for the numerical variables is given below



Leaving aside the factor `cover` (as it provides similar information to cover area), a sensible model for book weight is

$$\texttt{weight}_i = \beta_0 + \beta_1 \, \texttt{volume}_i + \beta_2 \, \texttt{area}_i + \varepsilon_i$$

The coefficient estimates are $\widehat{\beta}_0 = 22.41$ $(\mathrm{SE}(\widehat{\beta}_0) = 58.40)$, $\widehat{\beta}_1 = 0.708$ $(\mathrm{SE}(\widehat{\beta}_1) = 0.061)$, $\widehat{\beta}_2 = 0.468$ $(\mathrm{SE}(\widehat{\beta}_2) = 0.102)$, whereas the estimate for the noise standard deviation (the residual standard error) is $\widehat{\sigma} = 77.66$.

The low p-values for `volume` $(7.07 \cdot 10^{-8})$ and `area` (0.0006) highlights that they are both important predictors of book `weight`.

These results should be used informally, rather than as a basis for formal tests of significance, since the model parameter estimators, and the associated t-tests, are usually not independent.

The information on individual regression coefficients can readily be adapted to obtain a confidence interval for the coefficients and for the regression function and prediction intervals for the book weight.

Indeed, the multiple R^2 and the adjusted multiple R^2 are, respectively, 0.928 and 0.917.

An F-statistic can be defined to provide an overall test of significance on the regression, and in particular on the *global* significance of the regression coefficients, that is on

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

In the example, both this test and each of the individual t-tests on β_1 and β_2 are strongly significant. However, a significant F-test does not necessarily imply that all the individual t-tests are significant too.

The global test is not concerned with the intercept β_0. Despite what stated in the course textbook, testing about the significance of the intercept makes little sense.

The intercept is something which is convenient to include in a model, since the simplest possible model (*the null model*) includes *at least* the intercept.

Furthermore, without the intercept, the multiple R^2 measures are not interpretable.

The ANOVA results may be derived also for multiple linear regression, although the interpretation requires great care.

In this example, it is possible to describe the contribution of `volume` after fitting the intercept and then the contribution of `area` after fitting both the intercept and `volume`.

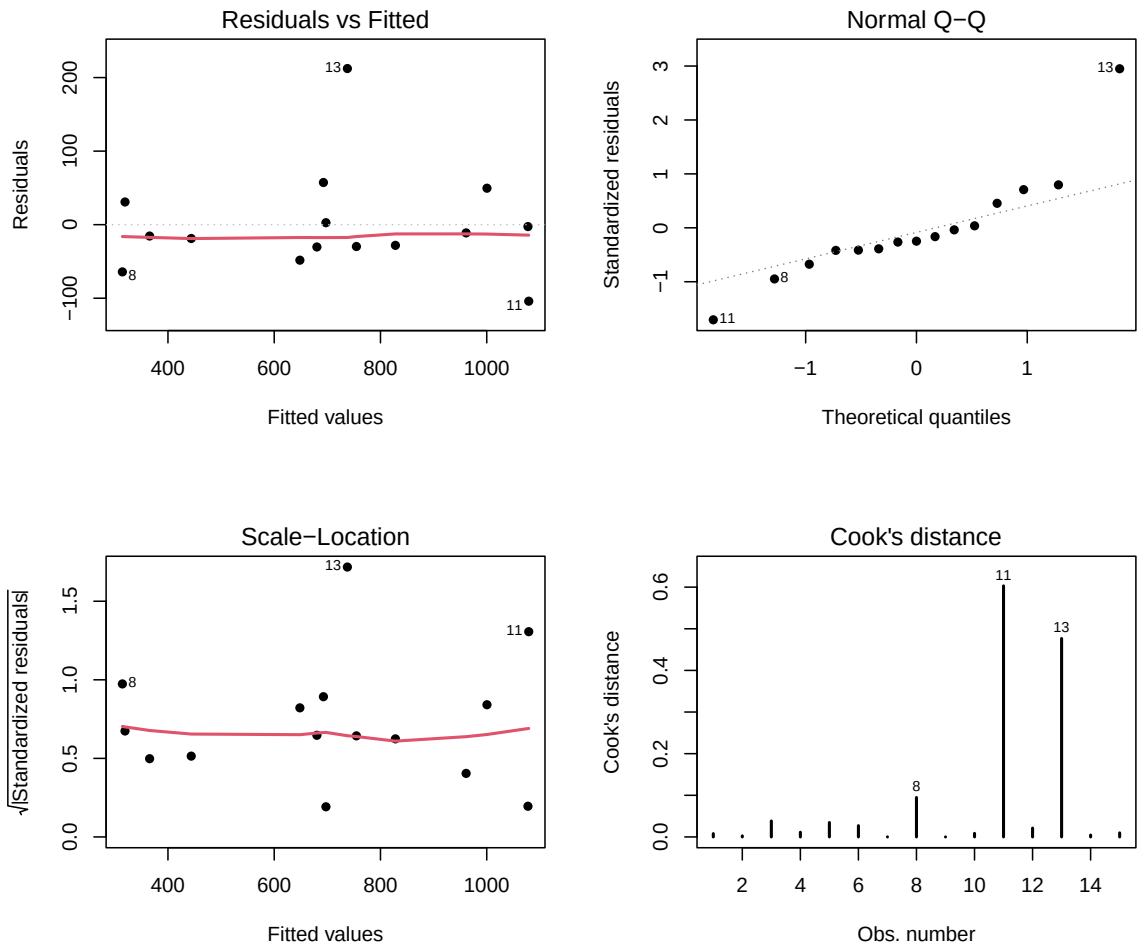
Thus the p-value for `area` agrees with that obtained using the t-test, since in both cases the p-values are computed in the model where `volume` is also included.

The p-value for `volume`, instead, differ from that obtained using the t-test, because is computed without considering `area`.

In this case, actually, the two p-values for `volume` are very close (both around $7 \cdot 10^{-8}$), but only because the correlation between `volume` and `area` is close to zero (0.0015).

This means that the two variables carry separate pieces of information.

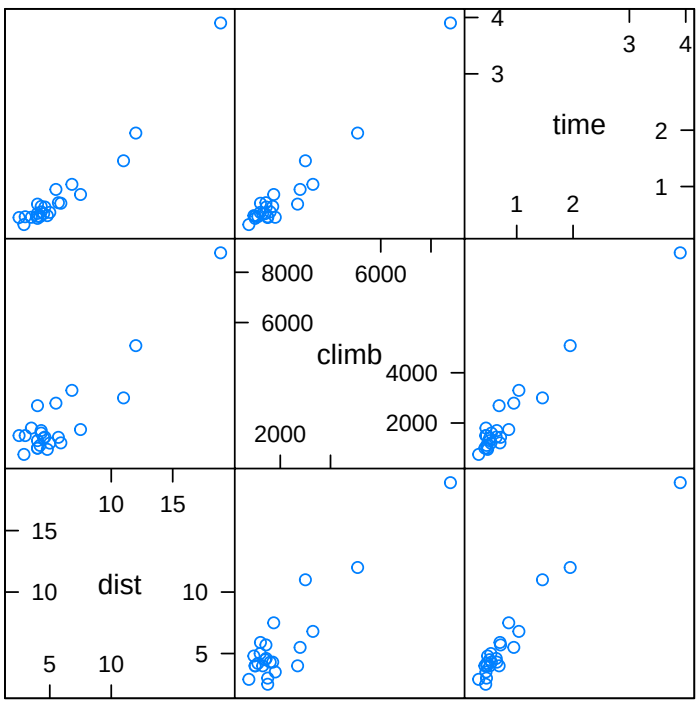
As for simple linear regression, diagnostic plots can be considered. They show that observations 11 and 13 correspond to the largest residuals, and are influential points. However, they seem legitimate observations, and with only 15 points it is better not remove them

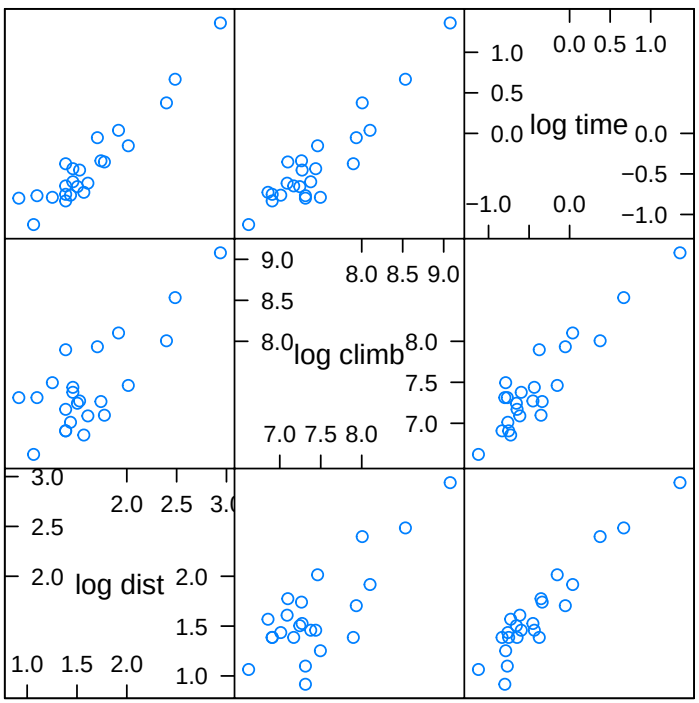


Example: hill races

Data on $n = 23$ hill races in Northern Ireland; the variables are: the distance `dist` (miles), the heights climbed `climb` (ft), male record time `time` (hours), female record time `timef` (hours).

Focusing on the male times only, the scatterplot matrix for original and log data reveal some linear relationships. Taking the logs seems preferable.





These considerations suggest fitting the model

$$\log(\texttt{time}_i) = \beta_0 + \beta_1 \log(\texttt{dist}_i) + \beta_2 \log(\texttt{climb}_i) + \varepsilon_i$$

This is equivalent to a model with a deterministic part described, in the original scale, by the power relationship

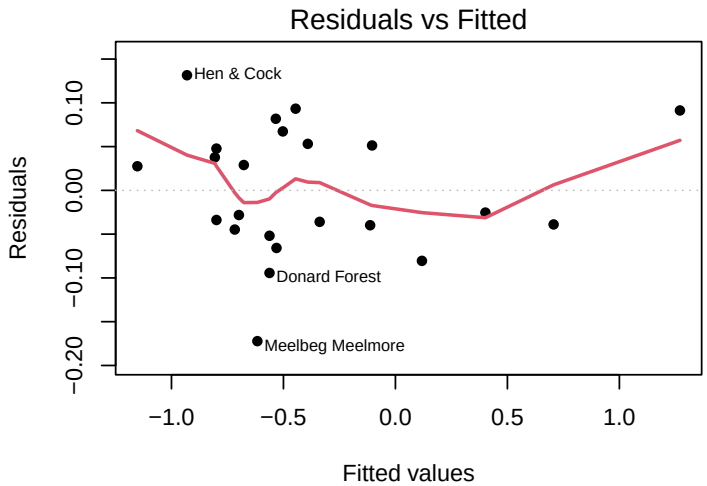
$$\texttt{time} = e^{\beta_0} \cdot \texttt{dist}^{\beta_1} \cdot \texttt{climb}^{\beta_2}$$

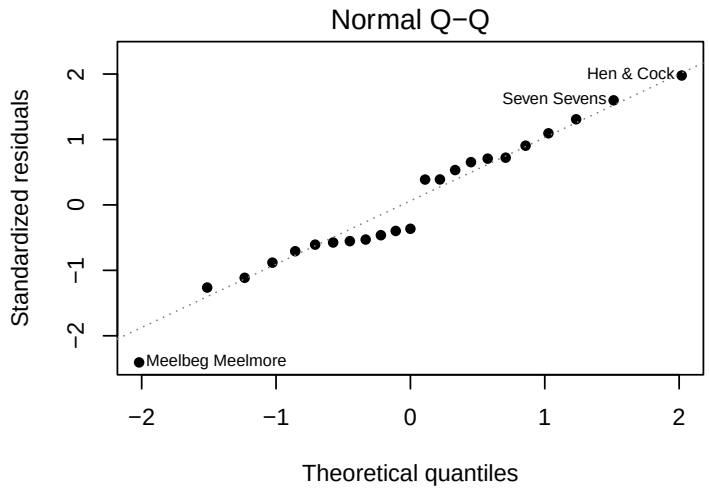
The coefficient estimates are $\widehat{\beta}_0 = -4.96$ ($\mathrm{SE}(\widehat{\beta}_0) = 0.273$), $\widehat{\beta}_1 = 0.68$ ($\mathrm{SE}(\widehat{\beta}_1) = 0.055$), $\widehat{\beta}_2 = 0.47$ ($\mathrm{SE}(\widehat{\beta}_2) = 0.045$), whereas the estimate for the noise standard deviation (the residual standard error) is $\widehat{\sigma} = 0.076$.

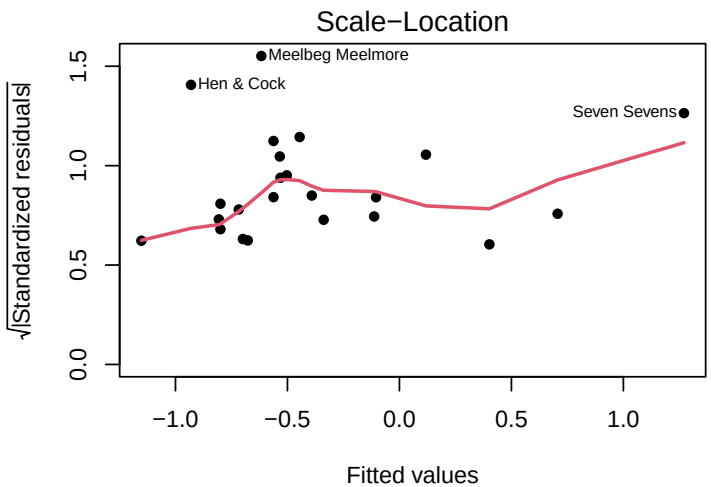
The low p-values for $\log(\texttt{dist})$ and $\log(\texttt{climb})$ highlights that they are both important predictors of $\log(\texttt{time})$. The *global* significance of the regression coefficients is confirmed by the F-test.

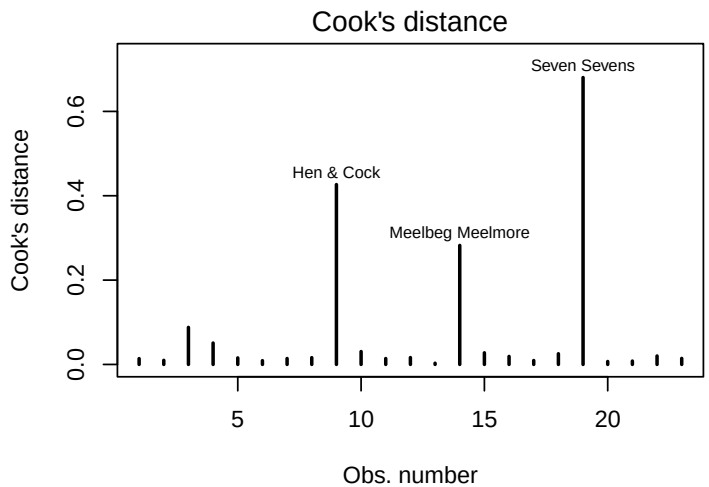
Indeed, the multiple R^2 and the adjusted multiple R^2 are, respectively, 0.983 and 0.981.

The diagnostic plots do not show any problem, apart from the moderately large residual associated with the Meelbeg Meelmore race









If the aim of the analysis is the interpretation of model coefficients, it is important to emphasizes that different formulations of the regression model, or different models, may serve different explanatory purposes.

To this regard, notice that the deterministic part of the fitted model is, on the original scale,

$$\texttt{time} = 0.007 \cdot \texttt{dist}^{0.68} \cdot \texttt{climb}^{0.47}$$

Surprisingly, the relative rate of increase in time is 68% of the relative rate of increase in distance, holding `climb` constant.

This implies that for a fixed value of `climb` the time is smaller for the second mile than for the first mile: indeed, setting `climb`=1500, the times are, respectively,

$$0.007 \cdot 1^{0.68} \cdot 1500^{0.47} = 0.218 \qquad 0.007 \cdot 2^{0.68} \cdot 1500^{0.47} = 0.349$$

This seems quite strange, but the explanation comes from the meaning of *keeping climb constant*, since short races will be steeper than long races.

The coefficient for $\log(\texttt{dist})$ is, reassuringly, greater than 1 if $\log(\texttt{time})$ is regressed on $\log(\texttt{dist})$ and $\log(\texttt{climb}/\texttt{dist})$, instead of $\log(\texttt{climb})$. Then,

$$\texttt{time} = 0.007 \cdot \texttt{dist}^{1.15} \cdot (\texttt{climb}/\texttt{dist})^{0.47}$$

Note that the two models provide the same fit, since they are different mathematical formulations of the same underlying model. Interpretability issues and application-specific considerations will drive the choice of a particular model form.

There is another, related, benefit in the second model. The correlation between $\log(\texttt{dist})$ and $\log(\texttt{climb}/\texttt{dist})$ is -0.065, negligible relative to the correlation of 0.78 between $\log(\texttt{dist})$ and $\log(\texttt{climb})$.

In designed experiments, uncorrelated regressors are usually considered. Even in observational studies, when possible, it is preferable to include terms in the model which have small cross-correlation.

In some sense, if two covariates x_1 and x_2 are correlated, the effect of x_1 on the response is not expressed only by $\widehat{\beta}_1$, but also by $\widehat{\beta}_2$ through the relation between x_1 and x_2.

Centering the covariates

- Centering the covariates amounts to subtraction of their sample mean before introducing them in the model. For example, the model

$$y_i = \beta_0 + \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \varepsilon_i$$

  can be written as

$$y_i = \alpha + \beta_1 \, (x_{i1} - \overline{x}_1) + \beta_2 \, (x_{i2} - \overline{x}_2) + \varepsilon_i$$

- This helps model interpretability in two ways:
  - the estimated intercept is $\widehat{\alpha} = \bar{y}$, and it is the fitted value obtained when both the covariates are equal to their sample mean;
  - the observed values can be written as

$$y_i = \bar{y} + \widehat{\beta}_1 \, (x_{i1} - \bar{x}_1) + \widehat{\beta}_2 \, (x_{i2} - \bar{x}_2) + \widehat{\varepsilon}_i = \bar{y} + t_{i1} + t_{i2} + \widehat{\varepsilon}_i$$

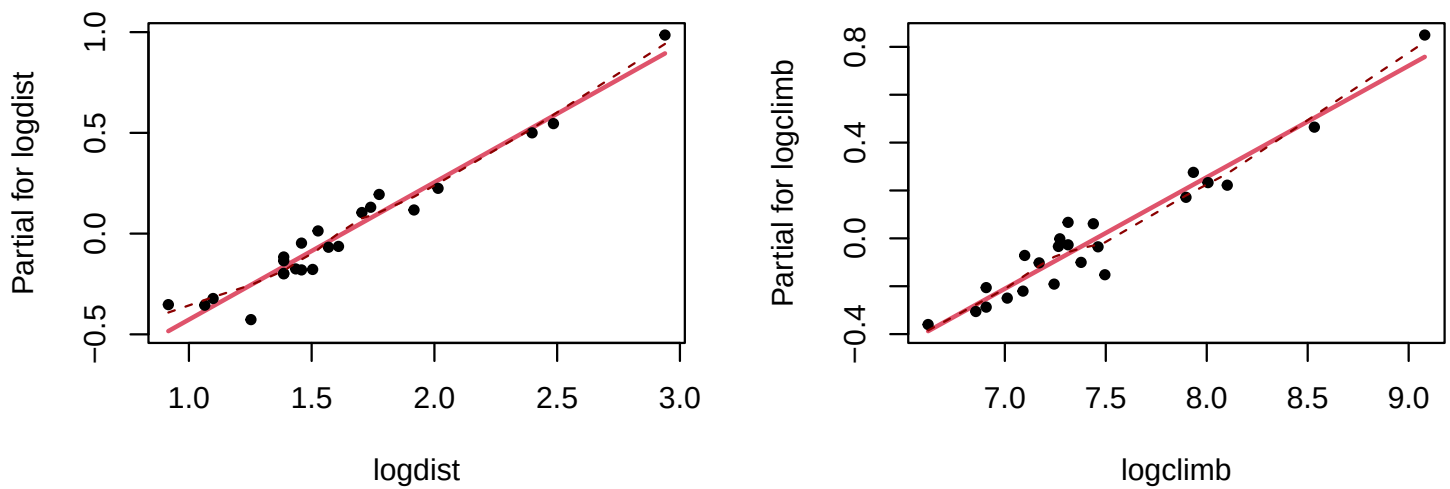
    where t_{i1} and t_{i2} are zero-sum contributions from each covariate.
- The terms t_{i1} and t_{i2} can be used for the **partial residual plots**.

Partial residual plot

The partial residual plot for the covariate x_j, given all the others, it is a scatterplot of $t_{ij} + \widehat{\varepsilon}_i$ vs x_{ij}. It accounts for that part of the response that is not explained by the covariates other than x_j.

For example, with two covariates, the partial residual plot for x_1 graphs $t_{i1} + \widehat{\varepsilon}_i = y_i - \bar{y} - t_{i2}$ against x_{i1}. It assesses whether the part of the response that is not explained by x_2 can be approximated by a linear function of x_1.

For the hill races data, using the model with log variables, the two partial residual plots, given below, are quite linear



Quadratic effect of a covariate

- In some cases, a certain covariate may have an evident nonlinear effect on the mean response; for instance, this can be highlighted by drawing a partial residual plot.

  In such case, it is convenient to include in the model more than a term to describe the effect of such covariate.

- In the simple case of a single covariate, a model with a **quadratic effect** of x will specify the mean response as

$$E(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

  Such model is not any longer in the class of simple linear regression models, being instead a special case of multiple linear regression, with two covariates, namely $x_{i1} = x_i$ and $x_{i2} = x_i^2$.

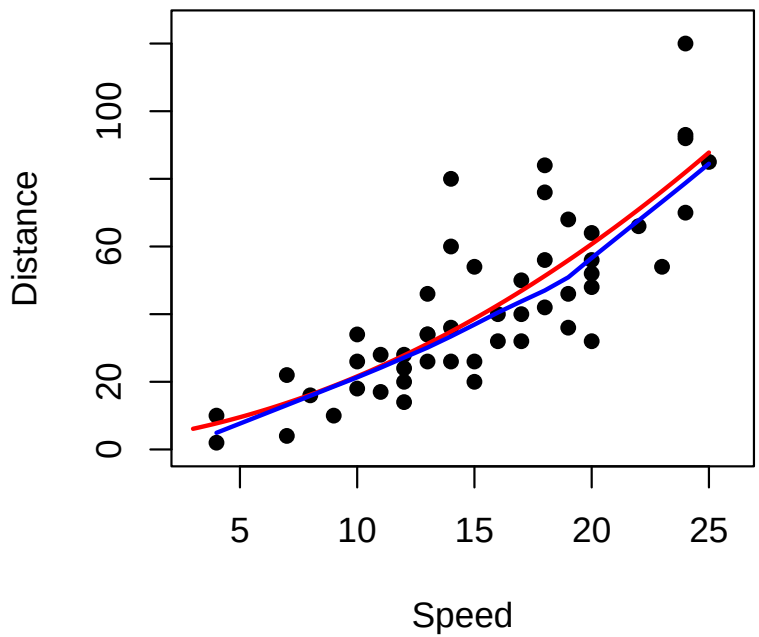
- **Polynomial regression models** generalize, in some sense, the multivariate linear model and they may contain squared, cross-terms and higher-order terms of the original predictor variables.

Example: cars

For the cars data, a preliminary statistical analysis, supported by physical considerations, suggests that the `distance` taken to stop should be a non-linear function of the `speed`. Then, a plausible model is

$$\texttt{distance}_i = \beta_0 + \beta_1 \texttt{speed}_i + \beta_2 \texttt{speed}_i^2 + \varepsilon_i$$

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



The estimated model parameters are $\widehat{\beta}_0 = 2.47$ ($\mathrm{SE}(\widehat{\beta}_0) = 14.82$), $\widehat{\beta}_1 = 0.91$ ($\mathrm{SE}(\widehat{\beta}_1) = 2.03$) and $\widehat{\beta}_2 = 0.10$ ($\mathrm{SE}(\widehat{\beta}_2) = 0.07$).

The p-values for all the model coefficients are very high, despite the fact that the p-value for the F test is $5.85 \cdot 10^{-12}$, indicating that there is a strong evidence that the assumed model is better than the constant one.

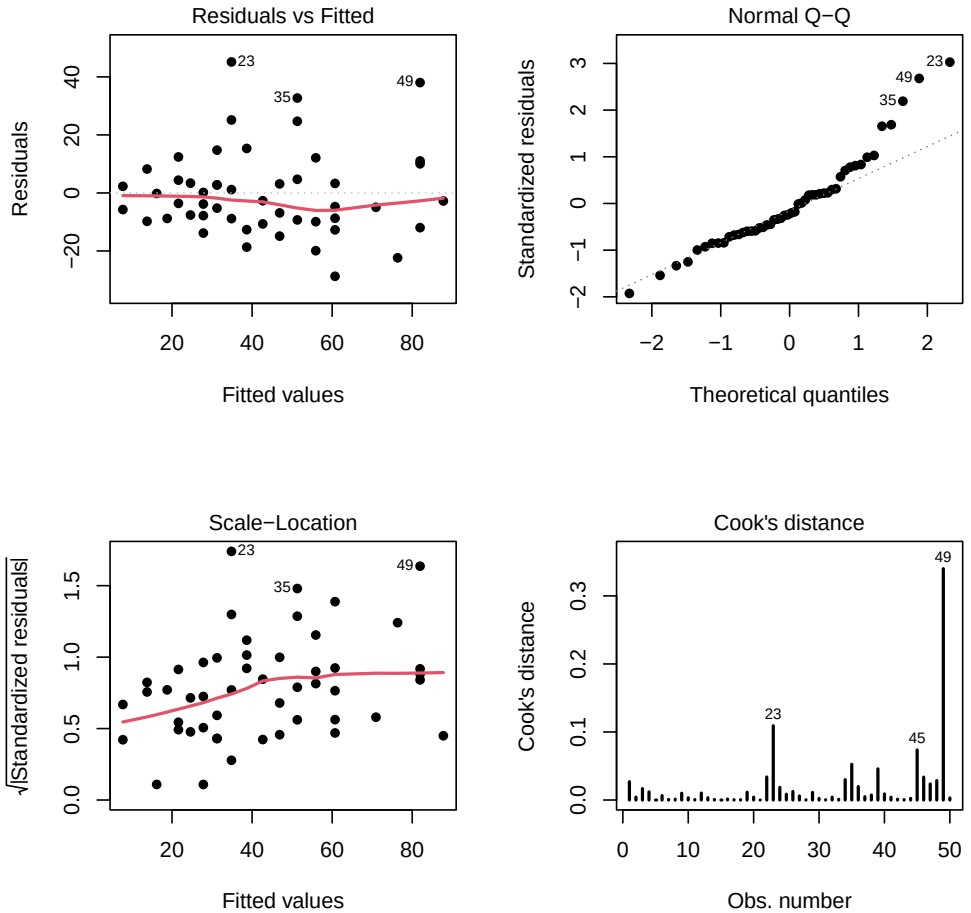
The p-values are testing whether the corresponding coefficients could really be zero given that the other terms remain in the model: they cannot be taken as an indication that all the terms can be dropped.

The point here is that the lack of independence between estimators creates difficulties in the interpretation of estimates.

The correlation between parameter estimators typically arises from correlation between the corresponding covariates.

In this case it is not possible to entirely separate out their effects on the response by examining the results of model fitting.

The diagnostic plots show some indication of non-constant variance (top left) and of a departure from normality in the residuals (top right)



An alternative model is one in which variability increases with speed; for example, assuming $\varepsilon_i \sim N(0, \sigma^2 \cdot \text{speed}_i)$. In this case, the parameter estimates are obtained by the **weighted least squares (WLS) method**; that is, minimizing $\sum \varepsilon_i^2 w_i$, with $w_i = 1/\text{speed}_i$.

Table of contents

Violation of model assumptions

- Each of the assumptions underlying multiple linear regression may be not plausible for a certain application. This may involve both the systematic part of the model and the random term.
- The assumption that the mean of Y_i is a linear combination of the xs is likely to be just an approximation.

  Important deviations may be the **nonlinear effect** of an explanatory variable, and the **lack of a certain one in the data set**. The latter is more serious, and requires careful consideration of the problem under investigation.
- The assumption of independence may be not plausible with **clustering** of the observations, or **repeated measurement** over time or space.
- The assumptions of constant variance and normality are often violated in practice. A careful choice of the scale for the response variable may make them more plausible.
- The presence of **outliers** is something to look into.

Checking on the residuals

- Any of the above problems may be fixed in practice, at least to a certain extent, provided they are readily detected.
- Although sometimes it is not easy or possible to detect failure of the assumptions, there are simple checks that, if the assumption fails, may indicate the nature of the failure.
- **Regression diagnostics** are a set of methods which assist such detection.
- Residual plots for multiple linear regression are interpreted in the same way as residual plots for simple linear regression. Similarly, it possible to check on **nonlinearity**, **constant variance** and (approximate) **normality**.
- To check nonlinearity in the multiple setting, it can be useful to plot residuals against predictors, rather than relying solely on the default plot concerning fitted values *vs* estimated residuals.

Outliers: leverage and influence

- Looking for outliers is more complicated when there are several explanatory variables. Two (or more) outliers, that are influential, may mask each other. The outliers are treated as in the simple linear regression model.
- To this extent, the concepts of leverage and influence are important.
- If the i-th observation y_i is changed into $y_i + \Delta_i$ (leaving all the other y-values unchanged), then the fitted value changes from $\widehat{y}_i$ to $\widehat{y}_i + h_{ii}\,\Delta_i$, and h_{ii} is called the **leverage** for the i-th point.
- The h_{ii} values are the diagonal element of the **hat** (**influence**) **matrix** $\mathbf{H} = \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, such that

$$\widehat{\mathbf{y}} = \mathbf{X}\,\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$$

  Large leverage values flag points away from the other points.
- In a model with p coefficients $\sum h_{ii} = p$, so that values h_{ii} larger than $2\,p/n$ or $3\,p/n$ can be considered as large.

- Data points that may alter the fitted values (if omitted) are **influential**. Such distortion, regarding the fitted response, is a combined effect of the size of the residual and its leverage.
- A common measure of influence is given by the **Cook's distance**. Given the **standardized residuals**

$$r_i = \frac{y_i - \widehat{y}_i}{s\sqrt{1 - h_{ii}}}$$

if the model has p coefficients, the Cook's distance for the i-th observation is

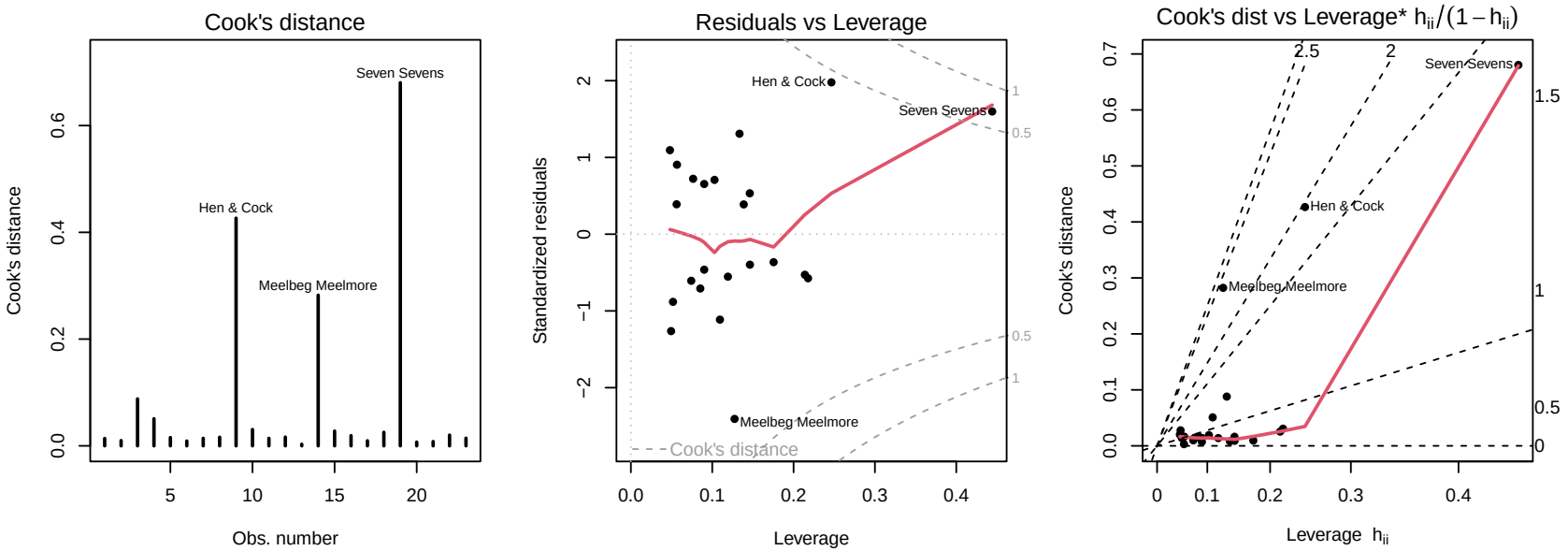
$$D_i = \frac{1}{p} r_i^2 \frac{h_{ii}}{1 - h_{ii}}$$

It measures the change in model estimates when the i-th observation is omitted.

- Values of D_i larger than 0.5 (or even 1.0) are suspicious, and may refer to points with a strong effect on the estimated coefficients and the estimated standard errors.
- The (standardized) effect of each observation on the estimates $\widehat{\boldsymbol{\beta}}$ can also be evaluated.

Example: hill races

For the log data on hill races in Northern Ireland, the following diagnostic plots describe Cook's distances and leverages. They do not indicate serious problems, apart a point with a Cook's distance larger than 0.5.



Evaluation of the (standardized) effect of each observation on the estimates shows that none of the three observations with the largest Cook's distance has a worrisome effect.

Table of contents

R^2 and adjusted R^2

- Goodness of fit for linear models is usually measured in terms of the proportion of response variability explained by the model, as quantified by the **coefficient of determination** R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  The R^2 tends to overestimate the goodness of fit. The adjusted R^2 is usually preferable, since it accounts for the degrees of freedom

$$\text{adjusted } R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2/(n-p-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}$$

- Both these measures are not suitable for comparisons between different studies, where the range of values of the explanatory variables may be different. They can be used instead for comparing **different models for the same data**.
- Values close to 1 indicate a good fit, but low values do not necessarily indicate a poor model; the data contain a substantial random component.

Model selection

- The ANOVA results, based on suitable F tests, can be considered for comparing **nested linear models**.
- An alternative procedure consists in a sequence of F tests comparing the full model with each of the models produced by dropping a single predictor.

  Starting from the full model, the model term with the highest p-value is repeatedly deleted (and the new full model refitted) until all p-values are below some threshold (*backward selection*).

  Starting from a simple model, the model term which has most evidence in an F test is repeatedly added until no more terms would lead to significant improvement (*forward selection*).
- There are also *backward-forward strategies*, in which cycles of backward and forward selection are alternated until convergence.

- Better alternatives, also useful for **non-nested models**, do exist and they aim at evaluating the predictive ability of alternative models.
- The information criteria are particularly well suited for choosing the model with the best capability of doing prediction for unobserved data. Criteria based on the cross-validation procedures can be also considered.
- The **AIC statistic** is perhaps the most commonly used, and for a linear regression model with p coefficients, is given by

$$\text{AIC} = n \, \log \left(\frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^2}{n} \right) + 2\,p + \text{const}$$

  where the constant term arises from the assumption of an i.i.d. normal distribution for the errors.
- The **BIC statistic**, which replaces $2p$ by $\log(n) \cdot p$, penalizes models with many parameters more strongly.
- Both the statistics are smaller for models with better predictive power.

- The **Mallow's C_p statistic** (in this framework, equivalent to the AIC) differs from the AIC statistic only by subtraction of n, and by omission of the constant term

$$C_p = n \log\left(\frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^2}{n}\right) + 2\,p - n$$

- At times, with a moderate number of predictors, these selection statistics can be used for semi-automatic model selection procedures, such as **stepwise model selection** (backward and/or forward).
- Another approach for variable selection, useful when there are large numbers of predictors relative to the number of data, involves the **lasso method**.

  It penalizes the model coefficients towards zero, in such a way that as the penalization increases, many of the coefficient estimates become zero.
- Alternatively, the variable selection problem can be solved using suitable **boosting algorithms**.

Example: cars

For the cars data, the following models are considered

$$\begin{aligned} \mathrm{M}_0 &: \texttt{distance}_i = \beta_0 + \beta_1 \texttt{speed}_i + \varepsilon_i \\ \mathrm{M}_1 &: \texttt{distance}_i = \beta_2 \texttt{speed}_i^2 + \varepsilon_i \\ \mathrm{M}_2 &: \texttt{distance}_i = \beta_0 + \beta_1 \texttt{speed}_i + \beta_2 \texttt{speed}_i^2 + \varepsilon_i \end{aligned}$$

In M_1 and M_2 the WLS method, with $w_i = 1/\texttt{speed}_i$, is considered.

The ANOVA comparison of the nested linear models M_0 and M_2 gives a p-value $2.2 \cdot 10^{-16}$ for the F test, indicating strong evidence against M_0.

The comparison between M_1 and M_2 gives a p-value 0.046 giving some evidence against M_1.

The AIC statistic suggests again the larger model, since

$$\mathrm{AIC}(\mathrm{M}_0) = 419.16, \ \mathrm{AIC}(\mathrm{M}_1) = 414.80, \ \mathrm{AIC}(\mathrm{M}_2) = 412.26$$

Conversely, the BIC statistic points to the model M_1, penalizing the larger model M_2. Model M_1 has also the larger values for R^2 and adjusted R^2.

Example: hill races

For the log data on hill races in Northern Ireland, the following models are considered

$$\begin{aligned} \mathrm{M}_1 : \log(\texttt{time}_i) &= \beta_0 + \beta_1 \log(\texttt{dist}_i) + \beta_2 \log(\texttt{climb}_i) + \varepsilon_i \\ \mathrm{M}_2 : \log(\texttt{time}_i) &= \beta_0 + \beta_1 \log(\texttt{dist}_i) + \beta_2 \log(\texttt{climb}_i) \\ &\quad + \beta_3 \left(\log(\texttt{dist}_i)\right)^2 + \varepsilon_i \end{aligned}$$

In the second one, a quadratic term for `log(dist)` is included.

The adjusted R^2 is about the same for the two models, but both the AIC and the BIC values are smaller for M_2:

$$\mathrm{AIC}(\mathrm{M}_1) = -48.13, \ \mathrm{AIC}(\mathrm{M}_2) = -49.83$$

$$\mathrm{BIC}(\mathrm{M}_1) = -43.58, \ \mathrm{BIC}(\mathrm{M}_2) = -44.15$$

Including the quadratic term seems a good idea even if the p-value for the quadratic coefficient is 0.084.

Some suggested steps for model fitting

- Examine the distribution of each of the explanatory variables, and of the dependent variable. Look for highly skew distributions, and outlying values.
- Examine the scatterplot matrix of all the explanatory variables, and also of the response variable.
- Note the ranges of each of the scatterplot variables, considering if they vary sufficiently to affect the response variable and if each of the explanatory variables is accurately measured.
- In case some pairwise plots hint to nonlinearity, consider the use of transformations to achieve more nearly linear relationships.
- Transformation of the response is advisable in case of skew distribution. The Box-Cox transformation, already introduced for simple linear regression, can be useful for suggesting the right scale.
- Pairs of explanatory variables that are so highly correlated that they appear to provide the same information should be analyzed. Background information may suggest which one should be retained.

Some diagnostic checks

- Plot residuals against fitted values. Check for patterns in the residuals, and changes in the variability.
- Quantile-quantile plots of residuals are also useful, but they should not be taken too seriously.
- For each explanatory variable, draw a partial residual plot, to check whether any of the explanatory variables require transformation.
- Examine the Cook's distance statistics. In case of doubt, it may be useful to examine the standardized effect of each observation on the model parameter estimates.
- In principle, outliers, influential or not, should never be disregarded. Their exclusion may be a result of use of the wrong model.
- If apparently genuine outliers remain excluded from the final fitted model, they must be noted in the eventual report and included, separately identified, in graphs.

Table of contents

Selecting the explanatory variables

- When there are several explanatory variables, selecting the variables that give the best prediction becomes an issue.
- Start from an *informed guess* about which variables are likely to be important. Some variables ought to be included in the model even when their contribution to the prediction of the response is limited.
- As a practical rule, one suggested rule is that there should be at least ten times as many observations as variables, before variable selection takes place.
- Any analysis should consider an explorative investigation of the available explanatory variables, leading at times to consider suitable transformations of some or all variables.
- Graphical scrutiny of the explanatory variables may lead to the omission of some variables.
- Beware of *spurious relationships*: two or more variables seem to be related, due to either coincidence or the presence of a third, unseen variable (`www.tylervigen.com/spurious-correlations`).

- Interaction effects between numerical variables (not factors) are often modeled by including pairwise products, namely $x_1 \cdot x_2$ as well as x_1 and x_2.
- Multivariate techniques, such as **Principal Components Analysis**, are sometimes useful for selecting low-dimensional combinations of several explanatory variables.

  These small number of combinations together account for most of the variation in the explanatory variables, thus reducing the dimension of the problem.
- Other methods, such as the **lasso**, the **least angle regression** or the **boosting**, allow for semi-automatic variable selection.
- Caution should be used with automatic selection techniques, such as stepwise regression and subset regression.

Multicollinearity

- Some explanatory variables are linearly related to another variable, or to combinations of other explanatory variables. This is known as **multicollinearity**, and it is very common with observational data.
- Multicollinearity implies that there are redundant variables. It alters the relation of the explanatory variables with the response. In extreme cases, it may lead to poorly estimated coefficients.
- A measure for quantifying the severity of multicollinearity is the **variance inflation factor** (**VIF**). For an explanatory variable x_j,

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

  where R_j^2 is the R^2 coefficient for the regression model having x_j as the response variable, and all the other explanatory variables as regressors.
- Thresholds usually adopted for VIF are 4 or 5, that suggest some multicollinearity. With a VIF > 10 multicollinearity is severe, and the model coefficients will be poorly estimated.

Remedies for multicollinearity

- Careful initial choice of variables, based on background information and careful scrutiny of exploratory plots, often will prevent the problem.
- Dropping one or more explanatory variables is the main route to address multicollinearity.
- An alternative route is to obtain a combination of the original explanatory variables that summarizes them without losing too much information: this is performed by the aforementioned PCA method.
- There are also some modern methods, such as **ridge regression** or the **lasso**, that are not affected by multicollinearity.

Example: coxite

The data set coxite gives the mineral compositions of $n = 25$ rock specimens of coxite type.

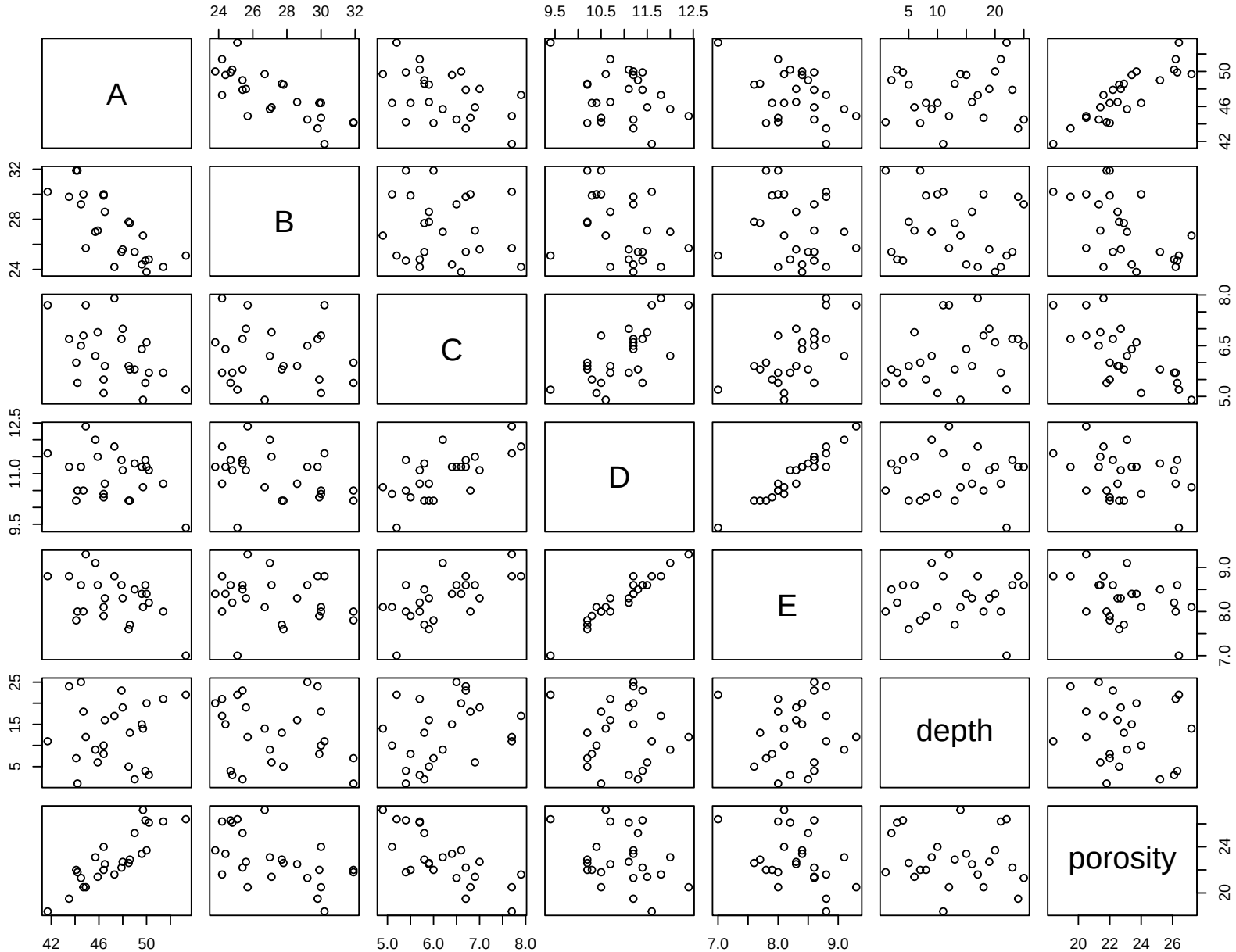
Each composition consists of the percentage by weight of five minerals, namely, albite `A`, blandite `B`, cornite `C`, daubite `D`, endite `E`.

Indeed, the recorded `depth` (m) of location of each specimen and the `porosity` (the percentage of void space that the specimen contains) are also provided.

Note that the percentages of the five minerals sum to 100.

The aim of regression analysis is to explain the response variable `porosity` as a function of mineral composition and `depth`.

The scatterplot matrix shows that `D` and `E` are strongly linearly related.


A
B
C
D
E
depth
porosity

Fitting the model with all six explanatory variables gives a coefficient for `E` equal to zero, since the five percentage sum to 100 and then `E` adds no additional information.

None of the individual coefficients is significantly different from zero (all the p-values are greater than 0.3).

However, the R^2 measures are high (0.935 and 0.919, for the adjusted version), and the F test of global significance of all terms is significant (the p-value is $1.18 \cdot 10^{-10}$).

These are clear symptoms of multicollinearity: indeed, omitting `E` (or one of `A`, `B`, `C`, `D`), the VIF values can be calculated and they are very large

A	B	C	D	depth
2717.8	2485	192.59	566.14	3.4166

Thus, it is unsurprising that none of the individual coefficients can be estimated meaningfully.

It is reasonable to try a model that uses those variables that, individually, correlate most strongly with `porosity`. Here are the correlations

A	B	C	D	E	depth
0.869	-0.551	-0.723	-0.320	-0.408	-0.147

The model with `A`, `B` and `C` as explanatory variables improves the previous one, but the coefficient for `A` is not significantly different from zero and yet the VIFs are all larger than 4

A	B	C
10.9360	8.5924	4.5551

The model with only `B` and `C` as explanatory variables is much better and it passes all the diagnostic checks. Indeed, the VIFs are both around 1.

Furthermore, the AIC for the full model is 56.50, whereas it is 52.47 for the final model.

Table of contents

- The model matrix $\mathbf{X}$ is fundamental to all calculations for a linear model. It carries the information needed to calculate the fitted values that correspond to any particular choice of coefficients.
- The columns of $\mathbf{X}$ contain the observed values of the (numeric) explanatory variables, perhaps after transformation, whereas the first column usually corresponds to the model intercept.
- However, explanatory variables are not always numeric, and actually factors are very common in many applied fields.
- Factors can be included in a model in a straightforward way, and it is possible to fit their effect on the response along with the effect of numerical variables.
- ANOVA models are just a special case of linear regression models where all the explanatory variables are factors.

Two-way ANOVA

- One-way ANOVA models have been previously introduced in the context of simple linear regression models, to deal with the situation where there is only one factor as explanatory variable.
- However, there might be more than one factor influencing the response variable, like in designed experiments.
- **Two-way ANOVA** is suitable with two factors, while **multi-way ANOVA** accounts for an arbitrary number of factors.
- With more than one factor ANOVA becomes more complex, as there may be **interaction** between the factors.
- Two-way ANOVA not only aims at assessing the main effect of each (categorical) variable on the response but also the potential effect due to the interaction between them.
- The methodology proceeds exactly as in the one-way case, with suitable extensions for considering the presence of two factors.

Example: warp breaks

The data set gives, as response variable, the `number of warp breaks` per a fixed length of yarn during weaving.

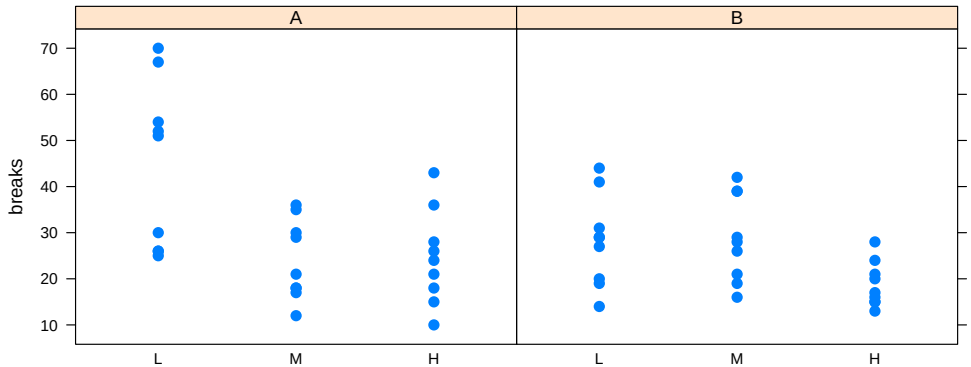
There are two experimental factors: the type of `wool` and the level of `tension`, with 2 (`A` or `B`) and 3 levels (`L`, `M`, `H`) respectively: there are 9 replications for each of the 6 combinations of factors levels.

The total sample size is then $n = 2 \times 3 \times 9 = 54$.

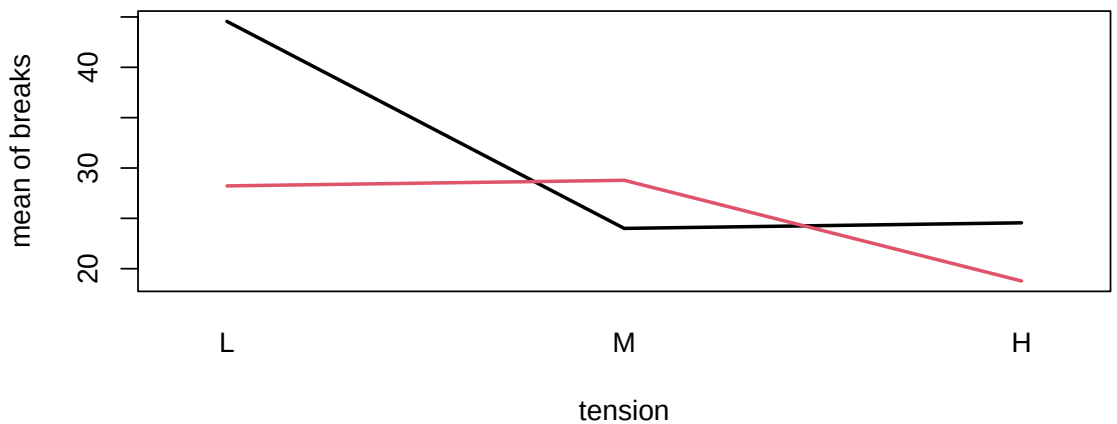
Also in this case, the data are balanced; unbalanced data are better dealt with by techniques based on linear regression models.

		replicate								
wool	tension	1	2	3	4	5	6	7	8	9
A	L	26	30	54	25	70	52	51	26	67
	M	18	21	29	17	12	18	35	30	36
	H	36	21	24	18	10	43	28	15	26
B	L	27	14	29	19	29	31	41	20	44
	M	42	26	19	16	39	28	21	39	29
	H	20	21	24	17	13	15	15	16	28

A **conditional plot** shows that the pattern of the response is not the same for the two levels of `wool`



A similar result can be obtained with an **interaction plot**, where the means (or another summary) for each combination of two factors are displayed: **level A** and level B of the factor `wool`



Statistical model for two-way ANOVA

- **Without interaction** between the two factors: each factor has the same effect on the mean response, regardless of the level of the other.
- The model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

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

  - i identifies the treatment level for the first factor;
  - j identifies the treatment level for the second factor;
  - k identifies the observation;
  - μ is the **general mean**;
  - α_i is the **main effect for the first factor**: the deviation from the general mean μ when the first factor is equal to the i-th category;
  - β_j is the **main effect for the second factor**: the deviation from μ when the second factor is equal to the j-th category;
  - ε_{ijk} i.i.d. $N(0, \sigma^2)$ distributed random errors.

- To test the main effect of the first factor

$$H_0: \quad \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$
$$H_1: \quad \alpha_i \neq 0 \text{ for at least one } i$$

  and similarly for the other factor.
- The test is performed exactly like for one-way ANOVA, by considering suitable sums of squares and d.f., that define an F test statistic.
- The results of the analysis can be trusted as long as the model assumptions are reasonably satisfied.
- Post-hoc analysis can also be performed, though they require more care than in the one-way case.

- When possible **interaction** is investigated, another set of coefficients γ_{ij} is introduced in the model, that becomes

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

  where γ_{ij} describes the **interaction effect for the two factors**: the deviation from the general mean μ due to the interaction of the i-th category of the first factor and the j-th category of the second factor.

- The test on the presence of interaction can be carried out similarly to the tests on main effects.
- *A hierarchical order should be followed in performing the various tests*:
  - first, the presence of interaction (if plausible) should be assessed;
  - in case the data support the presence of an interaction effect, *it makes little sense to investigate about the presence of main effects*, as both the both the two factors influence the mean response;
  - in case the interaction effect is not significant, the test on the two main effects can be performed.

Example: warp breaks

For the data set on `warp breaks`, the transformed response `sqrt(breaks)` is considered instead of the original observations, as the conditional plot shows some evidence of non-constant variance.

Indeed, ANOVA assumes a constant variance for the observations, so in case this is doubtful a **variance-stabilizing transformation**, such as the squared root or the log, should be investigated.

The p-value for the F test on the interaction effect is $p = 0.031$, showing an appreciable interaction effect, though not very large.

Moreover, it is consequently stated that the two main effects are present as-well, since both the two factors influence the mean of the response variable `warp breaks`.

Regression models with dummy variables

- More generally, in the context of multiple linear regression model, it is possible deal with the presence of factor explanatory variables.
- To include a factor in a model it is necessary to code its levels, and one possibility is to use **dummy variables** (regressors that assume only two values, usually, zero and one).
- The rule for coding a factor is quite simple. Coding a factor with h levels (categories) requires the usage of $(h-1)$ dummy variables.
- Consider, for example, the study of the relationship between `personal income` Y and `education level` X (a factor with $h = 3$ levels: `middle school`, `high school` and `university`).

  Two dummy variables x_1 and x_2 are needed, so that

factor level	x_1	x_2
`middle school`	0	0
`high school`	1	0
`university`	0	1

- The dummy variables used to represent the effect of different factor levels can then be included in the regression model.
- In the example,

$$y_i = \beta_0 + \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \varepsilon_i$$

  which corresponds to specify the following expressions for the mean of Y:
  - β_0 for `middle-school` level
  - $\beta_0 + \beta_1$ for `high-school` level
  - $\beta_0 + \beta_2$ for `university` level
- The coefficients used to represent the effect of a given factor represent the **main effect** of that factor.
- Such coefficients express the *differential effect* on the mean response that can be attributed to the different levels of that factor.
- One of the factor levels is set up as a baseline or reference, with the effects of other levels then measured from the baseline.

Models with both factors and numerical explanatory variables

- Often in applications there are both categorical and numerical explanatory variables.
- By using a suitable factor coding for categorical variables, it is possible to include both categorical and numerical explanatory variables in a model specification.
- Regression models with both factors and numerical regressors are also known as **analysis of covariance models**, and they essentially amount to fitting multiple lines.
- All the relevant inferential techniques are those already introduced for multiple regression models. Different models are compared by F tests, summarized by ANOVA-type results.

Fitting multiple lines

- In the example concerning `income` and `education level`, the numerical explanatory variable x_{i3}, summarizing the `number of years spent at work`, is also considered.
- The model including such variable and the `education level` is

$$y_i = \beta_0 + \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \beta_3 \, x_{i3} + \varepsilon_i$$

- This is apparently similar to the model with more than one numerical predictor, but there are some important differences.
- In this case, model fitting actually concerns **three parallel simple regression lines**, with intercepts changing with the level of education.
- Including also an interaction term between `education level` and `years spent at work` requires the fit of **three different simple regression lines** for the three groups of units with different level of education.

Example: regression on age and gender

A regression model for a response variable Y related to the `health status`, such as cholesterol or blood pressure, is defined.

Potential predictors are the numerical variable age x and a binary factor variable `gender`, coded by a dummy variable g (assuming 1 for female).

A model with the main effects and the interaction effect is

$$y_i = \beta_0 + \beta_1 \, x_i + \beta_2 \, g_i + \beta_3 \, x_i \, g_i + \varepsilon_i$$

This model assumes that Y depends linearly on age for males, with the regression line

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

and Y depends linearly on age for females, with the regression line

$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \, x$$

Then **two different simple regression lines** for the groups of males and females are defined.

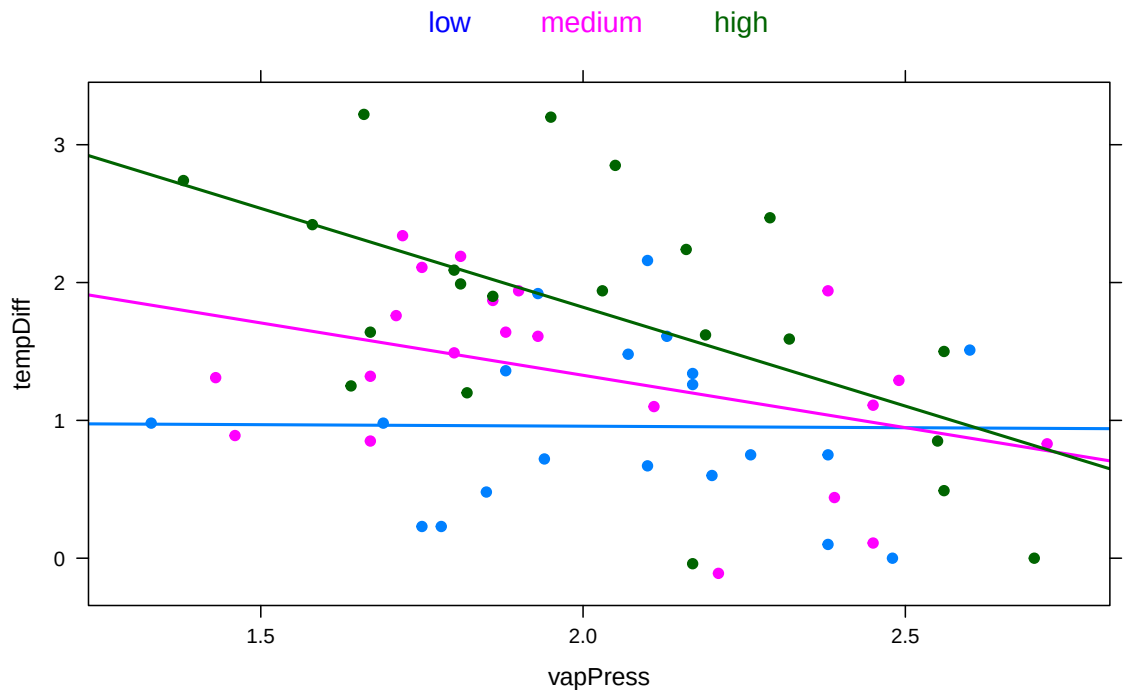
The following hypotheses may be of potential interest, and can be analyzed by means of suitable t tests or F tests.

Null/Alternative hypothesis	Translation
Relation between age and y does not depend on gender (Relation between gender and y does not depend on age)	$H_0 : \beta_3 = 0$
Gender influences the relation between y and age; (Age influences the relation between y and gender)	$H_1 : \beta_3 \neq 0$
y does not depend on age	$H_0 : \beta_1 = \beta_3 = 0$
y depends on age	$H_1 : \beta_1 \neq 0$ or $\beta_3 \neq 0$
y does not depend on gender	$H_0 : \beta_2 = \beta_3 = 0$
y depends on gender	$H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$
y does not depend on age and gender	$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
y depends on age or gender	$H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$

Example: leaf and air temperatures

The data set includes $n = 62$ environmental measures regarding the vapor pressure (`vapPress`) and the difference between leaf and air temperature (`tempDiff`), for three different levels (`low`, `medium`, `high`) of carbon dioxide (`CO2level`).

The scatterplot of `tempDiff` vs `vapPress` suggests that there may be three different regression lines, for the three different levels of `CO2level`, but this has to be confirmed more formally.



Denoting by x the numerical explanatory variable `vapPress`, and by z_1, z_2 the two dummies for the factor `CO2level`, four different models for the response y (`tempDiff`) may be defined:

M1 (constant response): $y = \beta_0 + \varepsilon$

M2 (single line): $y = \beta_0 + \beta_1 x + \varepsilon$

M3 (3 parallel lines): $y = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \varepsilon$

M4 (3 separate lines): $y = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x z_1 + \beta_5 x z_2 + \varepsilon$

Models M3 and M4 involve the factor `CO2level`, and the level `low` is considered as the baseline.

In model M3, the slope is β_1 and the three possible intercepts are β_0, $\beta_0 + \beta_2$ and $\beta_0 + \beta_3$ for the levels `low`, `medium` and `high`, respectively.

In model M4, the intercepts are as in M3 and the three possible slopes are β_1, $\beta_1 + \beta_4$ and $\beta_1 + \beta_5$ for the levels `low`, `medium` and `high`, respectively.

The analysis of variance table is helpful in making a choice between these model. The sequential analysis, using suitable F tests, on the four nested models in the increasing order, namely M1, M2, M3, M4, gives the p-values 0.0014, 0.0019, and 0.1112.

The analysis of variance results suggests use of the parallel line model M3, since the reduction in the mean square from M3 to M4 has a p-value equal to 0.1112.

The the diagnostic checking of M3 does not show any particular problem, and the final model fit is then

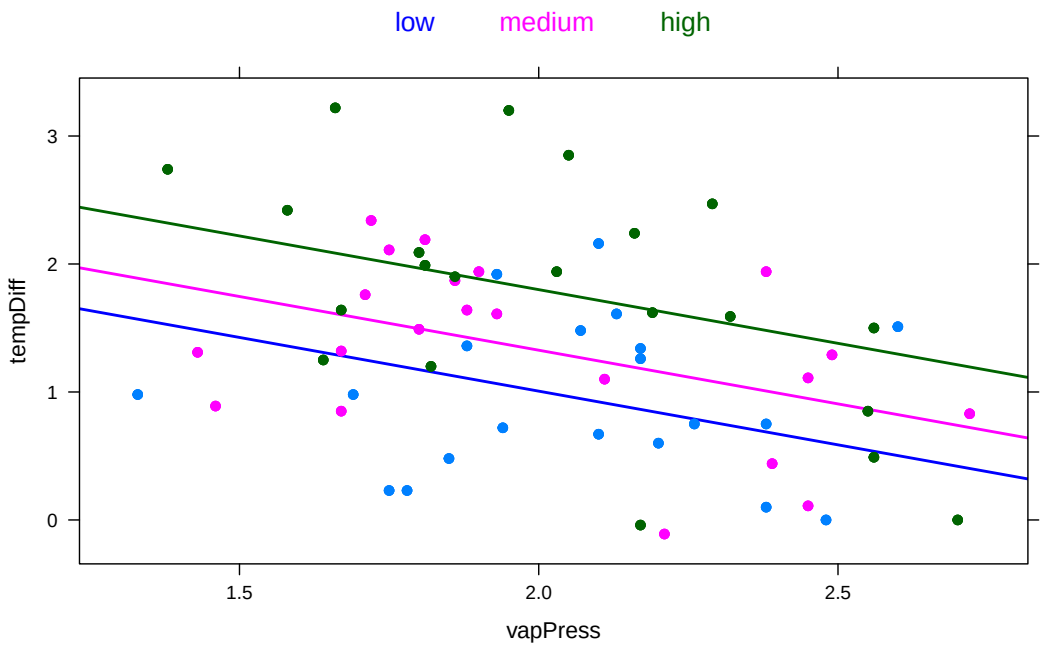


Table of contents

Non-Gaussian response

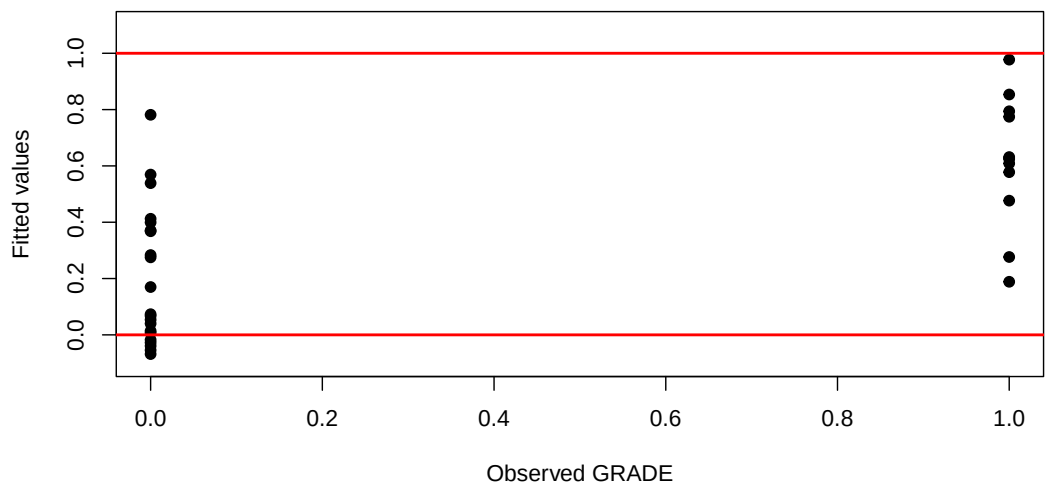
- Regression models may be extended to account for non-Gaussian response variables.
- In particular, the response might admit binary outcomes, that is only two values (usually coded as 0 and 1) described by a Bernoulli distribution, or more generally binomial distributed outcomes.
- In this case, using a linear regression model it is not attractive since predictions for the mean response, which corresponds to an outcome probability, can give off-scale values below 0 or above 1.
- It makes better sense to model the probabilities on a transformed scale; this is what is done in **logistic regression analysis**.
- Logistic regression models belongs to the class of **generalized linear models**.
- These models are characterized by their specific response distribution and by a link function, which transfers the mean value to a scale in which the relation to the explanatory variables is linear and additive.

Example: teaching program

Data from *Econometric Analysis* by W. Greene

Data on the effectiveness of a teaching program. For $n = 32$ students, four variables are observed: `GPA` (grade point average for the period), `TUCE` (test score on economics test), `PSI` (participation in program: yes, 1, and no, 0), `GRADE` (grade increase, 1, or decrease, 0, indicator).

To measure the effect of the explanatory variables on the response `GRADE`, it is tempting to fit a multiple linear regression model so that $E(Y_i) = \beta_0 + \beta_1 \cdot \texttt{GPA}_i + \beta_2 \cdot \texttt{TUCE}_i + \beta_3 \cdot \texttt{PSI}_i$.



Some fitted values are negative: it is unacceptable as $E(Y_i) = \mathrm{P}(Y_i = 1)$.

Generalized linear models

- Generalized Linear Models (GLMs) extend linear regression models so that
  - a more general form of expression for the mean response is allowed, using suitable link functions;
  - various types of distributions for the response can be considered.
- There naturally is quite a large overlap with the material on linear Gaussian models, but there are also some special issues concerning the specific response distribution and link function.
- The main application of GLMs is for modeling **proportions** (binomial data, including Bernoulli data) or **counts** (Poisson data).
- This class of models gives a unified theoretical and computational approach to models that had previously been treated as distinct.
- Here the focus will be on binomial data, but similar considerations apply to count data and other GLMs. Logistic regression models are perhaps the most widely used GLMs.

- In general, GLMs allow a transformation $f(\cdot)$ to the left-hand side of the regression equation. More precisely, instead of assuming

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

a linear model for $f\{E(Y_i)\}$ is specified, namely

$$f\{E(Y_i)\} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

where $f(\cdot)$ is a function usually called the **link function**, whereas $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is, as usual, the **linear predictor**.

- The link function transforms from the scale of Y to the scale of the linear predictor.
- In case of non-Gaussian response, there is no variance parameter. Extensions with a flexible specification of the variance are possible.
- Least squares estimation cannot be used for parameter estimation in GLMs. Estimation methods commonly used include maximum likelihood estimation and Bayesian methods.
- Maximizing the likelihood is equivalent to minimizing the *deviance*, which has a role similar to the residual sum of squares.

Table of contents

The analysis of binary data

- For binary (Bernoulli) data, it is not reasonable that the expected proportion will be a linear function of the explanatory variables. Then, a suitable link function, which goes from $[0, 1]$ to the real line, can be defined.
- The most commonly used one works on the *log odds scale* and it corresponds to the **logit (logistic) link** $f(u) = \log(u/(1-u))$.

  Odds are common for *bookmakers* in betting: if p is a probability, the corresponding odds and log(odds) are, respectively,

$$\text{odds} = \frac{p}{1-p}, \quad \log(\text{odds}) = \log(p/(1-p)) = \log(p) - \log(1-p)$$

  Furthermore,

$$p = \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}}$$

- GLMs for binary data that employ the logit link go under the name of **logistic (multiple) regression models**; they allow to model the log odds as a linear function of suitable explanatory variables.

- Other choices do exist: function $f(u) = \log(\log(1-u))$, which has a connection to survival analysis models, or **probit function** $f(u) = \Phi^{-1}(u)$ (the quantile function of the normal distribution).
- Similarities and differences between linear regression and logistic regression are summarized in the following table

Linear regression	**Logistic regression**
Estimates, std. errors, t-values	Estimates, std. errors, z-values
Sum of squares	Deviance
Residual standard error	–
Fit models by minimizing the residual sum of squares	Fit models by maximizing the log-lik. (minimizing the deviance)
Select models with smaller AIC	Select models with smaller AIC
Compare nested models via F tests	Compare nested models via χ^2 tests
Full set of diagnostic plots	Some diagnostic plots
Partial residual plots	Plots of explanatory variable contributions
R^2 and adjusted R^2	Predictive accuracy

Example: teaching program

Data set on the effectiveness of a teaching program with regard to $n = 32$ students. The 2×2 contingency table for the binary response GRADE (grade increase) and the factor predictor PSI (participation in program) is

		GRADE	
		0	1
PSI	0	15	3
	1	6	8

The observed proportion of GRADE=1 is 3/18=0.167, for students with PSI=0, and 8/14=0.571, for students with PSI=1, that is

$$\begin{aligned} \log(\text{odds}) &= \log(0.167/0.833) = -1.609 \quad \text{for PSI=0} \\ \log(\text{odds}) &= \log(0.571/0.428) = 0.288 \quad \text{for PSI=1} \end{aligned}$$

and the corresponding logistic regression model can be written as

$$\log(\text{odds}) = \beta_0 + \beta_1 \texttt{PSI} + \varepsilon$$

where odds $= P(\texttt{GRADE} = 1)/P(\texttt{GRADE} = 0)$ and the effect due to PSI is coded using a single dummy variable.

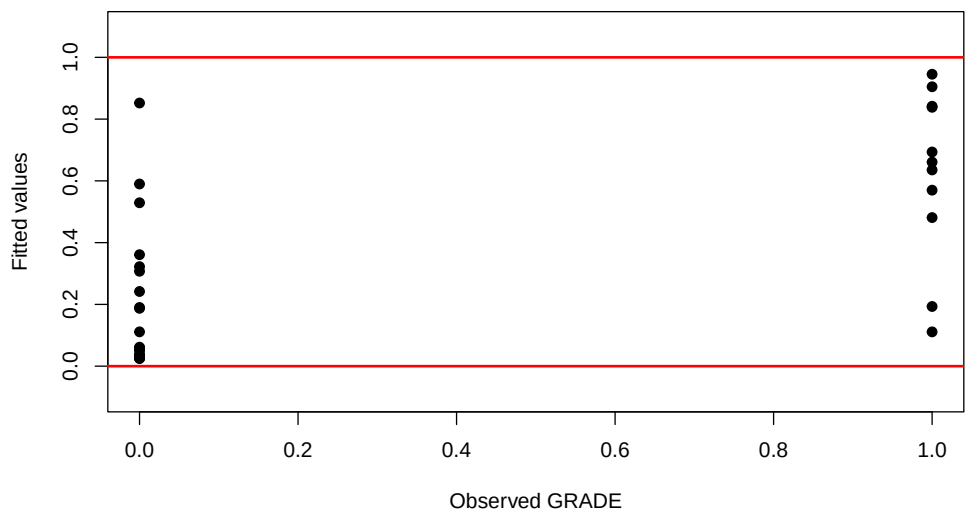
Fitting the model gives the estimates $\widehat{\beta}_0 = -1.61$ (0.632) and $\widehat{\beta}_1 = 1.90$ (0.832). According to the p-value for the z test, the actual significance of `PSI` is not so effective.

Effects due to `GPA` and `TUCE` can then be introduced by considering a logistic multiple regression model with both numerical and factor predictors. Not all the predictors induce a significant effect.

The fitted (values) probabilities from logistic regression are in $[0, 1]$, as

$$\widehat{P}(\texttt{GRADE}_i = 1) = \frac{e^{\log(\widehat{\texttt{odds}}_i)}}{1 + e^{\log(\widehat{\texttt{odds}}_i)}}$$

with $\log(\widehat{\texttt{odds}}_i) = \widehat{\beta}_0 + \widehat{\beta}_1 \texttt{PSI}_i + \widehat{\beta}_2 \texttt{TUCE}_i + \widehat{\beta}_3 \texttt{GPA}_i$.



Predictive accuracy

- For binary data, it makes sense to compare the observed data y_i with the predictions obtained from the model, that is

$$\widehat{y}_i = \begin{cases} 1 & \text{if } \widehat{P}(\texttt{GRADE}_i = 1) \geq 0.5 \\ 0 & \text{if } \widehat{P}(\texttt{GRADE}_i = 1) < 0.5 \end{cases}$$

- Predicted and observed values can be summarized in a 2×2 table. For the teaching program example, it corresponds to

		Observed values	
		0	1
Predicted values	0	18	3
	1	3	8

  An evaluation of the **predictive accuracy** is given by the percentage of correct classifications; in this case, it is equal to 26/32=0.812.

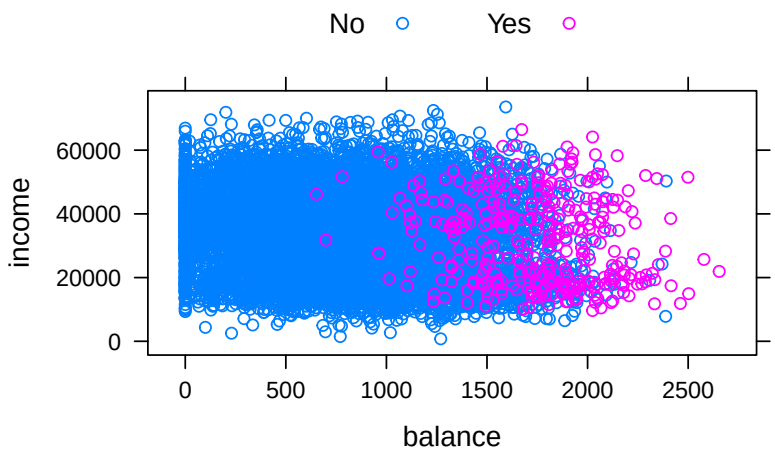
- As usual, when the same data are used twice, the performance of a given model is over-estimated. A more correct assessment uses a **cross-validation** procedure. For the current example, it returns a value around 0.688.

Example: credit card

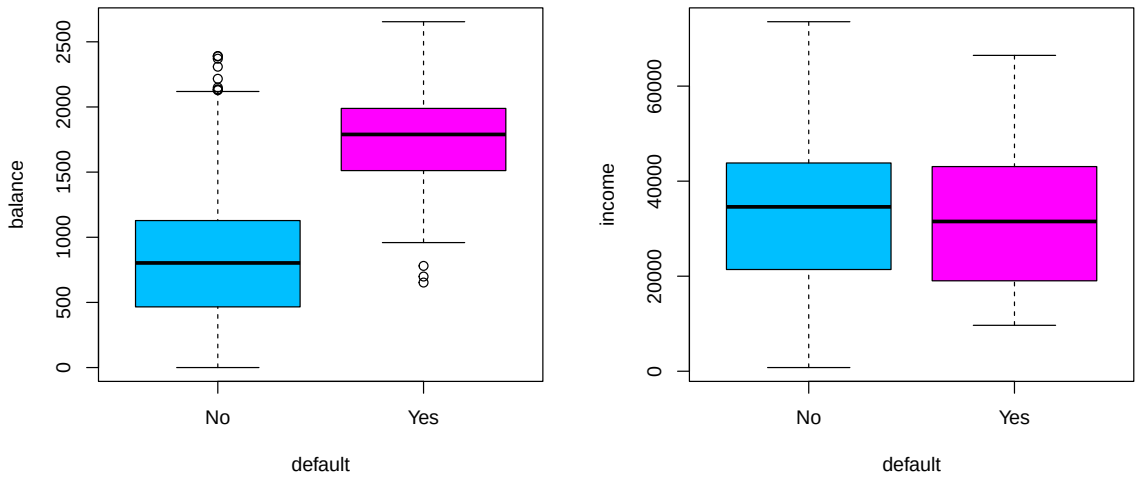
Data set on the defaults on credit card payments. For $n = 10000$ customers, four variables are observed: `DEFAULT` in a given period (yes, 1, and no, 0), `STUDENT` (the costumer is a student: yes, 1, and no, 0), `INCOME` (annual income), `BALANCE` (monthly credit card balance).

To describe the effect of the explanatory variables `STUDENT`, `INCOME` and `BALANCE` on the binary response `DEFAULT` and to predict whether an individual will default.

Only about 3% of people in the data set actually default and individuals who defaulted tended to have higher credit card balances than those who did not, as shown in the following plot



This is confirmed by the boxplots of BALANCE (left) and of INCOME (right) as a function of DEFAULT status



Then a sensible model for DEFAULT is the (simple) logistic regression model

$$\log(\text{odds}) = \beta_0 + \beta_1 \, \texttt{BALANCE} + \varepsilon$$

Fitting the model gives the estimate $\widehat{\beta}_1 = 0.0055$ (0.0002). According to the p-value for the z test, the actual significance of BALANCE is effective.

An increase in BALANCE is associated with an increase in the probability of DEFAULT: a one-unit increase in BALANCE increases the $\log(\text{odds})$ of DEFAULT by 0.0055 units.

Once the coefficients are estimated, it is possible to make predictions on DEFAULT by computing the probability of default for an individual with a given credit card balance x

$$\widehat{P}(\texttt{DEFAULT} = 1|x) = \frac{\exp\{-10.6513 + 0.0055 \cdot x\}}{1 + \exp\{-10.6513 + 0.0055 \cdot x\}}$$

The predicted probability of default for individuals with balances of \$1000 and \$2000 are 0.00576 and 0.586, respectively.

An alternative (simple) logistic regression model can be considered using, as explanatory variable, the qualitative variable STUDENT, coded as a dummy variable

$$\log(\text{odds}) = \beta_0 + \beta_1 \, \texttt{STUDENT} + \varepsilon$$

The estimated coefficient associate with the dummy variable is positive is $\widehat{\beta}_1 = 0.4049$ (0.1150) and the corresponding p-value is statistically significant.

Since $\widehat{\beta}_1 > 0$, students tend to have higher default probabilities than non-students: 0.0431 and 0.0292, respectively.

The joint effects of INCOME, BALANCE and STUDENT can be described using the following multiple logistic regression model

$$\log(\text{odds}) = \beta_0 + \beta_1 \, \texttt{INCOME} + \beta_2 \, \texttt{BALANCE} + \beta_3 \, \texttt{STUDENT} + \varepsilon$$

The estimated coefficients are $\widehat{\beta}_1 = 3.033 \cdot 10^{-6}$ (p-value=0.7115), $\widehat{\beta}_2 = 0.0057$ (p-value$< 2 \cdot 10^{-16}$) and $\widehat{\beta}_3 = -0.6468$ (p-value=0.0062).

Since $\widehat{\beta}_3 < 0$, for a fixed value of INCOME and BALANCE, students are less likely to default than non-students. This seems in contradiction with the previous conclusion.

Although the student default rate is usually below that of the non-student default rate for every value of BALANCE, the overall student default rate is higher (*confounding phenomenon*).

STUDENT and BALANCE are slightly correlated. Students tend to hold higher levels of credit card balances, which is associated with higher default rates.

A student is riskier than a non-student if no other information is available. However, that student is less risky than a non-student with the same credit card balance.

Example: UCB admissions

Data set on the admission at the University of California at Berkeley in the fall of 1973.

Three categorical variables: admission `Admit` (`Admitted`/`Rejected`), `Gender` (`Male`/`Female`) and department `Dept` (`A`, `B`, `C`, `D`, `E`, `F`).

The data set is related to the well-known *Berkeley gender bias case*: female appear discriminated, although no single department is strongly biased against woman.

The aggregate data and the department level data tell opposite stories about gender bias. Most departments have a slight female bias, while the difference on overall application and admission rates causes the aggregate bias to point in the other direction.

A logistic regression model is defined in order to describe the probability of admission. The factor explanatory variables `Dept` and `Gender` are set in this order; a potential interaction effect is also considered.

It is important, for present purposes, to fit `Dept`, thus adjusting for different admission rates in different departments, before fitting `Gender`.

The estimated coefficients are given below

Coefficients	Estimate	SE	p-value
Intercept	0.4921	0.0717	$6.94 \cdot 10^{-12}$
DeptB	0.0416	0.1132	0.71304
DeptC	-1.0276	0.1355	$3.34 \cdot 10^{-14}$
DeptD	-1.1961	0.1264	$< 2 \cdot 10^{-16}$
DeptE	-1.4491	0.1768	$2.49 \cdot 10^{-16}$
DeptF	-3.2619	0.2312	$< 2 \cdot 10^{-16}$
GenderFemale	1.0521	0.2627	$6.21 \cdot 10^{-5}$
DeptB:GenderFemale	-0.8321	0.5104	0.1031
DeptC:GenderFemale	-1.1770	0.2996	$8.53 \cdot 10^{-5}$
DeptD:GenderFemale	-0.9701	0.3026	0.0014
DeptE:GenderFemale	-1.2523	0.3303	0.0002
DeptF:GenderFemale	-0.8632	0.4027	0.0321

Comparison of the nested models using sequential χ^2 tests shows that there is a clear effect of `Dept` on the admission rate, while there is no detectable main effect of `Gender`.

The significant interaction term suggests that there are department-specific gender biases, which average out to reduce the main effect of `Gender` to close to zero.

Concerning the individual model coefficients:

- the first six coefficients relate to overall admission rates, for males, in the six departments;
- the strongly significant positive coefficient for `GenderFemale` indicates that $\log(\text{odds})$ is increased by 1.05, in department `A`, for females relative to males;
- in departments `C` and `E` the $\log(\text{odds})$ is reduced for females, relative to males.