Statistical Methods

Linear Models, Michaelmas Term, 2015

Marco Scutari

scutari@stats.ox.ac.uk
Department of Statistics
University of Oxford

February 23, 2016

Course Information

Lectures

Week 2:	Friday 9am/10am.
Week 3:	Tuesday 4pm/5pm, Friday 9am/10am.
Week 4:	Tuesday 4pm/5pm, Friday 9am/10am.
Week 5:	Tuesday 4pm/5pm.

Practicals

Week 3:	Friday 1:30/3/4:30pm (not assessed)
Week 5:	Thursday 1:30/3/4:30pm $\longrightarrow$ submit by 10am Monday Week 6

Reference Books (further references in the next slides)

DS	Draper NR, Smith H (1998). Applied Regression Analysis. Wiley, 3rd edition.
WB	Weisberg S (2013). Applied Linear Regression. Wiley, 4th edition.
M	Montgomery DC (2013) Design and Analysis of Experiments. Wiley, 8th edition.

Other Useful Books on Linear Models

- Cox DR and Reid N (2000) The Theory of the Design of Experiments. Chapman & Hall.
- Fahrmeir L, Knelb, T, Lang S, Marx B (2013). Regression: Models, Methods and Applications. Springer.
- Hastie T, Tibshirani R, Friedman J (2009). The Elements of Statistical Learning. Springer, 2nd edition.
- Burnham KP, Anderson DR (2002). Model Selection and Multimodel Inference. Springer.
- Cook RD, Weisberg S (1982). Residuals and Influence in Regression. Chapman & Hall.
- Puntatnen S, Styan GPH, Isotalo J (2013). Formulas Useful for Linear Regression Analysis and Related Matrix Theory. Springer.
- Stapleton JH (2009). Linear Statistical Models. Wiley, 2nd edition.
- Rao CR, Toutenburg H, Shalabh, Heumann C (2008). Linear Models and Generalizations. Springer, 3rd extended edition.

Overview

1. Definitions and Notation

   [DS, 1.1 & 4.1]

2. Simple and Multiple Linear Regression

   [DS 1.2; WB, 2 & 3.1–3.4]

3. Prediction and Model Diagnostics

   [DS 2; WB, 8 & 9]

4. Model Selection and Analysis of Variance

   [DS 1.3, 4.3, 22 & 23; WS, 3.5]

5. Experimental Design

   [M 3 & 4]

6. Robust and Advanced Regression methods

   [DS 25]

Definition and Notations

Linear Models: the Definition

Suppose we have, for each of $i = 1, \ldots, n$ observations,

- a numeric variable of interest y_i, the response;
- and a set of p explanatory variables or regressors $x_{i1}, x_{i2}, \ldots, x_{ip}$.

Then in a linear model we have that, for each observation,

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \varepsilon_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \quad (1)$$

where β_0 is the intercept (*e.g.* $x_{i0} = 1$ for all i), and ε_i is an error term with mean zero (*e.g.*. $\mathrm{E}(\varepsilon_i) = 0$). The model is linear in the parameters $\beta_0, \ldots, \beta_p$, which are called the regression coefficients.

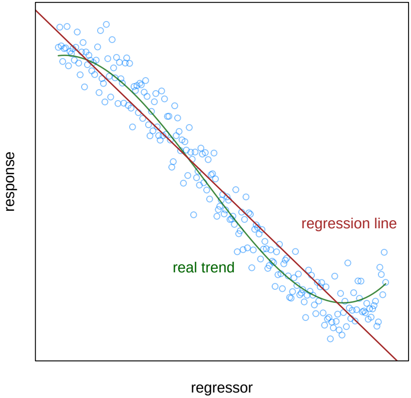
It is also commonly assumed that the ε_i are uncorrelated or independent, and often that they have the same variance σ^2_ε (homoscedasticity) as opposed to individual variances $\sigma^2_{\varepsilon_i}$ (heteroscedasticity).

Linear Models: What is Stochastic and What is Not

- The explanatory variables $x_{i1}, x_{i2}, \ldots, x_{ip}$ are assumed to be fixed effects, and the model is specified conditional on their values. This implies that they are observed without error, otherwise they would be stochastic, and that there is no missing value.
- The response variable y_i is assumed to be stochastic through the error term.
- The regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ and the variance σ^2_ε of the error term are unknown parameters, to be estimated from the data.

The role of the response and the regressors is not symmetric because of the conditioning.

Why Should We Care About Linear Models?



Real-world phenomena can rarely be expressed as linear functions of some parameters $\beta_0, \beta_1, \ldots, \beta_p$. However, we can view a linear model as a first-order approximation of more complicated models, and it is quite flexible in that we can transform the response and explanatory variables to make the real trend as linear as possible.

An Alternative Notation, from Linear Algebra

To make the notation more concise, we can stack the equations for the n observations into matrices and vectors,

$$\underset{n\times 1}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \underset{n\times(p+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \underset{(p+1)\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

so that we can write the whole system of equations as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2}$$

Using this notation reformulates the problem in the context of linear algebra, and thus makes it easier to apply results on manipulating and solving systems of linear equations in matrix form.

Fitting a Linear Model

The aim in fitting a linear model is estimating the regression coefficients $(\hat{\beta}_0, \ldots, \hat{\beta}_p)$, and subsequently the fitted values $(\hat{y}_i)$, the residuals $(\hat{\varepsilon}_i)$ and their variance $(\hat{\sigma}^2_\varepsilon)$. The three most fundamental ways of doing this are:

- least squares, from optimisation theory and linear algebra;
- and maximum likelihood, which unlike the first two requires a completely specified model.

Among more advanced techniques:

- penalised least squares and penalised likelihood;
- robust regression methods;
- weighted least squares;
- and splines.

Estimation: Least Squares

The simplest form of least squares estimation is ordinary least squares (OLS), which assumes:

- that the data are uncorrelated, *i.e.* $\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0$ if $i \neq j$;
- and that the residuals have mean zero and they all have the same variance, *i.e.* $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{VAR}(\varepsilon_i) = \sigma^2_\varepsilon$ for all observations.

The regression coefficients are estimated so as to minimise the residual sum of squares (RSS):

$$\begin{aligned}\{\hat{\beta}_0, \ldots, \hat{\beta}_p\} &= \operatorname*{argmin}_{\beta_0, \ldots, \beta_p} \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \operatorname*{argmin}_{\beta_0, \ldots, \beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \qquad (3)\end{aligned}$$

Estimation: Least Squares from Linear Algebra

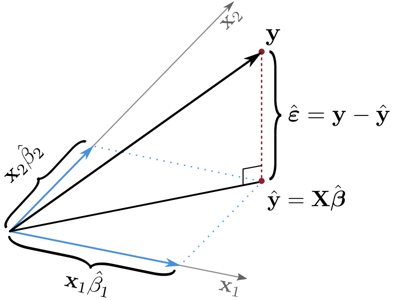
In matrix form, we can reformulate that a linear system and solve it as:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} \\ \mathbf{X}^T\mathbf{y} &= \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} &= \hat{\boldsymbol{\beta}} \end{aligned} \tag{4}$$

under the condition that $\mathbf{X}$ is full rank, that is, that the explanatory variables that make up the columns of $\mathbf{X}$ are linearly independent (orthogonal).

If that is not the case $\mathbf{X}^T\mathbf{X}$ is not invertible and the estimate requires advanced linear algebra techniques to compute (*e.g.* Moore-Penrose pseudoinverse).

Estimation: Least Squares, a Geometric Interpretation


$\mathbf{x}_2$
$\mathbf{y}$
$\hat{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}}$
$\mathbf{x}_2\hat{\beta}_2$
$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$
$\mathbf{x}_1\hat{\beta}_1$
$\mathbf{x}_1$

Least squares estimation can be intuitively explained as the orthogonal projection of the response $\mathbf{y}$ on the plane defined by the linear combination of the explanatory variables, with the residuals $\hat{\varepsilon}$ being orthogonal to the projection $\hat{\mathbf{y}}$.

In matrix terms, the projection matrix is $\mathbf{P}$,

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{P}\mathbf{y}, \qquad (5)$$

and $\hat{\varepsilon}$ is indeed orthogonal to $\hat{\mathbf{y}}$ because by definition $\mathbf{P}^2 = \mathbf{P}$:

$$\hat{\mathbf{y}}^T(\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{P}\mathbf{y})^T(\mathbf{y} - \mathbf{P}\mathbf{y}) = \cancel{\mathbf{yPy}} - \cancel{\mathbf{yPPy}} = 0. \qquad (6)$$

Estimation: Least Squares, a Summary

1. Estimate the regression coefficients:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{7}$$

2. Estimate the fitted values using the regression coefficients:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{8}$$

3. Estimate the residuals and their variance:

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} \qquad \text{and} \qquad \sigma^2_{\varepsilon} = \frac{1}{n}\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \tag{9}$$

Note that no distributional assumptions are involved, with the exception of the conditions on the ε_i, and thus this still is a nonparametric approach.

Estimation: Maximum Likelihood

If we assume that errors follow a normal distribution, that is $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$ and $\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0$, we have that the errors are now independent and identically distributed (iid). The likelihood for the model is

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2_\varepsilon; \mathbf{y}, \mathbf{X}) &= \frac{1}{(2\pi\sigma^2_\varepsilon)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2\sigma^2_\varepsilon} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \Sigma^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \qquad (10) \end{aligned}$$

where Σ is the covariance matrix of the errors

$$\Sigma = \sigma^2_\varepsilon \mathrm{I}_n = \begin{bmatrix} \sigma^2_\varepsilon & 0 & \cdots & 0 \\ 0 & \sigma^2_\varepsilon & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_\varepsilon \end{bmatrix}. \qquad (11)$$

Estimation: Maximum Likelihood and Least Squares

Since $|\Sigma| = (\sigma^2_\varepsilon)^n$ and $\Sigma^{-1} = \frac{1}{\sigma^2_\varepsilon}\mathrm{I}_n$, the log-likelihood has the form

$$l(\boldsymbol{\beta}, \sigma^2_\varepsilon; \mathbf{y}, \mathbf{X}) \propto -\frac{n}{2}\log\sigma^2_\varepsilon - \frac{1}{2\sigma^2_\varepsilon}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \qquad (12)$$

so maximising it to compute $\hat{\boldsymbol{\beta}}$ gives the same estimates as ordinary least squares:

$$\operatorname*{argmax}_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \sigma^2_\varepsilon; \mathbf{y}, \mathbf{X}) = \operatorname*{argmin}_{\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \qquad (13)$$

However, the estimate of the residual variance σ^2_ε is

$$\hat{\sigma^2_\varepsilon} = \frac{1}{n-p-1}\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} \neq \frac{1}{n}\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}, \qquad (14)$$

with the squared residuals divided by the sample size minus the number of regression coefficients plus the intercept.

Estimation: Why the Estimates of β Are the Same?

The only constraints set by ordinary least squares are $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{VAR}(\varepsilon_i) = \sigma^2_\varepsilon$. The maximum entropy distribution, which is the distribution that maximises the expected likelihood under these constraints, is the normal distribution. In other words, if the only information we have is $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{VAR}(\varepsilon_i) = \sigma^2_\varepsilon$, the distribution that on average has the highest likelihood is the normal distribution.

And among normal distributions, that with the highest likelihood minimises the sum of the squared residuals (*i.e.* $n\sigma^2_\varepsilon$) through the choice of $\hat{\boldsymbol{\beta}}$. Which is the same as estimating least squares:

$$\operatorname*{argmax}_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \sigma^2_\varepsilon; \mathbf{y}, \mathbf{X}) = \operatorname*{argmin}_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{15}$$

This idea is sometimes expressed saying that normal distribution is isoperimetric with the L_2 (aka Euclidean) norm, *e.g.* points having the same L_2 distance from the expectation have the same likelihood.

Estimation: Why the Estimate of σ^2_ε Is Not?

Now that we a completely specified distribution for the errors, we can compute the (exact) distribution of $\hat{\boldsymbol\beta}$, which is normal with

$$\mathrm{E}(\hat{\boldsymbol\beta}) = \mathrm{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) = \mathrm{E}(\cancel{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}}\boldsymbol\beta) = \boldsymbol\beta \qquad (16)$$

$$\mathrm{VAR}(\hat{\boldsymbol\beta}) = \mathrm{VAR}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) = \sigma^2_\varepsilon(\mathbf{X}^T\mathbf{X})^{-1} \qquad (17)$$

and the (exact) distribution of $\hat\sigma^2_\varepsilon$, which is χ^2 because $\varepsilon_i^2 \sim \sigma\chi^2_1$. The estimator we obtained from the ordinary least squares is biased so we use

$$\hat{\sigma^2_\varepsilon} = \frac{1}{n-p-1}\hat{\boldsymbol\varepsilon}^T\hat{\boldsymbol\varepsilon} \qquad \Rightarrow \qquad (n-p-1)\frac{\hat\sigma^2_\varepsilon}{\sigma^2_\varepsilon} \sim \chi^2_{n-p-1} \qquad (18)$$

in which the degrees of freedom are reduced by the presence of the regressors and the intercept as estimated parameters in the model.

Optimality of Least Squares and Maximum Likelihood

The Gauss-Markov theorem tells us that if we assume $\mathrm{E}(\varepsilon_i) = 0$ $\mathrm{VAR}(\varepsilon_i) = \sigma^2_\varepsilon$ and $\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0$, the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is that we obtain from ordinary least squares (OLS) estimator. In other words, it has the lowest $\mathrm{VAR}(\hat{\boldsymbol{\beta}})$, as compared to other unbiased linear estimates.

Compared to ordinary least square, the advantage of using the maximum likelihood approach is that we can make use of many parametric tests that have simple expressions and simple reference distributions as opposed advanced computational simulations.

Modelling Workflow: From the Data to Inference

0. Data Preprocessing: check whether the data can be modelled with a linear model, and whether a the classic estimators are sufficient or more advanced ones are needed due to missing values, non-numeric response, correlated or heteroscedastic observations, etc.
1. Model Selection: decide which variables to use as regressors for the response of interest, and whether to transform them and/or the response to make relationships linear.
2. Model Estimation: estimate the parameters of the modes.
3. Model Validation: check that the assumptions of the model are met, check that the model fit the data well and check whether the model predicts the data well (you can't have both so which is more important depends on the goals of the analysis).
4. Inference: use the model to answer relevant questions, either through closed form inference or (more often) simulation.

Simple and Multiple Linear Regression

The Simplest Regression Model

A simple linear regression has just a single regressor,

$$y_i = \beta_0 + x_{i1}\beta_i + \varepsilon_i \qquad \text{or} \qquad \mathbf{y} = \beta_0 + \mathbf{x}_1\beta_1 + \boldsymbol{\varepsilon}. \tag{19}$$

The value estimated by the model for y_i, called the fitted value and usually denoted $\hat{y}_i$, is

$$\mathrm{E}(y_i) = \mathrm{E}(\beta_0 + x_{i1}\beta_1 + \varepsilon_i) = \beta_0 + x_{i1}\beta_1 + \cancel{\mathrm{E}(\varepsilon_i)} = \beta_0 + x_{i1}\beta_1; \tag{20}$$

and the corresponding estimate for the error, the residual $\hat{\varepsilon}_i$, is

$$\hat{\varepsilon}_i = y_i - \mathrm{E}(y_i) = y_i - \hat{y}_i. \tag{21}$$

Along with the regression coefficients, they are the key quantities in the estimation and interpretation of the linear model.

Example: the Marks Data Set

The `marks` data set from the classic Mardia, Kent & Bigby’s book on “Multivariate Analysis” (1979) consists of the exam scores of 88 students across 5 different topics: mechanics (`MECH`), vectors (`VECT`), algebra (`ALG`), analysis (`ANL`) and statistics (`STAT`). The scores are on a 0 to 100 scale.

```
> marks = read.table("marks.txt", header = TRUE)
> str(marks)
’data.frame’: 88 obs. of  5 variables:
$ MECH: num  77 63 75 55 63 53 51 59 62 64 ...
$ VECT: num  82 78 73 72 63 61 67 70 60 72 ...
$ ALG : num  67 80 71 63 65 72 65 68 58 60 ...
$ ANL : num  67 70 66 70 70 64 65 62 62 62 ...
$ STAT: num  81 81 81 68 63 73 68 56 70 45 ...
```

Fitting a Simple Linear Regression

The `lm()` function is the standard tool to fit a linear regression model in R. It takes a formula of the form $\mathbf{y} \sim \mathbf{x}_1$ (the intercept is implicitly included as well) and a data set including the variables in the formula.

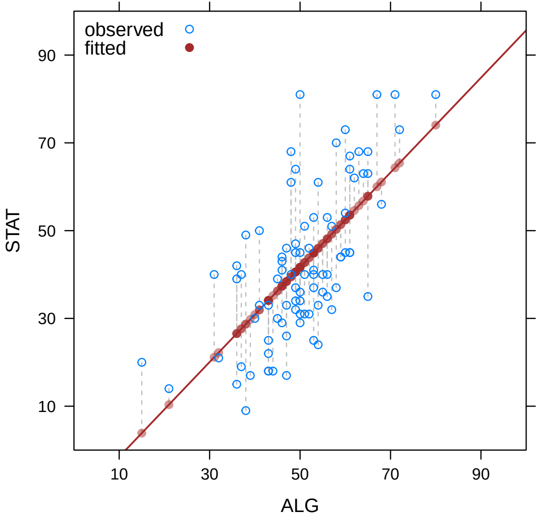
```
> m = lm(STAT ~ ALG, data = marks)
> m

Call:
lm(formula = STAT ~ ALG, data = marks)

Coefficients:
(Intercept)          ALG
     -12.32         1.08
```

The `m` objects contains many quantities that are of use in subsequent analyses, but it prints just the model formula and the regression coefficients.

The Regression Line



The Intercept and the Regression Coefficient

In the case of a simple linear regression model, the estimator for the intercept β_0 is the mean of the y_i adjusted for the mean of the x_i,

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y} - \hat{\beta}_1 \bar{x} \tag{22}$$

and the estimator for the regression coefficient β_1 is

$$\hat{\beta}_1 = \frac{\mathrm{COV}(\mathbf{x}_1, \mathbf{y})}{\mathrm{VAR}(\mathbf{x}_1)} \tag{23}$$

as the general expression in (4) simplifies due to the presence of a single explanatory variable. Note that

$$\hat{\beta}_1 \propto \mathrm{COR}(\mathbf{x}_1, \mathbf{y}) \quad \text{and}$$

$$\hat{\beta}_1 = \mathrm{COR}(\mathbf{x}_1, \mathbf{y}) \text{ if } \mathrm{VAR}(\mathbf{x}_1) = \mathrm{VAR}(\mathbf{y}) = 1. \tag{24}$$

Regression Coefficients as Correlations

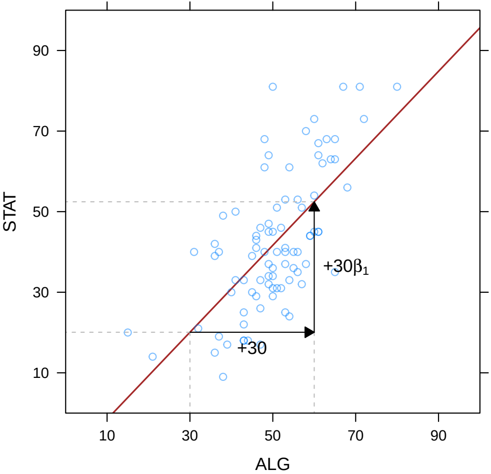
As a result of (24) we have that

$$\beta_1 = 0 \qquad \text{if and only if} \qquad \mathrm{COV}(\mathbf{x}_1, \mathbf{y}) = 0 \tag{25}$$

since $\beta_1 = \mathrm{E}(\hat{\beta}_1) \propto \mathrm{COV}(\mathbf{x}_1, \mathbf{y})$. So a regression coefficient is a function of the correlation between the response and the explanatory variable; and testing for zero correlation between $\mathbf{x}_1$ and $\mathbf{y}$ is equivalent to testing whether $\hat{\beta}_1$ is equal to zero.

The significance of a regression coefficient is often tested using the asymptotic (normal) distribution of the maximum likelihood estimator with mean from (16) and variance from (17). But we can also use the exact (Student's) t test and asymptotic (normal) Fisher's Z test for correlation among others.

Regression Coefficients as Slope



A second, geometric interpretation of a regression coefficient is that of the slope of the regression line.

Key Quantities Pre-Computed by `lm()`

- fitted values $\hat{y}_i$

```
> fitted(m)
      1       2       3       4       5       6
60.0096 74.0443 64.3279 55.6912 57.8504 65.4075 [...]
```

- residuals $\hat{\varepsilon}_i$

```
> resid(m)
      1       2       3       4      5      6
20.9903  6.9556 16.6720 12.3087 5.1495 7.5924 [...]
```

- intercept and regression coefficients $\hat{\boldsymbol{\beta}}$

```
> coef(m)
(Intercept)         ALG
  -12.32289     1.07959
```

Model Information from summary(m)

```
> summary(m)

Call:
lm(formula = STAT ~ ALG, data = marks)

Residuals:
    Min      1Q  Median      3Q     Max
-22.850  -8.741  -1.162   7.720  39.343

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.3229     6.7633  -1.822   0.0719 .
ALG           1.0796     0.1308   8.251 1.64e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 12.97 on 86 degrees of freedom
Multiple R-squared:  0.4419, Adjusted R-squared:  0.4354
F-statistic: 68.09 on 1 and 86 DF,  p-value: 1.638e-12
```

`summary(m)`: Residuals

First of all, `summary()` prints a few key quantiles of the residuals.

```
Residuals:
    Min      1Q  Median      3Q     Max
-22.850  -8.741  -1.162   7.720  39.343
```

They are assumed to be symmetric and centred at zero, so the median should be small and the 1st and 3rd quartiles should be symmetric. Maximum and minimum are usually not because it takes just a single misbehaving observation (*i.e* an outlier) to make them very different. The mean of the residuals is indeed very close to zero

```
> mean(resid(m))
[1] -2.317458e-16
```

and the maximum likelihood estimate of $\hat{\sigma}^2_{\varepsilon}$ (χ^2_{86}) is:

```
> var(resid(m)) * (nrow(marks) - 1) / (nrow(marks) - 2)
[1] 168.1175
```

The residual standard error is the corresponding standard deviation.

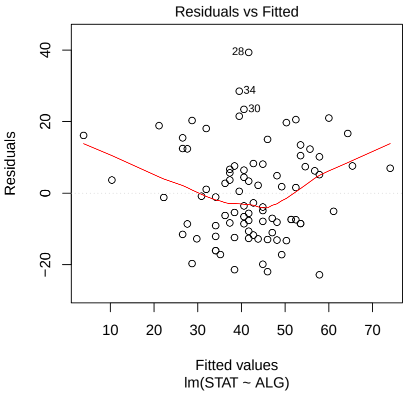
`plot(m)`: Graphical Diagnostics for the Residuals

`plot(m)` produces the most widely used graphical diagnostics to check whether the residuals violate the assumptions of the model.

- Residuals vs fitted values: we know from least squares that residuals and fitted values are orthogonal, so we should not be able to see any trend in the plot; and the range should remain constant because the residuals are homoscedastic.
- Quantile-quantile plot: the residuals are assumed to be normal, so we can check them against the theoretical quantiles.
- Cook's distance plot: Cook's distance measures the influence of each y_i on the model (the leverage) through the predicted $\hat{y}_i^*$ obtained by dropping y_i:

$$D_i = \sum_{j=1}^{n} \frac{(\hat{y}_j - \hat{y}_j^*)^2}{(1+p)\hat{\sigma}_\varepsilon^2}. \tag{26}$$

Residuals vs Fitted Values (Good)



There is no visible trend, the residuals are a shapeless cloud of points that is approximately symmetric around zero. The most extreme are labelled with the row number of the corresponding observation in $\mathbf{X}$. If we do not consider those points, the range of the residuals is $[-20, 20]$ which is indeed symmetric around zero.

The red line is the mean of the residuals at each point of the x axis; it only departs from zero at the far ends. That happens because there are few points with extreme fitted values, so the estimates of the corresponding mean residuals is very noisy.

More On Cooks' Distance

Despite appearances, computing D_i does not in fact require to estimate one linear model for each observation, which would be computationally demanding. It can be rewritten as

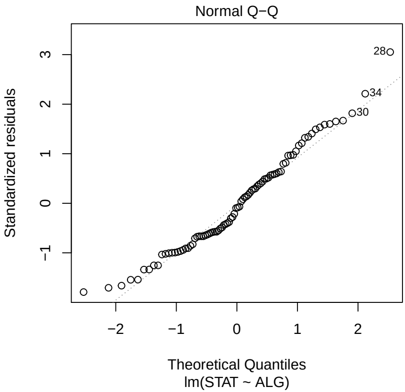
$$D_i = \sum_{j=1}^{n} \frac{(\hat{y}_j - \hat{y}_j^*)^2}{(1+p)\hat{\sigma}_\varepsilon^2} = \frac{\hat{\varepsilon}_i^2}{(p+1)\hat{\sigma}_\varepsilon^2} \cdot \frac{h_{ii}}{(1-h_{ii})^2} \qquad (27)$$

where h_{ii} is the (i,i) element of $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. In this context it is called the hat matrix, but it's just the projection matrix $\mathbf{P}$ from (5) in least squares definition and it's available from the original model fit. Plus,

$$\mathrm{VAR}(\hat{\boldsymbol{\varepsilon}}) = \hat{\sigma}_\varepsilon^2(\mathbf{I}_n - \mathbf{H}) \qquad (28)$$

which means that $\mathrm{VAR}(\hat{\varepsilon}_i) = \hat{\sigma}_\varepsilon^2(1-h_{ii})$, another useful quantity to check for homoscedasticity.

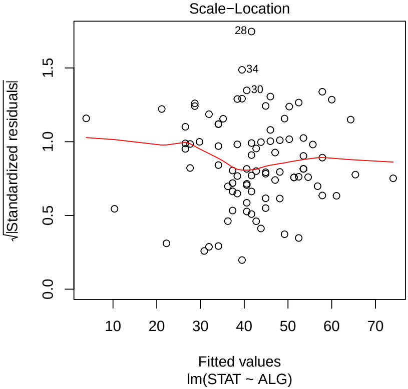
Quantile-Quantile Plot (Good)



The standardised quantiles of the residuals are on the y axis and the theoretical quantiles of a standard normal distribution are on the x axis.

Almost all the points are close to the diagonal of the plot (the grey dotted line), and there is no visible pattern. Only the two most extreme residuals at each end do not match the corresponding quantiles very well, which is expected and perfectly normal.

Standardised Residuals vs Fitted Values (Good)

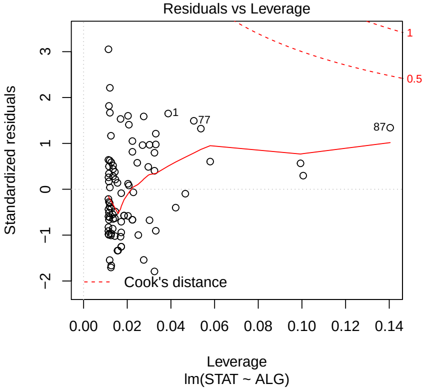


This plot is also called a spread-level plot.

The absolute values of the residuals (standardised with $\sqrt{\mathrm{VAR}(\hat{\varepsilon})}$) are useful to check whether there are patterns that were not visible in the previous plots.

The red line is the mean of the $\sqrt{\hat{\varepsilon_i}}$ around each point of the x axis. A flat, horizontal suggests the residuals are homoscedastic; any trend suggests that residuals are heteroschedastic.

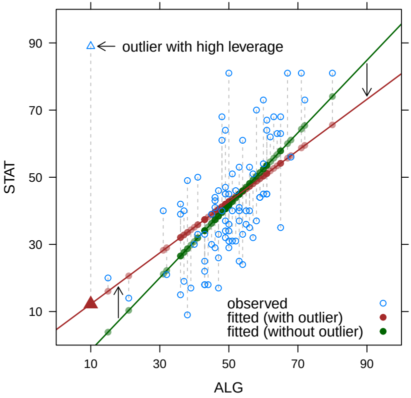
Residuals vs Leverage (Good)



Plotting standardised residuals against Cook's distance is useful to check which residuals may be problematic. A large standardised residual is not problematic per se; and even small residuals can have high leverage.

The dashed red lines are thresholds for Cook's distance; for residuals beyond the $D_i = 1$ threshold there is strong evidence the corresponding observation is an outlier. Residuals close to or beyond the $D_i = 0.5$ should be investigated as well.

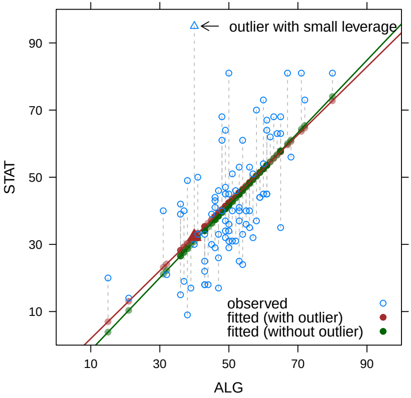
Outliers: High Leverage (Bad)



The green line is the model fitted from the `marks` data; the red line is the model fitted from the `marks` data plus an outlier (the blue triangle). The blue dots are the observations in the original `marks` data.

The outlier has high leverage because it has both an extreme `ALG` value and an extreme `STAT` value – both the response and the explanatory variable have values that are far from the respective averages.

Outliers: Small Leverage (Not as Bad)



On the other hand, in this case only `STAT` has an extreme value. So while it is still an outlier, it does have little leverage because `ALG` has a value that is near the centre of the range of the observed values.

Statistical Tests Used as Diagnostics for the Residuals

There are many possible ways for residuals to violate the assumptions required by a linear model; and it is impossible to cover all of them effectively by testing. However, two approaches are commonly suggested in classic literature.

- Testing for normality with Shapiro-Wilk's test or any other distribution test such as Anderson-Darling or Jarque-Bera.

```
> shapiro.test(resid(m))

    Shapiro-Wilk normality test

data:  resid(m)
W = 0.9723, p-value = 0.05602
```

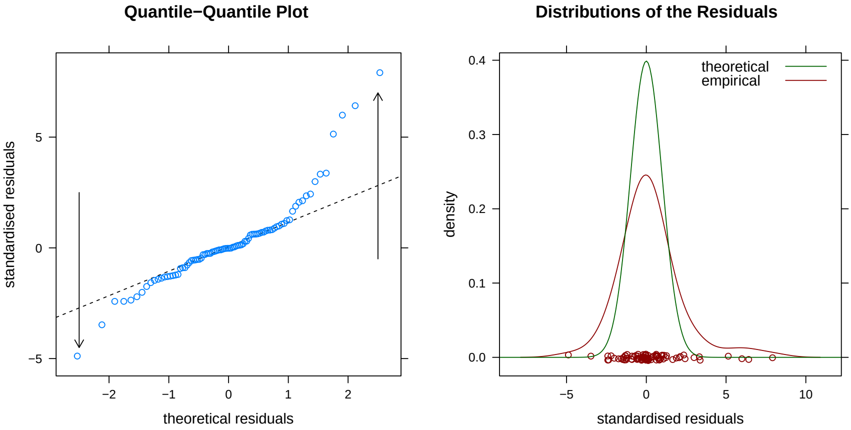
```
> library(nortest)
> ad.test(resid(m))

    Anderson-Darling normality test

data:  resid(m)
A = 0.7442, p-value = 0.05068
```

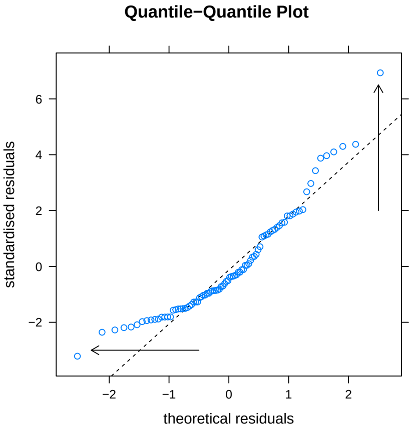
- Testing for correlation among the residuals, for example autocorrelation patterns similar to those in time series with the Durbin-Watson test.

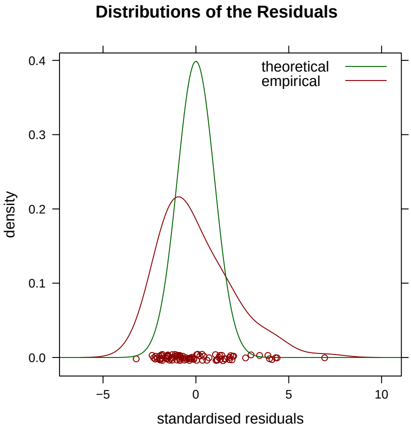
Departures from Normality: Heavy Tails (Bad)



With heavy tails we have symmetric departures on the tails of the quantile-quantile plot.

Departures from Normality: Skewness (Bad)





With skewness we have asymmetric departures on the tails of the quantile-quantile plot: up-and-left if the right tail is too heavy, down-and-right if the left tail is too heavy.

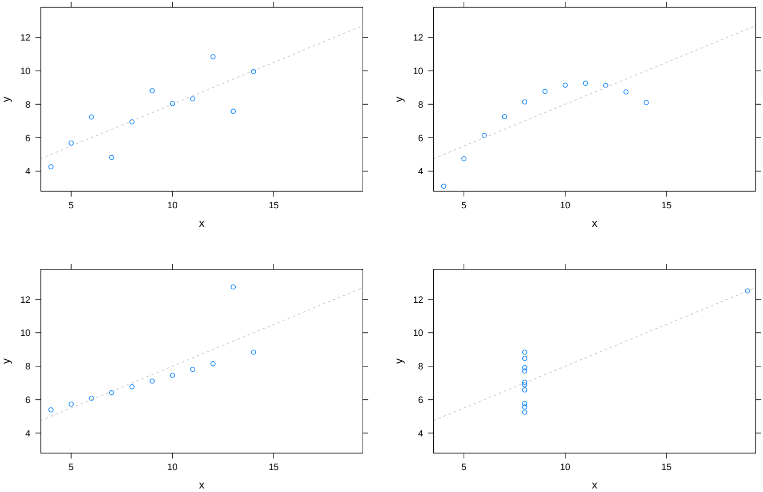
Why Graphical Diagnostics are Important

Examining graphical diagnostics for a linear model is important and complements the use of numeric diagnostics. In other words:

- some problems are easier to spot in a diagnostic plot;
- some problems are better assessed with numeric indicators.

A classic example of this fact is known as Anscombe's Quartet, a set of four simple regression models that have the same regression coefficients, residual standard error and R^2. However, only the first really fits the assumptions, the other are pathological cases that are apparent from even the simplest graphical diagnostics.

Anscombe's Quartet



Anscombe, FJ (1973) Graphs in statistical analysis. American Statistician, 27:17–21.

summary(m): Regression Coefficients

summary() then prints the regression coefficients and their significance.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.3229     6.7633  -1.822   0.0719 .
ALG           1.0796     0.1308   8.251 1.64e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

For each coefficient (including the intercept), summary() reports the p-value for $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$ computed with the (asymptotic normal or t_{n-2}) Wald test:

```
> 2 * pt(abs(-12.3229 / 6.7633), df = nrow(marks) - 2, lower.tail = FALSE)
[1] 0.07188737
```

Typically, we are not much interested in the significance of the intercept as much as in that of the regression coefficients.

Better Ways of Testing for the Significance of a Regressor

The Wald test is notoriously unreliable, so better options should be used in most cases. If $\rho = \mathrm{COR}(\mathbf{y}, \mathbf{x}_1)$, possible choices are:

- the loglikelihood ratio test

$$D = -n \log(1 - \rho^2) \sim \chi^2_1 \qquad \text{(asymptotic);} \qquad (29)$$

- the t test for correlation

$$t = \rho \sqrt{\frac{n-2}{1-\rho^2}} \sim t_{n-2} \qquad \text{(exact);} \qquad (30)$$

- Fisher's Z test

$$Z = \log\left(\frac{1+\rho}{1-\rho}\right) \frac{\sqrt{n-3}}{2} \sim N(0, 1) \qquad \text{(asymptotic).} \qquad (31)$$

Better Ways of Testing for the Significance of a Regressor

```
> -2 * (logLik(lm(STAT ~ 1, data = marks)) -
+        logLik(lm(STAT ~ ALG, data = marks)))
[1] 51.31894
> rho = cor(marks$STAT, marks$ALG)
> - nrow(marks) * log(1 - rho^2)
[1] 51.31894
> pchisq(51.31894, df = 1, lower.tail = FALSE)
[1] 7.851e-13 # loglikelihood ratio test
> cor.test(marks$STAT, marks$ALG) # exact t test

      Pearson's product-moment correlation

data:  marks$STAT and marks$ALG
t = 8.252, df = 86, p-value = 1.638e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5289 0.7673
> (log(1 + rho) - log(1 - rho)) * sqrt(nrow(marks) - 3)/2
[1] 7.387
> 2 * pnorm(7.387, lower.tail = FALSE)
[1] 1.502e-13 # Fisher's Z test
```

`summary(m)`: Goodness of Fit

Finally `summary()` prints three indicators of goodness of fit, *e.g.* how well the estimated model fits the data.

```
Residual standard error: 12.97 on 86 degrees of freedom
Multiple R-squared:  0.4419, Adjusted R-squared:  0.4354
F-statistic: 68.09 on 1 and 86 DF,  p-value: 1.638e-12
```

They are:

- the residual standard error, *e.g.* $\sqrt{\hat{\sigma}^2_{\varepsilon}}$;
- the R^2 coefficient, also known as coefficient of determination or multiple correlation coefficients, *e.g.* $R^2 = \mathrm{COR}(\mathbf{y}, \hat{\mathbf{y}})^2$;
- the F test for the fitted model against the null model including only the intercept.

The R^2 Coefficient

The R^2 coefficient can equivalently be estimated as the proportion of variability of $\mathbf{y}$ explained by the model, that is,

$$R^2 = \frac{\mathrm{VAR}(\hat{\mathbf{y}})}{\mathrm{VAR}(\mathbf{y})} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \in [0, 1]. \tag{32}$$

Clearly, the higher the R^2 the better the model fits the data. It's easy to show it is bound in $[0, 1]$:

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{total squares}} = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 =$$

$$= \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{\text{residual squares}} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{\text{regression squares}} \geqslant \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \tag{33}$$

The F Test for Nested Simple Linear Models

The F test is an overall goodness-of-fit test comparing the fitted model $\mathcal{M}_1$ with a (nested) barebone model $\mathcal{M}_0$ containing only the intercept,

$$\begin{aligned} H_0 &: \text{model is } \mathcal{M}_0, \mathbf{y} = \beta_0 \\ H_1 &: \text{model is } \mathcal{M}_1, \mathbf{y} = \beta_0 + \mathbf{x}_1\beta_1. \end{aligned} \tag{34}$$

The test statistic is

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\hat{\sigma}^2_\varepsilon} \sim F_{1,n-2}. \tag{35}$$

Rejecting the null hypothesis suggests that the response variable depends on $\mathbf{x}_1$; and the p-value of this test will be the same as the Wald test for β_1.

Multiple Linear Regression

A General Linear Regression Model

A multiple linear regression has $p > 1$ explanatory variables,

$$y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \qquad \text{or} \qquad \mathbf{y} = \mathbf{X}\boldsymbol{\beta}. \tag{36}$$

As we saw in the first lesson, this means that

$$\hat{y}_i = \beta_0 + \sum_{j=i}^{p} x_{ij}\beta_j \qquad \text{and} \qquad \hat{\varepsilon}_i = y_i - \beta_0 - \sum_{j=i}^{p} x_{ij}\beta_j. \tag{37}$$

We can fit the model with `lm()` as before, including all the topics in the `marks` data in the regression.

```
> m2 = lm(STAT ~ ALG + ANL + MECH + VECT, data = marks)
> m2

Call:
lm(formula = STAT ~ ALG + ANL + MECH + VECT, data = marks)

Coefficients:
(Intercept)          ALG          ANL         MECH         VECT
  -11.37822      0.72944      0.31293      0.02217      0.02574
```

Model Information from summary(m2)

```
> summary(m2)

Call:
lm(formula = STAT ~ ALG + ANL + MECH + VECT, data = marks)

Residuals:
    Min      1Q  Median      3Q     Max
-21.688  -9.925  -1.905   8.764  36.538

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.37822    6.98174  -1.630 0.106952
ALG           0.72944    0.20961   3.480 0.000802 ***
ANL           0.31293    0.13146   2.380 0.019581 *
MECH          0.02217    0.09895   0.224 0.823265
VECT          0.02574    0.13953   0.184 0.854092
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 12.75 on 83 degrees of freedom
Multiple R-squared:  0.4793, Adjusted R-squared:  0.4542
F-statistic:  19.1 on 4 and 83 DF,  p-value: 3.612e-11
```

Significance of the Regression Coefficients

Now that there are 4 topics as explanatory variables, `summary()` reports theirs Wald test statistics and the corresponding (t_{n-p}) p-values.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.37822    6.98174  -1.630 0.106952
ALG           0.72944    0.20961   3.480 0.000802 ***
ANL           0.31293    0.13146   2.380 0.019581 *
MECH          0.02217    0.09895   0.224 0.823265
VECT          0.02574    0.13953   0.184 0.854092
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The only tests that are significant are those for `ALG` ($p = 0.0008$) and `ANL` ($p = 0.0195$). However, it is important to note that the tests are not independent, because from (17) we know that $\mathrm{VAR}(\hat{\beta})$ is not diagonal.

Regression Coefficients and Multiple Testing Adjustment

Another point to consider is that performing multiple tests requires adjusting the significance threshold; otherwise some null hypotheses will be rejected just by chance due to type I errors. The easiest way of doing that is Bonferroni correction: dividing the threshold by (or multiplying the p-values by) the number of tests.

```
> p.adjust(c(ALG = 0.000802, ANL = 0.019581, MECH = 0.823265,
+   VECT = 0.854092), method = "bonferroni")
     ALG      ANL     MECH     VECT
0.003208 0.078324 1.000000 1.000000
```

After doing that, only `ALG` still has a p-value below 0.05. A state of the art method like FDR is less harsh, and flags `ANL` as borderline significant with a corrected p-value of $\simeq 0.04$.

```
> p.adjust(c(ALG = 0.000802, ANL = 0.019581, MECH = 0.823265,
+   VECT = 0.854092), method = "fdr")
     ALG      ANL     MECH     VECT
0.003208 0.039162 0.854092 0.854092
```

Testing Correlations in a Multiple Regression

In the case of simple linear regression we saw in (29), (30) and (31) that a regression coefficient is equal to zero if and only if the corresponding correlation is zero. This is still true for multiple regression, using partial correlation coefficients conditional on the other regressors: $\beta_i = 0$ if and only if $\rho = \mathrm{COR}(\mathbf{y}, \mathbf{x}_i \mid \text{all } \mathbf{x}_j, j \neq i)$.

- The loglikelihood ratio test

$$D = -n \log(1 - \rho^2) \sim \chi^2_1 \qquad \text{(asymptotic)}; \qquad (38)$$

- the t test for correlation

$$t = \rho \sqrt{\frac{n - p - 1}{1 - \rho^2}} \sim t_{n-p-1} \qquad \text{(exact)}; \qquad (39)$$

- Fisher's Z test

$$Z = \log\left(\frac{1 + \rho}{1 - \rho}\right) \frac{\sqrt{n - p - 2}}{2} \sim N(0, 1) \qquad \text{(asymptotic)}. \qquad (40)$$

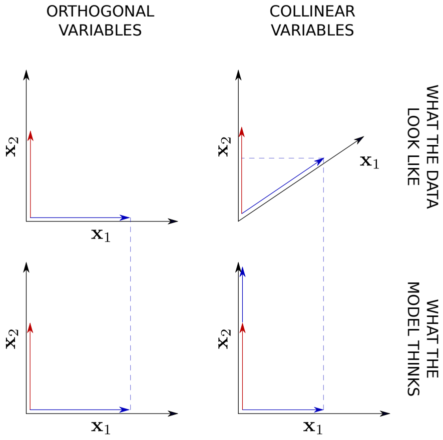
Regression Coefficients and Collinearity

Collinearity (the presence of correlated explanatory variables) can have a large impact on the values and the precision of the regression coefficients.

- In real-world problems, explanatory variables are rarely orthogonal and they are often linked by correlation patterns of varying intensity. Some collinearity is pretty common, but it does not necessarily bias (much) the regression coefficients even when it increases their variance.
- Multiple testing adjustment assumes the tests for the regression coefficients are independent, so it may not be valid. FDR has been proved to hold for weakly and moderately correlated tests, though.

It is important to note that apparent collinearity may be just an artefact due to the presence of outliers; and in that case it can be solved by identifying and removing those outliers.

Collinearity: A Geometric Interpretation



The blue arrow is β_1 (for the data) and $\hat{\beta}_1$ (for the model); the red arrow is β_2 (for the data) and $\hat{\beta}_2$ (for the model).

If $\mathbf{x}_1$ is orthogonal to $\mathbf{x}_2$, β_1 and β_2 are correctly estimated; if they are collinear, one of the regression coefficients ($\hat{\beta}_2$) may be inflated at the expense of the other ($\hat{\beta}_1$). They are both likely to have large $\mathrm{VAR}(\hat{\beta}_1)$ and $\mathrm{VAR}(\hat{\beta}_2)$.

In other words, $\mathbf{x}_1$ and $\mathbf{x}_2$ share some information about $\mathbf{y}$, as evidenced by the orthogonal projection (dashed blue line), and that may all get attributed to one the variables.

Collinearity: from Partial Correlation Coefficients

If $\mathbf{x}_1$ and $\mathbf{x}_2$ are orthogonal, then

$$\hat{\beta}_1 \propto \mathrm{COR}(\mathbf{y}, \mathbf{x}_1 \mid \mathbf{x}_2) = \mathrm{COR}(\mathbf{y}, \mathbf{x}_1) \tag{41}$$

$$\hat{\beta}_2 \propto \mathrm{COR}(\mathbf{y}, \mathbf{x}_2 \mid \mathbf{x}_1) = \mathrm{COR}(\mathbf{y}, \mathbf{x}_2) \tag{42}$$

but if they are not

$$\mathrm{COR}(\mathbf{y}, \mathbf{x}_1 \mid \mathbf{x}_2) = \frac{\mathrm{COR}(\mathbf{y}, \mathbf{x}_1) - \mathrm{COR}(\mathbf{x}_1, \mathbf{x}_2)\,\mathrm{COR}(\mathbf{y}, \mathbf{x}_2)}{\sqrt{1 - \mathrm{COR}(\mathbf{x}_1, \mathbf{x}_2)^2}\sqrt{1 - \mathrm{COR}(\mathbf{y}, \mathbf{x}_2)^2} \tag{43}$$

$$\mathrm{COR}(\mathbf{y}, \mathbf{x}_2 \mid \mathbf{x}_1) = \frac{\mathrm{COR}(\mathbf{y}, \mathbf{x}_2) - \mathrm{COR}(\mathbf{x}_1, \mathbf{x}_2)\,\mathrm{COR}(\mathbf{y}, \mathbf{x}_1)}{\sqrt{1 - \mathrm{COR}(\mathbf{x}_1, \mathbf{x}_2)^2}\sqrt{1 - \mathrm{COR}(\mathbf{y}, \mathbf{x}_1)^2} \tag{44}$$

which means that $\hat{\beta}_1$ and $\hat{\beta}_2$ may be biased upwards or downwards depending on the sign and the magnitude of the correlations.

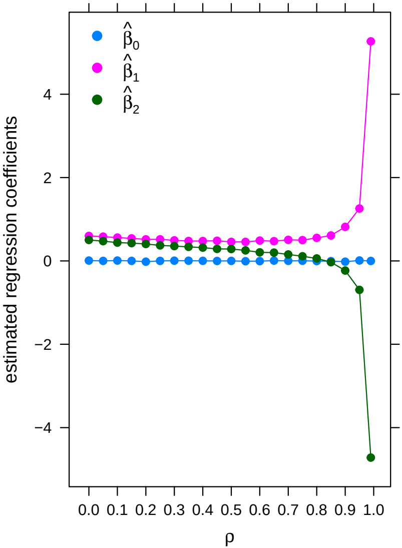
Collinearity: A Simple Simulation

```
> library(MASS)

> # a three-dimensional multivariate Gaussian.
> mu = rep(0, 3)
> R = matrix(c(1,   0.6, 0.5,
+              0.6, 1,   0,
+              0.5, 0,   1),
+       ncol = 3, dimnames = list(c("y", "x1", "x2"), c("y", "x1", "x2")))

> # gradually increase the correlation between the explanatory variables.
> for (rho in c(seq(from = 0, to = 0.95, by = 0.05), 0.99)) {
+
+   # update the correlation matrix and generate the data.
+   R[2, 3] = R[3, 2] = rho
+   data = as.data.frame(mvrnorm(10000, mu, R))
+   # fit the linear model.
+   beta = coef(lm(y ˜ x1 + x2, data = data))
+   # print the coefficients.
+   print(beta)
+
+ }#FOR
```

Collinearity: A Simple Simulation



The intercept is often largely unaffected by collinearity.

When `rho` is zero, $\hat{\beta}_1 = 0.6, \hat{\beta}_2 = 0.5$ because they coincide with the marginal correlations specified in the correlation matrix `R`.

As `rho` increases, both $\hat{\beta}_1$ and $\hat{\beta}_2$ start to drift apart, and they diverge in opposite directions as `rho` approaches 1.

Collinearity: the Variance Inflation Factor

One way to diagnose collinearity is to factorise $\mathbf{X}$ and investigate its eigenvalues or singular values. Another way is to compute the variance inflation factor (VIF) for each variable $\mathbf{x}_i$ as follows:

1. fit the regression

$$\mathbf{x}_i = \gamma_0 + \gamma_1 \mathbf{x}_1 + \ldots + \gamma_{i-1}\mathbf{x}_{i-1} + \gamma_{i+1}\mathbf{x}_{i+1} + \ldots + \gamma_p \mathbf{x}_p + \boldsymbol{\varepsilon}; \quad (45)$$

2. compute $\mathrm{VIF}(\hat{\beta}_i) = 1/(1 - R_i^2)$ where R_i^2 is the R^2 coefficient of the model in (45);
3. if $\mathrm{VIF}(\hat{\beta}_i) > 10$ then collinearity is considered to be high.

For a variable $\mathbf{x}_i$ that is perfectly orthogonal to all other $\mathbf{x}_j$, $R_i^2 = 0$ and then $\mathrm{VIF}(\hat{\beta}_i) = 1$. Otherwise, R_i^2 increases and $\mathrm{VIF}(\hat{\beta}_i) \to \infty$ as $\mathbf{x}_i$ is increasingly linearly related to the other $\mathbf{x}_j$.

Collinearity: VIFs for the Simulation and the `marks` Data

For the marks data, we have:

```
> vif(lm(m2, data = marks))
    MECH     VECT      ALG      ANL
1.602659 1.801461 2.655396 2.038807
```

This means, for example, that $\sqrt{\mathrm{VAR}(\hat{\beta}_{\texttt{VECT}})}$ is inflated by a factor of $\approx \sqrt{1.8} \approx 1.34$, which in turn means that any confidence interval for $\hat{\beta}_{\texttt{VECT}})$ will be 1.34 too wide.

In the simulation,

$\mathrm{VIF}(\hat{\beta}_1) = \mathrm{VIF}(\hat{\beta}_2) \approx 1$ for `rho` equal to 0;

$\mathrm{VIF}(\hat{\beta}_1) = \mathrm{VIF}(\hat{\beta}_2) \approx 1.2$ for `rho` equal to 0.40;

$\mathrm{VIF}(\hat{\beta}_1) = \mathrm{VIF}(\hat{\beta}_2) \approx 2$ for `rho` equal to 0.70;

$\mathrm{VIF}(\hat{\beta}_1) = \mathrm{VIF}(\hat{\beta}_2) \approx 50$ for `rho` equal to 0.99.

Comparing Nested Models

A better alternative to test multiple coefficients at the same time, *i.e.*

$$H_0 : \beta_{i_1} = \ldots = \beta_{i_k} = 0 \qquad \text{vs} \qquad H_1 : \beta_{i_1}, \ldots, \beta_{i_k} \neq 0 \tag{46}$$

is to perform a single overall goodness-of-fit test using the two nested models $\mathcal{M}_0$ (with the coefficients set to zero) and $\mathcal{M}_1$ (with the estimated coefficients). The likelihood ratio test coincides with the general form of the F test we saw in (35):

$$F = \frac{\left(\frac{\mathrm{RSS}(\mathcal{M}_0) - \mathrm{RSS}(\mathcal{M}_1)}{p_{\mathcal{M}_1} - p_{\mathcal{M}_0}}\right)}{\left(\frac{\mathrm{RSS}(\mathcal{M}_1)}{n - p_{\mathcal{M}_1}}\right)} \sim F_{p_{\mathcal{M}_1} - p_{\mathcal{M}_0}, n - p_{\mathcal{M}_1} - 1} \qquad \text{(exact)} \tag{47}$$

where $p_{\mathcal{M}_1}$ and $p_{\mathcal{M}_0}$ are the number of regressors of $\mathcal{M}_1$ and $\mathcal{M}_0$, and $\mathrm{RSS}(\mathcal{M}_1)$ and $\mathrm{RSS}(\mathcal{M}_0)$ are the corresponding residuals' variances.

Special Explanatory Variables: Categorical Variables

If an explanatory variable is a categorical variable (as opposed to a numerical variable), we are primarily interested in the mean response for each possible value (level) of the variable.
Say, for example, that the students assessed in the `marks` data belong to two different classes (say `A` and `B`). Then we would like to estimate

$$y_i = \beta_{0A} + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \qquad \text{for group A} \tag{48}$$

$$y_i = \beta_{0B} + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \qquad \text{for group B} \tag{49}$$

but to keep the original intercept β_0, we choose group `A` as a baseline ($\beta_0 = \beta_{0A}$) and we introduce a dummy variable to express the contrast ($\beta_B = \beta_{0B} - \beta_{0A}$) between group `B` and group `A`:

$$y_i = \beta_0 + \mathbb{1}(i \in \mathtt{B})\beta_{\mathtt{B}} + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \tag{50}$$

where $\mathbb{1}(i \in \mathtt{B})$ is 1 if the ith student belongs to group `B` and 0 if he belongs to group `A`.

Categorical Variables and Contrasts

We can create factor variables with `factor`, and the level for each student is a character string which in this case one of the two levels `A` and `B`.

```
> GROUP = factor(c(rep("A", 44), "B", rep("A", 7), rep("B", 36)))
> levels(GROUP)
[1] "A" "B"
```

The baseline is always the first level; since we have only two, the `contr.treatment()` function creates just one dummy variable: $\mathbb{1}(i \in \mathtt{B})$.

```
> contr.treatment(levels(GROUP))
  B
A 0
B 1
```

If we had three levels `A`, `B` and `C`, we would have both $\mathbb{1}(i \in \mathtt{B})$ and $\mathbb{1}(i \in \mathtt{C})$.

```
> contr.treatment(c("A", "B", "C"))
  B C
A 0 0
B 1 0
C 0 1
```

Categorical Variables in lm()

The `lm()` function recognises whether a variable is categorical and automatically takes care of generating the dummy variables for contrasts.

```
> lm(STAT ~ GROUP + ALG + ANL + MECH + VECT, data = marks)

Call:
lm(formula = STAT ~ GROUP + ALG + ANL + MECH + VECT, data = marks)

Coefficients:
(Intercept)       GROUPB          ALG          ANL         MECH         VECT
  -5.096880    -2.978552     0.705488     0.260411     0.021026    -0.000371
```

Note that this use of the dummy variables is just one of many possible choices; alternatives exist that make interpretation easier in specific settings. They all generate sets of orthogonal dummy variables, and the sum of the dummy variables is the same for all observations.

Categorical Variables in `summary()`

```
> summary(lm(STAT ~ GROUP + ALG + ANL + MECH + VECT, data = marks))

[...]

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.096880  12.327967   -0.41   0.6804
GROUPB      -2.978552   4.809454   -0.62   0.5374
ALG          0.705488   0.213922    3.30   0.0014 **
ANL          0.260411   0.156844    1.66   0.1007
MECH         0.021026   0.099333    0.21   0.8329
VECT        -0.000371   0.146258    0.00   0.9980
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 12.8 on 82 degrees of freedom
Multiple R-squared:  0.482,     Adjusted R-squared:  0.45
F-statistic: 15.2 on 5 and 82 DF,  p-value: 1.42e-10
```

Categorical Variables in `summary()`: Interpretation

- After including `GROUP` in the model, the regression coefficients of the other explanatory variables change. Recall that maximum likelihood estimates are correlated as $\mathrm{VAR}(\hat{\boldsymbol{\beta}}) = \sigma^2_\varepsilon(\mathbf{X}^T\mathbf{X})^{-1}$, so adding or removing terms from the model means all β has to be recomputed.
- Even if the new explanatory variable is perfectly orthogonal to all other variables (which never happens in practice), the coefficients still change because $\hat{\sigma}^2_\varepsilon$ will be smaller.
- The t test statistic for the contrast dummy variables is not indicative of whether the original categorical variable is significant, even when it has only two levels. It is preferable to use a technique called analysis of variance (ANOVA), which we will see later on.

Special Explanatory Variables: Polynomial Terms

Another interesting kind of explanatory variable are polynomial terms.

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \varepsilon_i \tag{51}$$

A linear model is linear in the parameters, not in the explanatory variables, so we may find that the response variable is best explained by including higher powers of some variable(s), say $\mathbf{x}_1$:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i1}^2\beta_{1.2} + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \varepsilon_i \tag{52}$$

Other transforms can be included in principle, but polynomial terms are by far the most common choice along with $\log(\mathbf{x}_1)$.

Polynomial Terms in `lm()`, the Wrong Way

```
> summary(lm(STAT ~ ALG + I(ALG^2) + ANL + MECH + VECT, data = marks))

[...]

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.844781  17.311003   1.377   0.1721
ALG         -0.726050   0.688595  -1.054   0.2948
I(ALG^2)     0.015226   0.006877   2.214   0.0296 *
ANL          0.303275   0.128544   2.359   0.0207 *
MECH         0.029088   0.096751   0.301   0.7644
VECT        -0.015349   0.137618  -0.112   0.9115
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 12.46 on 82 degrees of freedom
Multiple R-squared:  0.5087, Adjusted R-squared:  0.4787
F-statistic: 16.98 on 5 and 82 DF,  p-value: 1.719e-11
```

The `I()` function is used to make it clear to `lm()` that the transformation is applied to the `ALG` variable and is not part of the syntax of the model formula.

Polynomial Terms in `lm()`, the Right Way

The problem with simply including a polynomial terms like `ALG^2` is two-fold:

- terms will be correlated, which is may make the estimation of the regression coefficients problematic; and
- because of that it is difficult to test each term for significance, which is needed to assess the degree of the polynomial.

We see in the output of summary that `ALG` is not significant after adding `ALG^2`, exactly because $\mathrm{COR}(\texttt{ALG}, \texttt{ALG\^{}2}) = 0.98$.
A better way is to use contrasts again, because they are orthogonal; the function that creates them for polynomials is `contr.poly()`.

```
> contr.poly(3)
                .L         .Q
[1,] -7.071068e-01  0.4082483
[2,] -7.850462e-17 -0.8164966
[3,]  7.071068e-01  0.4082483
```

Polynomial Terms in lm() with poly()

```
> summary(lm(STAT ~ poly(ALG, 2) + ANL + MECH + VECT, data = marks))

[...]

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.79282    8.72183   3.187 0.002037 **
poly(ALG, 2)1 75.69314   20.35958   3.718 0.000366 ***
poly(ALG, 2)2 27.86824   12.58757   2.214 0.029609 *
ANL            0.30327    0.12854   2.359 0.020687 *
MECH           0.02909    0.09675   0.301 0.764444
VECT          -0.01535    0.13762  -0.112 0.911465
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 12.46 on 82 degrees of freedom
Multiple R-squared:  0.5087, Adjusted R-squared:  0.4787
F-statistic: 16.98 on 5 and 82 DF,  p-value: 1.719e-11
```

The poly() function calls contr.poly() and automatically sets up the model; p-values are now closer to those in the original model.

Special Explanatory Variables: Interaction Terms

A third type of explanatory variable that deserves special attention are interaction terms. In all the preceding models each regression coefficient was linked to a single explanatory variable; but there may be joint effects are not captured by the model, just marginal (main) effects.

Therefore, we may want to add to the model a regressor whose value is determined by two or more variables in such a way that:

- the interaction does not obscure the main effects of the underlying variables;
- the interpretation of the interactions is clear.

Interaction between Categorical Variables

Interactions between categorical variables are modelled using contrasts in the same way as for individual variables, but applying them to the configurations of the levels. So, if we have students examined in two SESSIONs S and T, an interaction with GROUP has 4 possible configurations:

$$y_i = \beta_{0A\&S} + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \quad \text{for group A in session S} \tag{53}$$

$$y_i = \beta_{0A\&T} + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \quad \text{for group A in session T} \tag{54}$$

$$y_i = \beta_{0B\&S} + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \quad \text{for group B in session S} \tag{55}$$

$$y_i = \beta_{0B\&T} + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \quad \text{for group B in session T} \tag{56}$$

and each configuration has its own intercept. From this set equations, we would like to have one set of contrasts for SESSION, one set for GROUP, and a third set over $\{\mathtt{A}, \mathtt{B}\} \times \{\mathtt{S}, \mathtt{T}\}$ defined to be orthogonal to the main effects.

Interactions and Contrasts

The model with the contrasts is:

$$y_i = \beta_0 + \underbrace{\mathbb{1}(i \in \mathtt{B})\beta_{\mathtt{B}}}_{\text{contrast for } \mathtt{GROUP}} + \underbrace{\mathbb{1}(i \in \mathtt{T})\beta_{\mathtt{T}}}_{\text{contrast for } \mathtt{SESSION}} + \underbrace{\mathbb{1}(i \in \mathtt{B}, i \in \mathtt{T})\beta_{\mathtt{BT}}}_{\text{contrasts for } \mathtt{GROUP} \times \mathtt{SESSION}} + \\ + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p. \qquad (57)$$

In R we code an interaction using "∗" instead of "+"; main effects are automatically included (":" should be used to add just the interaction).

```
> SESSION = factor(c(rep("S", 48), rep("T", 40)))
> m3 = lm(STAT ~ GROUP * SESSION + ALG + ANL + MECH + VECT, data = marks)
> m3

[...]

Coefficients:
    (Intercept)          GROUPB        SESSIONT             ALG
    -1.28725412      1.55863034     -7.58939653      0.69064381
            ANL            MECH            VECT  GROUPB:SESSIONT
     0.23344081      0.00190621      0.00000938      1.14435914
```

Interactions and Contrasts: How Many?

The joint distribution of GROUP and SESSION has 4 parameters, which correspond to the probabilities of the cells of the 2×2 table over $\{\mathtt{A}, \mathtt{B}\} \times \{\mathtt{S}, \mathtt{T}\}$. One of these 4 parameters is fixed because they sum up to one; this leaves 3 free. One is used by the main effect for GROUP, another is used by the main effect for SESSION, and the remaining one is available for the interaction.

	intercept	SESSION	GROUP	interaction
A and S	1	0	0	0
A and T	1	1	0	0
B and S	1	0	1	0
B and T	1	1	1	1

If we lay the components that go into the intercept, we can see how they are orthogonal to each other (and how we cannot have more than one contrast for the interaction).

Interaction between Categorical and Numerical Variables

An interaction term between a categorical and a numerical variable introduces a different regression coefficient for each level for the numerical variable, and a different intercept:

$$y_i = \beta_{0A} + x_{i1}\beta_{1A} + \ldots + x_{ip}\beta_p \qquad \text{for group A} \qquad (58)$$

$$y_i = \beta_{0B} + x_{i1}\beta_{1B} + \ldots + x_{ip}\beta_p \qquad \text{for group B.} \qquad (59)$$

In practice, we factor out the main effect of the two variables,

$$y_i = \beta_0 + \underbrace{\mathbb{1}(i \in \mathtt{B})\beta_{\mathtt{B}}}_{\text{contrast for }\mathtt{GROUP}} + \underbrace{x_{i1}\beta_1}_{\text{main effect for }\mathbf{x}_1} + \\ + \underbrace{x_{i1}\mathbb{1}(i \in \mathtt{B})\beta_{1\mathtt{B}}}_{\text{interaction}} + \ldots + x_{ip}\beta_p. \qquad (60)$$

Interactions, Contrasts and Coefficients

The syntax for fitting this kind of interaction is the same as in the previous case; here we model an interaction between ALG and GROUP.

```
> m4 = lm(STAT ~ GROUP * ALG + ANL + MECH + VECT, data = marks)
> m4

Call:
lm(formula = STAT ~ GROUP * ALG + ANL + MECH + VECT, data = marks)

Coefficients:
(Intercept)       GROUPB          ALG          ANL         MECH         VECT
   -22.6283      32.5563       1.0941       0.2305       0.0153      -0.0462
 GROUPB:ALG
    -0.7446
```

Interaction between Numerical Variables

An interaction between two numerical variables is simply their product, possibly after transforming them; as we can see below for ALG and ANL below, we get the same result by manually including ALG, ANL and I(ALG * ANL) in the model.

```
> lm(STAT ~ ALG * ANL + MECH + VECT, data = marks)

[...]

Coefficients:
(Intercept)          ALG          ANL         MECH         VECT      ALG:ANL
   24.90467     -0.00997     -0.46626      0.01530     -0.03967      0.01661

> lm(STAT ~ ALG + ANL + I(ALG * ANL) + MECH + VECT, data = marks)

[...]

Coefficients:
(Intercept)          ALG         ANL  I(ALG * ANL)         MECH         VECT
   24.90467     -0.00997    -0.46626       0.01661      0.01530     -0.03967
```

Model Selection and the Analysis of Variance

Model Selection: Which Explanatory Variables to Include?

In previous lessons we have estimated and examined models in which the set of explanatory variables was fixed — `ALG` for simple linear regression, and all topics apart from `STAT` (the response) for multiple regression.

This raises two questions:

1. Do we really need all topics as explanatory variables, or will a subset do just as well?
2. How do we pick which topics to include in the model?

Answering these questions is to perform model selection, performing inference on multiple models to pick that which is the most effective for the analysis.

What Does “The Most Effective” Mean?

What model is “most effective” depends on how the model will be used and on the goals of the analysis.

- Is the purpose of the model to provide an explanation of the phenomenon underlying the data, possibly highlighting the role of each explanatory variables?
- Is the purpose of the model to provide a tool to predict the value of the response variable for future observations, using as much information as possible as a black-box device?

In the first case we would like a model that is parsimonious, with a small number of explanatory variables we can easily interpret but fits the data in the sample well; in the second we focus on predictive ability at the expense of clarity and in-sample goodness-of-fit.

A Tradeoff: In-Sample versus Out-of-Sample

There is a tradeoff between fitting the observations in the sample well and providing good predictions for new observations.

On the one hand, including a larger number of explanatory variables produces $\hat{y}_i$ that are very close to the y_i — in the limit case, if we have $p = n - 1$ explanatory variables, all $\hat{y}_i = y_i$ and we have the saturated model. On the other hand, if we include fewer explanatory variables, the model is more likely to generalise because it will be able not pick up spurious patterns from the observed sample.

In both cases, then, we must beware of overfitting; but what counts as overfitting depends on the purpose of the model.

Model Selection: Which Selection Criterion?

- Statistical tests for nested models as in (46), evaluated through their p-values using a threshold; particularly simple if each pair of models differs by just one variable, we can use (38), (39), (40).
- Akaike Information Criterion (AIC), which approximates the expected loglikelihood ratio between the model and the unknown "true model" for the data:
$$\mathrm{AIC} = -2 \log L(\boldsymbol{\beta}, \sigma^2_{\varepsilon}; \mathbf{y}, \mathbf{X}) + 2K \tag{61}$$
where $L(\boldsymbol{\beta}, \sigma^2_{\varepsilon}; \mathbf{y}, \mathbf{X}) = n \log \left(\sum_{i=1}^{n} \hat{\varepsilon}^2_i / n\right)$ is the likelihood from (10) at its maximum and K is $p + 1$, the number of coefficients (including the intercept).
- the Bayesian Information Criterion (BIC), which approximates the (negated) posterior probability of the model as
$$\mathrm{BIC} = -2 \log L(\boldsymbol{\beta}, \sigma^2_{\varepsilon}; \mathbf{y}, \mathbf{X}) + \log(n) K. \tag{62}$$

Both AIC and BIC should be minimised to find the optimal model.

Comparing Selection Criteria: Pros and Cons

- Model selection through statistical tests is problematic because of the large number of tests typically required (requiring multiple testing correction). However, the resulting p-values are correlated and therefore multiple testing correction is also problematic.
- AIC focuses on getting close to the "true model" for the current data, and favours good in-sample fitted values over good predictions; BIC on the other hand provides better predictive model at the expense of poorer fitted values.
- Both AIC and BIC have penalties which increase with p and prevent model selection from overfitting; using just $L(\boldsymbol{\beta}, \sigma^2_\varepsilon; \mathbf{y}, \mathbf{X})$ always selects the saturated models (or all available explanatory variables).
- AIC can only be used to compare models fitted on the same sample; and the magnitude of both AIC and BIC depends on n.

Model Selection: Which Selection Strategy?

A large number of selection strategies have been proposed in the literature for these criteria; the simplest is a heuristic called stepwise selection. It is typically implemented as follows:

1. Take the empty regression model (*i.e.* with just the intercept) as the starting point $\mathcal{M}$ of the selection process.
2. Until no improvement is possible:
    2.1 Forward Selection: try to add one explanatory variable to $\mathcal{M}$, re-fit the model and estimate the selection criteria for the new model $\mathcal{M}^*$.
    2.2 Backward Selection: try to remove one explanatory variable from $\mathcal{M}$, re-fit the model and estimate the selection criteria for the new model $\mathcal{M}^*$.
    2.3 If no model $\mathcal{M}^*$ is better than $\mathcal{M}$, return $\mathcal{M}$.
    2.4 If any model $\mathcal{M}^*$ is better than $\mathcal{M}$, return the model $\mathcal{M}^*$ that has the best value for the selection criterion (lowest p-value or AIC/BIC).

Note that, being an heuristic, there is no guarantee that stepwise selection will return a globally optimal model; it can get stuck in a local optimum.

Stepwise Selection with AIC

```
> m = lm(STAT ~ 1, data = marks)
> step(m, trace = TRUE, scope = ~ ANL + ALG + VECT + MECH)
Start:  AIC=502.27
STAT ~ 1

       Df Sum of Sq   RSS    AIC
+ ALG   1   11446.6 14458 452.95
+ ANL   1    9550.0 16355 463.79
+ VECT  1    4934.5 20970 485.67
+ MECH  1    3921.9 21983 489.82
<none>              25905 502.27

Step:  AIC=452.95
STAT ~ ALG

       Df Sum of Sq   RSS    AIC
+ ANL   1     949.7 13508 448.97
<none>              14458 452.95
+ VECT  1      40.1 14418 454.70
+ MECH  1      24.3 14434 454.80
- ALG   1   11446.6 25905 502.27
```

```
Step:  AIC=448.97
STAT ~ ALG + ANL

       Df Sum of Sq   RSS    AIC
<none>              13508 448.97
+ MECH  1     14.80 13494 450.87
+ VECT  1     12.18 13496 450.89
- ANL   1    949.67 14458 452.95
- ALG   1   2846.23 16355 463.79

Call:
lm(formula = STAT ~ ALG + ANL, data = marks)

Coefficients:
(Intercept)          ALG          ANL
   -11.1920       0.7653       0.3164
```

Model Selection for Prediction

When we are interested in prediction, we are more interested in the prediction error (*e.g.* the residuals for observations not used in estimating the model) than in σ^2_ε (*e.g.* the residuals for the observations used in fitting the model).

The standard way of estimating the former is k-fold cross-validation:

1. Split the data into k random subsets (the folds) of size n/k, or as close as possible.
2. For each subset in turn:
   2.1 Fit the model using the other $k-1$ subsets.
   2.2 Predict the y_i for the observations in the kth subset from the fitted model, and save the $(\tilde{y}_i, y_i)$ pairs.
3. Estimate the prediction error as $\hat{\sigma}^2_p = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$.

In general $\sigma^2_p \gg \sigma^2_\varepsilon$, and we can use $\tilde{\varepsilon}_i = y_i - \tilde{y}_i$ to estimate AIC and BIC.

Cross-Validation and Linear Models

Cross-validation requires the `cv.glm()` function, which takes models fitted by `glm()` as generalised linear models. Models fitted by `lm()` are just a particular case of `glm` models.

```
> library(boot)
> m0 = glm(STAT ~ ALG, data = marks)
> m1 = glm(STAT ~ ALG + ANL, data = marks)
> m2 = glm(STAT ~ ALG + ANL + MECH + VECT, data = marks)
> cv.glm(marks, m0)$delta[2]
[1] 172.0457
> cv.glm(marks, m1)$delta[2]
[1] 164.8933
> cv.glm(marks, m2)$delta[2]
[1] 171.3003
```

As we can see above, the model including only `ALG` has a higher prediction error than that including `ALG` and `ANL`; so the latter is preferable for prediction. The model containing all the topics overfits the data and has a higher prediction error than the previous one.

What do Predictions Look Like: Confidence Intervals

In the case if a y_i and its $\mathbf{x}$ are observed, then similarly to (28) we have that

$$\mathrm{VAR}(\hat{y}_i) = \mathrm{VAR}(\mathbf{x}\hat{\boldsymbol{\beta}}) = \mathbf{x}^T\,\mathrm{VAR}(\hat{\boldsymbol{\beta}})\mathbf{x} = \sigma^2_\varepsilon(\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}) = \sigma^2_\varepsilon h_{ii}. \quad (63)$$

On the other hand, for a new observation we have that the variance of the prediction $\tilde{y}_i$ is greater because it also depends on the unobserved error $\tilde{\varepsilon}_i$:

$$\begin{aligned}\mathrm{VAR}(\tilde{y}_i) = \mathrm{VAR}(\tilde{\mathbf{x}}^T\hat{\boldsymbol{\beta}} + \tilde{\epsilon}_i) = \\ \tilde{\mathbf{x}}^T\,\mathrm{VAR}(\hat{\boldsymbol{\beta}})\tilde{\mathbf{x}} + \mathrm{VAR}(\tilde{\epsilon}_i) = \sigma^2_\varepsilon(1 + \tilde{\mathbf{x}}^T(\mathbf{X}^T\mathbf{X})^{-1}\tilde{\mathbf{x}}) \quad (64)\end{aligned}$$

The former can be used to construct a confidence interval, the latter a prediction interval:

$$\frac{\hat{y}_i - \mathbf{x}\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2_\varepsilon h_{ii}}} \sim t_{n-p-1} \qquad \text{and} \qquad \frac{\tilde{y}_i - \mathbf{x}\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2_\varepsilon(1 + \tilde{\mathbf{x}}^T(\mathbf{X}^T\mathbf{X})^{-1}\tilde{\mathbf{x}})}} \sim t_{n-p-1}. \quad (65)$$

Predictions, Intervals and `predict()`

The `predict()` function computes $\hat{y}_i$ for observed data points (same as `fitted`) and $\tilde{y}_i$ for new data points.

```
> m = lm(STAT ~ ALG + ANL, data = marks)
> predict(m, newdata = marks[1, ])
      1
61.2856
> predict(m, newdata = data.frame(ALG = 100, ANL = 85))
       1
92.23745
```

I has an `interval` argument which can compute both confidence and prediction intervals.

```
> predict(m, newdata = data.frame(ALG = 100, ANL = 85),
+   interval = "confidence")
       fit      lwr      upr
1 92.23745 79.16548 105.3094
> predict(m, newdata = data.frame(ALG = 100, ANL = 85),
+   interval = "prediction")
       fit      lwr      upr
1 92.23745 63.96853 120.5064
```

The Analysis of Variance

The analysis of variance (ANOVA) is a technique that analyses the contribution of each explanatory variable to the model by iteratively decomposing the sum of squared residuals into orthogonal (*i.e.* independent) components that can be tested separately using F tests.

Consider again (33):

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{total squares}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{residual squares}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{regression squares}} .$$

If we have more than one explanatory variables we can further decompose the last term into smaller sum-of-square contributions.

ANOVA: With One Categorical Variable

The simplest example of such a decomposition is for a model with a single categorical explanatory variable, and is known as one-way ANOVA.

Suppose we have two groups. If we denote the intercept for the ith observation as $\bar{y}_i$ (was $\beta_{0\mathtt{A}}$ or $\beta_{0\mathtt{B}}$ in (48) and (49) when we introduced contrasts), we can write

$$\underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{regression squares}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i + \bar{y}_i - \bar{y})^2 =$$

$$= \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}_{\text{within group variability}} + \underbrace{\sum_{i=1}^{n}(\bar{y}_i - \bar{y})^2}_{\text{between group variability}} \qquad (66)$$

and identify the component that pertains to the categorical variable.

ANOVA Decompositions with anova()

This procedure can be generalised to any kind of variables, including interactions, and is implemented in the anova() function.

```
> m = lm(STAT ~ GROUP + ALG + ANL, data = marks)
> anova(m)
Analysis of Variance Table

Response: STAT
          Df  Sum Sq Mean Sq F value    Pr(>F)
GROUP      1  8397.8  8397.8 52.5120 1.909e-10 ***
ALG        1  3622.4  3622.4 22.6510 8.000e-06 ***
ANL        1   451.2   451.2  2.8214   0.09673 .
Residuals 84 13433.4   159.9
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Note that the variables are decomposed and tested in the order they are specified in the model; different orderings can give different results if the variables are not perfectly orthogonal.

AIC, BIC and ANOVA

The difference between the AIC for two models $\mathcal{M}_0$ and $\mathcal{M}_1$ can be written as a loglikelihood ratio test

$$\begin{aligned} \mathrm{AIC}(\mathcal{M}_1) - \mathrm{AIC}(\mathcal{M}_0) &< 0 \\ -2\log L(\mathcal{M}_1) + 2K_{\mathcal{M}_1} + 2\log L(\mathcal{M}_0) - 2K_{\mathcal{M}_0} &< 0 \\ -2\log\frac{L(\mathcal{M}_1)}{L(\mathcal{M}_0)} &< -2(K_{\mathcal{M}_1} - K_{\mathcal{M}_0}) \end{aligned} \tag{67}$$

where the left hand side has distribution $\chi^2_{K_{\mathcal{M}_1} - K_{\mathcal{M}_0}}$. $\mathcal{M}_1$ is preferred over $\mathcal{M}_0$ only if the inequality holds; and provides an alternative threshold for significance. Equivalently, we can define similar relationships using BIC or the F test for ANOVA, since it is a likelihood ratio test.

Experimental Design

Data Collection: Observational vs Experimental

The data we use to fit linear regression (and other) models can be collected in different ways, which have a strong influence on what kind of conclusions we can draw from the models. Some approaches are:

- secondary analysis of already existing data;
- cross-sectional study, including descriptive sample survey;
- prospective or retrospective observational study;
- a designed experiment;
- a designed experiment with interventions.

There is quite a sharp distinction between an observational study and an experiment. In the latter the investigator has virtual control over the whole system, whereas in the former the investigator's choices are limited to deciding what to observe and to measure. As a result the conclusions that can be drawn from an observational study are much more limited that those we can draw from a designed experiment. In particular we cannot usually distinguish association from cause-effect relationships from observational data.

Different Fields, Different Terminology

- Applications to Medicine
  - Clinical Trials
  - Case-Control Studies
  - Cohort Studies
  - Epidemiological Studies
- Applications to Social Sciences
  - Opinion Polls
  - Surveys
  - Censuses
  - Questionnaires
- Applications to Industry
  - Taguchi Methods
  - Response-Surface Designs
  - Screening Designs

Key Steps in Data Collection

1. State the objective(s): which questions are to be answered?
2. Determine the "scope of inference": what is the reference population, how general we would like the conclusions to be?
3. Understand the system under study.
4. Decide how to measure the response(s).
5. Consider factor which could affect the response:
   - design factors to be varied as treatments, or to be kept fixed.
   - confounding factors to be controlled by design or by randomisation.
6. Plan the conduct of the study (time line).
7. Plan an (outline of) the statistical analysis: the alternative is to try one analysis after the other until one gives a statistically significant result. And then nobody can replicate your results (47 out of 53 landmark publications in cancer research!)
8. Determine the sample size: to have sufficient tests with enough power, so that we can tell null and alternative hypotheses apart.

Avoiding Systematic Bias

The conclusions we can draw from a designed experiment are affected by random errors (from the random noise present in the data) and systematic errors which introduce bias and possibly confounding.
Two ways to address and avoid the latter are:

- randomisation: some key variables in the design (typically treatment) are assigned at random to subjects to make the effect of those variables independent from the other variables in the model. Otherwise there may be patterns in how those key variables are assigned that make their effects confounded with the effects of other variables.
- retrospective adjustments: if sources of systematic errors are identified in data that have been previously collected, we can adjust our analysis to take those sources of errors into account. This typically done either by re-weighting observations or by adding interaction terms.

A common source of systematic errors is sampling or selection bias, when some members of the reference population are less likely to be included than others. Thus the resulting sample is biased because it is not representative of the reference population.

Block Designs to Reduce Haphazard Variation

In order to reduce the impact of random errors in estimating the effect of the key variables in the experiment, we would like to "compare like with like" as we do (for instance) with paired data.

A way to do that is to group similar observations into blocks that are relatively homogeneous for some other variables measured in the experiment. Thus we reduce known sources of variation that are not of interest to better focus on those that are of interest.

Sampling from the Reference Population

Depending on the goals of the experiments and the structure of the reference populations, we can choose to sample our observations in different ways.

- Simple random sampling: subjects are sampled at random with equal probability, without replacement (in small populations) or with replacement (in large population, where the probability of sampling the same subject twice is negligible). Unbiased but with high variance.
- Systematic sampling: subjects are sampled with a regular pattern (*e.g.* every kth widget coming out of a factory). If the order in which the subjects are considered is random, it is still a form of random sampling. Problematic with periodic data.
- Stratified sampling: subjects are sampled separately from disjoint subsets (the strata) of the reference population. Strata, like blocks, are chosen to minimise variability within each stratum and to maximise variability between strata. The sample size is chosen separately for each stratum.

Other possibilities include cluster sampling (*e.g.* sampling students as whole classes) and panel sampling (for longitudinal studies).

One-Way ANOVA

The one-way or single-factor ANOVA is the simplest experimental design: it includes a continuous response variable and a single (discrete) experimental factor. The latter is usually denoted as the treatment and is assumed to be randomised. The underlying model is a simple regression of the form

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2); i = 1, \ldots, k; j = 1, \ldots, r_k; \tag{68}$$

where μ is the overall mean and τ_i is the ith treatment effect. As we have seen before this model splits the sum of squares residual into between-treatments (the variability explained by the treatment) and within-treatment (residual variability).

In the general case, $\varepsilon_{ij} \sim N(0, \sigma_i^2)$ and the model above becomes a random-effect or hierarchical linear model.

A Plasma Etching Experiment

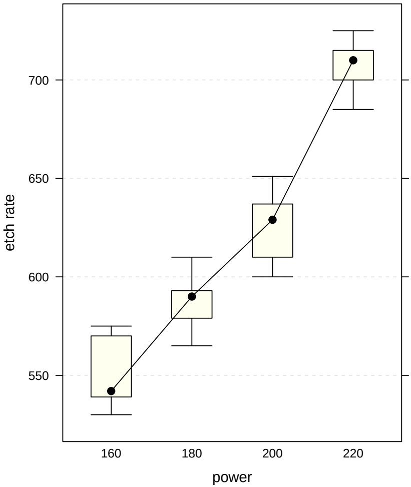
This experiment was set up by an engineer in an integrated circuits manufacturing plant to study the relationship between the power settings and the etch rate of the plasma machinery used to print circuits on silicon.

The experiment has just a single factor (the 4 power settings) and 5 replicates (for each power setting). Therefore, the design is balanced. The treatment is randomised, because measurements are assigned a random numeric ID and then sorted, so that the temperature is in fact randomly assigned each time.

```
> etch = read.table("plasma.etching.txt", header = TRUE)
> head(etch)
  random power rate
1  12417   200  600
2  18369   220  725
3  21238   220  700
4  24621   160  575
5  29337   160  542
6  32318   180  565
```

A simple `anova(lm(rate ~ power))` tells us the effect of power is highly significant with a p-value of $7.263 \cdot 10^{-10}$.

Etch Rate vs Power



Power			
160	180	200	220
575	565	600	725
542	593	651	700
530	590	610	715
539	579	637	685
570	610	629	710

Average			
551.2	587.4	625.4	707.0

Determining Sample Size

How do we know if we have enough replicates to have confidence in our ability of detecting a treatment effect? The power of the corresponding F test is

$$\begin{aligned} \beta &= \mathrm{P}(\text{reject } H_0 \mid H_0 \text{ is false}) \\ &= \mathrm{P}(\hat{F} > F_\alpha \mid H_0 \text{ is false}) \end{aligned} \tag{69}$$

where F_α is the $1 - \alpha$ quantile of a non-central F distribution with degrees of freedom $k, n - k$ and non-centrality parameter δ. Thus, for a fixed α and k we can compute β as a function of the overall sample size n. We can then set a minimum power level b we deem appropriate and find the lowest n for which $\beta(n) \geqslant b$; this is called a power calculation.

For most experimental designs this is done numerically, and subjects are then usually allocated in equal proportions to treatments for balance.

Two-Way ANOVA

Most experiments involve more than one factor of interest. Budget permitting, the most effective way of setting up such a trial is a (crossed) factorial design in which all configurations of the factors are present. The simplest of such models is the two-way ANOVA model

$$y_{ijk} = \mu + \tau_i + \upsilon_j + (\tau\upsilon)_{ij} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim N(0, \sigma^2). \tag{70}$$

Ideally, $k > 1$ replicates are available for each configuration to measure interaction terms (for $k = 1$ the model is saturated).

Note that:

- the model includes an interaction term because we are interested in both factors, and thus their interaction is also of interest;
- the treatment is randomised on both factors simultaneously;
- and that the factors are orthogonal by construction because the design includes all the configurations of their values.

Battery Life at Extreme Temperatures

An engineer decides to test batteries with 3 plate materials at 3 temperature levels that are consistent with the batteries' end-use environment. 4 batteries are tested at each combination of plate material and temperature, and all 36 tests are run in random order. Key questions are:

- What effects do material type and temperature have on the life of the battery?
- Is there a choice of material that would give uniformly long life regardless of temperature?

```
> battery = read.table("battery.life.txt", header = TRUE)
> battery$temperature =
+   factor(battery$temperature, levels = c("15F", "70F", "125F"))
> head(battery)
  material temperature life
1       M1         70F   75
2       M3        125F   60
3       M1         70F   34
4       M2         70F  122
5       M1        125F   58
6       M3         15F  160
```

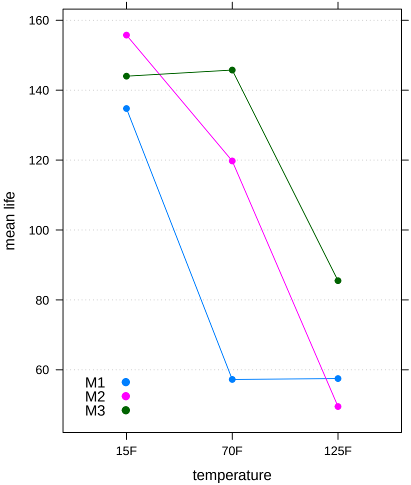
Material and Temperature

Studying the variance components tells the engineers that both material and temperature have significant effects, as does their interaction.

```
> m = lm(life ~ material * temperature, data = battery)
> anova(m)
Analysis of Variance Table

Response: life
                     Df Sum Sq Mean Sq F value    Pr(>F)
material              2  10684  5341.9  7.9114  0.001976 **
temperature           2  39119 19559.4 28.9677 1.909e-07 ***
material:temperature  4   9614  2403.4  3.5595  0.018611 *
Residuals            27  18231   675.2
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Which Material is Best Overall?



No material is always better than the others, as the lines for the materials overlap; but the all decrease with temperature. Batteries have longer life at lower temperatures.

The lines in the plot are not parallel because the interaction term is significant, so batteries built with different materials react differently at various temperatures.

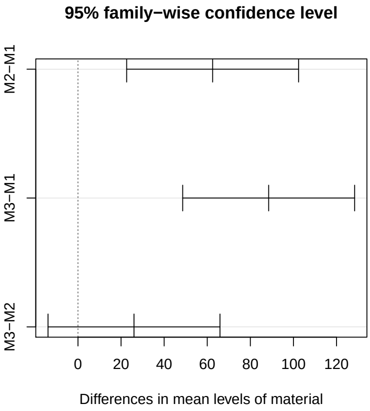
Are Two Material Equivalent at a Particular Temperature?

Are related question, which is often important in experimental design, is: are two treatments equivalent? In this example, the engineer may ask: which materials have equivalent mean battery lives at 70°F? Tukey's test performs a single-step multiple comparison procedure on all pairs of treatments to determine that.

```
> m70F = lm(life ˜ material, data =
+   battery[battery$temperature == "70F", ])
> TukeyHSD(aov(m70F), "material")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = m70F)

$material
      diff        lwr       upr     p adj
M2-M1 62.5  22.59911 102.40089 0.0045670
M3-M1 88.5  48.59911 128.40089 0.0004209
M3-M2 26.0 -13.90089  65.90089 0.2177840
```



(Balanced Complete) Block Designs

Previous experimental design included only variables of interest; but in many cases there are nuisance factors whose influence we want to remove. In other words, we are not interested in their influence on the response but we account for them in the model to remove their effects on treatment comparisons.

A complete randomised block design does this by having one block for each configuration of the nuisance variable, and each block contains all the variables of interest (treatments). The latter are randomised within each block.

In the simple case with a single treatment and a single nuisance variable, the underlying model is

$$y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim N(0, \sigma^2). \tag{71}$$

where the τ_i are the treatment effects and the β_j are the block factors. Note that there are no interaction terms involving both treatment and block factors, although there may be interactions between treatments. (Blocks are defined by the configurations of the nuisance variables, so they always have interaction terms.)

Quality Control for Medical Devices

A medical device manufacturer produces vascular grafts combining resin, lubricant and very thin tubes with some pressure machinery. He suspects that the amount of pressure may have an effect on production yield (in % of grafts without any defect). However, variability between different batches due to the inconsistent quality of the raw materials is also present and may be wrongly attributed to changes in pressure. Therefore, he tries 4 different pressure settings randomised within each of 6 batches of resin (the blocks).

```
> graft = read.table("medical.devices.txt", header = TRUE)
> head(graft)
  batch pressure yield
1     I  8500PSI  90.3
2     I  9100PSI  82.5
3     I  8900PSI  85.5
4     I  8700PSI  92.5
5    II  9100PSI  89.5
6    II  8900PSI  90.8
```

Note that there are no replicates but the model is not singular because it does not include any interaction term.

So, Is Low Yield the Resin's Fault?

From the p-value, we conclude that pressure settings affects mean yield ($p \approx 0.002$); and that there is a significant variability in different batches of resin ($p \approx 0.006$).

```
> anova(lm(yield ~ batch + pressure, data = graft))

Response: yield
          Df Sum Sq Mean Sq F value   Pr(>F)
batch      5 192.25  38.450  5.2487 0.005532 **
pressure   3 178.17  59.390  8.1071 0.001916 **
Residuals 15 109.89   7.326
```

Had we not accounted for the resin batches in the blocks, the estimate of the effect of pressure would be strongly biased ($0.023/0.002 = 12.24$).

```
> anova(lm(yield ~ pressure, data = graft))

Response: yield
          Df Sum Sq Mean Sq F value  Pr(>F)
pressure   3 178.17  59.390  3.9313 0.02345 *
Residuals 20 302.14  15.107
```

Which Pressure Differences are Significant?

This experimental design does not imply any interaction term, so we can just compute Tukey's test on the treatment (or in general, on the configurations of the treatment variables).

```
> TukeyHSD(aov(lm(yield ~ batch + pressure, data = graft)), "pressure")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm(yield ~ batch + pressure, data = graft))

$pressure
                     diff        lwr       upr     p adj
8700PSI-8500PSI -1.133333  -5.637161  3.370495 0.8854831
8900PSI-8500PSI -3.900000  -8.403828  0.603828 0.1013084
9100PSI-8500PSI -7.050000 -11.553828 -2.546172 0.0020883
8900PSI-8700PSI -2.766667  -7.270495  1.737161 0.3245644
9100PSI-8700PSI -5.916667 -10.420495 -1.412839 0.0086667
9100PSI-8900PSI -3.150000  -7.653828  1.353828 0.2257674
```

Differences of 200PSI are never significant. Moving from 8500 or 8700PSI to 9100PSI produces significant changes in the mean treatment effect; lower temperatures appear to produce higher yield.

(Balanced) Incomplete Block Designs

Sometimes we cannot have a complete design because we do not have enough experimental units to allocate all combinations of treatments and the block variables. We may still be able to set up a balanced incomplete block design in which any two treatments appear together the same number of times; there are tables in classic textbooks that detail such an allocation strategy for common experimental setups. For instance:

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
y_{1T_1}	y_{2T_1}	—	y_{4T_1}
—	y_{2T_2}	y_{3T_2}	y_{4T_2}
y_{1T_3}	y_{2T_3}	y_{3T_3}	—
y_{1T_4}	—	y_{3T_4}	y_{4T_4}

Treatments are still randomised inside each block, and in the design above each treatment appears three times with a single treatment missing from each block. Note that there are still enough observations (12) compared to the degrees of freedom of the model (7) even without replicates.

Conditions on Incomplete Block Designs

The combinations of numbers of treatments (v), numbers of replicates (r), numbers of blocks (b), and numbers of observations per block (k) are limited (see Cox & Reid's Experimental Design); sometimes it is not possible to achieve perfect balance. Intuitively, $n = rv = bk$ so fixing r, v, k we can determine b, or fixing b we can determine k. If the solution is an integer, a balanced incomplete block design is possible.

In the medical device example, with 3 observations per block we get:

	I	II	III	IV	V	VI	TOTAL
8500PSI	x		x	x	x		4
8700PSI	x	x		x	x	x	5
8900PSI	x	x	x		x	x	5
9100PSI		x	x	x		x	4
TOTAL	3	3	3	3	3	3	18

Medical Devices, Incomplete Blocks

If we apply the (not quite balanced) incomplete block design to the medical devices example, we see that both F tests are not as significant as before because of the smaller sample size (which means lower power). The batch effect is now clearly not significant, but the pressure effect is still significant.

```
> incomplete = graft[-c(2, 8, 12, 16, 20, 24), ]
> anova(lm(yield ~ batch + pressure, data = incomplete))
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value  Pr(>F)
batch      5 63.418  12.684  1.9120 0.18769
pressure   3 95.543  31.848  4.8008 0.02903 *
Residuals  9 59.704   6.634
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

However, the coefficients of the models from the complete and incomplete designs are quite similar. This suggests that the incomplete design is unbiased like the complete design.

Factorial Designs

A factorial design an experimental design that includes all possible combinations of the levels of the factors that are investigated. It includes as particular cases the one- and two-way ANOVA and 2^k designs (with k factors each with 2 levels). The underlying mathematical model is

$$\begin{aligned} y = \underbrace{\mu}_{\text{population mean}} + \underbrace{\tau_i + \upsilon_j + \dots}_{\text{experimental factors}} + \underbrace{(\tau\upsilon)_{ij} + \dots}_{\text{two-way interactions}} + \\ + \underbrace{(\beta\delta)_{kl} + \dots}_{\text{blocking factors}} + \varepsilon \quad \text{with } \varepsilon \sim N(0, \sigma^2) \qquad (72) \end{aligned}$$

for each observation, optionally with blocks. As before, configurations of the experimental factors are randomised (again possibly within each block). Two replicates are needed for each of them to be able to estimate interaction terms.

Detecting Targets on a Radar

An engineer is studying methods for improving the ability to detect targets on a radar scope with two different filters under the influence of background noise or ground clutter. To do that he designs a factorial experiment with 3 ground clutter levels (and the 2 filters); the response is the intensity of the target, which is increased until the operator observes it and then recorded. To account for experience levels of different operators, 4 operators are randomly selected as used as blocks. Once an operator is chosen, the order in which the six treatment combinations are run is randomised.

```
> radar = read.table("radar.txt", header = TRUE)
> head(radar)
  operator clutter filter intensity
1        I     LOW      A        90
2        I  MEDIUM      A       102
3        I    HIGH      A       114
4        I     LOW      B        86
5        I  MEDIUM      B        87
6        I    HIGH      B        93
```

Operators Have Very Different Skill Levels

The main effects of both experimental factors are significant, but their interaction is not. Note that there is no interaction between block and experimental factors.

```
> anova(lm(intensity ~ clutter * filter + operator, data = radar))
Analysis of Variance Table

Response: intensity
               Df  Sum Sq Mean Sq F value    Pr(>F)    
clutter         2  335.58  167.79 15.1315 0.0002527 ***
filter          1 1066.67 1066.67 96.1924 6.447e-08 ***
operator        3  402.17  134.06 12.0892 0.0002771 ***
clutter:filter  2   77.08   38.54  3.4757 0.0575066 .  
Residuals      15  166.33   11.09                      
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Thus, we conclude that both clutter level and the filter affect the operator's ability to detect targets, but their interaction is not supported by the data. There is also substantial variability in the performance of different operators.

Differences in Treatment Means in Factorial Designs

Since we have two experimental factors, the comparison of means that makes the most sense is that on their configurations.

```
> TukeyHSD(aov(lm(intensity ~ clutter * filter + operator,
+   data = radar)), "clutter:filter")
```

However, in this example the interaction is not significant and therefore it makes sense to compare (marginal) differences for the main effects.

```
> TukeyHSD(aov(lm(intensity ~ clutter * filter + operator,
+   data = radar)), "clutter")
> TukeyHSD(aov(lm(intensity ~ clutter * filter + operator,
+   data = radar)), "filter")
```

Most differences in the command above appear to be significant. Note that Tukey's test is an omnibus test, so multiplicity adjustment is not needed even when it returns a large number of p-values.

A Note About Confounding

A variable is a confounder if

1. it is not a block or experimental factors; and
2. correlates with both the response and a factor included in the experimental design.

Factors included in an experimental design are orthogonal by construction, and the interactions with block factors are eliminated through randomisation. In observational data, confounding may happen even with recorded explanatory variables if they are collinear or tend to vary with common patterns.

There is no general solution for confounding in experimental data, because it is difficult to ascertain the presence of a confounder and which factors it is related to. In observational data, when both the confounder and the confounded variables are available, the effect of the former can be separated with an instrumental variable model (also known as two-stage least squares model), which is a basic form of hierarchical linear model.

Best Practices in Designing Experiments

- Vary multiple (but not too many) design factors at once.
- Stratify the experimental units into blocks which are relatively homogeneous.
- Have enough replication.
- Strive for balance.
- Randomise all arbitrary choices.
- Blind the experiment so that neither the experimenter nor the subjects know which subjects received which treatment.

Robust and Advanced Regression Methods

Violating the Assumptions of Linear Models

As we have seen at the beginning of the course, linear models have one set of assumptions from ordinary least squares:

1. explanatory variables are not collinear,
2. residuals have mean zero ($\mathrm{E}(\varepsilon_i) = 0$),
3. residuals are homoscedastic ($\mathrm{VAR}(\varepsilon_i) = \sigma^2_\varepsilon$),
4. residuals are uncorrelated ($\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0$);

and in maximum likelihood estimation we assume that residuals are normally distributed, so that in the end $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$.

It may (often) happen that when trying to validate the model we find out that one or more of these assumptions are violated. In that case, we need to switch to more advanced forms of linear regression with weaker or more robust assumptions.

Heteroschedasticity: Weighted Least Squares

A linear model with heteroschedastic residuals can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \Sigma) \text{ with } \Sigma = \operatorname{diag}(\sigma^2_{\varepsilon_1}, \ldots, \sigma^2_{\varepsilon_n}); \tag{73}$$

or equivalently using weights so that $\sigma^2_{\varepsilon_i} = w_i \sigma^2_{\varepsilon}$. In that case model estimation minimises the weighted least squares

$$\operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \operatorname*{argmin}_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{74}$$

where $\mathbf{W} = \operatorname{diag}(w_1, \ldots, w_n)$, and estimates the regression coefficients as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}. \tag{75}$$

The weights can be estimated using a very general technique called iteratively re-weighted least squares (IRLS).

Heteroschedasticity: Linear Mixed Models

Another common way for residuals to be heteroschedastic is to depend on a categorical variable, say

$$y_i = \beta_{0\mathtt{A}} + x_{i1}\beta_1 + \ldots x_{ip}\beta_p \quad \text{with } \varepsilon_i \sim N(0, \sigma^2_{\mathtt{A}}) \quad \text{if } i \in \mathtt{A}, \tag{76}$$

$$y_i = \beta_{0\mathtt{B}} + x_{i1}\beta_1 + \ldots x_{ip}\beta_p \quad \text{with } \varepsilon_i \sim N(0, \sigma^2_{\mathtt{B}}) \quad \text{if } i \in \mathtt{B}. \tag{77}$$

Such models are called in many ways in the literature, such as mixed effects models, multilevel models or hierarchical models. They can in principle be estimated using weighted least squares, but there are specific techniques that allow a better interpretation of the results.

Robust Regression: Least Absolute Deviations

If the data include outliers, a more robust function to minimise is the sum of the absolute values of the residuals,

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - \beta_0 - x_{i1}\beta_1 - \ldots - x_{ip}\beta_p|, \tag{78}$$

instead of the squared residuals. So, this is called an L_1 regression as opposed to an L_2 regression such as ordinary least squares. It is known as least absolute deviations (LAD) regression.

While it has the advantage of being robust to outliers, small changes in the data can lead to very different estimates (because L_1 is not smooth); it may have multiple optimal solutions (because L_1 is convex but not strictly convex); and it lacks most of closed-form results available for least squares.

Generalised Linear Models

Assuming normally-distributed errors constrains the kinds of response variables we can model without transforming them. A general class of models that tackles this problem is generalised linear models (GLM), which assume the response has a distribution for the exponential family and regresses its expected value through a link function:

$$g(\mathrm{E}(y_i)) = \eta_i \qquad \text{where} \qquad \eta_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p. \tag{79}$$

For instance, setting $g(\cdot)$ to the identity function leads back to a classic linear regression; the logit function $g(\pi) = \log\{\pi/(1-\pi)\}$ allows to model binomial responses with success probability π.

Model estimation is done by maximising the likelihood of the model after substituting the regressors in the density through $g^{-1}(\eta_i)$.

Generalised Linear Models: Binomial Response

For a binary response, the natural assumption is the Binomial distribution. So

$$\mathrm{E}(y_i) = \pi_i \qquad \text{and} \qquad g(\pi_i) = \eta_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \tag{80}$$

and as a link function we need a $g : [0, 1] \to \mathbb{R}$. Popular candidates are:

- the logistic function or log-odds ratio

$$g(\pi) = \log \frac{\pi}{1 - \pi}; \tag{81}$$

- the probit function

$$g(\pi) = \Phi^{-1}(\pi), \qquad \text{where } \Phi() \text{ is the Normal CDF;} \tag{82}$$

- and the complementary log-log function

$$g(\pi) = \log(-\log(1 - \pi)). \tag{83}$$

Generalised Linear Models: the `glm()` Function

Fitting a generalised linear model using `glm()` is very similar to fitting a classic linear model with `lm()`. `coef()`, `resid()`, `fitted()`, `predict()` also work in the same way as before.

```
> glm(GROUP ~ ALG + STAT, data = marks, family = binomial(link = logit))

Call:  glm(formula = GROUP ~ ALG + STAT, family = binomial(link = logit),
    data = marks)

Coefficients:
(Intercept)          ALG         STAT
   24.32059     -0.43496     -0.07979

Degrees of Freedom: 87 Total (i.e. Null);  85 Residual
Null Deviance:      119.8
Residual Deviance: 43.98  AIC: 49.98
```

The additional argument `family` specifies which distribution we are assuming for the response and the link function $g()$.

Ridge Regression: the General Idea

For a linear model to be fitted, explanatory variables must not be collinear: otherwise $(\mathbf{X}^T\mathbf{X})$ is not invertible and we cannot estimate $\hat{\boldsymbol{\beta}}$. To make that matrix invertible we can add a penalty term to the least squares minimisation,

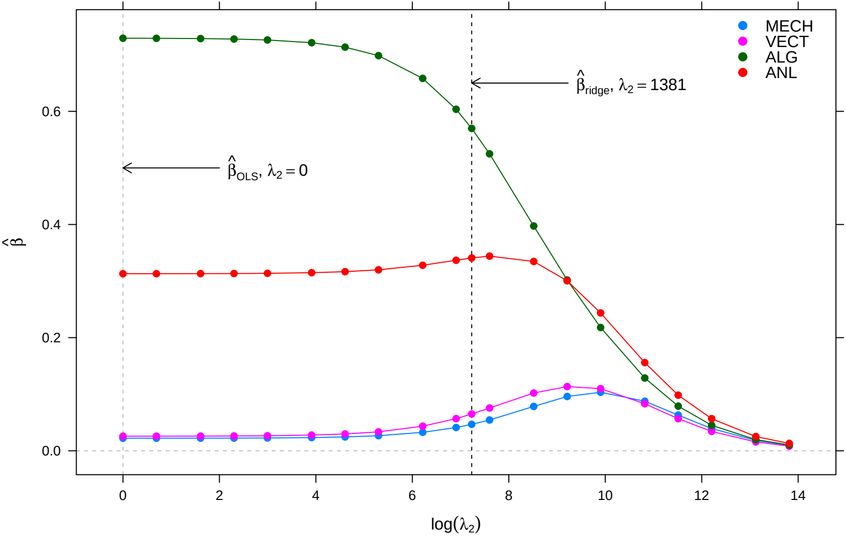
$$\operatorname*{argmin}_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_2 \sum_{i=0}^{p} \beta_i^2 \right\} \qquad \lambda_2 \geqslant 0, \tag{84}$$

so that the estimator for the regression coefficients becomes

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}. \tag{85}$$

We can always compute that because $\lambda_2\mathbf{I}_p$ is full rank, and thus $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p$ is always invertible if $\lambda_2 > 0$.

Ridge Regression: Penalty and Regression Coefficients



Ridge Regression: Choosing λ_2

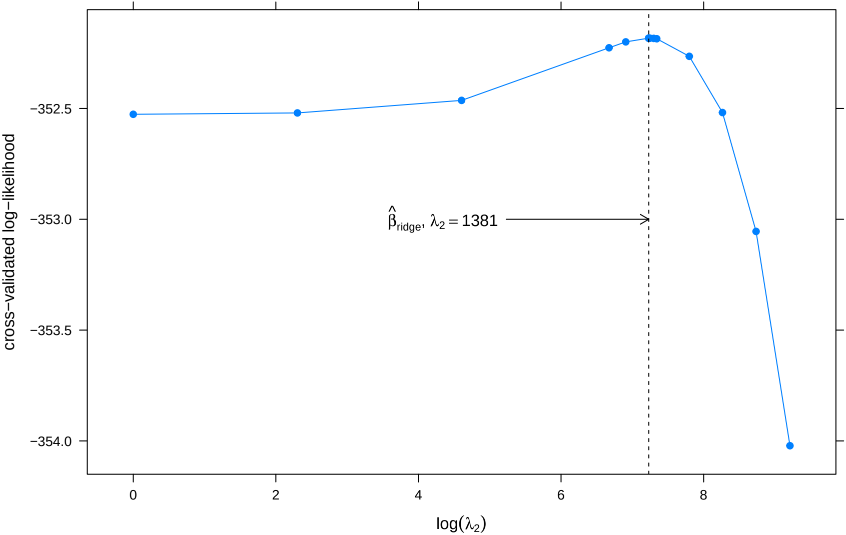
They key point in fitting ridge regression is tuning the model by choosing λ_2, typically using (multiple runs of) cross-validation to find the value that minimises residuals or maximises the likelihood.

```
> library(penalized)
> opt.lambda2 = optL2(response = marks[, "STAT"],
+                 penalized = marks[, c("MECH", "VECT", "ALG", "ANL")])
> penalized(response = marks[, "STAT"],
+   penalized = marks[, c("MECH", "VECT", "ALG", "ANL")],
+   lambda2 = opt.lambda2$lambda)
Penalized linear regression object
5 regression coefficients

Loglikelihood =  -346.6334
L2 penalty =  308.8962  at lambda2 =  1381.118
```

There are several packages to fit ridge regression in `R`, such as `penalized()` (slower, intuitive parameterisation) and `glmnet()` (much faster, different parameterisation).

Ridge Regression: Cross-Validated Log-Likelihood



Bias-Variance Trade-Off

Due to the L_2 penalty the estimated regression coefficients are biased

$$\hat{\boldsymbol{\beta}}_{\mathrm{RIDGE}} = (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y} = [\mathbf{I}_p - \lambda_2(\mathbf{X}^T\mathbf{X})^{-1}]\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \tag{86}$$

and so are the corresponding $\hat{\mathbf{y}}_{\mathrm{RIDGE}}$. On the other hand, introducing that bias may improves prediction because of the bias-variance trade-off:

$$\mathrm{E}[(\mathbf{y} - \hat{\mathbf{y}}_{\mathrm{RIDGE}})^2] = \underbrace{\mathrm{E}[(\mathbf{y} - \mathrm{E}[\hat{\mathbf{y}}_{\mathrm{RIDGE}}])^2]}_{\mathrm{Bias}^2(\hat{\mathbf{y}}_{\mathrm{RIDGE}})} + \underbrace{\mathrm{E}[(\hat{\mathbf{y}}_{\mathrm{RIDGE}} - \mathrm{E}[\hat{\mathbf{y}}_{\mathrm{RIDGE}}])^2]}_{\mathrm{VAR}(\hat{\mathbf{y}}_{\mathrm{RIDGE}})} \tag{87}$$

For new observations, introducing a little bit of bias (towards “simpler” models) can dramatically improve the accuracy of prediction at the cost of degrading the accuracy of fitted values somewhat.

How Many Degrees of Freedom has Ridge Regression?

Take the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. For ordinary least squares, if $\mathbf{X}$ is full rank, then

$$\operatorname{tr}(\mathbf{H}) = p + 1 \tag{88}$$

gives the number of parameters of the model, *i.e.* the degrees of freedom used by the model. For any model of the form $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, we can compute the effective degrees of freedom in the same way. For ridge regression:

$$\operatorname{tr}[\mathbf{H}_{\lambda_2}] = \operatorname{tr}[\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}_p)^{-1}\mathbf{X}^T] = \sum_{i=1}^{p} \frac{d_i^2}{d_i^2 + \lambda_2} \tag{89}$$

where the d_i are the singular values of $\mathbf{X}$ (from the singular value decomposition, SVD).

Does That Look Like Principal Components Again?

Using classic SVD notation, *i.e.* $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, we can write

$$\hat{\mathbf{y}}_{\text{RIDGE}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{RIDGE}} = \sum_{i=1}^{p} \left(\mathbf{u}_i \frac{d_i^2}{d_i^2 + \lambda_2} \mathbf{u}_i^T \right) \mathbf{y} \qquad (90)$$

which means that:

1. ridge regression projects the $\mathbf{y}$ on the principal components of $\mathbf{X}$; and
2. it shrinks low-variance components towards zero while leaving large-variance components relatively untouched.

P.S.: you do not shrink the intercept, you centre the $\mathbf{y}$ first and the and then add it back.

LASSO Regression

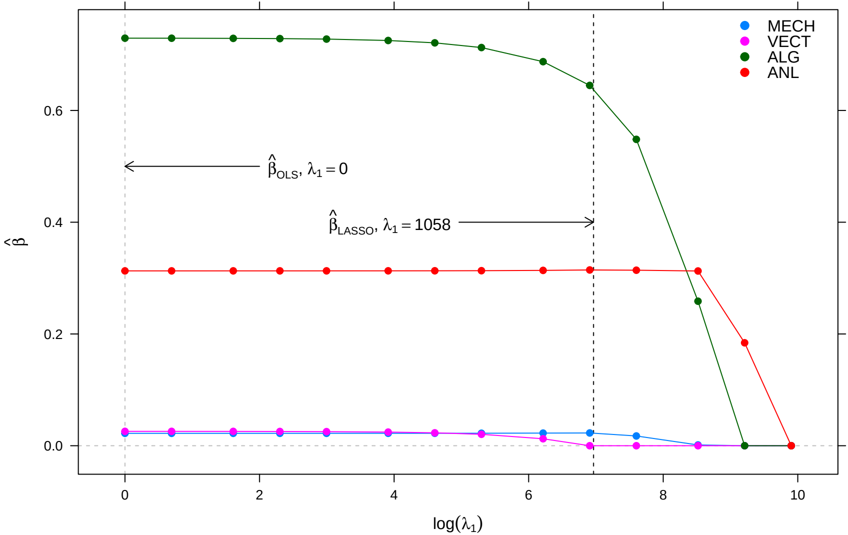
Another approach using penalised least squares is the least absolute shrinkage and selection operator (LASSO), which has a formulation similar to ridge regression:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \sum_{i=0}^{p} |\beta_i| \right\} \qquad \lambda_1 \geqslant 0. \tag{91}$$

The penalty term is based on the absolute values of the β_i (L_1 norm), not on their squares (L_2 norm); this makes the penalty non-smooth and forces coefficients to converge sharply to zero as opposed to getting asymptotically close as the penalty increases.

Since regression coefficients can be forced to be exactly zero, this means we are doing model selection at the same time as model estimation.

LASSO Regression: Penalty and Regression Coefficients



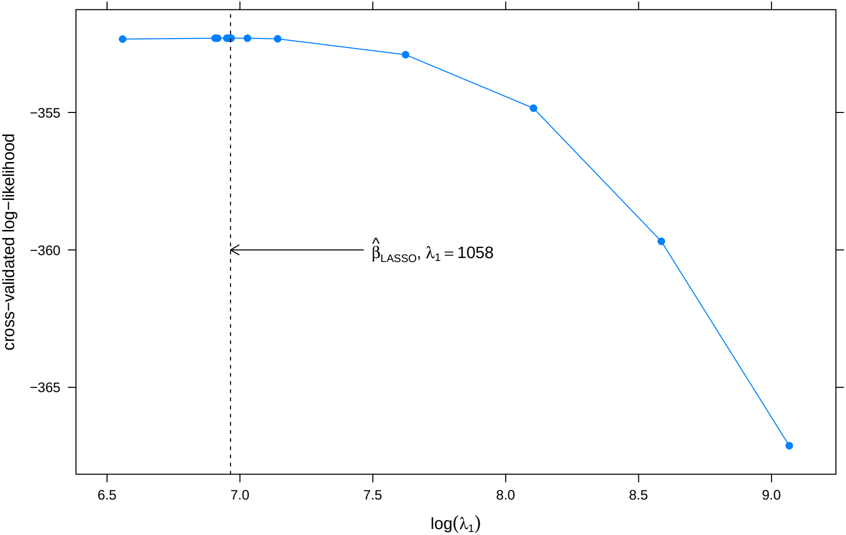
LASSO Regression: Choosing λ_1

Tuning the model is done in the same way as for ridge regression.

```
> library(penalized)
> opt.lambda1 = optL1(response = marks[, "STAT"],
+                 penalized = marks[, c("MECH", "VECT", "ALG", "ANL")])
> penalized(response = marks[, "STAT"],
+   penalized = marks[, c("MECH", "VECT", "ALG", "ANL")],
+   lambda1 = opt.lambda1$lambda)
Penalized linear regression object
5 regression coefficients of which 4 are non-zero

Loglikelihood =  -346.675
L1 penalty =  1033.313  at lambda1 =  1058.482
```

LASSO Regression: Cross-Validated Log-Likelihood



The Elastic Net Regression

Finally, a more flexible model is elastic net: it combines ridge regression and LASSO through their penalties and parameters λ_1 and λ_2:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \sum_{i=0}^{p} |\beta_i| + \lambda_2 \sum_{i=0}^{p} \beta_i^2 \right\} \quad \lambda_1, \lambda_2 \geqslant 0. \tag{92}$$

Ridge regression arises as a particular case when $\lambda_1 = 0$, and LASSO when $\lambda_2 = 0$. The general model is expensive to estimate because it must be tuned on λ_1 and λ_2 simultaneously, typically on a sensibly scaled and spaced grid of values.

Elastic Net: Fitting the Model

Elastic net is also found with a second parameterisation based on the one-to-one transformation:

$$\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2} \qquad \text{and} \qquad \lambda = \frac{\lambda_1 + \lambda_2}{1 + \lambda_2} \tag{93}$$

As we have seen, `penalized()` uses (λ_1, λ_2); `glmnet()` uses (α, λ) but it is easy to convert back and forth:

```
> alpha = lambda1 / (lambda1 + lambda2)
> lambda = (lambda1 + lambda2) / (1 + lambda2)
> glmnet(y = as.matrix(marks[, "STAT"]),
+   x = as.matrix(marks[, c("MECH", "VECT", "ALG", "ANL")]),
+   alpha = alpha, lambda = lambda)
```

or equivalently:

```
> penalized(response = marks[, "STAT"],
+   penalized = marks[, c("MECH", "VECT", "ALG", "ANL")],
+   lambda1 = lambda1, lambda2 = lambda2)
```

Splines Regression

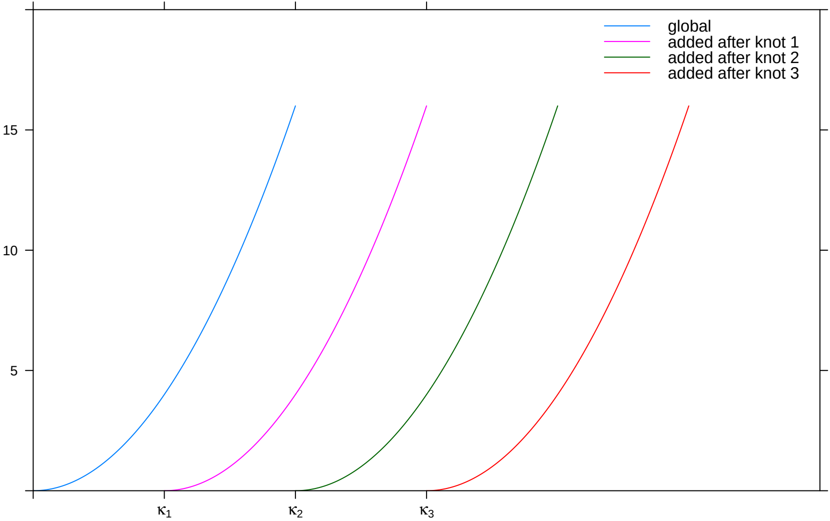
A very flexible way of using polynomial terms and dummy variables such as in (50) and (52) are polynomial or regression splines. They are constructed as follows:

1. divide the range the an explanatory variable in $m \geqslant 2$ intervals delimited on the left by points called knots, *e.g.* $[\min(x_1), \max(x_1)], [\kappa_1, \max(x_1)], \ldots, [\kappa_{m-1}, \max(x_1)]$;
2. set up one dummy variable for each interval, *e.g.* $\mathbb{1}_j(x_1 > \kappa_j)$;
3. decide on a degree l and set up a regression model with a global polynomial set of terms and add interactions with the dummy variables for the highest-order coefficient:

$$y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{i1}^l \beta_{1.l} + \sum_{j=1}^{m} \mathbb{1}_j(x_{i1} > \kappa_j)(x_{i1} - \kappa_j)^l \gamma_j; \quad (94)$$

4. to make the overall regression smooth, enforce smoothness at the knots by constraining it to be differentiable at the knots.

Splines: Knots and Polynomials



Splines: Smoothness and the Complexity

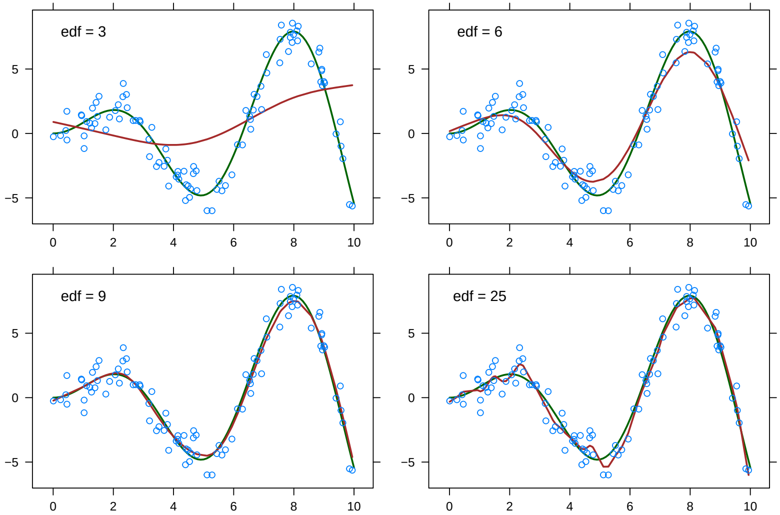
The complexity of a regression splines model is given by:

- the number of knots κ_i, which determines how many times and where the polynomial's trajectory is re-adjusted;
- the degree of the polynomial, which makes the polynomial more or less flexible in fitting the response;
- the constraints on the coefficients, such as those required by differentiability.

Therefore the degrees of freedom of the model in general are not equal to the number of parameters, and have to be estimated from the model as effective degrees of freedom from the $\mathbf{H}$ matrix or the analogous smoothing matrix $\mathbf{S}$ by analogy with classic linear regression, as

$$\mathrm{tr}(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = p + 1. \tag{95}$$

Splines: Equivalent Degrees of Freedom



An R command to fit a spline regression is `sm.spline()` in package `pspline`.

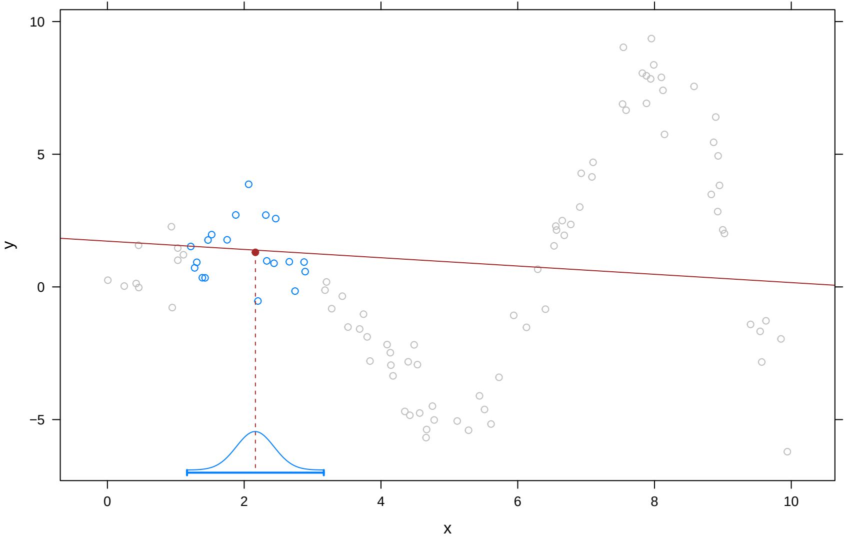
One Last Model: Locally Weighted Least Squares

Another model that can fit "wobbly" regression curves is locally weighted least squares (LOWESS), which is a local linear regression with weights defined as follows for a single explanatory variable:

1. for each y_i, take the corresponding x_{i1} and compute the weights for all other (y_j, x_{j1}) with a non-negative function $w(|x_{i1} - x_{j1}|)$ that decreases as x_{j1} gets farther from x_{i1};
2. estimate $\hat{y}_i$ using weighted linear regression with the weights above.

In practice, $w(\cdot)$ is usually chosen to be exactly zero outside a small interval centred on x_{i1}, whose length is called the bandwidth and determines the smoothness of the regression curve. This makes model estimation much faster since only a small number of observations is used to fit each (y_j, x_{j1}).

LOWESS: Local Weights and Regression



Fitting a LOWESS model

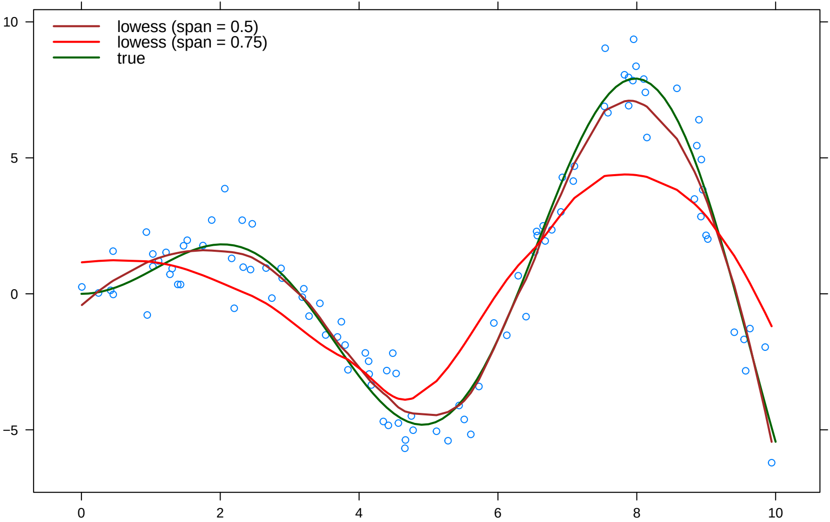
A commonly used function to fit a LOWESS model is `loess()`, which by default fits a second-order polynomial for each y_i using a fraction of the observations controlled by the `span` argument.

```
> x = sort(runif(100, 0, 10))
> y = x * sin(x) + rnorm(100, mean = 0, sd = 1)
> loess(y ~ x, span = 0.5)
Call:
loess(formula = y ~ x, span = 0.5)

Number of Observations: 100
Equivalent Number of Parameters: 6.27
Residual Standard Error: 1.009
> loess(y ~ x, span = 0.75)
Call:
loess(formula = y ~ x, span = 0.75)

Number of Observations: 100
Equivalent Number of Parameters: 4.34
Residual Standard Error: 1.845
```

LOWESS: Bandwidth and Smoothing



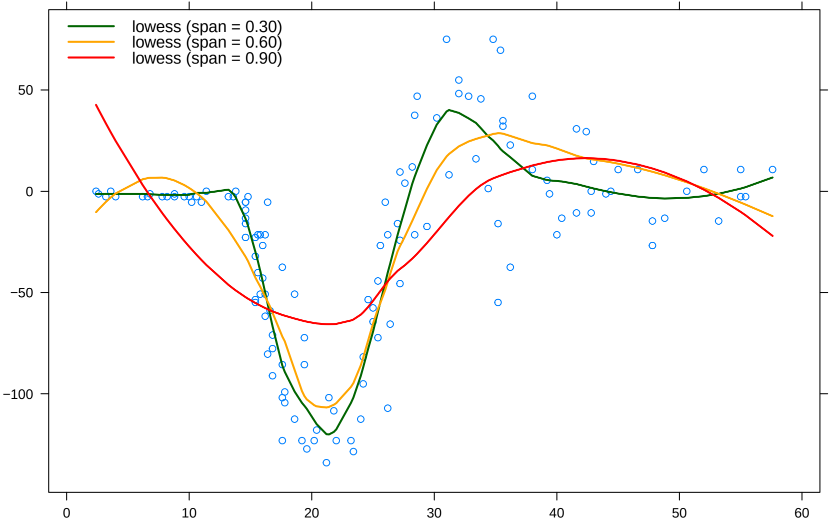
Motorcycle Accident Crash Test

This data set records 133 measurements of head acceleration (in multiples of g) and time after impact (in milliseconds) in a simulated motorcycle crash experiment on the efficacy of crash helmets.

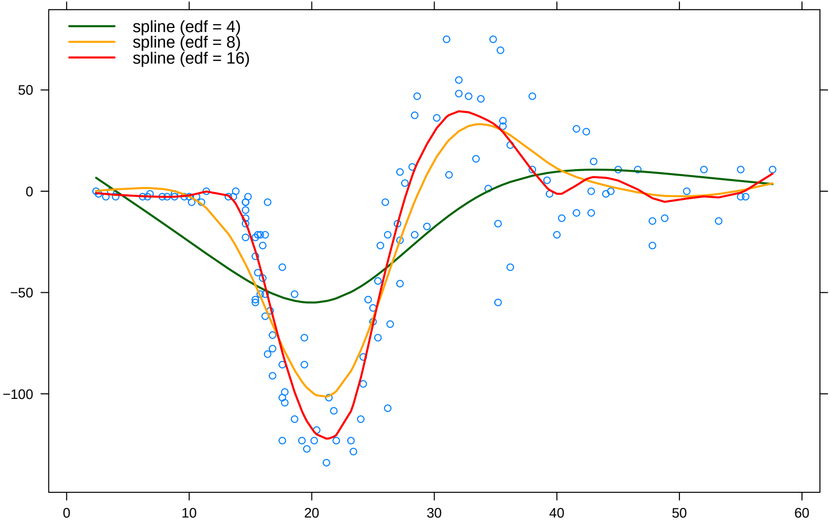
```
> library(MASS)
> head(mcycle)
  times accel
1   2.4   0.0
2   2.6  -1.3
3   3.2  -2.7
4   3.6   0.0
5   4.0  -2.7
6   6.2  -2.7
> cor(mcycle$times, mcycle$accel)
[1] 0.2964033
```

The two variables `times` and `accel` are strongly dependent, but not linearly; therefore their marginal correlation is small and a classic linear model has $R^2 \approx 0.08$.

Crash Test: a LOWESS Model



Crash Test: a Splines Model



The End