

Computational Statistics

Spring 2008

Peter Bühlmann and Martin Mächler

Seminar für Statistik
ETH Zürich

February 2008 (February 20, 2008)

Contents

1	Multiple Linear Regression	1
1.1	Introduction	1
1.2	The Linear Model	1
1.2.1	Stochastic Models	2
1.2.2	Examples	2
1.2.3	Goals of a linear regression analysis	3
1.3	Least Squares Method	4
1.3.1	The normal equations	4
1.3.2	Assumptions for the Linear Model	5
1.3.3	Geometrical Interpretation	6
1.3.4	Don't do many regressions on single variables!	7
1.3.5	Computer-Output from R : Part I	8
1.4	Properties of Least Squares Estimates	9
1.4.1	Moments of least squares estimates	9
1.4.2	Distribution of least squares estimates assuming Gaussian errors	10
1.5	Tests and Confidence Regions	10
1.5.1	Computer-Output from R : Part II	12
1.6	Analysis of residuals and checking of model assumptions	12
1.6.1	The Tukey-Anscombe Plot	12
1.6.2	The Normal Plot	13
1.6.3	Plot for detecting serial correlation	14
1.6.4	Generalized least squares and weighted regression	14
1.7	Model Selection	15
1.7.1	Mallows C_p statistic	15
2	Nonparametric Density Estimation	17
2.1	Introduction	17
2.2	Estimation of a density	17
2.2.1	Histogram	17
2.2.2	Kernel estimator	18
2.3	The role of the bandwidth	19
2.3.1	Variable bandwidths: k nearest neighbors	20
2.3.2	The bias-variance trade-off	20
2.3.3	Asymptotic bias and variance	20
2.3.4	Estimating the bandwidth	22
2.4	Higher dimensions	23
2.4.1	The curse of dimensionality	23

3	Nonparametric Regression	25
3.1	Introduction	25
3.2	The kernel regression estimator	25
3.2.1	The role of the bandwidth	27
3.2.2	Inference for the underlying regression curve	28
3.3	Local polynomial nonparametric regression estimator	29
3.4	Smoothing splines and penalized regression	30
3.4.1	Penalized sum of squares	30
3.4.2	The smoothing spline solution	30
3.4.3	Shrinking towards zero	31
3.4.4	Relation to equivalent kernels	31
4	Cross-Validation	33
4.1	Introduction	33
4.2	Training and Test Set	33
4.3	Constructing training-, test-data and cross-validation	34
4.3.1	Leave-one-out cross-validation	34
4.3.2	K -fold Cross-Validation	35
4.3.3	Random divisions into test- and training-data	35
4.4	Properties of different CV-schemes	36
4.4.1	Leave-one-out CV	36
4.4.2	Leave- d -out CV	36
4.4.3	K -fold CV; stochastic approximations	37
4.5	Computational shortcut for some linear fitting operators	37
5	Bootstrap	39
5.1	Introduction	39
5.2	Efron's nonparametric bootstrap	39
5.2.1	The bootstrap algorithm	40
5.2.2	The bootstrap distribution	41
5.2.3	Bootstrap confidence interval: a first approach	41
5.2.4	Bootstrap estimate of the generalization error	44
5.2.5	Out-of-bootstrap sample for estimation of the generalization error	45
5.3	Double bootstrap	45
5.4	Model-based bootstrap	47
5.4.1	Parametric bootstrap	47
5.4.2	Model structures beyond i.i.d. and the parametric bootstrap	50
5.4.3	The model-based bootstrap for regression	51
6	Classification	53
6.1	Introduction	53
6.2	The Bayes classifier	53
6.3	The view of discriminant analysis	54
6.3.1	Linear discriminant analysis	54
6.3.2	Quadratic discriminant analysis	55
6.4	The view of logistic regression	55
6.4.1	Binary classification	55
6.4.2	Multiclass case, $J > 2$	58

7	Flexible regression and classification methods	61
7.1	Introduction	61
7.2	Additive models	61
7.2.1	Backfitting for additive regression models	62
7.2.2	Additive model fitting in R	62
7.3	MARS	66
7.3.1	Hierarchical interactions and constraints	67
7.3.2	MARS in R	67
7.4	Neural Networks	68
7.4.1	Fitting neural networks in R	68
7.5	Projection pursuit regression	70
7.5.1	Projection pursuit regression in R	70
7.6	Classification and Regression Trees (CART)	71
7.6.1	Tree structured estimation and tree representation	71
7.6.2	Tree-structured search algorithm and tree interpretation	72
7.6.3	Pros and cons of trees	75
7.6.4	CART in R	75
7.7	Variable Selection, Regularization, Ridging and the Lasso	76
7.7.1	Introduction	76
7.7.2	Ridge Regression	77
7.7.3	The Lasso	78
7.7.4	Lasso extensions	79
8	Bagging and Boosting	83
8.1	Introduction	83
8.2	Bagging	83
8.2.1	The bagging algorithm	83
8.2.2	Bagging for trees	84
8.2.3	Subagging	84
8.3	Boosting	85
8.3.1	L_2 Boosting	85

Chapter 1

Multiple Linear Regression

1.1 Introduction

Linear regression is a widely used statistical model in a broad variety of applications. It is one of the easiest examples to demonstrate important aspects of statistical modelling.

1.2 The Linear Model

Multiple Regression Model:

Given is one **response variable**: up to some random errors it is a linear function of several **predictors** (or **covariables**).

The linear function involves unknown parameters. The goal is to estimate these parameters, to study their relevance and to estimate the error variance.

Model formula:

$$Y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i \quad (i = 1, \dots, n) \quad (1.1)$$

Usually we assume that $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. (independent, identically distributed) with $\mathbb{E}[\varepsilon_i] = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$.

Notations:

- $\mathbf{Y} = \{Y_i; i = 1, \dots, n\}$ is the vector of the **response variables**
- $\mathbf{x}^{(j)} = \{x_{ij}; i = 1, \dots, n\}$ is the vector of the j th predictor (covariable) ($j = 1, \dots, p$)
- $\mathbf{x}_i = \{x_{ij}; j = 1, \dots, p\}$ is the vector of predictors for the i th observation ($i = 1, \dots, n$)
- $\boldsymbol{\beta} = \{\beta_j; j = 1, \dots, p\}$ is the vector of the unknown parameters
- $\boldsymbol{\varepsilon} = \{\varepsilon_i; i = 1, \dots, n\}$ is the vector of the unknown random **errors**
- n is the sample size, p is the number of predictors

The parameters β_j and σ^2 are unknown and the errors ε_i are unobservable. On the other hand, the response variables Y_i and the predictors x_{ij} have been observed.

Model in vector notation:

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \dots, n)$$

Model in matrix form:

$$\begin{array}{c} \mathbf{Y} \\ n \times 1 \end{array} = \begin{array}{c} X \\ n \times p \end{array} \times \begin{array}{c} \boldsymbol{\beta} \\ p \times 1 \end{array} + \begin{array}{c} \boldsymbol{\varepsilon} \\ n \times 1 \end{array} \quad (1.2)$$

where X is a $(n \times p)$ -matrix with rows \mathbf{x}_i^\top and columns $\mathbf{x}^{(j)}$.

The first predictor variable is often a constant, i.e., $x_{i1} \equiv 1$ for all i . We then get an intercept in the model.

$$Y_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i.$$

We typically assume that the sample size n is larger than the number of predictors p , $n > p$, and moreover that the matrix X has full rank p , i.e., the p column vectors $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)}$ are linearly independent.

1.2.1 Stochastic Models

The linear model in (1.1) involves some stochastic (random) components: the error terms ε_i are random variables and hence the response variables Y_i as well. The predictor variables x_{ij} are here assumed to be non-random. In some applications, however it is more appropriate to treat the predictor variables as random.

The stochastic nature of the error terms ε_i can be assigned to various sources: for example, measurement errors or inability to capture all underlying non-systematic effects which are then summarized by a random variable with expectation zero. The stochastic modelling approach will allow to quantify uncertainty, to assign significance to various components, e.g. significance of predictor variables in model (1.1), and to find a good compromise between the size of a model and the ability to describe the data (see section 1.7).

The observed response in the data is always assumed to be realizations of the random variables Y_1, \dots, Y_n ; the x_{ij} 's are non-random and equal to the observed predictors in the data.

1.2.2 Examples

Two-sample model:

$$p = 2, \quad X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.$$

Main questions: $\mu_1 = \mu_2$? Quantitative difference between μ_1 and μ_2 ?

From introductory courses we know that one could use the two-sample t -test or two-sample Wilcoxon test.

Regression through the origin: $Y_i = \beta x_i + \varepsilon_i$ ($i = 1, \dots, n$).

$$p = 1, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \beta.$$

Simple linear regression: $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ ($i = 1, \dots, n$).

$$p = 2 \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Quadratic regression: $Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i$ ($i = 1, \dots, n$).

$$p = 3, \quad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Note that the fitted function is quadratic in the x_i 's but *linear* in the coefficients β_j and therefore a special case of the linear model (1.1).

Regression with transformed predictor variables:

$Y_i = \beta_1 + \beta_2 \log(x_{i2}) + \beta_3 \sin(\pi x_{i3}) + \varepsilon_i$ ($i = 1, \dots, n$).

$$p = 3, \quad X = \begin{pmatrix} 1 & \log(x_{12}) & \sin(\pi x_{13}) \\ 1 & \log(x_{22}) & \sin(\pi x_{23}) \\ \vdots & \vdots & \vdots \\ 1 & \log(x_{n2}) & \sin(\pi x_{n3}) \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Again, the model is *linear* in the coefficients β_j but nonlinear in the x_{ij} 's.

In summary:

The model in (1.1) is called linear because it is linear in the coefficients β_j . The predictor (and also the response) variables can be transformed versions of the original predictor and/or response variables.

1.2.3 Goals of a linear regression analysis

- **A good “fit”.** Fitting or estimating a (hyper-)plane over the predictor variables to explain the response variables such that the errors are “small”. The standard tool for this is the method of *least squares* (see section 1.3).
- **Good parameter estimates.** This is useful to describe the change of the response when varying some predictor variable(s).
- **Good prediction.** This is useful to predict a new response as a function of new predictor variables.

- **Uncertainties and significance for the three goals above.** Confidence intervals and statistical tests are useful tools for this goal.
- **Development of a good model.** In an interactive process, using methods for the goals mentioned above, we may change parts of an initial model to come up with a better model.

The first and third goal can become conflicting, see section 1.7.

1.3 Least Squares Method

We assume the linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. We are looking for a “good” estimate of $\boldsymbol{\beta}$. The least squares estimator $\hat{\boldsymbol{\beta}}$ is defined as

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{Y} - X\boldsymbol{\beta}\|^2, \quad (1.3)$$

where $\|\cdot\|$ denotes the Euclidean norm in \mathbb{R}^n .

1.3.1 The normal equations

The minimizer in (1.3) can be computed explicitly (assuming that X has rank p). Computing partial derivatives of $\|\mathbf{Y} - X\boldsymbol{\beta}\|^2$ with respect to $\boldsymbol{\beta}$ (p -dimensional gradient vector), evaluated at $\hat{\boldsymbol{\beta}}$, and setting them to zero yields

$$(-2) X^\top(\mathbf{Y} - X\hat{\boldsymbol{\beta}}) = \mathbf{0} \quad ((p \times 1) - \text{null-vector}).$$

Thus, we get the **normal equations**

$$X^\top X \hat{\boldsymbol{\beta}} = X^\top \mathbf{Y}. \quad (1.4)$$

These are p linear equations for the p unknowns (components of $\hat{\boldsymbol{\beta}}$).

Assuming that the matrix X has full rank p , the $p \times p$ matrix $X^\top X$ is invertible, the least squares estimator is unique and can be represented as

$$\boxed{\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y}.}$$

This formula is useful for theoretical purposes. For numerical computation it is much more stable to use the QR decomposition instead of inverting the matrix $X^\top X$.¹

From the *residuals* $r_i = Y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$, the usual estimate for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n r_i^2.$$

Note that the r_i 's are estimates for ε_i 's; hence the estimator is plausible, up to the somewhat unusual factor $1/(n-p)$. It will be shown in section 1.4.1 that due to this factor, $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.

¹Let $X = QR$ with orthogonal ($n \times p$) matrix Q and upper (Right) triangular ($p \times p$) R . Because of $X^\top X = R^\top Q^\top QR = R^\top R$, computing $\boldsymbol{\beta}$ only needs subsequent solution of two triangular systems: First solve $R^\top \mathbf{c} = X^\top \mathbf{Y}$ for \mathbf{c} , and then solve $R\hat{\boldsymbol{\beta}} = \mathbf{c}$.

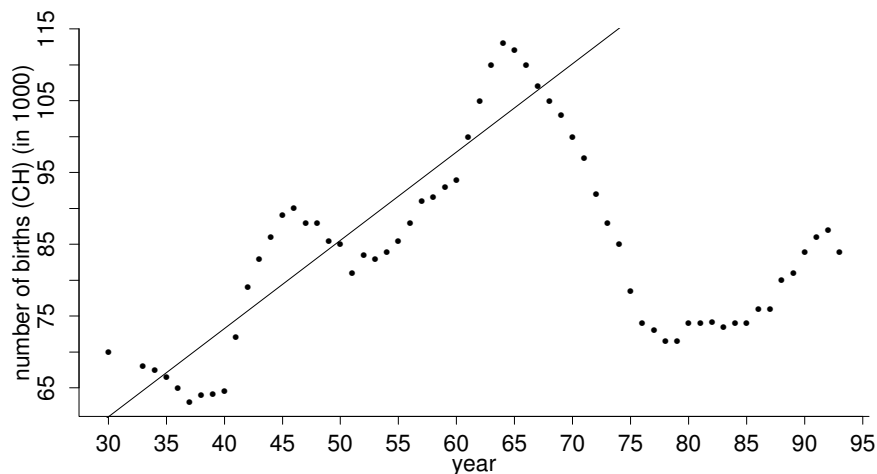


Figure 1.1: The pill kink.

1.3.2 Assumptions for the Linear Model

We emphasize here that we do not make any assumptions on the predictor variables, except that the matrix X has full rank $p < n$. In particular, the predictor variables can be continuous or discrete (e.g. binary).

We need some assumptions so that fitting a linear model by least squares is reasonable and that tests and confidence intervals (see 1.5) are approximately valid.

1. **The linear regression equation is correct.** This means: $\mathbb{E}[\varepsilon_i] = 0$ for all i .
2. **All x_i 's are exact.** This means that we can observe them perfectly.
3. **The variance of the errors is constant (“homoscedasticity”).** This means: $\text{Var}(\varepsilon_i) = \sigma^2$ for all i .
4. **The errors are uncorrelated.** This means: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.
5. **The errors $\{\varepsilon_i; i = 1, \dots, n\}$ are jointly normally distributed.** This implies that also $\{Y_i; i = 1, \dots, n\}$ are jointly normally distributed.

In case of violations of item 3, we can use weighted least squares instead of least squares. Similarly, if item 4 is violated, we can use generalized least squares. If the normality assumption in 5 does not hold, we can use robust methods instead of least squares. If assumption 2 fails to be true, we need corrections known from “errors in variables” methods. If the crucial assumption in 1 fails, we need other models than the linear model.

The following example shows violations of assumption 1 and 4. The response variable is the annual number of births in Switzerland since 1930, and the predictor variable is the time (year).

We see in Figure 1.1 that the data can be approximately described by a linear relation until the “pill kink” in 1964. We also see that the errors seem to be correlated: they are all positive or negative during periods of 10 – 20 years. Finally, the linear model is not representative after the pill kink in 1964. In general, it is dangerous to use a fitted model for extrapolation where no predictor variables have been observed (for example: if we

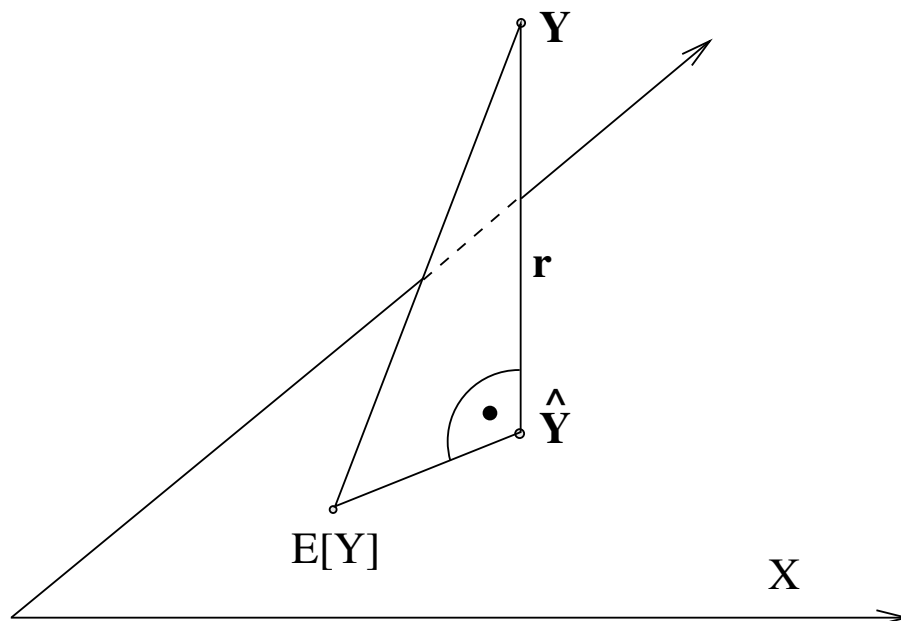


Figure 1.2: The residual vector \mathbf{r} is orthogonal to \mathcal{X} .

would have fitted the linear model in 1964 for prediction of number of births in the future until 2005).

1.3.3 Geometrical Interpretation

The response variable \mathbf{Y} is a vector in \mathbb{R}^n . Also, $X\boldsymbol{\beta}$ describes a p -dimensional subspace \mathcal{X} in \mathbb{R}^n (through the origin) when varying $\boldsymbol{\beta} \in \mathbb{R}^p$ (assuming that X has full rank p). The least squares estimator $\hat{\boldsymbol{\beta}}$ is then such that $X\hat{\boldsymbol{\beta}}$ is closest to \mathbf{Y} with respect to the Euclidean distance. But this means geometrically that

$$\boxed{X\hat{\boldsymbol{\beta}} \text{ is the orthogonal projection of } \mathbf{Y} \text{ onto } \mathcal{X}.$$

We denote the (vector of) fitted values by

$$\hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}}.$$

They can be viewed as an estimate of $X\boldsymbol{\beta}$.

The (vector of) residuals is defined by

$$\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

Geometrically, it is evident that the residuals are orthogonal to \mathcal{X} , because $\hat{\mathbf{Y}}$ is the orthogonal projection of \mathbf{Y} onto \mathcal{X} . This means that

$$\mathbf{r}^\top \mathbf{x}^{(j)} = 0 \text{ for all } j = 1, \dots, p,$$

where $\mathbf{x}^{(j)}$ is the j th column of X .

We can formally see why the map

$$\mathbf{Y} \mapsto \hat{\mathbf{Y}}$$

is an orthogonal projection. Since $\hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}} = X(X^\top X)^{-1}X^\top \mathbf{Y}$, the map can be represented by the matrix

$$P = X(X^\top X)^{-1}X^\top. \quad (1.5)$$

It is evident that P is symmetric ($P^\top = P$) and P is idem-potent ($P^2 = P$). Furthermore

$$\sum_i P_{ii} = \text{tr}(P) = \text{tr}(X(X^\top X)^{-1}X^\top) = \text{tr}((X^\top X)^{-1}X^\top X) = \text{tr}(I_{p \times p}) = p.$$

But these 3 properties characterize that P is an orthogonal projection from \mathbb{R}^n onto a p -dimensional subspace, here \mathcal{X} .

The residuals \mathbf{r} can be represented as

$$\mathbf{r} = (I - P)\mathbf{Y},$$

where $I - P$ is now also an orthogonal projection onto the orthogonal complement of \mathcal{X} , $\mathcal{X}^\perp = \mathbb{R}^n \setminus \mathcal{X}$, which is $(n - p)$ -dimensional. In particular, the residuals are elements of \mathcal{X}^\perp .

1.3.4 Don't do many regressions on single variables!

In general, it is not appropriate to replace multiple regression by many single regressions (on single predictor variables). The following (synthetic) example should help to demonstrate this point.

Consider two predictor variables $x^{(1)}, x^{(2)}$ and a response variable Y with the values

$x^{(1)}$	0	1	2	3	0	1	2	3
$x^{(2)}$	-1	0	1	2	1	2	3	4
Y	1	2	3	4	-1	0	1	2

Multiple regression yields the least squares solution which describes the data points exactly

$$Y_i = \hat{Y}_i = 2x_{i1} - x_{i2} \text{ for all } i \quad (\hat{\sigma}^2 = 0). \quad (1.6)$$

The coefficients 2 and -1, respectively, describe how y is changing when varying either $x^{(1)}$ or $x^{(2)}$ and keeping the other predictor variable constant. In particular, we see that Y decreases when $x^{(2)}$ increases.

On the other hand, if we do a simple regression of Y onto $x^{(2)}$ (while ignoring the values of $x^{(1)}$; and thus, we do not keep them constant), we obtain the least squares estimate

$$\hat{Y}_i = \frac{1}{9}x_{i2} + \frac{4}{3} \text{ for all } i \quad (\hat{\sigma}^2 = 1.72).$$

This least squares regression line describes how Y changes when varying $x^{(2)}$ while ignoring $x^{(1)}$. In particular, \hat{Y} increases when $x^{(2)}$ increases, in contrast to multiple regression!

The reason for this phenomenon is that $x^{(1)}$ and $x^{(2)}$ are strongly correlated: if $x^{(2)}$ increases, then also $x^{(1)}$ increases. Note that in the multiple regression solution, $x^{(1)}$ has a larger coefficient in absolute value than $x^{(2)}$ and hence, an increase in $x^{(1)}$ has a stronger influence for changing y than $x^{(2)}$. The correlation among the predictors in general makes also the interpretation of the regression coefficients more subtle: in the current setting, the coefficient β_1 quantifies the influence of $x^{(1)}$ on Y *after* having subtracted the effect of $x^{(2)}$ on Y , see also section 1.5.

Summarizing:

Simple least squares regressions on single predictor variables yield the multiple regression least squares solution, *only* if the predictor variables are orthogonal.

In general, *multiple* regression is the appropriate tool to include effects of more than one predictor variables simultaneously.

The equivalence in case of orthogonal predictors is easy to see algebraically. Orthogonality of predictors means $X^T X = \text{diag}(\sum_{i=1}^n x_{i1}^2, \dots, \sum_{i=1}^n x_{ip}^2)$ and hence the least squares estimator

$$\hat{\beta}_j = \sum_{i=1}^n x_{ij} Y_i / \sum_{i=1}^n x_{ij}^2 \quad (j = 1, \dots, p),$$

i.e., $\hat{\beta}_j$ depends only on the response variable Y_i and the j th predictor variable x_{ij} .

1.3.5 Computer-Output from R : Part I

We show here parts of the computer output (from R) when fitting a linear model to data about quality of asphalt.

```

y = LOGRUT : log("rate of rutting") = log(change of rut depth in inches
              per million wheel passes)
              ["rut" := 'Wagenspur', ausgefahrenes Geleise]
x1 = LOGVISC : log(viscosity of asphalt)
x2 = ASPH    : percentage of asphalt in surface course
x3 = BASE    : percentage of asphalt in base course
x4 = RUN     : '0/1' indicator for two sets of runs.
x5 = FINES   : 10* percentage of fines in surface course
x6 = VOIDS   : percentage of voids in surface course

```

The following table shows the least squares estimates $\hat{\beta}_j$ ($j = 1, \dots, 6$), some empirical quantiles of the residuals r_i ($i = 1, \dots, n$), the estimated standard deviation of the errors² $\sqrt{\hat{\sigma}^2}$ and the so-called *degrees of freedom* $n - p$.

Call:

```
lm(formula = LOGRUT ~ . , data = asphalt1)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.48348	-0.14374	-0.01198	0.15523	0.39652

Coefficients:

	Estimate
(Intercept)	-5.781239
LOGVISC	-0.513325
ASPH	1.146898
BASE	0.232809
RUN	-0.618893

² The term "residual standard error" is a misnomer with a long tradition, since "standard error" usually means $\sqrt{\text{Var}(\hat{\theta})}$ for an estimated parameter θ .

FINES 0.004343
VOIDS 0.316648

Residual standard error: 0.2604 on 24 degrees of freedom

1.4 Properties of Least Squares Estimates

As an introductory remark, we point out that the least squares estimates are random variables: for new data from the same data-generating mechanism, the data would look differently every time and hence also the least squares estimates. Figure 1.3 displays three least squares regression lines which are based on three different realizations from the same data-generating model (i.e., three simulations from a model). We see that the estimates are varying: they are random themselves!

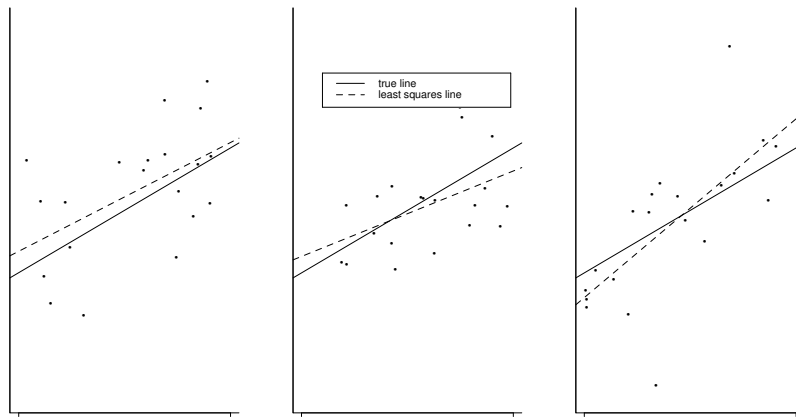


Figure 1.3: Three least squares estimated regression lines for three different data realizations from the same model.

1.4.1 Moments of least squares estimates

We assume here the usual linear model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}[\boldsymbol{\varepsilon}] = 0, \quad \text{Cov}(\boldsymbol{\varepsilon}) = \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2 I_{n \times n}. \quad (1.7)$$

This means that assumption 1.-4. from section 1.3.2 are satisfied.

It can then be shown that:

- (i) $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$: that is, $\hat{\boldsymbol{\beta}}$ is **unbiased**
- (ii) $\mathbb{E}[\hat{\mathbf{Y}}] = \mathbb{E}[\mathbf{Y}] = X\boldsymbol{\beta}$ which follows from (i). Moreover, $\mathbb{E}[\mathbf{r}] = 0$.
- (iii) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(X^\top X)^{-1}$
- (iv) $\text{Cov}(\hat{\mathbf{Y}}) = \sigma^2 P$, $\text{Cov}(\mathbf{r}) = \sigma^2(I - P)$

The residuals (which are estimates of the unknown errors ε_i) are also having expectation zero but they are not uncorrelated:

$$\text{Var}(r_i) = \sigma^2(1 - P_{ii}).$$

From this, we obtain

$$\mathbb{E}\left[\sum_{i=1}^n r_i^2\right] = \sum_{i=1}^n \mathbb{E}[r_i^2] = \sum_{i=1}^n \text{Var}(r_i) = \sigma^2 \sum_{i=1}^n (1 - P_{ii}) = \sigma^2(n - \text{tr}(P)) = \sigma^2(n - p).$$

Therefore, $\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}[\sum_{i=1}^n r_i^2 / (n - p)] = \sigma^2$ is **unbiased**.

1.4.2 Distribution of least squares estimates assuming Gaussian errors

We assume the linear model as in (1.7) but require in addition that $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. It can then be shown that:

- (i) $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(X^\top X)^{-1})$
- (ii) $\hat{\mathbf{Y}} \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 P)$, $\mathbf{r} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2(I - P))$
- (iii) $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p} \chi_{n-p}^2$.

The normality assumptions of the errors ε_i is often not (approximately) fulfilled in practice. We can then rely on the central limit theorem which implies that for large sample size n , the properties (i)-(iii) above are still approximately true. This is the usual justification in practice to use these properties for constructing confidence intervals and tests for the linear model parameters. However, it is often much better to use **robust methods** in case of non-Gaussian errors which we are not discussing here.

1.5 Tests and Confidence Regions

We assume the linear model as in (1.7) with $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ (or with ε_i 's i.i.d. and “large” sample size n). As we have seen above, the parameter estimates $\hat{\boldsymbol{\beta}}$ are normally distributed.

If we are interested whether the j th predictor variable is relevant, we can test the null-hypothesis $H_{0,j} : \beta_j = 0$ against the alternative $H_{A,j} : \beta_j \neq 0$. We can then easily derive from the normal distribution of $\hat{\beta}_j$ that

$$\frac{\hat{\beta}_j}{\sqrt{\sigma^2(X^\top X)^{-1}_{jj}}} \sim \mathcal{N}(0, 1) \text{ under the null-hypothesis } H_{0,j}.$$

Since σ^2 is unknown, this quantity is not useful, but if we substitute it with the estimate $\hat{\sigma}^2$ we obtain the so-called t -test statistic

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2(X^\top X)^{-1}_{jj}}} \sim t_{n-p} \text{ under the null-hypothesis } H_{0,j}, \quad (1.8)$$

which has a slightly different distribution than standard Normal $\mathcal{N}(0, 1)$. The corresponding test is then called the t -test. In practice, we can thus quantify the relevance of individual predictor variables by looking at the size of the test-statistics T_j ($j = 1, \dots, p$) or at the corresponding P -values which may be more informative.

The problem by looking at *individual* tests $H_{0,j}$ is (besides the multiple testing problem in general) that it can happen that all individual tests do not reject the null-hypotheses (say at the 5% significance level) although it is true that some predictor variables have

a significant effect. This “paradox” can occur because of correlation among the predictor variables.

An individual t -test for $H_{0,j}$ should be interpreted as quantifying the effect of the j th predictor variable after having subtracted the linear effect of all other predictor variables on Y .

To test whether there exists *any* effect from the predictor variables, we can look at the simultaneous null-hypothesis $H_0 : \beta_2 = \dots = \beta_p = 0$ versus the alternative $H_A : \beta_j \neq 0$ for at least one $j \in \{2, \dots, p\}$; we assume here that the first predictor variable is the constant $X_{i,1} \equiv 1$ (there are $p - 1$ (non-trivial) predictor variables). Such a test can be developed with an analysis of variance (*anova*) decomposition which takes a simple form for this special case:

$$\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2 = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$$

which decomposes the total squared error $\mathbf{Y} - \bar{\mathbf{Y}}$ around the mean $\bar{\mathbf{Y}} = n^{-1} \sum_{i=1}^n Y_i \cdot \mathbf{1}$ as a sum of the squared error due to the regression $\hat{\mathbf{Y}} - \bar{\mathbf{Y}}$ (the amount that the fitted values vary around the global arithmetic mean) and the squared residual error $\mathbf{r} = \mathbf{Y} - \hat{\mathbf{Y}}$. (The equality can be seen most easily from a geometrical point of view: the residuals \mathbf{r} are orthogonal to \mathcal{X} and hence to $\hat{\mathbf{Y}} - \bar{\mathbf{Y}}$). Such a decomposition is usually summarized by an ANOVA table (**AN**alysis **O**f **VA**riance).

	sum of squares	degrees of freedom	mean square	\mathbf{E} [mean square]
regression	$\ \hat{\mathbf{Y}} - \bar{\mathbf{Y}}\ ^2$	$p - 1$	$\ \hat{\mathbf{Y}} - \bar{\mathbf{Y}}\ ^2 / (p - 1)$	$\sigma^2 + \frac{\ \mathbf{E}[\mathbf{Y}] - \mathbf{E}[\bar{\mathbf{Y}}]\ ^2}{p - 1}$
error	$\ \mathbf{Y} - \hat{\mathbf{Y}}\ ^2$	$n - p$	$\ \mathbf{Y} - \hat{\mathbf{Y}}\ ^2 / (n - p)$	σ^2
total around global mean	$\ \mathbf{Y} - \bar{\mathbf{Y}}\ ^2$	$n - 1$	—	—

In case of the global null-hypothesis, there is no effect of any predictor variable and hence $\mathbb{E}[\mathbf{Y}] \equiv \text{const.} = \mathbb{E}[\bar{\mathbf{Y}}]$: therefore, the expected mean square equals σ^2 under H_0 . The idea is now to divide the mean square by the estimate $\hat{\sigma}^2$ to obtain a scale-free quantity: this leads to the so-called F -statistic

$$F = \frac{\|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 / (p - 1)}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 / (n - p)} \sim F_{p-1, n-p} \text{ under the global null-hypothesis } H_0.$$

This test is called the F -test (it is one among several other F -tests in regression).

Besides performing a global F -test to quantify the statistical significance of the predictor variables, we often want to describe the goodness of fit of the linear model for explaining the data. A meaningful quantity is the coefficient of determination, abbreviated by R^2 ,

$$R^2 = \frac{\|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2}.$$

This is the proportion of the total variation of \mathbf{Y} around $\bar{\mathbf{Y}}$ which is explained by the regression (see the ANOVA decomposition and table above).

Similarly to the t -tests as in (1.8), one can derive confidence intervals for the unknown parameters β_j :

$$\hat{\beta}_j \pm \sqrt{\hat{\sigma}^2 (X^\top X)_{jj}^{-1}} \cdot t_{n-p; 1-\alpha/2}$$

is a two-sided confidence interval which covers the true β_j with probability $1 - \alpha$; here, $t_{n-p; 1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of a t_{n-p} distribution.

1.5.1 Computer-Output from R: Part II

We consider again the dataset from section 1.3.5. We now give the complete list of summary statistics from a linear model fit to the data.

Call:

```
lm(formula = LOGRUT ~ ., data = asphalt1)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.48348	-0.14374	-0.01198	0.15523	0.39652

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-5.781239	2.459179	-2.351	0.027280	*
LOGVISC	-0.513325	0.073056	-7.027	2.90e-07	***
ASPH	1.146898	0.265572	4.319	0.000235	***
BASE	0.232809	0.326528	0.713	0.482731	
RUN	-0.618893	0.294384	-2.102	0.046199	*
FINES	0.004343	0.007881	0.551	0.586700	
VOIDS	0.316648	0.110329	2.870	0.008433	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2604 on 24 degrees of freedom

Multiple R-Squared: 0.9722, Adjusted R-squared: 0.9653

F-statistic: 140.1 on 6 and 24 DF, p-value: < 2.2e-16

The table displays the standard errors of the estimates $\sqrt{\widehat{\text{Var}}(\hat{\beta}_j)} = \sqrt{\hat{\sigma}^2(X^\top X)^{-1}_{jj}}$, the t -test statistics for the null-hypotheses $H_{0,j} : \beta_j = 0$ and their corresponding two-sided P -values with some abbreviation about strength of significance. Moreover, the R^2 and adjusted R^2 are given and finally also the F -test statistic for the null-hypothesis $H_0 : \beta_2 = \dots = \beta_p = 0$ (with the degrees of freedom) and its corresponding P -value.

1.6 Analysis of residuals and checking of model assumptions

The residuals $r_i = Y_i - \hat{Y}_i$ can serve as an approximation of the unobservable error term ε_i and for checking whether the linear model is appropriate.

1.6.1 The Tukey-Anscombe Plot

The Tukey-Anscombe is a graphical tool: we plot the residuals r_i (on the y -axis) versus the fitted values \hat{Y}_i (on the x -axis). A reason to plot against the fitted values \hat{Y}_i is that the sample correlation between r_i and \hat{Y}_i is always zero.

In the ideal case, the points in the Tukey-Anscombe plot “fluctuate randomly” around the horizontal line through zero: see also Figure 1.4. An often encountered deviation is non-constant variability of the residuals, i.e., an indication that the variance of ε_i increases with the response variable Y_i : this is shown in Figure 1.5 a)–c). If the Tukey-Anscombe plot shows a trend, there is some evidence that the linear model assumption is not correct

(the expectation of the error is not zero which indicates a systematic error): Figure 1.5d) is a typical example.

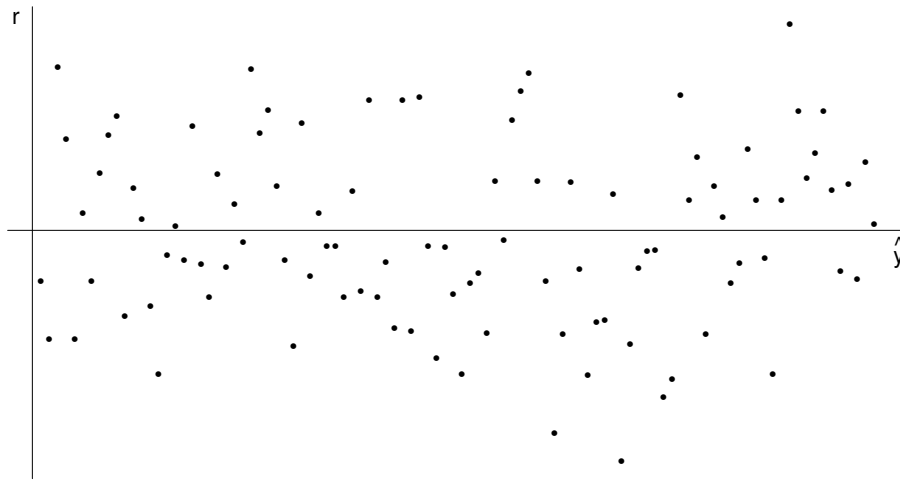


Figure 1.4: Ideal Tukey-Anscombe plot: no violations of model assumptions.

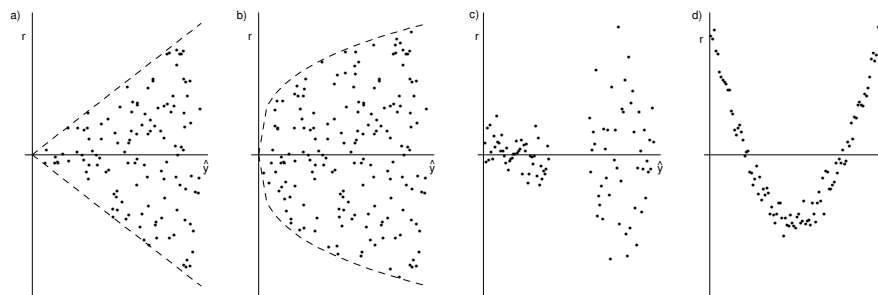


Figure 1.5: a) linear increase of standard deviation, b) nonlinear increase of standard deviation, c) 2 groups with different variances, d) missing quadratic term in the model.

In case where the Tukey-Anscombe plot exhibits a systematic relation of the variability on the fitted values \hat{Y}_i , we should either transform the response variable or perform a weighted regression (see Section 1.6.4). If the standard deviation grows linearly with the fitted values (as in Figure 1.5a)), the log-transform $Y \mapsto \log(Y)$ stabilizes the variance; if the standard deviation grows as the square root with the values \hat{Y}_i (as in Figure 1.5b)), the square root transformation $Y \mapsto \sqrt{Y}$ stabilizes the variance.

1.6.2 The Normal Plot

Assumptions for the distribution of random variables can be graphically checked with the QQ (quantile-quantile) plot. In the special case of checking for the normal distribution, the QQ plot is also referred to as a normal plot.

In the linear model application, we plot the empirical quantiles of the residuals (on the y axis) versus the theoretical quantiles of a $\mathcal{N}(0, 1)$ distribution (on the x axis). If the residuals would be normally distributed with expectation μ and variance σ^2 , the normal plot would exhibit an approximate straight line with intercept μ and slope σ . Figures 1.6 and 1.7 show some normal plots with exactly normally and non-normally distributed observations.

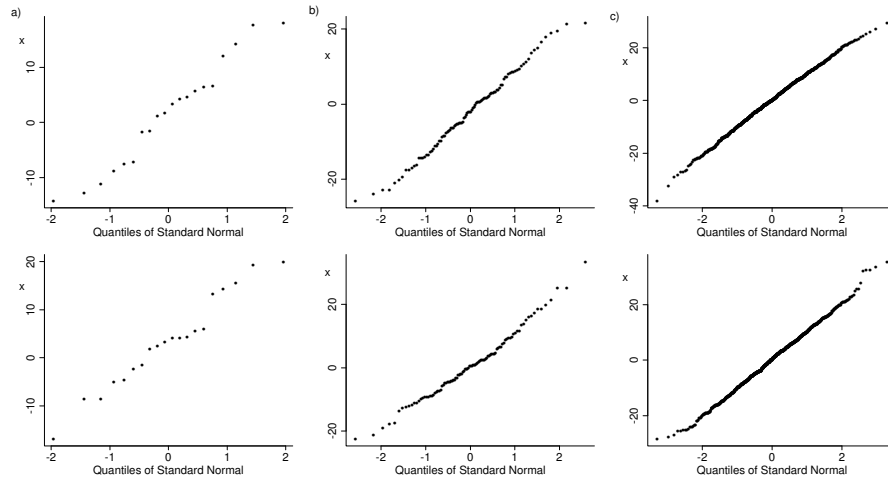


Figure 1.6: QQ-plots for i.i.d. normally distributed random variables. Two plots for each sample size n equal to a) 20, b) 100 and c) 1000.

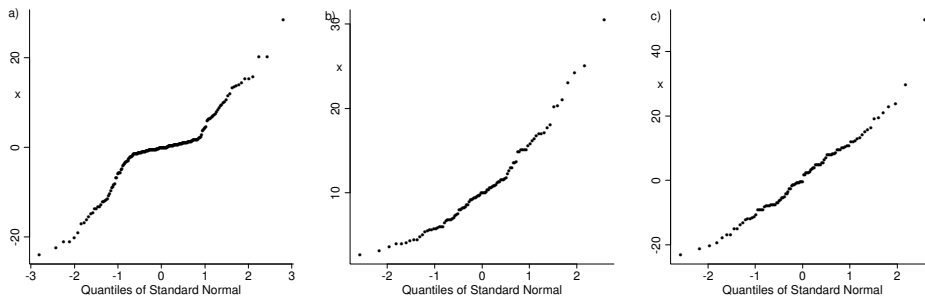


Figure 1.7: QQ-plots for a) long-tailed distribution, b) skewed distribution, c) dataset with outlier.

1.6.3 Plot for detecting serial correlation

For checking independence of the errors we plot the residuals r_i versus the observation number i (or if available, the time t_i of recording the i th observation). If the residuals vary randomly around the zero line, there are no indications for serial correlations among the errors ε_i . On the other hand, if neighboring (with respect to the x -axis) residuals look similar, the independence assumption for the errors seems violated.

1.6.4 Generalized least squares and weighted regression

In a more general situation, the errors are correlated with known covariance matrix,

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{with } \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}).$$

When $\boldsymbol{\Sigma}$ is known (and also in the case where $\boldsymbol{\Sigma} = \sigma^2 G$ with unknown σ^2), this case can be transformed to the i.i.d. one, using a “square root” C such that $\boldsymbol{\Sigma} = CC^\top$ (defined, e.g., via Cholesky factorization $\boldsymbol{\Sigma} = LL^\top$, L is uniquely determined lower triangular): If $\tilde{\mathbf{Y}} := C^{-1}\mathbf{Y}$ and $\tilde{X} := C^{-1}X$, we have $\tilde{\mathbf{Y}} = \tilde{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$, where $\tilde{\boldsymbol{\varepsilon}} \sim \mathcal{N}(\mathbf{0}, I)$. This leads to the *generalized least squares* solution $\hat{\boldsymbol{\beta}} = (X^\top \boldsymbol{\Sigma}^{-1} X)^{-1} X^\top \boldsymbol{\Sigma}^{-1} \mathbf{Y}$ with $\text{Cov}(\hat{\boldsymbol{\beta}}) = (X^\top \boldsymbol{\Sigma}^{-1} X)^{-1}$.

A special case where Σ is *diagonal*, $\Sigma = \sigma^2 \text{diag}(z_1, z_2, \dots, z_n)$ (with trivial inverse) is the *weighted* least squares problem $\min_{\beta} \sum_{i=1}^n w_i (Y_i - \mathbf{x}_i^\top \beta)^2$, with weights $w_i \equiv 1/z_i$.

1.7 Model Selection

We assume the linear model

$$Y_i = \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i \quad (i = 1, \dots, n),$$

where $\varepsilon_1, \dots, \varepsilon_n$ i.i.d., $\mathbb{E}[\varepsilon_i] = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$.

Problem: Which of the predictor variables should be used in the linear model? It may be that not all of the p predictor variables are relevant. In addition, every coefficient has to be estimated and thus is afflicted with variability: the individual variabilities for each coefficient sum up and **the variability of the estimated hyper-plane increases the more predictors are entered into the model, whether they are relevant or not**. The aim is often to look for the **optimal** or **best** - not the true - model.

What we just explained in words can be formalized a bit more. Suppose we are looking for optimizing the prediction

$$\sum_{r=1}^q \hat{\beta}_{j_r} x_{ij_r}$$

which includes q predictor variables with indices $j_1, \dots, j_q \in \{1, \dots, p\}$. The average mean squared error of this prediction is

$$\begin{aligned} & n^{-1} \sum_{i=1}^n \mathbb{E}[(m(\mathbf{x}_i) - \sum_{r=1}^q \hat{\beta}_{j_r} x_{ij_r})^2] \\ = & n^{-1} \sum_{i=1}^n (\mathbb{E}[\sum_{r=1}^q \hat{\beta}_{j_r} x_{ij_r}] - m(\mathbf{x}_i))^2 + \underbrace{n^{-1} \sum_{i=1}^n \text{Var}(\sum_{r=1}^q \hat{\beta}_{j_r} x_{ij_r})}_{= \frac{q}{n} \sigma^2}, \end{aligned} \quad (1.9)$$

where $m(\cdot)$ denotes the regression function in the full model with all the predictor variables. It is plausible that the systematic error (squared bias) $n^{-1} \sum_{i=1}^n (\mathbb{E}[\sum_{r=1}^q \hat{\beta}_{j_r} x_{ij_r}] - m(\mathbf{x}_i))^2$ decreases as the number of predictor variables q increases (i.e., with respect to bias, we have nothing to lose by using as many predictors as we can), but the variance term increases linearly in the number of predictors q (the variance term equals $q/n \cdot \sigma^2$ which is not too difficult to derive). This is the so-called **bias-variance trade-off** which is present in very many other situations and applications in statistics. Finding the best model thus means to optimize the bias-variance trade-off: this is sometimes also referred to as “regularization” (for avoiding a too complex model).

1.7.1 Mallows C_p statistic

The mean squared error in (1.9) is unknown: we do not know the magnitude of the bias term but fortunately, we can estimate the mean squared error.

Denote by $SSE(\mathcal{M})$ the residual sum of squares in a model \mathcal{M} : it is overly optimistic and not a good measure to estimate the mean squared error in (1.9). For example,

$SSE(\mathcal{M})$ becomes smaller the bigger the model \mathcal{M} and the biggest model under consideration has the lowest SSE (which generally contradicts the equation in (1.9)).

For any (sub-)model \mathcal{M} which involves some (or all) of the predictor variables, the mean squared error in (1.9) can be estimated by

$$n^{-1}SSE(\mathcal{M}) - \hat{\sigma}^2 + 2\hat{\sigma}^2|\mathcal{M}|/n,$$

where $\hat{\sigma}^2$ is the error variance estimate in the full model and $SSE(\mathcal{M})$ is the residual sum of squares in the submodel \mathcal{M} . (A justification can be found in the literature). Thus, in order to estimate the best model, we could search for the sub-model \mathcal{M} minimizing the above quantity. Since $\hat{\sigma}^2$ and n are constants with respect to submodels \mathcal{M} , we can also consider the well-known C_p statistic

$$C_p(\mathcal{M}) = \frac{SSE(\mathcal{M})}{\hat{\sigma}^2} - n + 2|\mathcal{M}|$$

and search for the sub-model \mathcal{M} minimizing the C_p statistic.

Other popular criteria to estimate the predictive potential of an estimated model are Akaike's information criterion (AIC) and the Bayesian information criterion (BIC).

Searching for the best model with respect to C_p

If the full model has p predictor variables, there are $2^p - 1$ sub-models (every predictor can be in or out but we exclude the sub-model \mathcal{M} which corresponds to the empty set).

Therefore, an exhaustive search for the sub-model \mathcal{M} minimizing the C_p statistic is only feasible if p is less than say 16 ($2^{16} - 1 = 65'535$ which is already fairly large). If p is "large", we can proceed with stepwise algorithms.

Forward selection.

1. Start with the smallest model \mathcal{M}_0 (location model) as the current model.
2. Include the predictor variable to the current model which reduces the residual sum of squares most.
3. Continue step 2. until all predictor variables have been chosen or until a large number of predictor variables has been selected. This produces a sequence of sub-models $\mathcal{M}_0 \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \dots$.
4. Choose the model in the sequence $\mathcal{M}_0 \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \dots$ which has smallest C_p statistic.

Backward selection.

1. Start with the full model \mathcal{M}_0 as the current model.
2. Exclude the predictor variable from the current model which increases the residual sum of squares the least.
3. Continue step 2. until all predictor variables have been deleted (or a large number of predictor variables). This produces a sequence of sub-models $\mathcal{M}_0 \supseteq \mathcal{M}_1 \supseteq \mathcal{M}_2 \supseteq \dots$.
4. Choose the model in the sequence $\mathcal{M}_0 \supseteq \mathcal{M}_1 \supseteq \mathcal{M}_2 \supseteq \dots$ which has smallest C_p statistic.

Backward selection is typically a bit better than forward selection but it is computationally more expensive. Also, in case where $p \geq n$, we don't want to fit the full model and forward selection is an appropriate way to proceed.