Chapter 3
Moving to Higher Dimensions

We have seen in the previous chapters how very simple graphical devices can help in understanding the structure and dependency of data. The graphical tools were based on either univariate (bivariate) data representations or on "slick" transformations of multivariate information perceivable by the human eye. Most of the tools are extremely useful in a modelling step, but unfortunately, do not give the full picture of the data set. One reason for this is that the graphical tools presented capture only certain dimensions of the data and do not necessarily concentrate on those dimensions or sub-parts of the data under analysis that carry the maximum structural information. In Part III of this book, powerful tools for reducing the dimension of a data set will be presented. In this chapter, as a starting point, simple and basic tools are used to describe dependency. They are constructed from elementary facts of probability theory and introductory statistics (for example, the covariance and correlation between two variables).

Sections 3.1 and 3.2 show how to handle these concepts in a multivariate setup and how a simple test on correlation between two variables can be derived. Since linear relationships are involved in these measures, Section 3.4 presents the simple linear model for two variables and recalls the basic t-test for the slope. In Section 3.5, a simple example of one-factorial analysis of variance introduces the notations for the well known F-test.

Due to the power of matrix notation, all of this can easily be extended to a more general multivariate setup. Section 3.3 shows how matrix operations can be used to define summary statistics of a data set and for obtaining the empirical moments of linear transformations of the data. These results will prove to be very useful in most of the chapters in Part III.

Finally, matrix notation allows us to introduce the flexible multiple linear model, where more general relationships among variables can be analysed. In Section 3.6, the least squares adjustment of the model and the usual test statistics are presented with their geometric interpretation. Using these notations, the ANOVA model is just a particular case of the multiple linear model.

W.K. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*,
DOI 10.1007/978-3-642-17229-8_3, © Springer-Verlag Berlin Heidelberg 2012

3.1 Covariance

Covariance is a measure of dependency between random variables. Given two (random) variables X and Y the (theoretical) covariance is defined by:

$$\sigma_{XY} = \mathsf{Cov}(X, Y) = \mathsf{E}(XY) - (\mathsf{E}\,X)(\mathsf{E}\,Y). \tag{3.1}$$

The precise definition of expected values is given in Chapter 4. If X and Y are independent of each other, the covariance $\mathsf{Cov}(X, Y)$ is necessarily equal to zero, see Theorem 3.1. The converse is not true. The covariance of X with itself is the variance:

$$\sigma_{XX} = \mathsf{Var}(X) = \mathsf{Cov}(X, X).$$

If the variable X is p-dimensional multivariate, e.g., $X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$, then the theoretical covariances among all the elements are put into matrix form, i.e., the covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_{X_1X_1} & \dots & \sigma_{X_1X_p} \\ \vdots & \ddots & \vdots \\ \sigma_{X_pX_1} & \dots & \sigma_{X_pX_p} \end{pmatrix}.$$

Properties of covariance matrices will be detailed in Chapter 4. Empirical versions of these quantities are:

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{3.2}$$

$$s_{XX} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{3.3}$$

For small n, say $n \le 20$, we should replace the factor $\frac{1}{n}$ in (3.2) and (3.3) by $\frac{1}{n-1}$ in order to correct for a small bias. For a p-dimensional random variable, one obtains the empirical covariance matrix (see Section 3.3 for properties and details)

$$\mathcal{S} = \begin{pmatrix} s_{X_1X_1} & \dots & s_{X_1X_p} \\ \vdots & \ddots & \vdots \\ s_{X_pX_1} & \dots & s_{X_pX_p} \end{pmatrix}.$$

For a scatterplot of two variables the covariances measure “how close the scatter is to a line”. Mathematical details follow but it should already be understood here that in this sense covariance measures only “linear dependence”.

Example 3.1 If $\mathcal{X}$ is the entire bank data set, one obtains the covariance matrix $\mathcal{S}$ as indicated below:

$$\mathcal{S} = \begin{pmatrix} 0.14 & 0.03 & 0.02 & -0.10 & -0.01 & 0.08 \\ 0.03 & 0.12 & 0.10 & 0.21 & 0.10 & -0.21 \\ 0.02 & 0.10 & 0.16 & 0.28 & 0.12 & -0.24 \\ -0.10 & 0.21 & 0.28 & 2.07 & 0.16 & -1.03 \\ -0.01 & 0.10 & 0.12 & 0.16 & 0.64 & -0.54 \\ 0.08 & -0.21 & -0.24 & -1.03 & -0.54 & 1.32 \end{pmatrix}. \tag{3.4}$$

The empirical covariance between X_4 and X_5, i.e., $s_{X_4X_5}$, is found in row 4 and column 5. The value is $s_{X_4X_5} = 0.16$. Is it obvious that this value is positive? In Exercise 3.1 we will discuss this question further.

If $\mathcal{X}_f$ denotes the counterfeit bank notes, we obtain:

$$\mathcal{S}_f = \begin{pmatrix} 0.123 & 0.031 & 0.023 & -0.099 & 0.019 & 0.011 \\ 0.031 & 0.064 & 0.046 & -0.024 & -0.012 & -0.005 \\ 0.024 & 0.046 & 0.088 & -0.018 & 0.000 & 0.034 \\ -0.099 & -0.024 & -0.018 & 1.268 & -0.485 & 0.236 \\ 0.019 & -0.012 & 0.000 & -0.485 & 0.400 & -0.022 \\ 0.011 & -0.005 & 0.034 & 0.236 & -0.022 & 0.308 \end{pmatrix}. \tag{3.5}$$

For the genuine $\mathcal{X}_g$, we have:

$$\mathcal{S}_g = \begin{pmatrix} 0.149 & 0.057 & 0.057 & 0.056 & 0.014 & 0.005 \\ 0.057 & 0.131 & 0.085 & 0.056 & 0.048 & -0.043 \\ 0.057 & 0.085 & 0.125 & 0.058 & 0.030 & -0.024 \\ 0.056 & 0.056 & 0.058 & 0.409 & -0.261 & -0.000 \\ 0.014 & 0.049 & 0.030 & -0.261 & 0.417 & -0.074 \\ 0.005 & -0.043 & -0.024 & -0.000 & -0.074 & 0.198 \end{pmatrix}. \tag{3.6}$$

Note that the covariance between X_4 (distance of the frame to the lower border) and X_5 (distance of the frame to the upper border) is negative in both (3.5) and (3.6). Why would this happen? In Exercise 3.2 we will discuss this question in more detail.

At first sight, the matrices $\mathcal{S}_f$ and $\mathcal{S}_g$ look different, but they create almost the same scatterplots (see the discussion in Section 1.4). Similarly, the common principal component analysis in Chapter 10 suggests a joint analysis of the covariance structure as in Flury and Riedwyl (1988).

Scatterplots with point clouds that are "upward-sloping", like the one in the upper left of Figure 1.14, show variables with positive covariance. Scatterplots with "downward-sloping" structure have negative covariance. In Figure 3.1 we show the scatterplot of X_4 vs. X_5 of the entire bank data set. The point cloud is upward-sloping. However, the two sub-clouds of counterfeit and genuine bank notes are downward-sloping.

Example 3.2 A textile shop manager is studying the sales of "classic blue" pullovers over 10 different periods. He observes the number of pullovers sold (X_1), variation

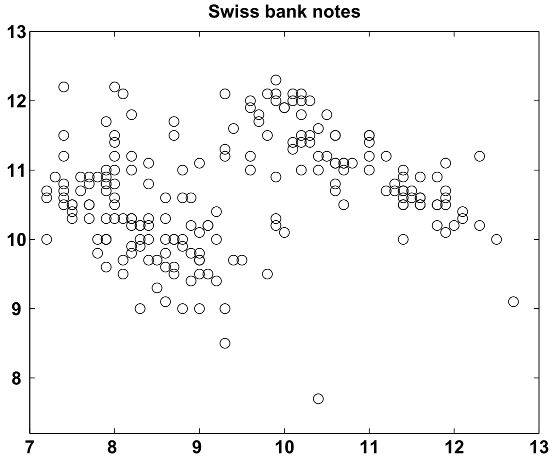


Fig. 3.1 Scatterplot of variables X_4 vs. X_5 of the entire bank data set MVAscabank45

in price (X_2, in EUR), the advertisement costs in local newspapers (X_3, in EUR) and the presence of a sales assistant (X_4, in hours per period). Over the periods, he observes the following data matrix:

$$\mathcal{X} = \begin{pmatrix} 230 & 125 & 200 & 109 \\ 181 & 99 & 55 & 107 \\ 165 & 97 & 105 & 98 \\ 150 & 115 & 85 & 71 \\ 97 & 120 & 0 & 82 \\ 192 & 100 & 150 & 103 \\ 181 & 80 & 85 & 111 \\ 189 & 90 & 120 & 93 \\ 172 & 95 & 110 & 86 \\ 170 & 125 & 130 & 78 \end{pmatrix}.$$

He is convinced that the price must have a large influence on the number of pullovers sold. So he makes a scatterplot of X_2 vs. X_1, see Figure 3.2. A rough impression is that the cloud is somewhat downward-sloping. A computation of the empirical covariance yields

$$s_{X_1X_2} = \frac{1}{9}\sum_{i=1}^{10}\left(X_{1i} - \bar{X}_1\right)\left(X_{2i} - \bar{X}_2\right) = -80.02,$$

a negative value as expected.

Note: The covariance function is scale dependent. Thus, if the prices in this example were in Japanese Yen (JPY), we would obtain a different answer (see Exercise 3.16). A measure of (linear) dependence independent of the scale is the correlation, which we introduce in the next section.

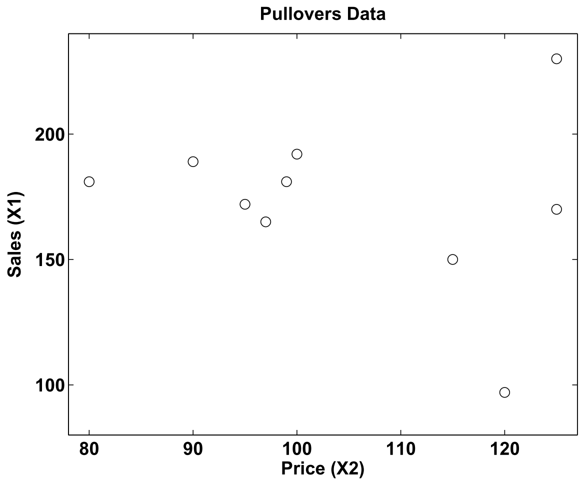
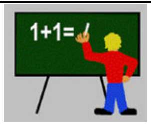


Fig. 3.2 Scatterplot of variables X_2 vs. X_1 of the pullovers data set `MVAscapull1`



Summary

↪ The covariance is a measure of dependence.

↪ Covariance measures only linear dependence.

↪ Covariance is scale dependent.

↪ There are nonlinear dependencies that have zero covariance.

↪ Zero covariance does not imply independence.

↪ Independence implies zero covariance.

↪ Negative covariance corresponds to downward-sloping scatterplots.

↪ Positive covariance corresponds to upward-sloping scatterplots.

↪ The covariance of a variable with itself is its variance $\mathsf{Cov}(X, X) = \sigma_{XX} = \sigma_X^2$.

↪ For small n, we should replace the factor $\frac{1}{n}$ in the computation of the covariance by $\frac{1}{n-1}$.

3.2 Correlation

The correlation between two variables X and Y is defined from the covariance as the following:

$$\rho_{XY} = \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)}}. \tag{3.7}$$

The advantage of the correlation is that it is independent of the scale, i.e., changing the variables' scale of measurement does not change the value of the correlation. Therefore, the correlation is more useful as a measure of association between two random variables than the covariance. The empirical version of ρ_{XY} is as follows:

$$r_{XY} = \frac{s_{XY}}{\sqrt{s_{XX}s_{YY}}}. \tag{3.8}$$

The correlation is in absolute value always less than 1. It is zero if the covariance is zero and vice-versa. For p-dimensional vectors $(X_1,\ldots,X_p)^\top$ we have the theoretical correlation matrix

$$\mathcal{P} = \begin{pmatrix} \rho_{X_1X_1} & \cdots & \rho_{X_1X_p} \\ \vdots & \ddots & \vdots \\ \rho_{X_pX_1} & \cdots & \rho_{X_pX_p} \end{pmatrix},$$

and its empirical version, the empirical correlation matrix which can be calculated from the observations,

$$\mathcal{R} = \begin{pmatrix} r_{X_1X_1} & \cdots & r_{X_1X_p} \\ \vdots & \ddots & \vdots \\ r_{X_pX_1} & \cdots & r_{X_pX_p} \end{pmatrix}.$$

Example 3.3 We obtain the following correlation matrix for the genuine bank notes:

$$\mathcal{R}_g = \begin{pmatrix} 1.00 & 0.41 & 0.41 & 0.22 & 0.05 & 0.03 \\ 0.41 & 1.00 & 0.66 & 0.24 & 0.20 & -0.25 \\ 0.41 & 0.66 & 1.00 & 0.25 & 0.13 & -0.14 \\ 0.22 & 0.24 & 0.25 & 1.00 & -0.63 & -0.00 \\ 0.05 & 0.20 & 0.13 & -0.63 & 1.00 & -0.25 \\ 0.03 & -0.25 & -0.14 & -0.00 & -0.25 & 1.00 \end{pmatrix}, \tag{3.9}$$

and for the counterfeit bank notes:

$$\mathcal{R}_f = \begin{pmatrix} 1.00 & 0.35 & 0.24 & -0.25 & 0.08 & 0.06 \\ 0.35 & 1.00 & 0.61 & -0.08 & -0.07 & -0.03 \\ 0.24 & 0.61 & 1.00 & -0.05 & 0.00 & 0.20 \\ -0.25 & -0.08 & -0.05 & 1.00 & -0.68 & 0.37 \\ 0.08 & -0.07 & 0.00 & -0.68 & 1.00 & -0.06 \\ 0.06 & -0.03 & 0.20 & 0.37 & -0.06 & 1.00 \end{pmatrix}. \tag{3.10}$$

As noted before for $\mathsf{Cov}(X_4, X_5)$, the correlation between X_4 (distance of the frame to the lower border) and X_5 (distance of the frame to the upper border) is negative. This is natural, since the covariance and correlation always have the same sign (see also Exercise 3.17).

Why is the correlation an interesting statistic to study? It is related to independence of random variables, which we shall define more formally later on. For the moment we may think of independence as the fact that one variable has no influence on another.

Theorem 3.1 *If X and Y are independent, then* $\rho(X, Y) = \mathsf{Cov}(X, Y) = 0$.

⚠ In general, the converse is not true, as the following example shows.

Example 3.4 Consider a standard normally-distributed random variable X and a random variable $Y = X^2$, which is surely not independent of X. Here we have

$$\mathsf{Cov}(X, Y) = \mathsf{E}(XY) - \mathsf{E}(X)\,\mathsf{E}(Y) = \mathsf{E}(X^3) = 0$$

(because $\mathsf{E}(X) = 0$ and $\mathsf{E}(X^2) = 1$). Therefore $\rho(X, Y) = 0$, as well. This example also shows that correlations and covariances measure only linear dependence. The quadratic dependence of $Y = X^2$ on X is not reflected by these measures of dependence.

Remark 3.1 For two normal random variables, the converse of Theorem 3.1 is true: zero covariance for two normally-distributed random variables implies independence. This will be shown later in Corollary 5.2.

Theorem 3.1 enables us to check for independence between the components of a bivariate normal random variable. That is, we can use the correlation and test whether it is zero. The distribution of r_{XY} for an arbitrary (X, Y) is unfortunately complicated. The distribution of r_{XY} will be more accessible if (X, Y) are jointly normal (see Chapter 5). If we transform the correlation by Fisher's Z-transformation,

$$W = \frac{1}{2}\log\left(\frac{1 + r_{XY}}{1 - r_{XY}}\right), \tag{3.11}$$

we obtain a variable that has a more accessible distribution. Under the hypothesis that $\rho = 0$, W has an asymptotic normal distribution. Approximations of the expectation and variance of W are given by the following:

$$\begin{aligned} \mathsf{E}(W) &\approx \frac{1}{2}\log\left(\frac{1 + \rho_{XY}}{1 - \rho_{XY}}\right) \\ \mathsf{Var}(W) &\approx \frac{1}{(n-3)}\cdot \end{aligned} \tag{3.12}$$

The distribution is given in Theorem 3.2.

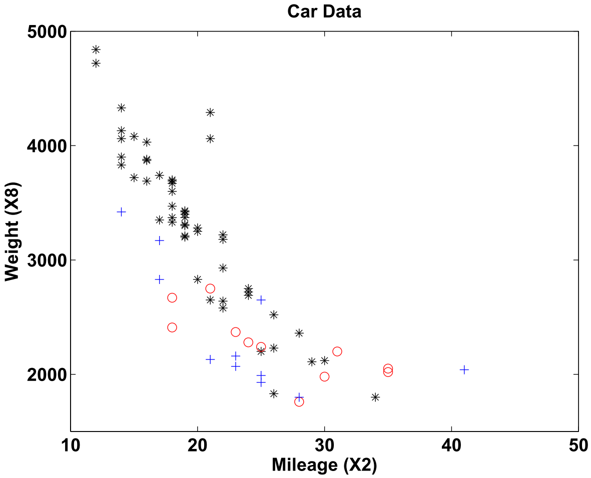


Fig. 3.3 Mileage (X_2) vs. weight (X_8) of U.S. (star), European (plus signs) and Japanese (circle) cars `MVAscacar`

Theorem 3.2

$$Z = \frac{W - \mathsf{E}(W)}{\sqrt{\mathsf{Var}(W)}} \xrightarrow{\mathcal{L}} N(0,1). \tag{3.13}$$

The symbol "$\xrightarrow{\mathcal{L}}$" denotes convergence in distribution, which will be explained in more detail in Chapter 4.

Theorem 3.2 allows us to test different hypotheses on correlation. We can fix the level of significance α (the probability of rejecting a true hypothesis) and reject the hypothesis if the difference between the hypothetical value and the calculated value of Z is greater than the corresponding critical value of the normal distribution. The following example illustrates the procedure.

Example 3.5 Let's study the correlation between mileage (X_2) and weight (X_8) for the car data set (B.3) where $n = 74$. We have $r_{X_2X_8} = -0.823$. Our conclusions from the boxplot in Figure 1.3 ("Japanese cars generally have better mileage than the others") needs to be revised. From Figure 3.3 and $r_{X_2X_8}$, we can see that mileage is highly correlated with weight, and that the Japanese cars in the sample are in fact all lighter than the others.

If we want to know whether $\rho_{X_2X_8}$ is significantly different from $\rho_0 = 0$, we apply Fisher's Z-transform (3.11). This gives us

$$w = \frac{1}{2}\log\left(\frac{1 + r_{X_2X_8}}{1 - r_{X_2X_8}}\right) = -1.166 \quad \text{and} \quad z = \frac{-1.166 - 0}{\sqrt{\frac{1}{71}}} = -9.825,$$

i.e., a highly significant value to reject the hypothesis that $\rho = 0$ (the 2.5% and 97.5% quantiles of the normal distribution are -1.96 and 1.96, respectively). If we

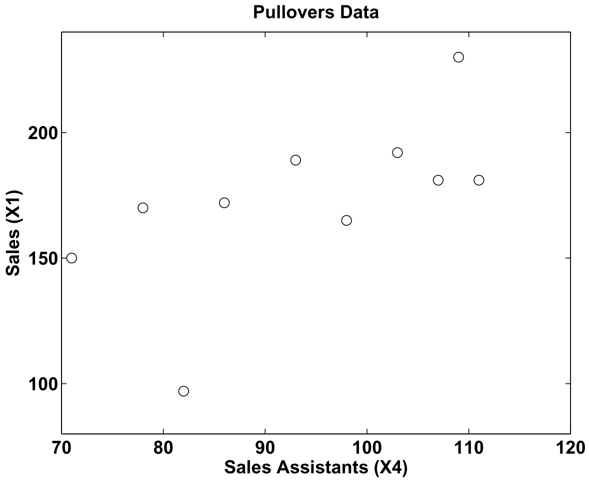


Fig. 3.4 Hours of sales assistants (X_4) vs. sales (X_1) of pullovers MVAscapull2

want to test the hypothesis that, say, $\rho_0 = -0.75$, we obtain:

$$z = \frac{-1.166 - (-0.973)}{\sqrt{\frac{1}{71}}} = -1.627.$$

This is a non-significant value at the $\alpha = 0.05$ level for z since it is between the critical values at the 5% significance level (i.e., $-1.96 < z < 1.96$).

Example 3.6 Let us consider again the pullovers data set from Example 3.2. Consider the correlation between the presence of the sales assistants (X_4) vs. the number of sold pullovers (X_1) (see Figure 3.4). Here we compute the correlation as

$$r_{X_1X_4} = 0.633.$$

The Z-transform of this value is

$$w = \frac{1}{2}\log_e\left(\frac{1 + r_{X_1X_4}}{1 - r_{X_1X_4}}\right) = 0.746. \tag{3.14}$$

The sample size is $n = 10$, so for the hypothesis $\rho_{X_1X_4} = 0$, the statistic to consider is:

$$z = \sqrt{7}(0.746 - 0) = 1.974 \tag{3.15}$$

which is just statistically significant at the 5% level (i.e., 1.974 is just a little larger than 1.96).

Remark 3.2 The normalising and variance stabilising properties of W are asymptotic. In addition the use of W in small samples (for $n \leq 25$) is improved by Hotelling's transform (Hotelling, 1953):

$$W^* = W - \frac{3W + \tanh(W)}{4(n-1)} \quad \text{with} \quad Var(W^*) = \frac{1}{n-1}.$$

The transformed variable W^* is asymptotically distributed as a normal distribution.

Example 3.7 From the preceding remark, we obtain $w^* = 0.6663$ and $\sqrt{10-1}w^* = 1.9989$ for the preceding Example 3.6. This value is significant at the 5% level.

Remark 3.3 Note that the Fisher's Z-transform is the inverse of the hyperbolic tangent function: $W = \tanh^{-1}(r_{XY})$; equivalently $r_{XY} = \tanh(W) = \frac{e^{2W}-1}{e^{2W}+1}$.

Remark 3.4 Under the assumptions of normality of X and Y, we may test their independence ($\rho_{XY} = 0$) using the exact t-distribution of the statistic

$$T = r_{XY}\sqrt{\frac{n-2}{1-r_{XY}^2}} \overset{\rho_{XY}=0}{\sim} t_{n-2}.$$

Setting the probability of the first error type to α, we reject the null hypothesis $\rho_{XY} = 0$ if $|T| \geq t_{1-\alpha/2;n-2}$.

Summary

↪ The correlation is a standardised measure of dependence.

↪ The absolute value of the correlation is always less than one.

↪ Correlation measures only linear dependence.

↪ There are nonlinear dependencies that have zero correlation.

↪ Zero correlation does not imply independence.

↪ Independence implies zero correlation.

↪ Negative correlation corresponds to downward-sloping scatterplots.

↪ Positive correlation corresponds to upward-sloping scatterplots.

↪ Fisher's Z-transform helps us in testing hypotheses on correlation.

↪ For small samples, Fisher's Z-transform can be improved by the transformation $W^* = W - \frac{3W+\tanh(W)}{4(n-1)}$.

3.3 Summary Statistics

This section focuses on the representation of basic summary statistics (means, covariances and correlations) in matrix notation, since we often apply linear transformations to data. The matrix notation allows us to derive instantaneously the corresponding characteristics of the transformed variables. The Mahalanobis transformation is a prominent example of such linear transformations.

Assume that we have observed n realisations of a p-dimensional random variable; we have a data matrix $\mathcal{X}(n \times p)$:

$$\mathcal{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}. \tag{3.16}$$

The rows $x_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ denote the i-th observation of a p-dimensional random variable $X \in \mathbb{R}^p$.

The statistics that were briefly introduced in Sections 3.1 and 3.2 can be rewritten in matrix form as follows. The "centre of gravity" of the n observations in $\mathbb{R}^p$ is given by the vector $\overline{x}$ of the means $\overline{x}_j$ of the p variables:

$$\overline{x} = \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{pmatrix} = n^{-1}\mathcal{X}^\top 1_n. \tag{3.17}$$

The dispersion of the n observations can be characterised by the covariance matrix of the p variables. The empirical covariances defined in (3.2) and (3.3) are the elements of the following matrix:

$$\mathcal{S} = n^{-1}\mathcal{X}^\top\mathcal{X} - \overline{x}\,\overline{x}^\top = n^{-1}(\mathcal{X}^\top\mathcal{X} - n^{-1}\mathcal{X}^\top 1_n 1_n^\top \mathcal{X}). \tag{3.18}$$

Note that this matrix is equivalently defined by

$$\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^\top.$$

The covariance formula (3.18) can be rewritten as $\mathcal{S} = n^{-1}\mathcal{X}^\top\mathcal{H}\mathcal{X}$ with the *centring matrix*

$$\mathcal{H} = \mathcal{I}_n - n^{-1}1_n 1_n^\top. \tag{3.19}$$

Note that the centring matrix is symmetric and idempotent. Indeed,

$$\begin{aligned} \mathcal{H}^2 &= (\mathcal{I}_n - n^{-1}1_n 1_n^\top)(\mathcal{I}_n - n^{-1}1_n 1_n^\top) \\ &= \mathcal{I}_n - n^{-1}1_n 1_n^\top - n^{-1}1_n 1_n^\top + (n^{-1}1_n 1_n^\top)(n^{-1}1_n 1_n^\top) \\ &= \mathcal{I}_n - n^{-1}1_n 1_n^\top = \mathcal{H}. \end{aligned}$$

As a consequence $\mathcal{S}$ is positive semidefinite, i.e.

$$\mathcal{S} \geq 0. \tag{3.20}$$

Indeed for all $a \in \mathbb{R}^p$,

$$\begin{aligned} a^\top \mathcal{S} a &= n^{-1} a^\top \mathcal{X}^\top \mathcal{H} \mathcal{X} a \\ &= n^{-1} (a^\top \mathcal{X}^\top \mathcal{H}^\top)(\mathcal{H}\mathcal{X}a) \quad \text{since } \mathcal{H}^\top \mathcal{H} = \mathcal{H}, \\ &= n^{-1} y^\top y = n^{-1} \sum_{j=1}^{p} y_j^2 \geq 0 \end{aligned}$$

for $y = \mathcal{H}\mathcal{X}a$. It is well known from the one-dimensional case that $n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ as an estimate of the variance exhibits a bias of the order n^{-1} (Breiman, 1973). In the multi-dimensional case, $\mathcal{S}_u = \frac{n}{n-1}\mathcal{S}$ is an unbiased estimate of the true covariance. (This will be shown in Example 4.15.)

The sample correlation coefficient between the i-th and j-th variables is $r_{X_i X_j}$, see (3.8). If $\mathcal{D} = \operatorname{diag}(s_{X_i X_i})$, then the correlation matrix is

$$\mathcal{R} = \mathcal{D}^{-1/2} \mathcal{S} \mathcal{D}^{-1/2}, \tag{3.21}$$

where $\mathcal{D}^{-1/2}$ is a diagonal matrix with elements $(s_{X_i X_i})^{-1/2}$ on its main diagonal.

Example 3.8 The empirical covariances are calculated for the pullover data set.

The vector of the means of the four variables in the dataset is $\overline{x} = (172.7, 104.6, 104.0, 93.8)^\top$.

The sample covariance matrix is $\mathcal{S} = \begin{pmatrix} 1037.2 & -80.2 & 1430.7 & 271.4 \\ -80.2 & 219.8 & 92.1 & -91.6 \\ 1430.7 & 92.1 & 2624 & 210.3 \\ 271.4 & -91.6 & 210.3 & 177.4 \end{pmatrix}$.

The unbiased estimate of the variance ($n = 10$) is equal to

$$\mathcal{S}_u = \frac{10}{9}\mathcal{S} = \begin{pmatrix} 1152.5 & -88.9 & 1589.7 & 301.6 \\ -88.9 & 244.3 & 102.3 & -101.8 \\ 1589.7 & 102.3 & 2915.6 & 233.7 \\ 301.6 & -101.8 & 233.7 & 197.1 \end{pmatrix}.$$

The sample correlation matrix is $\mathcal{R} = \begin{pmatrix} 1 & -0.17 & 0.87 & 0.63 \\ -0.17 & 1 & 0.12 & -0.46 \\ 0.87 & 0.12 & 1 & 0.31 \\ 0.63 & -0.46 & 0.31 & 1 \end{pmatrix}$.

Linear Transformation

In many practical applications we need to study linear transformations of the original data. This motivates the question of how to calculate summary statistics after such linear transformations.

Let $\mathcal{A}$ be a $(q \times p)$ matrix and consider the transformed data matrix

$$\mathcal{Y} = \mathcal{X}\mathcal{A}^\top = (y_1, \dots, y_n)^\top. \tag{3.22}$$

The row $y_i = (y_{i1}, \dots, y_{iq}) \in \mathbb{R}^q$ can be viewed as the i-th observation of a q-dimensional random variable $Y = \mathcal{A}X$. In fact we have $y_i = x_i \mathcal{A}^\top$. We immediately obtain the mean and the empirical covariance of the variables (columns) forming the data matrix $\mathcal{Y}$:

$$\overline{y} = \frac{1}{n}\mathcal{Y}^\top 1_n = \frac{1}{n}\mathcal{A}\mathcal{X}^\top 1_n = \mathcal{A}\overline{x} \tag{3.23}$$

$$\mathcal{S}_\mathcal{Y} = \frac{1}{n}\mathcal{Y}^\top \mathcal{H}\mathcal{Y} = \frac{1}{n}\mathcal{A}\mathcal{X}^\top \mathcal{H}\mathcal{X}\mathcal{A}^\top = \mathcal{A}\mathcal{S}_\mathcal{X}\mathcal{A}^\top. \tag{3.24}$$

Note that if the linear transformation is non-homogeneous, i.e.,

$$y_i = \mathcal{A}x_i + b \quad \text{where } b(q \times 1),$$

only (3.23) changes: $\overline{y} = \mathcal{A}\overline{x} + b$. The formulas (3.23) and (3.24) are useful in the particular case of $q = 1$, i.e., $y = \mathcal{X}a$, i.e. $y_i = a^\top x_i; i = 1, \dots, n$:

$$\overline{y} = a^\top \overline{x}$$
$$\mathcal{S}_y = a^\top \mathcal{S}_\mathcal{X} a.$$

Example 3.9 Suppose that $\mathcal{X}$ is the pullover data set. The manager wants to compute his mean expenses for advertisement (X_3) and sales assistant (X_4).

Suppose that the sales assistant charges an hourly wage of 10 EUR. Then the shop manager calculates the expenses Y as $Y = X_3 + 10X_4$. Formula (3.22) says that this is equivalent to defining the matrix $\mathcal{A}(4 \times 1)$ as:

$$\mathcal{A} = (0, 0, 1, 10).$$

Using formulas (3.23) and (3.24), it is now computationally very easy to obtain the sample mean $\overline{y}$ and the sample variance $\mathcal{S}_y$ of the overall expenses:

$$\overline{y} = \mathcal{A}\overline{x} = (0, 0, 1, 10)\begin{pmatrix} 172.7 \\ 104.6 \\ 104.0 \\ 93.8 \end{pmatrix} = 1042.0$$

$$\mathcal{S}_\mathcal{Y} = \mathcal{A}\mathcal{S}_\mathcal{X}\mathcal{A}^\top = (0, 0, 1, 10)\begin{pmatrix} 1152.5 & -88.9 & 1589.7 & 301.6 \\ -88.9 & 244.3 & 102.3 & -101.8 \\ 1589.7 & 102.3 & 2915.6 & 233.7 \\ 301.6 & -101.8 & 233.7 & 197.1 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 1 \\ 10 \end{pmatrix}$$
$$= 2915.6 + 4674 + 19710 = 27299.6.$$

Mahalanobis Transformation

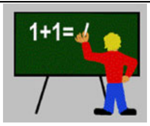
A special case of this linear transformation is

$$z_i = \mathcal{S}^{-1/2}(x_i - \overline{x}), \quad i = 1, \ldots, n. \tag{3.25}$$

Note that for the transformed data matrix $\mathcal{Z} = (z_1, \ldots, z_n)^\top$,

$$\mathcal{S}_\mathcal{Z} = n^{-1}\mathcal{Z}^\top \mathcal{H} \mathcal{Z} = \mathcal{I}_p. \tag{3.26}$$

So the Mahalanobis transformation eliminates the correlation between the variables and standardises the variance of each variable. If we apply (3.24) using $\mathcal{A} = \mathcal{S}^{-1/2}$, we obtain the identity covariance matrix as indicated in (3.26).



Summary

- ↪ The centre of gravity of a data matrix is given by its mean vector $\overline{x} = n^{-1}\mathcal{X}^\top 1_n$.
- ↪ The dispersion of the observations in a data matrix is given by the empirical covariance matrix $\mathcal{S} = n^{-1}\mathcal{X}^\top \mathcal{H} \mathcal{X}$.
- ↪ The empirical correlation matrix is given by $\mathcal{R} = \mathcal{D}^{-1/2}\mathcal{S}\mathcal{D}^{-1/2}$.
- ↪ A linear transformation $\mathcal{Y} = \mathcal{X}\mathcal{A}^\top$ of a data matrix $\mathcal{X}$ has mean $\mathcal{A}\overline{x}$ and empirical covariance $\mathcal{A}\mathcal{S}_\mathcal{X}\mathcal{A}^\top$.
- ↪ The Mahalanobis transformation is a linear transformation $z_i = \mathcal{S}^{-1/2}(x_i - \overline{x})$ which gives a standardised, uncorrelated data matrix $\mathcal{Z}$.

3.4 Linear Model for Two Variables

We have looked several times now at downward and upward-sloping scatterplots. What does the eye define here as a slope? Suppose that we can construct a line corresponding to the general direction of the cloud. The sign of the slope of this line would correspond to the upward and downward directions. Call the variable on the vertical axis Y and the one on the horizontal axis X. A slope line is a linear relationship between X and Y:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n. \tag{3.27}$$

Here, α is the intercept and β is the slope of the line. The errors (or deviations from the line) are denoted as ε_i and are assumed to have zero mean and finite variance σ^2. The task of finding (α, β) in (3.27) is referred to as a linear adjustment.

In Section 3.6 we shall derive estimators for α and β more formally, as well as accurately describe what a "good" estimator is. For now, one may try to find a "good" estimator $(\widehat{\alpha}, \widehat{\beta})$ via graphical techniques. A very common numerical and statistical technique is to use those $\widehat{\alpha}$ and $\widehat{\beta}$ that minimise:

$$(\widehat{\alpha}, \widehat{\beta}) = \arg \min_{(\alpha,\beta)} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2. \tag{3.28}$$

The solution to this task are the estimators:

$$\widehat{\beta} = \frac{s_{XY}}{s_{XX}} \tag{3.29}$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}. \tag{3.30}$$

The variance of $\widehat{\beta}$ is:

$$\mathsf{Var}(\widehat{\beta}) = \frac{\sigma^2}{n \cdot s_{XX}}. \tag{3.31}$$

The standard error (SE) of the estimator is the square root of (3.31),

$$SE(\widehat{\beta}) = \{\mathsf{Var}(\widehat{\beta})\}^{1/2} = \frac{\sigma}{(n \cdot s_{XX})^{1/2}}. \tag{3.32}$$

We can use this formula to test the hypothesis that $\beta = 0$. In an application the variance σ^2 has to be estimated by an estimator $\widehat{\sigma}^2$ that will be given below. Under a normality assumption of the errors, the t-test for the hypothesis $\beta = 0$ works as follows.

One computes the statistic

$$t = \frac{\widehat{\beta}}{SE(\widehat{\beta})} \tag{3.33}$$

and rejects the hypothesis at a 5% significance level if $|t| \geq t_{0.975;n-2}$, where the 97.5% quantile of the Student's t_{n-2} distribution is clearly the 95% critical value for the two-sided test. For $n \geq 30$, this can be replaced by 1.96, the 97.5% quantile of the normal distribution. An estimator $\widehat{\sigma}^2$ of σ^2 will be given in the following.

Example 3.10 Let us apply the linear regression model (3.27) to the "classic blue" pullovers. The sales manager believes that there is a strong dependence on the number of sales as a function of price. He computes the regression line as shown in Figure 3.5.

How good is this fit? This can be judged via goodness-of-fit measures. Define

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i, \tag{3.34}$$

as the predicted value of y as a function of x. With $\widehat{y}$ the textile shop manager in the above example can predict sales as a function of prices x. The variation in the response variable is:

$$n s_{YY} = \sum_{i=1}^{n} (y_i - \overline{y})^2. \tag{3.35}$$

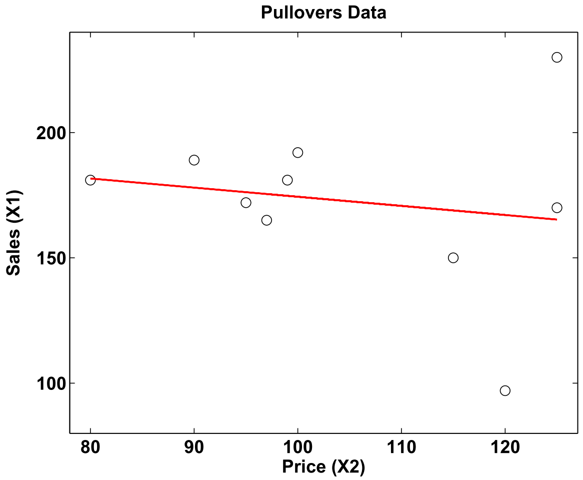


Fig. 3.5 Regression of sales (X_1) on price (X_2) of pullovers MVAregpull

The variation explained by the linear regression (3.27) with the predicted values (3.34) is:

$$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2. \tag{3.36}$$

The residual sum of squares, the minimum in (3.28), is given by:

$$RSS = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2. \tag{3.37}$$

An unbiased estimator $\widehat{\sigma}^2$ of σ^2 is given by $RSS/(n-2)$.

The following relation holds between (3.35)–(3.37):

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2, \tag{3.38}$$
$$\textit{Total variation} = \textit{Explained variation} + \textit{Unexplained variation}.$$

The *coefficient of determination* is r^2:

$$r^2 = \frac{\sum\limits_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\textit{explained variation}}{\textit{total variation}}. \tag{3.39}$$

The coefficient of determination increases with the proportion of explained variation by the linear relation (3.27). In the extreme cases where $r^2 = 1$, all of the variation is explained by the linear regression (3.27). The other extreme, $r^2 = 0$, is where the

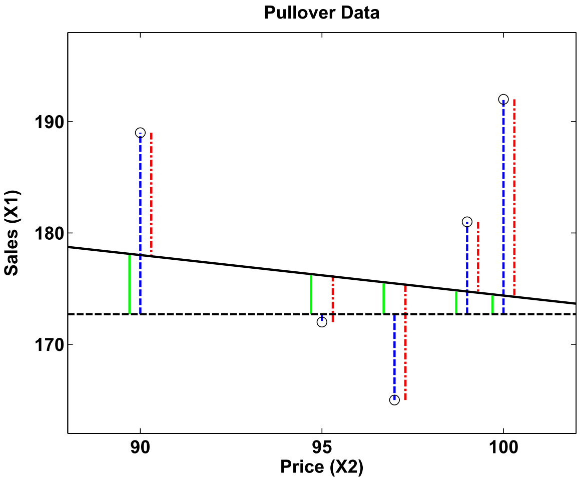


Fig. 3.6 Regression of sales (X_1) on price (X_2) of pullovers. The overall mean is given by the dashed line MVAregzoom

empirical covariance is $s_{XY} = 0$. The coefficient of determination can be rewritten as

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}. \tag{3.40}$$

From (3.39), it can be seen that in the linear regression (3.27), $r^2 = r^2_{XY}$ is the square of the correlation between X and Y.

Example 3.11 For the above pullover example, we estimate

$$\widehat{\alpha} = 210.774 \quad \text{and} \quad \widehat{\beta} = -0.364.$$

The coefficient of determination is

$$r^2 = 0.028.$$

The textile shop manager concludes that sales are not influenced very much by the price (in a linear way).

The geometrical representation of formula (3.38) can be graphically evaluated using Figure 3.6. This plot shows a section of the linear regression of the "sales" on "price" for the pullovers data. The distance between any point and the overall mean is given by the distance between the point and the regression line and the distance between the regression line and the mean. The sums of these two distances represent the total variance (solid blue lines from the observations to the overall mean), i.e., the explained variance (distance from the regression curve to the mean) and the unexplained variance (distance from the observation to the regression line), respectively.

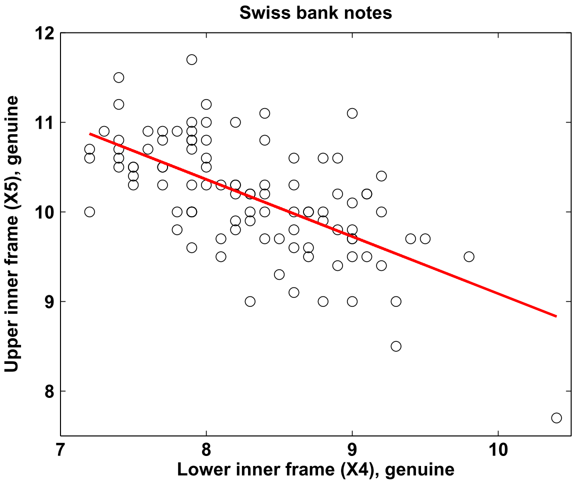


Fig. 3.7 Regression of X_5 (upper inner frame) on X_4 (lower inner frame) for genuine bank notes MVAregbank

In general the regression of Y on X is different from that of X on Y. We will demonstrate this, once again, using the Swiss bank notes data.

Example 3.12 The least squares fit of the variables X_4 (X) and X_5 (Y) from the genuine bank notes are calculated. Figure 3.7 shows the fitted line if X_5 is approximated by a linear function of X_4. In this case the parameters are

$$\widehat{\alpha} = 15.464 \quad \text{and} \quad \widehat{\beta} = -0.638.$$

If we predict X_4 by a function of X_5 instead, we would arrive at a different intercept and slope

$$\widehat{\alpha} = 14.666 \quad \text{and} \quad \widehat{\beta} = -0.626.$$

The linear regression of Y on X is given by minimising (3.28), i.e., the vertical errors ε_i. The linear regression of X on Y does the same but here the errors to be minimised in the least squares sense are measured horizontally. As seen in Example 3.12, the two least squares lines are different although both measure (in a certain sense) the slope of the cloud of points.

As shown in the next example, there is still one other way to measure the main direction of a cloud of points: it is related to the spectral decomposition of covariance matrices.

Example 3.13 Suppose that we have the following covariance matrix:

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Figure 3.8 shows a scatterplot of a sample of two normal random variables with such a covariance matrix (with $\rho = 0.8$).

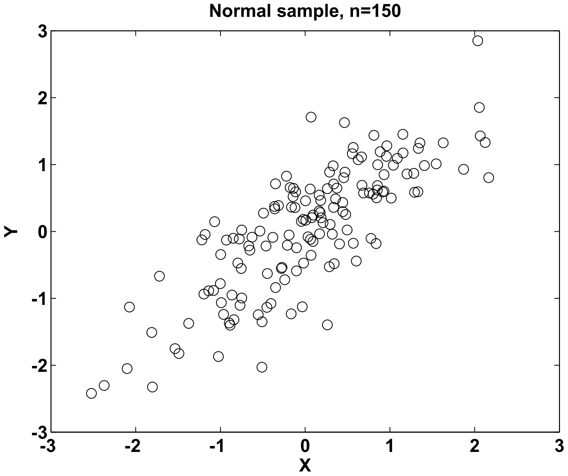


Fig. 3.8 Scatterplot for a sample of two correlated normal random variables (sample size $n = 150$, $\rho = 0.8$) MVAcorrnorm

The eigenvalues of Σ are, as was shown in Example 2.4, solutions to:

$$\begin{vmatrix} 1-\lambda & \rho \\ \rho & 1-\lambda \end{vmatrix} = 0.$$

Hence, $\lambda_1 = 1 + \rho$ and $\lambda_2 = 1 - \rho$. Therefore $\Lambda = \operatorname{diag}(1 + \rho, 1 - \rho)$. The eigenvector corresponding to $\lambda_1 = 1 + \rho$ can be computed from the system of linear equations:

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (1 + \rho) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

or

$$\begin{aligned} x_1 + \rho x_2 &= x_1 + \rho x_1 \\ \rho x_1 + x_2 &= x_2 + \rho x_2 \end{aligned}$$

and thus

$$x_1 = x_2.$$

The first (standardised) eigenvector is

$$\gamma_1 = \begin{pmatrix} 1\big/\sqrt{2} \\ 1\big/\sqrt{2} \end{pmatrix}.$$

The direction of this eigenvector is the diagonal in Figure 3.8 and captures the main variation in this direction. We shall come back to this interpretation in Chapter 10. The second eigenvector (orthogonal to γ_1) is

$$\gamma_2 = \begin{pmatrix} 1\big/\sqrt{2} \\ -1\big/\sqrt{2} \end{pmatrix}.$$

So finally

$$\Gamma = (\gamma_1, \gamma_2) = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

and we can check our calculation by

$$\Sigma = \Gamma \; \Lambda \; \Gamma^{\top}.$$

The first eigenvector captures the main direction of a point cloud. The linear regression of Y on X and X on Y accomplished, in a sense, the same thing. In general the direction of the eigenvector and the least squares slope are different. The reason is that the least squares estimator minimises either vertical or horizontal errors (in 3.28), whereas the first eigenvector corresponds to a minimisation that is orthogonal to the eigenvector (see Chapter 10).

Summary

↪ The linear regression $y = \alpha + \beta x + \varepsilon$ models a linear relation between two one-dimensional variables.

↪ The sign of the slope $\widehat{\beta}$ is the same as that of the covariance and the correlation of x and y.

↪ A linear regression predicts values of Y given a possible observation x of X.

↪ The coefficient of determination r^2 measures the amount of variation in Y which is explained by a linear regression on X.

↪ If the coefficient of determination is $r^2 = 1$, then all points lie on one line.

↪ The regression line of X on Y and the regression line of Y on X are in general different.

↪ The t-test for the hypothesis $\beta = 0$ is $t = \frac{\widehat{\beta}}{SE(\widehat{\beta})}$, where $SE(\widehat{\beta}) = \frac{\hat{\sigma}}{(n \cdot s_{XX})^{1/2}}$.

↪ The t-test rejects the null hypothesis $\beta = 0$ at the level of significance α if $|t| \geq t_{1-\alpha/2;n-2}$ where $t_{1-\alpha;n-2}$ is the $1 - \alpha/2$ quantile of the Student's t-distribution with $(n-2)$ degrees of freedom.

↪ The standard error $SE(\widehat{\beta})$ increases/decreases with less/more spread in the X variables.

↪ The direction of the first eigenvector of the covariance matrix of a two-dimensional point cloud is different from the least squares regression line.

Table 3.1 Observation structure of a simple ANOVA

Sample element	Factor levels l				
1	y_{11}	$\cdots$	y_{1l}	$\cdots$	y_{1p}
2	$\vdots$		$\vdots$		$\vdots$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
k	y_{k1}	$\cdots$	y_{kl}	$\cdots$	y_{kp}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$m = n/p$	y_{m1}	$\cdots$	y_{ml}	$\cdots$	y_{mp}

3.5 Simple Analysis of Variance

In a simple (i.e., one–factorial) analysis of variance (ANOVA), it is assumed that the average values of the response variable y are induced by one simple factor. Suppose that this factor takes on p values and that for each factor level, we have $m = n/p$ observations. The sample is of the form given in Table 3.1, where all of the observations are independent.

The goal of a simple ANOVA is to analyse the observation structure

$$y_{kl} = \mu_l + \varepsilon_{kl} \quad \text{for } k = 1, \ldots, m, \text{ and } l = 1, \ldots, p. \tag{3.41}$$

Each factor has a mean value μ_l. Each observation y_{kl} is assumed to be a sum of the corresponding factor mean value μ_l and a zero mean random error ε_{kl}. The linear regression model falls into this scheme with $m = 1$, $p = n$ and $\mu_i = \alpha + \beta x_i$, where x_i is the i-th level value of the factor.

Example 3.14 The "classic blue" pullover company analyses the effect of three marketing strategies

1 advertisement in local newspaper,
2 presence of sales assistant,
3 luxury presentation in shop windows.

All of these strategies are tried in 10 different shops. The resulting sale observations are given in Table 3.2.

There are $p = 3$ factors and $n = mp = 30$ observations in the data. The "classic blue" pullover company wants to know whether all three marketing strategies have the same mean effect or whether there are differences. Having the same effect means that all μ_l in (3.41) equal one value, μ. The hypothesis to be tested is therefore

$$H_0 : \mu_l = \mu \quad \text{for } l = 1, \ldots, p.$$

The alternative hypothesis, that the marketing strategies have different effects, can be formulated as

$$H_1 : \mu_l \neq \mu_{l'} \quad \text{for some } l \text{ and } l'.$$

This means that one marketing strategy is better than the others.

Table 3.2 Pullover sales as function of marketing strategy

Shop	Marketing strategy		
k	*factor l*		
	1	2	3
1	9	10	18
2	11	15	14
3	10	11	17
4	12	15	9
5	7	15	14
6	11	13	17
7	12	7	16
8	10	15	14
9	11	13	17
10	13	10	15

The method used to test this problem is to compute as in (3.38) the total variation and to decompose it into the sources of variation. This gives:

$$\sum_{l=1}^{p}\sum_{k=1}^{m}(y_{kl}-\bar{y})^2 = m\sum_{l=1}^{p}(\bar{y}_l-\bar{y})^2 + \sum_{l=1}^{p}\sum_{k=1}^{m}(y_{kl}-\bar{y}_l)^2. \tag{3.42}$$

The total variation (sum of squares = SS) is:

$$SS(\text{reduced}) = \sum_{l=1}^{p}\sum_{k=1}^{m}(y_{kl}-\bar{y})^2 \tag{3.43}$$

where $\bar{y} = n^{-1}\sum_{l=1}^{p}\sum_{k=1}^{m} y_{kl}$ is the overall mean. Here the total variation is denoted as SS(reduced), since in comparison with the model under the alternative H_1, we have a reduced set of parameters. In fact there is 1 parameter $\mu = \mu_l$ under H_0. Under H_1, the "full" model, we have three parameters, namely the three different means μ_l.

The variation under H_1 is therefore:

$$SS(\text{full}) = \sum_{l=1}^{p}\sum_{k=1}^{m}(y_{kl}-\bar{y}_l)^2 \tag{3.44}$$

where $\bar{y}_l = m^{-1}\sum_{k=1}^{m} y_{kl}$ is the mean of each factor l. The hypothetical model H_0 is called reduced, since it has (relative to H_1) fewer parameters.

The F-test of the linear hypothesis is used to compare the difference in the variations under the reduced model H_0 (3.43) and the full model H_1 (3.44) to the variation under the full model H_1:

$$F = \frac{\{SS(\text{reduced}) - SS(\text{full})\}/\{df(r) - df(f)\}}{SS(\text{full})/df(f)}. \tag{3.45}$$

Here $df(f)$ and $df(r)$ denote the degrees of freedom under the full model and the reduced model respectively. The degrees of freedom are essential in specifying the shape of the F-distribution. They have a simple interpretation: $df(\cdot)$ is equal to the number of observations minus the number of parameters in the model.

From Example 3.14, $p = 3$ parameters are estimated under the full model, i.e., $df(f) = n - p = 30 - 3 = 27$. Under the reduced model, there is one parameter to estimate, namely the overall mean, i.e., $df(r) = n - 1 = 29$. We can compute

$$SS(\text{reduced}) = 260.3$$

and

$$SS(\text{full}) = 157.7.$$

The F-statistic (3.45) is therefore

$$F = \frac{(260.3 - 157.7)/2}{157.7/27} = 8.78.$$

This value needs to be compared to the quantiles of the $F_{2,27}$ distribution. Looking up the critical values in a F-distribution shows that the test statistic above is highly significant. We conclude that the marketing strategies have different effects.

The F-test in a Linear Regression Model

The t-test of a linear regression model can be put into this framework. For a linear regression model (3.27), the reduced model is the one with $\beta = 0$:

$$y_i = \alpha + 0 \cdot x_i + \varepsilon_i .$$

The reduced model has $n - 1$ degrees of freedom and one parameter, the intercept α.

The full model is given by $\beta \neq 0$,

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

and has $n - 2$ degrees of freedom, since there are two parameters (α, β).

The SS(reduced) equals

$$SS(\text{reduced}) = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \textit{total variation}.$$

The SS(full) equals

$$SS(\text{full}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \text{RSS} = \textit{unexplained variation}.$$

The F-test is therefore, from (3.45),

$$F = \frac{(total\ variation - unexplained\ variation)/1}{(unexplained\ variation)/(n-2)} \tag{3.46}$$

$$= \frac{explained\ variation}{(unexplained\ variation)/(n-2)}. \tag{3.47}$$

Using the estimators $\hat{\alpha}$ and $\widehat{\beta}$ the explained variation is:

$$\begin{aligned}
\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 &= \sum_{i=1}^{n} \left(\hat{\alpha} + \widehat{\beta} x_i - \bar{y}\right)^2 \\
&= \sum_{i=1}^{n} \left\{(\bar{y} - \widehat{\beta}\bar{x}) + \widehat{\beta} x_i - \bar{y}\right\}^2 \\
&= \sum_{i=1}^{n} \widehat{\beta}^2 (x_i - \bar{x})^2 \\
&= \widehat{\beta}^2 n s_{XX}.
\end{aligned}$$

From (3.32) the F-ratio (3.46) is therefore:

$$F = \frac{\hat{\beta}^2 n s_{XX}}{RSS/(n-2)} \tag{3.48}$$

$$= \left(\frac{\widehat{\beta}}{SE(\widehat{\beta})}\right)^2. \tag{3.49}$$

The t-test statistic (3.33) is just the square root of the F- statistic (3.49).

Note, using (3.39) the F-statistic can be rewritten as

$$F = \frac{r^2/1}{(1-r^2)/(n-2)}.$$

In the pullover Example 3.11, we obtain $F = \frac{0.028}{0.972}\frac{8}{1} = 0.2305$, so that the null hypothesis $\beta = 0$ cannot be rejected. We conclude therefore that there is only a minor influence of prices on sales.

Summary

↪ Simple ANOVA models an output Y as a function of one factor.

↪ The reduced model is the hypothesis of equal means.

↪ The full model is the alternative hypothesis of different means.

	Summary (continued)
↪	The F-test is based on a comparison of the sum of squares under the full and the reduced models.
↪	The degrees of freedom are calculated as the number of observations minus the number of parameters.
↪	The F-statistic is $$F = \frac{\{SS(\text{reduced}) - SS(\text{full})\}/\{df(r) - df(f)\}}{SS(\text{full})/df(f)}.$$
↪	The F-test rejects the null hypothesis if the F-statistic is larger than the 95% quantile of the $F_{df(r)-df(f),df(f)}$ distribution.
↪	The F-test statistic for the slope of the linear regression model $y_i = \alpha + \beta x_i + \varepsilon_i$ is the square of the t-test statistic.

3.6 Multiple Linear Model

The simple linear model and the analysis of variance model can be viewed as a particular case of a more general linear model where the variations of one variable y are explained by p explanatory variables x respectively. Let y $(n \times 1)$ and $\mathcal{X}$ $(n \times p)$ be a vector of observations on the response variable and a data matrix on the p explanatory variables. An important application of the developed theory is the least squares fitting. The idea is to approximate y by a linear combination $\widehat{y}$ of columns of $\mathcal{X}$, i.e., $\widehat{y} \in C(\mathcal{X})$. The problem is to find $\widehat{\beta} \in \mathbb{R}^p$ such that $\widehat{y} = \mathcal{X}\widehat{\beta}$ is the best fit of y in the least-squares sense. The linear model can be written as

$$y = \mathcal{X}\beta + \varepsilon, \tag{3.50}$$

where ε are the errors. The least squares solution is given by $\widehat{\beta}$:

$$\widehat{\beta} = \arg\min_{\beta} \, (y - \mathcal{X}\beta)^\top (y - \mathcal{X}\beta) = \arg\min_{\beta} \, \varepsilon^\top \varepsilon. \tag{3.51}$$

Suppose that $(\mathcal{X}^\top \mathcal{X})$ is of full rank and thus invertible. Minimising the expression (3.51) with respect to β yields:

$$\widehat{\beta} = (\mathcal{X}^\top \mathcal{X})^{-1} \mathcal{X}^\top y. \tag{3.52}$$

The fitted value $\widehat{y} = \mathcal{X}\widehat{\beta} = \mathcal{X}(\mathcal{X}^\top \mathcal{X})^{-1}\mathcal{X}^\top y = \mathcal{P}y$ is the projection of y onto $C(\mathcal{X})$ as computed in (2.47).

The least squares residuals are

$$e = y - \widehat{y} = y - \mathcal{X}\widehat{\beta} = \mathcal{Q}y = (\mathcal{I}_n - \mathcal{P})y.$$

The vector e is the projection of y onto the orthogonal complement of $C(\mathcal{X})$.

Remark 3.5 A linear model with an intercept α can also be written in this framework. The approximating equation is:

$$y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i; \quad i = 1, \ldots, n.$$

This can be written as:

$$y = \mathcal{X}^* \beta^* + \varepsilon$$

where $\mathcal{X}^* = (1_n \ \mathcal{X})$ (we add a column of ones to the data). We have by (3.52):

$$\widehat{\beta}^* = \begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = (\mathcal{X}^{*\top} \mathcal{X}^*)^{-1} \mathcal{X}^{*\top} y.$$

Example 3.15 Let us come back to the "classic blue" pullovers example. In Example 3.11, we considered the regression fit of the sales X_1 on the price X_2 and concluded that there was only a small influence of sales by changing the prices. A linear model incorporating all three variables allows us to approximate sales as a linear function of price (X_2), advertisement (X_3) and presence of sales assistants (X_4) simultaneously. Adding a column of ones to the data (in order to estimate the intercept α) leads to

$$\widehat{\alpha} = 65.670 \quad \text{and} \quad \widehat{\beta}_1 = -0.216, \quad \widehat{\beta}_2 = 0.485, \quad \widehat{\beta}_3 = 0.844.$$

The coefficient of determination is computed as before in (3.40) and is:

$$r^2 = 1 - \frac{e^\top e}{\sum (y_i - \overline{y})^2} = 0.907.$$

We conclude that the variation of X_1 is well approximated by the linear relation.

Remark 3.6 The coefficient of determination is influenced by the number of regressors. For a given sample size n, the r^2 value will increase by adding more regressors into the linear model. The value of r^2 may therefore be high even if possibly irrelevant regressors are included. A corrected coefficient of determination for p regressors and a constant intercept ($p + 1$ parameters) is

$$r^2_{\text{adj}} = r^2 - \frac{p(1 - r^2)}{n - (p + 1)}. \tag{3.53}$$

Example 3.16 The corrected coefficient of determination for Example 3.15 is

$$\begin{aligned} r^2_{\text{adj}} &= 0.907 - \frac{3(1 - 0.907^2)}{10 - 3 - 1} \\ &= 0.818. \end{aligned}$$

This means that 81.8% of the variation of the response variable is explained by the explanatory variables.

Note that the linear model (3.50) is very flexible and can model nonlinear relationships between the response y and the explanatory variables x. For example, a quadratic relation in one variable x could be included. Then $y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ could be written in matrix notation as in (3.50), $y = \mathcal{X}\beta + \varepsilon$ where

$$\mathcal{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}.$$

Properties of $\widehat{\beta}$

When y_i is the i-th observation of a random variable Y, the errors are also random. Under standard assumptions (independence, zero mean and constant variance σ^2), inference can be conducted on β. Using the properties of Chapter 4, it is easy to prove:

$$\begin{aligned} \mathsf{E}(\widehat{\beta}) &= \beta \\ \mathsf{Var}(\widehat{\beta}) &= \sigma^2(\mathcal{X}^\top\mathcal{X})^{-1}. \end{aligned}$$

The analogue of the t-test for the multivariate linear regression situation is

$$t = \frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}.$$

The standard error of each coefficient $\widehat{\beta}_j$ is given by the square root of the diagonal elements of the matrix $\mathsf{Var}(\widehat{\beta})$. In standard situations, the variance σ^2 of the error ε is not known. One may estimate it by

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)}(y-\hat{y})^\top(y-\hat{y}),$$

where $(p+1)$ is the dimension of β. In testing $\beta_j = 0$ we reject the hypothesis at the significance level α if $|t| \geq t_{1-\alpha/2;n-(p+1)}$. More general issues on testing linear models are addressed in Chapter 7.

The ANOVA Model in Matrix Notation

The simple ANOVA problem (Section 3.5) may also be rewritten in matrix terms. Recall the definition of a vector of ones from (2.1) and define a vector of zeros as 0_n. Then construct the following $(n \times p)$ matrix, (here $p = 3$),

$$\mathcal{X} = \begin{pmatrix} 1_m & 0_m & 0_m \\ 0_m & 1_m & 0_m \\ 0_m & 0_m & 1_m \end{pmatrix}, \tag{3.54}$$

where $m = 10$. Equation (3.41) then reads as follows.

The parameter vector is $\beta = (\mu_1, \mu_2, \mu_3)^\top$. The data set from Example 3.14 can therefore be written as a linear model $y = \mathcal{X}\beta + \varepsilon$ where $y \in \mathbb{R}^n$ with $n = m \cdot p$ is the stacked vector of the columns of Table 3.1. The projection into the column space $C(\mathcal{X})$ of (3.54) yields the least-squares estimator $\hat{\beta} = (\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top y$. Note that $(\mathcal{X}^\top\mathcal{X})^{-1} = (1/10)\mathcal{I}_3$ and that $\mathcal{X}^\top y = (106, 124, 151)^\top$ is the sum $\sum_{k=1}^m y_{kj}$ for each factor, i.e., the 3 column sums of Table 3.1. The least squares estimator is therefore the vector $\hat{\beta}_{H_1} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (10.6, 12.4, 15.1)^\top$ of sample means for each factor level $j = 1, 2, 3$. Under the null hypothesis of equal mean values $\mu_1 = \mu_2 = \mu_3 = \mu$, we estimate the parameters under the same constraints. This can be put into the form of a linear constraint:

$$\begin{aligned} -\mu_1 + \mu_2 &= 0 \\ -\mu_1 + \mu_3 &= 0. \end{aligned}$$

This can be written as $\mathcal{A}\beta = a$, where

$$a = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$\mathcal{A} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

The constrained least-squares solution can be shown (Exercise 3.24) to be given by:

$$\hat{\beta}_{H_0} = \hat{\beta}_{H_1} - (\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{A}^\top\{\mathcal{A}(\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{A}^\top\}^{-1}(\mathcal{A}\hat{\beta}_{H_1} - a). \tag{3.55}$$

It turns out that (3.55) amounts to simply calculating the overall mean $\bar{y} = 12.7$ of the response variable y: $\hat{\beta}_{H_0} = (12.7, 12.7, 12.7)^\top$.

The F-test that has already been applied in Example 3.14 can be written as

$$F = \frac{\{||y - \mathcal{X}\hat{\beta}_{H_0}||^2 - ||y - \mathcal{X}\hat{\beta}_{H_1}||^2\}/2}{||y - \mathcal{X}\hat{\beta}_{H_1}||^2/27} \tag{3.56}$$

which gives the same significant value 8.78. Note that again we compare the RSS_{H_0} of the reduced model to the RSS_{H_1} of the full model. It corresponds to comparing the lengths of projections into different column spaces. This general approach in testing linear models is described in detail in Chapter 7.

Summary

↪ The relation $y = \mathcal{X}\beta + e$ models a linear relation between a one-dimensional variable Y and a p-dimensional variable X. $\mathcal{P}y$ gives the best linear regression fit of the vector y onto $C(\mathcal{X})$. The least squares parameter estimator is $\widehat{\beta} = (\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top y$.

Summary (continued)
↪ The simple ANOVA model can be written as a linear model.
↪ The ANOVA model can be tested by comparing the length of the projection vectors.
↪ The test statistic of the F-Test can be written as $$\frac{\{\|y-\mathcal{X}\hat{\beta}_{H_0}\|^2-\|y-\mathcal{X}\hat{\beta}_{H_1}\|^2\}/\{df(r)-df(f)\}}{\|y-\mathcal{X}\hat{\beta}_{H_1}\|^2/df(f)}.$$
↪ The adjusted coefficient of determination is $$r^2_{\text{adj}}=r^2-\frac{p(1-r^2)}{n-(p+1)}.$$

3.7 Boston Housing

The main statistics presented so far can be computed for the data matrix $\mathcal{X}(506\times 14)$ from our Boston Housing data set. The sample means and the sample medians of each variable are displayed in Table 3.3. The table also provides the unbiased estimates of the variance of each variable and the corresponding standard deviations. The comparison of the means and the medians confirms the assymmetry of the components of $\mathcal{X}$ that was pointed out in Section 1.9.

Table 3.3 Descriptive statistics for the Boston Housing data set
MVAdescbh

X	$\overline{x}$	median(X)	Var(X)	std(X)
X_1	3.61	0.26	73.99	8.60
X_2	11.36	0.00	543.94	23.32
X_3	11.14	9.69	47.06	6.86
X_4	0.07	0.00	0.06	0.25
X_5	0.55	0.54	0.01	0.12
X_6	6.28	6.21	0.49	0.70
X_7	68.57	77.50	792.36	28.15
X_8	3.79	3.21	4.43	2.11
X_9	9.55	5.00	75.82	8.71
X_{10}	408.24	330.00	28405.00	168.54
X_{11}	18.46	19.05	4.69	2.16
X_{12}	356.67	391.44	8334.80	91.29
X_{13}	12.65	11.36	50.99	7.14
X_{14}	22.53	21.20	84.59	9.20

The (unbiased) sample covariance matrix is given by the following (14×14) matrix $\mathcal{S}_n$:

$$\begin{pmatrix}
73.99 & -40.22 & 23.99 & -0.12 & 0.42 & -1.33 & 85.41 & -6.88 & 46.85 & 844.82 & 5.40 & -302.38 & 27.99 & -30.72 \\
-40.22 & 543.94 & -85.41 & -0.25 & -1.40 & 5.11 & -373.90 & 32.63 & -63.35 & -1236.45 & -19.78 & 373.72 & -68.78 & 77.32 \\
23.99 & -85.41 & 47.06 & 0.11 & 0.61 & -1.89 & 124.51 & -10.23 & 35.55 & 833.36 & 5.69 & -223.58 & 29.58 & -30.52 \\
-0.12 & -0.25 & 0.11 & 0.06 & 0.00 & 0.02 & 0.62 & -0.05 & -0.02 & -1.52 & -0.07 & 1.13 & -0.10 & 0.41 \\
0.42 & -1.40 & 0.61 & 0.00 & 0.01 & -0.02 & 2.39 & -0.19 & 0.62 & 13.05 & 0.05 & -4.02 & 0.49 & -0.46 \\
-1.33 & 5.11 & -1.89 & 0.02 & -0.02 & 0.49 & -4.75 & 0.30 & -1.28 & -34.58 & -0.54 & 8.22 & -3.08 & 4.49 \\
85.41 & -373.90 & 124.51 & 0.62 & 2.39 & -4.75 & 792.36 & -44.33 & 111.77 & 2402.69 & 15.94 & -702.94 & 121.08 & -97.59 \\
-6.88 & 32.63 & -10.23 & -0.05 & -0.19 & 0.30 & -44.33 & 4.43 & -9.07 & -189.66 & -1.06 & 56.04 & -7.47 & 4.84 \\
46.85 & -63.35 & 35.55 & -0.02 & 0.62 & -1.28 & 111.77 & -9.07 & 75.82 & 1335.76 & 8.76 & -353.28 & 30.39 & -30.56 \\
844.82 & -1236.45 & 833.36 & -1.52 & 13.05 & -34.58 & 2402.69 & -189.66 & 1335.76 & 28404.76 & 168.15 & -6797.91 & 654.71 & -726.26 \\
5.40 & -19.78 & 5.69 & -0.07 & 0.05 & -0.54 & 15.94 & -1.06 & 8.76 & 168.15 & 4.69 & -35.06 & 5.78 & -10.11 \\
-302.38 & 373.72 & -223.58 & 1.13 & -4.02 & 8.22 & -702.94 & 56.04 & -353.28 & -6797.91 & -35.06 & 8334.75 & -238.67 & 279.99 \\
27.99 & -68.78 & 29.58 & -0.10 & 0.49 & -3.08 & 121.08 & -7.47 & 30.39 & 654.71 & 5.78 & -238.67 & 50.99 & -48.45 \\
-30.72 & 77.32 & -30.52 & 0.41 & -0.46 & 4.49 & -97.59 & 4.84 & -30.56 & -726.26 & -10.11 & 279.99 & -48.45 & 84.59
\end{pmatrix},$$

and the corresponding correlation matrix $\mathcal{R}(14 \times 14)$ is:

$$\begin{pmatrix}
1.00 & -0.20 & 0.41 & -0.06 & 0.42 & -0.22 & 0.35 & -0.38 & 0.63 & 0.58 & 0.29 & -0.39 & 0.46 & -0.39 \\
-0.20 & 1.00 & -0.53 & -0.04 & -0.52 & 0.31 & -0.57 & 0.66 & -0.31 & -0.31 & -0.39 & 0.18 & -0.41 & 0.36 \\
0.41 & -0.53 & 1.00 & 0.06 & 0.76 & -0.39 & 0.64 & -0.71 & 0.60 & 0.72 & 0.38 & -0.36 & 0.60 & -0.48 \\
-0.06 & -0.04 & 0.06 & 1.00 & 0.09 & 0.09 & 0.09 & -0.10 & -0.01 & -0.04 & -0.12 & 0.05 & -0.05 & 0.18 \\
0.42 & -0.52 & 0.76 & 0.09 & 1.00 & -0.30 & 0.73 & -0.77 & 0.61 & 0.67 & 0.19 & -0.38 & 0.59 & -0.43 \\
-0.22 & 0.31 & -0.39 & 0.09 & -0.30 & 1.00 & -0.24 & 0.21 & -0.21 & -0.29 & -0.36 & 0.13 & -0.61 & 0.70 \\
0.35 & -0.57 & 0.64 & 0.09 & 0.73 & -0.24 & 1.00 & -0.75 & 0.46 & 0.51 & 0.26 & -0.27 & 0.60 & -0.38 \\
-0.38 & 0.66 & -0.71 & -0.10 & -0.77 & 0.21 & -0.75 & 1.00 & -0.49 & -0.53 & -0.23 & 0.29 & -0.50 & 0.25 \\
0.63 & -0.31 & 0.60 & -0.01 & 0.61 & -0.21 & 0.46 & -0.49 & 1.00 & 0.91 & 0.46 & -0.44 & 0.49 & -0.38 \\
0.58 & -0.31 & 0.72 & -0.04 & 0.67 & -0.29 & 0.51 & -0.53 & 0.91 & 1.00 & 0.46 & -0.44 & 0.54 & -0.47 \\
0.29 & -0.39 & 0.38 & -0.12 & 0.19 & -0.36 & 0.26 & -0.23 & 0.46 & 0.46 & 1.00 & -0.18 & 0.37 & -0.51 \\
-0.39 & 0.18 & -0.36 & 0.05 & -0.38 & 0.13 & -0.27 & 0.29 & -0.44 & -0.44 & -0.18 & 1.00 & -0.37 & 0.33 \\
0.46 & -0.41 & 0.60 & -0.05 & 0.59 & -0.61 & 0.60 & -0.50 & 0.49 & 0.54 & 0.37 & -0.37 & 1.00 & -0.74 \\
-0.39 & 0.36 & -0.48 & 0.18 & -0.43 & 0.70 & -0.38 & 0.25 & -0.38 & -0.47 & -0.51 & 0.33 & -0.74 & 1.00
\end{pmatrix}.$$

Analyzing $\mathcal{R}$ confirms most of the comments made from examining the scatterplot matrix in Chapter 1. In particular, the correlation between X_{14} (the value of the house) and all the other variables is given by the last row (or column) of $\mathcal{R}$. The highest correlations (in absolute values) are in decreasing order X_{13}, X_6, X_{11}, X_{10}, etc.

Using the Fisher's Z-transform on each of the correlations between X_{14} and the other variables would confirm that all are significantly different from zero, except the correlation between X_{14} and X_4 (the indicator variable for the Charles River). We know, however, that the correlation and Fisher's Z-transform are not appropriate for binary variable.

The same descriptive statistics can be calculated for the transformed variables (transformations were motivated in Section 1.9). The results are given in Table 3.4 and as can be seen, most of the variables are now more symmetric. Note that the covariances and the correlations are sensitive to these nonlinear transformations. For example, the correlation matrix is now

Table 3.4 Descriptive statistics for the Boston Housing data set after the transformation MVAdescbh

$\widetilde{X}$	$\overline{\widetilde{x}}$	median($\widetilde{X}$)	Var($\widetilde{X}$)	std($\widetilde{X}$)
$\widetilde{X}_1$	−0.78	−1.36	4.67	2.16
$\widetilde{X}_2$	1.14	0.00	5.44	2.33
$\widetilde{X}_3$	2.16	2.27	0.60	0.78
$\widetilde{X}_4$	0.07	0.00	0.06	0.25
$\widetilde{X}_5$	−0.61	−0.62	0.04	0.20
$\widetilde{X}_6$	1.83	1.83	0.01	0.11
$\widetilde{X}_7$	5.06	5.29	12.72	3.57
$\widetilde{X}_8$	1.19	1.17	0.29	0.54
$\widetilde{X}_9$	1.87	1.61	0.77	0.87
$\widetilde{X}_{10}$	5.93	5.80	0.16	0.40
$\widetilde{X}_{11}$	2.15	2.04	1.86	1.36
$\widetilde{X}_{12}$	3.57	3.91	0.83	0.91
$\widetilde{X}_{13}$	3.42	3.37	0.97	0.99
$\widetilde{X}_{14}$	3.03	3.05	0.17	0.41

$$\begin{pmatrix}
1.00 & -0.52 & 0.74 & 0.03 & 0.81 & -0.32 & 0.70 & -0.74 & 0.84 & 0.81 & 0.45 & -0.48 & 0.62 & -0.57 \\
-0.52 & 1.00 & -0.66 & -0.04 & -0.57 & 0.31 & -0.53 & 0.59 & -0.35 & -0.31 & -0.35 & 0.18 & -0.45 & 0.36 \\
0.74 & -0.66 & 1.00 & 0.08 & 0.75 & -0.43 & 0.66 & -0.73 & 0.58 & 0.66 & 0.46 & -0.33 & 0.62 & -0.55 \\
0.03 & -0.04 & 0.08 & 1.00 & 0.08 & 0.08 & 0.07 & -0.09 & 0.01 & -0.04 & -0.13 & 0.05 & -0.06 & 0.16 \\
0.81 & -0.57 & 0.75 & 0.08 & 1.00 & -0.32 & 0.78 & -0.86 & 0.61 & 0.67 & 0.34 & -0.38 & 0.61 & -0.52 \\
-0.32 & 0.31 & -0.43 & 0.08 & -0.32 & 1.00 & -0.28 & 0.28 & -0.21 & -0.31 & -0.32 & 0.13 & -0.64 & 0.61 \\
0.70 & -0.53 & 0.66 & 0.07 & 0.78 & -0.28 & 1.00 & -0.80 & 0.47 & 0.54 & 0.38 & -0.29 & 0.64 & -0.48 \\
-0.74 & 0.59 & -0.73 & -0.09 & -0.86 & 0.28 & -0.80 & 1.00 & -0.54 & -0.60 & -0.32 & 0.32 & -0.56 & 0.41 \\
0.84 & -0.35 & 0.58 & 0.01 & 0.61 & -0.21 & 0.47 & -0.54 & 1.00 & 0.82 & 0.40 & -0.41 & 0.46 & -0.43 \\
0.81 & -0.31 & 0.66 & -0.04 & 0.67 & -0.31 & 0.54 & -0.60 & 0.82 & 1.00 & 0.48 & -0.43 & 0.53 & -0.56 \\
0.45 & -0.35 & 0.46 & -0.13 & 0.34 & -0.32 & 0.38 & -0.32 & 0.40 & 0.48 & 1.00 & -0.20 & 0.43 & -0.51 \\
-0.48 & 0.18 & -0.33 & 0.05 & -0.38 & 0.13 & -0.29 & 0.32 & -0.41 & -0.43 & -0.20 & 1.00 & -0.36 & 0.40 \\
0.62 & -0.45 & 0.62 & -0.06 & 0.61 & -0.64 & 0.64 & -0.56 & 0.46 & 0.53 & 0.43 & -0.36 & 1.00 & -0.83 \\
-0.57 & 0.36 & -0.55 & 0.16 & -0.52 & 0.61 & -0.48 & 0.41 & -0.43 & -0.56 & -0.51 & 0.40 & -0.83 & 1.00
\end{pmatrix}.$$

Notice that some of the correlations between $\widetilde{X}_{14}$ and the other variables have increased.

If we want to explain the variations of the price $\widetilde{X}_{14}$ by the variation of all the other variables $\widetilde{X}_1, \dots, \widetilde{X}_{13}$ we could estimate the linear model

$$\widetilde{X}_{14} = \beta_0 + \sum_{j=1}^{13} \beta_j \widetilde{X}_j + \varepsilon. \tag{3.57}$$

The result is given in Table 3.5.

The value of r^2 (0.765) and r^2_{adj} (0.759) show that most of the variance of X_{14} is explained by the linear model (3.57).

Again we see that the variations of $\widetilde{X}_{14}$ are mostly explained by (in decreasing order of the absolute value of the t-statistic) $\widetilde{X}_{13}, \widetilde{X}_8, \widetilde{X}_{11}, \widetilde{X}_{10}, \widetilde{X}_{12}, \widetilde{X}_6, \widetilde{X}_9, \widetilde{X}_4$

Table 3.5 Linear regression results for all variables of Boston Housing data set MVAlinregbh

Variable	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	t	p-value
constant	4.1769	0.3790	11.020	0.0000
$\widetilde{X}_1$	−0.0146	0.0117	−1.254	0.2105
$\widetilde{X}_2$	0.0014	0.0056	0.247	0.8051
$\widetilde{X}_3$	−0.0127	0.0223	−0.570	0.5692
$\widetilde{X}_4$	0.1100	0.0366	3.002	0.0028
$\widetilde{X}_5$	−0.2831	0.1053	−2.688	0.0074
$\widetilde{X}_6$	0.4211	0.1102	3.822	0.0001
$\widetilde{X}_7$	0.0064	0.0049	1.317	0.1885
$\widetilde{X}_8$	−0.1832	0.0368	−4.977	0.0000
$\widetilde{X}_9$	0.0684	0.0225	3.042	0.0025
$\widetilde{X}_{10}$	−0.2018	0.0484	−4.167	0.0000
$\widetilde{X}_{11}$	−0.0400	0.0081	−4.946	0.0000
$\widetilde{X}_{12}$	0.0445	0.0115	3.882	0.0001
$\widetilde{X}_{13}$	−0.2626	0.0161	−16.320	0.0000

and $\widetilde{X}_5$. The other variables $\widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3$ and $\widetilde{X}_7$ seem to have little influence on the variations of $\widetilde{X}_{14}$. This will be confirmed by the testing procedures that will be developed in Chapter 7.

3.8 Exercises

Exercise 3.1 The covariance $s_{X_4X_5}$ between X_4 and X_5 for the entire bank data set is positive. Given the definitions of X_4 and X_5, we would expect a negative covariance. Using Figure 3.1 can you explain why $s_{X_4X_5}$ is positive?

Exercise 3.2 Consider the two sub-clouds of counterfeit and genuine bank notes in Figure 3.1 separately. Do you still expect $s_{X_4X_5}$ (now calculated separately for each cloud) to be positive?

Exercise 3.3 We remarked that for two normal random variables, zero covariance implies independence. Why does this remark not apply to Example 3.4?

Exercise 3.4 Compute the covariance between the variables

$$X_2 = \text{miles per gallon},$$

$$X_8 = \text{weight}$$

from the car data set (Table B.3). What sign do you expect the covariance to have?

Exercise 3.5 Compute the correlation matrix of the variables in Example 3.2. Comment on the sign of the correlations and test the hypothesis

$$\rho_{X_1 X_2} = 0.$$

Exercise 3.6 Suppose you have observed a set of observations $\{x_i\}_{i=1}^n$ with $\overline{x} = 0$, $s_{XX} = 1$ and $n^{-1} \sum_{i=1}^n (x_i - \overline{x})^3 = 0$. Define the variable $y_i = x_i^2$. Can you immediately tell whether $r_{XY} \neq 0$?

Exercise 3.7 Find formulas (3.29) and (3.30) for $\widehat{\alpha}$ and $\widehat{\beta}$ by differentiating the objective function in (3.28) w.r.t. α and β.

Exercise 3.8 How many sales does the textile manager expect with a "classic blue" pullover price of $x = 105$?

Exercise 3.9 What does a scatterplot of two random variables look like for $r^2 = 1$ and $r^2 = 0$?

Exercise 3.10 Prove the variance decomposition (3.38) and show that the coefficient of determination is the square of the simple correlation between X and Y.

Exercise 3.11 Make a boxplot for the residuals $\varepsilon_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i$ for the "classic blue" pullovers data. If there are outliers, identify them and run the linear regression again without them. Do you obtain a stronger influence of price on sales?

Exercise 3.12 Under what circumstances would you obtain the same coefficients from the linear regression lines of Y on X and of X on Y?

Exercise 3.13 Treat the design of Example 3.14 as if there were thirty shops and not ten. Define x_i as the index of the shop, i.e., $x_i = i, i = 1, 2, \ldots, 30$. The null hypothesis is a constant regression line, $EY = \mu$. What does the alternative regression curve look like?

Exercise 3.14 Perform the test in Exercise 3.13 for the shop example with a 0.99 significance level. Do you still reject the hypothesis of equal marketing strategies?

Exercise 3.15 Compute an approximate confidence interval for $\rho_{X_2 X_8}$ in Example 3.2. Hint: start from a confidence interval for $\tanh^{-1}(\rho_{X_2 X_8})$ and then apply the inverse transformation.

Exercise 3.16 In Example 3.2, using the exchange rate of 1 EUR = 106 JPY, compute the same empirical covariance using prices in Japanese Yen rather than in Euros. Is there a significant difference? Why?

Exercise 3.17 Why does the correlation have the same sign as the covariance?

Exercise 3.18 Show that $\text{rank}(\mathcal{H}) = \text{tr}(\mathcal{H}) = n - 1$.

Exercise 3.19 Show that $\mathcal{X}_* = \mathcal{H}\mathcal{X}\mathcal{D}^{-1/2}$ is the standardized data matrix, i.e., $\overline{x}_* = 0$ and $\mathcal{S}_{\mathcal{X}_*} = \mathcal{R}_{\mathcal{X}}$.

Exercise 3.20 Compute for the pullovers data the regression of X_1 on X_2, X_3 and of X_1 on X_2, X_4. Which one has the better coefficient of determination?

Exercise 3.21 Compare for the pullovers data the coefficient of determination for the regression of X_1 on X_2 (Example 3.11), of X_1 on X_2, X_3 (Exercise 3.20) and of X_1 on X_2, X_3, X_4 (Example 3.15). Observe that this coefficient is increasing with the number of predictor variables. Is this always the case?

Exercise 3.22 Consider the ANOVA problem (Section 3.5) again. Establish the constraint Matrix $\mathcal{A}$ for testing $\mu_1 = \mu_2$. Test this hypothesis via an analog of (3.55) and (3.56).

Exercise 3.23 Prove (3.52). (Hint, let $f(\beta) = (y - x\beta)^\top(y - x\beta)$ and solve $\frac{\partial f(\beta)}{\partial \beta} = 0$.)

Exercise 3.24 Consider the linear model $Y = \mathcal{X}\beta + \varepsilon$ where $\hat{\beta} = \arg\min_\beta \varepsilon^\top\varepsilon$ is subject to the linear constraints $\mathcal{A}\widehat{\beta} = a$ where $\mathcal{A}(q \times p), (q \le p)$ is of rank q and a is of dimension $(q \times 1)$. Show that

$$\widehat{\beta} = \widehat{\beta}_{\text{OLS}} - (\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{A}^\top\left(\mathcal{A}(\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{A}^\top\right)^{-1}\left(\mathcal{A}\widehat{\beta}_{\text{OLS}} - a\right)$$

where $\widehat{\beta}_{\text{OLS}} = (\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top y$. (Hint, let $f(\beta, \lambda) = (y - x\beta)^\top(y - x\beta) - \lambda^\top(\mathcal{A}\beta - a)$ where $\lambda \in \mathbb{R}^q$ and solve $\frac{\partial f(\beta,\lambda)}{\partial \beta} = 0$ and $\frac{\partial f(\beta,\lambda)}{\partial \lambda} = 0$.)

Exercise 3.25 Compute the covariance matrix $\mathcal{S} = \mathsf{Cov}(\mathcal{X})$ where $\mathcal{X}$ denotes the matrix of observations on the counterfeit bank notes. Make a Jordan decomposition of $\mathcal{S}$. Why are all of the eigenvalues positive?

Exercise 3.26 Compute the covariance of the counterfeit notes after they are linearly transformed by the vector $a = (1, 1, 1, 1, 1, 1)^\top$.