Introduction 1

Sir Francis Galton (1822–1911) was a diverse researcher, who did pioneering work in many disciplines. Among statisticians, he is especially known for the Galton board which demonstrates the binomial distribution. At the end of the nineteenth century, Galton was mainly interested in questions regarding heredity. Galton collected extensive data illustrating body height of parents and their grown children. He examined the *relationship* between body heights of the children and the average body height of both parents. To adjust for the natural height differences across gender, the body height of women was multiplied by a factor of 1.08. In order to better examine this relationship, he listed all his data in a contingency table (Table 1.1). With the help of this table, he was able to make the following discoveries:

- Column-wise, i.e., for given average heights of the parents, the heights of the adolescents approximately follow a normal distribution.
- The normal distributions in each column have a common variance.
- When examining the relationship between the height of the children and the average height of the parents, an approximate linear trend was found with a slope of 2/3. A slope with value less than one led Galton to the conclusion that children of extremely tall (short) parents are usually shorter (taller) than their parents. In either case there is a tendency towards the population average, and Galton referred to this as *regression* towards the mean.

Later, Galton illustrated the data in the form of a scatter plot showing the heights of the children and the average height of the parents (Fig. 1.1). He visually added the trend or the *regression line*, which provides the average height of children as (average) parent height is varied.

Galton is viewed as a pioneer of regression analysis, because of his regression analytic study of heredity. However, Galton's mathematical capabilities were limited. His successors, especially Karl Pearson (1857–1936), Francis Ysidro Edgeworth (1845–1926), and George Udny Yule (1871–1951) formalized his work. Today, linear regression models are part of every introductory statistics book. In modern terms, Galton studied the systematic influence of the *explanatory variable*

L. Fahrmeir et al., *Regression*, DOI 10.1007/978-3-642-34333-9_1,

Table 1.1 Galton heredity data: contingency table between the height of 928 adult children and the average height of their 205 set of parents

Height of	Average height of parents											
children	64.0	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	73.0	Total
73.7	0	0	0	0	0	0	5	3	2	4	0	14
73.2	0	0	0	0	0	3	4	3	2	2	3	17
72.2	0	0	1	0	4	4	11	4	9	7	1	41
71.2	0	0	2	0	11	18	20	7	4	2	0	64
70.2	0	0	5	4	19	21	25	14	10	1	0	99
69.2	1	2	7	13	38	48	33	18	5	2	0	167
68.2	1	0	7	14	28	34	20	12	3	1	0	120
67.2	2	5	11	17	38	31	27	3	4	0	0	138
66.2	2	5	11	17	36	25	17	1	3	0	0	117
65.2	1	1	7	2	15	16	4	1	1	0	0	48
64.2	4	4	5	5	14	11	16	0	0	0	0	59
63.2	2	4	9	3	5	7	1	1	0	0	0	32
62.2	–	1	0	3	3	0	0	0	0	0	0	7
61.7	1	1	1	0	0	1	0	1	0	0	0	5
Total	14	23	66	78	211	219	183	68	43	19	4	928

The unit of measurement is inch which has already been used by Galton (1 inch corresponds to 2.54 cm)
Source: Galton (1889)

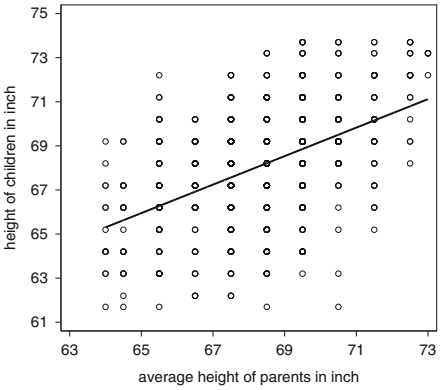

Fig. 1.1 Galton heredity data: scatter plot including a regression line between the height of children and the average height of their parents

$x =$ "average size of the parents" on the *response variable* $y =$ "height of grown-up children." Explanatory variables are also known as *independent variables*, *regressors*, or *covariates*. Response variables are also known as *dependent variables* or *target variables*. The fact that the linear relationship is not exact, but rather depends

on random errors, is a main characteristic for regression problems. Galton assumed the most simple regression model,

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

where the systematic component $\beta_0 + \beta_1 x$ is linear and ε constitutes the random error. While Galton determined the parameters β_0 and β_1 of the regression line in an ad hoc manner, nowadays these regression parameters are estimated via the *method of least squares*. The parameters β_0 and β_1 are estimated using the data pairs (y_i, x_i), $i = 1, \ldots, n$, so that the sum of the squared deviations

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

of the observations y_i from the regression line $\beta_0 + \beta_1 x_i$ is minimized. If we apply this principle to Galton's data, the estimated slope of the regression line is 0.64, a value that is fairly close to Galton's visually determined slope of 2/3.

Interestingly, the method of least squares was already discovered prior to Galton's study of heredity. The first publication by the mathematician Adrien Marie Legendre (1752–1833) appeared in 1805 making the method of least squares one of the oldest general estimation concepts in statistics. In the eighteenth and nineteenth century, the method was primarily used to predict the orbits of asteroids. Carl Friedrich Gauß (1777–1855) became famous for the prediction of the orbit of the asteroid Ceres, which was discovered in the year 1801 by the astronomer Giuseppe Piazzi. After forty days of observation, the asteroid disappeared behind the sun. Since an exact calculation of the asteroid's orbit was very difficult at that time, it was impossible to relocate the asteroid. By using the method of least squares, the twenty-four-year-old Gauß was able to give a feasible prediction of the asteroid's orbit. In his book "Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium" (1809), Gauß claimed the discovery of the method of least squares. Sometime later, Gauß even stated to have used this method since 1795 (as an eighteen year old), which provoked a priority dispute between Legendre and Gauß. Fact is that Gauß's work is the basis of the modern linear regression model with Gaussian errors.

Since the discovery of the method of least squares by Legendre and Gauß and the first regression analysis by Francis Galton, the methodology of regression has been improved and developed in many ways, and is nowadays applied in almost all scientific disciplines. The aim of this book is to give a modern introduction of the most important techniques and models of regression analysis and their application. We will address the following models in detail:

- *Regression models:* In Chap. 2, we present the different model classes without technical details; the subsequent chapters provide a thorough presentation of each of these models.

- *Linear models:* In Chaps. 3 and 4, we present a comprehensive introduction into linear regression models, including recent developments.
- *Generalized linear models:* In Chaps. 5 and 6, we discuss generalized linear models. These are especially suitable for problems where the response variables do not follow a normal distribution, including categorical response variables or count data.
- *Mixed models:* In Chap. 7, we present mixed models (models with random effects) for clustered data. A main focus in this chapter will be the analysis of panel and longitudinal data.
- *Univariate, bivariate, and spatial smoothing:* In Chap. 8, we introduce univariate and bivariate smoothing (nonparametric regression). These semiparametric and nonparametric methods are suitable to estimate complex nonlinear relationships including an automatic determination of the required amount of nonlinearity. Methods of spatial smoothing will also be discussed in detail.
- *Structured additive regression:* In Chap. 9, we present a unifying framework that combines the methods presented in Chap. 8 into one all-encompassing model. Structured additive regression models include a variety of special cases, for example, nonparametric and semiparametric regression models, additive models, geoadditive models, and varying-coefficient models. This chapter also illustrates how these models can be put into practice using a detailed case study.
- *Quantile regression:* Chapter 10 presents an introduction to quantile regression. While the methods of the previous chapters are more or less restricted to estimating the (conditional) mean depending on covariates, quantile regression allows to estimate the (conditional) quantiles of a response variable depending on covariates.

For the first time, this book presents a comprehensive and practical presentation of the most important models and methods of regression analysis. Chapter 2 is especially innovative, since it illustrates all model classes in a unified setting without focusing on the (often complicated) estimation techniques. The chapter gives the reader an overview of modern methods of regression and, at the same time, serves as a guide for choosing the appropriate model for each individual problem.

The following section illustrates the versatility of modern regression models to examine scientific questions in a variety of disciplines.

1.1 Examples of Applications

This book illustrates models and techniques of regression analysis via several applications taken from a variety of disciplines. The following list gives an overview:

- *Development economics:* Analysis of socioeconomic determinants of childhood malnutrition in developing countries
- *Hedonic prices:* Analysis of retail prices of the VW-Golf model
- *Innovation research:* Examination of the probability of opposition against patents granted by the European patent office

- *Credit scoring:* Analysis of the creditability of private bank customers
- *Marketing research:* Analysis of the relationship between the weekly unit sales of a product and sales promotions, particularly price variations
- *Rent index:* Analysis of the dependence between the monthly rent and the type, location, and condition of the rented apartment
- *Calculation of risk premium:* Analysis of claim frequency and claim size of motor vehicle insurance in order to calculate the risk premium
- *Ecology:* Analysis of the health status of trees in forests
- *Neuroscience:* Determination of the active brain area when solving certain cognitive tasks
- *Epidemiologic studies and clinical trials:*
  - Impact of testosterone on the growth of rats
  - Analysis of the probability of infection after Caesarean delivery
  - Study of the impairment to pulmonary function
  - Analysis of the life span of leukemia patients
- *Social science:* Analysis of speed dating data

Some of the listed examples will play a central role in this book and will now be discussed in more detail.

Example 1.1 Munich Rent Index

Many cities and communities in Germany establish rent indices in order to provide the renter and landlord with a market review for the "typical rent for the area." The basis for this index is a law in Germany that defines the "typical rent for the area" as the common remuneration that has been stipulated or changed over the last few years for price-maintained living area of comparable condition, size, and location within a specific community. This means that the average rent results from the apartment's characteristics, size, condition, etc. and therefore constitutes a typical regression problem. We use the net rent—the monthly rental price, which remains after having subtracted all running costs and incidentals—as the response variable. Alternatively, we can use the net rent per square meter as the response.

Within the scope of this book and due to data confidentiality, we confine ourselves to a fraction of the data and variables, which were used in the rent index for Munich in the year 1999. We use the 1999 data since more recent data is either not publicly available or less adequate for illustration purposes. The current rent index of Munich including documentation can be found at www.mietspiegel.muenchen.de (in German only).

Table 1.2 includes names and descriptions of the variables used in the subsequent analyses. The data of more than 3,000 apartments were collected by representative random sampling.

The goal of a regression analysis is to model the impact of explanatory variables (living area, year of construction, location, etc.) on the response variable of net rent or net rent per square meter. In a final step, we aim at representing the estimated effect of each explanatory variable in a simpler form by appropriate tables in a brochure or on the internet.

In this book, we use the Munich rent index data mainly to illustrate regression models with continuous responses (see Chaps. 2–4, 9, and 10). In doing so, we use simplified models for illustration purposes. This implies that the results do not always correspond to the official rent index.

△

Example 1.2 Malnutrition in Zambia

The World Health Organization (WHO) has decided to conduct representative household surveys (demographic and health surveys) in developing countries on a regular basis. Among others, these surveys consist of information regarding malnutrition, mortality, and

Table 1.2 Munich rent index: description of variables including summary statistics

Variable	Description	Mean/ frequency in %	Std.-dev.	Min/max
rent	Net rent per month (in Euro)	459.43	195.66	40.51/1,843.38
rentsqm	Net rent per month per square meter (in Euro)	7.11	2.44	0.41/17.72
area	Living area in square meters	67.37	23.72	20/160
yearc	Year of construction	1,956.31	22.31	1918/1997
location	Quality of location according to an expert assessment			
	1 = average location	58.21		
	2 = good location	39.26		
	3 = top location	2.53		
bath	Quality of bathroom			
	0 = standard	93.80		
	1 = premium	6.20		
kitchen	Quality of kitchen			
	0 = standard	95.75		
	1 = premium	4.25		
cheating	Central heating			
	0 = without central heating	10.42		
	1 = with central heating	89.58		
district	District in Munich			

health risks for children. The American institute Macro International collects data from over 50 countries. This data is freely available at www.measuredhs.com for research purposes. In this book, we look at an exemplary profile of a data set for Zambia taken in the year 1992 (4,421 observations in total). The Republic of Zambia is located in the south of Africa and is one of the poorest and most underdeveloped countries of the world.

One of the most serious problems of developing countries is the poor and often catastrophic nutritional condition of a high proportion of the population. Immediate consequences of malnutrition are reduced productivity and high mortality. Within the scope of this book, we will analyze the nutritional condition of children who are between 0 and 5 years old. The nutritional condition of children is usually determined by an anthropometric measure called Z-score. A Z-score compares the anthropometric status of a child, for example, a standardized age-specific body height, with comparable measures taken from a reference population. Until the age of 24 months, this reference population is based on white US-American children from wealthy families with a high socioeconomic status. After 24 months, the reference population changes and then consists of a representative sample taken from all US-American children. Among several possible anthropometric indicators, we use a measure for chronic malnutrition, which is based on body height as indication for the long-term development of the nutritional condition. This measure is defined as

$$zscore_i = \frac{h_i - mh}{\sigma},$$

for a child i, where h_i represents the height of the child, mh represents the median height of children belonging to the reference population of the same age group, and σ refers to the corresponding standard deviation for the reference population.

Table 1.3 Malnutrition in Zambia: description of variables including summary statistics

Variable	Description	Mean/ frequency in %	Std-dev.	Min/max
zscore	Child's Z-score	−171.19	139.34	−600/503
c_gender	Gender			
	1 = male	49.02		
	0 = female	50.98		
c_breastf	Duration of breast-feeding in months	11.11	9.42	0/46
c_age	Child's age in months	27.61	17.08	0/59
m_agebirth	Mother's age at birth in years	26.40	6.87	13.16/48.66
m_height	Mother's height in centimeter	158.06	5.99	134/185
m_bmi	Mother's body mass index	21.99	3.32	13.15/39.29
m_education	Mother's level of education			
	1 = no education	18.59		
	2 = primary school	62.34		
	3 = secondary school	17.35		
	4 = higher education	1.72		
m_work	Mother's work status			
	1 = mother working	55.25		
	0 = mother not working	44.75		
region	Region of residence in Zambia			
	1 = Central	8.89		
	2 = Copperbelt	21.87		
	3 = Eastern	9.27		
	4 = Luapula	8.91		
	5 = Lusaka	13.78		
	6 = Northern	9.73		
	7 = North western	5.88		
	8 = Southern	14.91		
	9 = Western	6.76		
district	District of residence in Zambia (55 districts)			

The primary goal of the statistical analysis is to determine the effect of certain socioeconomic variables of the child, the mother, and the household on the child's nutritional condition. Examples for socioeconomic variables are the duration of breastfeeding (*c_breastf*), the age of the child (*c_age*), the mother's nutritional condition as measured by the body mass index (*m_bmi*), and the mother's level of education as well as her work status (*m_education* and *m_work*). The data record also includes geographic information such as region or district where the mother's place of residence is located. A description of all available variables can be found in Table 1.3.

With the help of the regression models presented in this book, we will be able to pursue the aforementioned goals. Geoadditive models (see Sect. 9.2) are employed in particular. These also allow an adequate consideration of spatial information in the data. The data are analyzed within a comprehensive case study (see Sect. 9.8), which illustrates in detail the practical application of many techniques and methods presented in this book.

△

Table 1.4 Patent opposition: description of variables including summary statistics

Variable	Description	Mean/ frequency in %	Std-dev.	Min/max
opp	Patent opposition			
	1 = yes	41.49		
	0 = no	58.51		
biopharm	Patent from biotech/pharma sector			
	1 = yes	44.31		
	0 = no	55.69		
ustwin	US twin patent exists			
	1 = yes	60.85		
	0 = no	39.15		
patus	Patent holder from the USA			
	1 = yes	33.74		
	0 = no	66.26		
patgsgr	Patent holder from Germany, Switzerland, or Great Britain			
	1 = yes	23.49		
	0 = no	76.51		
year	Grant year			
	1980	0.18		
	⋮	⋮		
	1997	1.62		
ncit	Number of citations for the patent	1.64	2.74	0/40
ncountry	Number of designated states for the patent	7.8	4.12	1/17
nclaims	Number of claims	13.13	12.09	1/355

Example 1.3 Patent Opposition

The European Patent Office is able to protect a patent from competition for a certain period of time. The Patent Office has the task to examine inventions and to declare patent if certain prerequisites are fulfilled. The most important requirement is that the invention is something truly new. Even though the office examines each patent carefully, in about 80 % of cases competitors raise an objection against already assigned patents. In the economic literature the analysis of patent opposition plays an important role as it allows to (indirectly) investigate a number of economic questions. For instance, the frequency of patent opposition can be used as an indicator for the intensity of the competition in different market segments.

In order to analyze objections against patents, a data set with 4,866 patents from the sectors biotechnology/pharmaceutics and semiconductor/computer was collected. Table 1.4 lists the variables contained in this data set. The goal of the analysis is to model the probability of patent opposition, while using a variety of explanatory variables for the binary response variable "patent opposition" (yes/no). This corresponds to a regression problem with a binary response.

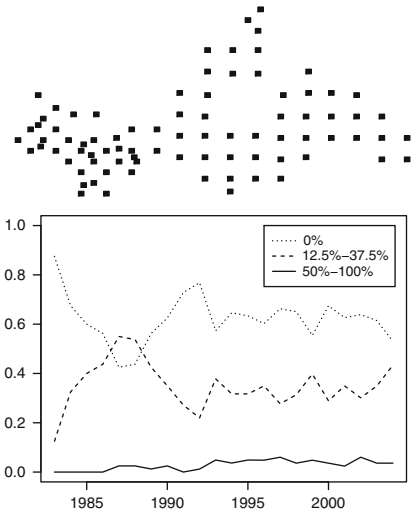

Fig. 1.2 Forest health status: The *top panel* shows the observed tree locations where the center constitutes the town of Rothenbuch. The *bottom panel* displays the temporal trend of defoliation degree

A possible explanatory variable is how often a patent has been cited in succeeding patents (variable *ncit*). Citations of patents are somewhat comparable to citations of scientific papers. Empirical experience and economic arguments indicate that the probability of an objection against a patent increases the more often it is cited. Regression models for binary response variables can formulate and examine this particular and other hypotheses. In this book the data set on patent opposition is primarily used to illustrate regression models with binary responses; see Chaps. 2 and 5.

△

Example 1.4 Forest Health Status

Knowledge about the health status of trees in a forest and its influencing factors is important from an ecological and economical point of view. This is the reason why Germany (and many other countries) conducts annual surveys regarding the condition of the forest. The data in our example come from a specific project in the forest of Rothenbuch (Spessart), which has been carried out by Axel Göttlein (Technical University, Munich) since 1982. In comparison to the extensive official land surveys, the observations, i.e., the locations of the examined trees, are much closer to each other. Figure 1.2 visualizes the 83 examined locations in Rothenbuch forest. Five tree species are part of this survey: beech, oak, spruce, larch, and pine. Here we will restrict ourselves to beech trees. Every year, the condition of beech trees is categorized by the response variable "defoliation" (*defol*) into nine ordinal categories 0 %, 12.5 %, 25 %, 37.5 %, 50 %, 62.5 %, 75 %, 87.5 %, and 100 %. Whereas the category 0 % signifies that the beech tree is healthy, the category 100 % implies that the tree is dead.

In addition to the (ordinal) response variable, explanatory variables are collected every year as well. Table 1.5 includes a selection of these variables including summary statistics. The mean values and frequencies (in percent) have been averaged over the years (1982–2004) and the observation points.

The goal of the analysis is to determine the effect of explanatory variables on the degree of defoliation. Moreover, we aim at quantifying the temporal trend and the spatial effect of geographic location, while adjusting for the effects of the other regressors. Additionally to the observed locations Fig. 1.2 presents the temporal trend of relative frequencies for the degree of defoliation of three (aggregated) categories.

Table 1.5 Forest health status: description of variables including summary statistics

Variable	Description	Mean/ frequency in %	Std-dev.	Min/max
id	Location identification number			
year	Year of data collection	1,993.58	6.33	1983/2004
defol	Degree of defoliation, in nine ordinal categories			
	0 %	62.07		
	12.5 %	24.26		
	25 %	7.03		
	37.5 %	3.79		
	50 %	1.62		
	62.5 %	0.89		
	75 %	0.33		
	87.5 %	0.00		
	100 %	0.00		
x	x-coordinate of location			
y	y-coordinate of location			
age	Average age of trees at the observation plot in years	106.17	51.38	7/234
canopyd	Canopy density in percent	77.31	23.70	0/100
gradient	Gradient of slope in percent	15.45	11.27	0/46
alt	Altitude above see level in meter	387.04	58.86	250/480
depth	Soil depth in cm	24.63	9.93	9/51
ph	pH-value in 0–2 cm depth	4.29	0.34	3.28/6.05
watermoisture	Level of soil moisture in three categories			
	1 = moderately dry	11.04		
	2 = moderately moist	55.16		
	3 = moist or temporarily wet	33.80		
alkali	Fraction of alkali ions in soil in four categories			
	1 = very low	19.63		
	2 = low	55.10		
	3 = moderate	17.18		
	4 = high	8.09		
humus	Thickness of humus layer in five categories			
	0 = 0 cm	25.71		
	1 = 1 cm	28.56		
	2 = 2 cm	21.58		
	3 = 3 cm	14.84		
	4 = more than 3 cm	9.31		

(continued)

Table 1.5 (continued)

Variable	Description	Mean/ frequency in %	Std-dev.	Min/max
type	Type of forest			
	0 = deciduous forest	50.31		
	1 = mixed forest	49.69		
fert	Fertilization			
	0 = not fertilized	80.87		
	1 = fertilized	19.13		

To analyze the data we apply regression models for multi-categorical response variables that can simultaneously accommodate nonlinear effects of the continuous covariates, as well as temporal and spatial trends. Such complex categorical regression models are illustrated in Chaps. 6 and 9.

△

The next section shows the first exploratory steps of regression analysis, which are illustrated using the data on the Munich rent index and the Zambia malnutrition data.

1.2 First Steps

1.2.1 Univariate Distributions of the Variables

The first step when conducting a regression analysis (and any other statistical evaluation) is to get an overview of the variables in the data set. We pursue the following goals for the initial descriptive and graphical univariate analysis:

- Summary and exploration of the distribution of the variables
- Identification of extreme values and outliers
- Identification of incorrect variable coding

To achieve these goals, we can use descriptive statistics, as well as graphical visualization techniques. The choice of appropriate methods depends on the individual type of variable. In general, we can differentiate between continuous and categorical variables.

We can get a first overview of continuous variables by determining some descriptive summary statistics, in particular the arithmetic mean and the median as typical measures of location, the standard deviation as a measure of variation, and the minimum and maximum of variables. Furthermore, it is useful to visualize the data. Histograms and box plots are most frequently used, but smooth nonparametric density estimators such as kernel densities are useful alternatives to histograms. Many introductory books, e.g., Veaux, Velleman, and Bock (2011) or Agresti and Finlay (2008), give easily accessible introductions to descriptive and exploratory statistics.

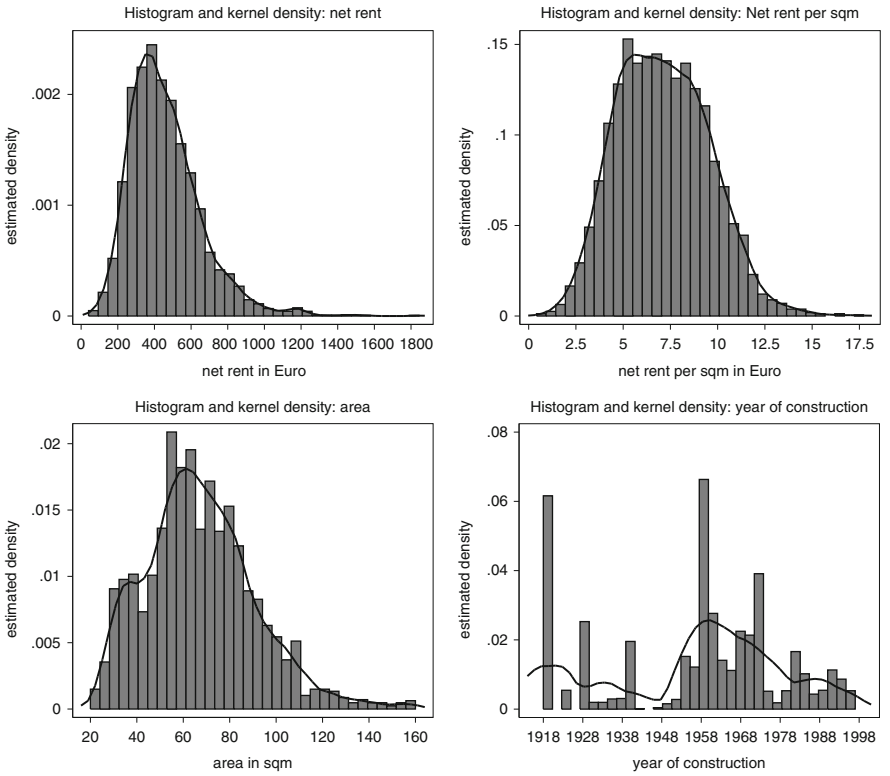

Fig. 1.3 Munich rent index: histograms and kernel density estimators for the continuous variables *rent*, *rentsqm*, *area* and *yearc*

Compared to continuous variables, it is easier to get an overview of the distribution of categorical variables. Here, we can use simple frequency tables or their graphical counterparts, particularly bar graphs.

Example 1.5 Munich Rent Index—Univariate Distributions

Summary statistics for the continuous variables *rent*, *rentsqm*, *area*, and *yearc* are already listed in Table 1.2 (p. 6). Figure 1.3 displays histograms and kernel density estimators for these variables. To give an example, we interpret summary statistics and graphical representations for the two variables "net rent" and "year of construction":

The monthly net rent roughly varies between 40 and 1,843 Euro with an average rent of approximately 459 Euro. For the majority of apartments, the rent varies between 50 and 1,200 Euro. For only a few apartments the monthly rent is higher than 1,200 Euro. This implies that any inference from a regression analysis regarding expensive apartments is comparably uncertain, when compared to the smaller and more modest sized apartments. Generally, the distribution of the monthly net rent is asymmetric and skewed towards the right.

The distribution of the year of construction is highly irregular and multimodal, which is in part due to historical reasons. Whereas the data basis for apartments for the years of the economic crises during the Weimar Constitution and the Second World War is rather

limited, there are much more observations for the later years of reconstruction (mode near 1960). Starting in the mid-1970s the construction boom stopped again. Altogether the data range from 1918 until 1997. Obviously, the 1999 rent index does not allow us to draw conclusions about new buildings after 1997 since there is a temporal gap of more than one year between data collection and the publication of the rent index. Particularly striking is the relative accumulation of apartments constructed in 1918. However, this is a data artifact since all apartments that were built prior to 1918 are antedated to the year 1918.

We leave the interpretation of the distribution of the other continuous variables in the data set to the reader.

Table 1.2 also shows frequency tables for the categorical variables. We observe, for example, that most of the apartments (58 %) are located in an average location. Only about 3 % of the apartments are to be found in top locations.

△

Example 1.6 Malnutrition in Zambia—Univariate Distributions

In addition to Table 1.3 (p. 7), Fig. 1.4 provides a visual overview of the distribution of the response variable and selected continuous explanatory variables using histograms and kernel density estimators. We provide detailed interpretations in our case study in Sect. 9.8. Note that for some variables (duration of breast-feeding and child's age in months) the kernel density estimate shows artifacts in the sense that the density is positive for values lower than zero. However, for the purpose of getting an overview of the variables, this somewhat unsatisfactory behavior is not problematic.

△

1.2.2 Graphical Association Analysis

In a second step, we can graphically investigate the relationship between the response variable and the explanatory variables, at least for continuous responses. By doing so, we get a first overview regarding the type (e.g., linear versus nonlinear) and strength of the relationship between the response variable and the explanatory variables. In most cases, we focus on bivariate analyses (between the response and one explanatory variable). In the following we assume a continuous response variable.

The appropriateness of graphical tools depends on whether the explanatory variable is continuous or categorical.

Continuous Explanatory Variables

As already used by Galton at the end of the nineteenth century, simple scatter plots can provide useful information about the relationship between the response variable and the explanatory variables.

Example 1.7 Munich Rent Index—Scatter Plots

Figure 1.5 shows for the rent index data scatter plots between net rent or net rent per square meter and the continuous explanatory variables living area and year of construction. A first impression is that the scatter plots are not very informative which is a general problem with large sample sizes (in our case more than 3,000 observations). We do find some evidence of an approximately linear relationship between net rent and living area. We also notice that the variability of the net rent increases with an increased living area. The relationship

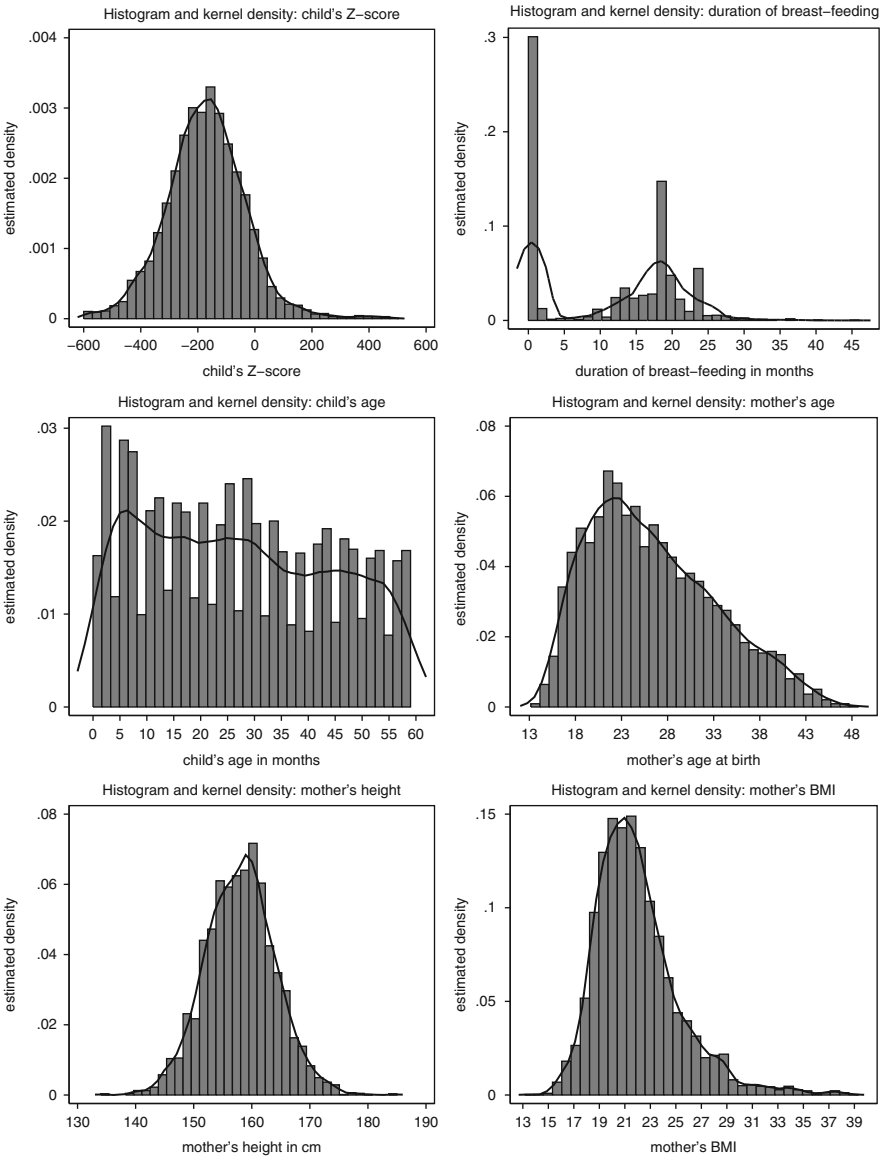

Fig. 1.4 Malnutrition in Zambia: histograms and kernel density estimators for the continuous variables

between net rent per square meter and living area is more difficult to determine. Generally the net rent per square meter for larger apartments seems to decrease. It is however difficult to assess the type of relationship (linear or nonlinear). The relationship of either of the two response variables and the year of construction is again hardly visible (if it exists at all), but

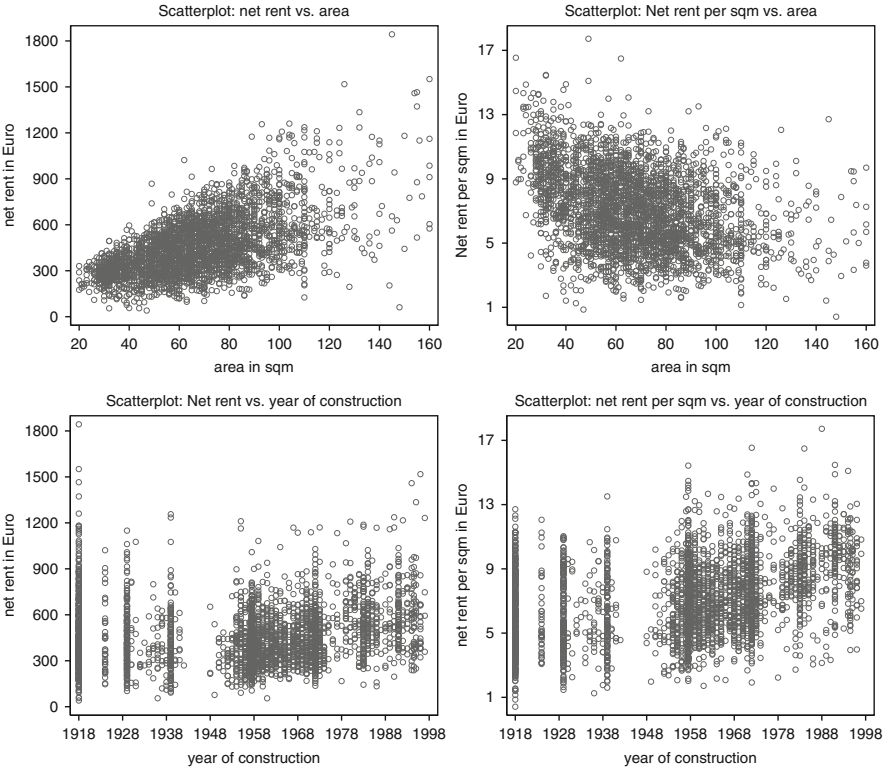

Fig. 1.5 Munich rent index: scatter plots between net rent (*left*) / net rent per sqm (*right*) and the covariates area and year of construction

there is at least evidence for a monotonic increase of rents (and rents per square meter) for flats built after 1948.

△

The preceding example shows that for large sample sizes simple scatter plots do not necessarily contain much information. In this situation, it can be useful to *cluster* the data. If the number of *different* values of the explanatory variable is relatively small in comparison to the sample size, we can summarize the response with the mean value and the corresponding standard deviation for each observed level of the explanatory variable and then visualize these in a scatter plot. Alternatively we could visualize the cluster medians together with the 25 % and 75 % quantiles (or any other combination of quantiles). The resulting data reduction often makes it easier to detect relationships. If the number of different levels of the explanatory variables is large relative to the sample size, it can be useful to cluster or categorize the data. More specifically, we divide the range of values of the explanatory variable into small intervals and calculate mean and standard deviation of the aggregated

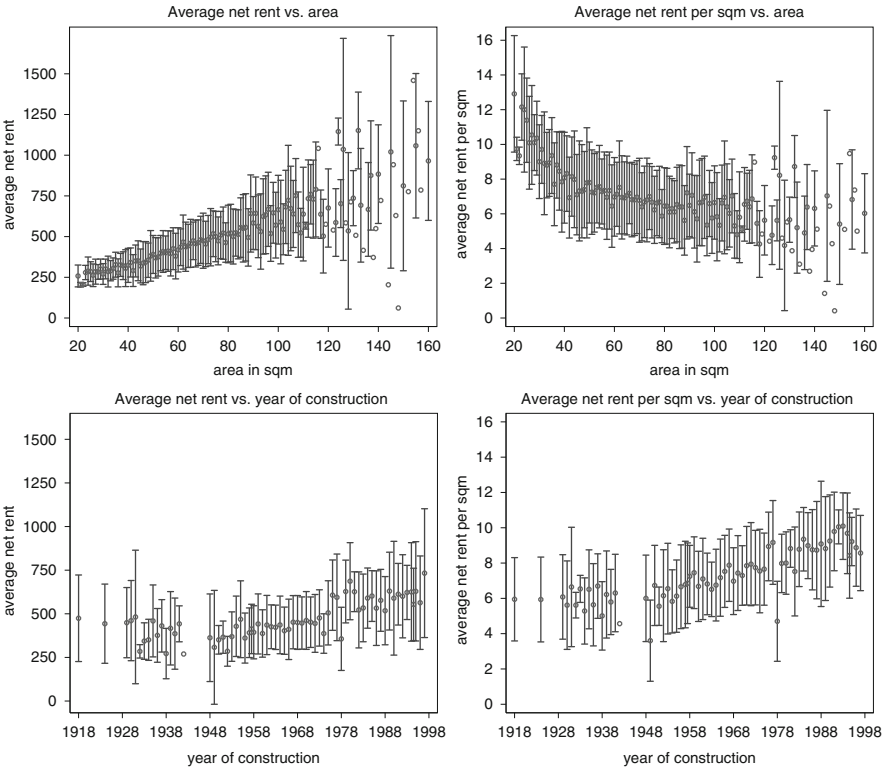

Fig. 1.6 Munich rent index: average net rent (*left*) and net rent per sqm (*right*) plus/minus one standard deviation versus area and year of construction

response for each interval separately. The cluster mean plus/minus one standard deviation is next combined into a scatter plot.

Example 1.8 Munich Rent Index—Clustered Scatter Plots

Living area and year of construction are measured in square meters and years, respectively. In both cases the units of measurement provide a natural basis for clustering. It is thus possible to calculate and visualize the mean values and standard deviations for either of the net rent responses clustered either by living area or year of construction (see Fig. 1.6). Compared to Fig. 1.5 it is now easier to make statements regarding possible relationships that may exist. If we take, e.g., the net rent per square meter as the response variable, a clear nonlinear and monotonically decreasing relationship with the living area becomes apparent. For large apartments (120 square meters or larger), we can also see a clear increase in the variability of average rents.

It also appears that there exists a relationship between the year of construction and the net rent per square meter, even though the relationship seems to be much weaker. Again the relationship is nonlinear: for apartments that were constructed prior to 1940, the rent per square meter is relatively constant (about 6 Euro). On average the rent appears somewhat lower for the few apartments from the sample taken from the years of the war. After

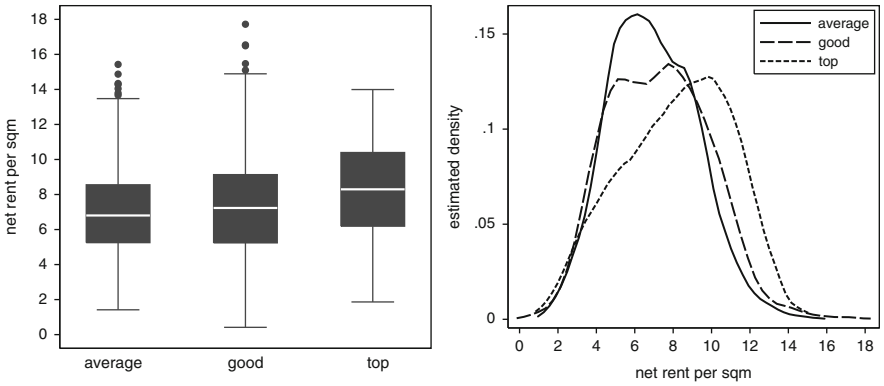

Fig. 1.7 Munich rent index: distribution of net rent per sqm clustered according to location

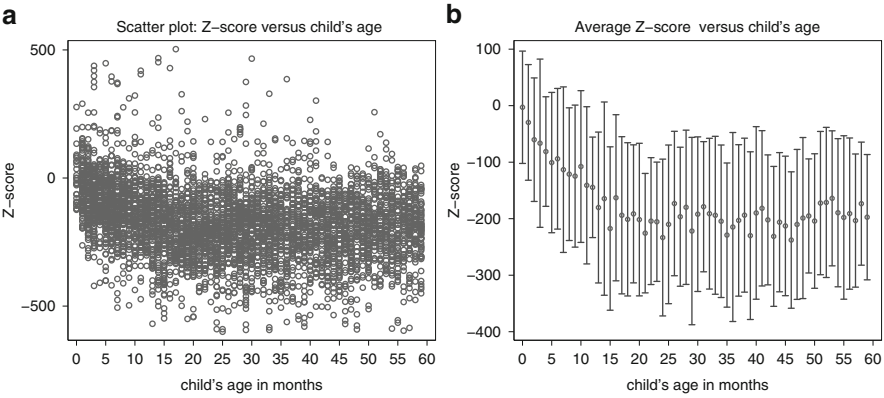

Fig. 1.8 Malnutrition in Zambia: different visualizations of the relationship between Z-score and child's age

1945, the average rent per square meter shows a linearly increasing trend with year of construction.

△

Categorical Explanatory Variables

Visualizing the relationship between a continuous response variable and categorical explanatory variables can be obtained by summarizing the response variable at each level of the categorical variable. Histograms, box plots, and (kernel) density estimators are all adequate means of illustration. In many cases, box plots are best suited as differences in mean values (measured through the median) can be well detected.

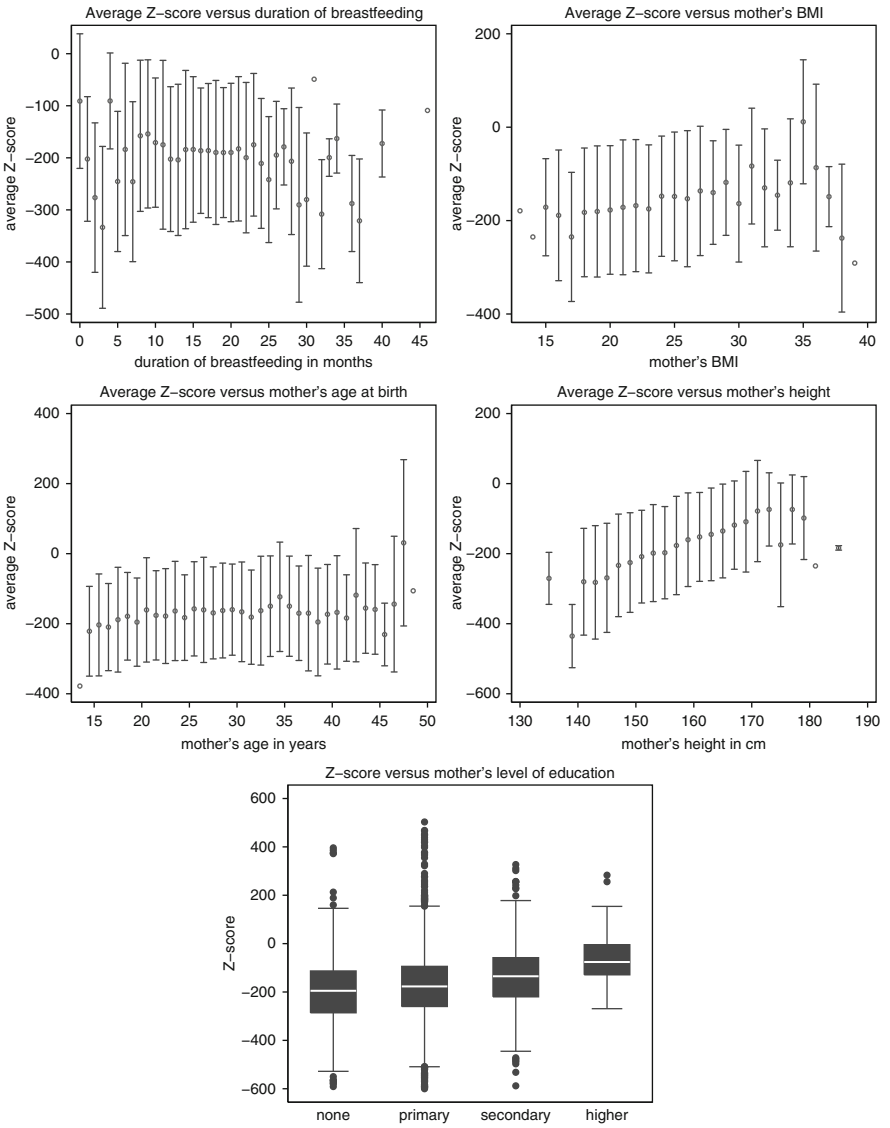

Fig. 1.9 Malnutrition in Zambia: visualization of the relationship between Z-score and selected explanatory variables

Example 1.9 Munich Rent Index—Categorical Explanatory Variables

Figure 1.7 illustrates the distribution of the net rent per square meter as the location (average, good, top) of the apartment is varied. The left panel uses box plots for illustration, and the right panel uses kernel density estimators. The box plots clearly show how the median rent (as well as the variation) increases as the location of the apartment improves.

Even though the smooth density estimators offer similar information, the visualization of these findings is not as obvious as for box plots.

△

Example 1.10 Malnutrition in Zambia—Graphical Association Analysis

Figures 1.8 and 1.9 offer a graphical illustration of the relationship between Z-scores and various explanatory variables. Similar to the rent data, the relationship between the Z-score and the age of the child is difficult to visualize (Fig. 1.8, left panel). A better choice of illustration is obtained when clustering the Z-scores by monthly age of the children (0 to 59 months). For each month, the mean plus/minus standard deviation of Z-scores is computed and plotted (right panel), which provides a much clearer picture of the relationship between Z-score and age. This type of illustration is also used for the other continuous explanatory variables, see Fig. 1.9. We will provide detailed interpretations of Figs. 1.8 and 1.9 in our case study on malnutrition in Zambia in Sect. 9.8.

△

1.3 Notational Remarks

Before we give an overview of regression models in the next chapter some remarks on notation are in order.

In introductory textbooks on statistics authors usually distinguish notationally between random variables and their realizations (the observations). Random variables are denoted by upper case letters, e.g., X, Y, while realizations are denoted by lower case letters, e.g., x, y. However, in more advanced textbook, in particular books on regression analysis, random variables and their realizations are usually *not* distinguished and both denoted by lower case letters, i.e., x, y. It then depends on the context whether y denotes the random variable or the realization. In this book we will keep this convention with the exception of Appendix B which introduces some concepts of probability and statistics. Here we will distinguish between random variables and realizations notationally in the way described above, i.e., by denoting random variables as capital letters and realizations as lower case letters.