1

Regression and the Normal Distribution

Chapter Preview. Regression analysis is a statistical method that is widely used in many fields of study, with actuarial science being no exception. This chapter provides an introduction to the role of the normal distribution in regression, the use of logarithmic transformations in specifying regression relationships, and the sampling basis that is critical for inferring regression results to broad populations of interest.

1.1 What Is Regression Analysis?

Statistics is about data. As a discipline, it is about the collection, summarization, and analysis of data to make statements about the real world. When analysts collect data, they are really collecting information that is quantified, that is, transformed to a numerical scale. There are easy, well-understood rules for reducing the data, through either numerical or graphical summary measures. These summary measures can then be linked to a theoretical representation, or model, of the data. With a model that is calibrated by data, statements about the world can be made.

Statistics is about the collection, summarization, and analysis of data to make statements about the real world.

Statistical methods have had a major impact on several fields of study:

- In the area of data collection, the careful design of *sample surveys* is crucial to market research groups and to the auditing procedures of accounting firms.
- *Experimental design* is a subdiscipline devoted to data collection. The focus of experimental design is on constructing methods of data collection that will extract information in the most efficient way possible. This is especially important in fields such as agriculture and engineering where each observation is expensive, possibly costing millions of dollars.
- Other applied statistical methods focus on managing and predicting data. *Process control* deals with monitoring a process over time and deciding when intervention is most fruitful. Process control helps manage the quality of goods produced by manufacturers.
- *Forecasting* is about extrapolating a process into the future, whether it be sales of a product or movements of an interest rate.

Regression analysis is a statistical method used to analyze data. As we will see, the distinguishing feature of this method is the ability to make statements about

Table 1.1 Galton's 1885 Regression Data

Height of Adult Child in Inches	Parents' Height											
	<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	—	—	—	—	—	—	5	3	2	4	—	14
73.2	—	—	—	—	—	3	4	3	2	2	3	17
72.2	—	—	1	—	4	4	11	4	9	7	1	41
71.2	—	—	2	—	11	18	20	7	4	2	—	64
70.2	—	—	5	4	19	21	25	14	10	1	—	99
69.2	1	2	7	13	38	48	33	18	5	2	—	167
68.2	1	—	7	14	28	34	20	12	3	1	—	120
67.2	2	5	11	17	38	31	27	3	4	—	—	138
66.2	2	5	11	17	36	25	17	1	3	—	—	117
65.2	1	1	7	2	15	16	4	1	1	—	—	48
64.2	4	4	5	5	14	11	16	—	—	—	—	59
63.2	2	4	9	3	5	7	1	1	—	—	—	32
62.2	—	1	—	3	3	—	—	—	—	—	—	7
<61.2	1	1	1	—	—	1	—	1	—	—	—	5
Total	14	23	66	78	211	219	183	68	43	19	4	928

Source: Stigler (1986).

variables after having controlled for values of known explanatory variables. As important as other methods are, it is regression analysis that has been the most influential one. To illustrate, an index of business journals, ABI/INFORM, lists more than 24,000 articles using regression techniques over the thirty-year period 1978–2007. And these are only the applications that were considered innovative enough to be published in scholarly reviews!

Regression analysis of data is so pervasive in modern business that it is easy to overlook the fact that the methodology is barely more than 120 years old. Scholars attribute the birth of regression to the 1885 presidential address of Sir Francis Galton to the anthropological section of the British Association of the Advancement of Sciences. In that address, described in Stigler (1986), Galton provided a description of regression and linked it to *normal curve* theory. His discovery arose from his studies of properties of natural selection and inheritance.

® **Empirical** *Filename is "Galton"*

To illustrate a dataset that can be analyzed using regression methods, Table 1.1 displays some data included in Galton's 1885 paper. The table displays the heights of 928 adult children, classified by an index of their parents' height. Here, all female heights were multiplied by 1.08, and the index was created by taking the average of the father's height and rescaled mother's height. Galton was aware that the parents' and the adult child's height could each be adequately approximated by a normal curve. In developing regression analysis, he provided a single model for the joint distribution of heights.

Regression analysis is a method to quantify the relationship between a variable of interest and explanatory variables.

Table 1.1 shows that much of the information concerning the height of an adult child can be attributed to, or "explained," in terms of the parents' height. Thus, we use the term *explanatory variable* for measurements that provide information

Figure 1.1 Ten deutsche mark – German currency featuring the scientist Gauss and the normal curve.

about a variable of interest. Regression analysis is a method to quantify the relationship between a variable of interest and explanatory variables. The methodology used to study the data in Table 1.1 can also be used to study actuarial and other risk management problems, the thesis of this book.

1.2 Fitting Data to a Normal Distribution

Historically, the normal distribution had a pivotal role in the development of regression analysis. It continues to play an important role, although we will be interested in extending regression ideas to highly "nonnormal" data.

Formally, the normal curve is defined by the function

$$\mathrm{f}(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right). \tag{1.1}$$

This curve is a probability density function with the whole real line as its domain. From equation (1.1), we see that the curve is symmetric about μ (the mean and median). The degree of peakedness is controlled by the parameter σ^2. These two parameters, μ and σ^2, are known as the *location* and *scale parameters*, respectively. Appendix A3.1 provides additional details about this curve, including a graph and tables of its cumulative distribution that we will use throughout the text.

Appendix A3.1 provides additional details about the normal curve, including a graph and distribution table.

The normal curve is also depicted in Figure 1.1, a display of an out-of-date German currency note, the ten Deutsche Mark. This note contains the image of the German Carl Gauss, an eminent mathematician whose name is often linked with the normal curve (it is sometimes referred to as the *Gaussian curve*). Gauss developed the normal curve in connection with the theory of least squares for fitting curves to data in 1809, about the same time as related work by the French scientist Pierre LaPlace. According to Stigler (1986), there was quite a bit of acrimony between these two scientists about the priority of discovery! The normal curve was first used as an approximation to histograms of data around 1835 by Adolph Quetelet, a Belgian mathematician and social scientist. As with many good things, the normal curve had been around for some time, since about 1720, when Abraham de Moivre derived it for his work on modeling games of

Table 1.2 Summary Statistics of Massachusetts Automobile Bodily Injury Claims

Variable	Number	Mean	Median	Standard Deviation	Minimum	Maximum	25th Percentile	75th Percentile
Claims	272	0.481	0.793	1.101	−3.101	3.912	−0.114	1.168

Note: Data are in logs of thousands of dollars.

chance. The normal curve is popular because it is easy to use and has proved successful in many applications.

® **Empirical**
Filename is "MassBodilyInjury"

Example: Massachusetts Bodily Injury Claims. For our first look at fitting the normal curve to a set of data, we consider data from Rempala and Derrig (2005). They considered claims arising from automobile bodily injury insurance coverages. These are amounts incurred for outpatient medical treatments that arise from automobile accidents, typically sprains, broken collarbones, and the like. The data consist of a sample of 272 claims from Massachusetts that were closed in 2001 (by "closed," we mean that the claim is settled and no additional liabilities can arise from the same accident). Rempala and Derrig were interested in developing procedures for handling mixtures of "typical" claims and others from providers who reported claims fraudulently. For this sample, we consider only those typical claims, ignoring the potentially fraudulent ones.

Table 1.2 provides several statistics that summarize different aspects of the distribution. Claim amounts are in units of logarithms of thousands of dollars. The average logarithmic claim is 0.481, corresponding to \$1,617.77 (=1000 exp(0.481)). The smallest and largest claims are −3.101 (\$45) and 3.912 (\$50,000), respectively.

For completeness, here are a few definitions. The *sample* is the set of data available for analysis, denoted by $y_1, \ldots, y_n$. Here, n is the number of observations, y_1 represents the first observation, y_2 the second, and so on up to y_n for the nth observation. Here are a few important summary statistics.

Basic Summary Statistics

(i) The *mean* is the average of observations, that is, the sum of the observations divided by the number of units. Using algebraic notation, the mean is

$$\overline{y} = \frac{1}{n}(y_1 + \cdots + y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

(ii) The *median* is the middle observation when the observations are ordered by size. That is, it is the observation at which 50% are below it (and 50% are above it).

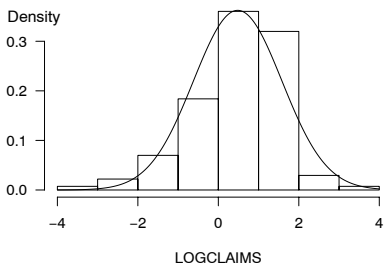

Figure 1.2 Bodily injury relative frequency with normal curve superimposed.

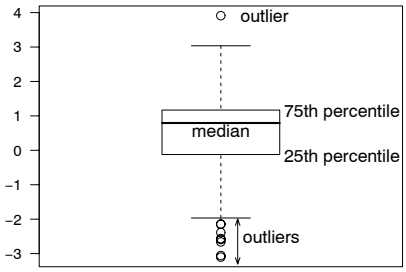

Figure 1.3 Box plot of bodily injury claims.

(iii) The *standard deviation* is a measure of the spread, or scale, of the distribution. It is computed as

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

(iv) A *percentile* is a number at which a specified fraction of the observations is below it, when the observations are ordered by size. For example, the 25th percentile is the number so that 25% of observations are below it.

To help visualize the distribution, Figure 1.2 displays a *histogram* of the data. Here, the height of the each rectangle shows the relative frequency of observations that fall within the range given by its base. The histogram provides a quick visual impression of the distribution; it shows that the range of the data is approximately (−4,4), that the central tendency is slightly greater than zero, and that the distribution is roughly symmetric.

Normal Curve Approximation

Figure 1.2 also shows a normal curve superimposed, using $\overline{y}$ for μ and s_y^2 for σ^2. With the normal curve, only two quantities (μ and σ^2) are required to summarize the entire distribution. For example, Table 1.2 shows that 1.168 is the 75th percentile, which is approximately the 204th ($= .75 \times 272$) largest observation from the entire sample. From the equation (1.1) normal distribution, we see that $z = (y - \mu)/\sigma$ is a standard normal, of which 0.675 is the 75th percentile. Thus, $\overline{y} + 0.675 s_y = 0.481 + 0.675 \times 1.101 = 1.224$ is the 75th percentile using the normal curve approximation.

Box Plot

A quick visual inspection of a variable's distribution can reveal some surprising features that are hidden by statistics: numerical summary measures. The *box plot*, also known as a box-and-whiskers plot, is one such graphical device. Figure 1.3 illustrates a box plot for the bodily injury claims. Here, the box captures the

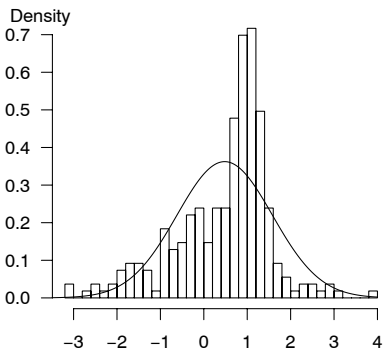

Figure 1.4 Redrawing of Figure 1.2 with an increased number of rectangles.

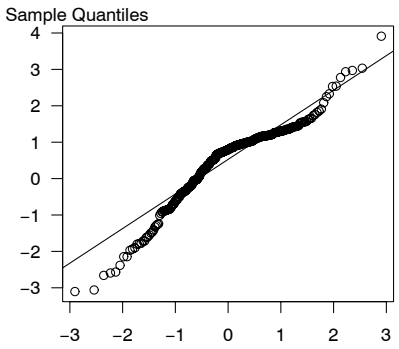

Figure 1.5 A *qq* plot of bodily injury claims, using a normal reference distribution.

middle 50% of the data, with the three horizontal lines corresponding to the 75th, 50th, and 25th percentiles, reading from top to bottom. The horizontal lines above and below the box are the "whiskers." The upper whisker is 1.5 times the *interquartile range* (the difference between the 75th and 25th percentiles) above the 75th percentile. Similarly, the lower whisker is 1.5 times the interquartile range below the 25th percentile. Individual observations outside the whiskers are denoted by small circular plotting symbols and are referred to as "outliers."

Graphs are powerful tools; they allow analysts to readily visualize nonlinear relationships that are hard to comprehend when expressed verbally or by mathematical formula. However, by their very flexibility, graphs can also readily deceive the analyst. Chapter 21 will underscore this point. For example, Figure 1.4 is a redrawing of Figure 1.2; the difference is that Figure 1.4 uses more, and finer, rectangles. This finer analysis reveals the asymmetric nature of the sample distribution that was not evident in Figure 1.2.

Quantile-Quantile Plots

Increasing the number of rectangles can unmask features that were not previously apparent; however, there are, in general, fewer observations per rectangle, meaning that the uncertainty of the relative frequency estimate increases. This represents a trade-off. Instead of forcing the analyst to make an arbitrary decision about the number of rectangles, an alternative is to use a graphical device for comparing a distribution to another known as a *quantile-quantile*, or *qq*, plot.

Points in a qq *plot close to a straight line suggest agreement between the sample and the reference distributions.*

Figure 1.5 illustrates a *qq* plot for the bodily injury data using the normal curve as a reference distribution. For each point, the vertical axis gives the quantile using the sample distribution. The horizontal axis gives the corresponding quantity using the normal curve. For example, earlier we considered the 75th percentile point. This point appears as (1.168, 0.675) on the graph. To interpret a *qq* plot, if the quantile points lie along the superimposed line, then the sample and the normal reference distribution have the same shape. (This line is defined by connecting the 75th and 25th percentiles.)

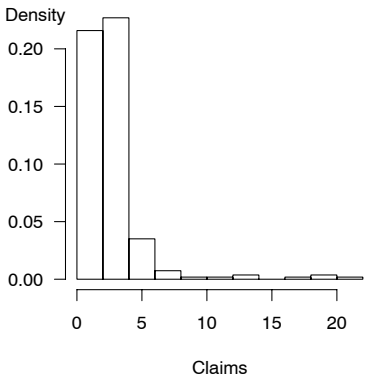

Figure 1.6 Distribution of bodily injury claims. Observations are in (thousands of) dollars, with the largest observation omitted.

In Figure 1.5, the small sample percentiles are consistently smaller than the corresponding values from the standard normal, indicating that the distribution is skewed to the left. The difference in values at the ends of the distribution are due to the outliers noted earlier that can also be interpreted as the sample distribution having larger tails than the normal reference distribution.

1.3 Power Transforms

In the Section 1.2 example, we considered claims without justifying the use of the logarithmic scaling. When analyzing variables such as assets of firms, wages of individuals, and housing prices of households in business and economic applications, it is common to consider logarithmic units instead of the original units. A log transform retains the original ordering (e.g., large wages remain large on the log wage scale) but serves to pull in extreme values of the distribution.

To illustrate, Figure 1.6 shows the bodily injury claims distribution in (thousands of) dollars. To graph the data meaningfully, the largest observation ($50,000) was removed prior to making this plot. Even with this observation removed, Figure 1.6 shows that the distribution is heavily lopsided to the right, with several large values of claims appearing.

A right-skewed distribution has long tails on the right and a concentration of mass on the left. Many insurance claims distributions are right skewed.

Distributions that are lopsided in one direction or the other are known as *skewed*. Figure 1.6 is an example of a distribution skewed to the right, or positively skewed. Here, the tail of the distribution on the right is longer, and there is a greater concentration of mass to the left. In contrast, a left-skewed, or negatively skewed, distribution has a longer tail on the left and a greater concentration of mass to the right. Many insurance claims distributions are right skewed (see Klugman, Panjer, and Willmot, 2008, for extensive discussions). As we saw in Figures 1.4 and 1.5, a logarithmic transformation yields a distribution that is only mildly skewed to the left.

Logarithmic transformations are used extensively in applied statistics work. One advantage is that they serve to symmetrize distributions that are skewed. More generally, we consider *power transforms*, also known as the *Box-Cox family of*

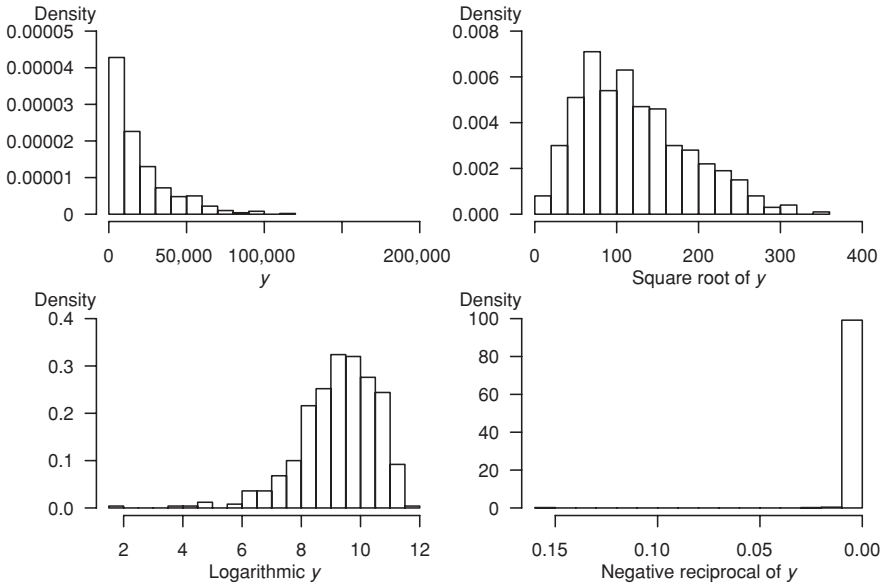

Figure 1.7 500 simulated observations from a chi-square distribution. The upper-left panel is based on the original distribution. The upper-right panel corresponds to the square root transform, the lower left to the log transform, and the lower right to the negative reciprocal transform.

transforms. In this family of transforms, in lieu of using the response y, we use a transformed, or rescaled version, y^{λ}. Here, the power λ (lambda, a Greek letter "el") is a number that may be user specified. Typical values of λ that are used in practice are $\lambda = 1, 1/2, 0$, or -1. When we use $\lambda = 0$, we mean $\ln(y)$, that is, the natural logarithmic transform. More formally, the Box-Cox family can be expressed as

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(y) & \lambda = 0 \end{cases}.$$

As we will see, because regression estimates are not affected by location and scale shifts, in practice, we do not need to subtract 1 or divide by λ when rescaling the response. The advantage of the foregoing expression is that, if we let λ approach 0, then $y^{(\lambda)}$ approaches $\ln(y)$, from some straightforward calculus arguments.

To illustrate the usefulness of transformations, we simulated 500 observations from a chi-square distribution with two degrees of freedom. Appendix A3.2 introduces this distribution (which we will encounter again in studying the behavior of test statistics). The upper-left panel of Figure 1.7 shows that the original distribution is heavily skewed to the right. The other panels in Figure 1.7 show the data rescaled using the square root, logarithmic, and negative reciprocal transformations. The logarithmic transformation, in the lower-left panel, provides the best approximation to symmetry for this example. The negative reciprocal transformation is based on $\lambda = -1$, and then multiplying the rescaled observations by -1, so that large observations remain large.

1.4 Sampling and the Role of Normality

A *statistic* is a summary measure of data, such as a mean, median, or percentile. Collections of statistics are very useful for analysts, decision makers, and

everyday consumers for understanding massive amounts of data that represent complex situations. To this point, our focus has been on introducing sensible techniques to summarize variables; techniques that will be used repeatedly thought this text. However, the true usefulness of the discipline of statistics is its ability to say something about the unknown, not merely to summarize information already available. To this end, we need to make some fairly formal assumptions about the manner in which the data are observed. As a science, a strong feature of the discipline of statistics is the ability to critique these assumptions and offer improved alternatives in specific situations.

A statistic is a summary measure of a sample. Statistics, as a discipline, can be used to infer behavior about a larger population from a sample.

It is customary to assume that the data are drawn from a larger population that we are interested in describing. The process of drawing the data is known as the *sampling*, or *data generating*, *process*. We denote this sample as $\{y_1, \ldots, y_n\}$. So that we may critique, and modify, these sampling assumptions, we list them here in detail:

Basic Sampling Assumptions

1. E $y_i = \mu$.
2. Var $y_i = \sigma^2$.
3. $\{y_i\}$ are independent.
4. $\{y_i\}$ are normally distributed.

Assumption 4 is not required for many statistical inference procedures because central limit theorems provide approximate normality for many statistics of interest.

In this basic setup, μ and σ^2 serve as *parameters* that describe the location and scale of the parent population. The goal is to infer something sensible about them on the basis of statistics such as $\overline{y}$ and s_y^2. For the third assumption, we assume independence among the draws. In a sampling scheme, this may be guaranteed by taking a simple random sample from a population. The fourth assumption is not required for many statistical inference procedures because central limit theorems provide approximate normality for many statistics of interest. However, a formal justification of some statistics, such as t-statistics, requires this additional assumption.

Section 1.9 provides an explicit statement of one version of the central limit theorem, giving conditions in which $\overline{y}$ is approximately normally distributed. This section also discusses a related result, known as an *Edgeworth approximation*, that shows that the quality of the normal approximation is better for symmetric parent populations when compared to skewed distributions.

Linear regression is the study of weighted averages.

How does this discussion apply to the study of regression analysis? After all, so far we have focused only on the simple arithmetic average $\overline{y}$. In subsequent chapters, we will emphasize that linear regression is the study of weighted averages; specifically, many regression coefficients can be expressed as weighted averages with appropriately chosen weights. Central limit and Edgeworth approximation theorems are available for weighted averages – these results will ensure approximate normality of regression coefficients. To use normal curve approximations in a regression context, we will often transform variables to achieve approximate symmetry.

Table 1.3 Terminology for Regression Variables

y-Variable	x-Variable
Outcome of interest	Explanatory variable
Dependent variable	Independent variable
Endogenous variable	Exogenous variable
Response	Treatment
Regressand	Regressor
Left-hand-side variable	Right-hand-side variable
Explained variable	Predictor variable
Output	Input

1.5 Regression and Sampling Designs

We often transform variables to achieve approximate symmetry to use normal curve approximations in a regression context.

Approximating normality is an important issue in practical applications of linear regression. Parts I and II of this book focus on linear regression, where we will learn basic regression concepts and sampling design. Part III will focus on *nonlinear* regression, involving binary, count, and fat-tailed responses, where the normal is not the most helpful reference distribution. Ideas concerning basic concepts and design are also used in the nonlinear setting.

In regression analysis, we focus on one measurement of interest: the *dependent variable*. Other measurements are used as *explanatory variables*. A goal is to compare differences in the dependent variable in terms of differences in the explanatory variables. As noted in Section 1.1, regression is used extensively in many scientific fields. Table 1.3 lists alternative terms that you may encounter as you read regression applications.

In the latter part of the nineteenth century and early part of the twentieth century, statistics was beginning to make an important impact on the development of experimental science. Experimental sciences often use *designed studies*, where the data are under the control of an analyst. Designed studies are performed in laboratory settings, where there are tight physical restrictions on every variable that a researcher thinks may be important. Designed studies also occur in larger field experiments, where the mechanisms for control are different than in laboratory settings. Agriculture and medicine use designed studies. Data from a designed study are said to be *experimental data*.

In designed studies, the data are under the control of an analyst. Data from a designed study are said to be experimental data.

To illustrate, a classic example is to consider the yield of a crop such as corn, where each of several parcels of land (the observations) are assigned various levels of fertilizer. The goal is ascertain the effect of fertilizer (the explanatory variable) on the corn yield (the response variable). Although researchers attempt to make parcels of land as much alike as possible, differences inevitably arise. Agricultural researchers use *randomization techniques* to assign different levels of fertilizer to each parcel of land. In this way, analysts can explain the variation in corn yields in terms of the variation of fertilizer levels. Through the use of randomization techniques, researchers using designed studies can infer that the treatment has a *causal effect* on the response. Chapter 6 discusses causality further.