11

Regression: Advanced Topics

11.1 The Theory Behind Linear Regression

This section provides some theoretical results about linear least-squares estimation. The study of linear regression is facilitated by the use of matrices. Equation (9.1) can be written more succinctly as

$$Y_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n \tag{11.1}$$

where $\boldsymbol{x}_i = (1\ X_{i,1}\ \cdots\ X_{i,p})^{\mathsf{T}}$ and $\boldsymbol{\beta} = (\beta_0\ \beta_1\ \cdots\ \beta_p)^{\mathsf{T}}$. Let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}, \text{ and } \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Then, the n equations in (11.1) can be expressed as

$$\underbrace{\boldsymbol{Y}}_{n\times 1} = \underbrace{\boldsymbol{X}}_{n\times(p+1)}\underbrace{\boldsymbol{\beta}}_{(p+1)\times 1} + \underbrace{\boldsymbol{\epsilon}}_{n\times 1}, \tag{11.2}$$

with the matrix dimensions indicated by underbraces.

The least-squares estimate of $\boldsymbol{\beta}$ minimizes

$$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y} - 2\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Y} + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{\beta}. \tag{11.3}$$

By setting the derivatives of (11.3) with respect to $\beta_0, \ldots, \beta_p$ equal to 0 and simplifying the resulting equations, one finds that the least-squares estimator is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Y}. \tag{11.4}$$

D. Ruppert, D.S. Matteson, *Statistics and Data Analysis for Financial Engineering*, Springer Texts in Statistics,
DOI 10.1007/978-1-4939-2614-5_11

Using (7.9), one can find the covariance matrix of $\widehat{\boldsymbol{\beta}}$:

$$\begin{aligned}\mathrm{COV}(\widehat{\boldsymbol{\beta}}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) &= (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\mathrm{COV}(\boldsymbol{Y}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\\ &= (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}(\sigma_\epsilon^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\\ &= \sigma_\epsilon^2(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1},\end{aligned}$$

since $\mathrm{COV}(\boldsymbol{Y}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = \mathrm{COV}(\boldsymbol{\epsilon}) = \sigma_\epsilon^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix. Therefore, the standard error of $\widehat{\beta}_j$ is the square root of the jth diagonal element of $\sigma_\epsilon^2(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}$.

The vector of fitted values is

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \{\boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\}\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y},$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}$ is the *hat matrix.* The leverage of the ith observation is H_{ii}, the ith diagonal element of $\boldsymbol{H}$.

11.1.1 Maximum Likelihood Estimation for Regression

In this section, we assume a linear regression model with noise that may not be normally distributed and independent.

For example, consider the special case of i.i.d. errors. It is useful to put the scale parameter explicitly into the regression model, so we assume that

$$Y_i = \boldsymbol{x}_i^{\mathsf{T}}\beta + \sigma\epsilon_i,$$

where $\{\epsilon_i\}$ are i.i.d. with a known density f that has variance equal to 1 and σ is the unknown noise standard deviation. For example, f could be a standardized t-density. Then the likelihood of $Y_1,\ldots,Y_n$ is

$$\prod_{i=1}^n \frac{1}{\sigma} f\left\{\frac{Y_i - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}{\sigma}\right\}.$$

The maximum likelihood estimator maximizes the log-likelihood

$$L(\boldsymbol{\beta},\sigma) = -n\log(\sigma) + \sum_{i=1}^n \log\left[f\left\{\frac{Y_i - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}{\sigma}\right\}\right].$$

For normally distributed errors, $\log\{f(x)\} = -\frac{1}{2}x^2 - \frac{1}{2}\log(2\pi)$, and for the purpose of maximization, the constant $-\frac{1}{2}\log(2\pi)$ can be ignored. Therefore, the log-likelihood is

$$L^{\mathrm{GAUSS}}(\boldsymbol{\beta},\sigma) = -n\log(\sigma) - \frac{1}{2}\sum_{i=1}^n \left(\frac{Y_i - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}{\sigma}\right)^2.$$

It should be obvious that the least-squares estimator is the MLE of $\boldsymbol{\beta}$. Also, maximizing $L^{\mathrm{GAUSS}}(\widehat{\boldsymbol{\beta}},\sigma)$ in σ, where $\boldsymbol{\beta}$ has been replaced by the least-squares estimate, is a standard calculus exercise and the result is

$$\widehat{\sigma}^2_{\mathrm{MLE}} = n^{-1} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^{\mathsf{T}} \widehat{\boldsymbol{\beta}})^2.$$

In can be shown that $\widehat{\sigma}^2_{\mathrm{MLE}}$ is biased but that the bias is eliminated if n^{-1} is replaced by $\{n - (p+1)\}^{-1}$ where $p+1$ is the dimension of $\boldsymbol{\beta}$. This give us the estimator (9.16).

Now assume that $\boldsymbol{\epsilon}$ has a covariance matrix $\boldsymbol{\Sigma}$ and, for some function f, density

$$|\boldsymbol{\Sigma}|^{-1/2} f\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\}.$$

Then the log-likelihood is

$$-\frac{1}{2} \log |\boldsymbol{\Sigma}| + \log \left[f\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\} \right].$$

In the important special case where $\boldsymbol{\epsilon}$ has a mean-zero multivariate normal distribution, the density of $\boldsymbol{\epsilon}$ is

$$\left[\frac{1}{|\boldsymbol{\Sigma}|^{1/2} (2\pi)^{p/2}} \right] \exp\left\{ -\frac{1}{2} \boldsymbol{\epsilon}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\epsilon} \right\}, \tag{11.5}$$

If $\boldsymbol{\Sigma}$ is known, then the MLE of $\boldsymbol{\beta}$ minimizes

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

and is called the *generalized least-squares estimator* (GLS estimator). If $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated but with possibly different variances, then $\boldsymbol{\Sigma}$ is the diagonal matrix of these variances and the generalized least-squares estimator is the weighted least-squares estimator (10.4).

The GLS estimator is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (\boldsymbol{X}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}. \tag{11.6}$$

Typically, $\boldsymbol{\Sigma}$ is unknown and must be replaced by an estimate, for example, from an ARMA model for the errors.

11.2 Nonlinear Regression

Often we can derive a theoretical model relating predictor variables and a response, but the model we derive is not linear. In particular, models derived from economic theory are commonly used in finance and many are not linear.

The nonlinear regression model is

$$Y_i = f(\boldsymbol{X}_i; \boldsymbol{\beta}) + \epsilon_i, \tag{11.7}$$

where Y_i is the response measured on the ith observation, $\boldsymbol{X}_i$ is a vector of observed predictor variables for the ith observation, $f(\cdot\,;\cdot)$ is a *known* function, $\boldsymbol{\beta}$ is an unknown parameter vector, and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. with mean 0 and variance σ_ϵ^2. The least-squares estimate $\widehat{\boldsymbol{\beta}}$ minimizes

$$\sum_{i=1}^{n} \{Y_i - f(\boldsymbol{X}_i; \boldsymbol{\beta})\}^2 .$$

The predicted values are $\widehat{Y}_i = f(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}})$ and the residuals are $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i$.

Since the model is nonlinear, finding the least-squares estimate requires nonlinear optimization. Because of the importance of nonlinear regression, almost every statistical software package will have routines for nonlinear least-squares estimation. This means that most of the difficult programming has already been done for us. However, we do need to write an equation that specifies the model we are using.[1] In contrast, when using linear regression only the predictor variables need to be specified.

Example 11.1. Simulated bond prices

Consider prices of par \$1000 zero-coupon bonds issued by a particular borrower, perhaps the Federal government or a corporation. Suppose that there are several times to maturity, the ith being denoted by T_i. Suppose also that the yield to maturity is a constant, say r. The assumption that $Y_T = r$ for all T is not realistic and is used only to keep this example simple. In Sect. 11.3 more realistic models will be used.

The rate r is determined by the market and can be estimated from prices. Under the assumption of a constant value of r, the present price of a bond with maturity T_i is

$$P_i = 1000 \exp(-rT_i). \tag{11.8}$$

There is some random variation in the observed prices. One reason is that the price of a bond can only be determined by the sale of the bond, so the observed prices have not been determined simultaneously. Prices that may no longer reflect current market values are called *stale*. Each bond's price was determined at the time of the last trade of a bond of that maturity, and r may have had a somewhat different value then. It is only as a function of time to maturity that r is assumed constant, so r may vary with calendar time. Thus, we augment model (11.8) by including a noise term to obtain the regression model

$$P_i = 1000 \exp(-rT_i) + \epsilon_i. \tag{11.9}$$

An estimate of r can be determined by least squares, that is, by minimizing over r the sum of squares:

$$\sum_{i=1}^{n} \left\{ P_i - 1{,}000 \exp(-rT_i) \right\}^2 .$$

[1] Even this work can sometimes be avoided, since some nonlinear regression software has many standard models already programmed.

The least-squares estimator is denoted by $\widehat{r}$.

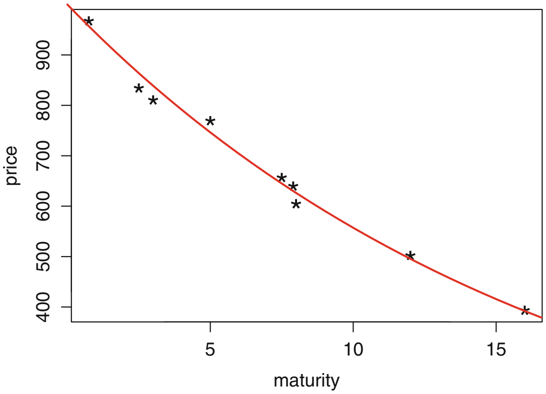

Fig. 11.1. *Plot of bond prices against maturities with the predicted price from the nonlinear least-squares fit.*

Since it is unlikely that market data will have a constant r, this example uses simulated data. The data were generated with r fixed at 0.06 and plotted in Fig. 11.1. The nonlinear least-squares estimate of r was found using R's `nls()` function. Nonlinear optimization requires starting values for the parameters, and a starting value of 0.04 was used for r.

```
bondprices = read.table("bondprices.txt", header = TRUE)
attach(bondprices)
fit = nls(price ~ 1000 * exp(-r * maturity), start = list(r = 0.04))
summary(fit)
detach(bondprices)
```

The output is:

```
Formula: price ~ 1000 * exp(-r * maturity)

Parameters:
  Estimate Std. Error t value Pr(>|t|)
r  0.05850    0.00149    39.3  1.9e-10 ***
---

Residual standard error: 20 on 8 degrees of freedom

Number of iterations to convergence: 4
Achieved convergence tolerance: 5.53e-08
```

Notice that $\widehat{r} = 0.0585$ and the standard error of this estimate is 0.00149. The predicted price curve using nonlinear regression is shown in Fig. 11.1. □

As mentioned, in *nonlinear regression* the form of the regression function is nonlinear but *known* up to a few unknown parameters. For example, the regression function has an exponential form in model (11.9). For this reason, nonlinear regression would best be called *nonlinear parametric regression* to distinguish it from nonparametric regression, where the regression function is also nonlinear but not of a known parametric form. Nonparametric regression is discussed in Chap. 21.

Polynomial regression may appear to be nonlinear since polynomials are nonlinear functions. For example, the quadratic regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \tag{11.10}$$

is nonlinear in X_i. However, by defining X_i^2 as a second predictor variable, this model is linear in (X_i, X_i^2) and therefore is an example of multiple *linear* regression. What makes model (11.10) linear is that the right-hand side is a linear function of the parameters β_0, β_1, and β_2, and therefore can be interpreted as a linear regression with the appropriate definition of the variables. In contrast, the exponential model

$$Y_i = \beta_0 e^{\beta_1 X_i} + \epsilon_i$$

is nonlinear in the parameter β_1, so it cannot be made into a linear model by redefining the predictor variable.

Example 11.2. Estimating default probabilities

This example illustrates both nonlinear regression and the detection of heteroskedasticity by residual plotting.

Credit risk is the risk to a lender that a borrower will default on contractual obligations, for example, that a loan will not be repaid in full. A key parameter in the determination of credit risk is the probability of default. Bluhm, Overbeck, and Wagner (2003) illustrate how one can calibrate Moody's credit rating to estimate default probabilities. These authors use observed default frequencies for bonds in each of 16 Moody's ratings from Aaa (best credit rating) to B3 (worse rating). They convert the credit ratings to a 1 to 16 scale (Aaa = 1, . . . , B3 = 16). Figure 11.2a shows default frequencies (as fractions, not percentages) plotted against the ratings. The data are from Bluhm, Overbeck, and Wagner (2003). The relationship is clearly nonlinear. Not surprisingly, Bluhm, Overbeck, and Wagner used a nonlinear model, specifically

$$Pr\{\text{default}|\text{rating}\} = \exp\{\beta_0 + \beta_1 \text{rating}\}. \tag{11.11}$$

To use this model they fit a linear function to the logarithms of the default frequencies. One difficulty with doing this is that six of the default frequencies are zero giving a log transformation of $-\infty$.

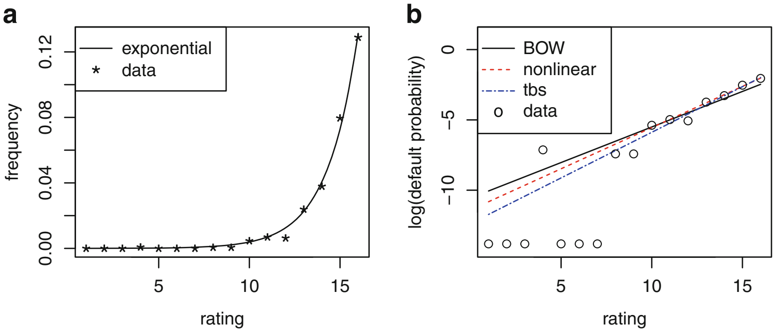

Fig. 11.2. *(a) Default frequencies with an exponential fit. "Rating" is a conversion of the Moody's rating to a* 1 *to* 16*-point scale as follows:* $1 = Aaa$, $2 = Aa1$, $3 = Aa3$, $4 = A1$, $\ldots$, $16 = B3$. *(b) Estimation of default probabilities by Bluhm, Overbeck, and Wagner's* (2003) *linear regression with ratings removed that have no observed defaults (BOW) and by nonlinear regression with all data (nonlinear). Because some default frequencies are zero, when plotting the data on a semilog plot,*10^{-6} *was added to the default frequencies. This constant was* not *added when estimating default frequencies, only for plotting the raw data. The six observations along the bottom of the plot are the ones removed by Bluhm, Overbeck, and Wagner. "TBS" is the transform-both-sides estimate, which will be discussed soon.*

Bluhm, Overbeck, and Wagner (2003) address this issue by labeling default frequencies equal to zero as "unobserved" and not using them in the estimation process. The problem with their technique is that they have deleted the data with the lowest observed default frequencies. This biases their estimates of default probabilities in an upward direction. Bluhm, Overbeck, and Wagner argue that an observed default frequency of zero does not imply that the true default probability is zero. This is certainly true. However, the default frequencies, even when they are zero, are unbiased estimates of the true default probabilities. There is no intent here to be critical of their book, which is well-written and useful. However, one can avoid the bias of their method by using nonlinear regression with model (11.11). The advantage of fitting (11.11) by nonlinear regression is that it avoids the use of a logarithm transformation thus allowing the use of all the data, even data with a default frequency of zero. The fits by the Bluhm, Overbeck, and Wagner method and by nonlinear regression using model (11.11) are shown in Fig. 11.2b with a log scale on the vertical axis so that the fitted functions are linear. Notice that at good credit ratings the estimated default probabilities are lower using nonlinear regression compared to Bluhm, Overbeck, and Wagner's biased method. The differences between the two sets of estimated default probabilities can be substantial. Bluhm, Overbeck, and Wagner estimate the default probability of an Aaa bond as 0.005 %. In contrast, the unbiased estimate by nonlinear regression is only 40 % of that figure, specifically, 0.0020 %. Thus, the bias in the Bluhm, Overbeck, and Wagner estimate leads to a substantial overestimate of the credit risk of Aaa bonds and similar overestimation at other good credit ratings.

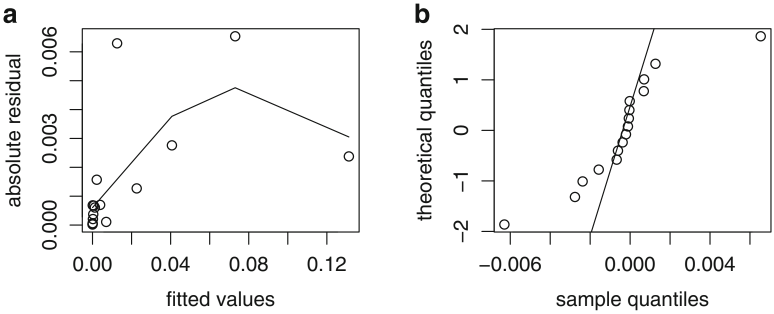

Fig. 11.3. *(a) Residuals for estimation of default probabilities by nonlinear regression. Absolute studentized residuals plotted against fitted values with a loess smooth. Substantial heteroskedasticity is indicated because the data on the left side are less scattered than elsewhere. (b) Normal probability plot of the residuals. Notice the outliers caused by the nonconstant variance.*

A plot of the absolute residuals versus the fitted values in Fig. 11.3a gives a clear indication of heteroskedasticity. Heteroskedasticity does not cause bias but it does cause inefficient estimates. In Sect. 11.4, this problem is fixed by a variance-stabilizing transformation. Figure 11.3b is a normal probability plot of the residuals. Outliers with both negative and positive values can be seen. These are due to the nonconstant variance and are not necessarily a sign of nonnormality. This plot illustrates the danger of attempting to interpret a normal plot when the data have a nonconstant variance. One should apply a variance-stabilizing transformation first before checking for normality. □

11.3 Estimating Forward Rates from Zero-Coupon Bond Prices

In practice, the forward-rate function $r(t)$ is unknown. Only bond prices are known. If the prices $P(T_i)$ of zero-coupon bonds are available on a relatively fine grid of values of $T_1 < T_2 < \cdots < T_n$, then using (3.24) we can estimate the forward-rate curve at T_i with

$$-\frac{\Delta \log\{P(T_i)\}}{\Delta T_i} = -\frac{\log\{P(T_i)\} - \log\{P(T_{i-1})\}}{T_i - T_{i-1}}. \qquad (11.12)$$

We will call these the *empirical forward-rate estimates*. Figure 11.4 shows prices and empirical forward-rate estimates from data to be described soon in Example 11.3. As can be seen in the plot, the empirical forward-rate estimates can be rather noisy when the denominators in (11.12) are small because the maturities are spaced closely together. If the maturities were more widely spaced, then bias rather than variance would be the major problem. Despite these difficulties, the empirical forward-rate estimates give a general impression of the forward-rate curve and are useful for comparing with estimates from parametric models, which are discussed next.

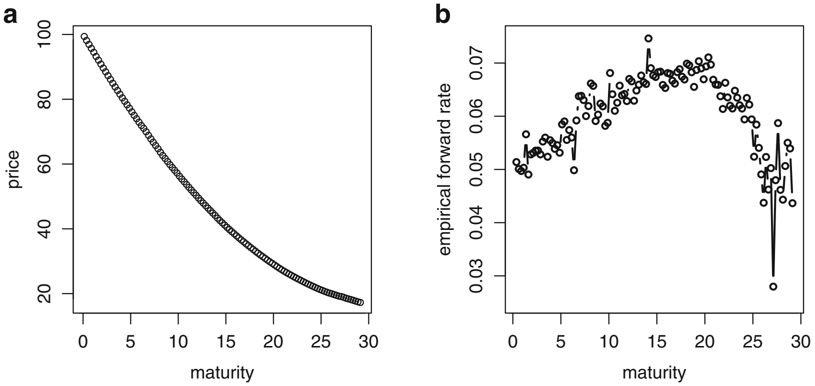

Fig. 11.4. *(a) U.S. STRIPS prices. (b) Empirical forward-rate estimates from the prices.*

We can estimate $r(t)$ from the bond prices using nonlinear regression. An example of estimating $r(t)$ was given in Sect. 11.2 assuming that $r(t)$ was constant and using as data the prices of zero-coupon bonds of different maturities. In this section, we estimate $r(t)$ without assuming it is constant.

Parametric estimation of the forward-rate curves starts with a parametric family $r(t;\boldsymbol{\theta})$ of forward rates and the correspond yield curves

$$y_T(\boldsymbol{\theta}) = T^{-1}\int_0^T r(t;\boldsymbol{\theta})\,dt$$

and model for the price of a par-\$1 bond:

$$P_T(\boldsymbol{\theta}) = \exp\{-Ty_T(\boldsymbol{\theta})\} = \exp\left(-\int_0^T r(t;\boldsymbol{\theta})\,dt\right).$$

For example, suppose that $r(t;\boldsymbol{\theta})$ is a pth-degree polynomial, so that

$$r(t;\boldsymbol{\theta}) = \theta_0 + \theta_1 t + \cdots + \theta_p t^p$$

for some unknown parameters $\theta_0, \ldots, \theta_p$. Then

$$\int_0^T r(t;\boldsymbol{\theta})\,dt = \theta_0 T + \theta_1 \frac{T^2}{2} + \cdots + \theta_p \frac{T^{p+1}}{p},$$

and the yield curve is

$$y_T = T^{-1} \int_0^T r(t;\boldsymbol{\theta})dt = \theta_0 + \theta_1 \frac{T}{2} + \cdots + \theta_p \frac{T^p}{p}.$$

A popular model is the Nelson–Siegel family with forward-rate and yield curves

$$\begin{aligned} r(t;\boldsymbol{\theta}) &= \theta_0 + (\theta_1 + \theta_2 t)\exp(-\theta_3 t), \\ y_t(\boldsymbol{\theta}) &= \theta_0 + \left(\theta_1 + \frac{\theta_2}{\theta_3}\right)\frac{1-\exp(-\theta_3 t)}{\theta_3 t} - \frac{\theta_2}{\theta_3}\exp(-\theta_3 t). \end{aligned}$$

The six-parameter Svensson model extends the Nelson–Siegel model by adding the term $\theta_4 t \exp(-\theta_5 t)$ to the forward rate.

The nonlinear regression model for estimating the forward-rate curve states that the price of the ith bond in the sample with maturity T_i expressed as a fraction of par value is

$$P_i = D(T_i) + \epsilon_i = \exp\left(-\int_0^{T_i} r(t;\boldsymbol{\theta})\,dt\right) + \epsilon_i, \qquad (11.13)$$

where D is the discount function and ϵ_i is an "error" due to problems such as prices being somewhat stale and the bid–ask spread.[2]

Example 11.3. Estimating forward rates from STRIPS prices

We now look at an example using data on U.S. STRIPS, a type of zero-coupon bond. STRIPS is an acronym for "Separate Trading of Registered Interest and Principal of Securities." The interest and principal on Treasury bills, notes, and bonds are traded separately through the Federal Reserve's book-entry system, in effect creating zero-coupon bonds by repackaging coupon bonds.[3]

The data are from December 31, 1995. The prices are given as a percentage of par value. Price is plotted against maturity in years in Fig. 11.4a. There are 117 prices and the maturities are nearly equally spaced from 0 to 30 years. We can see that the price drops smoothly with maturity and that there is not much noise in the price data. The empirical forward-rate estimates in Fig. 11.4b are much noisier than the prices.

[2] A bond dealer buys bonds at the bid price and sells them at the ask price, which is slightly higher than the bid price. The difference is called the bid–ask spread and covers the trader's administrative costs and profit.

[3] Jarrow(2002, p. 15).

Three models for the forward curve were fit: quadratic polynomial, cubic polynomial, and quadratic polynomial spline with a knot at $T = 15$. The latter splices two quadratic functions together at $T = 15$ so that the resulting curve is continuous and with a continuous first derivative. The spline's second derivative jumps at $T = 15$. One way to write the spline is

$$r(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 (t - 15)_+^2, \tag{11.14}$$

where the positive-part function is $x_+ = x$ if $x \geq 0$ and $x_+ = 0$ if $x < 0$. Also, x_+^2 means $(x_+)^2$, that is, take the positive part first. See Chap. 21 for further information about splines. From (11.14), one obtains

$$\int_0^T r(t)\,dt = \beta_0 T + \beta_1 \frac{T^2}{2} + \beta_2 \frac{T^3}{3} + \beta_3 \frac{(T-15)_+^3}{3}, \tag{11.15}$$

and therefore the yield curve is

$$y_T = \beta_0 + \beta_1 \frac{T}{2} + \beta_2 \frac{T^2}{3} + \beta_3 \frac{(T-15)_+^3}{3T}. \tag{11.16}$$

From (11.15), the model for a bond price (as a percentage of par) is

$$100 \exp\left\{ - \left(\beta_0 T + \beta_1 \frac{T^2}{2} + \beta_2 \frac{T^3}{3} + \beta_3 \frac{(T-15)_+^3}{3} \right) \right\}. \tag{11.17}$$

R code to fit the quadratic spline and plot its forward-rate estimate is

```
fitSpline = nls(price ~ 100 * exp(-beta0 * T
     - (beta1 * T^2)/2 - (beta2 * T^3) / 3
     - (T > 15) * (beta3 * (T - 15)^3) / 3), data = dat,
     start = list(beta0 = 0.03, beta1 = 0, beta2 = 0, beta3 = 0))
coefSpline = summary(fitSpline)$coef[ , 1]
forwardSpline = coefSpline[1] + (coefSpline[2] * t) +
    (coefSpline[3] * t^2)  + (t > 15) * (coefSpline[4] * (t - 15)^2)
plot(t, forwardSpline, lty = 2, lwd = 2)
```

Only slight changes in the code are needed to fit the quadratic or cubic polynomial models.

Figure 11.5 contains all three estimates of the forward rate and the empirical forward rates. The cubic polynomial and quadratic spline models follow the empirical forward rates much more closely than the quadratic polynomial model. The cubic polynomial and quadratic spline fits both use four parameters and are similar to each other, though the spline has a slightly smaller residual sum of squares. The summary of the spline model's fit is

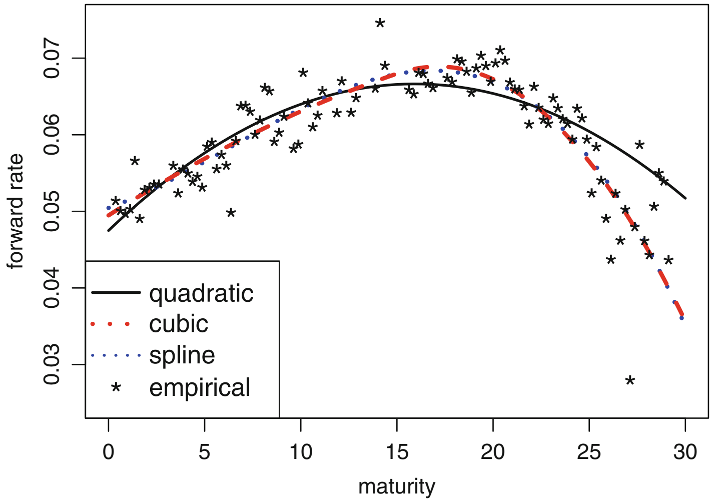

Fig. 11.5. *Polynomial and spline estimates of forward rates of U.S. Treasury bonds. The empirical forward rates are also shown.*

```
> summary(fitSpline)

Formula: price ~ 100 * exp(-beta0 * T - (beta1 * T^2)/2
  - (beta2 * T^3)/3 - (T > 15) * (beta3 * (T - 15)^3)/3)

Parameters:
        Estimate Std. Error t value Pr(>|t|)
beta0  4.947e-02  9.221e-05  536.52   <2e-16 ***
beta1  1.605e-03  3.116e-05   51.51   <2e-16 ***
beta2 -2.478e-05  1.820e-06  -13.62   <2e-16 ***
beta3 -1.763e-04  5.755e-06  -30.64   <2e-16 ***
---

Residual standard error: 0.0667 on 113 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 1.181e-07
```

Notice that all coefficients have very small p-values. The small p-value of `beta3` is further evidence that the spline model fits better than the quadratic polynomial model, since the two models differ only in that `beta3` is 0 for the quadratic model.

R's `nls` function could not find the least-squares estimator for the Nelson–Siegel model, but the least-squares estimator was found using the `optim` nonlinear optimization function with the sum of squares as the objective function. The fit of the Nelson–Siegel model was noticeably inferior to that of the cubic polynomial and quadratic spline models. In fact, the Nelson–Siegel model did not fit even as well as the quadratic polynomial model.

The Svensson model is likely to fit better than the Nelson–Siegel model, but the four-parameter cubic polynomial and quadratic spline models fit sufficiently well that it did not seem worthwhile to try the six-parameter Svensson model. □

11.4 Transform-Both-Sides Regression

Suppose we have a theoretical model that states that in the absence of any noise,

$$Y_i = f(\boldsymbol{X}_i; \boldsymbol{\beta}). \tag{11.18}$$

Model (11.18) is identical to the model

$$h\{Y_i\} = h\{f(\boldsymbol{X}_i; \boldsymbol{\beta})\}, \tag{11.19}$$

where h is *any* one-to-one function, such as, a strictly increasing function. In the absence of noise, one choice of h is as good as any other and one might as well stick with model (11.18), but when noise exists, this is no longer true.

When we have noisy data, Eq. (11.19) can be converted to the nonlinear regression model

$$h\{Y_i\} = h\{f(\boldsymbol{X}_i; \boldsymbol{\beta})\} + \epsilon_i. \tag{11.20}$$

Model (11.20) is called *the transform-both-sides (TBS) regression model* because both sides of Eq. (11.19) have been transformed by the same function h. Typically, h will be one of the Box–Cox transformations and h is chosen to stabilize the variation and to induce nearly normally distributed errors. To estimate $\boldsymbol{\beta}$ for a fixed h, one minimizes

$$\sum_{i=1}^{n} \left[h\{Y_i\} - h\left\{ f(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}}) \right\} \right]^2. \tag{11.21}$$

Various choices of h can be compared by residual plots. The h that gives approximately normally distributed residuals with a constant variance is used for the final analysis.

Example 11.4. TBS regression for the default frequency data

TBS regression was applied to the default frequency data. The Box–Cox transformation $h(y) = y^{(\alpha)}$ was tried with various positive values of α. It was found that $\alpha = 1/2$ gave residuals that appeared normally distributed with a constant variance, so the square-root transformation was used for estimation; see Fig. 11.6. With this transformation, $\boldsymbol{\beta}$ is estimated by minimizing

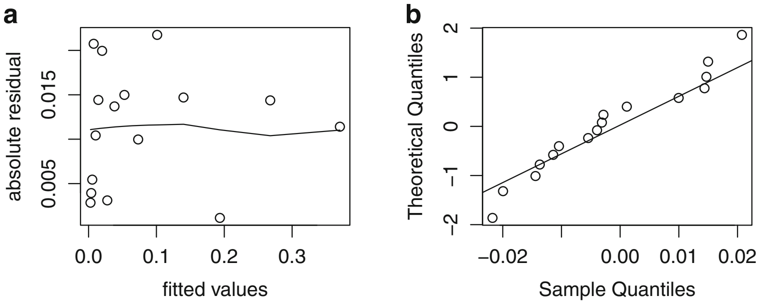

Fig. 11.6. (a) Transform-both-sides regression (TBS) with $h(y) = \sqrt{y}$. Absolute studentized residuals plotted against fitted values with a loess smooth. (b) Normal plot of the studentized residuals.

$$\sum_{i=1}^{n} \left[\sqrt{Y_i} - \exp\{\beta_0/2 + (\beta_1/2)X_i\} \right]^2, \tag{11.22}$$

where Y_i is the ith default frequency and X_i is the ith rating. The square-root transformation of the model is accomplished by dividing β_0 and β_1 by 2.

The R code to fit the TBS model and create Fig. 11.6 is below. The fitted values fit_tbs are computed by subtracting the residuals from the responses; this is done because the function summary() does not return the fitted values.

```
DefaultData = read.table("DefaultData.txt", header = TRUE)
attach(DefaultData)
freq2 = freq / 100
fit_tbs = nls(sqrt(freq2) ~ exp(b1 / 2 + b2 * rating / 2),
   start = list(b1 = -6, b2 = 0.5))
sum_tbs = summary(fit_tbs)
par(mfrow = c(1, 2))
fitted_tbs = sqrt(freq2) - sum_tbs$resid
plot(fitted_tbs,abs(sum_tbs$resid), xlab = "fitted values",
   ylab = "absolute residual")
fit_loess_tbs  = loess( abs(sum_tbs$resid) ~ fitted_tbs,
   span = 1, deg = 1)
ord_tbs = order(fitted_tbs)
lines(fitted_tbs[ord_tbs], fit_loess_tbs$fit[ord_tbs])
qqnorm(sum_tbs$resid, datax = TRUE, main = "")
qqline(sum_tbs$resid, datax = TRUE)
detact(DefaultData)
```

Using TBS regression, the estimated default probability of Aaa bonds is 0.0008 %, only 16 % of the estimate given by Bluhm, Overbeck, and Wagner (2003) and only 40 % of the estimate given by nonlinear regression without a transformation. Of course, a reduction in estimated risk by 84 % is a huge change. This shows how proper statistical modeling—e.g., using all the data and an appropriate transformation—can have a major impact on financial risk analysis. TBS allows one to use all the data (for unbiasedness) and, as described next, to effectively weight the data by the reciprocals of their variances for high efficiency.

□

11.4.1 How TBS Works

TBS in effect weights the data. To appreciate this, we use a Taylor series linearization[4] to obtain

$$\sum_{i=1}^{n}\left[h(Y_i)-h\left\{f(\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}})\right\}\right]^2=\sum_{i=1}^{n}\left[h^{(1)}\left\{f(\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}})\right\}\right]^2\left\{Y_i-f(\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}})\right\}^2.$$

The weight of the ith observation is $\left[h^{(1)}\{f(\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}})\}\right]^2$. Since the best weights are inverse variances, the most appropriate transformation h solves

$$\mathrm{Var}(Y_i|\boldsymbol{X}_i)\propto\left[h^{(1)}\{f(\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}})\}\right]^{-2}. \qquad (11.23)$$

For example, if $h(y)=\log(y)$, then $h^{(1)}(y)=1/y$ and (11.23) becomes

$$\mathrm{Var}(Y_i|\boldsymbol{X}_i)\propto\{f(\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}})\}^2, \qquad (11.24)$$

so that the conditional standard deviation of the response is proportional to its conditional mean. This occurs frequently. For example, if the response is exponentially distributed then (11.24) must hold. Equation (11.24) holds also if the response is lognormally distributed and the log-variance is constant. In this case, it is not surprising that the log transformation is best since the log transforms to i.i.d. normal noise.

The *coefficient of variation* of a random variable is the ratio of its standard deviation to its expected value. When (11.24) holds, the response has a constant coefficient of variation.

A transformation that causes that conditional variance to be constant is called the *variance-stabilizing transformation*. We have just shown that when the coefficient of variation is constant, then the variance-stabilizing transformation is the logarithm.

Example 11.5. Poisson responses

Assume $Y_i|\boldsymbol{X}_i$ is Poisson distributed with mean $f(\boldsymbol{X}_i;\boldsymbol{\beta})$, as might, for example, happen if Y_i were of the number of companies declaring bankruptcy

[4] A Taylor series linearization of the function h about the point x is $h(y)\approx h(x)+h^{(1)}(x)(y-x)$, where $h^{(1)}$ is the first derivative of h. See any calculus textbook for further discussion of Taylor series.

in a year, with $f(\boldsymbol{X}_i;\boldsymbol{\beta})$ modeling how that expected number depends on macroeconomic variables in $\boldsymbol{X}_i$. The variance equals the mean for the Poisson distribution, so

$$\mathrm{Var}(Y_i|\boldsymbol{X}_i) = f(\boldsymbol{X}_i;\boldsymbol{\beta}).$$

Using the same type of reasoning as in the previous example, it follows that one should use $\alpha = 1/2$; the square-root transformation is the variance-stabilizing transformation for Poisson-distributed responses. □

11.5 Transforming Only the Response

The so-called Box–Cox transformation model is

$$Y_i^{(\alpha)} = \beta_0 + X_{i,1}\beta_1 + \cdots + X_{i,p}\beta_p + \epsilon_i, \tag{11.25}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, \sigma_\epsilon^2)$ for some σ_ϵ. In contrast to the TBS model, only the response is transformed. The goal of transforming the response is to achieve three objectives:

1. a simple model: $Y_i^{(\alpha)}$ is linear in predictors $X_{i,1}, \ldots, X_{i,p}$ and in the parameters $\beta_1, \ldots, \beta_p$;
2. constant residual variance; and
3. Gaussian noise.

In contrast, 2 and 3 but *not* 1 are the goals of the TBS model.

Model (11.25) was introduced by Box and Cox (1964) who suggested estimation of α by maximum likelihood. The function `boxcox()` in `R`'s `MASS` package will compute the profile log-likelihood for α along with a confidence interval. Usually, $\widehat{\alpha}$ is taken to be some round number, e.g., -1, $-1/2$, 0, 1/2, or 1, in the confidence interval. The reason for selecting one of these numbers is that then the transformation is readily interpretable, that is, it is the square root, log, inverse, or some other familiar function. Of course, one can use the value of α that maximizes the profile log-likelihood if one is not concerned with having a familiar transformation. After $\widehat{\alpha}$ has been selected in this way, $\beta_0, \ldots, \beta_p$ and σ_ϵ^2 can be estimated by regressing $Y_i^{(\widehat{\alpha})}$ on $X_{i,1}, \ldots, X_{i,p}$.

Example 11.6. Simulated data—Box Cox transformation

This example uses the simulated data introduced in Example 10.6. The model is

$$Y_i^{(\alpha)} = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,1}^2 + \beta_3 X_{i,2} + \epsilon_i. \tag{11.26}$$

The profile likelihood for α was produced by the `boxcox()` function in `R` and is plotted in Fig. 11.7. The code to produce the figure is:

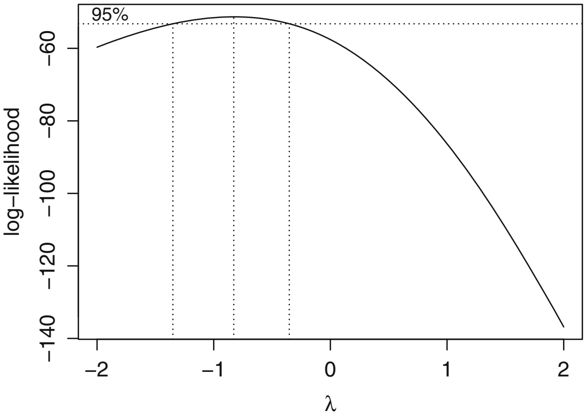

Fig. 11.7. *Profile likelihood for the Box–Cox model applied to the simulated data.*

```
boxcox(y ~ poly(x1,2) + x2, ylab = "log-likelihood")
```

We see that the MLE is near −1 and −1 is well within the confidence interval; these results suggest that we use $-1/Y_i$ as the response.

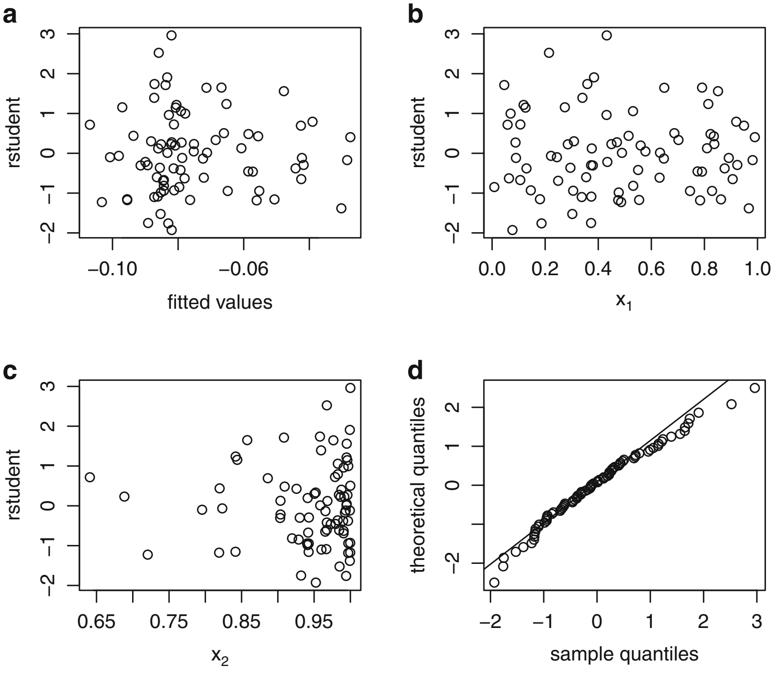

Fig. 11.8. *Residuals for the Box–Cox model applied to the simulated data.*

Residual plots with response $-1/Y_i$ are shown in Fig. 11.8. We see in panel (a) that there is no sign of heteroskedasticity, since the vertical scatter of the residuals does not change from left to right. In panels (b) and (c) we see uniform vertical scatter which shows that the model that is quadratic in X_1 and linear in X_2 fits $-1/Y_i$ well. Finally, in panel (d), we see that the residuals appear normally distributed. □

11.6 Binary Regression

A binary response Y can take only two values, 0 or 1, which code two possible outcomes, for example, that a company goes into default on its loans or that it does not default. Binary regression models the conditional probability that a binary response is 1, given the values of the predictors $X_{i,1}, \ldots, X_{i,p}$. Since a probability is constrained to lie between 0 and 1, a linear model is not appropriate for a binary response. However, linear models are so convenient that one would like a model that has many of the features of a linear model. This has motivated the development of *generalized linear models*, often called GLMs.

Generalized linear models for binary responses are of the form

$$P(Y_i = 1 | X_{i,1}, \ldots, X_{i,p}) = H(\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}) = H(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}),$$

where $H(x)$ is a function that increases from 0 to 1 as x increases from $-\infty$ to ∞, so that $H(x)$ is a CDF, and the last expression uses the vector notation of (11.1). The most common GLMs for a binary responses are probit regression, where $H(x) = \Phi(x)$, the $N(0,1)$ CDF; and logistic regression, where $H(x)$ is the logistic CDF, which is $H(x) = 1/\{1 + \exp(-x)\}$. The parameter vector $\boldsymbol{\beta}$ can be estimated by maximum likelihood. Assume that conditional on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ the binary responses $Y_1, \ldots, Y_n$ are mutually independent. Then, using (A.8), the likelihood (conditional on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$) is

$$\prod_{i=1}^{n} H\left(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}\right)^{Y_i} \left\{1 - H(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta})\right\}^{1-Y_i}. \tag{11.27}$$

The MLEs can be found by standard software, e.g., the function `glm()` in `R`.

Example 11.7. Who gets a credit card?

In this example, we will analyze the data in the `CreditCard` data set in `R`'s `AER` package. The following variables are included in the data set:

1. `card` = Was the application for a credit card accepted?
2. `reports` = Number of major derogatory reports
3. `income` = Yearly income (in USD 10,000)
4. `age` = Age in years plus 12ths of a year

5. `owner` = Does the individual own his or her home?
6. `dependents` = Number of dependents
7. `months` = Months living at current address
8. `share` = Ratio of monthly credit card expenditure to yearly income
9. `selfemp` = Is the individual self-employed?
10. `majorcards` = Number of major credit cards held
11. `active` = Number of active credit accounts
12. `expenditure` = Average monthly credit card expenditure

The first variable, `card`, is binary and will be the response. Variables 2–8 will be used as predictors. The goal of the analysis is to discover which of the predictors influences the probability that an application is accepted. `R`'s documentation mentions that there are some values of the variable `age` under one year. These cases must be in error and they were deleted from the analysis. Figure 11.9 contains histograms of the predictors. The variable `share` is highly right-skewed, so `log(share)` will be used in the analysis. The variable `reports`

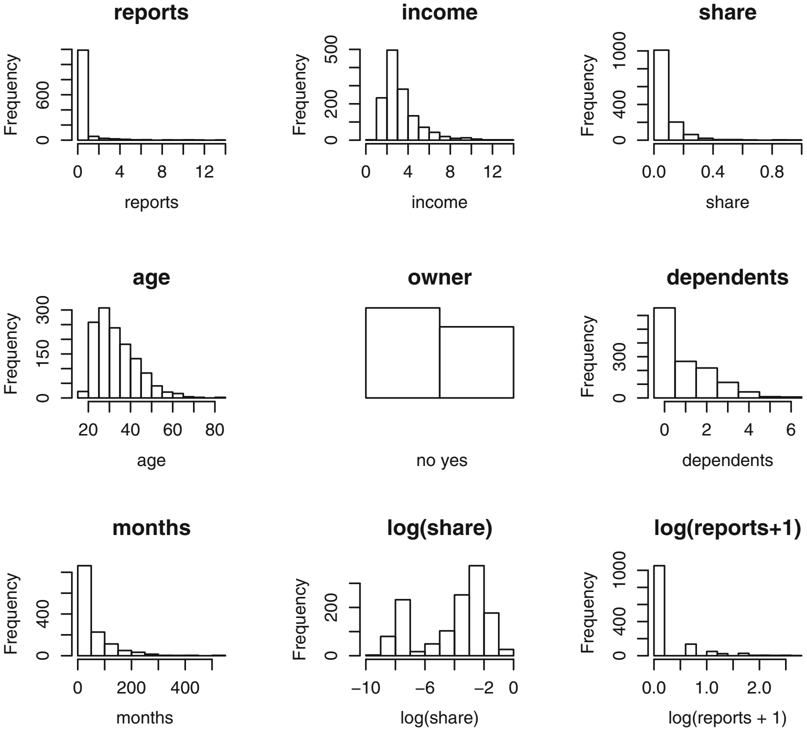

Fig. 11.9. *Histograms of variables for potential use in a model to predict whether a credit card application will be accepted.*

is also extremely right-skewed; most values of reports are 0 or 1 but the maximum value is 14. To reduce the skewness, log(reports+1) will be used instead of reports. The "1" is added to avoid taking the logarithm of 0. There are no assumptions in regression about the distributions of the predictors, so skewed predictor variables can, in principle, be used. However, highly skewed predictors have high-leverage points and are less likely to be linearly related to the response. It is a good idea at least to consider transformation of highly skewed predictors. In fact, the logistic model was also fit with reports and share untransformed, but this increased AIC by more than 3 compared to using the transformed predictors.

First, a logistic regression model is fit with all seven predictors using the glm() function. The R code is:

```
library("AER")
library("faraway")
data("CreditCard")
CreditCard_clean = CreditCard[CreditCard$age > 18, ]
names(CreditCard)
fit1 = glm(card ~ log(reports + 1) + income + log(share) + age
   + owner + dependents + months,
   family = "binomial", data = CreditCard_clean)
summary(fit1)
stepAIC(fit1)

Call:
glm(formula = card ~ log(reports + 1) + income + log(share) +
    age + owner + dependents + months, family = "binomial",
    data = CreditCard_clean)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      21.473930   3.674325   5.844 5.09e-09 ***
log(reports + 1) -2.908644   1.097604  -2.650  0.00805 **
income            0.903315   0.189754   4.760 1.93e-06 ***
log(share)        3.422980   0.530499   6.452 1.10e-10 ***
age               0.022682   0.021895   1.036  0.30024
owneryes          0.705171   0.533070   1.323  0.18589
dependents       -0.664933   0.267404  -2.487  0.01290 *
months           -0.005723   0.003988  -1.435  0.15130
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1398.53  on 1311  degrees of freedom
Residual deviance:  139.79  on 1304  degrees of freedom
AIC: 155.79

Number of Fisher Scoring iterations: 11
```

Several of the regressors have large p-values, so `stepAIC()` was used to find a more parsimonious model. The final step where no more variables were deleted is

```
Step:  AIC=154.22
card ~ log(reports + 1) + income + log(share) + dependents

                   Df Deviance     AIC
<none>                  144.22  154.22
- dependents        1   150.28  158.28
- log(reports + 1)  1   164.18  172.18
- income            1   173.62  181.62
- log(share)        1  1079.61 1087.61
```

Below is the fit using the model selected by `stepAIC()`. For convenience later, each of the regressors was mean-centered; "`_c`" appended to a variable name indicates centering.

```
glm(formula = card ~ log_reports_c + income_c + log_share_c +
    dependents_c, family = "binomial", data = CreditCard_clean)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     9.5238     1.7213   5.533 3.15e-08 ***
log_reports_c  -2.8953     1.0866  -2.664  0.00771 **
income_c        0.8717     0.1724   5.056 4.28e-07 ***
log_share_c     3.3102     0.4942   6.698 2.11e-11 ***
dependents_c   -0.5506     0.2505  -2.198  0.02793 *
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1398.53  on 1311  degrees of freedom
Residual deviance:  144.22  on 1307  degrees of freedom
AIC: 154.22

Number of Fisher Scoring iterations: 11
```

It is important to understand what the logistic regression model is telling us about the probability of an application being accepted. Qualitatively, we see that the probability of having an application accepted increases with `income` and `share` and decreases with `reports` and `dependents`. To understand these effects quantitatively, first consider the intercept. Since the predictors have been mean-centered, the probability of an application being accepted when all variables are at their mean is simply $H(9.5238) = 0.999927$. Since `reports` and `dependents` are integer-valued and cannot exactly equal their means, this probability only provides an idea of what the intercept 9.5238 signifies. Figure 11.10 plots the probability that a credit card application is accepted as functions of `reports`, `income`, `log(share)`, and `dependents`. In each plot, the other variables are fixed at their means. Clearly, the variable with the largest effect is `share`, the ratio of monthly credit card expenditure to yearly income. We see that applicants who spend little of their income through credit cards are unlikely to have their applications accepted.

In Fig. 11.11, panel (a) is a plot of `card`, which takes value 0 if an application is rejected and 1 if it is accepted, versus `log(share)`. It should be emphasized that panel (a) is a plot of the data, not a fit from the model. We see that an application is always accepted if `log(share)` exceeds −6, which translates into `share` exceeding 0.0025. Thus, in this data set, among the group of applicants whose average monthly credit card expenses exceeded 0.25 % of yearly income, all credit card applications were accepted. How do these applicants look on the other variables? Panels (b)–(d) plot `reports`, `income`, and `majorcards` versus `log(share)`. The variable `majorcards` was not used in the logistic regression analysis but is included here.

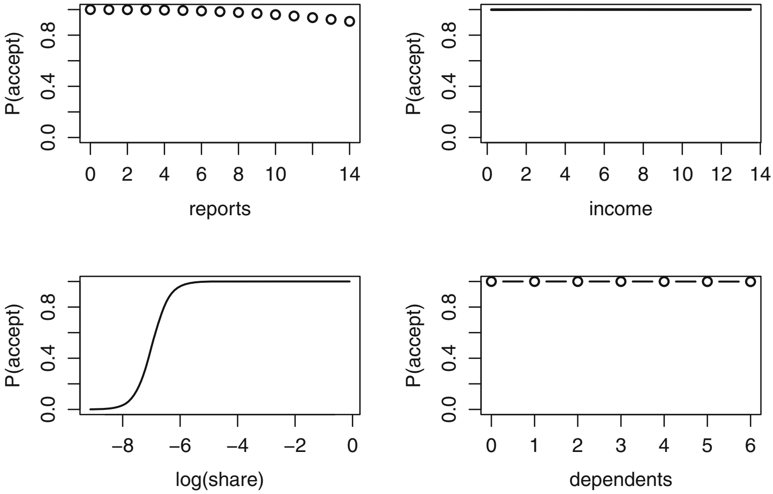

Fig. 11.10. *Plots of probabilities of a credit card application being accepted as functions of single predictors with other predictors fixed at their means. The variables vary over their ranges in the data.*

An odd feature in Fig. 11.11c is a group of points following a smooth curve. This is a group of 316 applications who had the product of `share` times `income` exactly equal to 0.0012, the minimum value of this product. Oddly, `share` is never 0. Perhaps because of some coding artifact, these 316 had 0 credit card expenditures rather than the reported values. Another interesting feature of the data is that among these 316 applications, only 21 were accepted. Among the remaining 996 applications, all were accepted.

Besides illustrating logistic regression, this example demonstrates that real-world data often contain errors, or perhaps we should call them idiosyncracies, and that a thorough graphical analysis of the data is always a good thing. □

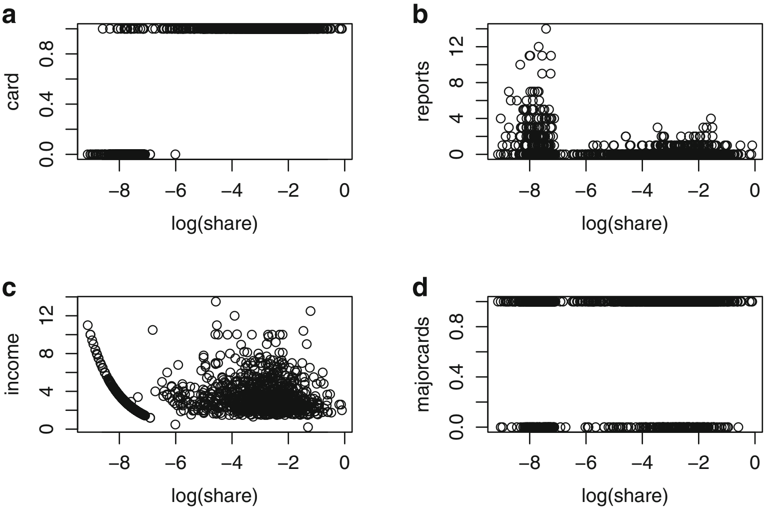

Fig. 11.11. *Plots of* `log(share)` *versus other variables.*

11.7 Linearizing a Nonlinear Model

Sometimes a nonlinear model can be linearized by applying a transformation to both the model and the response. In such cases, should one use a linearizing transformation or, instead, apply nonlinear regression to the original model? The answer is that linearization can sometimes be a good thing, but not always. Fortunately, residual analysis can help us decide whether a linearizing transformation should be used.

For example, consider the model

$$Y_i = \beta_1 \exp(\beta_2 X_i). \tag{11.28}$$

This model is "equivalent" to the linear model

$$\log(Y_i) = \alpha + \beta_2 X_i, \tag{11.29}$$

where $\alpha = \log(\beta_1)$. "Equivalent" is in quotes, because the two models are no longer equivalent when noise is present.

Suppose (11.28) has i.i.d. additive noise, so that

$$Y_i = \beta_1 \exp(\beta_2 X_i) + \epsilon_i, \tag{11.30}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Then applying the log transformation to (11.29) gives us the model

$$\log(Y_i) = \log\{\beta_1 \exp(\beta_2 X_i) + \epsilon_i\} \tag{11.31}$$

with nonadditive noise. Because the noise is not additive, the variation of $\log(Y_i)$ about the model $\log\{\beta_1 \exp(\beta_2 X_i)\}$ will have nonconstant variation and skewness, even if $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Gaussian.

Example 11.8. Linearizing transformation—Simulated data

Figure 11.12a shows a simulated sample from model (11.28) with $\beta_1 = 1$, $\beta_2 = -1$, and $\sigma_\epsilon = 0.02$. The X_i are equally spaced from -1 to 2.5 by increments of 0.025. Panel (b) shows $\log(Y_i)$ plotted against X_i. One can see that the transformation has linearized the relationship between the variables but has introduced nonconstant residual variation. Panels (c) and (d) show residual plots using the linearized model. Notice the nonlinear normal plot and the severe nonconstant variance. □

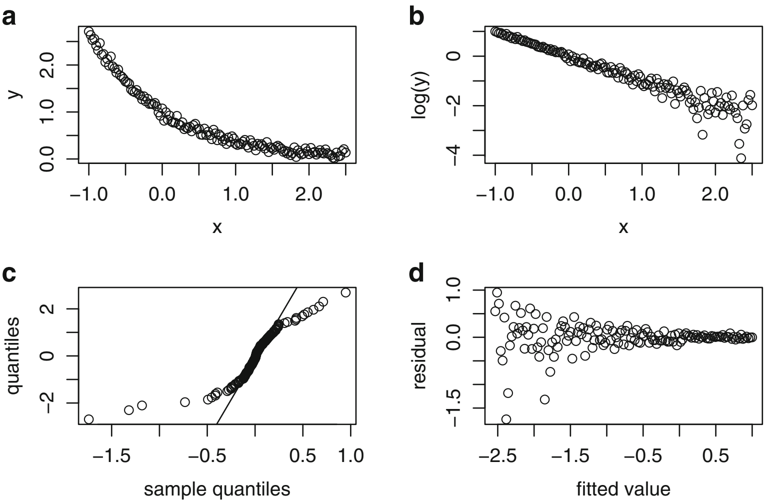

Fig. 11.12. *Example where the log transformation linearizes a model but induces substantial heteroskedasticity and skewness.* (a) *Raw data.* (b) *Data after log transformation of the response.* (c) *Normal plot of residuals after linearization.* (d) *Absolute residual plot after linearization.*

Linearizing is not always a bad thing. Suppose the noise is multiplicative and lognormal so that (11.28) becomes

$$Y_i = \beta_1 \exp(\beta_2 X_i) \exp(\epsilon_i) = \beta_1 \exp(\beta_2 X_i + \epsilon_i), \tag{11.32}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Gaussian. Then the log transformation converts (11.32) to

$$\log(Y_i) = \alpha + \beta_2 X_i + \epsilon_i, \tag{11.33}$$

which is a linear model satisfying all of the usual assumptions.

In summary, a linearizing transformation may or may not cause the data to better follow the assumptions of regression analysis. Residual analysis can help one decide whether a transformation is appropriate.

11.8 Robust Regression

A robust regression estimator should be relatively immune to two types of outliers. The first are *bad data*, meaning *contaminants* that are not part of the population, for example, due to undetected recording errors. The second are outliers due to the noise distribution having heavy tails. There are a large number of robust regression estimators, and their sheer number has been an impediment to their use. Many data analysts are confused as to which robust estimator is best and consequently are reluctant to use any. Rather than describe many of these estimators, which might contribute to this problem, we mention just one, the *least-trimmed sum of squares estimator*, often called the *LTS*.

Recall the trimmed mean, a robust estimator of location for a univariate sample. The trimmed mean is simply the mean of the sample after a certain percentage of the largest observations and the same percentage of the smallest observations have been removed. This trimming removes some non-outliers, which, under the ideal conditions of no outliers, causes some loss of precision, but not an unacceptable amount. The trimming also removes outliers, and this causes the estimator to be robust. Trimming is easy for a univariate sample because we know which observations to trim, the very largest and the very smallest. This is not the case in regression. Consider the data in Fig. 11.13. There are 26 observations that fall closely along a line plus two *residual outliers* that are far from this line. Notice that the residual outliers have neither extreme X-values nor extreme Y-values. They are outlying only relative to the linear regression fit to the other data.

The residual outliers are obvious in Fig. 11.13 because there is only a single predictor. When there are many predictors, outliers can only be identified when we have a model *and* good estimates of the parameters in that model. The difficulty, then, is that estimation of the parameters requires the identification of the outliers, and vice versa. One can see from the figure that the least-squares line is changed by including the residual outliers in the data used for estimation. In some cases, e.g., Fig. 10.1b, the effect of a residual outlier can be so severe that it totally changes the least-squares estimates. This is likely to happen if the residual outlier occurs at a high-leverage point.

The LTS estimator simultaneously identifies residual outliers and estimates robustly the parameters of a model. Let $0 < \alpha \leq 1/2$ be the trimming proportion and let k equal $n\alpha$ rounded to an integer. The trimmed sum of squares about a set of values of the regression parameters is defined as follows: Form the residuals from the model evaluated at these parameters, square the residuals, then order the squared residuals and remove the k largest, and finally sum the remaining squared residuals. The LTS estimates are the set of parameter values that minimize the trimmed sum of squares. The LTS estimator can be computed using the `ltsReg()` function in `R`'s `robust` package.

If the noise distribution is heavy-tailed, then an alternative to a robust regression analysis is to use a heavy-tailed distribution as a model for the noise and then to estimate the parameters by maximum likelihood. For example, one could assume that the noise has a double-exponential or t-distribution. In the latter case, one could either estimate the degrees of freedom or simply fix the degrees of freedom at a low value, which implies heavier tails; see Lange, Little, and Taylor (1989). This strategy is called *robust modeling* rather than robust estimation. The distinction is that in robust estimation one assumes a fairly restrictive model such as a normal noise distribution, but finds a robust alternative to maximum likelihood. In robust modeling, one uses a more flexible model so that maximum likelihood estimation is itself robust. When there is a single gross residual outlier, particularly at a high-leverage point, robust regression is a better alternative than the MLE with a heavy-tailed noise distribution; see the next example.

Another possibility is that residual outliers are due to nonconstant standard deviations, with the outliers mainly in the data with a higher noise standard deviation. The remedy to this problem is to apply a variance stabilization transformation or to model the nonconstant standard deviation, say by one of the GARCH models discussed in Chap. 14.

Example 11.9. Simulated data in Example 10.1—Robust regression

Figure 11.14 compares least-squares fit, the LTS fit, and the MLE assuming t-distributed noise for the simulated data in Example 10.1. In panel (a) with no residuals outliers, the three fits coincide. In panels (b) and (c), the LTS fits are not affected by the residual outliers and fit the nonoutlying data very well. In these panels, the LS and MLE fits are highly affected by the outlier and nearly identical. For these examples, the MLE assuming t-distributed noise is not robust. □

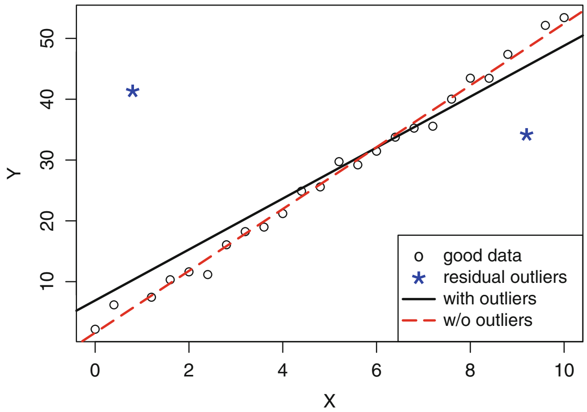

Fig. 11.13. *Straight-line regression with two residual outliers showing least-squares fits with and without the outliers.*

11.9 Regression and Best Linear Prediction

11.9.1 Best Linear Prediction

Often we observe a random variable X and we want to predict an unobserved random variable Y that is related to X. For example, Y could be the future price of an asset and X might be the most recent change in that asset's price. Prediction has many practical uses, and it is also important in theoretical studies.

The predictor of Y that minimizes the expected squared prediction error is $E(Y|X)$ (see Appendix A.19), but $E(Y|X)$ is often a nonlinear function of X and difficult to compute. A common solution to this difficulty it to consider only linear functions of X as possible predictors. This is called *linear prediction*. In this section, we will show that linear prediction is closely related to linear regression.

A linear predictor of Y based on X is a function $\beta_0 + \beta_1 X$ where β_0 and β_1 are parameters that we can choose. *Best linear prediction* means finding β_0 and β_1 so that expected squared prediction error, which is given by

$$E\{Y - (\beta_0 + \beta_1 X)\}^2, \tag{11.34}$$

is minimized. Doing this makes the predictor as close as possible, on average, to Y. The expected squared prediction error can be rewritten as

$$\begin{aligned} &E\{Y - (\beta_0 + \beta_1 X)\}^2 \\ &\quad = E(Y^2) - 2\beta_0 E(Y) - 2\beta_1 E(XY) + \beta_0^2 + 2\beta_0\beta_1 E(X) + \beta_1^2 E(X^2). \end{aligned}$$

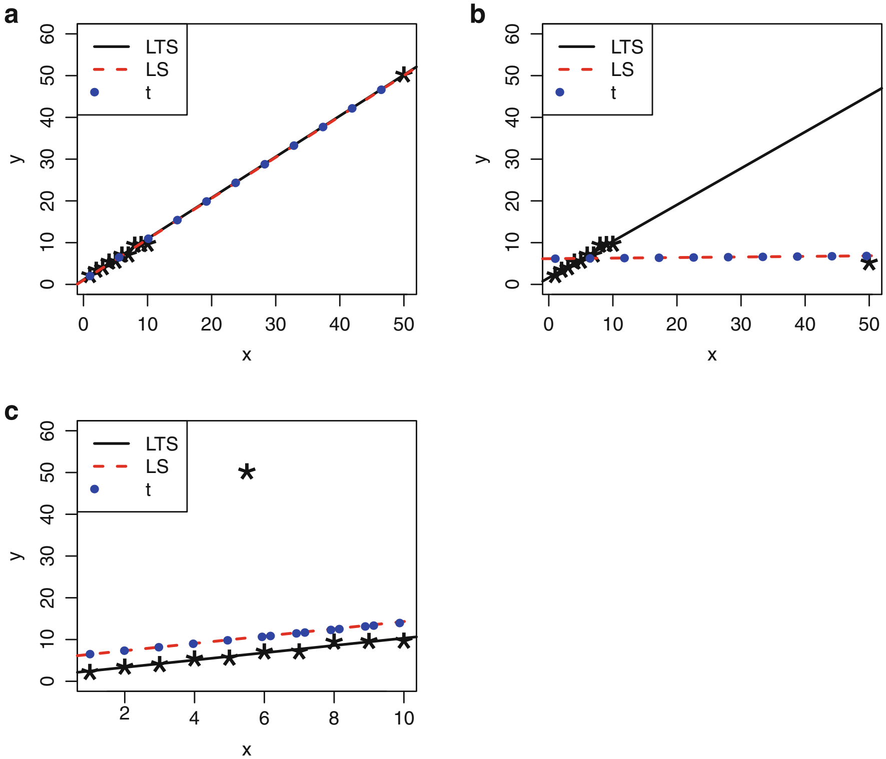

Fig. 11.14. *Simulated data in Example* 10.1 *with LS fits (dashed red) and LTS fits (solid black) and MLEs assuming t-distributed noise (dotted blue). In (a) the three fits are too close together to distinguish between them. In (b) and (c) the LS and t fits are nearly identical and difficult to distinguish.*

To find the minimizers, we set the partial derivatives of this expression to zero to obtain

$$0 = -E(Y) + \beta_0 + \beta_1 E(X), \tag{11.35}$$

$$0 = -E(XY) + \beta_0 E(X) + \beta_1 E(X^2). \tag{11.36}$$

After some algebra we find that

$$\beta_1 = \sigma_{XY}/\sigma_X^2 \tag{11.37}$$

and

$$\beta_0 = E(Y) - \beta_1 E(X) = E(Y) - \sigma_{XY}/\sigma_X^2\, E(X). \tag{11.38}$$

One can check that the matrix of second derivatives of (11.34) is positive definite so that the solution (β_0, β_1) to (11.35) and (11.36) minimizes (11.34). Thus, the best linear predictor of Y is

$$\widehat{Y}^{\text{Lin}}(X) = \beta_0 + \beta_1 X = E(Y) + \frac{\sigma_{XY}}{\sigma_X^2}\{X - E(X)\}. \tag{11.39}$$

In practice, (11.39) cannot be used directly unless $E(X)$, $E(Y)$, σ_{XY}, and σ_X^2 are known, which is often not the case. Linear regression analysis is essentially the use of (11.39) with these unknown parameters replaced by least-squares estimates—see Sect. 11.9.3.

11.9.2 Prediction Error in Best Linear Prediction

In this section, assume that $\widehat{Y}$ is the best linear predictor of Y. The *prediction error* is $Y - \widehat{Y}$. It is easy to show that $E\{Y - \widehat{Y}\} = 0$ so that the prediction is unbiased. With a little algebra we can show that the expected squared prediction error is

$$E\{Y - \widehat{Y}\}^2 = \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2} = \sigma_Y^2(1 - \rho_{XY}^2). \tag{11.40}$$

How much does X help us predict Y? To answer this question, notice first that if we do not observe X, then we must predict Y using a constant, which we denote by c. It is easy to show that the best predictor has c equal to $E(Y)$. Notice first that the expected squared prediction error is $E(Y - c)^2$. Some algebra shows that

$$E(Y - c)^2 = \text{Var}(Y) + \{c - E(Y)\}^2, \tag{11.41}$$

which, since $\text{Var}(Y)$ does not depend on c, shows that the expected squared prediction error is minimized by $c = E(Y)$. Thus, when X is unobserved, the best predictor of Y is $E(Y)$ and the expected squared prediction error is σ_Y^2, but when X is observed, then the expected squared prediction error is smaller, $\sigma_Y^2(1 - \rho_{XY}^2)$. Therefore, ρ_{XY}^2 is the fraction by which the prediction error is reduced when X is known. This is an important fact that we will see again.

Result 11.1 Prediction when Y is independent of all available information:

If Y is independent of all presently available information, that is, Y is independent of all random variables that have been observed, then the best predictor of Y is $E(Y)$ and the expected value of the squared prediction error is σ_Y^2. We say that Y "cannot be predicted" when there exists no predictor better than its expected value.

11.9.3 Regression Is Empirical Best Linear Prediction

For the case of a single predictor, note the similarity between the best linear predictor,

$$\widehat{Y} = E(Y) + \frac{\sigma_{XY}}{\sigma_X^2}\{X - E(X)\},$$

and the least-squares line,

$$\widehat{Y} = \overline{Y} + \frac{s_{XY}}{s_X^2}(X - \overline{X}).$$

The least-squares line is a sample version of the best linear predictor. Also, ρ_{XY}^2, the squared correlation between X and Y, is the fraction of variation in Y that can be predicted using the linear predictor, and the sample version of ρ_{XY}^2 is $R^2 = r_{XY}^2 = r_{\widehat{Y}Y}^2$.

11.9.4 Multivariate Linear Prediction

So far we have assumed that there is only a single random variable, X, available to predict Y. More commonly, Y is predicted using a set of observed random variables, $X_1, \ldots, X_n$.

Let $\boldsymbol{Y}$ and $\boldsymbol{X}$ by $p \times 1$ and $q \times 1$ random vectors. As before in Sect. 7.3.1, define

$$\boldsymbol{\Sigma}_{Y,X} = E\{\boldsymbol{Y} - E(\boldsymbol{Y})\}\{\boldsymbol{X} - E(\boldsymbol{X})\}^{\mathsf{T}},$$

so that the i,jth element of $\boldsymbol{\Sigma}_{Y,X}$ is the covariance between Y_i and X_j. Then the best linear predictor of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is

$$\widehat{\boldsymbol{Y}} = E(\boldsymbol{Y}) + \boldsymbol{\Sigma}_{Y,X}\boldsymbol{\Sigma}_X^{-1}\{\boldsymbol{X} - E(\boldsymbol{X})\}. \tag{11.42}$$

Note the similarity between (11.39) and (11.42), the best linear predictors in the univariate and multivariate cases.

The sample analog of multivariate linear prediction is multiple regression.

11.10 Regression Hedging

An interesting application of regression is determining the optimal hedge of a bond position. Market makers buy securities at a *bid price* and make a profit by selling them at a higher *ask price*. Suppose a market maker has just purchased a bond from a pension fund. Ideally, the market maker would sell the bond immediately after purchasing it. However, many bonds are illiquid, so it may take some time before the bond can be sold. During the period that a market maker is holding a bond, the market maker is at risk that the bond price could drop due to a change in interest rates. The change could wipe out the profit due to the small bid–ask spread. The market maker would prefer to hedge this risk by assuming another risk which is likely to be in the opposite direction. To hedge the interest-rate risk of the bond being held, the market maker can sell other, more liquid, bonds short. Suppose that the market maker decides to sell short a 30-year Treasury bond, which is more liquid.

Regression hedging determines the optimal amount of the 30-year Treasury bonds to sell short to hedge the risk of the bond just purchased. The goal is that the price of the portfolio long in the first bond and short in the Treasury bond changes as little as possible as yields change. Suppose the first bond has a maturity of 25 years. One can determine the sensitivity of price to yield changes using results from Sect. 3.8. Let y_{30} be the yield on 30-year bonds, let P_{30} be the price of \$1 in face amount of 30-year bonds, and let DUR_{30} be the duration. The change in price, $\varDelta P_{30}$, and the change in yield, $\varDelta y_{30}$, are related by

$$\varDelta P_{30} \approx -P_{30}\,\mathrm{DUR}_{30}\,\varDelta y_{30}$$

for small values of $\varDelta y_{30}$. A similar result holds for 25-year bonds.

Consider a portfolio that holds face amount F_{25} in 25-year bonds and is short face amount F_{30} in 30-year bonds. The value of the portfolio is

$$F_{25}P_{25} - F_{30}P_{30}.$$

If $\varDelta y_{25}$ and $\varDelta y_{30}$ are the changes in the yields, then the change in value of the portfolio is approximately

$$\{F_{30}P_{30}\,\mathrm{DUR}_{30}\,\varDelta y_{30} - F_{25}P_{25}\,\mathrm{DUR}_{25}\,\varDelta y_{25}\}. \tag{11.43}$$

Suppose that the regression of $\varDelta y_{30}$ on $\varDelta y_{25}$ is

$$\varDelta y_{30} = \widehat{\beta}_0 + \widehat{\beta}_1 \varDelta y_{25} \tag{11.44}$$

and $\widehat{\beta}_0 \approx 0$, as is usually the case for regression of changes in interest rates, as in Example 9.1. Substituting (11.44) into (11.43), the change in price of the portfolio is approximately

$$\{F_{30}P_{30}\,\mathrm{DUR}_{30}\widehat{\beta}_1 - F_{25}P_{25}\,\mathrm{DUR}_{25}\}\varDelta y_{25}. \tag{11.45}$$

This change is approximately zero for all values of $\varDelta y_{25}$ if

$$F_{30} = F_{25}\frac{P_{25}\,\mathrm{DUR}_{25}}{P_{30}\,\mathrm{DUR}_{30}\widehat{\beta}_1}. \tag{11.46}$$

Equation (11.46) tells us how much face value of the 30-year bond to sell short in order to hedge F_{25} face value of the 25-year bond. All quantities on the right-hand side of (11.46) are known or readily calculated: F_{25} is the current position in the 25-year bond, P_{25} and P_{30} are known bond prices, calculation of DUR_{25} and DUR_{30} is discussed in Chap. 3, and $\widehat{\beta}_1$ is the slope of the regression of $\varDelta y_{30}$ on $\varDelta y_{25}$.

The higher the R^2 of the regression, the better the hedge works. Hedging with two or more liquid bonds, say a 30-year and a 10-year, can be done by multiple regression and might produce a better hedge.

11.11 Bibliographic Notes

Atkinson (1985) has nice coverage of transformations and residual plotting and many good examples. For more information on nonlinear regression, see Bates and Watts (1988) and Seber and Wild (1989). Graphical methods for detecting a nonconstant variance, transform-both-sides regression, and weighting are discussed in Carroll and Ruppert (1988). Hosmer and Lemeshow (2000) is an in-depth treatment of logistic regression. Faraway (2006) covers generalized linear models including logistic regression. See Tuckman (2002) for more discussion of regression hedging.

The Nelson–Siegel and Svensson models are from Nelson and Siegel (1985) and Svensson (1994).

11.12 R Lab

11.12.1 Nonlinear Regression

In this section, you will be fitting short-rate models. Let r_t be the short rate (the risk-free rate for short-term borrowing) at time t. It is assumed that the short rate satisfies the stochastic differential equation

$$dr_t = \mu(t, r_t)\, dt + \sigma(t, r_t)\, dW_t, \tag{11.47}$$

where $\mu(t, r_t)$ is a drift function, $\sigma(t, r_t)$ is a volatility function, and W_t is a standard Brownian motion. We will use a discrete approximation to (11.47):

$$(r_t - r_{t-1}) = \mu(t-1, r_{t-1}) + \sigma(t-1, r_{t-1})\, \epsilon_{t-1} \tag{11.48}$$

where $\epsilon_1, \ldots, \epsilon_{n-1}$ are i.i.d. $N(0,1)$.

We will start with the Chan, Karolyi, Longstaff, and Sanders (1992) (CKLS) model, which assumes that

$$\mu(t, r) = \mu(r) = a\,(\theta - r) \tag{11.49}$$

for some unknown parameters a and θ, and

$$\sigma(t, r) = \sigma r^{\gamma} \tag{11.50}$$

for some σ and γ. Be careful to distinguish between the volatility function $\sigma(t, r)$ and the constant volatility parameter σ.

We will use the `Irates` data set in the `Ecdat` package. This data set has interests rates for maturities from 1 to 120 months. We will use the first column, which has the one-month maturity rates, since we want the short rate.

Run the following code to input the data, compute the lagged and differenced short-rate series, and construct some basic plots.

```
library(Ecdat)
data(Irates)
r1 = Irates[,1]
n = length(r1)
lag_r1 = lag(r1)[-n]
delta_r1 = diff(r1)
n = length(lag_r1)
par(mfrow = c(3, 2))
plot(r1, main = "(a)")
plot(delta_r1, main = "(b)")
plot(delta_r1^2, main = "(c)")
plot(lag_r1, delta_r1, main = "(d)")
plot(lag_r1, delta_r1^2, main = "(e)")
```

Problem 1 *What is the maturity of the interest rates in the first column? What is the sampling frequency of this data set—daily, weekly, monthly, or quarterly? What country are the data from? Are the rates expressed as percentages or fractions (decimals)?*

In the plot you have just created, panels (a), (b), and (c) show how the short rate, changes in the short rate, and squared changes in the short rate depend on time. The plots of changes in the short rate are useful for choosing the drift $\mu(t-1, r_{t-1})$ while squared changes in the short rate are helpful for selecting the volatility $\sigma(t-1, r_{t-1})$.

Problem 2 *Model (11.49) states that $\mu(t, r) = \mu(r)$, that is, that the drift does not depend on t. Use your plots to discuss whether this assumption seems valid. Assuming for the moment that this assumption is valid, any trend in the plot in panel (d) would give us information about the form of $\mu(r)$. Do you see any trend?*

Now run the following code to fit model (11.49) and fill in the first two panels of a figure. This figure will be continued next.

```
#  CKLS (Chan, Karolyi, Longstaff, Sanders)

nlmod_CKLS = nls(delta_r1 ~ a * (theta-lag_r1),
   start=list(theta = 5,   a = 0.01),
   control = list(maxiter = 200))
param = summary(nlmod_CKLS)$parameters[ , 1]
par(mfrow = c(2, 2))
t = seq(from = 1946, to = 1991 + 2 / 12, length = n)
plot(lag_r1, ylim = c(0, 16), ylab = "rate and theta",
   main = "(a)", type = "l")
abline(h = param[1], lwd = 2, col = "red")
```

Problem 3 *What are the estimates of a and θ and their 95 % confidence intervals?*

Note that the nonlinear regression analysis estimates $\sigma^2(r)$, not $\sigma(r)$, since the response variable is the squared residual. Here $A = \sigma^2$ and $B = 2\gamma$.

```
res_sq = residuals(nlmod_CKLS)^2
nlmod_CKLS_res <- nls(res_sq ~  A*lag_r1^B,
   start = list(A = 0.2, B = 1/2))
param2 = summary(nlmod_CKLS_res)$parameters[ , 1]
plot(lag_r1, sqrt(res_sq), pch = 5, ylim = c(0, 6),
   main = "(b)")
attach(as.list(param2))
curve(sqrt(A * x^B), add = T, col = "red", lwd = 3)
```

Problem 4 *What are the estimates of σ and γ and their 95 % confidence intervals?*

Finally, refit model (11.49) using weighted least squares.

```
nlmod_CKLS_wt = nls(delta_r1 ~ a * (theta-lag_r1),
   start = list(theta = 5,  a = 0.01),
   control = list(maxiter = 200),
   weights = 1 / fitted(nlmod_CKLS_res))

plot(lag_r1, ylim = c(0, 16), ylab = "rate and theta",
   main = "(c)", type = "l")
param3 = summary(nlmod_CKLS_wt)$parameters[ , 1]
abline(h = param3[1], lwd = 2, col = "red")
```

Problem 5 *How do the unweighted estimate of θ shown in panel (a) and the weighted estimate plotted in panel (d) differ? Why do they differ in this manner?*

11.12.2 Response Transformations

This section uses the `HousePrices` data set in the `AER` package. This is a cross-sectional data set on house prices and other features, e.g., the number of bedrooms of houses in Windsor, Ontario. The data were gathered during the summer of 1987. Accurate modeling of house prices is important for the mortgage industry. Run the code below to read the data and regress `price` on the other variables; the period on the right-hand side of the formula "`price~ .`" specifies that the predictors should include all variables except, of course, the response.

```
library(AER)
data(HousePrices)
fit1 = lm(price ~ ., data = HousePrices)
summary(fit1)
```

Next construct a profile log-likelihood plot for the transformation parameter α in model (11.25)

```
library(MASS)
fit2 = boxcox(fit1, xlab = expression(alpha))
```

Problem 6 *What is the MLE of α? (Hint: Type* `?boxcox` *to learn what is returned by this function.)*

Next, fit a linear model with `price` transformed by $\widehat{\alpha}$ (the MLE). Here the function `bcPower()` in the `AER` package computes a Box–Cox transformation for a given value of α and must be distinguished from `boxcox()`, which computes the profile log-likelihood for α. In the following code, replace 1/2 by the MLE of α.

```
library(car)
alphahat = 1/2
fit3 = lm(bcPower(price, alphahat) ~ ., data = HousePrices)
summary(fit3)
AIC(fit1)
AIC(fit3)
```

Problem 7 *Does the Box–Cox transformation offer a substantial improvement in fit compared to the regression with no transformation of* `price`*?*

Problem 8 *Would it be worthwhile to check the residuals for correlation?*

11.12.3 Binary Regression: Who Owns an Air Conditioner?

This section uses the `HousePrices` data set used in Sect. 11.12.2. The goal here is to investigate how the presence or absence of air conditioning is related to the other variables. The code below fits a logistic regression model to all potential predictor variables and then uses `stepAIC()` to find a parsimonious model.

```
library(AER)
data(HousePrices)
fit1 = glm(aircon ~ ., family = "binomial",
   data = HousePrices)
summary(fit1)
library(MASS)
fit2 = stepAIC(fit1)
summary(fit2)
```

Problem 9 *Which variables are most useful for predicting whether a home has air conditioning? Describe qualitatively the relationships between these variables and the variable* `aircon`*. Are there any variables in the model selected by* `stepAIC()` *that you think might be dropped?*

Problem 10 *Estimate the probability that a house will have air conditioning if it has the following characteristics:*

```
price lotsize bedrooms bathrooms stories driveway recreation
42000    5850        3         1       2      yes         no
fullbase gasheat garage prefer
 yes       no      1     no
```

(Hint: The `R` *function* `plogis()` *computes the logistic function.)*

11.13 Exercises

1. When we were finding the best linear predictor of Y given X, we derived the equations
$$\begin{aligned} 0 &= -E(Y) + \beta_0 + \beta_1 E(X) \\ 0 &= -E(XY) + \beta_0 E(X) + \beta_1 E(X^2). \end{aligned}$$
Show that their solution is
$$\beta_1 = \frac{\sigma_{XY}}{\sigma_X^2}$$
and
$$\beta_0 = E(Y) - \beta_1 E(X) = E(Y) - \frac{\sigma_{XY}}{\sigma_X^2} E(X).$$
2. Suppose one has a long position of F_{20} face value in 20-year Treasury bonds and wants to hedge this with short positions in both 10- and 30-year Treasury bonds. The prices and durations of 10-, 20-, and 30-year Treasury bonds are P_{10}, DUR_{10}, P_{20}, DUR_{20}, P_{30}, and DUR_{30} and are assumed to be known. A regression of changes in the 20-year yield on changes in the 10- and 30-year yields is $\Delta y_{20} = \widehat{\beta}_0 + \widehat{\beta}_1 \Delta y_{10} + \widehat{\beta}_2 \Delta y_{30}$. The p-value of $\widehat{\beta}_0$ is large and it is assumed that β_0 is close enough to zero to be ignored. What face amounts F_{10} and F_{30} of 10- and 30-year Treasury bonds should be shorted to hedge the long position in 20-year Treasury bonds? (Express F_{10} and F_{30} in terms of the known quantities P_{10}, P_{20}, P_{30}, DUR_{10}, DUR_{20}, DUR_{30}, $\widehat{\beta}_1$, $\widehat{\beta}_2$, and F_{20}.)
3. The maturities (T) in years and prices in dollars of zero-coupon bonds are in file `ZeroPrices.txt` on the book's website. The prices are expressed

as percentages of par. A popular model is the Nelson–Siegel family with forward rate

$$r(T;\theta_1,\theta_2,\theta_3,\theta_4) = \theta_1 + (\theta_2 + \theta_3 T)\exp(-\theta_4 T).$$

Fit this forward rate to the prices by nonlinear regression using R's `optim()` function.

(a) What are your estimates of $\theta_1,\ \theta_2,\ \theta_3$, and θ_4?

(b) Plot the estimated forward rate and estimated yield curve on the same figure. Include the figure with your work.

4. Least-squares estimators are unbiased in linear models, but in nonlinear models they can be biased. Simulation studies (including bootstrap resampling) can be used to estimate the amount of bias. In Example 11.1, the data were simulated with $r = 0.06$ and $\widehat{r} = 0.0585$. Do you think this is a sign of bias or simply due to random variability? Justify your answer.

References

Atkinson, A. C. (1985) *Plots, Transformations and Regression*, Clarendon, Oxford.

Bates, D. M., and Watts, D. G. (1988) *Nonlinear Regression Analysis and Its Applications*, Wiley, New York.

Bluhm, C., Overbeck, L., and Wagner, C. (2003) *An Introduction to Credit Risk Modelling*, Chapman & Hall/CRC, Boca Raton, FL.

Box, G. E. P., and Dox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26** 211–246.

Carroll, R. J., and Ruppert, D. (1988) *Transformation and Weighting in Regression*, Chapman & Hall, New York.

Chan, K. C., Karolyi, G. A., Longstaff, F. A., and Sanders, A. B. (1992) An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance*, **47**, 1209–1227.

Faraway, J. J. (2006) *Extending the Linear Model with R*, Chapman & Hall, Boca Raton, FL.

Hosmer, D., and Lemeshow, S. (2000) *Applied Logistic Regression*, 2nd ed., Wiley, New York.

Jarrow, R. (2002) *Modeling Fixed-Income Securities and Interest Rate Options, 2nd Ed.*, Stanford University Press, Stanford, CA.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989) Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association*, **84**, 881–896.

Nelson, C. R., and Siegel, A. F. (1985) Parsimonious modelling of yield curves. *Journal of Business*, **60**, 473–489.

Seber, G. A. F., and Wild, C. J. (1989) *Nonlinear Regression*, Wiley, New York.

Svensson, L. E. (1994) Estimating and interpreting forward interest rates: Sweden 1992–94, Working paper. International Monetary Fund, 114.

Tuckman, B. (2002) *Fixed Income Securities*, 2nd ed., Wiley, Hoboken, NJ.