22

The method of least squares

The maximum likelihood principle provides a way to estimate parameters. The applicability of the method is quite general but not universal. For example, in the simple linear regression model, introduced in Section 17.4, we need to know the distribution of the response variable in order to find the maximum likelihood estimates for the parameters involved. In this chapter we will see how these parameters can be estimated using the method of least squares. Furthermore, the relation between least squares and maximum likelihood will be investigated in the case of normally distributed errors.

22.1 Least squares estimation and regression

Recall from Section 17.4 the simple linear regression model for a bivariate dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. In this model $x_1, x_2, \ldots, x_n$ are nonrandom and $y_1, y_2, \ldots, y_n$ are realizations of random variables $Y_1, Y_2, \ldots, Y_n$ satisfying

$$Y_i = \alpha + \beta x_i + U_i \qquad \text{for} \quad i = 1, 2, \ldots, n,$$

where $U_1, U_2, \ldots, U_n$ are independent random variables with zero expectation and variance σ^2. How can one obtain estimates for the parameters α, β, and σ^2 in this model?

Note that we cannot find maximum likelihood estimates for these parameters, simply because we have no further knowledge about the distribution of the U_i (and consequently of the Y_i). We want to choose α and β in such a way that we obtain a line that fits the data best. A classical approach to do this is to consider the sum of squared distances between the observed values y_i and the values $\alpha + \beta x_i$ on the regression line $y = \alpha + \beta x$. See Figure 22.1, where these distances are indicated. The *method of least squares* prescribes to choose α and β such that the sum of squares

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

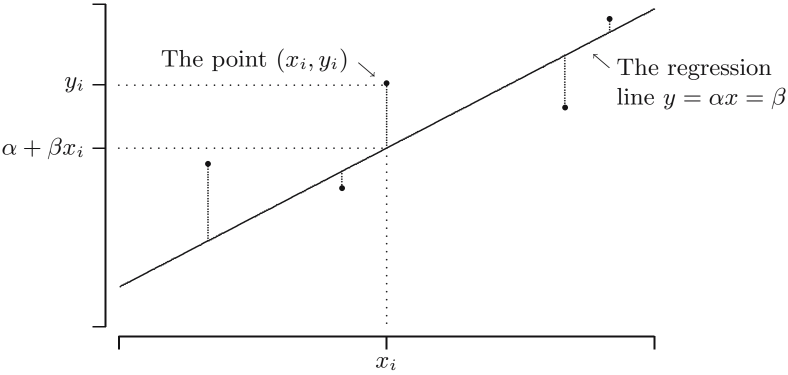

Fig. 22.1. The observed value y_i corresponding to x_i and the value $\alpha + \beta x_i$ on the regression line $y = \alpha + \beta x$.

is minimal. The ith term in the sum is the squared distance in the vertical direction from (x_i, y_i) to the line $y = \alpha + \beta x$. To find these so-called *least squares estimates*, we differentiate $S(\alpha, \beta)$ with respect to α and β, and we set the derivatives equal to 0:

$$\frac{\partial}{\partial \alpha} S(\alpha, \beta) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial}{\partial \beta} S(\alpha, \beta) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)\, x_i = 0.$$

This is equivalent to

$$n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

For example, for the timber data from Table 15.5 we would obtain

$$36\,\alpha + 1646.4\,\beta = 52\,901$$

$$1646.4\,\alpha + 81750.02\,\beta = 2\,790\,525.$$

These are two equations with two unknowns α and β. Solving for α and β yields the solutions $\hat{\alpha} = -1160.5$ and $\hat{\beta} = 57.51$. In Figure 22.2 a scatterplot of the timber dataset, together with the estimated regression line $y = -1160.5 + 57.51x$, is depicted.

Quick exercise 22.1 Suppose you are given a piece of Australian timber with density 65. What would you choose as an estimate for the Janka hardness?

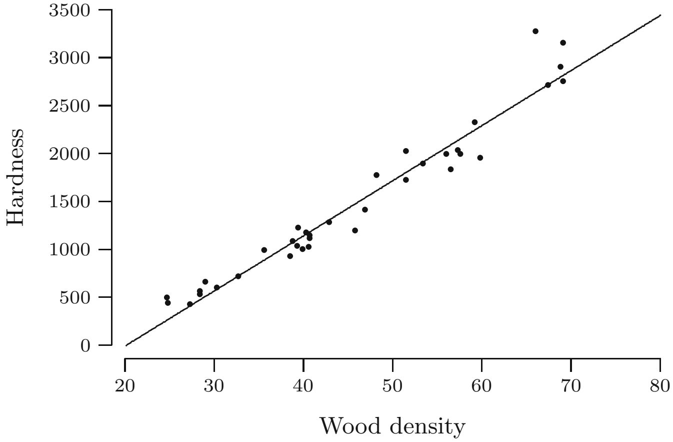

Fig. 22.2. Scatterplot and estimated regression line for the timber data.

In general, writing $\sum$ instead of $\sum_{i=1}^{n}$, we find the following formulas for the estimates $\hat{\alpha}$ (the *intercept*) and $\hat{\beta}$ (the *slope*):

$$\hat{\beta} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} \tag{22.1}$$

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n. \tag{22.2}$$

Since $S(\alpha, \beta)$ is an elliptic paraboloid (a "vase"), it follows that $(\hat{\alpha}, \hat{\beta})$ is the unique minimum of $S(\alpha, \beta)$ (except when all x_i are equal).

Quick exercise 22.2 Check that the line $y = \hat{\alpha} + \hat{\beta}x$ always passes through the "center of gravity" $(\bar{x}_n, \bar{y}_n)$.

Least squares estimators are unbiased

We denote the least squares *estimates* by $\hat{\alpha}$ and $\hat{\beta}$. It is quite common to also denote the least squares *estimators* by $\hat{\alpha}$ and $\hat{\beta}$:

$$\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n, \qquad \hat{\beta} = \frac{n \sum x_i Y_i - (\sum x_i)(\sum Y_i)}{n \sum x_i^2 - \left(\sum x_i\right)^2}.$$

In Exercise 22.12 it is shown that $\hat{\beta}$ is an unbiased estimator for β. Using this and the fact that $\mathrm{E}[Y_i] = \alpha + \beta x_i$ (see page 258), we find for $\hat{\alpha}$:

$$\begin{aligned}
\mathrm{E}[\hat{\alpha}] &= \mathrm{E}\left[\bar{Y}_n\right] - \bar{x}_n \mathrm{E}\left[\hat{\beta}\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[Y_i] - \bar{x}_n\beta \\
&= \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i) - \bar{x}_n\beta = \alpha + \beta\bar{x}_n - \bar{x}_n\beta \\
&= \alpha.
\end{aligned}$$

We see that $\hat{\alpha}$ is an unbiased estimator for α.

An unbiased estimator for σ^2

In the simple linear regression model the assumptions imply that the random variables Y_i are independent with variance σ^2. Unfortunately, one cannot apply the usual estimator $(1/(n-1))\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2$ for the variance of the Y_i (see Section 19.4), because different Y_i have different expectations. What would be a reasonable estimator for σ^2? The following quick exercise suggests a candidate.

Quick exercise 22.3 Let $U_1, U_2, \ldots, U_n$ be independent random variables, each with expected value zero and variance σ^2. Show that

$$T = \frac{1}{n}\sum_{i=1}^{n} U_i^2$$

is an unbiased estimator for σ^2.

At first sight one might be tempted to think that the unbiased estimator T from this quick exercise is a useful tool to estimate σ^2. Unfortunately, we only observe the x_i and Y_i, not the U_i. However, from the fact that $U_i = Y_i - \alpha - \beta x_i$, it seems reasonable to try

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \tag{22.3}$$

as an estimator for σ^2. Tedious calculations show that the expected value of this random variable equals $\frac{n-2}{n}\sigma^2$. But then we can easily turn it into an unbiased estimator for σ^2.

An unbiased estimator for σ^2. In the simple linear regression model the random variable

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

is an unbiased estimator for σ^2.

22.2 Residuals

A way to explore whether the simple linear regression model is appropriate to model a given bivariate dataset is to inspect a scatterplot of the so-called *residuals* r_i against the x_i. The ith residual r_i is defined as the vertical distance between the ith point and the estimated regression line:

$$r_i = y_i - \hat{\alpha} - \hat{\beta}x_i, \qquad i = 1, 2, \ldots, n.$$

When a linear model is appropriate, the scatterplot of the residuals r_i against the x_i should show truly random fluctuations around zero, in the sense that it should not exhibit any trend or pattern. This seems to be the case in Figure 22.3, which shows the residuals for the black cherry tree data from Exercise 17.9.

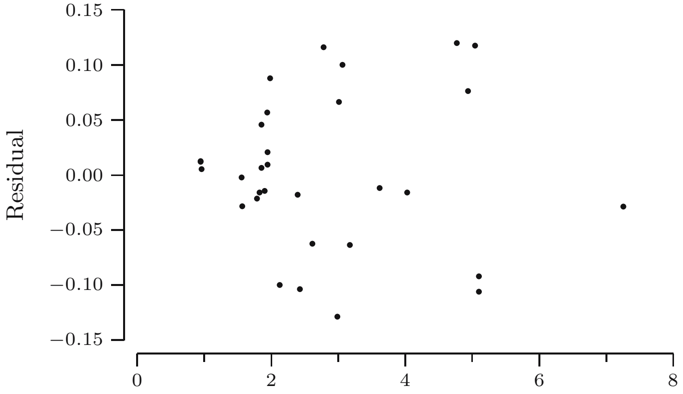

Fig. 22.3. Scatterplot of r_i versus x_i for the black cherry tree data.

Quick exercise 22.4 Recall from Quick exercise 22.2 that $(\bar{x}_n, \bar{y}_n)$ is on the regression line $y = \hat{\alpha} + \hat{\beta}x$, i.e., that $\bar{y}_n = \hat{\alpha} + \hat{\beta}\bar{x}_n$. Use this to show that $\sum_{i=1}^{n} r_i = 0$, i.e., that the sum of the residuals is zero.

In Figure 22.4 we depicted r_i versus x_i for the timber dataset. In this case a slight parabolic pattern can be observed. Figures 22.2 and 22.4 suggest that

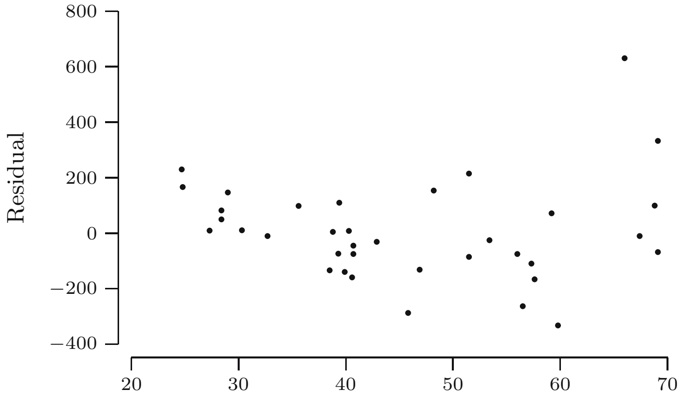

Fig. 22.4. Scatterplot of r_i versus x_i for the timber data with the simple linear regression model $Y_i = \alpha + \beta x_i + U_i$.

for the timber dataset a better model might be

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + U_i \qquad \text{for} \quad i = 1, 2, \ldots, n.$$

In this new model the residuals are

$$r_i = y_i - \hat{\alpha} - \hat{\beta} x_i - \hat{\gamma} x_i^2,$$

where $\hat{\alpha}, \hat{\beta}$, and $\hat{\gamma}$ are the least squares estimates obtained by minimizing

$$\sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i - \gamma x_i^2\right)^2.$$

In Figure 22.5 we depicted r_i versus x_i. The residuals display no trend or pattern, except that they "fan out"—an example of a phenomenon called *heteroscedasticity.*

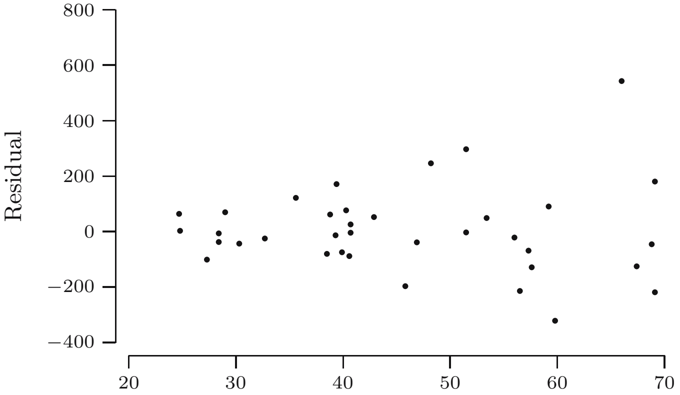

Fig. 22.5. Scatterplot of r_i versus x_i for the timber data with the model $Y_i = \alpha + \beta x_i + \gamma x_i^2 + U_i$.

Heteroscedasticity

The assumption of equal variance of the U_i (and therefore of the Y_i) is called *homoscedasticity.* In case the variance of Y_i depends on the value of x_i, we speak of *heteroscedasticity.* For instance, heteroscedasticity occurs when Y_i with a large expected value have a larger variance than those with small expected values. This produces a "fanning out" effect, which can be observed in Figure 22.5. This figure strongly suggests that the timber data are heteroscedastic. Possible ways out of this problem are a technique called weighted least squares or the use of variance-stabilizing transformations.

22.3 Relation with maximum likelihood

To apply the method of least squares no assumption is needed about the type of distribution of the U_i. In case the type of distribution of the U_i is known, the maximum likelihood principle can be applied. Consider, for instance, the classical situation where the U_i are independent with an $N(0, \sigma^2)$ distribution. What are the maximum likelihood estimates for α and β?

In this case the Y_i are independent, and Y_i has an $N(\alpha + \beta x_i, \sigma^2)$ distribution. Under these assumptions and assuming that the linear model is appropriate to model a given bivariate dataset, the r_i should look like the realization of a random sample from a normal distribution. As an example a histogram of the residuals r_i of the cherry tree data of Exercise 17.9 is depicted in Figure 22.6.

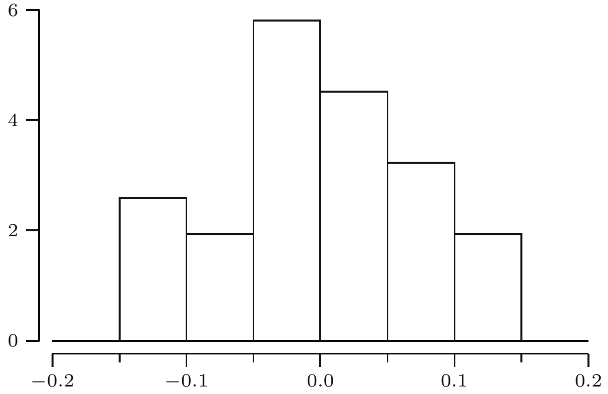

Fig. 22.6. Histogram of the residuals r_i for the black cherry tree data.

The data do not exhibit strong evidence against the assumption of normality. When Y_i has an $N(\alpha + \beta x_i, \sigma^2)$ distribution, the probability density of Y_i is given by

$$f_i(y) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(y-\alpha-\beta x_i)^2/(2\sigma^2)} \qquad \text{for} \quad -\infty < y < \infty.$$

Since

$$\ln\left(f_i(y_i)\right) = -\ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y_i - \alpha - \beta x_i}{\sigma}\right)^2,$$

the loglikelihood is:

$$\begin{aligned} \ell(\alpha, \beta, \sigma) &= \ln\left(f_1(y_1)\right) + \cdots + \ln\left(f_n(y_n)\right) \\ &= -n\ln(\sigma) - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2. \end{aligned}$$

Note that for any fixed $\sigma > 0$, the loglikelihood $\ell(\alpha, \beta, \sigma)$ attains its maximum precisely when $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ is minimal. Hence, in case the U_i are independent with an $N(0, \sigma^2)$ distribution, the maximum likelihood principle and the least squares method yield the *same* estimators.

To find the maximum likelihood estimate of σ we differentiate $\ell(\alpha, \beta, \sigma)$ with respect to σ:

$$\frac{\partial}{\partial \sigma}\ell(\alpha, \beta, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

It follows (from the invariance principle on page 321) that the maximum likelihood estimator of σ^2 is given by

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta} x_i)^2,$$

which is the estimator from (22.3).

22.4 Solutions to the quick exercises

22.1 We can use the estimated regression line $y = -1160.5 + 57.51x$ to predict the Janka hardness. For density $x = 65$ we find as a prediction for the Janka hardness $y = 2577.65$.

22.2 Rewriting $\hat{\alpha} = \bar{y}_n - \hat{\beta}$, it follows that $\bar{y}_n = \hat{\alpha} + \hat{\beta}\bar{x}_n$, which means that $(\bar{x}_n, \bar{y}_n)$ is a point on the estimated regression line $y = \hat{\alpha} + \hat{\beta} x$.

22.3 We need to show that $\mathrm{E}[T] = \sigma^2$. Since $\mathrm{E}[U_i] = 0$, $\mathrm{Var}(U_i) = \mathrm{E}\left[U_i^2\right]$, so that:

$$\mathrm{E}[T] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} U_i^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[U_i^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(U_i) = \sigma^2.$$

22.4 Since $r_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$ for $i = 1, 2, \ldots, n$, it follows that the sum of the residuals equals

$$\begin{aligned}\sum r_i &= \sum y_i - \left(n\hat{\alpha} + \hat{\beta}\sum x_i\right)\\ &= n\bar{y}_n - \left(n\hat{\alpha} + n\hat{\beta}\bar{x}_n\right) = n\left(\bar{y}_n - (\hat{\alpha} + \hat{\beta}\bar{x}_n)\right) = 0,\end{aligned}$$

because $\bar{y}_n = \hat{\alpha} + \hat{\beta}\bar{x}_n$, according to Quick exercise 22.2.

22.5 Exercises

22.1 ⊞ Consider the following bivariate dataset:

$$(1, 2) \quad (3, 1.8) \quad (5, 1).$$

a. Determine the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters of the regression line $y = \alpha + \beta x$.

b. Determine the residuals r_1, r_2, and r_3 and check that they add up to 0.

c. Draw in one figure the scatterplot of the data and the estimated regression line $y = \hat{\alpha} + \hat{\beta}x$.

22.2 Adding one point may dramatically change the estimates of α and β. Suppose one extra datapoint is added to the dataset of the previous exercise and that we have as dataset:

$$(0, 0) \quad (1, 2) \quad (3, 1.8) \quad (5, 1).$$

Determine the least squares estimate of $\hat{\beta}$. A point such as $(0, 0)$, which dramatically changes the estimates for α and β, is called a *leverage point*.

22.3 Suppose we have the following bivariate dataset:

$$(1, 3.1) \quad (1.7, 3.9) \quad (2.1, 3.8) \quad (2.5, 4.7) \quad (2.7, 4.5).$$

a. Determine the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters of the regression line $y = \alpha + \beta x$. You may use that $\sum x_i = 10$, $\sum y_i = 20$, $\sum x_i^2 = 21.84$, and $\sum x_i y_i = 41.61$.

b. Draw in one figure the scatterplot of the data and the estimated regression line $y = \hat{\alpha} + \hat{\beta}x$.

22.4 We are given a bivariate dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_{100}, y_{100})$. For this bivariate dataset it is known that $\sum x_i = 231.7$, $\sum x_i^2 = 2400.8$, $\sum y_i = 321$, and $\sum x_i y_i = 5189$. What are the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters of the regression line $y = \alpha + \beta x$?

22.5 ⊞ For the timber dataset it seems reasonable to leave out the intercept α ("no hardness without density"). The model then becomes

$$Y_i = \beta x_i + U_i \qquad \text{for} \quad i = 1, 2, \ldots, n.$$

Show that the least squares estimator $\hat{\beta}$ of β is now given by

$$\hat{\beta} = \frac{\displaystyle\sum_{i=1}^{n} x_i Y_i}{\displaystyle\sum_{i=1}^{n} x_i^2}$$

by minimizing the appropriate sum of squares.

22.6 ⊡ (Quick exercise 22.1 and Exercise 22.5 continued). Suppose we are given a piece of Australian timber with density 65. What would you choose as an estimate for the Janka hardness, based on the regression model with no intercept? Recall that $\sum x_i y_i = 2790525$ and $\sum x_i^2 = 81750.02$ (see also Section 22.1).

22.7 Consider the dataset

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

where $x_1, x_2, \ldots, x_n$ are nonrandom and $y_1, y_2, \ldots, y_n$ are realizations of random variables $Y_1, Y_2, \ldots, Y_n$, satisfying

$$Y_i = \mathrm{e}^{\alpha + \beta x_i} + U_i \qquad \text{for} \quad i = 1, 2, \ldots, n.$$

Here $U_1, U_2, \ldots, U_n$ are independent random variables with zero expectation and variance σ^2. What are the least squares estimates for the parameters α and β in this model?

22.8 ⊡ Which simple regression model has the larger *residual sum of squares* $\sum_{i=1}^{n} r_i^2$, the model with intercept or the one without?

22.9 For some datasets it seems reasonable to leave out the slope β. For example, in the jury example from Section 6.3 it was assumed that the score that juror i assigns when the performance deserves a score g is $Y_i = g + Z_i$, where Z_i is a random variable with values around zero. In general, when the slope β is left out, the model becomes

$$Y_i = \alpha + U_i \quad \text{for } i = 1, 2, \ldots, n.$$

Show that $\bar{Y}_n$ is the least squares estimator $\hat{\alpha}$ of α.

22.10 ⊡ In the method of least squares we choose α and β in such a way that the sum of squared residuals $S(\alpha, \beta)$ is minimal. Since the ith term in this sum is the squared vertical distance from (x_i, y_i) to the regression line $y = \alpha + \beta x$, one might also wonder whether it is a good idea to replace this squared distance simply by the distance. So, given a bivariate dataset

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

choose α and β in such a way that the sum

$$A(\alpha, \beta) = \sum_{i=1}^{n} |y_i - \alpha - \beta x_i|$$

is minimal. We will investigate this by a simple example. Consider the following bivariate dataset:

$$(0, 2), (1, 2), (2, 0).$$

a. Determine the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$, and draw in one figure the scatterplot of the data and the estimated regression line $y = \hat{\alpha} + \hat{\beta}x$. Finally, determine $A(\hat{\alpha}, \hat{\beta})$.

b. One might wonder whether $\hat{\alpha}$ and $\hat{\beta}$ also minimize $A(\alpha, \beta)$. To investigate this, choose $\beta = -1$ and find α's for which $A(\alpha, -1) < A(\hat{\alpha}, \hat{\beta})$. For which α is $A(\alpha, -1)$ minimal?

c. Find α and β for which $A(\alpha, \beta)$ is minimal.

22.11 Consider the dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where the x_i are nonrandom and the y_i are realizations of random variables $Y_1, Y_2, \ldots, Y_n$ satisfying

$$Y_i = g(x_i) + U_i \quad \text{for } i = 1, 2, \ldots, n,$$

where $U_1, U_2, \ldots, U_n$ are independent random variables with zero expectation and variance σ^2. Visual inspection of the scatterplot of our dataset in

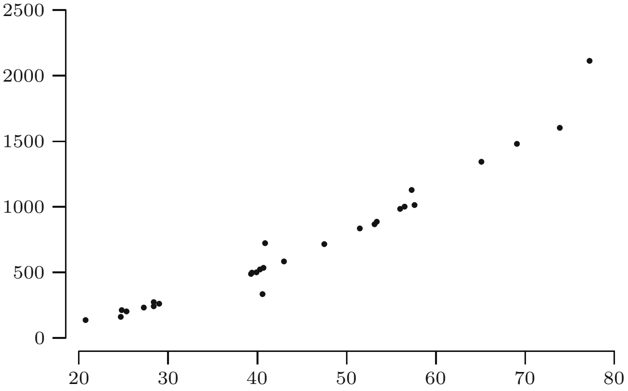

Fig. 22.7. Scatterplot of y_i versus x_i.

Figure 22.7 suggests that we should model the Y_i by

$$Y_i = \beta x_i + \gamma x_i^2 + U_i \quad \text{for } i = 1, 2, \ldots, n.$$

a. Show that the least squares estimators $\hat{\beta}$ and $\hat{\gamma}$ satisfy

$$\beta \sum x_i^2 + \gamma \sum x_i^3 = \sum x_i y_i,$$
$$\beta \sum x_i^3 + \gamma \sum x_i^4 = \sum x_i^2 y_i.$$

b. Infer from **a**—for instance, by using linear algebra—that the estimators $\hat{\beta}$ and $\hat{\gamma}$ are given by

$$\hat{\beta} = \frac{(\sum x_i Y_i)(\sum x_i^4) - (\sum x_i^3)(\sum x_i^2 Y_i)}{(\sum x_i^2)(\sum x_i^4) - (\sum x_i^3)^2}$$

and

$$\hat{\gamma} = \frac{(\sum x_i^2)(\sum x_i^2 Y_i) - (\sum x_i^3)(\sum x_i Y_i)}{(\sum x_i^2)(\sum x_i^4) - (\sum x_i^3)^2}.$$

22.12 ⊞ The least square estimator $\hat{\beta}$ from (22.1) is an unbiased estimator for β. You can show this in four steps.

a. First show that

$$\mathrm{E}\left[\hat{\beta}\right] = \frac{n \sum x_i \mathrm{E}\left[Y_i\right] - (\sum x_i)(\sum \mathrm{E}\left[Y_i\right])}{n \sum x_i^2 - \left(\sum x_i\right)^2}.$$

b. Next use that $\mathrm{E}\left[Y_i\right] = \alpha + \beta x_i$, to obtain that

$$\mathrm{E}\left[\hat{\beta}\right] = \frac{n \sum x_i(\alpha + \beta x_i) - (\sum x_i)\left[n\alpha + \beta \sum x_i\right]}{n \sum x_i^2 - \left(\sum x_i\right)^2}.$$

c. Simplify this last expression to find

$$\mathrm{E}\left[\hat{\beta}\right] = \frac{n\alpha \sum x_i + n\beta \sum x_i^2 - n\alpha \sum x_i - \beta(\sum x_i)^2}{n \sum x_i^2 - \left(\sum x_i\right)^2}.$$

d. Finally, conclude that $\hat{\beta}$ is an unbiased estimator for β.