



## Chapter 5

# Linear Regression and Least Squares

Linear mathematics is widely used as a basic modeling tool throughout the mathematical sciences. It will therefore come as not surprise that the most basic statistical model used to study the relationship between a dependent variable and a set of independent variables is a linear model, specifically the “linear regression model”. In this chapter, we will introduce the basic regression model, establish its close relationship to the theory of least squares, and describe a number of modeling extensions to the basic regression model. These different modeling extensions should serve as a powerful illustration of the fact that when one statistically analyzes a given data set, one is indeed involved in a modeling exercise. Changing the modeling assumptions can have a profound impact on the statistical conclusions that one reaches from a particular data set.

### 5.1 The Method of Least Squares

The method of least squares is generally credited to Carl Friedrich Gauss, who developed the basic idea in 1795 at the age of eighteen, but who did not publish the method until 1809. In 1801, Gauss used his least squares method to help relocate the asteroid Ceres after it emerged from the glare of the sun, based on 40 days of earlier tracking data.

Least squares arises in the analysis of a data set  $\{(x_i, y_i) : 1 \leq i \leq n\}$  for which one believes that the dependent variable  $y$  can be explained as a linear function of the independent variable  $x$ . We wish to use the data set to “fit” a linear model to the relationship between  $y$  and  $x$ . When the number of data points is great than or equal to 3, such an empirical data set will rarely fit exactly on a straight line.

In view of this lack of exact fit, one can choose to fit the slope  $a$  and intercept  $b$  associated with the linear model so as to minimize a suitable chosen error criterion. One natural error criterion involves selecting  $a$  and  $b$  so as to solve the optimization problem

$$\min_{a,b} \sum_{i=1}^n |y_i - ax_i - b|^p \tag{5.1}$$

for  $p \geq 1$ . The minimizing  $\hat{a}$  and  $\hat{b}$  then satisfy the first-order optimality conditions

$$\begin{aligned} \sum_{i=1}^n |y_i - \hat{a}x_i - \hat{b}|^{p-1} \mathbf{I} \left\{ y_i \geq \hat{a}x_i + \hat{b} \right\} (-x_i) + \sum_{i=1}^n |y_i - \hat{a}x_i - \hat{b}|^{p-1} \mathbf{I} \left\{ \hat{a}x_i + \hat{b} \geq y_i \right\} x_i &= 0, \\ \sum_{i=1}^n |y_i - \hat{a}x_i - \hat{b}|^{p-1} \mathbf{I} \left\{ y_i \geq \hat{a}x_i + \hat{b} \right\} (-1) + \sum_{i=1}^n |y_i - \hat{a}x_i - \hat{b}|^{p-1} \mathbf{I} \left\{ \hat{a}x_i + \hat{b} \geq y_i \right\} (+1) &= 0. \end{aligned}$$

Gauss observed that when  $p = 2$ , the above equations simplify to a pair of simultaneous linear equations in  $\hat{a}$  and  $\hat{b}$

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix}.$$

When the  $x$ -values  $x_1, x_2, \dots, x_n$  are distinct, the above linear system has a unique solution

$$\hat{a} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}_n \bar{y}_n}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}_n^2}, \quad \hat{b} = \bar{y}_n - \hat{a} \bar{x}_n, \quad (5.2)$$

where  $\bar{y}_n = (y_1 + \dots + y_n)/n$  and  $\bar{x}_n = (x_1 + \dots + x_n)/n$ .

Choosing  $p = 2$  in (5.1) (and hence selecting  $a$  and  $b$  so as to minimize the sum of squared errors) leads to even greater computational simplification when the number of explanatory variables is large. In particular, consider the data set  $\{(\tilde{x}_i, y_i) : 1 \leq i \leq n\}$  in which  $\tilde{x}_i \in \mathbb{R}^d$  (written a column vector) and  $y_i$  is the real-valued dependent variable. The method of least squares suggests choosing  $\tilde{a} \in \mathbb{R}^d$  (written as a column vector) and  $b \in \mathbb{R}$  so as to minimize

$$\sum_{i=1}^n (y_i - \tilde{x}_i^T \tilde{a} - b)^2 \quad (5.3)$$

over  $\tilde{a}$  and  $b$ . Note that is we set  $x_i^T = (\tilde{x}_i^T, 1)$  and  $a^T = (\tilde{a}^T, b)$ , we can remove the intercept term  $b$  from (5.3), yielding

$$\sum_{i=1}^n (y_i - x_i^T a)^2.$$

This can be re-written using matrix-vector notation as

$$(y - Xa)^T (y - Xa), \quad (5.4)$$

where  $y = (y_1, \dots, y_n)^T$  and

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}.$$

The minimizing  $\hat{a}$  of the quadratic form (5.4) then satisfies the linear system

$$X^T X \hat{a} = X^T y,$$

so that if  $X^T X$  is non-singular, we arrive at the least squares solution

$$\hat{a} = (X^T X)^{-1} X^T y.$$

The fact that the least squares formulation leads to a linear system for the minimizing  $\hat{a}$  makes this method exceptionally appealing from a computational viewpoint, and largely explains the method's popularity. In the next section we shall see that there is very natural statistical model that leads to the estimator  $\hat{a}$  that coincides with that produced by least squares. This statistical interpretation of the least squares estimator offers significant additional benefits relative to the least squares development provided above. In particular, a statistical formulation of the problem of fitting a linear model to observed data allows one to:

1. Develop "error bars" (i.e. confidence intervals) that describe the amount of sampling uncertainty that is present in the estimator  $\hat{a}$ ;
2. Test hypothesis (such as whether a simplified model involving fewer explanatory variables adequately explains the observed data);

3. Extrapolate the model to permit predictions outside the observed range of the data, with "error bars" (i.e. prediction intervals) that indicate the associated uncertainty;
4. Develop improved least squares estimates, in which the improvements are suggested by appropriately refining the underlying statistical assumptions (e.g. "weighted least squares" arising as a consequence of a model in which the variance is non-constant across the observations).

## 5.2 The Basic Linear Regression Model

We start by describing the most basic linear model considered by statisticians, in which there is a single explanatory variable. Consider, for example, a clinical trial in which a drug treatment has been administered at dosage levels  $x_1, x_2, \dots, x_n$  with corresponding treatment responses  $Y_1, Y_2, \dots, Y_n$ . We postulate the following linear model for the data:

$$Y_i = a^* x_i + b^* + \epsilon_i,$$

for some constant  $a^*$  and  $b^*$ , where the "residual errors"  $\epsilon_1, \epsilon_2, \dots, \epsilon_n$  are iid  $N(0, \sigma_*^2)$  rv's.

Estimating the parameters  $a^*$ ,  $b^*$  and  $\sigma_*^2$  via maximum likelihood is straightforward for this model; the likelihood function is

$$L_n(a, b, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - ax_i - b)^2}{2\sigma^2}\right),$$

so that the log-likelihood is

$$\mathcal{L}_n(a, b, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - ax_i - b)^2.$$

The MLE's are then given by

$$\hat{a}_n = \frac{\frac{1}{n} \sum_{i=1}^n x_i Y_i - \bar{x}_n \bar{Y}_n}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}_n^2}, \quad \hat{b}_n = \bar{Y}_n - \hat{a}_n \bar{x}_n, \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{a}_n x_i - \hat{b}_n)^2,$$

where  $\bar{Y}_n = (Y_1 + \dots + Y_n)/n$ . Note that the estimators for  $a^*$  and  $b^*$  are exactly identical to those derived earlier from Gauss' least squares argument; see (5.2). So, this statistical model recovers the least squares estimators obtained in Section 5.1. However, because we have imposed additional statistical structure on the residual errors, we can exploit this structure to answer various questions about the underlying model.

1. *Confidence intervals for  $a^*$  and  $b^*$ :* Let

$$\hat{\sigma}_n^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{a}_n x_i - \hat{b}_n)^2 \quad \text{and} \quad \tilde{s}_{xx}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}_n^2.$$

Then

$$\sqrt{n} \frac{\hat{a}_n - a^*}{\sqrt{\hat{\sigma}_n^2 / \tilde{s}_{xx}^2}} \stackrel{\mathcal{D}}{=} t_{n-2} \quad \text{and} \quad \sqrt{n} \frac{(\hat{a}_n - a^*)}{\sqrt{\hat{\sigma}_n^2 (1 + \bar{x}_n^2 / \tilde{s}_{xx}^2)}} \stackrel{\mathcal{D}}{=} t_{n-2},$$

where  $t_{n-2}$  is a Student- $t$  rv with  $n-2$  degrees of freedom (and is a so-called "tabulated distribution"). It follows that

$$\left[ \hat{a}_n - z \sqrt{\frac{\hat{\sigma}_n^2}{n \tilde{s}_{xx}^2}}, \hat{a}_n + z \sqrt{\frac{\hat{\sigma}_n^2}{n \tilde{s}_{xx}^2}} \right]$$

and

$$\left[ \hat{b}_n - z \sqrt{\frac{\hat{\sigma}_n^2}{n} (1 + \bar{x}_n^2 / \tilde{s}_{xx}^2)}, \hat{b}_n + z \sqrt{\frac{\hat{\sigma}_n^2}{n} (1 + \bar{x}_n^2 / \tilde{s}_{xx}^2)} \right]$$

are *exact*  $100(1 - \delta)\%$  confidence intervals for  $a^*$  and  $b^*$ , respectively, provided  $z$  is selected so that  $P\{-z \leq t_{n-2} \leq z\} = 1 - \delta$ .

2. *Hypothesis testing*: Suppose that we wish to see whether the data set  $\{(x_i, Y_i) : 1 \leq i \leq n\}$  can be adequately explained by the simple model in which  $a^* = 0$ . In other words, we wish to test:

$$H_0 : a^* = 0$$

versus

$$H_1 : a^* \neq 0.$$

Noting that

$$\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2,$$

we expect that if  $a^* \neq 0$ , then  $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$  will be small compared to  $\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2$ . So we expect

$$\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2}{\hat{\sigma}_n^2}$$

will tend to be large when  $a^* \neq 0$  and small when  $a^* = 0$ . More precisely, it turns out that when  $a^* = 0$ ,

$$\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2}{\hat{\sigma}_n^2} \stackrel{\mathcal{D}}{=} F_{1, n-2},$$

where  $F_{1, n-2}$  has an  $F$  distribution with 1 and  $n-2$  degrees of freedom (and is a “tabulated distribution”). It follows that if we set  $z$  so that  $P\{F_{1, n-2} \geq z\} = \alpha$  and reject  $H_0$  if

$$\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2}{\hat{\sigma}_n^2} \geq z,$$

then the Type 1 error is  $\alpha$  (i.e. the probability of incorrectly rejecting  $H_0$  when  $H_0$  is true is  $\alpha$ ).

3. *Prediction intervals*: Suppose that we wish to extrapolate the model to an  $x$ -value at which no observations have yet been taken. Let  $Y|x$  be the value of the dependent variable at  $x$ . Note that

$$Y|x \stackrel{\mathcal{D}}{=} a^*x + b^* + N(0, \sigma_*^2),$$

so that the interval

$$[a^*x + b^* - \sigma_*z, a^*x + b^* + \sigma_*z]$$

contains  $Y|x$  with probability  $1 - \delta$ , provided that  $z$  is selected so that  $P\{-x \leq N(0, 1) \leq z\} = 1 - \delta$ .

The above interval presumes knowledge of  $a^*$ ,  $b^*$  and  $\sigma_*^2$ . Since the parameters must be estimated from the existing data, the corresponding “prediction interval” must incorporate the parameter uncertainty. Suppose we select  $z$  so that  $P\{-z \leq t_{n-2} \leq z\} = 1 - \delta$ . Then,

$$\left[ \hat{a}_n + \hat{b}_n - \hat{\sigma}_n z \sqrt{1 + \frac{(1 + (x - \bar{x}_n / \hat{s}_{xx}^2)^2)}{n}}, \hat{a}_n + \hat{b}_n + \hat{\sigma}_n z \sqrt{1 + \frac{(1 + (x - \bar{x}_n / \hat{s}_{xx}^2)^2)}{n}} \right]$$

is the desired interval.

**Exercise 5.1:** Prove that  $\hat{a}_n$ ,  $\hat{b}_n$  and  $\hat{\sigma}_n^2$  are unbiased estimators for  $a^*$ ,  $b^*$  and  $\sigma_*^2$  (when  $n \geq 2$ ).

**Exercise 5.2:** Suppose that the “design points” ( $x_i : i \geq 1$ ) are chosen so that there exists  $c$  for which

$$-\infty < \inf_i x_i \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i I\{x_i \leq c\} < \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i I\{x_i > c\} \leq \sup_i x_i < \infty.$$

Prove that

$$\hat{a}_n \xrightarrow{p} a^*$$

as  $n \rightarrow \infty$ .

**Exercise 5.3:** The *coefficient of determination* is defined by

$$R^2 = \frac{(\sum_{i=1}^n x_i Y_i - n\bar{x}_n \bar{Y}_n)^2}{(\sum_{i=1}^n x_i^2 - n\bar{x}_n^2)(\sum_{i=1}^n Y_i^2 - n\bar{Y}_n^2)}.$$

Show that  $R^2$  can be rewritten as

$$R^2 = \frac{\sum_{i=1}^n (Y_i - \hat{a}_n x_i - \hat{b}_n)^2}{\sum_{i=1}^n Y_i^2 - n\bar{Y}_n^2}$$

and that  $R^2 \in [0, 1]$ . (Note that when  $R^2$  is close to 1, this suggests that most of the variation in the dependant variables can be explained by the regression.)

### 5.3 Weighted Least Squares

In many settings, the assumption the residual error has a common variance across all  $x$ -values is unrealistic. (The assumption of common variance is called “homoscedacity” by statisticians.) Suppose that we instead postulate that the error at  $x$  follows a Gaussian distribution with mean zero and variance  $\sigma_*^2 = \delta(x)$ , where  $\delta(\cdot)$  is a known positive function and  $\sigma_*^2$  is an unknown parameter. For example, if we assume (as is reasonable in many contexts) that the error scales in proportion to  $x$  (and  $Y|x$ ), then  $\delta(x) = x^2$ .

In the presence of such a model,

$$Y_i = a^* x_i + b^* + \epsilon_i,$$

where there  $\epsilon_i$ 's are independent and  $\epsilon_i$  is  $N(0, \delta(x_i)\sigma_*^2)$ . The likelihood is again straightforward to compute:

$$L_n(a, b, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2\delta(x_i)}} \exp\left(-\frac{(Y_i - ax_i - b)^2}{2\sigma^2\delta(x_i)}\right).$$

The MLE estimator  $\hat{a}_n$  and  $\hat{b}_n$  for  $a^*$  and  $b^*$  now are minimizers of

$$\min_{a,b} \sum_{i=1}^n \frac{(Y_i - ax_i - b)^2}{\delta(x_i)},$$

and hence minimize a “weighted least squares” problem. The estimators  $\hat{a}_n$  and  $\hat{b}_n$  “down-weight” the less reliable observations (for which the variance  $\sigma_*^2\delta(x_i)$  is large), as makes sense intuitively. Note that the statistical interpretation of the “weights”  $1/\delta(x_i)$  provide guidance to the modeler in appropriately selecting the weights.

In this setting,  $\hat{a}_n$  and  $\hat{b}_n$  satisfy the linear system

$$\begin{pmatrix} \sum_{i=1}^n x_i^2/\delta(x_i) & \sum_{i=1}^n x_i/\delta(x_i) \\ \sum_{i=1}^n x_i/\delta(x_i) & \sum_{i=1}^n 1/\delta(x_i) \end{pmatrix} \begin{pmatrix} \hat{a}_n \\ \hat{b}_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n Y_i x_i/\delta(x_i) \\ \sum_{i=1}^n Y_i/\delta(x_i) \end{pmatrix}$$

and are given by

$$\hat{a}_n = \frac{\frac{\sum_{i=1}^n \frac{Y_i x_i}{\delta(x_i)}}{\sum_{i=1}^n \frac{1}{\delta(x_i)}} - \bar{x}_n \bar{Y}_n}{\frac{\sum_{i=1}^n \frac{x_i^2}{\delta(x_i)}}{\sum_{i=1}^n \frac{1}{\delta(x_i)}} - \bar{x}_n^2} \quad \text{and} \quad \hat{b}_n = \bar{Y}_n - \hat{a}_n \bar{x}_n,$$

where

$$\bar{x}_n = \frac{\sum_{i=1}^n \frac{x_i}{\delta(x_i)}}{\sum_{i=1}^n \frac{1}{\delta(x_i)}} \quad \text{and} \quad \bar{Y}_n = \frac{\sum_{i=1}^n \frac{Y_i}{\delta(x_i)}}{\sum_{i=1}^n \frac{1}{\delta(x_i)}}.$$

Furthermore, the ML estimator for  $\sigma_*^2$  is

$$\tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{a}_n x_i - \hat{b}_n)^2 / \delta(x_i).$$

**Exercise 5.4:** Prove that  $\hat{a}_n$  and  $\hat{b}_n$  are unbiased estimators for  $a^*$  and  $b^*$ .

**Exercise 5.5:** Suppose that the design  $(x_i : i \geq 1)$  satisfies the assumption of Exercise 5.2 and, in addition

$$0 < \liminf_{n \rightarrow \infty} \delta(x_i) < \limsup_{n \rightarrow \infty} \delta(x_n) < \infty.$$

Prove that

$$\hat{a}_n \xrightarrow{p} a^*$$

as  $n \rightarrow \infty$ .

## 5.4 Applying Linear Regression in the Presence of Nonlinearity

Suppose that we observe a biological population over time that is subject to Malthusian growth. We then expect that the population  $y$ , as a function of time  $t$ , approximately satisfies,

$$y(t) \approx c^* e^{a^* t}.$$

It follows that

$$\log y(t) \approx a^* t + b^*,$$

where  $b^* = \log c^*$ . In other words,  $\log y(t)$  can be approximately explained as a linear function of  $t$ .

This suggests that if  $\{(Y(t_i), t_i) : 1 \leq i \leq n\}$  is the corresponding data set, a natural statistical model is to hypothesize that

$$\log Y(t_i) = a^* t_i + b^* + \epsilon_i \tag{5.5}$$

where there  $\epsilon_i$ 's are iid and follow a  $N(0, \sigma_*^2)$  distribution. One then applies the linear regression ideas of Section 5.2 to  $\{(\log Y(t_i), t_i) : 1 \leq i \leq n\}$ .

Of course, the final inferences to be drawn will typically concern  $Y(\cdot)$ , not the transformed  $\log Y(\cdot)$ . For example, we are likely to be more interested in a prediction interval for  $Y(t)$  than for  $\log Y(t)$ . However, if  $[L_n, R_n]$  is a  $100(1 - \delta)\%$  prediction interval for  $\log Y(t)$  (as obtained through the methods of Section 5.2), then  $[\exp(L_n), \exp(R_n)]$  is the approximate prediction interval for  $Y(t)$ .

It should further be noted that the model (5.5) is equivalent to requiring that

$$Y(t_i) = Z_i c^* e^{a^* t_i} \tag{5.6}$$

for  $1 \leq i \leq n$  where the  $Z_i$ 's are iid log-normal rv's with parameters 0 and  $\sigma_*^2$ . Thus, the model (5.6), when expressed in terms of the  $Y(t_i)$ 's, assumes multiplicative noise and log-normal errors. If one insists on assuming an additive error model for the data set  $\{(Y(t_i), t_i) : 1 \leq i \leq n\}$  one is led to the statistical model in which

$$Y(t_i) = c^* e^{a^* t_i} + \epsilon_i,$$

where the  $\epsilon_i$ 's are iid  $N(0, \sigma_*^2)$  rv's. In this setting, the ML estimators  $\hat{a}_n$  and  $\hat{c}_n$  for  $a^*$  and  $c^*$  are minimizers of

$$\min_{a, c} \sum_{i=1}^n (Y(t_i) - c e^{a t_i})^2,$$

so that  $\hat{a}_n$  and  $\hat{c}_n$  solve a nonlinear least-squares optimization problem (and the corresponding statistical model is an example of a non-linear regression problem).

We conclude this section with a brief discussion of the case in which the dependent variable  $y$  is believed to approximately follow a power law relationship as a function of the independent variable  $x$ , so that

$$y \approx c^* x^{a^*}.$$

In this case,

$$\log y \approx a^* \log x + b^*,$$

where  $b^* = \log c^*$ . In other words,  $\log y$  can be approximately explained by a linear function of  $\log x$ . Given a data set  $\{(Y_i, x_i) : 1 \leq i \leq n\}$ , the associated “transformed linear regression model” is

$$\log Y_i = a^* \log x_i + b^* + \epsilon_i,$$

where the  $\epsilon_i$ 's are iid  $N(0, \sigma_*^2)$  rv's. This, of course, is equivalent to the multiplicative noise model for which

$$Y_i = Z_i c^* x_i^{a^*} \tag{5.7}$$

for  $i \geq 1$ , where the  $Z_i$ 's are iid log-normal rv's with parameters 0 and  $\sigma_*^2$ . As in the setting of exponential growth, a assumption of additive error directly on the  $Y_i$ 's (rather than the  $\log Y_i$ 's) leads to a non-linear least squares problem, rather than a (linear) least squares problem.

**Exercise 5.6:**

1. Show that  $E[Z_i] > 1$  in (5.6) and (5.7).
2. Does this concern you? Explain your answer.

**Exercise 5.7:** Find a data set corresponding to the magnitude of the global human population recorded over time since 1850. Predict the global population in 2050 using both the model (5.5) and the model (5.6).

## 5.5 The Bootstrap for Regression Problems

We now consider a variant of the linear regression model in which we relax the assumption that the residual errors follow a Gaussian distribution. Specifically, we consider a data set  $\{(Y_i, x_i) : 1 \leq i \leq n\}$  for which the design points  $(x_i : i \geq 1)$  are viewed as deterministic (as in the previous models considered in this chapter) and for which

$$Y_i = a^* x_i + b^* + \epsilon_i, \tag{5.8}$$

where the  $\epsilon_i$ 's are iid rv's with  $E[\epsilon_i] = 0$  and  $\sigma_*^2 = \text{var}(\epsilon_i) < \infty$ . To estimate  $a^*$  and  $b^*$ , we consider estimators  $\hat{a}_n$  and  $\hat{b}_n$  that are linear in the  $Y_i$ 's, so that

$$\hat{a}_n = \sum_{i=1}^n w_i Y_i \quad \text{and} \quad \hat{b}_n = \sum_{i=1}^n \tilde{w}_i Y_i.$$

We want the estimators  $\hat{a}_n$  and  $\hat{b}_n$  to be unbiased, so that

$$E[\hat{a}_n] = a^* \quad \text{and} \quad E[\hat{b}_n] = b^*.$$

But,

$$E[\hat{a}_n] = a^* \sum_{i=1}^n w_i x_i + b^* \sum_{i=1}^n w_i,$$

so that unbiasedness of  $\hat{a}_n$  requires that

$$a^* \left( \sum_{i=1}^n w_i x_i - 1 \right) + b^* \sum_{i=1}^n w_i = 0$$

for all  $a^*$  and  $b^*$ . This implies that the  $w_i$ 's must satisfy

$$\sum_{i=1}^n w_i x_i = 1 \quad \text{and} \quad \sum_{i=1}^n w_i = 0 \quad (5.9)$$

The *best linear unbiased estimator* (BLUE) for  $a^*$  corresponds to the choice  $(w_1^*, w_2^*, \dots, w_n^*)$  that minimizes the mean square error

$$\mathbb{E} \left[ \left( \sum_{i=1}^n w_i Y_i - a^* \right)^2 \right].$$

Observe that if the  $w_i$ 's satisfy (5.9), then

$$\begin{aligned} \mathbb{E} \left[ \left( \sum_{i=1}^n w_i Y_i - a^* \right)^2 \right] &= \text{var} \left( \sum_{i=1}^n w_i Y_i \right) \\ &= \sum_{i=1}^n w_i^2 \text{var}(Y)_i \\ &= \sigma_*^2 \sum_{i=1}^n w_i^2. \end{aligned}$$

Hence,  $(w_1^*, w_2^*, \dots, w_n^*)$  is the minimizer of

$$\begin{aligned} \min_{w_1, \dots, w_n} \quad & \sum_{i=1}^n w_i^2 \\ \text{s.t.} \quad & \sum_{i=1}^n w_i x_i = 1 \\ & \sum_{i=1}^n w_i = 0. \end{aligned}$$

If  $\lambda_1$  and  $\lambda_2$  are the Lagrange multipliers corresponding to the equality constraints (5.9), then  $(w_1^*, \dots, w_n^*)$  must satisfy

$$2w_i^* - \lambda_1 x_i - \lambda_2 = 0 \quad (5.10)$$

for  $1 \leq i \leq n$ . Summing (5.10) over  $i$  and using the second equality in (5.9) yields

$$\lambda_1 \sum_{i=1}^n x_i + \lambda_2 n = 0. \quad (5.11)$$

Similarly, multiplying through (5.10) by  $x_i$  and summing over  $i$  leads to

$$\lambda_1 \sum_{i=1}^n x_i^2 + \lambda_2 \sum_{i=1}^n x_i = 2. \quad (5.12)$$

Assuming that  $\sum_{i=1}^n x_i^2 - n\bar{x}_n^2$  (with  $\bar{x}_n = n^{-1}(x_1 + \dots + x_n)$ ) is non-zero, the unique solution of (5.11) and (5.12) is

$$\lambda_1 = \frac{2}{\sum_{i=1}^n x_i^2 - n\bar{x}_n^2} \quad \text{and} \quad \lambda_2 = -\frac{2\bar{x}_n}{\sum_{i=1}^n x_i^2 - n\bar{x}_n^2}.$$

From (5.10) we conclude that the BLUE  $\hat{a}_n$  for  $a^*$  is

$$\hat{a}_n = \frac{\sum_{i=1}^n Y_i x_i - n \bar{Y}_n \bar{x}_n}{\sum_{i=1}^n x_i^2 - n \bar{x}_n^2},$$

which coincides with the least squares estimator for  $a^*$  obtained in Section 5.1. Note that the BLUE theory just presented is an alternative statistical justification for the method of least squares that requires no normality assumption.

Given the non-Gaussian residual error model hypothesized in this section, the confidence interval procedures of Section 5.2 are no longer valid. However, the bootstrap can be applied here to produce asymptotically valid confidence interval procedures for  $a^*$  and  $b^*$ . As usual, the idea is to generate bootstrap samples from a linear regression model that is as similar as possible to the original model (5.8). Note that under mild assumptions on the design points ( $x_i : i \geq 1$ ) and residual error distribution,  $\hat{a}_n \rightarrow a^*$  and  $\hat{b}_n \rightarrow b^*$  as  $n \rightarrow \infty$ ; see Exercise Exercise 5.9. It follows that the estimated residuals  $\hat{\epsilon}_i \triangleq Y_i - \hat{a}_n x_i - \hat{b}_n \stackrel{\mathcal{D}}{=} \epsilon_i$  for  $1 \leq i \leq n$ , and

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}\{\hat{\epsilon}_i \leq x\} \stackrel{\mathcal{D}}{=} F(x) = \mathbf{P}\{\epsilon_i \leq x\}.$$

Thus, if  $\hat{\epsilon}_1^*, \dots, \hat{\epsilon}_n^*$  is an iid sample of size  $n$  from  $F_n$ , the regression model

$$\tilde{Y}_i = \hat{a}_n x_i + \hat{b}_n + \hat{\epsilon}_i^*$$

should behave similarly to (5.8). In particular, note that  $\mathbf{E}[\hat{\epsilon}_1^*] = 0$ . This suggests the following bootstrap procedure for computing an approximate  $100(1 - \delta)\%$  confidence interval for  $a^*$ ; an analogous procedure produces confidence intervals for  $b^*$ .

#### *Confidence Interval Bootstrap Procedure for $a^*$*

1. Generate a bootstrap sample  $(\hat{\epsilon}_i^* : 1 \leq i \leq n)$  of iid observations from  $F_n$ .
2. Set  $\tilde{Y}_i = \hat{a}_n x_i + \hat{b}_n + \hat{\epsilon}_i^*$  for  $1 \leq i \leq n$ .
3. Compute the least squares estimator  $\hat{a}_{1n}^*$  for the slope, based on the data set  $\{(\tilde{Y}_i, x_i) : 1 \leq i \leq n\}$ .
4. Repeat the first three steps  $m$  iid times, thereby generating  $\hat{a}_{1n}^*, \dots, \hat{a}_{mn}^*$ .
5. Find  $z_1$  and  $z_2$  such that

$$\frac{1}{m} \sum_{i=1}^m \mathbf{I}\{\hat{a}_{in}^* - \hat{a}_n > z_2\} \approx \frac{\delta}{2}$$

and

$$\frac{1}{m} \sum_{i=1}^m \mathbf{I}\{\hat{a}_{in}^* - \hat{a}_n < z_1\} \approx \frac{\delta}{2},$$

and output  $[\hat{a}_n - z_2, \hat{a}_n - z_1]$  as an  $100(1 - \delta)\%$  confidence interval for  $a^*$ .

**Exercise 5.8:** Find the BLUE for  $b^*$ , assuming the model (5.8).

**Exercise 5.9:** Suppose that the design  $(x_i : i \geq 1)$  satisfies the assumptions of Exercise 5.5 and that the  $\epsilon_i$ 's are bounded rv's. Prove that  $\hat{a}_n \xrightarrow{p} a^*$  and  $\hat{b}_n \xrightarrow{p} b^*$  as  $n \rightarrow \infty$ .

**Exercise 5.10:** Explain how you would use the bootstrap to test the hypothesis  $H_0 : a^* = 0$  versus  $H_1 : a^* \neq 0$ .

**Exercise 5.11:** Explain how you would use the bootstrap to construct a prediction interval for  $Y|x$ , assuming the model (5.8).

## 5.6 Regression Models with Randomness in the Independent Variable

In our above discussion of linear regression we have modeled the design points  $(x_i : i \geq 1)$  as deterministic, reflecting the fact that in a typical experiment environment, the experimenter can carefully choose the  $x$ -values at which to collect observations.

But in other settings in which linear regression ideas are used, one may be interested in studying linear relationships between two quantities in which one has no experimental control over either variable. For examples, suppose that one wishes to study the possibility of a linear relationship between the number of hours of exercise and income. The data is gathered by randomly polling individuals and collecting both pieces of information from them. In this context, an appropriate model is a set of pairs  $\{(X_i, Y_i) : 1 \leq i \leq n\}$ , where  $n$  is the number of individuals polled. Suppose that we assume that the  $(X_i, Y_i)$ 's are iid and bivariate Gaussian with

$$Y_i = a^* X_i + b^* + \epsilon_i, \quad (5.13)$$

where  $\epsilon_i$  is a normal rv, independent of  $X_i$ , with  $E[\epsilon_i] = 0$  and  $\text{var}(\epsilon_i) = \sigma_*^2$ . The model is an obvious generalization of the simple linear model of Section 5.2, with the principal modification being that the  $x$ -values are no longer modeled as deterministic, but are now themselves rv's. In this setting, note that the likelihood of the sample depends not only on  $a, b$  and  $\sigma^2$  but also on the mean  $\mu_X$  and  $\text{var}(\sigma_X^2)$  of the rv  $X_i$ :

$$\begin{aligned} L_n(a, b, \sigma^2, \mu_X, \sigma_X^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(X_i - \mu_X)^2}{2\sigma_X^2}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - aX_i - b)^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi)^n} \frac{1}{(\sigma_X^2 \sigma^2)^{\frac{n}{2}}} \exp\left(-\sum_{i=1}^n \frac{(X_i - \mu_X)^2}{2\sigma_X^2} - \sum_{i=1}^n \frac{(Y_i - aX_i - b)^2}{2\sigma^2}\right) \end{aligned}$$

The MLE's for  $a^*$ ,  $b^*$ , and  $\sigma_*^2$  are then given by:

$$\hat{a}_n = \frac{\frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X}_n \bar{Y}_n}{\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2}, \quad \hat{b}_n = \bar{Y}_n - \hat{a}_n \bar{X}_n \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{a}_n X_i - \hat{b}_n)^2,$$

where  $\bar{X}_n = n^{-1}(X_1 + \dots + X_n)$  and  $\bar{Y}_n = n^{-1}(Y_1 + \dots + Y_n)$ . In other words, the MLE's for  $a^*$ ,  $b^*$  and  $\sigma_*^2$  coincide with those derived earlier, and merely involve directly substituting the  $X_i$ 's in the place of the  $x_i$ 's.

We can further study a non-parametric version of (5.13). Specifically, suppose that  $(X_1, Y_1), \dots, (X_n, Y_n)$  is an iid sample from a population  $(X, Y)$ , in which

$$Y_i = a^* X_i + b^* + \epsilon_i, \quad (5.14)$$

where  $\epsilon_i$  is uncorrelated with  $X_i$  and  $E[\epsilon_i] = 0$  and  $\text{var}(\epsilon_i) = \sigma_*^2$ . Note that

$$a^* = \text{cov}(X_i, Y_i) / \text{var}(X_i) \quad \text{and} \quad b^* = E[Y_i] - a^* E[X_i]. \quad (5.15)$$

In this context,  $\hat{Y}_i = a^* X_i + b^*$  is the best linear predictor of  $Y_i$ , based on  $X_i$ . Hence, the estimator of  $a^*$  and  $b^*$  from the sample  $\{(X_i, Y_i) : 1 \leq i \leq n\}$  is precisely what is required in this setting to determine the coefficients corresponding to the best linear predictor.

Given the expressions (5.15) for  $a^*$  and  $b^*$ , estimators based on use of the sample covariances and variances seem particularly natural:

$$\hat{a}_n = \frac{\frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X}_n \bar{Y}_n}{\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2} \quad \text{and} \quad \hat{b}_n = \bar{Y}_n - \hat{a}_n \bar{X}_n.$$

To produce an approximate  $100(1 - \delta)\%$  confidence interval for  $a^*$ , we can utilize a version of the bootstrap appropriate to this setting (which recognizes that the iid structure in (5.14) comes from the  $(X_i, Y_i)$ 's, rather than the  $\epsilon_i$ 's (as in the bootstrap of Section 5.5)); an analogous procedure produces confidence intervals for  $b^*$ .

*Confidence Interval Bootstrap for  $a^*$*

1. Generate a bootstrap sample  $\{(X_i^*, Y_i^*) : 1 \leq i \leq n\}$  by sampling each of the observations  $(X_i, Y_i)$  ( $1 \leq i \leq n$ ) with probability  $n^{-1}$   $n$  iid times.

2. Compute

$$\hat{a}_n^* = \frac{\frac{1}{n} \sum_{i=1}^n X_i^* Y_i^* - \frac{1}{n} \sum_{i=1}^n X_i^* \cdot \frac{1}{n} \sum_{i=1}^n Y_i^*}{\frac{1}{n} \sum_{i=1}^n X_i^{*2} - \left(\frac{1}{n} \sum_{i=1}^n X_i^*\right)^2}.$$

3. Repeat the first two steps  $m$  iid times, thereby generating  $\hat{a}_{1n}^*, \dots, \hat{a}_{mn}^*$  (i.e.  $m$  iid copies of  $\hat{a}_n^*$ ).

4. Find  $z_1$  and  $z_2$  such that

$$\frac{1}{m} \sum_{i=1}^m \mathbf{I}\{\hat{a}_{in}^* - \hat{a}_n > z_2\} \approx \frac{\delta}{2}$$

and

$$\frac{1}{m} \sum_{i=1}^m \mathbf{I}\{\hat{a}_{in}^* - \hat{a}_n < z_1\} \approx \frac{\delta}{2},$$

and output  $[\hat{a}_n - z_2, \hat{a}_n - z_1]$  as an approximate  $100(1 - \delta)\%$  confidence interval for  $a^*$ .

**Exercise 5.12:** Suppose that we collect  $n$  iid samples  $\{(X_i, Y_i) : 1 \leq i \leq n\}$ , and wish to use this data set to construct a linear predictor for  $Y_{n+1}$  based on the observation  $X_{n+1}$ .

1. What would you use as your predictor  $\hat{Y}_{n+1}$ ?
2. What would your estimate of the mean squared prediction error be? (Justify your estimator on the basis of its behavior as  $n \rightarrow \infty$ .)
3. Suppose that we strengthen the assumption in (5.14) to require that  $\epsilon_i$  is independent of  $X_i$ . Show that  $\hat{Y}_i = a^* X_i + b^*$  is the best non-linear predictor.
4. (continuation of above) Provide a bootstrap algorithm for computing an approximate  $100(1 - \delta)\%$  prediction interval for  $Y_{n+1}$ , based on  $X_{n+1}$  (and the previously observed  $(X_i, Y_i)$ 's).

**Exercise 5.13:** As “error in variables” linear regression model presumes that

$$\eta_i = a^* \xi_i + b^*, \quad 1 \leq i \leq n,$$

where  $(\eta_1, \xi_1), \dots, (\eta_n, \xi_n)$  are assumed iid (but unobserved). The observed quantities  $(X_1, Y_1), \dots, (X_n, Y_n)$  are related to the  $(\eta_i, \xi_i)$ 's via

$$Y_i = \eta_i + \epsilon_i \quad \text{and} \quad X_i = \xi_i + \delta_i,$$

where  $\epsilon_i$  and  $\delta_i$  are viewed as “observation errors”. It is assumed that  $E[\epsilon_i] = 0 = E[\delta_i]$  and that  $(\epsilon_i, \delta_i)$ 's are iid and independent of the  $(\eta_i, \xi_i)$ 's. Discuss the problem of parameter estimation for this model when the  $(\eta_i, \xi_i)$ 's and  $(\epsilon_i, \delta_i)$ 's are Gaussian.

## 5.7 Multiple Linear Regressions

As discussed in Section 5.1, many applications require regression the dependent variable  $Y$  on a collection of explanatory variables  $x$ . This leads naturally to the consideration of the statistical model

$$Y_i = x_i^T a^* + \epsilon_i, \quad 1 \leq i \leq n \quad (5.16)$$

where the  $\epsilon_i$ 's are iid normal rv's with mean zero and variance  $\sigma_*^2$  and the  $x_i$ 's are vector-valued. In this setting, the likelihood is given by

$$L_n(x, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2} \exp\left(-\frac{(Y_i - x_i^T a)^2}{2\sigma^2}\right)},$$

so that the MLE  $\hat{a}_n$  for  $a^*$  is the minimizer of

$$\min_a \sum_{i=1}^n (Y_i - x_i^T a)^2 = \min_a (Y - Xa)^T (Y - Xa)$$

where

$$Y = (Y_1, \dots, Y_n)^T$$

and

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \dots \\ x_n^T \end{pmatrix}.$$

It follows that if the design points ( $x_i : i \geq 1$ ) are selected so that  $X^T X$  is non-singular then

$$\hat{a}_n = (X^T X)^{-1} X^T Y \quad (5.17)$$

so that  $\hat{a}_n$  coincides with the least squares estimate of Section 5.1. As for the simple linear regression model, we can use the Gaussian structure of the model to construct *exact* confidence regions for  $a^*$  and  $\sigma_*^2$ , prediction intervals and hypothesis testing.

The estimator (5.17) can also be statistically justified in the presence of the more general model

$$Y_i = x_i^T a^* + \epsilon_i, \quad 1 \leq i \leq n, \quad (5.18)$$

in which the  $\epsilon_i$ 's are iid with  $E[\epsilon_i] = 0$  and  $\text{var}(\epsilon_i) = \sigma_*^2 < \infty$ . The model (5.18) can be re-written as

$$Y = Xa^* + \epsilon,$$

where  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$  has a covariance matrix equal to  $\sigma_*^2$  times the identity matrix. A linear estimator for  $a^*$  is one that can be expressed as  $BY$  for some (deterministic) choice of  $B$ . Unbiasedness of the linear estimator for all choices of  $a^*$  requires that

$$BX = I \quad (5.19)$$

(Note that  $B$  and  $X$  are not square matrices, so (5.19) does not assert that  $B = X^{-1}$ .) It turns out that setting

$$B = (X^T X)^{-1} X^T$$

is the choice that minimizes

$$E \left[ (\lambda^T BY - \lambda^T a^*)^2 \right]$$

over all (deterministic) vectors  $\lambda$ , so that (5.17) is the BLUE for the model (5.18). Construction of approximate confidence regions for  $a^*$  for the non-Gaussian formulation (5.18) can be implemented via the

bootstrap, as discussed in Section 5.5 for the simple regression model.

A useful generalization of the model (5.18) is to postulate that

$$Y = Xa^* + \epsilon, \quad (5.20)$$

where  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$  has a covariance matrix of the form  $\sigma_*^2 C$ , where  $C$  is a known covariance matrix and (as usual)  $\sigma_*^2$  is an unknown statistical parameter. Note that the model (5.20) includes, as a special class, that of “weighted least squares” (in which case  $C$  is take to be a diagonal matrix). The model (5.20) can be easily reduced to the model (5.18). Assume that  $C$  is positive definite, write  $C$  as  $C = LL^T$  (so that  $L$  is a lower Cholesky factor of  $C$ ). Pre-multiplying both sides of (5.20) by  $L^{-1}$  yield

$$\tilde{Y} = \tilde{X}a^* + \tilde{\epsilon},$$

where  $\tilde{Y} = L^{-1}Y$ ,  $\tilde{X} = L^{-1}X$  and  $\tilde{\epsilon} = L^{-1}\epsilon$ . Since the covariance matrix of  $\tilde{\epsilon}$  is precisely of the form required by the model (5.18), it follows that

$$\hat{a}_n = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}$$

is the BLUE for  $a^*$ . But this is identical to

$$\hat{a}_n = (X^T C^{-1} X)^{-1} X^T C^{-1} Y,$$

yielding a complete solution to the problem of computing the BLUE for the more general model (5.20).

**Exercise 5.14:** Prove that setting  $B = (X^T X)^{-1} X^T$  yields the BLUE for the model (5.18).

**Exercise 5.15:** The best mean-square predictor for  $Y$ , based on  $X$ , is  $E[Y|X]$ . Suppose that  $E[Y^2] < \infty$  and  $X$  is bounded. Assume that one observes  $n$  iid observations  $(X_1, Y_1), \dots, (X_n, Y_n)$  from the distribution of  $(X, Y)$ , and finds the solution  $\hat{a}_1, \dots, \hat{a}_p$  of the least squares problem

$$\min_{a_1, \dots, a_p} \sum_{i=1}^n (Y_i - \sum_{j=1}^p a_j X_i^j)^2.$$

When  $n$  and  $p$  are large, what can be said about the connection of the least squares problem to the best mean-square predictor?

**Exercise 5.16:**

1. Prove that

$$\hat{\sigma}_n^2 = \frac{1}{n-d} \sum_{i=1}^n (Y_i - x_i^T \hat{a}_n)^2$$

is an unbiased estimator for  $\sigma_g^2$  for the model (5.16), where  $d$  is the dimension of  $a^*$ .

2. What is an unbiased estimator for  $\sigma_*^2$  for the model (5.20)?

## 5.8 Bayesian Linear Regression

Suppose that we have taken measurements  $Y_1, \dots, Y_n$  at design points  $x_1, \dots, x_n \in \mathbb{R}^d$ , and model the data as

$$Y_i = x_i^T a + \epsilon_i, \quad (5.21)$$

where the  $\epsilon_i$ 's are iid  $N(0, \sigma^2)$  rv's. In many settings, we may have prior information on  $a$  and / or  $\sigma^2$  that we wish to incorporate into the model. Such prior information may, for example, come from previously

conducted similar observational studies.

We start with the simplest Bayesian formulation, in which  $\sigma^2$  is assumed to be a known quantity, and a prior is imposed only on the regression coefficients  $a$ . Specifically, we impose a conjugate prior on  $a$  so that  $a$  is Gaussian with mean  $\tilde{a}$  and positive definite covariance matrix  $\Lambda$ , so that the prior density is given by

$$p(a) = \frac{1}{\sqrt{(2\pi)^d |\det \Lambda|}} \exp\left(-\frac{(a - \tilde{a})^T \Lambda^{-1} (a - \tilde{a})}{2}\right).$$

Note that (5.21) asserts that the density of the observations  $Y = (Y_1, \dots, Y_n)^T$ , given  $a$ , is

$$f(Y_1, \dots, Y_n | a) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{(Y - Xa)^T (Y - Xa)}{2\sigma^2}\right).$$

Hence, the posterior density of  $a$ , given  $Y$  is proportional to

$$\exp\left(-\frac{(Y - Xa)^T (Y - Xa)}{2\sigma^2} - \frac{(a - \tilde{a})^T \Lambda^{-1} (a - \tilde{a})}{2}\right). \quad (5.22)$$

Let  $U$  be a square root of  $\Lambda^{-1}$ , so that  $\Lambda^{-1} = U^T U$ . Then, the exponent of (5.22) can be re-written as

$$\frac{1}{2} \left( \left( \frac{Y}{\sigma} - \frac{Xa}{\sigma} \right)^T, (U\tilde{a} - Ua)^T \right) \begin{pmatrix} \frac{Y}{\sigma} - \frac{Xa}{\sigma} \\ U\tilde{a} - Ua \end{pmatrix} = \frac{1}{2} (Z - Wa)^T (Z - Wa) \quad (5.23)$$

where

$$Z = \begin{pmatrix} Y/\sigma \\ U\tilde{a} \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} X/\sigma \\ U \end{pmatrix}.$$

The minimizer of the least squares problem (5.23) is then

$$\hat{a} = (W^T W)^{-1} W^T z = \left( \frac{1}{\sigma^2} X^T X + \Lambda^{-1} \right)^{-1} \left( \frac{1}{\sigma^2} X^T Y + \Lambda^{-1} \tilde{a} \right).$$

Furthermore, we can write the quadratic for (5.23) as

$$-\frac{1}{2} ((Z - W\hat{a}) + W(\hat{a} - a))^T ((Z - W\hat{a}) + W(\hat{a} - a)) = \frac{1}{2} (Z - W\hat{a})^T (Z - W\hat{a}) + \frac{1}{2} (\hat{a} - a)^T W^T W (\hat{a} - a), \quad (5.24)$$

by using the fact that  $W^T W \hat{a} = W^T Z$ . It follows that the posterior density of  $a$  is proportional to

$$\exp\left(-\frac{1}{2} (a - \hat{a})^T W^T W (a - \hat{a})\right),$$

so that the posterior must be the Gaussian density

$$\frac{1}{(2\pi)^{\frac{d}{2}} |\det(X^T X/\sigma^2 + \Lambda^{-1})|^{-1}} \exp\left(-\frac{1}{2} (a - \hat{a})^T \left( \frac{X^T X}{\sigma^2} + \Lambda^{-1} \right) (a - \hat{a})\right). \quad (5.25)$$

In other words, the posterior distribution for  $a$  is normal with mean  $\hat{a}$  and covariance matrix given by  $(X^T X/\sigma^2 + \Lambda^{-1})^{-1}$ .

We can now use the posterior distribution to easily compute a credible set for  $a$  (i.e. the Bayesian analog to a confidence region), and to produce predictive intervals. For the latter, note that if we are asked to predict the value of  $Y|x$  for a given  $x$ , then (5.21) asserts that conditional on  $a$

$$Y|x \stackrel{\mathcal{D}}{=} N(x^T a, \sigma^2).$$

Given the observations  $Y_1, \dots, Y_n$ ,  $a$  has the posterior distribution (5.25). Hence, the conditional prediction distribution of  $Y|x$ , given  $Y_1, \dots, Y_n$  is normal with mean

$$x^T(X^T X/\sigma^2 + \Lambda^{-1})^{-1}(X^T Y/\sigma^2 + \Lambda^{-1}\hat{a})$$

and variance

$$x^T(X^T X/\sigma^2 + \Lambda^{-1})^{-1}x + \sigma^2.$$

This Bayesian formulation is connected to the idea of *Tikhonov regularization*. When the mean of the prior  $\tilde{a}$  is zero, then

$$\hat{a} = (X^T X/\sigma^2 + \Lambda^{-1})X^T Y/\sigma^2. \quad (5.26)$$

Furthermore,  $\hat{a}$  is the minimizer of the least squares problem

$$\min_a \frac{1}{\sigma^2}(Y - Xa)^T(Y - Xa) + a^T \Lambda^{-1}a. \quad (5.27)$$

In view of (5.27), we may therefore view  $\hat{a}$  either as a particular Bayes estimator or as a solution to a least squares problem in which an additional quadratic form  $a^T \Lambda^{-1}a$  is added to the objective function, so as to penalize solutions  $a$  with a large  $\Lambda^{-1}$  norm. The estimator  $\hat{a}$  given by (5.26) is sometimes call the *ridge regression* estimator.

We conclude this section with a brief discussion of the posterior distribution on  $a$  and  $\sigma^2$  when both  $a$  and  $\sigma^2$  are given their conjugate priors. We postulate that the prior on  $\sigma^2$  is given by an inverse gamma distribution, so that the prior density on  $\sigma^2$  is

$$p(\sigma^2) = \frac{\beta^\alpha \sigma^{-2\alpha-2} e^{-\frac{\beta}{\sigma^2}}}{\Gamma(\alpha)}$$

for  $\beta > 0$  and  $\alpha > 2$ . The mean  $E[\sigma^2]$  for this prior is  $\beta(\alpha - 1)^{-1}$  and the variance is  $\beta^2(\alpha - 1)^{-2}(\alpha - 2)^{-1}$ . Conditional on  $\sigma^2$ , the prior on  $a$  is then given by

$$p(a|\sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^d |\det \Lambda|}} \exp\left(-\frac{(a - \tilde{a})\Lambda^{-1}(a - \tilde{a})}{2\sigma^2}\right)$$

(so that  $a$  is conditionally Gaussian with mean  $\tilde{a}$  and covariance  $\sigma^2\Lambda$ ). In this case, the joint posterior density on  $a$  and  $\sigma^2$ , conditional on  $Y = (Y_1, \dots, Y_n)^T$ , is proportional to

$$\sigma^{-2\alpha-2-d-n} \exp\left(-\frac{\beta}{\sigma^2} - \frac{1}{2\sigma^2}(Y - Xa)^T(Y - Xa) - \frac{1}{2\sigma^2}(a - \tilde{a})^T \Lambda^{-1}(a - \tilde{a})\right).$$

In view of (5.24), the joint posterior density on  $a$  and  $\sigma^2$  factorizes in to the product of an inverse gamma density and a Gaussian density, namely

$$\frac{\hat{\beta}^{\hat{\alpha}} \sigma^{-2\hat{\alpha}-2} \exp(-\hat{\beta}/\sigma^2)}{\Gamma(\hat{\alpha})} \frac{1}{\sqrt{(2\pi\sigma^2)^d |\det(X^T X + \Lambda^{-1})|^{-1}}} \exp\left(-\frac{(a - \hat{a})^T(X^T X + \Lambda^{-1})(a - \hat{a})}{2\sigma^2}\right)$$

where

$$\hat{\beta} = \beta + \frac{(Z - W\hat{a})^T(Z - W\hat{a})}{2}$$

and

$$\hat{\alpha} = \alpha + \frac{n}{2}.$$

This posterior distribution can now be utilized as in the previous setting in which only a prior on  $a$  was specified.

## 5.9 $L^1$ Regression and the Lasso

Much of our preceding discussion of linear regression has presumed that the residual errors are Gaussian distributed. But a quite different theory arises when one makes alternative parametric assumptions about the error distribution. In particular, suppose that we assume that

$$Y_i = x_i^T a^* + \epsilon_i, \quad 1 \leq i \leq n, \quad (5.28)$$

where  $\epsilon_1, \dots, \epsilon_n$  is a sequence of iid rv's having Laplace distribution with common density

$$f(x) = \frac{1}{2} \lambda^* e^{-\lambda^* |x|}$$

for  $\lambda^* > 0$ . The unknown parameters in the model are  $a^*$  and  $\lambda^*$ .

For the model (Exercise 5.17), the likelihood function is

$$\begin{aligned} L_n(a, \lambda) &= \prod_{i=1}^n \left( \frac{\lambda}{2} \right) e^{-\lambda |Y_i - x_i^T a|} \\ &= 2^{-n} \lambda^n e^{-\lambda \sum_{i=1}^n |Y_i - x_i^T a|}. \end{aligned}$$

The MLE estimator  $\hat{a}_n$  for  $a^*$  is then the minimizer of the optimization problem

$$\min_a \sum_{i=1}^n |Y_i - x_i^T a|, \quad (5.29)$$

and the MLE for  $\lambda^*$  is

$$\hat{\lambda}_n = \frac{n}{\sum_{i=1}^n |Y_i - x_i^T \hat{a}_n|}.$$

Since  $\hat{a}_n$  is obtained by minimizing the sum of the absolute values of the regression errors,  $\hat{a}_n$  is often called the  $L^1$  regression estimator for  $a^*$ . Note that (5.29) can be formulated as a linear program, namely

$$\min_{a, w_1^+, \dots, w_n^+, w_1^-, \dots, w_n^-} \sum_{i=1}^n (w_i^+ w_i^-)$$

subject to

$$\begin{aligned} -w_i^- &\leq Y_i - x_i^T a \leq w_i^+, \quad 1 \leq i \leq n \\ w_1^+, \dots, w_n^+ &\geq 0 \\ w_1^-, \dots, w_n^- &\geq 0 \end{aligned}$$

and hence (5.29) can typically be solved efficiently. One important advantage of  $L^1$  regression solutions to the  $L^2$  regression) based on minimizing a sum of squares) discussed earlier is when the number of explanatory variables is large (i.e. the dimension of the  $x_i$ 's is large),  $L^1$  regression will often put non-zero coefficients on only some of the components of the  $x$ -vector (so that many components of  $\hat{a}_n$  will be zero). On the other hand,  $L^2$  regression almost always generates an estimator for  $a^*$  that is non-zero in all components, leading to a “non-sparse” representation of the data.

The so-called “Lasso” regression estimator similarly attempts to build a regression model that is reasonably sparse. It estimates  $a^*$  via the minimization of the problem

$$\min_a \sum_{i=1}^n (Y_i - x_i^T a)^2 \quad (5.30)$$

$$\text{s.t. } \sum_{j=1}^d |a_j| \leq s, \quad (5.31)$$

where  $a = (a_1, \dots, a_d)^T$ . The closely related Lagrangian version of this problem is to select the minimizer of the unconstrained problem

$$\min_a \sum_{i=1}^n (Y_i - x_i^T a)^2 + \gamma \sum_{j=1}^d |a_j| \quad (5.32)$$

as an estimator for  $a^*$ . Exercise Exercise 5.18 establishes that (5.32) arises when considering the regression model (5.21) with a Laplace prior on the  $a_i$ 's.

**Exercise 5.17:** Discuss how you would construct a confidence region for  $a^*$  for the model (Exercise 5.17).

**Exercise 5.18:** Suppose that conditional on  $a$ , the  $Y_i$ 's are governed by the model (5.21). Assume that the prior on the coefficients  $a_1, \dots, a_d$  takes the form

$$p(a_1, \dots, a_d) = \prod_{i=1}^d \frac{\lambda}{2} \lambda e^{-\lambda |a_i|}.$$

Prove that if you choose to estimate  $a$  via the posterior mode, we are led to a problem of the form (5.32).

**Exercise 5.19:** Discuss how you would construct a predictive interval for  $Y|x$ , in the presence of the model described in Exercise 5.29.

## 5.10 Logistic Regression

Suppose that we are asked to develop a marketing strategy, based on data  $(Y_1, x_1), \dots, (Y_n, x_n)$ , in which  $Y_i$  is 1 or 0 depending on whether or not the  $i$ 'th individual in a marketing study bought the product, and  $x_i$  represents a vector describing various socio-economic attributes of the individual. In creating such a strategy, it may be useful to build a statistical model to help correlate the marketing outcome with the available socio-economic predictors. Because the  $Y_i$ 's are binary rv's, standard regression methods are no longer applicable. *Logistic regression* is a statistical tool that is intended to deal with such settings.

The model that underlies logistic regression is to presume that for  $1 \leq i \leq n$

$$P(Y_i = 1) = \frac{e^{x_i^T a^*}}{1 + e^{x_i^T a^*}}$$

for some  $a^*$ . As usual, we assume that the  $Y_i$ 's are independent. The resulting likelihood function is then

$$L_n(a) = \prod_{i=1}^n \frac{(e^{x_i^T a})^{Y_i}}{1 + e^{x_i^T a}},$$

so that the MLE  $\hat{a}_n$  is the maximizer of

$$\max_a \sum_{i=1}^n \left( x_i^T a Y_i - \log(1 + e^{x_i^T a}) \right).$$

This non-linear optimization problem is then solved numerically to compute  $\hat{a}_n$ .

**Exercise 5.20:** Discuss how you would use the bootstrap to construct confidence intervals for coefficients of  $a^*$ .

References:

Dougherty, F.R. *Probability and Statistics in the Engineering, Computing and Physical Sciences*. Prentice Hall, Englewood Cliffs, New Jersey, 1990.

Rao, C.R. *Linear Statistical Inference and its Application*. John Wiley, New York, 1973.