Chapter 10

Linear Regression

10.1 Introduction

The problem of predicting a real-valued Y from a covariate $X \in \mathbb{R}^d$ is called *regression*. The term "regression" was originally coined by Francis Galton in the nineteenth century, in an attempt to find a mathematical law for some of the phenomena of heredity. In particular, it was observed that the heights of descendants of tall ancestors tend to regress downwards towards a normal average.

Let us recall of few of the basics; see Section **??** for more details. If we predict Y with $g(X)$, then the *prediction risk* with respect to the squared error loss is defined as

$$R(g) = \mathbb{E}(Y - g(X))^2. \tag{10.1}$$

The prediction risk is minimized by setting g equal to the *regression function* $m(x) = \mathbb{E}(Y \mid X = x)$ (see Exercise 1). Thus, $R(m) \leq R(g)$ for all g. Let $\sigma^2 = \mathbb{E}(Y - m(X))^2$ and let P be the distribution of X. $R(g)$ can then be written as

$$R(g) = \mathbb{E}\left(Y - m(X) + m(X) - g(X)\right)^2 = \int \left(g(x) - m(x)\right)^2 dP(x) + \sigma^2 \tag{10.2}$$

where the expectation is over the random test point (X, Y) and we have used

$$\begin{aligned}
\mathbb{E}\left\{\left(Y - m(X)\right)\left(m(X) - g(X)\right)\right\} &= \mathbb{E}\Big\{\mathbb{E}\left\{\left(Y - m(X)\right)\left(m(X) - g(X)\right) \mid X\right\}\Big\} && (10.3)\\
&= \mathbb{E}\Big\{\left(m(X) - g(X)\right)\mathbb{E}\left\{\left(Y - m(X)\right) \mid X\right\}\Big\} && (10.4)\\
&= \mathbb{E}\Big\{\left(m(X) - g(X)\right)\left(m(X) - m(X)\right)\Big\} && (10.5)\\
&= 0. && (10.6)
\end{aligned}$$

The second term σ^2 in the right hand side of Equation (10.2) is the unavoidable error due to the fact that we are predicting a random quantity.

Suppose now that we have n pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i = (x_{i1}, \ldots, x_{id})^T$ denotes the d-dimensional covariate vector for the i^{th} observation. Let

$\mathbf{y} = (y_1, \ldots, y_n)^T$ and define the $n \times p$ *design matrix*

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1d} \\ x_{21} & x_{22} & \ldots & x_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nd} \end{pmatrix}. \tag{10.7}$$

Each row of $\mathbf{X}$ is one observation; each column corresponds to one of the d covariates.

We denote $\widehat{m}_n$ be an estimate of the regression function and let (X, Y) denote a new observation. Since $\widehat{m}_n$ is based on the observed data $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we define two types of prediction risks:

$$\text{(Conditional prediction risk)} : R(\widehat{m}_n) = \mathbb{E}\left[(Y - \widehat{m}_n(X))^2 \,\middle|\, \mathcal{D}_n\right] \tag{10.8}$$

$$\text{(Expected prediction risk)} : \mathcal{R}(\widehat{m}_n) = \mathbb{E}\big[R(\widehat{m}_n)\big]. \tag{10.9}$$

One thing to note is that the expectation in (10.8) is conditioned on the data $\mathcal{D}_n$, while the expectation in (10.9) is the average over all possible data sets. Therefore the conditional prediction risk is still a random quantity but the expected prediction risk is deterministic. A good regression estimate $\widehat{m}_n$ should try to minimize the expected prediction risk $\mathcal{R}\left(\widehat{m}_n\right)$. If both $\widehat{m}_n(x)$ and the regression function $m(x)$ are integrable, $\mathcal{R}\left(\widehat{m}_n\right)$ has the following *bias-variance decomposition*:

$$\mathcal{R}(\widehat{m}_n) = \int \mathbb{E}\left(\widehat{m}_n(x) - m(x)\right)^2 dP(x) + \sigma^2 = \int b_n^2(x) dP(x) + \int v_n(x) dP(x) + \sigma^2 \tag{10.10}$$

where $b_n(x) = \mathbb{E}(\widehat{m}_n(x)) - m(x)$ is the bias and $v_n(x) = \text{Var}(\widehat{m}_n(x))$ is the variance.

The *predicted values* or *fitted values* are defined to be $\widehat{\mathbf{y}} = (\widehat{y}_1, \ldots, \widehat{y}_n)^T$ where $\widehat{y}_i = \widehat{m}_n(x_i)$. The *residuals* are $r_i = y_i - \widehat{y}_i$. The *training error* is

$$\widetilde{R} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2. \tag{10.11}$$

The training error is a biased estimate of the expected prediction risk. To see this, let $\overline{m}(x) = \mathbb{E}\left(\widehat{m}_n(x)\right)$, where the expectation is over $\mathcal{D}_n$. We then compute

$$\begin{aligned} \mathbb{E}(y_i - \widehat{y}_i)^2 &= \mathbb{E}\left(y_i - m(x_i) + m(x_i) - \overline{m}(x_i) + \overline{m}(x_i) - \widehat{y}_i\right)^2 && (10.12) \\ &= \sigma^2 + \mathbb{E}\left(m(x_i) - \overline{m}(x_i)\right)^2 + \mathbb{E}\left(\widehat{m}_n(x_i) - \overline{m}(x_i)\right)^2 - 2\text{Cov}(\widehat{y}_i, y_i) && (10.13) \\ &= \sigma^2 + \int b_n^2(x) dP(x) + \int v_n(x) dP(x) - 2\text{Cov}(\widehat{y}_i, y_i). && (10.14) \end{aligned}$$

Hence, from (10.10),

$$\mathbb{E}(\widetilde{R}) = \mathcal{R}(\widehat{m}_n) - \frac{2}{n} \sum_{i=1}^{n} \text{Cov}(\widehat{y}_i, y_i). \tag{10.15}$$

Typically, $\text{Cov}(\widehat{y}_i, y_i) > 0$ and so $\widetilde{R}$ underestimates the expected prediction risk. We discuss better methods for estimating the risk in the next chapter.

Summary of Notation

true regression function	$m(x) = \mathbb{E}(Y \mid X = x)$
estimated regression function	$\widehat{m}_n(x)$
data	$\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
conditional prediction risk	$R(\widehat{m}_n) = \mathbb{E}\left[(Y - \widehat{m}_n(X))^2 \mid \mathcal{D}_n\right]$
expected prediction risk	$\mathcal{R}(\widehat{m}_n) = \mathbb{E}\big[R(\widehat{m}_n)\big]$
response vector	$\mathbf{y} = (y_1, \ldots, y_n)^T$
design matrix	$\mathbf{X}$
predicted values	$\widehat{\mathbf{y}} = (\widehat{y}_1, \ldots, \widehat{y}_n)^T$ where $\widehat{y}_i = \widehat{m}_n(x_i)$
residual	$r_i = y_i - \widehat{y}_i$
training error	$n^{-1} \sum_{i=1}^n (y_i - \widehat{y}_i)^2$

10.2 Linear Predictors and Least Squares

The best predictor of Y, among all functions of x, is $m(x) = \mathbb{E}(Y \mid X = x)$. In this chapter we restrict ourselves to linear predictors, that is, predictors of the form $m_\beta(x) = \beta_0 + \sum_j \beta_j x_j$. For convenience, we define $x_1 = 1$ so we can write $m_\beta(x) = \beta^T x$ where now β_1 represents the intercept. Thus, the first column of the design matrix $\mathbf{X}$ is $(1, 1, \ldots, 1)^T$. It is important to emphasize that we do not assume that the true regression function m is linear; rather, we seek the best predictor among the class of all linear predictors.

The best linear predictor, or the *linear oracle*, is $m_*(x) = \beta_*^T x$ where

$$\beta_* = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \mathbb{E}\left(Y - \beta^T X\right)^2 . \tag{10.16}$$

We call $R_* = \mathbb{E}(Y - \beta_*^T X)^2$ the *oracle risk* with respect to the class of linear predictors. Let $\boldsymbol{\Sigma} = \mathbb{E}(XX^T)$ and let $C = \mathbb{E}(YX)$. Note that $\boldsymbol{\Sigma}$ is a $d \times d$ matrix and C is a vector of length d.

10.17 Theorem. *If $\boldsymbol{\Sigma}$ is invertible then the best linear predictor is $m_*(x) = \beta_*^T x$ where $\beta_* = \boldsymbol{\Sigma}^{-1} C$.*

Exercise 2 asks you to prove this theorem. To approximate the best linear predictor, we use the least squares method.

The *least squares estimator* $\widehat{\beta}$ is the value of β that minimizes $\displaystyle\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta^T x_i)^2$.

10.18 Theorem. *Suppose that* $\mathbf{X}^T\mathbf{X}$ *is invertible. Then the least squares estimator* $\widehat{\beta}$ *exists, is unique and is given by*

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \widehat{\mathbf{\Sigma}}^{-1}\widehat{C} \tag{10.19}$$

where $\widehat{\mathbf{\Sigma}} = \mathbf{X}^T\mathbf{X}/n$ *and* $\widehat{C} = \mathbf{X}^T\mathbf{y}/n$.

Proof. See Exercise 3. □

Let $R(\beta) = \mathbb{E}\left(Y - \beta^T X\right)^2$, for large n we expect that $\widehat{\mathbf{\Sigma}} \approx \mathbf{\Sigma}$ and $\widehat{C} \approx C$. This suggests that $R(\widehat{\beta}) \approx R(\beta_*)$.

10.20 Theorem. *If* X *and* Y *have finite variances and* $\mathbf{\Sigma} = \mathbb{E}(XX^T)$ *is invertible, then*

$$R(\widehat{\beta}_n) - R(\beta_*) \xrightarrow{\mathrm{P}} 0 \tag{10.21}$$

as $n \to \infty$. *Here we use the notation* $\widehat{\beta}_n$ *to emphasize the fact that* $\widehat{\beta}$ *is a function of* $\mathcal{D}_n$.

Proof Idea. Since

$$\begin{aligned} R(\widehat{\beta}_n) &= \mathbb{E}\left[\left(Y - X^T\widehat{\beta}_n\right)^2 \,|\, \mathcal{D}_n\right] & (10.22)\\ &= \mathbb{E}\left[\left(Y - X^T\beta_* + X^T\beta_* - X^T\widehat{\beta}_n\right)^2 \,|\, \mathcal{D}_n\right], & (10.23)\end{aligned}$$

we have

$$R(\widehat{\beta}_n) - R(\beta_*) = \left(\widehat{\beta}_n - \beta_*\right)^T \mathbf{\Sigma}\left(\widehat{\beta}_n - \beta_*\right) + 2\left(\beta_* - \widehat{\beta}_n\right)^T (C - \mathbf{\Sigma}\beta_*)\,. \tag{10.24}$$

It's then sufficient to show the right hand side of Equation (10.24) converges to zero in probability as n goes to infinity. The proof is left to Exercise 4. □

According to Theorem 10.20, we get closer and closer to the oracle risk as the sample size increases. Hence, the least squares predictor is a consistent approximation to the linear oracle predictor. This theorem is no longer true if we let d grow with n; we discuss this point in the next chapter.

10.25 Example (HIV Drug Resistance). In this example we consider which amino acids in a virus predict drug resistance. The data consist of 496 viruses, where we can represent each virus as a string of amino acids of length 99. Thus, there are 99 covariates, corresponding to the identity of the amino acid at each location. Amino acids come in 21 varieties, and X_j refers to the identity of the amino acid at location j. The outcome Y is a measure of drug resistance to the drug *Indinavir*. The data are available at `http://hivdb.stanford.edu`. Some analyses of these data are given in Rhee et al. (2006) and Percival et al. (2009).

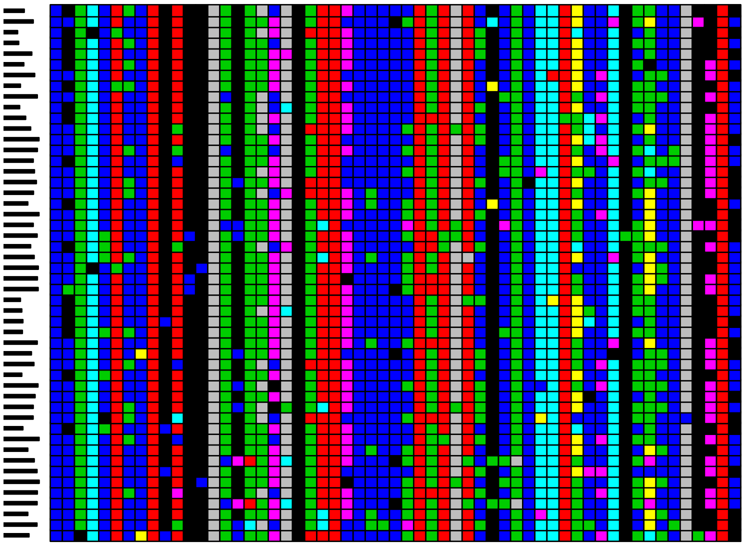

Figure 10.1. *The first 50 data points of the HIV drug resistance data. Each row is one data point, with different colors corresponding to different amino acids. Each column represents one position on the virus. The black sticks in the left column represent the the relative magnitudes of the outcome.*

We first eliminated all covariates which are constant, or nearly constant, leaving 57 covariates. Figure 10.1 shows the first 50 data points. Each row is one data point, and different colors correspond to different amino acids. Each column represents one position on the virus. The black sticks in the left column represent the relative magnitudes of the outcome Y. The goal is to predict Y from the X_js. The covariates are categorical since X_j can take 21 different values corresponding to the 21 different amino acids, but we simplified each X_j to be a binary variable so that $x_{ij} = 1$ if virus i has the most common amino acid at location j and $x_{ij} = 0$ otherwise. The training error after regressing Y on these 57 covariates is 0.55. The error if no covariates are included in the model is 1.57. It thus appears that we get improved prediction by regressing on the covariates, but keep in mind that the training error is a biased estimate of the expected prediction risk. The top left plot of Figure 10.2 shows the $\widehat{\beta}_j$s.

It is also possible to regress Y on each of the X_js separately. This is called marginal regression. The top right plot of Figure 10.2 shows the marginal regression coefficients which we call $\widehat{\alpha}_j$. There is some similarity between the $\widehat{\alpha}_j$s and the $\widehat{\beta}_j$s but there are also many differences. In general, it is possible to have $\widehat{\alpha}_j \approx 0$ while $\widehat{\beta}_j$ is large, and it is also possible to have $\widehat{\alpha}_j$ large while $\widehat{\beta}_j \approx 0$.

The bottom left plot of Figure 10.2 shows the residuals $y_i - \widehat{y}_i$ versus the fitted values $\widehat{y}_i$. The reason for looking at this plot is to see if there are suspicious patterns. One hopes to see points randomly scattered around 0. For example, a nonlinear trend in the residuals suggests that the nonparametric methods we discuss later might lead to better predictions. If the X_js are continuous, you should also plot the residuals versus each X_j. In our case, there

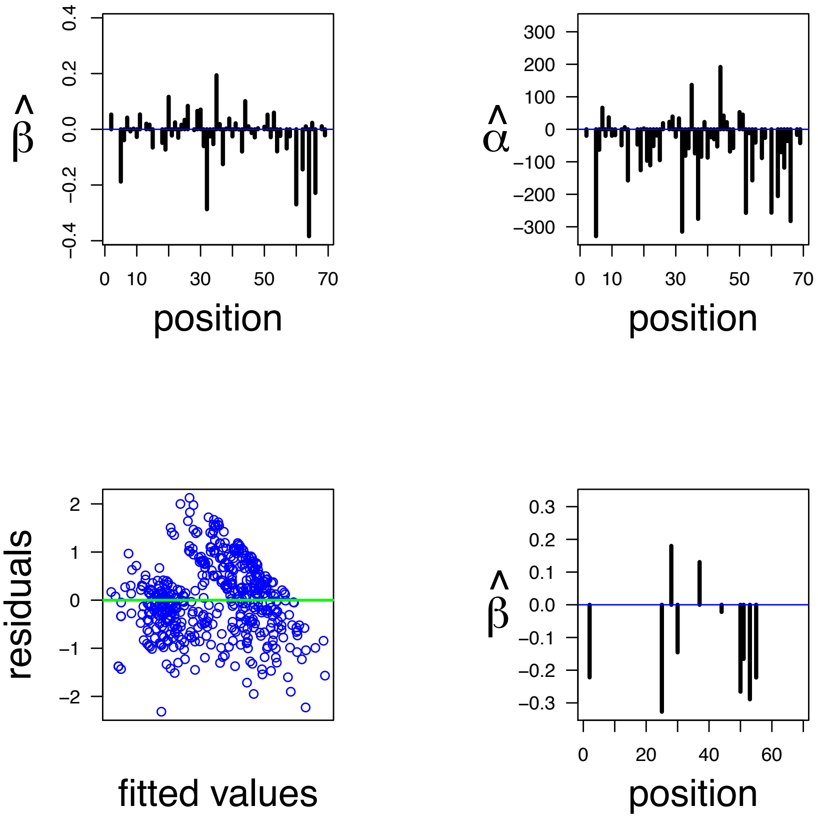

Figure 10.2. *The HIV drug resistance data. Top left: regression coefficients $\widehat{\beta}_j$. Top right: marginal regression coefficients $\widehat{\alpha}_j$. Bottom left: plot of residuals versus fitted values. Bottom right: $\widehat{\beta}_j$s in a smaller model using only 10 of the covariates.*

is some sort of pattern but this is probably an artifact of the discreteness of the covariates.

The bottom right plot of Figure 10.2 shows the $\widehat{\beta}_j$s when we we regress Y on only 10 of the covariates. In this case the training error is 0.62, which is only slightly larger than 0.55. Since the training error is biased, the smaller model might have a better prediction performance than the larger model. Can we find a small subset of the covariates that does have smaller expected predictive risk? This is a question we address in the next chapter. □

10.3 The Geometry of Least Squares

The *fitted values* (or *predicted values*) are defined to be $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta}$. From (10.19) it follows that $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{10.26}$$

is called the *hat matrix*. Define the *column space* $\mathcal{L}$ of $\mathbf{X}$ to be the set of vectors that can be obtained as linear combinations of the columns of $\mathbf{X}$.

The *hat matrix* $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ has the following properties:

1. $\mathbf{H}\mathbf{X} = \mathbf{X}$.
2. $\mathbf{H}$ is symetric: $\mathbf{H} = \mathbf{H}^T$.
3. $\mathbf{H}$ is idempotent: $\mathbf{H}^2 = \mathbf{H}$.
4. $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ is the projection of $\mathbf{y}$ onto the column space $\mathcal{L}$.
5. $\mathrm{rank}(\mathbf{X}) = \mathrm{tr}(\mathbf{H}) = d$. [a]

[a] The last property can be easily proved using the equality $\mathrm{tr}(\mathbf{A}\mathbf{B}) = \mathrm{tr}(\mathbf{B}\mathbf{A})$.

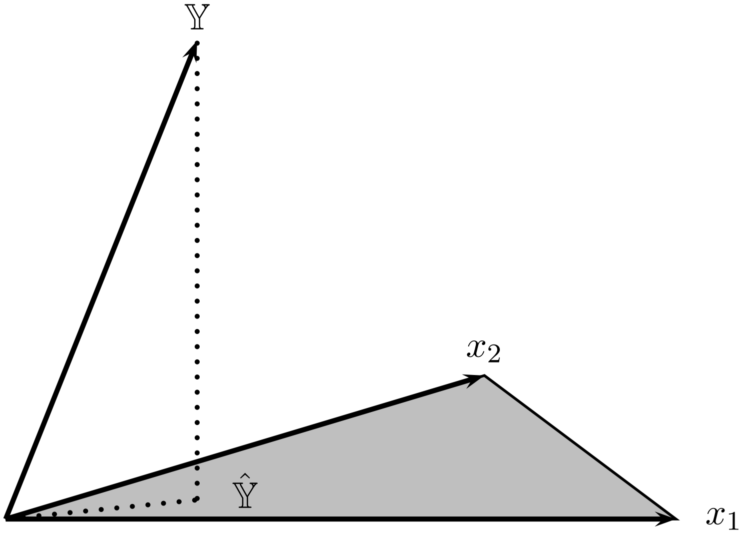

Figure 10.3. *The predicted values* $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ *are the projection of* $\mathbf{y}$ *onto the column space* $\mathcal{L}$ *of the design matrix* $\mathbf{X}$.

We therefore have the following geometric interpretation of least squares: The vector $\widehat{\mathbf{y}}$ is the closest vector in $\mathcal{L}$ to the data vector $\mathbf{y}$. The case where the $\mathbf{X}$ matrix has two columns is depicted in Figure 10.3.

10.4 When the True Regression Function is Linear

Let us now *temporarily* assume that the true regression function is linear. This is not a safe assumption to make in general, but it is interesting to see what happens when m is in fact linear.

First, the least squares estimator is the *maximum conditional likelihood estimator* if the errors have a normal distribution. The phrase "maximum conditional likelihood estimator" deserves some comments. The likelihood function is

$$\prod_{i=1}^{n} p(x_i, y_i) = \prod_{i=1}^{n} p_X(x_i) \times \prod_{i=1}^{n} p_{Y\,|\,X}(y_i\,|\,x_i). \tag{10.27}$$

where $p(x, y)$ is the joint density function of (X, Y) with p_X and $p_{Y\,|\,X}$ as its marginal and conditional density functions. The first term $\prod_{i=1}^{n} p_X(x_i)$ does not involve the parameter β. The second term is called the *conditional likelihood* and a maximum conditional likelihood estimator $\widehat{\beta}$ is defined to be the maximizer of this term.

10.28 Theorem. *Suppose that $y_i = \beta^T x_i + \epsilon_i$ and $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$. The conditional likelihood function $L(\beta)$ is*

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{\sum_{i=1}^{n}(y_i - x_i^T\beta)^2}{2\sigma^2} \right\} \tag{10.29}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{rss}{2\sigma^2}\right\} \exp\left\{-\frac{(\widehat{\beta}-\beta)^T \mathbf{X}^T\mathbf{X}(\widehat{\beta}-\beta)}{2\sigma^2}\right\} \tag{10.30}$$

where $\widehat{\beta}$ is the least squares estimator and $rss = \|\mathbf{y} - \mathbf{X}\widehat{\beta}\|^2$. The least squares estimator $\widehat{\beta}$ is also the maximum conditional likelihood estimator. The maximum likelihood estimator for σ^2 is $\widehat{\sigma}^2 = rss/n$.

Proof. See Exercise 9. □

Another property of least squares is minimaxity. Let us regard the $x_1, \ldots, x_n$ as fixed. Let $\mu = (m(x_1), \ldots, m(x_n))^T$, $\mu(\beta) = (\beta^T x_1, \ldots, \beta^T x_n)^T$ and $\mathcal{M} = \{\mu(\beta) : \beta \in \mathbb{R}^p\}$. Recall that $\widehat{\mathbf{y}} = \mu(\widehat{\beta})$. A linear estimator of μ is defined to be any estimator of the form $\widehat{\mu} = \mathbf{C}\mathbf{y}$ where $\mathbf{C}$ is a $n \times n$ matrix.[1] The linear minimax risk is defined as

$$\inf_{\mathbf{C}} \sup_{\mu\in\mathcal{M}} \mathbb{E}\|\mathbf{C}\mathbf{y} - \mu\|^2. \tag{10.31}$$

10.32 Theorem. *The least squares predictor $\widehat{\mathbf{y}}$ is linear minimax:*

$$\sup_{\mu\in\mathcal{M}} \mathbb{E}\|\widehat{\mathbf{y}} - \mu\|^2 = \inf_{\mathbf{C}} \sup_{\mu\in\mathcal{M}} \mathbb{E}\|\mathbf{C}\mathbf{y} - \mu\|^2. \tag{10.33}$$

[1] The word linear here refers to the fact that $\mathbf{C}(a_1\mathbf{y}_1 + a_2\mathbf{y}_2) = a_1\mathbf{C}\mathbf{y}_1 + a_2\mathbf{C}\mathbf{y}_2$.

A proof of this theorem can be found in Blaker (2001).

10.5 Summary

1. The linear oracle, or best linear predictor, is $\beta_*^T x$ where $\beta_* = \mathbb{E}\left\{(X^T X)^{-1}\right\} \mathrm{Cov}(Y, X)$. An estimate of β_* is $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

2. The predicted values are $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta} = \mathbf{H}\mathbf{y}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat matrix which projects $\mathbf{y}$ onto the column space of $\mathbf{X}$.

10.6 Bibliographic Remarks

A succinct book on linear regression is Weisberg (1985). A data-mining view of regression is given in Hastie et al. (2001). A comprehensive, theoretical treatment of regression can be found in Györfi et al. (2002).

Exercises

10.1 Suppose you predict Y with $g(X)$. Show that the prediction error $\mathbb{E}(Y - g(X))^2$ is minimized among all functions of x by taking $g(x) = m(x) \equiv \mathbb{E}(Y \mid X = x)$.

10.2 Prove Theorem 10.17. You could start by calculating the first-order optimality condition of (10.16).

10.3 Prove Theorem 10.18.

10.4 Prove that the right hand side of Equation (10.24) converges to zero in probability:

$$\left(\widehat{\beta} - \beta_*\right)^T \mathbf{\Sigma} \left(\widehat{\beta} - \beta_*\right) + 2\left(\beta_* - \widehat{\beta}\right)^T (C - \mathbf{\Sigma}\beta_*) \xrightarrow{\mathrm{P}} 0. \tag{10.34}$$

10.5 Prove Equation (10.10) using the following fact: for a random variable X with finite variance,

$$\mathbb{E}X^2 = (\mathbb{E}X)^2 + \mathrm{Var}(X). \tag{10.35}$$

10.6 Prove the formulas for the standard errors in Theorem 10.28. You should regard the x_is as fixed constants.

10.7 Consider the univariate *regression through the origin* model:

$$y_i = \beta x_i + \epsilon, \quad x_i \in \mathbb{R} \quad \text{for } i = 1, \ldots, n.$$

Find the least squares estimate for β. Find the standard error of the estimate. Find as weak as possible the conditions that guarantee the estimate to be consistent.

10.8 Prove that (10.29) equals (10.30).

10.9 Prove Theorem 10.28.

10.10 Get the automobile data from the UCI repository at

`http://archive.ics.uci.edu/ml/datasets/Auto+MPG`

(a) Use the 8 covariates to build a linear predictor of `mpg` (miles per gallon), using least squares. After fitting the model, plot the residuals versus all 8 covariates. If the model fits well, these residuals should look random. Comment on the residuals.
(b) Now define 64 new covariates of the form $X_1^2, X_2^2, \ldots, X_8^2, X_1 \cdot X_2, X_1 \cdot X_3 \ldots$. The intercept, the original 8 covariates and the 64 new covariates now give you a total of 73 covariates. Do you think this model will predict better or worse than the original model? Fit the model and compare the fit to the original model.
(c) Repeat part (b) but set aside half of the data as a training set and half as a test set. Use the training data to fit a linear model as in (a) and a bigger linear model as in (b). Now use the two models to predict the test data and compare the sum of squared errors. Which model is a better linear predictor?

10.11 Let $n = 100$ and $d = 50$. Simulate $(y_1, x_1), \ldots, (y_n, x_n)$ as follows. For $i = 1, \ldots, n$, draw x_i uniformly from the cube $[0,1]^{50}$. Now take

$$y_i = \beta_1 x_{i1} + \cdots + \beta_{50} x_{i,50} + \epsilon_i$$

where $\epsilon_i \sim N(0,1)$ and $\beta_i = i$.
(a) Estimate the β_is by least squares.
(b) Estimate the β_js by marginal regression. Thus we get $\widetilde{\beta}_j$ by regressing Y on X_j only. Compare with the estimates obtained in (a).
(c) In general, if $\widetilde{\beta}_j$ is the marginal estimator then $\mathbb{E}(\widetilde{\beta}_j) \neq \beta_j$. However, in this particular example we do have $\mathbb{E}(\widetilde{\beta}_j) = \beta_j$. Prove this.
(d) Construct an example where $\mathbb{E}(\widetilde{\beta}_j) \neq \beta_j$. Simulate data from the model and compare the marginal estimator to the least squares estimator.

10.12 Prove that the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the projection matrix onto the column space $\mathcal{L}$ of $\mathbf{X}$. In other words, show that $\mathbf{H}$ satisfies:

1. $\mathbf{H}$ is symmetric.
2. $\mathbf{H}$ is idempotent: $\mathbf{H}^2 = \mathbf{H}$.
3. If $v \in \mathcal{L}$ then $\mathbf{H}v = v$.

10.13 (Oracle and Consistency) Let $X \in \mathbb{R}$ and

$$Y = \gamma X^2 + \epsilon \tag{10.36}$$

where $\mathbb{E}(\epsilon) = 0$.

(a) Find an expression for the oracle linear predictor. In other words, find β_* such that $m(x) = \beta_* x$ minimizes the predictive risk.

(b) We are given n observations $(x_1, y_1), \ldots, (x_n, y_n)$ from (10.36). Give an estimator $\widehat{\beta}_n$ for β_* and show that it is consistent.

10.14 Suppose that the data $(x_1, y_1), \ldots, (x_n, y_n)$ come from the model $y_i = \sum_{j=1}^{d} \beta_j x_{ij} + \epsilon_i$ where $\mathbb{E}(\epsilon_i) = 0$ and the ϵ_i's are independent of the X's and Y's. Here, $x_i = (x_{i1}, \ldots, x_{id})^T$. We will further assume that $\mathbb{E}(X_j) = 0$, $\mathbb{E}(X_j^2) = 1$. Define the marginal regression estimator $\widehat{\beta}_j = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i \, x_{ij}$.

(a) Find the mean and variance of $\widehat{\beta}_j$. What is the bias of the estimator? (Treat the x_{ij} as random, not fixed.)

(b) Now suppose that the random variables x_{ij} are all independent. Show that, in this case, $\widehat{\beta}_j$ is a consistent estimator of β_j.

(b) Now suppose that the true model is $y_i = m(x_i) + \epsilon_i$ for some arbitrary function $m(x)$. What does $\widehat{\beta}_j$ converge to as $n \to \infty$?