13
Linear and Logistic Regression

Regression is a method for studying the relationship between a **response variable** Y and a **covariate** X. The covariate is also called a **predictor variable** or a **feature.** [1] One way to summarize the relationship between X and Y is through the **regression function**

$$r(x) = \mathbb{E}(Y|X = x) = \int y\, f(y|x)dy. \tag{13.1}$$

Our goal is to estimate the regression function $r(x)$ from data of the form

$$(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y}.$$

In this Chapter, we take a parametric approach and assume that r is linear. In Chapters 20 and 21 we discuss nonparametric regression.

13.1 Simple Linear Regression

The simplest version of regression is when X_i is simple (one-dimensional) and $r(x)$ is assumed to be linear:

$$r(x) = \beta_0 + \beta_1 x.$$

[1] The term "regression" is due to Sir Francis Galton (1822-1911) who noticed that tall and short men tend to have sons with heights closer to the mean. He called this "regression towards the mean."

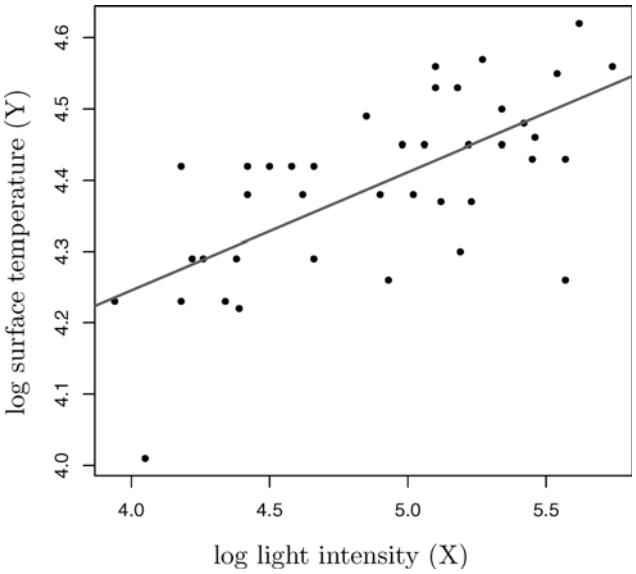

FIGURE 13.1. Data on nearby stars. The solid line is the least squares line.

This model is called the **the simple linear regression model**. We will make the further simplifying assumption that $\mathbb{V}(\epsilon_i|X = x) = \sigma^2$ does not depend on x. We can thus write the linear regression model as follows.

13.1 Definition. The Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{13.2}$$

where $\mathbb{E}(\epsilon_i|X_i) = 0$ *and* $\mathbb{V}(\epsilon_i|X_i) = \sigma^2$.

13.2 Example. Figure 13.1 shows a plot of log surface temperature (Y) versus log light intensity (X) for some nearby stars. Also on the plot is an estimated linear regression line which will be explained shortly. ■

The unknown parameters in the model are the intercept β_0 and the slope β_1 and the variance σ^2. Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ denote estimates of β_0 and β_1. The **fitted line** is

$$\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x. \tag{13.3}$$

The **predicted values** or **fitted values** are $\widehat{Y}_i = \widehat{r}(X_i)$ and the **residuals** are defined to be

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \Big(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\Big). \tag{13.4}$$

The **residual sums of squares** or RSS, which measures how well the line fits the data, is defined by $\text{RSS} = \sum_{i=1}^n \widehat{\epsilon}_i^2$.

13.3 Definition. *The* **least squares estimates** *are the values* $\widehat{\beta}_0$ *and* $\widehat{\beta}_1$ *that minimize* $\text{RSS} = \sum_{i=1}^n \widehat{\epsilon}_i^2$.

13.4 Theorem. *The least squares estimates are given by*

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}, \tag{13.5}$$

$$\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n. \tag{13.6}$$

An unbiased estimate of σ^2 *is*

$$\widehat{\sigma}^2 = \left(\frac{1}{n-2}\right) \sum_{i=1}^n \widehat{\epsilon}_i^2. \tag{13.7}$$

13.5 Example. Consider the star data from Example 13.2. The least squares estimates are $\widehat{\beta}_0 = 3.58$ and $\widehat{\beta}_1 = 0.166$. The fitted line $\widehat{r}(x) = 3.58 + 0.166\, x$ is shown in Figure 13.1. ■

13.6 Example (The 2001 Presidential Election)**.** Figure 13.2 shows the plot of votes for Buchanan (Y) versus votes for Bush (X) in Florida. The least squares estimates (omitting Palm Beach County) and the standard errors are

$$\begin{aligned} \widehat{\beta}_0 &= 66.0991 \quad \widehat{\mathsf{se}}(\widehat{\beta}_0) = 17.2926 \\ \widehat{\beta}_1 &= 0.0035 \quad \widehat{\mathsf{se}}(\widehat{\beta}_1) = 0.0002. \end{aligned}$$

The fitted line is

$$\text{Buchanan} = 66.0991 + 0.0035\, \text{Bush}.$$

(We will see later how the standard errors were computed.) Figure 13.2 also shows the residuals. The inferences from linear regression are most accurate when the residuals behave like random normal numbers. Based on the residual plot, this is not the case in this example. If we repeat the analysis replacing votes with log(votes) we get

$$\begin{aligned} \widehat{\beta}_0 &= -2.3298 \quad \widehat{\mathsf{se}}(\widehat{\beta}_0) = 0.3529 \\ \widehat{\beta}_1 &= 0.730300 \quad \widehat{\mathsf{se}}(\widehat{\beta}_1) = 0.0358. \end{aligned}$$

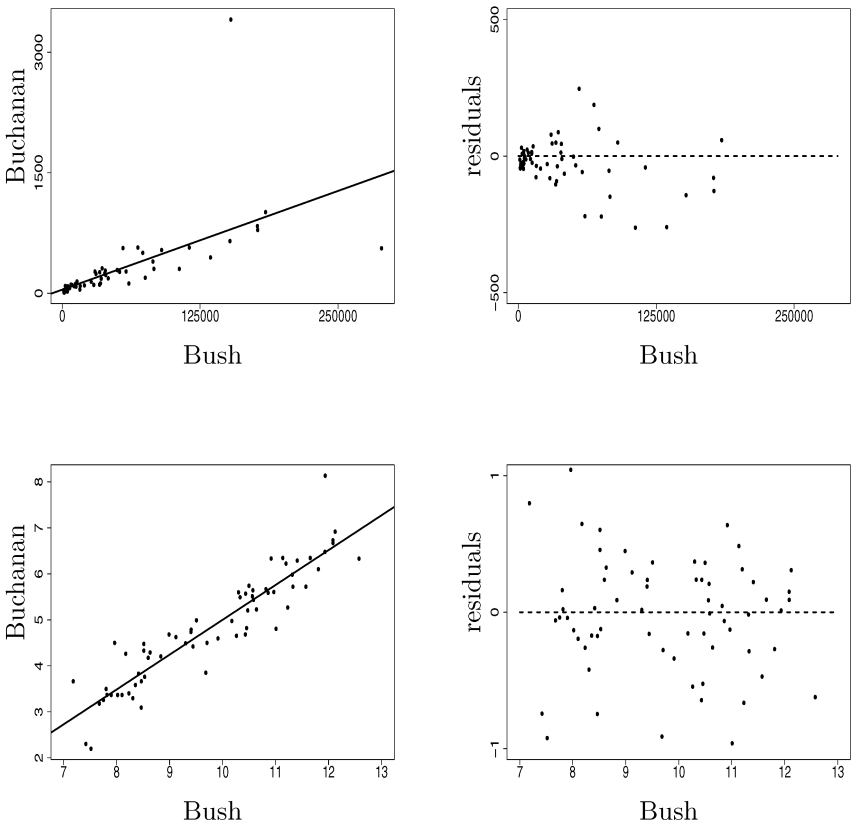

FIGURE 13.2. Voting Data for Election 2000. See example 13.6.

This gives the fit

$$\log(\text{Buchanan}) = -2.3298 + 0.7303 \ \log(\text{Bush}).$$

The residuals look much healthier. Later, we shall address the following question: how do we see if Palm Beach County has a statistically plausible outcome? ■

13.2 Least Squares and Maximum Likelihood

Suppose we add the assumption that $\epsilon_i | X_i \sim N(0, \sigma^2)$, that is,

$$Y_i | X_i \sim N(\mu_i, \sigma^2)$$

where $\mu_i = \beta_0 + \beta_1 X_i$. The likelihood function is

$$\begin{aligned}\prod_{i=1}^n f(X_i, Y_i) &= \prod_{i=1}^n f_X(X_i) f_{Y|X}(Y_i|X_i) \\ &= \prod_{i=1}^n f_X(X_i) \times \prod_{i=1}^n f_{Y|X}(Y_i|X_i) \\ &= \mathcal{L}_1 \times \mathcal{L}_2\end{aligned}$$

where $\mathcal{L}_1 = \prod_{i=1}^n f_X(X_i)$ and

$$\mathcal{L}_2 = \prod_{i=1}^n f_{Y|X}(Y_i|X_i). \tag{13.8}$$

The term $\mathcal{L}_1$ does not involve the parameters β_0 and β_1. We shall focus on the second term $\mathcal{L}_2$ which is called the **conditional likelihood**, given by

$$\mathcal{L}_2 \equiv \mathcal{L}(\beta_0, \beta_1, \sigma) = \prod_{i=1}^n f_{Y|X}(Y_i|X_i) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (Y_i - \mu_i)^2\right\}.$$

The conditional log-likelihood is

$$\ell(\beta_0, \beta_1, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n \Big(Y_i - (\beta_0 + \beta_1 X_i)\Big)^2. \tag{13.9}$$

To find the MLE of (β_0, β_1) we maximize $\ell(\beta_0, \beta_1, \sigma)$. From (13.9) we see that maximizing the likelihood is the same as minimizing the RSS $\sum_{i=1}^n \Big(Y_i - (\beta_0 + \beta_1 X_i)\Big)^2$. Therefore, we have shown the following:

13.7 Theorem. *Under the assumption of Normality, the least squares estimator is also the maximum likelihood estimator.*

We can also maximize $\ell(\beta_0, \beta_1, \sigma)$ over σ, yielding the MLE

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i \widehat{\epsilon}_i^2. \tag{13.10}$$

This estimator is similar to, but not identical to, the unbiased estimator. Common practice is to use the unbiased estimator (13.7).

13.3 Properties of the Least Squares Estimators

We now record the standard errors and limiting distribution of the least squares estimator. In regression problems, we usually focus on the properties of the estimators conditional on $X^n = (X_1, \ldots, X_n)$. Thus, we state the means and variances as conditional means and variances.

13.8 Theorem. *Let $\widehat{\beta}^T = (\widehat{\beta}_0, \widehat{\beta}_1)^T$ denote the least squares estimators. Then,*

$$\begin{aligned} \mathbb{E}(\widehat{\beta}|X^n) &= \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \\ \mathbb{V}(\widehat{\beta}|X^n) &= \frac{\sigma^2}{n\, s_X^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix} \end{aligned} \tag{13.11}$$

where $s_X^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

The estimated standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are obtained by taking the square roots of the corresponding diagonal terms of $\mathbb{V}(\widehat{\beta}|X^n)$ and inserting the estimate $\widehat{\sigma}$ for σ. Thus,

$$\widehat{\mathsf{se}}(\widehat{\beta}_0) = \frac{\widehat{\sigma}}{s_X\sqrt{n}} \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}} \tag{13.12}$$

$$\widehat{\mathsf{se}}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}. \tag{13.13}$$

We should really write these as $\widehat{\mathsf{se}}(\widehat{\beta}_0|X^n)$ and $\widehat{\mathsf{se}}(\widehat{\beta}_1|X^n)$ but we will use the shorter notation $\widehat{\mathsf{se}}(\widehat{\beta}_0)$ and $\widehat{\mathsf{se}}(\widehat{\beta}_1)$.

13.9 Theorem. *Under appropriate conditions we have:*

1. *(Consistency): $\widehat{\beta}_0 \xrightarrow{\mathrm{P}} \beta_0$ and $\widehat{\beta}_1 \xrightarrow{\mathrm{P}} \beta_1$.*

2. *(Asymptotic Normality):*

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\mathsf{se}}(\widehat{\beta}_0)} \rightsquigarrow N(0,1) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathsf{se}}(\widehat{\beta}_1)} \rightsquigarrow N(0,1).$$

3. *Approximate $1-\alpha$ confidence intervals for β_0 and β_1 are*

$$\widehat{\beta}_0 \pm z_{\alpha/2}\, \widehat{\mathsf{se}}(\widehat{\beta}_0) \quad \text{and} \quad \widehat{\beta}_1 \pm z_{\alpha/2}\, \widehat{\mathsf{se}}(\widehat{\beta}_1). \tag{13.14}$$

4. *The Wald test* [2] *for testing* $H_0 : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$ *is: reject* H_0 *if* $|W| > z_{\alpha/2}$ *where* $W = \widehat{\beta}_1/\widehat{\mathsf{se}}(\widehat{\beta}_1)$.

13.10 Example. For the election data, on the log scale, a 95 percent confidence interval is $.7303 \pm 2(.0358) = (.66, .80)$. The Wald statistics for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ is $|W| = |.7303 - 0|/.0358 = 20.40$ with a p-value of $\mathbb{P}(|Z| > 20.40) \approx 0$. This is strong evidence that that the true slope is not 0. ■

13.4 Prediction

Suppose we have estimated a regression model $\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$ from data $(X_1, Y_1), \ldots, (X_n, Y_n)$. We observe the value $X = x_*$ of the covariate for a new subject and we want to predict their outcome Y_*. An estimate of Y_* is

$$\widehat{Y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_*. \tag{13.15}$$

Using the formula for the variance of the sum of two random variables,

$$\mathbb{V}(\widehat{Y}_*) = \mathbb{V}(\widehat{\beta}_0 + \widehat{\beta}_1 x_*) = \mathbb{V}(\widehat{\beta}_0) + x_*^2 \mathbb{V}(\widehat{\beta}_1) + 2x_* \mathsf{Cov}(\widehat{\beta}_0, \widehat{\beta}_1).$$

Theorem 13.8 gives the formulas for all the terms in this equation. The estimated standard error $\widehat{\mathsf{se}}(\widehat{Y}_*)$ is the square root of this variance, with $\widehat{\sigma}^2$ in place of σ^2. However, the confidence interval for Y_* is **not** of the usual form $\widehat{Y}_* \pm z_{\alpha/2}\widehat{\mathsf{se}}$. The reason for this is explained in Exercise 10. The correct form of the confidence interval is given in the following theorem.

13.11 Theorem (Prediction Interval). *Let*

$$\widehat{\xi}_n^2 = \widehat{\sigma}^2 \left(\frac{\sum_{i=1}^n (X_i - X_*)^2}{n \sum_i (X_i - \overline{X})^2} + 1 \right). \tag{13.16}$$

An approximate $1 - \alpha$ *prediction interval for* Y_* *is*

$$\widehat{Y}_* \pm z_{\alpha/2}\, \widehat{\xi}_n. \tag{13.17}$$

[2] Recall from equation (10.5) that the Wald statistic for testing $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$ is $W = (\widehat{\beta} - \beta_0)/\widehat{\mathsf{se}}(\widehat{\beta})$.

13.12 Example (Election Data Revisited)**.** On the log scale, our linear regression gives the following prediction equation:

$$\log(\text{Buchanan}) = -2.3298 + 0.7303 \log(\text{Bush}).$$

In Palm Beach, Bush had 152,954 votes and Buchanan had 3,467 votes. On the log scale this is 11.93789 and 8.151045. How likely is this outcome, assuming our regression model is appropriate? Our prediction for log Buchanan votes -2.3298 + .7303 (11.93789)=6.388441. Now, 8.151045 is bigger than 6.388441 but is it "significantly" bigger? Let us compute a confidence interval. We find that $\widehat{\xi}_n = .093775$ and the approximate 95 percent confidence interval is (6.200,6.578) which clearly excludes 8.151. Indeed, 8.151 is nearly 20 standard errors from $\widehat{Y}_*$. Going back to the vote scale by exponentiating, the confidence interval is (493,717) compared to the actual number of votes which is 3,467. ■

13.5 Multiple Regression

Now suppose that the covariate is a vector of length k. The data are of the form

$$(Y_1, X_1), \ldots, (Y_i, X_i), \ldots, (Y_n, X_n)$$

where

$$X_i = (X_{i1}, \ldots, X_{ik}).$$

Here, X_i is the vector of k covariate values for the i^{th} observation. The linear regression model is

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \tag{13.18}$$

for $i = 1, \ldots, n$, where $\mathbb{E}(\epsilon_i | X_{1i}, \ldots, X_{ki}) = 0$. Usually we want to include an intercept in the model which we can do by setting $X_{i1} = 1$ for $i = 1, \ldots, n$. At this point it will be more convenient to express the model in matrix notation. The outcomes will be denoted by

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

and the covariates will be denoted by

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix}.$$

Each row is one observation; the columns correspond to the k covariates. Thus, X is a $(n \times k)$ matrix. Let

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \text{and} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Then we can write (13.18) as

$$Y = X\beta + \epsilon. \tag{13.19}$$

The form of the least squares estimate is given in the following theorem.

13.13 Theorem. *Assuming that the $(k \times k)$ matrix $X^T X$ is invertible,*

$$\widehat{\beta} = (X^T X)^{-1} X^T Y \tag{13.20}$$

$$\mathbb{V}(\widehat{\beta}|X^n) = \sigma^2 (X^T X)^{-1} \tag{13.21}$$

$$\widehat{\beta} \approx N(\beta, \sigma^2 (X^T X)^{-1}). \tag{13.22}$$

The estimate regression function is $\widehat{r}(x) = \sum_{j=1}^k \widehat{\beta}_j x_j$. An unbiased estimate of σ^2 is

$$\widehat{\sigma}^2 = \left(\frac{1}{n-k}\right) \sum_{i=1}^n \widehat{\epsilon}_i^2$$

where $\widehat{\epsilon} = X\widehat{\beta} - Y$ is the vector of residuals. An approximate $1-\alpha$ confidence interval for β_j is

$$\widehat{\beta}_j \pm z_{\alpha/2} \widehat{\mathsf{se}}(\widehat{\beta}_j) \tag{13.23}$$

where $\widehat{\mathsf{se}}^2(\widehat{\beta}_j)$ is the j^{th} diagonal element of the matrix $\widehat{\sigma}^2 (X^T X)^{-1}$.

13.14 Example. Crime data on 47 states in 1960 can be obtained from
http://lib.stat.cmu.edu/DASL/Stories/USCrime.html.
If we fit a linear regression of crime rate on 10 variables we get the following:

Covariate	$\widehat{\beta}_j$	$\widehat{\mathsf{se}}(\widehat{\beta}_j)$	t value	p-value
(Intercept)	-589.39	167.59	-3.51	0.001 **
Age	1.04	0.45	2.33	0.025 *
Southern State	11.29	13.24	0.85	0.399
Education	1.18	0.68	1.7	0.093
Expenditures	0.96	0.25	3.86	0.000 ***
Labor	0.11	0.15	0.69	0.493
Number of Males	0.30	0.22	1.36	0.181
Population	0.09	0.14	0.65	0.518
Unemployment (14–24)	-0.68	0.48	-1.4	0.165
Unemployment (25–39)	2.15	0.95	2.26	0.030 *
Wealth	-0.08	0.09	-0.91	0.367

This table is typical of the output of a multiple regression program. The "t-value" is the Wald test statistic for testing $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. The asterisks denote "degree of significance" and more asterisks denote smaller p-values. The example raises several important questions: (1) should we eliminate some variables from this model? (2) should we interpret these relationships as causal? For example, should we conclude that low crime prevention expenditures cause high crime rates? We will address question (1) in the next section. We will not address question (2) until Chapter 16. ■

13.6 Model Selection

Example 13.14 illustrates a problem that often arises in multiple regression. We may have data on many covariates but we may not want to include all of them in the model. A smaller model with fewer covariates has two advantages: it might give better predictions than a big model and it is more parsimonious (simpler). Generally, as you add more variables to a regression, the bias of the predictions decreases and the variance increases. Too few covariates yields high bias; this called **underfitting**. Too many covariates yields high variance; this called **overfitting**. Good predictions result from achieving a good balance between bias and variance.

In model selection there are two problems: (i) assigning a "score" to each model which measures, in some sense, how good the model is, and (ii) searching through all the models to find the model with the best score.

Let us first discuss the problem of scoring models. Let $S \subset \{1, \ldots, k\}$ and let $\mathcal{X}_S = \{X_j : \ j \in S\}$ denote a subset of the covariates. Let β_S denote the coefficients of the corresponding set of covariates and let $\widehat{\beta}_S$ denote the least squares estimate of β_S. Also, let X_S denote the X matrix for this subset of

covariates and define $\widehat{r}_S(x)$ to be the estimated regression function. The predicted values from model S are denoted by $\widehat{Y}_i(S) = \widehat{r}_S(X_i)$. The **prediction risk** is defined to be

$$R(S) = \sum_{i=1}^{n} \mathbb{E}(\widehat{Y}_i(S) - Y_i^*)^2 \tag{13.24}$$

where Y_i^* denotes the value of a future observation of Y_i at covariate value X_i. Our goal is to choose S to make $R(S)$ small.

The **training error** is defined to be

$$\widehat{R}_{\text{tr}}(S) = \sum_{i=1}^{n} (\widehat{Y}_i(S) - Y_i)^2.$$

This estimate is very biased as an estimate of $R(S)$.

13.15 Theorem. *The training error is a downward-biased estimate of the prediction risk:*

$$\mathbb{E}(\widehat{R}_{\text{tr}}(S)) < R(S).$$

In fact,

$$\mathsf{bias}(\widehat{R}_{\text{tr}}(S)) = \mathbb{E}(\widehat{R}_{\text{tr}}(S)) - R(S) = -2\sum_{i=1} \mathsf{Cov}(\widehat{Y}_i, Y_i). \tag{13.25}$$

The reason for the bias is that the data are being used twice: to estimate the parameters and to estimate the risk. When we fit a complex model with many parameters, the covariance $\mathsf{Cov}(\widehat{Y}_i, Y_i)$ will be large and the bias of the training error gets worse. Here are some better estimates of risk.

Mallow's C_p statistic is defined by

$$\widehat{R}(S) = \widehat{R}_{\text{tr}}(S) + 2|S|\widehat{\sigma}^2 \tag{13.26}$$

where $|S|$ denotes the number of terms in S and $\widehat{\sigma}^2$ is the estimate of σ^2 obtained from the full model (with all covariates in the model). This is simply the training error plus a bias correction. This estimate is named in honor of Colin Mallows who invented it. The first term in (13.26) measures the fit of the model while the second measure the complexity of the model. Think of the C_p statistic as:

$$\text{lack of fit } + \text{ complexity penalty.}$$

Thus, **finding a good model involves trading off fit and complexity.**

A related method for estimating risk is **AIC (Akaike Information Criterion).** The idea is to choose S to maximize

$$\ell_S - |S| \tag{13.27}$$

where ℓ_S is the log-likelihood of the model evaluated at the MLE. [3] This can be thought of "goodness of fit" minus "complexity." In linear regression with Normal errors (and taking σ equal to its estimate from the largest model), maximizing AIC is equivalent to minimizing Mallow's C_p; see Exercise 8. The appendix contains more explanation about AIC.

Yet another method for estimating risk is **leave-one-out cross-validation**. In this case, the risk estimator is

$$\widehat{R}_{CV}(S) = \sum_{i=1}^{n}(Y_i - \widehat{Y}_{(i)})^2 \tag{13.28}$$

where $\widehat{Y}_{(i)}$ is the prediction for Y_i obtained by fitting the model with Y_i omitted. It can be shown that

$$\widehat{R}_{CV}(S) = \sum_{i=1}^{n}\left(\frac{Y_i - \widehat{Y}_i(S)}{1 - U_{ii}(S)}\right)^2 \tag{13.29}$$

where $U_{ii}(S)$ is the i^{th} diagonal element of the matrix

$$U(S) = X_S(X_S^T X_S)^{-1} X_S^T. \tag{13.30}$$

Thus, one need not actually drop each observation and re-fit the model. A generalization is **k-fold cross-validation.** Here we divide the data into k groups; often people take $k = 10$. We omit one group of data and fit the models to the remaining data. We use the fitted model to predict the data in the group that was omitted. We then estimate the risk by $\sum_i (Y_i - \widehat{Y}_i)^2$ where the sum is over the the data points in the omitted group. This process is repeated for each of the k groups and the resulting risk estimates are averaged.

For linear regression, Mallows C_p and cross-validation often yield essentially the same results so one might as well use Mallows' method. In some of the more complex problems we will discuss later, cross-validation will be more useful.

Another scoring method is BIC (Bayesian information criterion). Here we choose a model to maximize

$$\text{BIC}(S) = \ell_S - \frac{|S|}{2}\log n. \tag{13.31}$$

[3] Some texts use a slightly different definition of AIC which involves multiplying the definition here by 2 or -2. This has no effect on which model is selected.

The BIC score has a Bayesian interpretation. Let $\mathcal{S} = \{S_1, \ldots, S_m\}$ denote a set of models. Suppose we assign the prior $\mathbb{P}(S_j) = 1/m$ over the models. Also, assume we put a smooth prior on the parameters within each model. It can be shown that the posterior probability for a model is approximately,

$$\mathbb{P}(S_j|\text{data}) \approx \frac{e^{BIC(S_j)}}{\sum_r e^{BIC(S_r)}}.$$

Hence, choosing the model with highest BIC is like choosing the model with highest posterior probability. The BIC score also has an information-theoretic interpretation in terms of something called minimum description length. The BIC score is identical to Mallows C_p except that it puts a more severe penalty for complexity. It thus leads one to choose a smaller model than the other methods.

Now let us turn to the problem of model search. If there are k covariates then there are 2^k possible models. We need to search through all these models, assign a score to each one, and choose the model with the best score. If k is not too large we can do a complete search over all the models. When k is large, this is infeasible. In that case we need to search over a subset of all the models. Two common methods are **forward and backward stepwise regression.** In forward stepwise regression, we start with no covariates in the model. We then add the one variable that leads to the best score. We continue adding variables one at a time until the score does not improve. Backwards stepwise regression is the same except that we start with the biggest model and drop one variable at a time. Both are greedy searches; nether is guaranteed to find the model with the best score. Another popular method is to do random searching through the set of all models. However, there is no reason to expect this to be superior to a deterministic search.

13.16 Example. We applied backwards stepwise regression to the crime data using AIC. The following was obtained from the program R. This program uses a slightly different definition of AIC. With their definition, we seek the smallest (not largest) possible AIC. This is the same is minimizing Mallows C_p.

The full model (which includes all covariates) has AIC= 310.37. In ascending order, the AIC scores for deleting one variable are as follows:

variable	Pop	Labor	South	Wealth	Males	U1	Educ.	U2	Age	Expend
AIC	308	309	309	309	310	310	312	314	315	324

For example, if we dropped Pop from the model and kept the other terms, then the AIC score would be 308. Based on this information we drop "pop-

ulation" from the model and the current AIC score is 308. Now we consider dropping a variable from the current model. The AIC scores are:

variable	South	Labor	Wealth	Males	U1	Education	U2	Age	Expend
AIC	308	308	308	309	309	310	313	313	329

We then drop "Southern" from the model. This process is continued until there is no gain in AIC by dropping any variables. In the end, we are left with the following model:

$$\begin{aligned}\text{Crime} \;=\; & 1.2\ \text{Age} \;+\; .75\ \text{Education} \;+\; .87\ \text{Expenditure} \\ & +\; .34\ \text{Males} \;-\; .86\ \text{U1} \;+\; 2.31\ \text{U2}.\end{aligned}$$

Warning! This does not yet address the question of which variables are **causes** of crime. ■

There is another method for model selection that avoids having to search through all possible models. This method, which is due to Zheng and Loh (1995), does not seek to minimize prediction errors. Rather, it assumes some subset of the β_j's are exactly equal to 0 and tries to find the true model, that is, the smallest sub-model consisting of nonzero β_j terms. The method is carried out as follows.

Zheng-Loh Model Selection Method [4]

1. Fit the full model with all k covariates and let $W_j = \widehat{\beta}_j / \widehat{\mathsf{se}}(\widehat{\beta}_j)$ denote the Wald test statistic for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$.
2. Order the test statistics from largest to smallest in absolute value:
$$|W_{(1)}| \geq |W_{(2)}| \geq \cdots \geq |W_{(k)}|.$$
3. Let $\widehat{j}$ be the value of j that minimizes
$$\text{RSS}(j) + j\,\widehat{\sigma}^2\,\log n$$
where $\text{RSS}(j)$ is the residual sums of squares from the model with the j largest Wald statistics.
4. Choose, as the final model, the regression with the $\widehat{j}$ terms with the largest absolute Wald statistics.

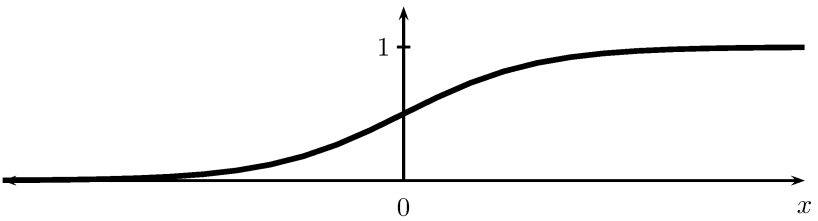

FIGURE 13.3. The logistic function $p = e^x/(1+e^x)$.

Zheng and Loh showed that, under appropriate conditions, this method chooses the true model with probability tending to one as the sample size increases.

13.7 Logistic Regression

So far we have assumed that Y_i is real valued. **Logistic regression** is a parametric method for regression when $Y_i \in \{0,1\}$ is binary. For a k-dimensional covariate X, the model is

$$p_i \equiv p_i(\beta) \equiv \mathbb{P}(Y_i = 1|X = x) = \frac{e^{\beta_0+\sum_{j=1}^k \beta_j x_{ij}}}{1+e^{\beta_0+\sum_{j=1}^k \beta_j x_{ij}}} \tag{13.32}$$

or, equivalently,

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} \tag{13.33}$$

where

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right). \tag{13.34}$$

The name "logistic regression" comes from the fact that $e^x/(1+e^x)$ is called the logistic function. A plot of the logistic for a one-dimensional covariate is shown in Figure 13.3.

Because the Y_i's are binary, the data are Bernoulli:

$$Y_i|X_i = x_i \sim \text{Bernoulli}(p_i).$$

Hence the (conditional) likelihood function is

$$\mathcal{L}(\beta) = \prod_{i=1}^n p_i(\beta)^{Y_i}(1-p_i(\beta))^{1-Y_i}. \tag{13.35}$$

[4] This is just one version of their method. In particular, the penalty $j \log n$ is only one choice from a set of possible penalty functions.

The MLE $\widehat{\beta}$ has to be obtained by maximizing $\mathcal{L}(\beta)$ numerically. There is a fast numerical algorithm called reweighted least squares. The steps are as follows:

Reweighted Least Squares Algorithm

Choose starting values $\widehat{\beta}^0 = (\widehat{\beta}_0^0, \ldots, \widehat{\beta}_k^0)$ and compute p_i^0 using equation (13.32), for $i = 1, \ldots, n$. Set $s = 0$ and iterate the following steps until convergence.

1. Set
$$Z_i = \text{logit}(p_i^s) + \frac{Y_i - p_i^s}{p_i^s(1 - p_i^s)}, \quad i = 1, \ldots, n.$$
2. Let W be a diagonal matrix with (i, i) element equal to $p_i^s(1 - p_i^s)$.
3. Set
$$\widehat{\beta}^s = (X^T W X)^{-1} X^T W Y.$$
This corresponds to doing a (weighted) linear regression of Z on Y.
4. Set $s = s + 1$ and go back to the first step.

The Fisher information matrix I can also be obtained numerically. The estimate standard error of $\widehat{\beta}_j$ is the (j, j) element of $J = I^{-1}$. Model selection is usually done using the AIC score $\ell_S - |S|$.

13.17 Example. The Coronary Risk-Factor Study (CORIS) data involve 462 males between the ages of 15 and 64 from three rural areas in South Africa, (Rousseauw et al. (1983)). The outcome Y is the presence ($Y = 1$) or absence ($Y = 0$) of coronary heart disease. There are 9 covariates: systolic blood pressure, cumulative tobacco (kg), ldl (low density lipoprotein cholesterol), adiposity, famhist (family history of heart disease), typea (type-A behavior), obesity, alcohol (current alcohol consumption), and age. A logistic regression yields the following estimates and Wald statistics W_j for the coefficients:

Covariate	$\widehat{\beta}_j$	$\widehat{\mathsf{se}}$	W_j	p-value
Intercept	-6.145	1.300	-4.738	0.000
sbp	0.007	0.006	1.138	0.255
tobacco	0.079	0.027	2.991	0.003
ldl	0.174	0.059	2.925	0.003
adiposity	0.019	0.029	0.637	0.524
famhist	0.925	0.227	4.078	0.000
typea	0.040	0.012	3.233	0.001
obesity	-0.063	0.044	-1.427	0.153
alcohol	0.000	0.004	0.027	0.979
age	0.045	0.012	3.754	0.000

Are you surprised by the fact that systolic blood pressure is not significant or by the minus sign for the obesity coefficient? If yes, then you are confusing association and causation. This issue is discussed in Chapter 16. The fact that blood pressure is not significant does not mean that blood pressure is not an important *cause* of heart disease. It means that it is not an important *predictor* of heart disease relative to the other variables in the model. ■

13.8 Bibliographic Remarks

A succinct book on linear regression is Weisberg (1985). A data-mining view of regression is given in Hastie et al. (2001). The Akaike Information Criterion (AIC) is due to Akaike (1973). The Bayesian Information Criterion (BIC) is due to Schwarz (1978). References on logistic regression include Agresti (1990) and Dobson (2001).

13.9 Appendix

The Akaike Information Criterion (AIC). Consider a set of models $\{M_1, M_2, \ldots\}$. Let $\widehat{f}_j(x)$ denote the estimated probability function obtained by using the maximum likelihood estimator of model M_j. Thus, $\widehat{f}_j(x) = \widehat{f}(x; \widehat{\beta}_j)$ where $\widehat{\beta}_j$ is the MLE of the set of parameters β_j for model M_j. We will use the loss function $D(f, \widehat{f})$ where

$$D(f, g) = \sum_x f(x) \log\left(\frac{f(x)}{g(x)}\right)$$

is the Kullback-Leibler distance between two probability functions. The corresponding risk function is $R(f, \widehat{f}) = \mathbb{E}(D(f, \widehat{f})$. Notice that $D(f, \widehat{f}) = c -$

$A(f, \widehat{f})$ where $c = \sum_x f(x) \log f(x)$ does not depend on $\widehat{f}$ and

$$A(f, \widehat{f}) = \sum_x f(x) \log \widehat{f}(x).$$

Thus, minimizing the risk is equivalent to maximizing $a(f, \widehat{f}) \equiv \mathbb{E}(A(f, \widehat{f}))$.

It is tempting to estimate $a(f, \widehat{f})$ by $\sum_x \widehat{f}(x) \log \widehat{f}(x)$ but, just as the training error in regression is a highly biased estimate of prediction risk, it is also the case that $\sum_x \widehat{f}(x) \log \widehat{f}(x)$ is a highly biased estimate of $a(f, \widehat{f})$. In fact, the bias is approximately equal to $|M_j|$. Thus:

13.18 Theorem. *$AIC(M_j)$ is an approximately unbiased estimate of $a(f, \widehat{f})$.*

13.10 Exercises

1. Prove Theorem 13.4.

2. Prove the formulas for the standard errors in Theorem 13.8. You should regard the X_i's as fixed constants.

3. Consider the **regression through the origin** model:

   $$Y_i = \beta X_i + \epsilon.$$

   Find the least squares estimate for β. Find the standard error of the estimate. Find conditions that guarantee that the estimate is consistent.

4. Prove equation (13.25).

5. In the simple linear regression model, construct a Wald test for $H_0 : \beta_1 = 17\beta_0$ versus $H_1 : \beta_1 \neq 17\beta_0$.

6. Get the passenger car mileage data from

   http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html

   (a) Fit a simple linear regression model to predict MPG (miles per gallon) from HP (horsepower). Summarize your analysis including a plot of the data with the fitted line.

   (b) Repeat the analysis but use log(MPG) as the response. Compare the analyses.

7. Get the passenger car mileage data from

   http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html

   (a) Fit a multiple linear regression model to predict MPG (miles per gallon) from the other variables. Summarize your analysis.

   (b) Use Mallow C_p to select a best sub-model. To search through the models try (i) forward stepwise, (ii) backward stepwise. Summarize your findings.

   (c) Use the Zheng-Loh model selection method and compare to (b).

   (d) Perform all possible regressions. Compare C_p and BIC. Compare the results.

8. Assume a linear regression model with Normal errors. Take σ known. Show that the model with highest AIC (equation (13.27)) is the model with the lowest Mallows C_p statistic.

9. In this question we will take a closer look at the AIC method. Let $X_1, \ldots, X_n$ be IID observations. Consider two models $\mathcal{M}_0$ and $\mathcal{M}_1$. Under $\mathcal{M}_0$ the data are assumed to be $N(0,1)$ while under $\mathcal{M}_1$ the data are assumed to be $N(\theta, 1)$ for some unknown $\theta \in \mathbb{R}$:

$$\begin{aligned} \mathcal{M}_0 : \; X_1, \ldots, X_n &\sim N(0,1) \\ \mathcal{M}_1 : \; X_1, \ldots, X_n &\sim N(\theta, 1), \quad \theta \in \mathbb{R}. \end{aligned}$$

   This is just another way to view the hypothesis testing problem: $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Let $\ell_n(\theta)$ be the log-likelihood function. The AIC score for a model is the log-likelihood at the MLE minus the number of parameters. (Some people multiply this score by 2 but that is irrelevant.) Thus, the AIC score for $\mathcal{M}_0$ is $AIC_0 = \ell_n(0)$ and the AIC score for $\mathcal{M}_1$ is $AIC_1 = \ell_n(\widehat{\theta}) - 1$. Suppose we choose the model with the highest AIC score. Let J_n denote the selected model:

$$J_n = \begin{cases} 0 & \text{if } AIC_0 > AIC_1 \\ 1 & \text{if } AIC_1 > AIC_0. \end{cases}$$

   (a) Suppose that $\mathcal{M}_0$ is the true model, i.e. $\theta = 0$. Find

$$\lim_{n \to \infty} \mathbb{P}\left(J_n = 0\right).$$

   Now compute $\lim_{n\to\infty} \mathbb{P}\left(J_n = 0\right)$ when $\theta \neq 0$.

(b) The fact that $\lim_{n\to\infty} \mathbb{P}(J_n = 0) \neq 1$ when $\theta = 0$ is why some people say that AIC "overfits." But this is not quite true as we shall now see. Let $\phi_\theta(x)$ denote a Normal density function with mean θ and variance 1. Define

$$\widehat{f}_n(x) = \begin{cases} \phi_0(x) & \text{if } J_n = 0 \\ \phi_{\widehat{\theta}}(x) & \text{if } J_n = 1. \end{cases}$$

If $\theta = 0$, show that $D(\phi_0, \widehat{f}_n) \xrightarrow{p} 0$ as $n \to \infty$ where

$$D(f, g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

is the Kullback-Leibler distance. Show also that $D(\phi_\theta, \widehat{f}_n) \xrightarrow{p} 0$ if $\theta \neq 0$. Hence, AIC consistently estimates the true density even if it "overshoots" the correct model.

(c) Repeat this analysis for BIC which is the log-likelihood minus $(p/2)\log n$ where p is the number of parameters and n is sample size.

10. In this question we take a closer look at prediction intervals. Let $\theta = \beta_0 + \beta_1 X_*$ and let $\widehat{\theta} = \widehat{\beta}_0 + \widehat{\beta}_1 X_*$. Thus, $\widehat{Y}_* = \widehat{\theta}$ while $Y_* = \theta + \epsilon$. Now, $\widehat{\theta} \approx N(\theta, \mathsf{se}^2)$ where

$$\mathsf{se}^2 = \mathbb{V}(\widehat{\theta}) = \mathbb{V}(\widehat{\beta}_0 + \widehat{\beta}_1 x_*).$$

Note that $\mathbb{V}(\widehat{\theta})$ is the same as $\mathbb{V}(\widehat{Y}_*)$. Now, $\widehat{\theta} \pm 2\sqrt{\mathbb{V}(\widehat{\theta})}$ is an approximate 95 percent confidence interval for $\theta = \beta_0 + \beta_1 x_*$ using the usual argument for a confidence interval. But, as you shall now show, it is not a valid confidence interval for Y_*.

(a) Let $s = \sqrt{\mathbb{V}(\widehat{Y}_*)}$. Show that

$$\begin{aligned} \mathbb{P}(\widehat{Y}_* - 2s < Y_* < \widehat{Y}_* + 2s) &\approx \mathbb{P}\left(-2 < N\left(0, 1 + \frac{\sigma^2}{s^2}\right) < 2\right) \\ &\neq 0.95. \end{aligned}$$

(b) The problem is that the quantity of interest Y_* is equal to a parameter θ plus a random variable. We can fix this by defining

$$\xi_n^2 = \mathbb{V}(\widehat{Y}_*) + \sigma^2 = \left[\frac{\sum_i (x_i - x_*)^2}{n\sum_i (x_i - \overline{x})^2} + 1\right]\sigma^2.$$

In practice, we substitute $\widehat{\sigma}$ for σ and we denote the resulting quantity by $\widehat{\xi}_n$. Now consider the interval $\widehat{Y}_* \pm 2\,\widehat{\xi}_n$. Show that

$$\mathbb{P}(\widehat{Y}_* - 2\widehat{\xi}_n < Y_* < \widehat{Y}_* + 2\widehat{\xi}_n) \approx \mathbb{P}(-2 < N(0,1) < 2) \approx 0.95.$$

11. Get the Coronary Risk-Factor Study (CORIS) data from the book web site. Use backward stepwise logistic regression based on AIC to select a model. Summarize your results.