# 12 Linear Regression

Consider data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Our goal is to study the relationship between the *response* $Y$ and the *covariate* $X$. Sometimes $X$ is called a *feature* or a *predictor*. One way to study the relationship between $X$ and $Y$ is to estimate the *regression function* $s(x) = E(Y|X = x)$. A related problem is *prediction* in which we try to predict a new $Y$ based on its covariate value $X$. When $Y \in \{0, 1\}$, prediction is called *classification*.

## 12.1 Introduction to Linear Regression

We will begin with the special case called *linear regression* where we assume that

$$s(x) = E(Y|X = x) = \beta_0 + \beta_1 x.$$

Since we are interested in the mean of $Y$ given $x$, we shall now treat $x_1, \ldots, x_n$ as fixed numbers. Let $\epsilon_i = Y_i - s(x_i)$. Then,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent and $E(\epsilon_i) = 0$.

Given estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ define the *fitted line* by

$$\widehat{s}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

The *predicted values* or *fitted values* are $\widehat{Y}_i = \widehat{s}(x_i)$. The *residuals* are defined to be

$$r_i = Y_i - \widehat{Y}_i = Y_i - [\widehat{\beta}_0 + \widehat{\beta}_1 x_i].$$

Let $Q = \sum_i r_i^2$ be the sum of the squared residuals. The *least squares estimates* are the values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize $Q$. We can find these by solving

$$\frac{\partial Q}{\partial \beta_0} = 0 \qquad \frac{\partial Q}{\partial \beta_1} = 0.$$

The solution is

$$\widehat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_i (x_i - \overline{x})^2} \quad \text{and} \quad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}.$$

**EXAMPLE 12.1 (The 2001 Presidential Election.)** *Figure 12.1 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*

$$\begin{aligned}\widehat{\beta}_0 &= 66.0991 \quad se(\widehat{\beta}_0) = 17.2926 \\ \widehat{\beta}_1 &= 00.0035 \quad se(\widehat{\beta}_1) = 0.0002\end{aligned}$$

*so the prediction line is*

$$\text{Buchanon} = 66.0991 + .0035\,\text{Bush}.$$

*(We will see shortly how to compute the standard errors.) I also plotted the residuals. Linear regression works well when the residuals behave like random normal numbers. Based on the residual plot, this is not the case in this example. I repeated the analysis replacing votes with log(votes) giving*

$$\begin{aligned}\widehat{\beta}_0 &= -2.3298 \quad se(\widehat{\beta}_0) = .3529 \\ \widehat{\beta}_1 &= 0.730300 \quad se(\widehat{\beta}_1) = 0.0358.\end{aligned}$$

*This gives the fit*

$$\log(\text{Buchanon}) = -2.3298 + .7303\ \log(\text{Bush}).$$

*The residuals look much healthier.*

*Later, we shall address two interesting questions: (1) how do we see if Palm Beach County has a statistically plausible outcome? (2) how do we do this problem nonparametrically?*

## 12.2 Least Squares and Maximum Likelihood

The linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i$ are independent and $E(\epsilon_i) = 0$. Let's now make the stronger assumption that $\epsilon_i \sim N(0, \sigma^2)$. In other words, we are assuming that

$$Y_i \sim N(\mu_i, \sigma_i^2)$$

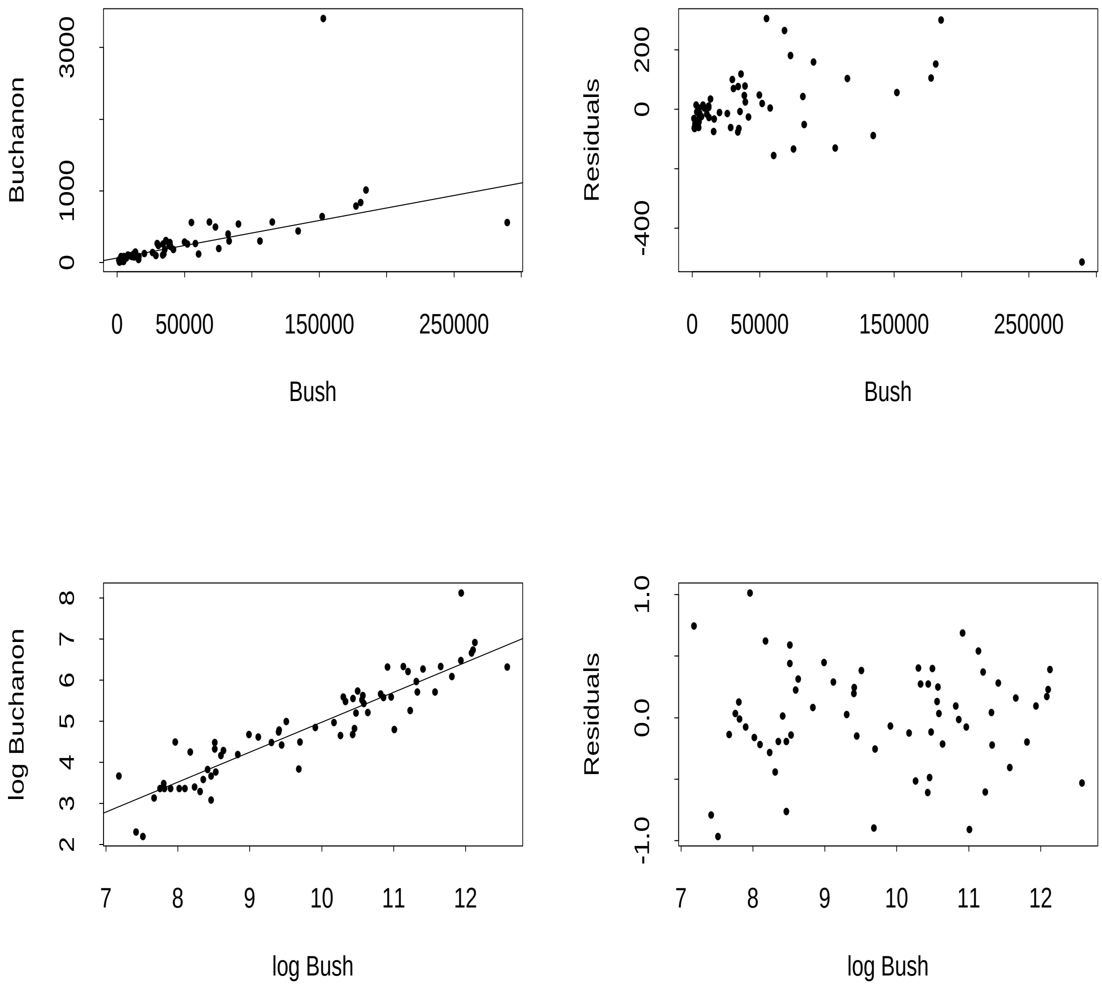


Figure 12.1. Voting Data for Election 2000.

where $\mu_i = \beta_0 + \beta_1 x_i$. Let $f(y_i|x_i;\beta_0,\beta_1,\sigma)$ denote this Normal density. The likelihood function is

$$\mathcal{L}(\beta_0,\beta_1,\sigma) = \prod_i f(y_i|x_i;\beta_0,\beta_1,\sigma) \propto \sigma^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (Y_i - \mu_i)^2\right\}$$

and the log-likelihood is

$$\ell(\beta_0,\beta_1,\sigma) = -\frac{n}{2}\log\sigma - \frac{\sum_i [Y_i - (\beta_0+\beta_1 x_i)]^2}{2\sigma^2}.$$

To find the mle of $(\beta_0,\beta_1)$ we maximize $\ell(\beta_0,\beta_1,\sigma)$ which is the same is minimizing $\sum_i [Y_i - (\beta_0 + \beta_1 x_i)]^2$. This gives the least squares estimates. Therefore, under this model, the least squares method and the maximum likelihood method are identical. We can also maximize $\ell(\beta_0,\beta_1,\sigma)$ over $\sigma$. This yields

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i r_i^2$$

where $r_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$ is the $i^{th}$ residual. Some people prefer an unbiased estimator which is

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_i r_i^2.$$

## 12.3 Standard Errors of the Least Squares Estimators

The estimator $\widehat{\beta}_1$ can be written as

$$\widehat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_i (x_i - \overline{x})^2} = \frac{\sum_i (x_i - \overline{x})Y_i}{\sum_i (x_i - \overline{x})^2} = \sum_i w_i Y_i$$

where

$$w_i = \frac{(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2}.$$

Note that $\sum_i w_i = 0$. Hence,

$$\begin{aligned} E(\widehat{\beta}_1) &= \sum_i w_i E(Y_i) \\ &= \sum_i w_i (\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_i w_i + \beta_1 \sum_i x_i w_i \end{aligned}$$

$$
\begin{aligned}
&= \beta_1 \sum_i x_i w_i \\
&= \beta_1 \frac{\sum_i x_i(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2} \\
&= \beta_1 \frac{\sum_i (x_i - \overline{x})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2} \\
&= \beta_1.
\end{aligned}
$$

So $\widehat{\beta}_1$ is an unbiased estimator of $\beta_1$. Also,

$$\mathrm{Var}(\widehat{\beta}_1) = \sum_i w_i^2 Var(Y_i) = \sigma^2 \sum_i w_i^2 = \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2}.$$

Thus $\mathrm{Var}(\widehat{\beta}_1) \to 0$ if $\sum_i (x_i - \overline{x})^2 \to \infty$. We shall assume that $\sum_i (x_i - \overline{x})^2 \to \infty$ from now on. Hence,

$$MSE(\widehat{\beta}_1) = \mathrm{bias}^2 + \mathrm{Variance} = \mathrm{Variance} \to 0$$

as $n \to \infty$ and therefore $\widehat{\beta}_1 \xrightarrow{p} \beta_1$. The central limit theorem also applies and so

$$\widehat{\beta}_1 \approx N(\beta_1, \widehat{se}^2)$$

where

$$\widehat{se}^2 = Var(\widehat{\beta}_1) = \frac{\widehat{\sigma}^2}{\sum_i (x_i - \overline{x})^2}.$$

A $1-\alpha$ per cent confidence interval for $\beta_1$ is $\widehat{\beta}_1 \pm z_{\alpha/2} \widehat{se}(\widehat{\beta}_1)$. Using similar arguments we can show that $E(\widehat{\beta}_0) = \beta_0$ and that

$$\widehat{se}^2(\widehat{\beta}_0) = \frac{\widehat{\sigma}^2 \sum_i x_i^2}{n \sum_i (x_i - \overline{x})^2}.$$

Later we shall also need $Cov(\widehat{\beta}_0, \widehat{\beta}_1)$. The estimate of this is given by

$$\widehat{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{\overline{x}\, \widehat{\sigma}^2}{\sum_i (x_i - \overline{x})^2}.$$

For the election data, on the log scale, a 95 per cent confidence interval is $.7303 \pm 2(.0358) = [.66, .80]$. The fact that the interval excludes 0 is regarded as evidence that the true slope is not 0.

## 12.4 Prediction From Regression

Consider predicting a new outcome $Y_*$ given $x_*$, before seeing the outcome. Since we haven't seen $Y_*$ we can think if it as a parameter. Our estimate of $Y_*$ is

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

To compute a confidence interval we need the variance of $\widehat{Y}_*$. This is

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

We previously computed all the terms that go into this formula.

**EXAMPLE 12.2 (Election Data Revisited.)** *On the log-scale, our linear regression gives the following prediction equation:* $log(Buchanon) = -2.3298 + .7303 log(Bush)$. *In Palm Beach, Bush had 152954 votes and Buchanan had 3467 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 is "significantly" bigger?*

We can answer the last question by considering a confidence interval for $Y_*$. But **the usual confidence interval does not work for prediction.** In other words, $\widehat{Y}_* \pm 2\sqrt{\mathrm{Var}(\widehat{Y}_*)}$ is **not** a 95 per cent interval. The reason is that the quantity we want to estimate, $Y_*$, is not a fixed parameter, it is a random variable.

To understand this point better, 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, $\widehat{\theta} \approx N(\theta, se^2)$ where

$$se^2 = \mathrm{Var}(\widehat{\theta}) = \mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_*).$$

Note that $Var(\widehat{\theta})$ is the same as $Var(\widehat{Y}_*)$. Now, $\widehat{\theta} \pm 2\sqrt{Var(\widehat{\theta})}$ is a 95 per cent confidence interval for $\theta = \beta_0 + \beta_1 x_*$ using the usual argument for a confidence interval. It is not a valid confidence interval for $Y_*$. To see why, let's compute the probability that $\widehat{Y}_* \pm 2\sqrt{Var(\widehat{Y}_*)}$ contains $Y_*$. Let $s = \sqrt{Var(\widehat{Y}_*)}$. Then,

$$\begin{aligned} P(\widehat{Y}_* - 2s < Y_* < \widehat{Y}_* + 2s) &= P\left(-2 < \frac{\widehat{Y}_* - Y_*}{s} < 2\right) \\ &= P\left(-2 < \frac{\widehat{\theta} - \theta - \epsilon}{s} < 2\right) \end{aligned}$$

$$
\begin{aligned}
&= P\left(-2 < \frac{\widehat{\theta} - \theta}{s} - \frac{\epsilon}{s} < 2\right) \\
&\approx P\left(-2 < N(0,1) - N\left(0, \frac{\sigma^2}{s^2}\right) < 2\right) \\
&= P\left(-2 < N\left(0, 1 + \frac{\sigma^2}{s^2}\right) < 2\right) \\
&\neq .95.
\end{aligned}
$$

The problem is that the quantity of interest $Y_*$ is equal to a parameter $\theta$ plus a random variable. We can fix this by defining

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

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

$$
\begin{aligned}
P(\widehat{Y}_* - 2\widehat{\xi}_n < Y_* < \widehat{Y}_* + 2\widehat{\xi}_n) &= P\left(-2 < \frac{\widehat{Y}_* - Y_*}{\widehat{\xi}_n} < 2\right) \\
&= P\left(-2 < \frac{\widehat{\theta} - \theta - \epsilon}{\widehat{\xi}_n} < 2\right) \\
&\approx P\left(-2 < \frac{N(0, s^2 + \sigma^2)}{\widehat{\xi}_n} < 2\right) \\
&\approx P\left(-2 < \frac{N(0, s^2 + \sigma^2)}{\xi_n} < 2\right) \\
&= P\left(-2 < N(0,1) < 2\right) = .95.
\end{aligned}
$$

Of course, a $1 - \alpha$ interval is given by $\widehat{Y}_* \pm z_{\alpha/2}\widehat{\xi}_n$.

**EXAMPLE 12.3 (Election Data Again.)** *In our example, we find that* $\widehat{\xi}_n = .093775$ *and the 95 per cent 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 3467. This is not a definitive analysis of this problem but hopefully it gives you a flavor for linear regression.*

## 12.5 The Regression Fallacy

In the 1880's Galton noticed that tall men tended to have sons shorter than themselves and short men tended to have sons taller than themselves. he called this "regression toward the mean" and this is where the term regression comes from. This seems to suggest that with each generation, men get closer and closer to the mean height. Eventually, everyone should be the same height! The same thing happens in sports; people have a good first year and then do less well the second year, often called a sophomore jinx.

Let's take a look at this. I will show you that the conditional mean of $Y$ given $X$ can be closer to the overall mean without the marginal distribution of $Y$ becoming more concentrated. Consider the following example from DeGroot. The scores of students on two exams are as shown in Table 1. For simplicity we pretend there are only three possible scores.

| | Test 1 | | | |
|---|---|---|---|---|
| Test 2 | 30 | 60 | 90 | |
| 90 | 0 | $\frac{1}{9}$ | $\frac{2}{9}$ | $\frac{1}{3}$ |
| 60 | $\frac{1}{9}$ | $\frac{1}{9}$ | $\frac{1}{9}$ | $\frac{1}{3}$ |
| 30 | $\frac{2}{9}$ | $\frac{1}{9}$ | 0 | $\frac{1}{3}$ |
| | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ | |

Note that the marginal distribution of test score $Y$ on exam 2 is the same as the marginal distribution of test score $X$ on exam 1. So the distribution of test scores has not become more concentrated on the second exam. Note that $\mu = E(Y) = E(X) = 60$. However, $E(Y|X = 90) = (2/3)90 + (1/3)60 = 80$ and $E(Y|X = 30) = (2/3)30 + (1/3)60 = 40$. Hence, $|E(Y|X = x) - \mu| < |x - \mu|$. The conditional mean is closer to the average than $X$ is but the distribution is not becoming more concentrated.