Linear Regression

April 15, 2008

1 Simple linear regression

We now consider two dimensional data. The values of the first variable $x_1, x_2, \ldots, x_n$ are assumed known and in an experiment are often set by the experimenter. These variables are called the **explanatory variable** and in a two dimensional **scatterplot** of the data are values on the horizontal axis. The values $y_1, y_2 \ldots, y_m$, taken from observations with input $x_1, x_2, \ldots, x_n$ are called the **response variable** and are values on the vertical axis. In **linear regression**, the response variable linearly related to the explanatory variable, but is subject to error.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

The errors $\{\epsilon_i; 1 \leq i \leq n\}$ are assumed to be independent mean zero random variable. The most common assumption is that they are normal with an unknown variance.

Thus, simple linear regression is a three parameter model

$$\Theta = \{(\beta_0, \beta_1, \sigma^2); \beta_0, \beta_1 \in \mathbb{R}, \sigma^2 \geq 0\}.$$

The likelihood,

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2 | \mathbf{x}, \mathbf{y}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_1 - \beta_0 - \beta_1 x_1)^2\right) \cdots \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_n - \beta_0 - \beta_1 x_n)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right) \end{aligned}$$

or

$$\ln L(\beta_0, \beta_1, \sigma^2 | \mathbf{x}, \mathbf{y}) = -\frac{n}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

The **least square regression line** uses the maximum likelihood estimates $\hat{\beta}_0, \hat{\beta}_1$ to make a **prediction** $\widehat{y}_i$ based on x_i, the value of the explanatory variable.

$$\widehat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

to maximize the likelihood,we take derivatives.

$$\frac{\partial}{\partial \beta_0} \ln L(\beta_0, \beta_1, \sigma^2 | \mathbf{x}, \mathbf{y}) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-1)$$

and at the maximum likelihood estimate,

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \quad \bar{y} = \hat{\beta}_0 + \beta_1 \bar{x}. \tag{1}$$

Next,

$$\frac{\partial}{\partial \beta_1} \ln L(\beta_0, \beta_1, \sigma^2 | \mathbf{x}, \mathbf{y}) = -\frac{1}{\sigma^2} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-x_i)$$

and

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_1 = 0, \quad \overline{xy} = \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \overline{x^2}. \tag{2}$$

Now, multiply equation (1) by $\bar{x}$ and subtract from equation (2) to obtain

$$\overline{xy} - \bar{x}\bar{y} = \beta_1(\overline{x^2} - (\bar{x})^2, \quad \mathrm{Cov}(x, y) = \hat{\beta}_1 \mathrm{Var}(x)$$

Write s_x^2 and s_y^2 for the sample variance of $\mathbf{x}$ and $\mathbf{y}$. Recall that the correlation

$$r = \frac{\mathrm{Cov}(x, y)}{s_x s_y}.$$

Then,

$$\hat{\beta}_1 = r\frac{s_y}{s_x}.$$

Equation (1) tells us that the center of mass$(\bar{x}, \bar{y})$ lies on the regression line. Consequently, we can write the regression line in point slope form as

$$\hat{y} - \bar{y} = r\frac{s_y}{s_x}(x - \bar{x}), \quad \text{or} \quad \frac{\hat{y} - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x}$$

In words, if we standardize the explanatory and response variables, the the regression line contains the origin and has slope equal to the correlation of the explanatory and response variables.

For the maximum likelihood for the variance of the ϵ_i, we have

$$\frac{\partial}{\partial \sigma^2} \ln L(\beta_0, \beta_1, \sigma^2 | \mathbf{x}, \mathbf{y}) = -\frac{n}{2\sigma^2} + \frac{1}{(2\sigma^2)^2} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

and, therefore, the maximum likelihood estimator.

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n} \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

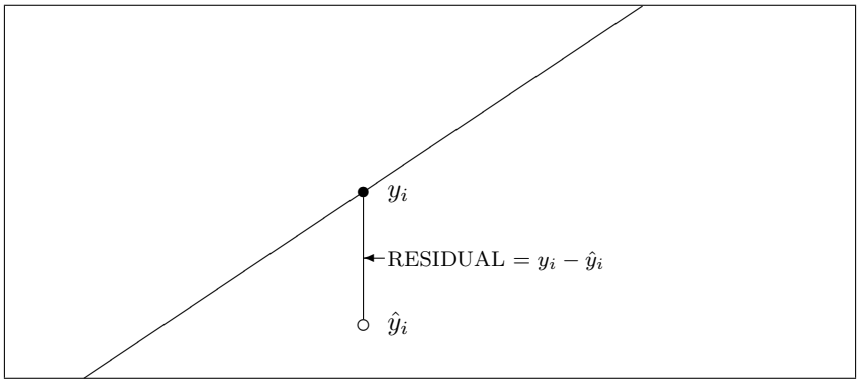
The **unbiased estimator**

$$\widehat{\sigma^2}_U = \frac{1}{n-2} \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

For a given data entry, the difference

$$e_i = y_i - \widehat{y_i} = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

is called the **residual**.



We generate 50 pairs of independent standard normal random variable with correlation $r = 0.9$.

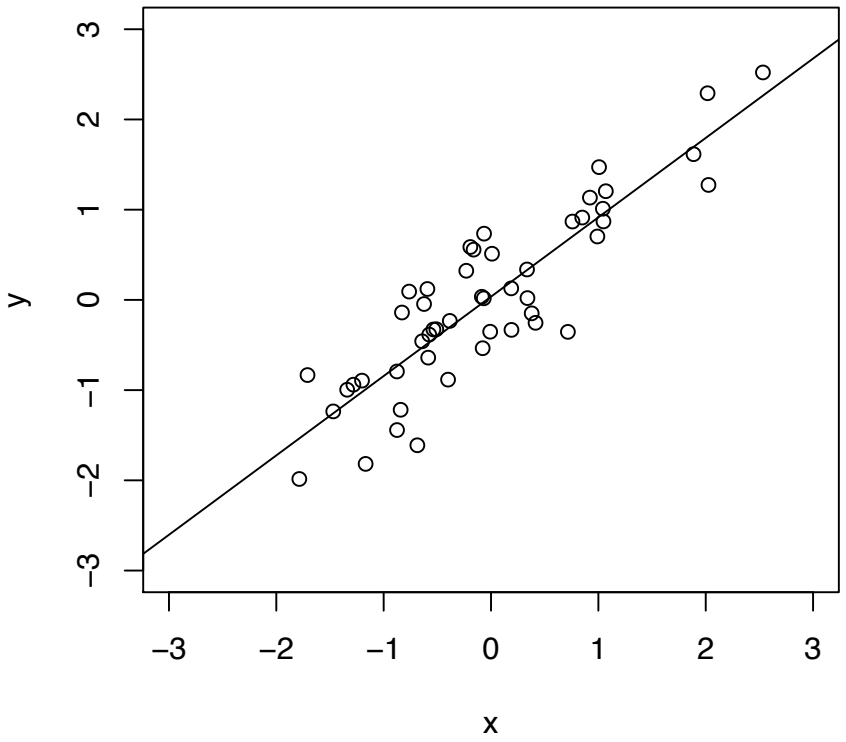
```
> x<-rnorm(50)
> z<-rnorm(50)
> r = 0.9
> y<-r*x +sqrt(1-r^2)*z
> plot(x,y,xlim=c(-3,3),ylim=c(-3,3))
> myline.fit <- lm(y ~ x)
> summary(myline.fit)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04313 -0.36433  0.06956  0.28016  0.75557

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03577    0.06672   0.536    0.594
x            0.87970    0.06805  12.927   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4715 on 48 degrees of freedom
Multiple R-Squared: 0.7769,Adjusted R-squared: 0.7722
F-statistic: 167.1 on 1 and 48 DF,  p-value: < 2.2e-16
```



2 Multiple Linear Regression

Now, we consider the case in which the explanatory variables $\mathbf{x}_1, \cdots, \mathbf{x}_n$ where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$.

The response variable linearly related to the explanatory variable, but is subject to error.

$$y_i = \beta_0 + x_{ii}\beta_1 + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

The errors $\{\epsilon_i; 1 \leq i \leq n\}$ are again assumed to be independent mean zero normal random variables with an unknown variance, σ^2. Now, regression is a $k+2$ parameter model

$$\Theta = \{(\beta, \sigma^2); \beta \in \mathbb{R}^{k+1}, \sigma^2 \geq 0\}, \quad \beta = (\beta_0, \beta_1, \ldots, \beta_k) \tag{3}.$$

The likelihood,

$$L(\beta, \sigma^2 | \mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - x_{ii}\beta_1 - \cdots - \beta_k x_{ik})^2 \right).$$

We can write this in matrix notation. Set

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}.$$

Then

$$\ln L(\beta, \sigma^2|\mathbf{x}, \mathbf{y}) = -\frac{n}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)$$

and

$$\nabla_\beta \ln L(\beta, \sigma^2|\mathbf{x}, \mathbf{y}) = (\mathbf{y} - X\beta)^T X.$$

To find the maximim likelihood estimator $\hat{\beta}$

$$(\mathbf{y} - X\hat{\beta})^T X = 0, \quad \mathbf{y}X = \hat{\beta}^T X^T X.$$

If $X^T X$ is invertible, then we take the transpose to obtain

$$X^T X\hat{\beta} = X^T\mathbf{y}, \quad \hat{\beta} = (X^T X)^{-1}X^T\mathbf{y}.$$

The maximum likelihood for the variance of the ϵ_i is similar as the one-dimensional case.

$$\frac{\partial}{\partial\sigma^2} \ln L(\beta, \sigma^2|\mathbf{x}, \mathbf{y}) = -\frac{n}{2\sigma^2} + \frac{1}{(2\sigma^2)^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

and, therefore, the maximum likelihood estimator.

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta).$$

The **unbiased estimator**

$$\widehat{\sigma^2}_U = \frac{1}{n-k-1}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta).$$

3 Properties of least square estimates

Property 1. The likelihood function is maximized at $\hat{\beta}$. Equivalently, the least square $(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)$ is minimized at this point.

Property 2. $\hat{\beta}$ is an unbiased estimator of β.

From equation (3), we find that the expected value

$$E_{\beta,\sigma^2}Y = X\beta.$$

Then

$$E_{\beta,\sigma^2}\hat{\beta} = E_{\beta,\sigma^2}[(X^T X)^{-1}X^T Y] = (X^T X)^{-1}X^T E_{\beta,\sigma^2}Y = (X^T X)^{-1}X^T X\beta = \beta.$$

For a vector valued random variable, $\xi = (\xi_1, \ldots, \xi_n)$, define the **covariance matrix** $\mathrm{Cov}(\xi)$ to be the $n \times n$ matrix having entries

$$\mathrm{Cov}(\xi)_{ij} = \mathrm{Cov}(\xi_i, \xi_j).$$

If ξ has mean vector μ, we can write this in matrix form as

$$\mathrm{Cov}(\xi) = E(\xi - \mu)(\xi - \mu)^T.$$

Check that for any $n \times n$ matrix A.

$$\mathrm{Cov}(A\xi) = A\mathrm{Cov}(\xi)A^T.$$

We also have that $\hat{\beta}_i$ is a uniformly minimum variance unbiased estimator.

Property 3. $\mathrm{Cov}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$.

Again, returning to equation (3), we see that

$$\mathrm{Cov}(Y_i, Y_j) = \begin{cases} \sigma^2 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Consequently, $\mathrm{Cov}(Y) = \sigma^2 I$, where I is the $n \times n$ identity matrix.

Now

$$\begin{aligned} \mathrm{Cov}(\hat{\beta}) &= \mathrm{Cov}((X^TX)^{-1}X^TY) = (X^TX)^{-1}X^T\mathrm{Cov}(Y)X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}. \end{aligned}$$