Stat 20: Discussion at section

B. M. Bolstad, bolstad@stat.berkeley.edu

Nov 10, 2003

The Homoskedastic Normal Linear model

We will now assume that Y is normally distributed with mean $\mu(\mathbf{x})$ and standard deviation σ. Note that $\mathbf{x} = (x_1, \ldots, x_m)$ is a set of known factor variables. In the context of linear regression we write $\mu(\mathbf{x})$ (this is referred to as the regression function) as a linear combination of functions of $\mathbf{x}$. In particular

$$\mu(\mathbf{x}) = \beta_0 g_0(\mathbf{x}) + \beta_1 g_1(\mathbf{x}) + \cdots + \beta_p g_p(\mathbf{x})$$

where $g_0(\mathbf{x}), \ldots, g_p(\mathbf{x})$ are referred to as basis functions. For example suppose that $\mathbf{x} = (x_1, x_2)$ then some possible basis functions are $g_j(\mathbf{x}) = 1, g_j(\mathbf{x}) = x_1, g_j(\mathbf{x}) = x_1^2, g_j(\mathbf{x}) = x_1 x_2, g_j(\mathbf{x}) = \sin(x_2), \ldots$. In practice the basis functions you will use will depend on the model that you are interested in fitting to your data. Sometimes the following notation is used to represent the model

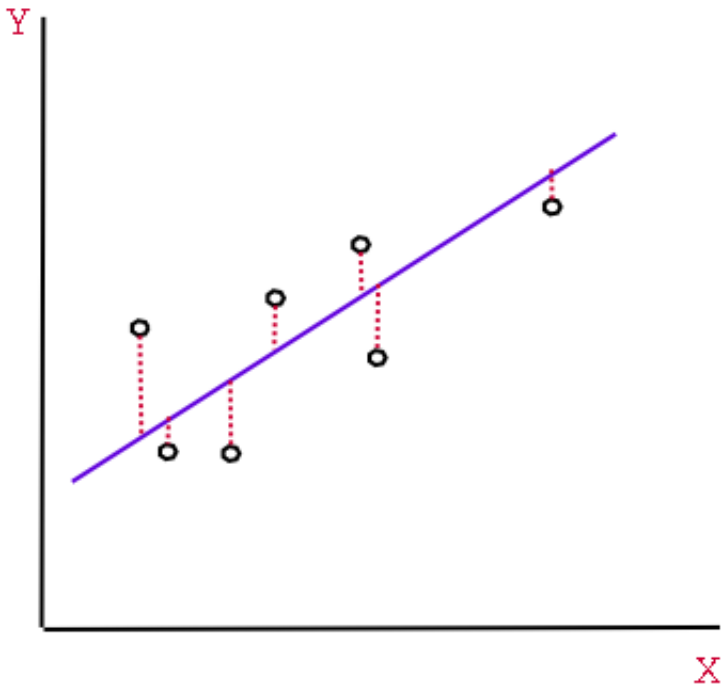
$$Y = \mu(\mathbf{x}) + \epsilon$$

where ϵ is a random error term with normal distribution with mean 0 and standard deviation σ.

In practice we know Y_i and $\mathbf{x}_i$ for each observation, $i = 1, \ldots, n$ but we do not know the value of the parameters $\beta_0, \beta_1, \ldots, \beta_p$. How do we estimate these parameters? The method that we will use is called *Least Squares*. In particular we seek the values of $\beta_0, \ldots, \beta_p$ which minimizes the Sum of Square Errors (SSE).

$$\text{SSE} = \sum_{i=1}^{n} \left(Y_i - \hat{\mu}(\mathbf{x})\right)^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 g_0(\mathbf{x}) - \beta_1 g_1(\mathbf{x}) - \cdots - \beta_p g_p(\mathbf{x})\right)^2$$

The following diagram demonstrates what we are seeking to minimize (for the special case of a line in 2 dimensions). the blue line is the regression line that we wish to fit and the vertical dotted lines are the distances between our observed Y and the line. Our goal is to find the values of β_0, β_1 which parameterize the line which has the minimum sum of the squared distances between itself and the Y's.



The least squares process is as follows: First we differentiate the SSE with respect to each of $\beta_0, \ldots, \beta_p$ and set each of these equations equal to zero. This gives us a system of $p+1$ linear equations in $p+1$ unknowns. This system of equations is referred to as the *Normal equations*. Solving the system of equations for $\beta_0, \ldots, \beta_p$ gives us estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$.

Lets consider the simple linear regression model

$$\mu(x) = \beta_0 + \beta_1 x$$

and so the SSE is

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

differentiating the SSE with respect to β_0 and setting equal to 0 gives

$$\sum_{i=1}^{n} 2\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-1) = 0$$

with a little simplification this becomes

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \tag{1}$$

similarly taking the derivative of the SSE with respect to β_1 and setting equal to 0 gives

$$\sum_{i=1}^{n} 2\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-x_i) = 0$$

with a little simplification this becomes

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{2}$$

rearranging (1) above gives $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, then substituting this into (2) gives

$$\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

solving for $\hat{\beta}_1$ gives

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

As you can see it is a lot of work to solve these equations, and with even more terms in our regression model we would have even more equations to solve.

We say that two basis functions are orthogonal for the data if

$$\sum_{i=1}^{n} g_j\left(\mathbf{x}_i\right) g_k\left(\mathbf{x}_i\right) = 0$$

When the basis functions are all *orthogonal* for the data there is a simplified formula that we can use to estimate the parameters

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} g_j\left(\mathbf{x}_i\right) y_i}{\sum_{i=1}^{n} g_j^2\left(\mathbf{x}_i\right)} \text{ for } j = 0, \ldots, p$$

Question 1

Suppose that X is the change in pH (from Neutral) and Y is crop yield

Y	X
13.5	-0.8
13.2	-0.4
13.7	-0.2
13.7	0.0
13.9	0.2
14.1	0.4
14.3	0.8

We are interested in the regression model $\mu(x) = \beta_0 + \beta_1 x$.

(a) Are the basis functions 1 and x orthogonal?

(b) Fit the regression line

(c) Interpret β_0,β_1 for this model

(d) Would it be safe to use the fitted regression line to estimate the crop yield at x=2?

Answer:

(a) First we need to check that $\sum i = 1^7(1)(x_i) = 0$. If this is true then 1 and x are an orthogonal basis.

$$\sum_{i=1}^{7}(1)(x_i) = \sum_{i=1}^{7} x_i = -0.8 + -0.4 + -0.2 + 0 + 0.2 + 0.4 + 0.8 = 0$$

so we can conclude that the basis functions 1 and x are orthogonal.

(b) Since 1 and x are orthogonal. We may use the formula above to estimate each of the two parameters

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{7}(1)y_i}{\sum_{i=1}^{7}(1)^2} = \frac{\sum_{i=1}^{7} y_i}{n} = \bar{y}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{7} x_i y_i}{\sum_{i=1}^{7} x_i^2}$$

Computing the summary statistics

$$\sum_{i=1}^{7} y_i = 96.4$$

$$\sum_{i=1}^{7} x_i^2 = 1.68$$

$$\sum_{i=1}^{7} x_i y_i = 1.04$$

and so

$$\hat{\beta}_0 = 96.4/7 = 13.77$$

$$\hat{\beta}_1 = 1.04/1.68 = 0.6190$$

and so our fitted regression line is

$$y = 13.77 + 0.6190x$$

(c) $\beta_0 = \bar{y}$ so β_0 is the mean crop yield (Alternatively you could view it as the crop yield at x=0).

Suppose we consider increasing x by 1 (for some value of x). So initially we have

$$y_{\text{init}} = \beta_0 + \beta_1 x$$

and after increasing x by 1 we have

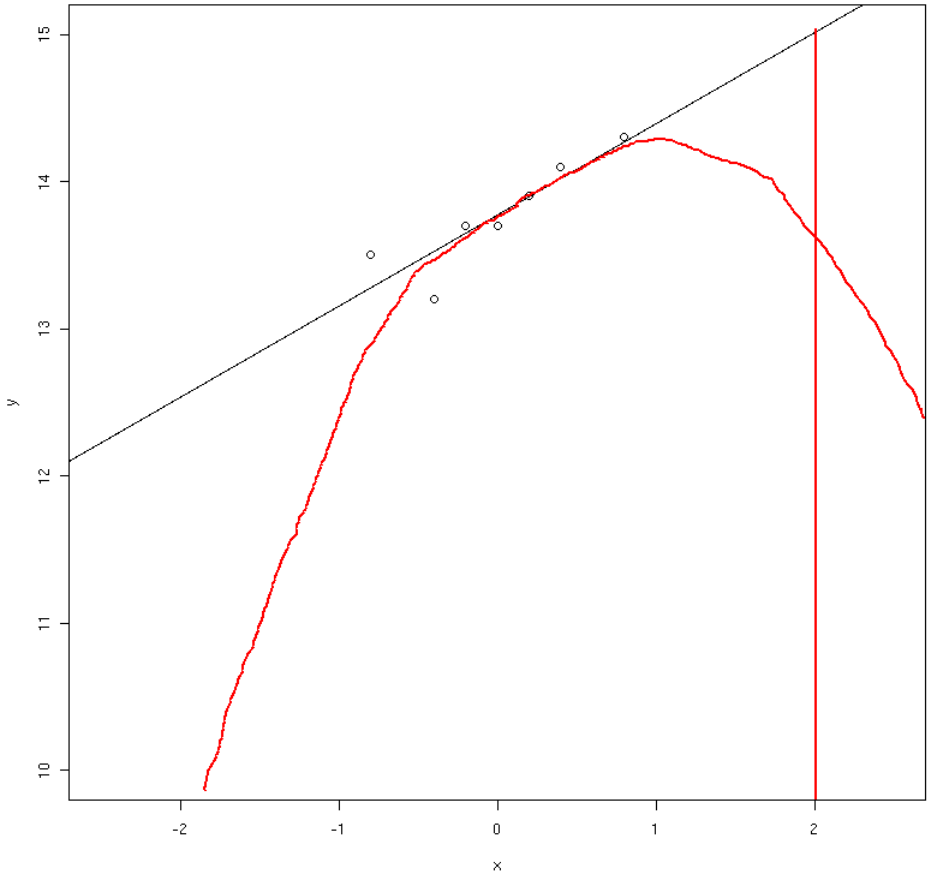
$$y_{\text{final}} = \beta_0 + \beta_1(x+1)$$

consider $y_{\text{final}} - y_{\text{init}} = (\beta_0 + \beta_1(x+1)) - (\beta_0 + \beta_1 x) = \beta_1$ and so we can interpret β_1 as the change in crop yield if we increase x by 1.

(d) If $x = 2$ then our fitted regression model would estimate the crop yield as

$$\hat{y} = 13.77 + 0.6190(2) = 15.008$$

is this going to be problematic? The answer is yes. The following diagram illustrates why



The red curve indicates the "truth" and so what we see is that while the relationship between the change in concentration is linear in the range of the observed data it is clearly not linear. It is dangerous to use a fitted regression line to predict outside the range of the observed data (this process is called extrapolation). Note that we would have no difficulties predicting yield at say $x = 0.1$.

Question 2

Consider the following experimental data

Y	X_1	X_2
21	-1	-1
14	0	-1
13	1	-1
23	-1	0
20	0	0
11	+1	0
31	-1	1
20	0	1
27	1	1

and further consider the following regression model $\mu(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

(a) Show that 1, x_1, x_2 are an orthogonal basis

(b) Determine the least squares estimates

(c) Interpret β_0,β_1 and β_2

Answer:

(a) $\sum_{i=1}^{n} 1(x_{i2}) = \sum_{i=1}^{9} x_{i1} = -3 + 3 = 0$ and $\sum_{i=1}^{n} 1(x_{i2}) = \sum_{i=1}^{9} x_{i2} = -3 + 3 = 0$ and finally $\sum_{i=1}^{n} x_{i1}x_{i2} = 1 + 0 - 1 + 0 + 0 + 0 - 1 + 0 + 1 = 0$ and so we can conclude 1,x_1 and x_2 form an orthogonal basis.

(b) Since 1,x_1 and x_2 are an orthogonal basis we can use the formula to get each of the parameter estimates.

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{9} 1(y_i)}{\sum_{i=1}^{9} 1^2} = \frac{\sum_{i=1}^{9} y_i}{9} = \bar{y}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{9} x_{i1}y_i}{\sum_{i=1}^{9} x_{i1}^2}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{9} x_{i2}y_i}{\sum_{i=1}^{9} x_{i2}^2}$$

using the data we compute $\sum_{i=1}^{9} y_i = 180$, $\sum_{i=1}^{9} x_{i1}^2 = 6$, $\sum_{i=1}^{9} x_{i2}^2 = 6$, $\sum_{i=1}^{9} x_{i1} y_i = -24$ and $\sum_{i=1}^{9} x_{i2} y_i = 30$. Therefore

$$\hat{\beta}_0 = 180/9 = 20$$

$$\hat{\beta}_1 = -24/6 = -4$$

$$\hat{\beta}_2 = 30/6 = 5$$

ans so our estimated regression model is $\hat{\mu}(x_1, x_2) = 20 - 4x_1 + 5x_2$

(c) We would interpret β_0 as the mean value of Y. And β_1 would be interpreted as the change in Y is we increase x_1 by 1 and hold x_2 fixed. Similarly β_2 would be the change in Y if we increase x_2 by 1 and hold x_1 fixed.