# Stat 20: Discussion at section

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

Nov 12, 2003

This section we will see what to do when the basis is not orthogonal (and thus we cannot use the formula presented last time).

## What happens if my basis is not orthogonal?

If this is the case, then it is possible to create a new set of basis functions $\tilde{g}_0(\mathbf{x}), \ldots, \tilde{g}_p(\mathbf{x})$ that are linear combinations of the original set of basis functions $\tilde{g}_0(\mathbf{x}), \ldots, g_p(\mathbf{x})$. eg suppose $t_{ij}$ are constants for $i, j = 0, \ldots, p$. Then we may write

$$\tilde{g}_i(\mathbf{x}) = t_{i0}g_0(\mathbf{x}) + t_{i1}g_1(\mathbf{x}) + \cdots + t_{ip}g_p(\mathbf{x}) \text{ for } i = 0, \ldots p$$

### Example of transforming data

We know that $g_0(\mathbf{x}) = 1$, $g_1(\mathbf{x}) = x$ are orthogonal if and only if $0 = \sum_{i=1}^n (1)(x_i) = \sum_{i=1}^n x_i$. But what if, for our data set $\sum_{i=1}^n x_i \neq 0$. If this is the case then we can use an orthogonal basis $\tilde{g}_0(\mathbf{x}) = 1, \tilde{g}_1(\mathbf{x}) = x - \bar{x}$ then it is easy to see that $\sum_{i=1}^n (1)(x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = \sum_{i=1}^n x_i - \sum_{i=1}^n x_i = 0$. Thus the alternative basis is orthogonal. We then fit the model

$$\mu(x) = \tilde{\beta}_0 + \tilde{\beta}_1(x - \bar{x})$$

How does this related to the original model $\beta_0 + \beta_1 x$? it is easy to see that $\beta_1 = \tilde{\beta}_1$ and $\beta_0 + \beta_1 \bar{=} \tilde{\beta}_0$.

See Professor Stones supplement for an example of the transformation (and conditions) required when we regress $Y$ on $X_1$ and $X_2$ to transform to an orthogonal basis.

### An example using data

Lets return to the crop yield example we discussed in a previous section Suppose that $X$ is the pH and $Y$ is crop yield. We wish to fit $\mu(x) = \beta_0 + \beta_1 x$

| $Y$ | $X$ | $X - \bar{X}$ |
|---|---|---|
| 13.5 | 6.2 | -0.8 |
| 13.2 | 6.6 | -0.4 |
| 13.7 | 6.8 | -0.2 |
| 13.7 | 7.0 | 0.0 |
| 13.9 | 7.2 | 0.2 |
| 14.1 | 7.4 | 0.4 |
| 14.3 | 7.8 | 0.8 |

Note that $X - \bar{X}$ is the explanatory variable that we used last time. Clearly in this case $\sum x_i > 0$ and so we do not have an orthogonal basis 1 and $X$. Now, we could take the transformation and estimate the parameters to get (as we did last time) the fitted model $13.77 + 0.6190(X - \bar{X})$, but instead we will estimate the parameters directly.

Your text gives you the following formulas for estimating the parameters of this model $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = r\frac{s_y}{s_x}$ where $r$ is the correlation, $s_y$ and $s_x$ are the standard deviations of $y$ and $x$. Calculating the means

$$\bar{x} = \frac{6.2 + \cdots + 7.8}{7} = 7$$

and

$$\bar{y} = \frac{13.5 + \cdots + 14.3}{7} = 13.77$$

similarly

$$\sum_{i=1}^{n} x_i^2 = 6.2^2 + \cdots + 7.8^2 = 344.68$$

and

$$\sum_{i=1}^{n} y_i^2 = 13.5^2 + \cdots + 14.3^2 = 1328.38$$

and so

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} x_i - n\bar{x}^2}{n-1}} = \sqrt{\frac{344.68 - 7(7)^2}{6}} = 0.5292$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} y_i - n\bar{x}^2}{n-1}} = \sqrt{\frac{1328.38 - 7(13.77)^2}{6}} = 0.3684$$

Finally we want the correlation

$$r = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

It can be found that

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 1.04$$

and so

$$r = \frac{1}{6}\frac{1.04}{0.5292(0.3684)} = 0.8891$$

therefore

$$\hat{\beta}_1 = 0.8891\frac{0.3684}{0.5292} = 0.6190$$

and

$$\hat{\beta}_0 = 13.77 - 0.6190(7) = 9.44$$

and so our fitted model is

$$9.44 + 0.6190X$$

It is easy to verify that our previous claim is true ie $\beta_1 = \tilde{\beta}_1$ (0.6190 =0.6190) and $\beta_0 + \beta_1\bar{x} = \tilde{\beta}_0$ (9.44 + 0.6190(7) = 13.77).

# Some notes on hand computing correlation

Just as we saw that there is formulas for more quickly hand calculating standard deviation there are formulas which are quicker for correlation. In particular consider the numerator in the formula for correlation used above

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\bar{y}) \\
&= \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum x = 1^n x_i + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - 2n\bar{x}\bar{y} + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}
\end{aligned}$$

and so we can compute the correlation using

$$r = \frac{1}{n-1}\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{s_x s_y}$$

Note that the correlation should always be $-1 \leq r \leq 1$.

# Checking agreement in methods of estimating parameters

In the previous section we should using least squares that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

and

$$\beta_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

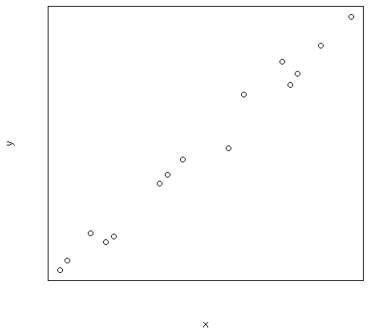
Clearly the first term agrees with what we used this time, but does the second term agree with $\hat{\beta}_1 = r\frac{s_y}{s_x}$? Lets check

$$\begin{aligned}
\hat{\beta}_1 &= r\frac{s_y}{s_x} \\
&= \frac{1}{n-1}\frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{s_x s_y}\frac{s_y}{s_x} \\
&= \frac{1}{n-1}\frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{s_x^2} \\
&= \frac{1}{n-1}\frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\frac{sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}} \\
&= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}
\end{aligned}$$

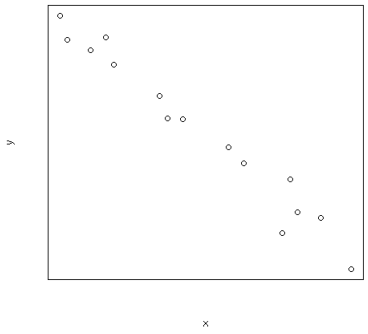
which agrees with our previous result.

# Interpreting scatterplots

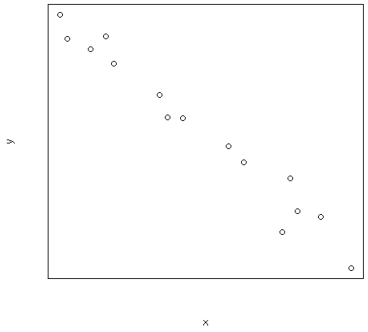
We will concentrate on interpreting the scatter plots in the content of the model $\mu(x) = \beta_0 + \beta_1 x$.



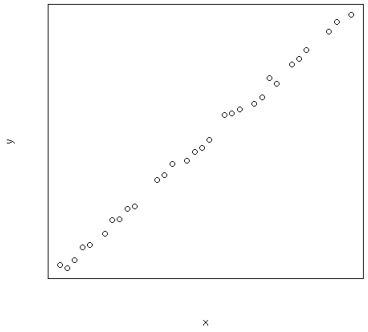

$\beta_1 > 0$, increasing (positive) slope, positive relationship between $x$ and $y$



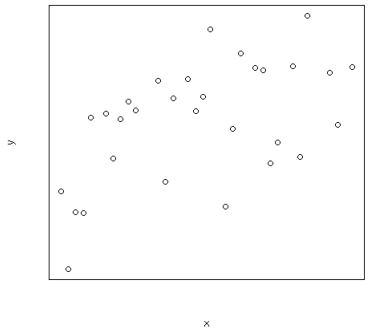

$\beta_1 < 0$, decreasing (positive) slope, negative relationship between $x$ and $y$



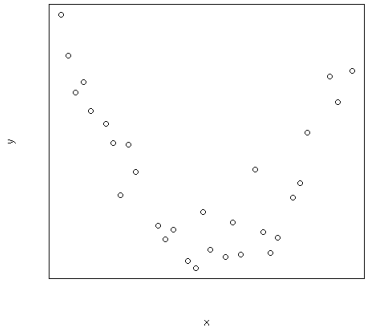

$\beta_1 = 0$, no relationship between $x$ and $y$



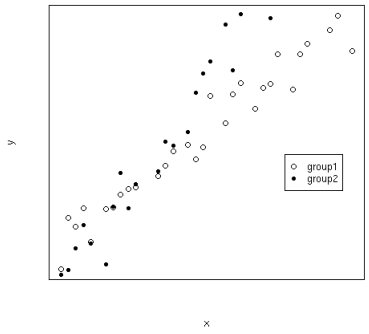

Strong relationship between $x$ and $y$. correlation between $x$ and $y$ is high



Weak relationship between $x$ and $y$. correlation between $x$ and $y$ is low



there is a non linear relationship between $x$ and $y$. A better model to fit would be $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$.




there is a different linear relationship for each of group 1 and group 2.