M358K

THE LEAST SQUARES REGRESSION LINE and R^2

I. Recall from p. 136 that “the **least squares regression line of y on x** is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.” The book (which is not written for math majors) then just gives formulas for the coefficients of this line on p. 137. As math majors, you should know why these are the right coefficients, so here is a proof, with blanks for you to fill in. I intend to go through this in class, calling on students to fill in the missing items A, B, …

Notation: The least squares line has equation y = a + bx. The data points are (x_1,y_1), ... , (x_n, y_n).

A. Draw a picture showing the line, the i^{th} data point (x_i, y_i), and the vertical distance d_i from (x_i, y_i) to the line y = ax + b.

B. Express d_i in terms of x_i, y_i, a, and b: d_i = _____________

C. Use your answer to (B) to find an expression for the sum Q of the squares of the distances of the data points from the line:

Q = ____________________________________

We want to minimize Q.

D. The independent variables in this expression for Q are ___________________

We need the partials of Q with respect to these variables to be zero

E. Why? ____

So we have

(1) $0 = \partial Q/\partial a = -\Sigma 2(y_i - a - bx_i)$

(2) $0 = \partial Q/\partial b = -\Sigma 2x_i(y_i - a - bx_i)$

Let’s work first with (1). Dividing by -2 and distributing the sum gives

(3) $\Sigma y_i - \Sigma a - \Sigma bx_i = 0.$

Now $\Sigma y_i = n\bar{y}$, $\Sigma a = na$, and $\Sigma bx_i = b\Sigma x_i = b(n\bar{x})$, so (3) simplifies to

(4) $n\bar{y}$-na - nb$\bar{x}$.

Canceling the n’s and solving for $\bar{y}$ gives

(5) $\bar{y} = \mathrm{a} + \mathrm{b}\bar{x}$.

This tells us two things:

a) *The point (* $\bar{x}$ *,* $\bar{y}$ *) lies on the least squares regression line.*

b) *Once we find b, we can find a by* $a = \bar{y} - b\bar{x}$.

Now we'll work with (2). First divide by -2 to get

(6) $\Sigma x_i(y_i - a - bx_i) = 0$.

Solve (5) for a and substitute in (6):

(7) $\Sigma x_i(y_i - \bar{y} + b\bar{x} - bx_i) = 0$.

Regroup this to give

(8) $\Sigma x_i(y_i - \bar{y}) - b\,\Sigma x_i\,(x_i - \bar{x}) = 0$.

Now solve for b:

(9) $b = [\ \Sigma x_i(y_i - \bar{y})]/[\ \Sigma x_i\,(x_i - \bar{x})]$,

which doesn't look like the formula on p. 137 for b. Let's look, however, at the book's formula:

(10) $b = r(s_y/s_x)$. (formula from book)

Remembering the formula for r, namely,

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

(10) becomes

(12) $$b = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}\left(\frac{s_y}{s_x}\right)$$ (from book formula)

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x^{\ 2}} \qquad \text{(from book formula)}$$

Using the definition of s_x (and canceling n-1's) now gives us

(13) $b = [\ \Sigma(x_i - \bar{x})(y_i - \bar{y})]/[\ \Sigma\ (x_i - \bar{x})^2]$, (from book formula)

which still isn't the same as (9), but looks at least somewhat similar. We will show that it is in fact the same; statistics has lots of formulas that are somewhat like trig identities, so your answer may not look like "the answer in the back of the book" unless you do some algebra first.

What we will use is the fact that "the sum of the deviations from the mean is zero." What this means is that

(14) $\Sigma(y_i - \bar{y}) = \Sigma y_i - \Sigma\bar{y} = n\bar{y} - n\bar{y}$ (F. Why?___________________)

$= 0.$

Using (14) can make the numerator of (13) look like the numerator of (9):

(15) $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i(y_i - \bar{y}) - \Sigma\bar{x}\ (y_i - \bar{y})$

$= \Sigma x_i(y_i - \bar{y}) - \bar{x}\Sigma(y_i - \bar{y})$

$= \Sigma x_i(y_i - \bar{y})$ by (14)

Similarly, using the x version of (14) we can make the denominator of (14) look like the denominator of (9). So the two formulas, (9) and (10), really say the same thing.

Comment: The original equations (1) and (2) (rather, the equations once we divide by -2) have further uses. In the terminology of Section 2.4, $y_i - \hat{y}_i$ (= $y_i - a - bx_i$) is called a *residual*. Since it is the residual corresponding to the i^{th} data point (x_i, y_i), we will call it the *i^{th} residual* and denote it e_i. Thus:

$e_i = y_i - \hat{y}_i = y_i - a - bx_i$.

Equation (1) cleaned up then says that

(16) $\Sigma e_i = 0.$

That is, the sum of the residuals is zero.

Equation (2) in cleaned up form (i.e., equation (6)) says

(17) $\Sigma x_i e_i = 0.$

(This can be thought of as saying that the sum of the residuals weighted by the x observations is zero.)

Using these, we also have

(18) $\Sigma \hat{y}_i e_i = \Sigma(a + bx_i)e_i$

$= a\Sigma e_i + b\ \Sigma x_i e_i$

$= 0$ (by (16) and (17))

(Thus the sum of the residuals weighted by the predicted values is zero.)

II. The book also makes an assertion about the connection of r^2 with regression that we will now prove.

First, we need to find $\bar{\hat{y}}$, the mean of the predicted values. In fact,

(19) $\bar{\hat{y}} = (1/n)\Sigma\ \hat{y}_i$

$= (1/n)\Sigma\ (a + bx_i)$

$= (1/n)\Sigma\ a + (1/n)(b\Sigma x_i)$

$= a + b\bar{x}$

$= \bar{y}$ by (5).

In other words, the *y observations and their predicted values have the same mean*!

Now use the formulas to re-express the least squares line:

$$
\begin{aligned}
\hat{y} &= a + bx \\
&= (\bar{y} \text{ -} b) + bx \\
&= \bar{y} + b(x - \bar{x}),
\end{aligned}
$$

so

$$
\begin{aligned}
\hat{y} - \bar{y} &= b(x - \bar{x}) \\
&= r(s_y/s_x)(x - \bar{x})
\end{aligned}
$$

Applying this to the data points, we have in particular that for each i,

$$\hat{y}_i - \bar{y} = r(s_y/s_x)(x_i - \bar{x}).$$

Summing over i, dividing by n-1, and using (19), we get

$$\text{var}(\hat{y}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 = \frac{1}{n-1} r^2 \frac{s_y^2}{s_x^2}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

$$= r^2 \frac{s_y^2}{s_x^2}\text{var}(x) = \text{r}^2(\text{s}_\text{y})^2 = \text{r}^2\text{var(y)}$$

So

(20) $\text{r}^2 = \text{var}(\hat{y})/\text{var(y)}$.

Since r^2 is at most 1, we see in particular that the variance of $\hat{y}$ is at most the variance of y. We can think of $\hat{y}$ as the part of y that can be explained by regression of y on x. Hence we can say that "*r^2 is the fraction of the variance of y that is explained by regression of y on x*." (This is not quite the same as the assertion on p. 141, but the two assertions are equivalent, since "the variation in y" refers to (n-1)var(y).)

Comment: With a little more work it is possible to show something even stronger, namely:

(21) $\text{var(y)} = \text{var}(\hat{y}) + \text{var(e)}$.

In words: *The variance of the observations is the sum of the variance of the predicted values and the variance of the residuals.* When this is written out as summations (and multiplied by n-1 to get less messy expressions), it looks like a typographical error or arithmetic mistake:

(22) $\Sigma[(\hat{y}_i - \bar{y}) + \text{e}_\text{i}]^2 = \Sigma\,(\hat{y}_i - \bar{y})^2 + \Sigma\text{e}_\text{i}^{\,2}$

However, if you look at it another way, you might see that it is something like a Pythagorean Theorem.

Equation (21) says that there are two contributions to var(y): the contribution var($\hat{y}$) from the predicted values -- that is, from the regression along x -- and the contribution var(e) from the residuals. (This idea of "contributions to variance from different sources" occurs in some advanced topics in statistics, such as Analysis of Variance and Multivariate Analysis.)