Using the SER

When using the regression equation to predict an *individual* y-value from an observed x-value, we can say that y is likely to be within one (or two) SER(s) of $\hat{y}(x)$.

Typically, we want to have actual numbers, so it is nice to know that there is a **shortcut** for computing the SER:

$$SER = \sqrt{1 - r_{xy}^2} \cdot SD_y.$$

Example. A large (hypothetical) study of the effect of smoking on the cardiac health of men, involved 2709 men aged 25 - 45, and obtained the following statistics,

$$\overline{x} = 17,\ SD_x = 8,\ \overline{y} = 129,\ SD_y = 7,\ r_{xy} = 0.64,$$

The SER (for predicting BP from cigarette consumption) is

$$SER = \sqrt{1 - (0.64)^2} \cdot 7 \approx 5.38.$$

Question: John is a 31-year old man who smokes 30 cigarettes a day. What is John's predicted blood pressure.

Answer: Our best guess for John is the average blood pressure of men who smoke 30 cigarettes a day. Since 30 is $13 = 13/8 \cdot SD_x$ above $\overline{x}$, the regression method predicts that John's blood pressure will be about

$$r_{x,y} \cdot \left(\frac{13}{8} \cdot SD_y\right) = 0.64 \cdot \left(\frac{13}{8} \cdot 7\right) \approx 7.28$$

mmHg *above average* — about 136.28 mmHg.

Now we can also add a margin of error, namely the SER, so we can say that John's blood pressure is likely to be in the range

$$136.28 \pm 5.38 \text{ mmHg.}$$

From the **regression method**...

> *Given a set of paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with **summary statistics***
>
> $$\overline{x},\ SD_x,\ \overline{y},\ SD_y \ \ \textit{and} \ \ r_{xy}$$
>
> *and if x_i is $k \cdot SD_x$ above (or below) $\overline{x}$, then the average of all the y-values corresponding to x_i is approximately $r_{xy} \cdot k \cdot SD_y$ above (or below) $\overline{y}$.*

If $\hat{y}(x_i)$ is the estimate for the average of all the y-values corresponding to $x = x_i$, then the regression method says that

$$\hat{y}(x_i) - \overline{y} = r_{xy} \cdot \left(\frac{x_i - \overline{x}}{SD_x} \right) \cdot SD_y$$

or

$$\hat{y}(x_i) = \overline{y} + r_{xy} \cdot \left(\frac{x_i - \overline{x}}{SD_x} \right) \cdot SD_y = \left(\frac{r \cdot SD_y}{SD_x} \right) \cdot x_i + \left(\overline{y} - \left(\frac{r \cdot SD_y}{SD_x} \right) \cdot \overline{x} \right)$$

... to the **regression *equation***. Renaming things, we can write

$$\hat{y}(x_i) = \beta_0 + \beta_1 x_i$$

- $\beta_1 = \dfrac{r_{xy} \cdot SD_y}{SD_x}$ is the *slope* of the regression line.
- $\beta_0 = \overline{y} - \beta_1 \overline{x}$, is the y-intercept of the regression line.

Example: A study of education and income is done for men age 30 - 35. A representative sample of 2317 men in this age group is surveyed, and the following statistics are collected:

$$\overline{E} = 13 \quad SD_E = 1.5$$

$$\overline{I} = 46 \quad SD_I = 10 \quad r_{I,E} = 0.45$$

where E = years of education and I = annual income, in \$1000s.

(*) Reg coeff.s: $\beta_1 = 0.45 \cdot (10/1.5) = 3$ and $\beta_0 = 46 - 3 \cdot 13 = 7$.

(*) Regression equation:

$$\hat{I} = 7 + 3E.$$

(*) A 32-year old man is observed, our best guess for his income is ...

(*) A 32-year old man is observed, our best guess for his income is ... the average income for all men in this age group, $\overline{I} = 46$.

(*) More informatively: his income is likely to fall in the range

$$\overline{I} \pm SD_I = 46 \pm 10.$$

(*) A 32-year old man with 16 years of education is observed, our best guess for his income is ... the average income of all men in this age group with 16 years of education, which we estimate using the regression equation:

$$\overline{I}(16) \approx \hat{I}(16) = 7 + 3 \cdot 16 = 55.$$

(*) The R.M.S. error of regression in this case is $\sqrt{1 - 0.45^2} \cdot 10 \approx 8.93$, so we can therefore predict the income of the 32 year old with 16 years of school (more informatively) as

$$55 \pm 8.93$$

Comments:

(*) When computing the regression coefficients and the SER of the regression of ***x on y***,

$$\hat{x} = \gamma_0 + \gamma_1 y,$$

the roles of SD_y, SD_x, $\overline{x}$ and $\overline{y}$ all switch, but r_{xy} plays the same role:

$$\gamma_1 = \frac{r_{xy} \cdot SD_x}{SD_y}, \ \gamma_0 = \overline{x} - \gamma_1 \overline{y} \ \text{ and } \ SER = \sqrt{1 - r_{xy}^2} \cdot SD_x.$$

Example: Regression for predicting educational level from income:

$$\gamma_1 = 0.45 \cdot \frac{1.5}{10} = 0.0675 \ \text{ and } \ \gamma_0 = 13 - 0.0675 \cdot 46 \approx 9.9.$$

$$\Longrightarrow \hat{E} = 9.9 + 0.0675 I$$

and the SER for predicting education from income is

$$SER = \sqrt{1 - (0.45)^2} \cdot 1.5 \approx 1.34.$$

Comments:

(*) By using the (same) SER to estimate the spread around $\hat{y}(x)$ for *all* the (different) x-values, we are making a fairly strong assumption — we are assuming that the spread in each vertical column through the scatterplot is about the same. If this is true, then the data is said to be ***homoscedastic***. Data that is *not* homoscedastic is said to be ***heteroscedastic***.

(*) There are statistical procedures used to test for heteroscedasticity, but we can also use common sense.

⇒ For example, data for education and income is likely to be (quite) heteroscedastic — the more education an individual has the bigger the range of possible incomes.

⇒ On the other hand, the data for heights of sons and heights of fathers can be expected to be (approximately) homoscedastic — the factors that result in variations in sons heights for fathers of the same height are likely to be very similar across all (fathers') heights.

(*) Another assumption that is commonly made is that of *normality* — that the y-values in each individual vertical strip are (approximately) normally distributed.

Example. Heights of sons on heights of fathers.

$$\overline{F} = 68 \quad SD_F = 2.7 \quad \overline{S} = 69 \quad SD_S = 2.7 \quad r = 0.5$$

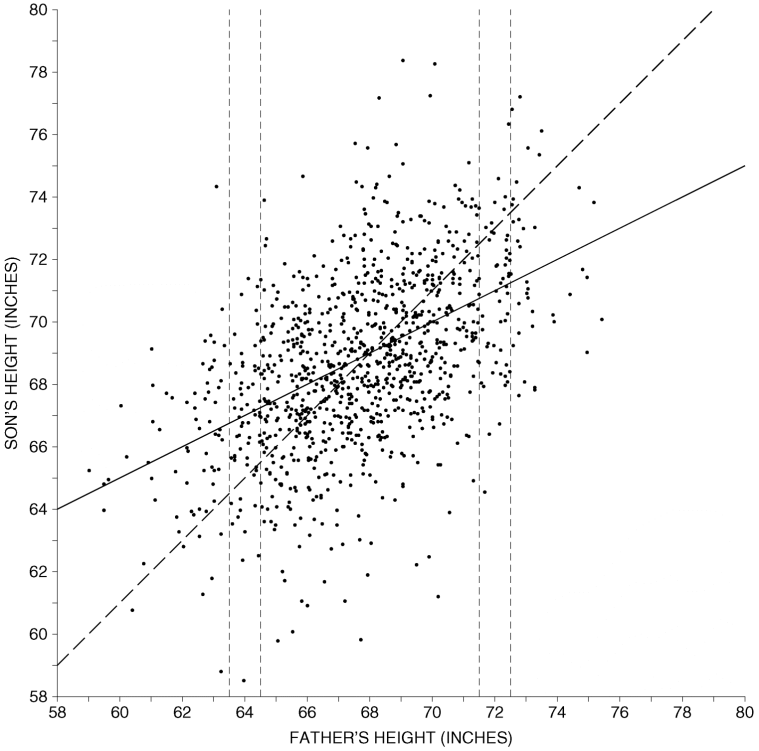

Figure 5., p.171 in FPP, sons and fathers' heights, with SD line and regression line.

The regression coefficients for predicting sons' heights from fathers' heights are

$$\beta_1 = 0.5 \cdot \frac{2.7}{2.7} = 0.5 \quad \text{and} \quad \beta_0 = 69 - 0.5 \cdot 68 = 35,$$

so the regression equation for predicting sons' heights from fathers' heights is $\hat{S} = 35 + 0.5F$ and the SER for this regression is

$$SER = \sqrt{1 - (0.5)^2} \cdot 2.7 \approx 2.34.$$

Example. The average height of men whose fathers were 72 inches tall is predicted to be about $\hat{S} = 35 + 0.5 \cdot 72 = 71$ inches. Assuming normality and using the SER as a proxy for the SD of the height distribution of these men, we can say that roughly 68% of the men whose fathers are 72 inches tall will have heights in the range

$$71 \pm 2.34 \text{ inches}$$

(average ±1 SD).

Least-squares line of 'best fit'.

We introduced the R.M.S error of regression

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \qquad (= \sqrt{1-r^2}\cdot SD_y)$$

to quantify the *vertical* spread of the data in a scatter-plot around the regression line. In this formula, $\hat{y}_j$ is the y-coordinate of the point on the regression line with the same x-coordinate as the point (x_j, y_j) in the data.

We can quantify the *vertical* spread of the data around the SD-line in the same way, and there is a similar shortcut formula:

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \tilde{y}_j)^2} \qquad \left(= \sqrt{2-2|r|}\cdot SD_y\right).$$

In this case, $\tilde{y}_j$ is the y-coordinate of the point on the <u>SD-line</u> with the same x-coordinate as the point (x_j, y_j) in the data.

Observation:

$$(2-2|r|)-(1-r^2)=r^2-2|r|+1=|r|^2-2|r|+1=(|r|-1)^2\geq 0,$$

so

$$2-2|r|\geq 1-r^2,$$

and this means that

$$\sqrt{2-2|r|}\cdot SD_y\geq\sqrt{1-r^2}\cdot SD_y,$$

This means that the vertical spread around the regression line is smaller than the vertical spread around the SD-line. I.e., the regression line does a better job (on average) than the SD-line of predicting y values from given x values.

In fact, in this sense, the regression line is better than ***any other*** straight line.

Specifically, given a set of paired data $\{(x_1, y_y), (x_2, y_2), \ldots, (x_n, y_n)\}$, then for any line, with equation $y = ax + b$ say, we can calculate a R.M.S. error,

$$R.M.S. \text{ error for the line } y = ax + b \quad = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - (ax_j + b))^2},$$

to measure the vertical spread of the data around this line.

Fundamental fact: *The regression-line is the line for which the R.M.S error is the smallest possible:*

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - (ax_j + b))^2} \geq \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

for any line $y = ax + b$. For this reason, the regression line is also called the *least-squares line of best fit.*