Linear regression for numerical bivariate data (Part II)

BEA140 Quantitative Methods - Module 2

UNIVERSITY *of* TASMANIA

Example of numerical bivariate data

You may recall that in part I we calculated the regression line of the following numerical bivariate data for temperature and ice cream sales.

temp (X)	ice cream sales (Y)	X^2	Y^2	XY
29	200	841	40000	5800
22	160	484	25600	3520
28	170	784	28900	4760
19	120	361	14400	2280
25	120	625	14400	3000
24	180	576	32400	4320
20	100	400	10000	2000
33	230	1089	52900	7590
200	1280	5160	218600	33270

The line we ended up with was $Y = -38.4375 + 7.9375X$.

Variance of errors (${s_e}^2$)

The **variance of errors** is to the regression line of bivariate data what the variance is to the mean of univariate data, and is given by:

$$ {s_e}^2 = \frac{\text{SSE}}{n-2} = \frac{\Sigma(Y - Y_c)^2}{n-2} \text{ (definitional form)} $$

$$ {s_e}^2 = \frac{\text{SSE}}{n-2} = \frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{n-2} \text{ (computational form)} $$

Standard error of the estimate (s_e)

The **standard error of the estimate** is to the regression line of bivariate data what the standard deviation was to the mean of univariate data, and is given by $s_e = \sqrt{{s_e}^2}$.

Reminder: It is important to always use full precision for calculations as excessive rounding can compound in to huge errors. Excessive rounding in the exam, assignments and tests will likely incur a penalty.

Example calculation for standard error

Going back to our bivariate data for temperature and ice cream sales.

temp (X)	ice cream sales (Y)	X^2	Y^2	XY
29	200	841	40000	5800
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
33	230	1089	52900	7590
200	1280	5160	218600	33270

$a = -38.4375$ and $b = 7.9375$

$$
\begin{aligned}
{s_e}^2 &= \frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{n-2} \\
&= \frac{218600 - (-38.4375)(1280) - 7.9375(33270)}{6} \\
&= 619.8958 \text{ (to 4 dp).}
\end{aligned}
$$

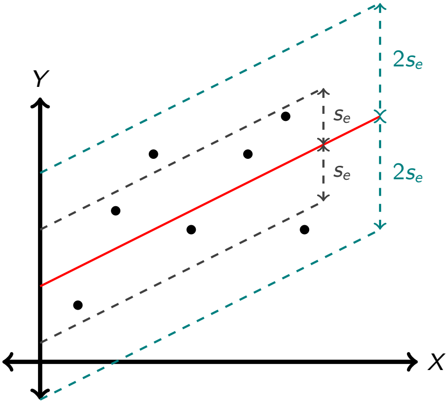
$$\Rightarrow s_e = \sqrt{619.8958} = 24.8977 \text{ (to 4 dp).}$$

Empirical rule for s_e

The empirical rule applies to the standard error for bivariate data in a similar way to how it applies to the standard deviation for univariate data. I.e.:

(i) around 68% of the data will fall (vertically) within one standard error of the regression line; and

(ii) around 95% of the data will fall within two standard errors of the regression line.



Standard error vs. standard deviation

Note: The standard deviation of temperatures is 44.4008. Applying the empirical rule:

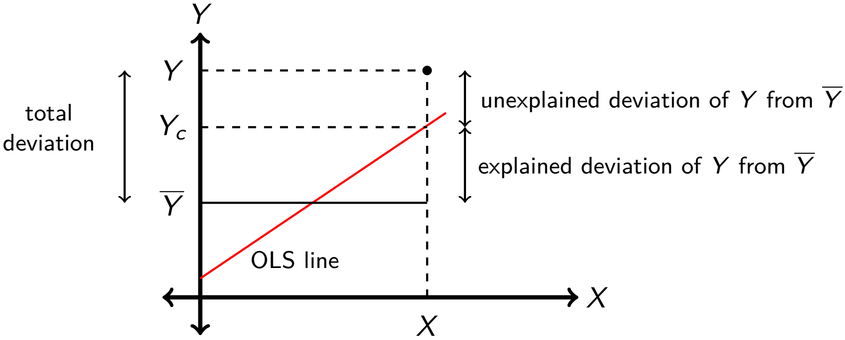
(i) approximately 68% of the data will fall within 44.4008 degrees of the mean temperature $\overline{Y}$;and

(ii) approximately 68% of the data for temperature and ice cream sales will fall (vertically) within 24.8977 degrees of the regression line.

I.e. in moving from a univariate/mean perspective to a bivariate/regression perspective improves our predictive capability, and hence our ability to plan and manage resources.

Note: If you do further courses in statistics, you will likely learn about regressions for data sets with an arbitrary number of variables, which can further improve the predictive capability/accuracy.

Deviation of data from $\overline{Y}$



(i) sum of explained/regression deviations $= \Sigma(Y_c - \overline{Y})$;

(ii) sum of unexplained/error deviations $= \Sigma(Y - Y_c)$; and

(iii) sum of total deviations $= \Sigma(Y - \overline{Y})$.

Note: For the regression line of a bivariate data set, the above sums are all equal to zero (as already noted for the sum of unexplained/error deviations when discussing the properties of regression lines).

Sum of squared deviations of data from $\overline{Y}$

(i) sum of squared regression deviations $= \text{SSR} = \Sigma(Y_c - \overline{Y})^2$;
(ii) sum of squared error deviations $= \text{SSE} = \Sigma(Y - Y_c)^2$; and
(iii) sum of squared total deviations $= \text{SST} = \Sigma(Y - \overline{Y})^2$.

Note: It can be shown mathematically that:

(i) $\text{SST} = \text{SSR} + \text{SSE}$;
(ii) $\text{SST} = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$; and
(iii) $\text{SSE} = \Sigma Y^2 - a\Sigma Y - b\Sigma XY$.

Coefficient of determination (r^2)

The **coefficient of determination** is the proportion of the deviation of data around $\overline{Y}$ that can be explained by the regression line.

$$\begin{array}{c}\text{coefficient of}\\ \text{determination}\end{array} = r^2 = \frac{\text{explained deviation}}{\text{total deviation}} = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Note: r^2 can be used to compare strengths of alternative relationships, i.e. which model best fits the data.

Sanity checks:

(i) $SST > 0$ and $SSE > 0$;

(ii) $0 \leq r^2 \leq 1$; and

(iii) does your calculation differ wildly from how well the line visually fits the data?

Note: Larger values of r^2 typically correspond to lower values of s_e and vice versa.

Example calculation for the coefficient of determination

Going back to our bivariate data for temperature and ice cream sales:

temp (X)	ice cream sales (Y)	X^2	Y^2	XY
29	200	841	40000	5800
⋮	⋮	⋮	⋮	⋮
33	230	1089	52900	7590
200	1280	5160	218600	33270

$$a = -38.4375 \text{ and } b = 7.9375$$

$$\begin{aligned}\text{SSE} &= \Sigma Y^2 - a\Sigma Y - b\Sigma XY \\ &= 218600 - (-38.4375)(1280) - 7.9375(33270) \\ &= 3719.375\end{aligned}$$

$$\text{SST} = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 218600 - \frac{1280^2}{8} = 13800$$

$$\Rightarrow r = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{3719.375}{13800} = 0.7305 \text{ (to 4 dp).}$$

I.e. For the given data set, 73.05% of the deviation of temperature data around $\overline{Y}$ can be explained by the regression line.

Pearson's coefficient of correlation (r)

Pearson's coefficeint of correlation is a summary measure that can take values from -1 (when all points lie on a negatively sloped line) to $+1$ (when all points lie on a positively sloped line).
The 'long formula' is

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{(\Sigma X^2 - \frac{(\Sigma X)^2}{n})(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n})}}$$

However it can also be calculated directly from r^2 using the following 'short formula':

$$r = \begin{cases} \sqrt{r^2} & \text{when } b \geq 0 \text{ (i.e. positive slope); and} \\ -\sqrt{r^2} & \text{when } b < 0 \text{ (i.e. negative slope).} \end{cases}$$

Example calculation for the coefficient of correlation

Using the bivariate data for temperatures and ice cream sales again, the 'long formula' calculation of r is:

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{(\Sigma X^2 - \frac{(\Sigma X)^2}{n})(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n})}} = \frac{33270 - \frac{200*1280}{8}}{\sqrt{(5160 - \frac{200^2}{8})(218600 - \frac{1280^2}{8})}}$$

$= 0.8547$ (to 4 dp),

and since the slope of the regression is positive, the 'short formula' calculation is:

$$r = +\sqrt{r^2} = \sqrt{0.7304800725} = 0.8547 \text{ (to 4 dp)}.$$

I.e. there is a reasonably strong positive correlation for our temperature and ice cream sale data.

Note: It is important to remember that if the slope of a regression line that you are working with is negative, then the coefficient of correlation is instead $r = -\sqrt{r^2}$.

. . . that's it for now, thanks for watching!

Don't forget that you can ask questions via:

(i) face-to-face lectures;

(ii) workshops or tutorials;

(iii) consultation hours; or

(iv) email.