Statistics

WHAT IS THE COEFFICIENT OF DETERMINATION?

As we know, the standard deviation squared is called VARIANCE. So, the VARIANCE can be calculated using the formula:

$$\frac{\sum(y_i - \bar{y})^2}{(N-1)}$$

That is why we will call VARIATION at the sum:

$$\sum(y_i - \bar{y})^2$$

This expression can be re-written this way:

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$$

If we distribute the binomial we get:

$$\sum(y_i - \bar{y})^2 = \sum[(y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2]$$

Or

$$\sum(y_i - \bar{y})^2 = \sum[(y_i - \hat{y}_i)^2 + \sum 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum(\hat{y}_i - \bar{y})^2]$$

Now, I will show that the sum $\sum 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ is equal zero. That means that the sum $\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ should be zero.

We know that $\hat{y}_i = a + bx_i$, so:

$$\sum(y_i - a - bx_i)(a - bx_i - \bar{y})$$

We also know that, $\bar{y} = a + b\bar{x}$, so

$$\sum(y_i - a - bx_i)(a - bx_i - a - b\bar{x})$$

And taking b common factor:

$$\sum(y_i - a - bx_i)b(x_i - \bar{x})$$

But b cannot be zero, so the sum:

$$\sum(y_i - a - bx_i)(x_i - \bar{x})$$

Should be zero, or if we distribute the product:

$$\sum y_i x_i - N\bar{y}\bar{x} - bN(\overline{x^2} - \bar{x}^2)$$

Should be zero.

By the definition of standard deviation we get,

$$(N-1){S_x}^2 = \sum(x_i - \bar{x})^2$$

Or, what is equivalent:

$$(\overline{x^2} - \bar{x}^2) = \frac{(N-1)}{N} {S_x}^2$$

So the expression (2) transform into:

$$(2) \qquad \sum y_i x_i - N\bar{y}\bar{x} - b(N-1)\,{S_x}^2$$

Now, from the definition of correlation coefficient r:

$$R = \frac{\sum(y_i - \bar{y})(x_i - \bar{x}),}{(N-1)S_x S_y},$$

From here we get that

$$\sum y_i x_i - N\bar{y}\bar{x} = r(N-1)S_x S_y$$

Now, if we take into account that $b = rS_y/S_x$ and from (2) we get, the expression

$r(N-1)S_x S_y - b(N-1)\,{S_x}^2$ transform into:

$r(N-1)S_x S_y - r(N-1)S_x S_y$ which is ZERO!

So, we can re-write

$$\sum(y_i - \bar{y})^2 = \sum[(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y}\,\bar{y})^2]$$

As we know, the difference $(y_i - \hat{y}_i)$ is called the RESIDUAL. The residual is defined as the difference between the observed

and the expected values. That means, the difference between the experimental value and the value predicted by our model. The bigger the residual, the worse will be the prediction power of our model. That is why we will call UN-EXPLAINED VARIATION to the term:

$$\sum (y_i - \hat{y}_i)^2$$

And we will call EXPLAINED VARIATION to the other term:

$$\sum (\hat{y}_i - \bar{y}\,\bar{y})^2$$

Now, I will introduce the PROPORTION OF EXPLAINED VARIATION (R^2) as the ratio:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y}\bar{y})^2}{\sum (y_i - \bar{y}\,\bar{y})^2}$$

Now I will simplify this expression using the definition of $\hat{y}_i = a + bx_i$, and the value of $\bar{y} = a + b\bar{x}$.

$$R^2 = \frac{\sum (a + bx_i - a - b\bar{x})^2}{\sum (y_i - \bar{y}\bar{y})^2}$$

$$R^2 = b^2 \frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y}\,\bar{y})^2}$$

But , b = $r\frac{S_y}{S_x}$ and,

$$\sum(x_i - \bar{x})^2 = (N-1)\,{S_x}^2 \text{ and,}$$

$$\sum(\,y_i - \bar{y})^2 = (N-1){S_y}^2$$

So, we conclude that

$$\mathbf{R^2 = r^2}$$

So, the coefficient of determination is equal to the coefficient of correlation squared!

Remember this: the coefficient of determination is the proportion of explained variation"

BACK