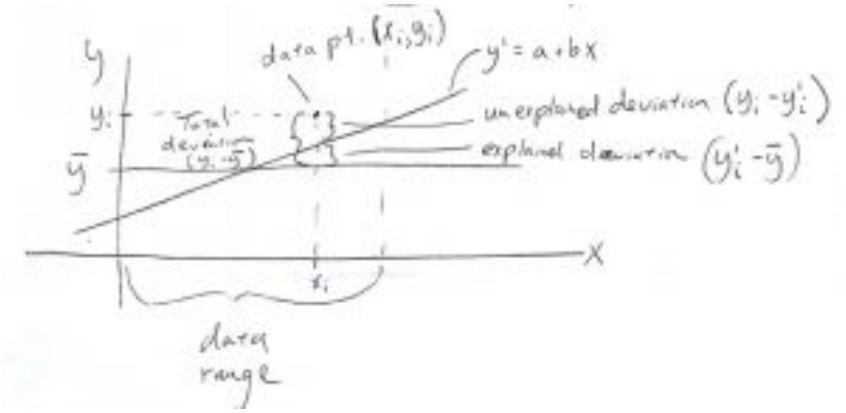
14.5: r² and the Standard Error of the Estimate of y′

Consider the deviations :



Looking at the picture we see that

$$\begin{aligned} \text{total deviation} &= \text{explained deviation} + \text{unexplained deviation} \\ (y_i - \bar{y}_i) &= (y_i' - \bar{y}_i) + (y_i - y_i') \end{aligned}$$

Remember that variance is the sum of the squared deviations (divided by degrees of freedom), so squaring the above and summing gives:

$$\sum_{i=1}^{n} (y_i - \bar{y}_i)^2 = \sum_{i=1}^{n} (y_i' - \bar{y}_i)^2 + \sum_{i=1}^{n} (y_i - y_i')^2 \tag{14.5.1}$$

(the cross terms all cancel because y' is the least square solution and $a = \bar{y} - b\bar{x}$, see Section 14.6.1, below, for details). This is also a sum of squares statement:

$$\text{SS}_T = \text{SS}_R + \text{SS}_E \tag{14.5.2}$$

where $\text{SS}_E = \sum (y_i - y_i')^2$, $\text{SS}_T = \sum (y_i - \bar{y})^2$ and $\text{SS}_R = \sum (y_i' - \bar{y})^2$ are the sum of squares — error, sum of squares — total and sum of squares — regression (explained) respectively.

Dividing by the degrees of freedom, which is $n - 2$ in this {\em bivariate} situation, we get:

$$\begin{aligned} \frac{\sum (y_i - \bar{y}_i)^2}{n-2} &= \frac{\sum (y_i' - \bar{y}_i)^2}{n-2} + \frac{\sum (y_i - y_i')^2}{n-2} \\ \text{total variance} &= \text{explained variance} + \text{unexplained variance} \\ &= \text{signal (or model)} + \text{noise} \end{aligned}$$

It turns out that

$$r^2 = \frac{\text{explained variance}}{\text{total variance}} = \frac{\text{SS}_R}{\text{SS}_T} \tag{14.5.3}$$

The quantity r^2 is called the *coefficient of determination* and gives the ***the fraction of variance explained by the model*** (here the model is the equation of a line). The quantity r^2 appears with many statistical models. For example with ANOVA it turns out that the “effect size” eta-squared is the fraction of variance explained by the ANOVA model[1], $\eta^2 = r^2$.

The *standard error of the estimate* is the standard deviation of the noise (the square root of the unexplained variance) and is given by

$$s_{\text{est}} = \sqrt{\frac{\sum (y - y')^2}{n-2}} = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}} \tag{14.5.4}$$

Example 14.4: Continuing with the data of Example 14.3, we had

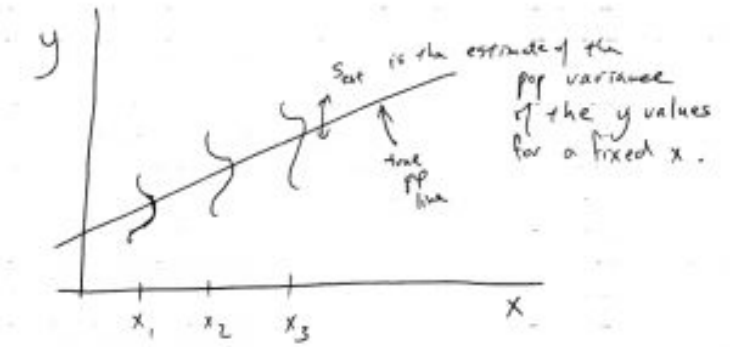
$$\sum y = 511 \quad \sum y^2 = 38993 \quad \sum xy = 3745 \quad a = 102.493 \quad b = -3.622 \quad n = 7 \tag{14.5.5}$$

so

$$\begin{aligned} s_{\text{est}} &= \sqrt{\frac{(38993) - (102.493)(511) - (-3.622)(3745)}{5}} \\ s_{\text{est}} &= \sqrt{\frac{38993 - 52373.923 + 13564.39}{5}} \\ s_{\text{est}} &= 6.06 \end{aligned}$$

□

Here is a graphical interpretation of s_{est} :



The assumption for computing confidence intervals for is that s_{est} is independent of x. This is the assumption of homoscedasticity. You can think of the regression situation as a generalized one-way ANOVA where instead of having a finite number of discrete populations for the IV, we have an infinite number of (continuous) populations. All the populations have the same variance σ^2 (and they are assumed to be normal) and s_{est^2} is the pooled estimate of that variance.

14.6.1: **Details: from deviations to variances

Squaring both sides of

$$(y_i - \bar{y}_i) = (y'_i - \bar{y}_i) + (y_i - y'_i) \tag{14.5.6}$$

and summing gives

$$\sum (y_i - \bar{y}_i)^2 = \sum (y'_i - \bar{y}_i)^2 + \sum (y_i - y'_i)^2 + \sum 2(y'_i - \bar{y})(y_i - y'_i) \tag{14.5.7}$$

Working on that cross term, using $a = \bar{y} - b\bar{x}$, we get

$$\begin{aligned} \sum 2(y'_i - \bar{y})(y_i - y'_i) &= \sum 2((\bar{y} - b\bar{x} + bx_i) - \bar{y})(y_i - y'_i) \\ &= \sum 2((\bar{y} + b(x_i - \bar{x})) - \bar{y})(y_i - y'_i) \\ &= \sum 2(b(x_i - \bar{x}))(y_i - y'_i) \\ &= \sum 2b(x_i - \bar{x})(y_i - (\bar{y} + b(x_i - \bar{x}))) \\ &= \sum 2b((y_i - \bar{y})(x_i - \bar{x}) - b(x_i - \bar{x})^2) \\ &= 2b \sum ((y_i - \bar{y})(x_i - \bar{x}) - (y_i - \bar{y})(x_i - \bar{x})) = 0 \end{aligned}$$

where

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{14.5.8}$$

was used in the last line.

1. In ANOVA the ``model'' is the difference of means between the groups. We will see more about this aspect of ANOVA in Chapter 17. ↵