Relationship between the Ratio of the Linear Estimate to the Estimate's Standard Error and the Regression Coefficient

Philip A. Shea

10th September, 2024

Abstract

An unexpected relationship was observed between the ratio $\widehat{\beta}/S_{\widehat{\beta}}$ and the regression coefficient r^2. The relationship was discovered to depend on small values of the estimated slope, and a constant set of independent variables.

1 Purpose

Regression is used to determine if a linear relationship exists between variables, and the existence of such a relationship is rejected unless it can be shown to exist with statistical significance. That is, the slope must be (statistically) significantly greater than the error in estimating the slope. The hypothesis are:

$$H_0\text{: } \widehat{\beta} = 0$$
$$H_1\text{: } |\widehat{\beta}| > 0$$

Statisticians have developed a statistic to prove that 1: there is correlation, and 2: that the slope of the regression line is significant (these are not independent). Ordinarily, H_0 is declared correct unless the statistic is violated

by a desired degree of probability. Our task, however, is the opposite: ensure that the slope is zero with statistical significance. Fig. 1 shows the Students t probability distribution, which the statistics will follow. We can choose the center region and declare that the slope is zero or (equivalently) there is no correlation if either or both statistics fall in this region[1]. However, this is a very narrow region as it occupies the center, and the whole curve represents the likelihood of the statistics *when* H_0 *is true*; even when true, we will only declare it so 5% of the time.

The t statistic for the estimate of the slope is given by the following, where β is the actual slope (zero in this analysis), $\widehat{\beta}$ is the estimated slope, and $S_{\widehat{\beta}}$ is the standard error of the estimated slope.

$$t_\beta = \frac{\widehat{\beta} - \beta}{S_{\widehat{\beta}}}$$

The t statistic for the regression coefficient is given by the following, where n is the number of samples, and r is the regression coefficient:

$$t_r = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

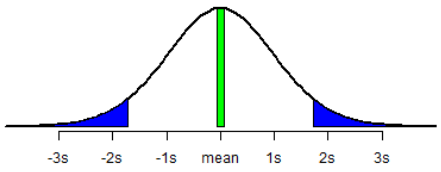


Figure 1 – Student t distribution with 20 degrees of freedom (i.e., 22 samples), showing the $\pm 5\%$ tails (blue) and 5% middle (green). The coordinates for the tails are $\pm 1.725s$ and the middle region is $\pm 0.0635s$. The Student t distribution applies to both the regression coefficient and the estimate of the slope.

2 Problem

These statistics, which, if the samples are drawn from a normal distribution, have a students-t distribution with $n-2$ degrees of freedom (where n is the

[1]There are two 5% tails because a real slope will either be positive of negative, and if we reject the null hypothesis with 5% chance of error we can say the detected slope is real with 95% confidence. Accepting the null hypothesis requires the 5% surround zero, as so if we want reject H_1 we can say the slope is non-zero with 95% confidence.

number of samples). This statistic may be thought of as simply the number of standard deviations the estimate (either the slope or r) is above zero. Thus a t statistic of 2 indicates the estimate is two standard deviations above zero.

This was being examined for possible use in determining when a slope could be safely assumed to be zero. In the process, the plot of Fig. 2 was produced from real data which exhibited random correlations, and the shape was startling, not so much for the quadratic behaviour, but from the complete lack of randomness. Basically, $r^2 \sim t^2$, or $r \sim t$. A brief search of books on statistics and regression failed to reveal any such established relationship. So where is this coming from?

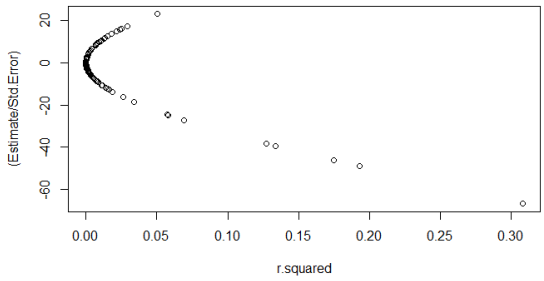


Figure 2 – Example of surprising relationship between $t_\beta = \widehat{\beta}/S_{\widehat{\beta}}$ and r^2. Each of the 100 data points was created from a fit of 10,000 samples.

3 Definitions

In the equations below, we will use a hat (e.g., $\widehat{\beta}$) to signify an estimate of a quantity, with the exception of simple means, where we will use the familiar over-bar (e.g., $\bar{x} = \widehat{\mu}_x$). The equation for the estimate is[2]:

$$\begin{aligned}
\widehat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})}{nS_x} \\
&= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)S_x} \\
&= \frac{S_{xy}}{S_x^2}
\end{aligned} \tag{1}$$

[2]Most of the equations are copied from Wikipedia's article on Simple Linear Regression.

with $\bar{x} = \sum x_i/n$ the estimate of the mean of x, $S_x^2 = \sum(x_i - \bar{x})^2/(n-1)$ the estimate of the standard deviation of x, and $S_{xy} = \left(\sum x_i y_i - n\bar{x}\bar{y}\right)/(n-1)$ the covariance of x and y. The standard error of the estimate is:

$$S_{\widehat{\beta}} = \sqrt{\frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{2}$$

where $\widehat{\varepsilon}_i = y_i - \hat{y}_i$. The estimate can be simply related to the regression coefficient as:

$$\widehat{\beta} = \frac{S_{xy}}{S_x^2} = r_{xy}\frac{S_y}{S_x} \tag{3}$$

where r_{xy} is the square root of the regression coefficient r^2 (see below). Finally, we have the regression coefficient:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \tag{4}$$

where S_x and S_y are the standard deviations of x and y.

4 Algebra

The square root in $S_{\widehat{\beta}}$ is difficult to work with, so we will work with the square of the statistic $\widehat{\beta}/S_{\widehat{\beta}}$ from equations 1 and 2:

$$\left(\frac{\widehat{\beta}}{S_{\widehat{\beta}}}\right)^2 = \frac{\frac{S_{xy}^2}{S_x^4}}{\frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{(n-2)\sum_{i=1}^n (x_i-\bar{x})^2}}$$
$$= \frac{(n-2)S_{xy}^2 \sum_{i=1}^n (x_i-\bar{x})^2}{S_x^4 \sum_{i=1}^n \widehat{\varepsilon}_i^2}$$
$$= \frac{(n-2)(n-1)S_{xy}^2 S_x^2}{S_x^4 \sum_{i=1}^n \widehat{\varepsilon}_i^2}$$
$$= \frac{(n-2)(n-1)S_{xy}^2}{S_x^2 \sum_{i=1}^n \widehat{\varepsilon}_i^2}$$

4.1 Expanding the Sum of Squared Errors

The expression for the sum of the errors can be expanded recognizing that $\hat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$:

$$\sum_{i=1}^n \widehat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^n (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2$$
$$= \sum_{i=1}^n \left(y_i^2 - 2\widehat{\alpha} y_i - 2\widehat{\beta} x_i y_i + \widehat{\alpha}^{\,2} + 2\widehat{\alpha}\widehat{\beta} x_i + \widehat{\beta}^{\,2} x_i^2 \right)$$
$$= \sum_{i=1}^n y_i^2 - 2\widehat{\alpha} \sum_{i=1}^n y_i - 2\widehat{\beta} \sum_{i=1}^n x_i y_i + n\widehat{\alpha}^{\,2} + 2\widehat{\alpha}\widehat{\beta} \sum_{i=1}^n x_i + \widehat{\beta}^{\,2} \sum_{i=1}^n x_i^2$$

Using the definitions for the means and for S_{xy}, S_x^2, and S_y^2 (in particular, the variances expressed as, for example, $\sum y_i^2 = (n-1)S_y^2 + \bar{y}^2$, and the covariance $\sum x_i y_i = (n-1)S_{xy} + n\bar{x}\bar{y}$), the above becomes:

$$\sum_{i=1}^n \widehat{\varepsilon}_i^2 = n\widehat{\alpha}^{\,2} + \bar{y}^2 - 2n\widehat{\alpha}\bar{y} + \widehat{\beta}^{\,2} \left[\bar{x}^2 + -\bar{x}\bar{y} - (n-1)S_x^2 \right] +$$
$$2\widehat{\beta}(n\widehat{\alpha}\bar{x} + S_{xy}(1-n)S_{xy}) + (n-1)S_y^2$$

Perhaps we should have done this earlier, but we need to use the definitions of $\widehat{\alpha}$ (i.e, $\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}$) and $\widehat{\beta}$ (S_{xy}/S_x^2, see equation 1). This complicated expression was simplified through *Mathematica*'s `FullSimplify` function.

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \frac{(1-n)\left[\bar{y}^2 S_x^4 - 2\bar{x}\bar{y}S_x^2 S_{xy} + (\bar{x}^2 + S_x^2)S_{xy}^2 - S_x^4 S_y^2\right]}{S_x^4}$$

4.2 Putting them Together

The ratio we started with can now be written as the following:

$$\left(\frac{\widehat{\beta}}{S_{\widehat{\beta}}}\right)^2 = \frac{(n-2)S_x^2 S_{xy}^2}{\bar{y}^2 S_x^4 - 2\bar{x}\bar{y}S_x^2 S_{xy} + (\bar{x}^2 + S_x^2)S_{xy}^2 - S_x^4 S_y^2} \tag{5}$$

which doesn't appear at all helpful. The relationship we suspect exists is that $\left(\frac{\widehat{\beta}}{S_{\widehat{\beta}}}\right)^2 \sim r_{xy}^2$. We can divide equation 5 by equation 4 squared and see how they are related.

$$\left(\frac{\widehat{\beta}}{S_{\widehat{\beta}}}\right)^2 \Big/ \left(\frac{S_{xy}}{S_x S_y}\right)^2 = \frac{(n-2)S_x^4 S_y^2}{\bar{y}^2 S_x^4 - 2\bar{x}\bar{y}S_x^2 S_{xy} + (\bar{x}^2 + S_x^2)S_{xy}^2 - S_x^4 S_y^2} \tag{6}$$

At first glance, that is perhaps even less helpful. Each data point in the data which produced 2 had the same set of x_i ($x_i = 1 : 10,000$) and the same n (9,998), so we can recast the above with constants to see what true variables remain:

$$\left(\frac{\widehat{\beta}}{S_{\widehat{\beta}}}\right)^2 \Big/ \left(\frac{S_{xy}}{S_x S_y}\right)^2 = \frac{(n-2)c_1 S_y^2}{c_1\bar{y}^2 - c_2\bar{y}S_{xy} + c_3 S_{xy}^2 - c_1 S_y^2} \tag{7}$$

where $c_1 = S_x^4$, $c_2 = 2\bar{x}S_x^2$, and $c_3 = \bar{x}^2 + S_x^2$. These have been computed and are shown below in Tab. 1.

Constant	Value
c_1	6.946×10^{13}
c_2	8.335×10^{10}
c_3	3.333×10^{7}

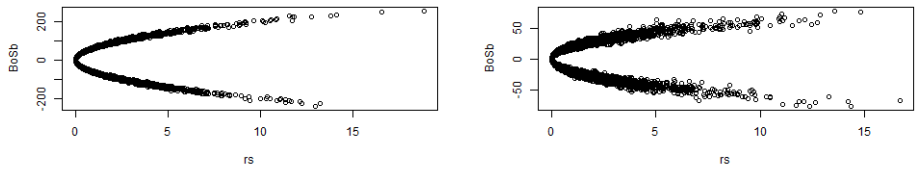
Table 1 – Constants from figure 2 and equation 7

It is clear now what is happening: the terms with c_1 are dominating. The ratio can be written as

$$\left(\frac{\widehat{\beta}}{S_{\widehat{\beta}}}\right)^2 / \left(\frac{S_{xy}}{S_x S_y}\right)^2 \approx \frac{(n-2)}{\bar{y} - S_y^2}$$

5 Comparison to Completely Random Data

To show what would happen in data whose independent variable was random, a small program was written to generate 10,000 linear fits, each of 1,000 points, and whose linear parameters where themselves random. The results are displayed below in Figure 3. This shows that surprise at the lack of randomness was warranted, although a strong quadratic relationship exists. The strong dependence on the number of points is evident in figure b below, where each of the 10,000 fits was to only 100 points.



(a) $n = 1,000$ **(b)** $n = 100$

Figure 3 – Comparing $\frac{\widehat{\beta}}{S_{\widehat{\beta}}}$ for different values of n.