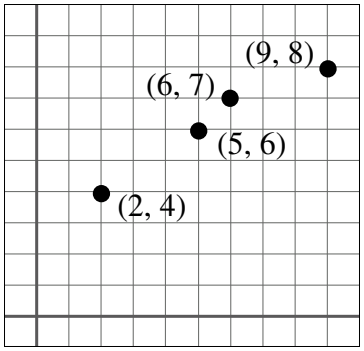
Stat 401, section 12.2 Estimating Model Parameters

notes prepared by Tim Pilachowski

In the real world, the true population regression is not known (in contrast to our omniscient knowledge in section 12.1). Instead, data is collected—observations are made about phenomenon. The point estimate of the regression equation is found by finding values that are the best fit for the data.

Hopefully you remember how to find the equation of a line given two points: calculate slope (m), substitute to find the y-coordinate of the y-intercept (b), and write the equation ($y = mx + b$). When there are more than two points, and they don't conveniently line up for us, we use partial derivatives to minimize the difference (error) between the observed data and the **estimated regression line** or **least squares line**. The process is called **regression analysis**. We'll use the same convention as before and use the notations $\hat{\beta}_0$ and $\hat{\beta}_1$ for the **least squares estimates** of the linear parameters, with b_0 and b_1 denoting the calculated values of our point estimates.



Example O: A manufacturer has collected preliminary data relating number of units produced (x, measured in hundreds) and cost (y, measured in \$1000's). Find the equation that best represents cost as a function of number of units produced. *answer*: $y = 3.06 + 058x$

The data are pictured in the scatterplot to the right. The line of regression will go "through the middle", but the question becomes, "Of all the lines that we might draw that seem to fit the data, which is the one that has the least error between the observed (actual) y-value and the regression's (predicted) y-value = $b_0 + b_1 x$?"

x_i	observed y_i	regression $\hat{y}_i = b_0 - b_1 x_i$	error $y_i - (b_0 + b_1 x_i)$	(error)^2
2	4			
5	6			
6	7			
9	8			

Since some of the points will lie above the regression line, and some will lie below it, some of the errors will be positive and some will be negative. Finding the sum of the errors would result in a "canceling" effect. Instead, to keep the value of each error and retain its effect we'll find the sum of the squared errors.

This is the function for which we want a minimum. We'll use the technique of partial derivatives. (We'll also need the chain rule.)

(Example O continued)

The process above involved only four data points. If we had 400, or 4000, the same process would become very unwieldy very quickly. We can develop a general formula which can be applied to N = any number of points.

x_i	observed y_i	regression y	error $y_i-(b_0+b_1x_i)$	(error)^2
x_1	y_1	$b_0+b_1x_1$	$y_1-b_0-b_1x_1$	$(y_1-b_0-b_1x_1)^2$
x_2	y_2	$b_0+b_1x_2$	$y_2-b_0-b_1x_2$	$(y_2-b_0-b_1x_2)^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_N	y_N	$b_0+b_1x_n$	$y_n-b_0-b_1x_n$	$(y_n-b_0-b_1x_n)^2$

The least-squares error function we want to minimize is

$$\text{sum of squared vertical deviations} = f(b_0,b_1)=\sum (y_i-b_0-b_1x_i)^2$$

Note: I really *should* write $f(b_0,b_1)=\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2$, but chose the version above for simplicity's sake.

Using the sum rule and chain rule, we get

$$\frac{\partial f}{\partial b_0}=\sum 2(y-b_0-b_1x)(-1)=0$$

$$\frac{\partial f}{\partial b_1}=\sum 2(y-b_0-b_1x)(-x)=0$$

Note that since we have n points and thus have n terms in our sum, we can replace $\sum b_0$ with nb_0.

Dividing by 2 to get easier numbers and rearranging gives us the system of **normal equations**.

$$\begin{cases}-\sum y + nb_0 + b_1\sum x = 0\\ -\sum xy + b_0\sum x + b_1\sum x^2 = 0\end{cases}\Rightarrow\begin{cases}nb_0 + b_1\sum x = \sum y\\ b_0\sum x + b_1\sum x^2 = \sum xy\end{cases}$$

Solving the first equation for b_0 we get $b_0=\dfrac{\sum y - b_1\sum x}{n}=\bar{y}-b_1\bar{x}$ [random variable $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$].

Substituting into the second equation and solving for b_1 we get $b_1=\dfrac{\sum xy-\dfrac{\sum x*\sum y}{n}}{\sum x^2-\dfrac{(\sum x)^2}{n}}=\dfrac{S_{xy}}{S_{xx}}$

[computational formula for random variable $\hat{\beta}_1$]

(The numerator and denominator above are the text's computational formulas for

$$S_{xy}=\sum (x-\bar{x})(y-\bar{y}),\ S_{xx}=\sum (x-\bar{x})^2\ .)$$

(Go to Examples A-1 and A-2 below.)

Fitted (or **predicted**) **values** for random variable Y are obtained by successively substituting x-values into the equation of the estimated regression line: $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$, $\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2$, … , $\hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$. The **residuals** are the differences between the observed and fitted y values: $y_1 - \hat{y}_1$, $y_2 - \hat{y}_2$, … , $y_n - \hat{y}_n$.

In much the same way that the deviations from the mean in a one-sample situation were combined to obtain the estimate $s^2 = \dfrac{\sum (x - \bar{x})^2}{n-1}$, the estimate of σ^2 in regression analysis is based on squaring and summing the residuals (as we did earlier in Example O). The text still uses the symbol s^2 for this estimated variance; you'll have to figure out from the context that it is not the same formula as for a single-sample variance.

$$S_{xy} = \sum xy - \frac{\sum x * \sum y}{n}, \quad S_{yy} = \sum (y - \bar{y})^2 = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

The **error sum of squares** (equivalently, **residual sum of squares**), is given by

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum \left(y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 x_i\right]\right)^2 = S_{yy} - \hat{\beta}_1 S_{xy}$$

and the estimate of σ^2 is given by

$$\hat{\sigma}^2 = s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2}.$$

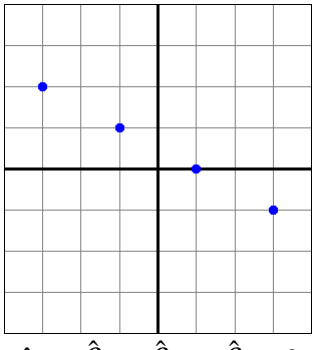
Important: Rounding errors can compound quickly. Carry as many decimal places as possible when calculating SSE and s^2.

The divisor $n - 2$ in s^2 is the number of degrees of freedom associated with SSE and the estimate s^2. This is a result of needing to estimate the two parameters β_0 and β_1, which results in a loss of two degrees of freedom. (In a similar fashion, μ had to be estimated in one-sample problems, resulting in an estimated variance based on $n - 1$ degrees of freedom.)

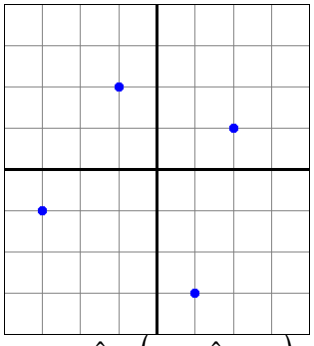
It can be shown that the random variable S^2 is an unbiased estimator for σ^2 (although the estimator S is not an unbiased estimator for σ).

(Go to Example A-3 below.)

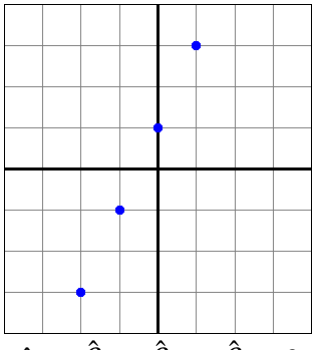
Different scatterplots exhibit different variability.



$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $\hat{\beta}_1 < 0$



$\hat{y}_i = \hat{\beta}_0$, (i.e. $\hat{\beta}_1 = 0$)



$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $\hat{\beta}_1 > 0$

The points in the first and third scatterplots all fall exactly on the estimated regression line. In these two cases, all (100%) of the sample variation in y can be attributed to the fact that x and y are linearly related in combination with variation in x. That is, the 1st and 3rd illustrate a deterministic rather than a probabilistic relationship. In the middle scatterplot, none (0%) of the sample variation in y can be attributed to a linear relationship – the best estimate we ever have for y is the expected value of y, i.e. $\hat{y}_i = \hat{\beta}_0 = \bar{y}$ for all indices i.

Of course, in actuality, we'd almost never encounter any of these extremes.

A quantitative measure of the total amount of variation in observed *y* values is given by the **total sum of squares**: $\text{SST} = S_{yy} = \sum \left(y - \bar{y}\right)^2 = \sum y^2 - \dfrac{\left(\sum y\right)^2}{n}$. (Yes, we saw this same formula earlier.)

Total sum of squares is the sum of squared deviations about the sample mean of the observed *y* values. That is, SST is the sum of squared deviations about the horizontal line at height $\bar{y}$.

In contrast, because the estimated regression line is the "line of best fit", the sum of squared deviations about the least squares line will always be smaller than the sum of squared deviations about *any* other line. In particular, SSE < SST, unless (as in the middle scatterplot above) $\hat{y}_i = \bar{y}$ is itself the least squares line.

The ratio $\dfrac{\text{SSE}}{\text{SST}}$ is the proportion of total variation that *cannot* be explained by the simple linear regression model, and $r^2 = 1 - \dfrac{\text{SSE}}{\text{SST}}$ (a number between 0 and 1) is the proportion of observed *y* variation which *can* be explained by the linear regression model. The higher the value of r^2, called the **coefficient of determination**, the more successful is the simple linear regression model in explaining variation in *y*.

The coefficient of determination can be written in a slightly different way by introducing a third sum of squares, **regression sum of squares**, given by $\text{SSR} = \sum \left(\hat{y} - \bar{y}\right)^2 = \text{SST} - \text{SSE}$.

Then, we have $r^2 = 1 - \dfrac{\text{SSE}}{\text{SST}} = \dfrac{\text{SST}}{\text{SST}} - \dfrac{\text{SSE}}{\text{SST}} = \dfrac{\text{SSR}}{\text{SST}}$.

You may wonder, "Why is the coefficient of determination denoted by r^2?"

And, "Is there any connection to the correlation coefficient *r* from back in chapter 5?"

The answer is, "Yes", and the connection is in the notation.

Specifically, coefficient of determination = (correlation coefficient)^2, $r^2 = (r)^2$.

The methods and conclusions of regression analysis can be applied both when the values of the independent variable are fixed in advance and when the values of the independent variable are random.

The derivations and interpretations are more straightforward in the former case, but can, for this same reason, lead to erroneous assumptions about cause and effect.

Recall from chapter 5: "Correlation" means "connection" or "relationship".
Correlation *does not* necessarily mean "cause".

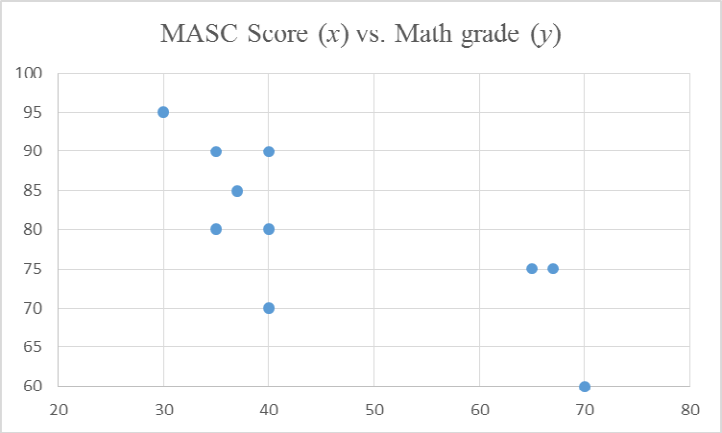
There *may* be a cause-and-effect relationship, but regression analysis alone will not tell us whether there is or not. A much more extensive and rigorous statistical analysis is required.

Example A: A paper in *Measurement and Evaluation in Counseling and Development* (Oct 90, pp. 121–127) discussed a survey instrument called the *Mathematics Anxiety Scale for Children* (MASC). Suppose the MASC was administered to ten fifth graders with the following results:

MASC Score	67	37	70	40	35	65	40	35	30	40
Math grade (%)	75	85	60	90	80	75	70	90	95	80

A-1. Use the formulas derived above to find the least-squares regression equation.

	x	y	xy	x^2	y^2
	67	75			
	37	85			
	70	60			
	40	90			
	35	80			
	65	75			
	40	70			
	35	90			
	30	95			
	40	80			
$\Sigma =$					



$S_{xy} =$

$S_{xx} =$

$b_1 =$

$b_0 =$

The least squares (estimated) regression equation is

Plot the estimated regression equation on the scatterplot above.

The estimated regression equation line can be used to 1) obtain a point estimate of the expected value of Y for a chosen $x = x^*$, or 2) predict the Y value for a single new observation made at a chosen $x = x^*$.

A-2. Estimate the expected value for the Math grade of students whose MASC score is 50.

In practice, it is necessary to avoid predicting Y values for x-values much outside the range of the sample. The best-fit line is based on the data at hand, and may not extrapolate outside that original data.

A-3. Estimate the variance for Math grades versus MASC score.

A-4. Calculate the coefficient of determination for Math grades versus MASC score.

Example O revisited: Earlier we found the estimated linear regression equation the long way. For yourself for practice recalculate the estimated linear regression equation using the shortcut formulas, and then find the variance and the coefficient of determination. *answers*: $\hat{y} = 0.58x + 3.06$; ≈ 0.17; ≈ 0.96