Regression Estimation and Nonlinear Transformations

Joel S Steele
Portland State University

Regression through Calculus with Partial Derivatives

Level-Level bivariate regression example

Within the regression framework, we are most interested in using a linear combination of parameters and variables to explain variance in our outcome of interest. The basic model takes the form of a line.

$$y = ax + b$$

or a more common expression in regression,

$$y_i = b_0 + b_1 x_i + \epsilon_i$$

Where b_0 and b_1 represent the intercept and slope respectively.

Parameter estimation

Below we will use the *least squares* criteria to find the optimal estimates of both the intercept and slope. While we most often use computers to do this, it may be instructive to see a small example of exactly such a function is *minimized* by hand.

Hand computation with Calculus

Example: Say that you are interested in whether or not a mother's level of education relates to her child's high school GPA.

The data:

- Mother's education: $X = [0, 1, 3, 4]$
- HS GPA: $Y = [3.0, 3.2, 3.3, 3.7]$

1. $(0, 3.0)$
2. $(1, 3.2)$
3. $(3, 3.3)$
4. $(4, 3.7)$

We know that the equation for a line is $y = ax + b$. If we only had two points we could solve for each parameter a and b exactly. However, above we have 4 points. Thus we have more information (*observations*) that we do unknowns (*parameters*). Technically, this means our hypothesized system is over-identified and to get an exact solution would require more parameters. But in this case, let's say that we like our line, and what's more, we believe in it (remember, parsimony is preferred). So, in practice we need to introduce some criteria with which to judge our linear model relative to the data. Thus, we need to define the error term, we will use expected, $ax + b$, minus observed y, which makes the error equation,

$$\epsilon_i = ax_i + b - y_i.$$

To minimize the sum of squared error we take this function and square it

$$\sum_i \epsilon_i^2 = \sum_i (ax_i + b - y_i)^2$$

Using our data this sum of squared errors can now be expressed as:

$$\begin{aligned} SS_e &= [(0a + b - 3.0)^2 && \text{values from point 1} \\ &\quad +(1a + b - 3.2)^2 && \text{values from point 2} \\ &\quad +(3a + b - 3.3)^2 && \text{values from point 3} \\ &\quad +(4a + b - 3.7)^2] && \text{values from point 4} \end{aligned}$$

Simplify and expand

$$\begin{aligned} SS_e &= [(b - 3.0)(b - 3.0)+ \\ &\quad (a + b - 3.2)(a + b - 3.2)+ \\ &\quad (3a + b - 3.3)(3a + b - 3.3)+ \\ &\quad (4a + b - 3.7)(4a + b - 3.7)] \end{aligned}$$

Multiply through and collect similar terms within each sub-expression

$$\begin{aligned} SS_e &= [(b^2 - 6b + 9)+ \\ &\quad (a^2 + 2ab - 6.4a + b^2 - 6.4b + 10.24)+ \\ &\quad (9a^2 + 6ab - 19.8a + b^2 - 6.6b + 10.89)+ \\ &\quad (16a^2 + 8ab - 29.6a + b^2 - 7.4b + 13.69)] \end{aligned}$$

Combine all sub-expressions and collect common terms

$$\begin{aligned} SS_e &= a^2 + 9a^2 + 16a^2 \\ &\quad +b^2 + b^2 + b^2 + b^2 \\ &\quad -6.4a - 19.8a - 29.6a \\ &\quad -6b - 6.4b - 6.6b - 7.4b \\ &\quad +2ab + 6ab + 8ab \\ &\quad +9 + 10.24 + 10.89 + 13.69 \end{aligned}$$

Simplify common terms

$$\begin{aligned} SS_e &= 26a^2 \\ &\quad +4b^2 \\ &\quad -55.8a \\ &\quad -26.4b \\ &\quad +16ab \\ &\quad +43.82 \end{aligned}$$

This is the equation for the sum of squared errors for our four observed points

$$SS_e = 26a^2 + 4b^2 - 55.8a - 26.4b + 16ab + 43.82$$

Take the partial derivative of this equation with respect to each parameter. For example, taking the partial derivative of the function SS_e with respect to a is presented below. It is important to note that since we are differentiating the equation based on the parameter a we only need to consider those terms that have the term a in them. We will be using the power rule, which states $\frac{d}{dx} = (x^n) = n \cdot x^{n-1}$.

$$\begin{aligned}
SS_e \text{ w.r.t. } a \quad &= 26a^2 && -55.8a && +16ab \\
\tfrac{\partial SS_e}{\partial a} \quad &= 26\left(2 \cdot a^1\right) && -55.8\left(1 \cdot a^0\right) && +16b\left(1 \cdot a^0\right) \\
\tfrac{\partial SS_e}{\partial a} \quad &= 26(2 \cdot a) && -55.8(1 \cdot 1) && +16b(1 \cdot 1)
\end{aligned}$$

$$\tfrac{\partial SS_e}{\partial a} = 52a - 55.8 + 16b$$

We rearrange it to look like our equation for a line and set this equal to zero, this gives us the minimum point for the equation, or where the change stops.

$$\begin{aligned}
\tfrac{\partial SSe}{\partial a} &= 52a + 16b - 55.8 \\
0 &= 52a + 16b - 55.8
\end{aligned}$$

Repeat for the parameter b

$$SS_e \text{ w.r.t. } b \quad = 4b^2 - 26.4b + 16ab$$

$$\begin{aligned}
\tfrac{\partial SSe}{\partial b} &= 8b - 26.4 + 16a \\
\tfrac{\partial SSe}{\partial b} &= 16a + 8b - 26.4 \\
0 &= 16a + 8b - 26.4
\end{aligned}$$

How the function changes with respect to a.

$$0 = 52a + 16b - 55.8 \tag{1}$$

How the function changes with respect to b.

$$0 = 16a + 8b - 26.4 \tag{2}$$

Solve for a in Equation 1

$$\frac{(55.8 - 16b)}{52} = a \tag{3}$$

Plug a into Equation 2 and solve for b

$$\begin{aligned}
0 &= 16 \cdot \left(\tfrac{(55.8-16b)}{52}\right) + 8b - 26.4 \\
0 &= 16 \cdot \tfrac{55.8}{52} - 16 \cdot \tfrac{16b}{52} + 8b - 26.4
\end{aligned}$$

move all of the b terms to one side of the equation

$$\begin{aligned}
26.4 - 16 \cdot \tfrac{55.8}{52} &= -16 \cdot \tfrac{16b}{52} + \tfrac{416b}{52} \\
9.23077 &= \tfrac{160b}{52} \\
9.23077 \cdot 52 &= 160b \\
480 &= 160b \\
\tfrac{480}{160} &= b
\end{aligned}$$

the intercept estimate

$$3 = b$$

Plug our estimate of b into Equation 3 and solve for a.

$$\begin{aligned}
\tfrac{(55.8-16\cdot 3)}{52} &= a \\
\tfrac{(55.8-48)}{52} &= a \\
\tfrac{7.8}{52} &= a \\
0.15 &= a
\end{aligned}$$

So the best fitting line is

$$y = 0.15x + 3$$

Or as commonly expressed in regression

$$\hat{y} = 3 + .015x$$

Why do we care? Well, another way to think about regression coefficients are as the partial derivatives with respect to each input. So in the equation,

$$GPA_{HS} \sim \beta_0 + \beta_1 ED_{mom} + \epsilon$$

The β_1 coefficient is equal to

$$\frac{\partial GPA_{HS}}{\partial ED_{mom}} = \beta_1,$$

which, from our model was estimated to be 0.15. So for an increase of one unit in a mother's level of education there would be a corresponding increase of 0.15 in GPA_{HS}. This type of regression is sometimes referred to as *Level-Level* regression, because, a change in the level of the input x results in a change in the level of the output y, while holding everything else (that is the other inputs) constant. We will see other types of regression below, but first a bit of a review.

Quick and dirty Power rules

Raising a number, say a to a power, b, then raising that quantity to the power c is the same as multiplying the powers together, thus $(a^b)^c = a^{bc}$. For example,

$$(2^2)^3 = 2^{2\times 3} = 2^6 = (2 \times 2) \times (2 \times 2) \times (2 \times 2) = 64.$$

Also of note, is that the product of the same value, or base, let's say 2, raised to different powers is equal to the base raised to the sum of the powers. For example,

$$2^2 \times 2^3 = 2^{2+3} = 2^5 = (2 \times 2) \times (2 \times 2 \times 2) = 32.$$

Quick and dirty Log rules

Remember that logs are meant to show the number of times a number, the *base*, is to be multiplied by itself to get a particular value. Put another way, what power of the *base* is needed to get the answer. Let's take an easy example. We will use 100, which can be expressed the following **equivalent** ways.

$$\begin{aligned} 100 &= 10^2 \\ &= 10 \times 10 \\ &= 1000/10 \end{aligned}$$

Now, if we work with *log* with a base of 10 we are interested in what power to raise 10 to in order to produce the result of 100

$$\begin{aligned} \text{if} \quad & 10^2 = 100 \\ \text{then} \quad & log_{10}(100) = 2 \end{aligned}$$

As we can see, 2 is the answer for base 10. We will just assume base 10 for the following rules:

power rule

$log(A^n) = n \times log(A)$

- $log(10^2) = 2 \times log(10) = 2 \times 1 = 2$

product rule

$log(A \times B) = log(A) + log(B)$

- $log(10 \times 10) = log(10) + log(10) = 1 + 1 = 2$

quotient rule

$log(\frac{A}{B}) = log(A) - log(B)$

- $log(\frac{1000}{10}) = log(1000) - log(10) = 3 - 1 = 2$

Nonlinear models

From here out, we will be looking at nonlinear trends and some of the ways that we approach modeling them. It is important to keep in mind that one of the major assumptions of regression is that the variables are linearly related. For the most part we will be trying to transform relations of the form,

$$y = x^{\beta} + \epsilon \tag{4}$$

into something that look more like the lines that we know, specifically $y = ax + b$. Figure 1 illustrates three possible trajectories for different powers.

Log-Level regression example

For this model the outcome will be transformed in order to make the model linear. In particular the hypothesized model takes the the form

$$y \sim \alpha e^{\beta x}. \tag{5}$$

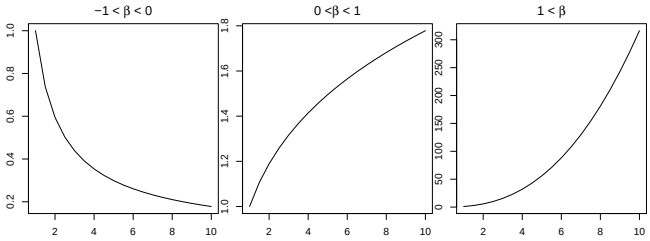

Figure 1. Nonlinear trajectories for different powers

Based on our knowledge of logs, in particular the **power rule** from above, we can take the log of both sides of Equation 5 to get [1],

$$\ln(y) \sim \ln(\alpha) + \beta x. \tag{6}$$

Table 1 contains example data that will be used to present this model [2].

Table 1
Log-level example data

	x	y
1	45.00	33.00
2	99.00	72.00
3	31.00	19.00
4	57.00	27.00
5	37.00	23.00
6	85.00	62.00
7	21.00	24.00
8	64.00	32.00
9	17.00	18.00
10	41.00	36.00
11	103.00	76.00

Now the question is how to interpret the resulting estimates. At first glace we are dealing with log changes in the outcome y for corresponding unit changes in x. This is where the term *Log-Level* comes from, and put simply, we can expect a 0.02% change in y for a 1 unit change in x.

[1] Here we take the natural log, or log base e.

[2] These example data, and others from http://www.real-statistics.com/regression/

Table 2
Log-level regression results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.64	0.12	21.83	0.00
x	0.02	0.00	8.19	0.00

A better understanding, at least in terms of raw scale units can be gained from *back transformation*. In particular, we must exponentiate our estimates if we want to get back to raw score levels. Thus, our equation estimates

$$\ln(y) \sim 2.64 + 0.02x, \tag{7}$$

would need to be transformed back as

$$\begin{aligned} y &\sim e^{2.64+0.02x} \\ &\quad e^{2.64} \times e^{0.02x} \end{aligned} \tag{8}$$

$$y \sim 14.0132 \times 1.0202^{x} \tag{9}$$

Now, the intercept term is the expected level of $\ln(y)$ when $x = 0$. In our equation above, the value is 14.0132, however the mean of our outcome is actually 38.3636. Let's see what happens when we mean center our predictor.

Table 3
Log-level with mean centered predictor

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.52	0.06	62.55	0.00
x_c	0.02	0.00	8.19	0.00

Based on this model the expected mean of the outcome is 33.7982. This is closer to our reported mean of 38.3636 but not exact... why? This is because here we are dealing with what is called the *Geometric* mean, rather than the mean we normally use. The *Geometric* mean is computed as,

$$\left(\prod_{i=1}^{N} x_i \right)^{1/N}. \tag{10}$$

When we compute the geometric mean of y we get 33.7982, which matches our estimate based on the model. Figure 2 compares the untransformed trajectory and the transformed trajectory.

Log-Log regression example

In this model, both the outcome and the predictor are log-transformed. That is because the predictor is raised to the power of a parameter, specifically

$$y = \alpha x^{\beta} \tag{11}$$

Thus, taking the log of both sides results in,

$$\ln(y) = \ln(\alpha) + \beta \ln(x) \tag{12}$$

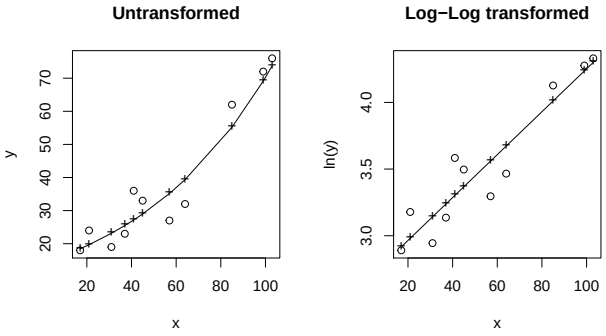

Figure 2. Log-level Observed vs predicted

Table 4
Log-Log example data

	x	y
1	8.10	33.00
2	69.90	49.00
3	4.20	19.00
4	14.10	27.00
5	5.60	23.00
6	52.10	51.00
7	44.60	34.00
8	19.60	32.00
9	33.00	28.00
10	6.70	36.00
11	30.10	43.00

Parameter interpretation

Interpretation of the model estimates in Table 5 is pretty straight-forward. We are dealing with percent changes in both the outcome and the predictor. In economics this is referred to as *elasticity*. So, based on the model a 1% change in x would result in an 0.23% change in y. Figure 3 compares the untransformed trajectory and the transformed trajectory.

A caution about nonlinear transformations

Above, we've discussed power-*ish* transformations, notice that in Figure 1 the basic equation was $y = x^{\beta}$. The transformations just illustrated only really make sense if the ratio of largest to smallest value on the raw scale is large. If it's not, then something like the natural log will have little effect on

Table 5
Log-Log regression results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.81	0.21	13.65	0.00
log(x)	0.23	0.07	3.44	0.01

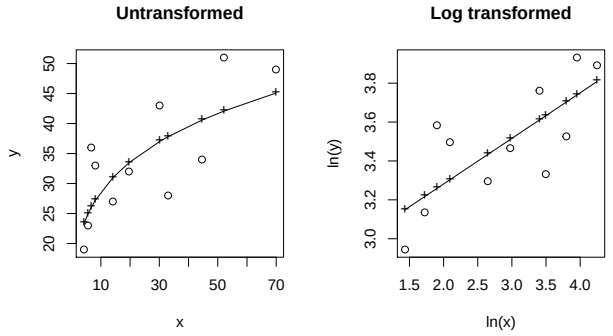

Figure 3. Log-Log Observed vs predicted

the relation.

Also, if the values are negative, it will be necessary to add a constant before taking the log of the values.

Lastly, these transformations should only be used when the relationship is *monotonic*, meaning it passes the horizontal line test, or is a *one-to-one* function. This means that quadratic trends or trends that oscillate are not good candidates for transformation. When relations are *monotonic* these transformations will not change the *rank-order* of the observations, just the spaces in between successive values.

Proportions and the Binomial models

There is a special case involved around **binary** outcomes. In general power transformations don't work well when the data values are near 0 or 1, which is exactly the case for binary data. Think about coding *pass-fail* or *True-False* outcomes. In this case we need to develop a way in which to transform these 0/1 values into something manageable.

The Logistic curve. We will be using a function called the *logistic curve* which has the functional specification of,

$$y = \frac{e^{\theta}}{1 + e^{\theta}} \tag{13}$$

Where θ represents all of the possible inputs of interest for predicting y. From our *level-level* regression example, *theta* may be equal to $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ thus making the equation

$$y = \frac{e^{(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}} \tag{14}$$

With this approach we are modeling proportions. So, instead of trying to use either 0 or 1 as an outcome directly, we will be looking at the total number of 1s out of all responses. A little later we will specify this as $Pr(Y = 1)$ or the probability of scoring a 1 on the outcome. For binary data this follows our logistic curve.

However, we can still model proportions directly as well, as shown below.

Dose-Response example

These data are a reproduction of data from:

C.I. Bliss (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology, 22*(1), 134-167. [3]

The Data Beetles were exposed to carbon disulphide at varying concentrations for 5 hours.

- dose = mf/L concentration of CS2
- nexp = number of beetles exposed
- ndied = number of beetles killed
- prop = proportion of dead to exposed beetles

Table 6
Beetle data

	dose	nexp	ndied	prop	nalive
1	49.10	59.00	6.00	0.10	53.00
2	53.00	60.00	13.00	0.22	47.00
3	56.90	62.00	18.00	0.29	44.00
4	60.80	56.00	28.00	0.50	28.00
5	64.80	63.00	52.00	0.82	11.00
6	68.70	59.00	53.00	0.90	6.00
7	72.60	62.00	61.00	0.98	1.00
8	76.50	60.00	60.00	1.00	0.00

The Logistic Model

Run a logistic regression of the proportion of dead to living beetles as a function of the dose of CS2 gas. Our model specification is,

$$\frac{n_{died}}{n_{alive}} \sim b_0 + b_1 dose \tag{15}$$

[3] Many thanks to Thaddeus Tarpey at Wright University. Check out his cite for this and more http://www.wright.edu/ thaddeus.tarpey/

Table 7
Logistic model results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-14.8230	1.2896	-11.49	0.0000
dose	0.2494	0.0214	11.66	0.0000

we may be interested in finding the concentration of CS2 gas that is lethal 50% of the time, the **LD50**.

Note that if we have a function with multiple predictors we can solve for each variable using something similar. For example if,

$$y \sim b_0 + b_1(x_1) + b_2(x_2) + b_3(x_3)$$

is the model. Then to find a specific value for one of the predictors $(x1, x2, x3)$ that corresponds to a desired probability y.

- $x_1 = (-b_0 - b_2 - b_3 + log\left(\frac{-y}{(y-1)}\right))/b1$
- $x_2 = (-b_0 - b_1 - b_3 + log\left(\frac{-y}{(y-1)}\right))/b2$
- $x_3 = (-b_0 - b_1 - b_2 + log\left(\frac{-y}{(y-1)}\right))/b3$

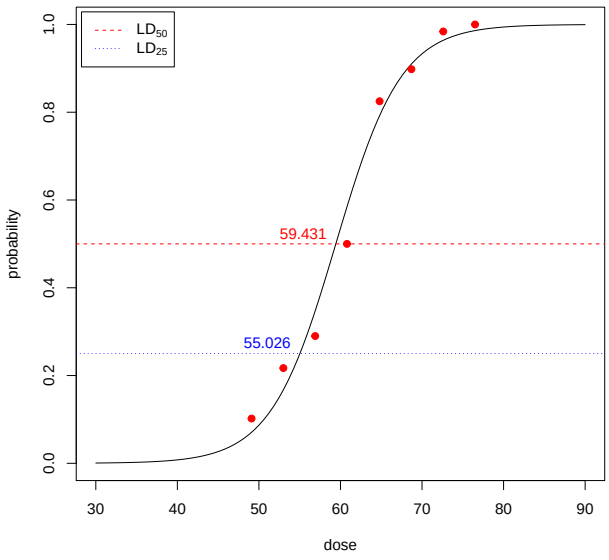

Figure 4. Example Dose-Response curve

Below is a table to help you understand the different types of transformations available and how to interpret them.

Name	Outcome	Input	Form	β_1	interpretation
Level-Level	Y	X	$y \sim \beta_0 + \beta_1 x + \epsilon$	$\Delta y = \beta_1 \Delta x$	1 unit change in x give β_1 unit change in y
Level-Log	Y	$ln(X)$	$y \sim \beta_0 + \beta_1 ln(x) + \epsilon$	$\Delta y = \beta_1 \%\Delta x$	1% change in x give β_1 unit change in y
Log-Level	$ln(Y)$	X	$ln(y) \sim \beta_0 + \beta_1 x + \epsilon$	$\%\Delta y = \beta_1 \Delta x$	1 unit change in x gives β_1% change in y
Log-Log	$ln(Y)$	$ln(X)$	$ln(y) \sim \beta_0 + \beta_1 ln(x) + \epsilon$	$\%\Delta y = \beta_1 \%\Delta x$	1% change in x give β_1% change in y

(16)