15
Exponential growth, power laws, and the log transformation

Exponential growth and decay

Beginning in March 2014, West Africa experienced the largest outbreak of the Ebola virus in history. Guinea, Liberia, Niger, Sierra Leone, and Senegal were all hit hard by the epidemic. Figure 15.1 shows the number of laboratory-confirmed cases of Ebola in these five countries over time, beginning on March 25.

If we wanted to fit a model to describe how the number of Ebola infections grew over time, we might be tempted to fit a polynomial function (since a linear model clearly won't work well here). However, basic biology tells us that the transmission rate of a disease through a population is reasonably well described by an exponential growth model: 1 infection leads to 2, which lead to 4, which lead to 8, to 16, and so on. The equation for an exponential-growth model is

$$y = \alpha \cdot e^{\beta t}, \tag{15.1}$$

where y is the expected number of cases and t is the number of time intervals (e.g. weeks or days) since the start of the outbreak.

It turns out that we can use least squares to fit an exponential growth model of this form, using a new trick: *take the logarithm of the response variable* and fit a linear model to this new transformed variable. We can see why this works if we take the logarithm of y in the equation for exponential growth (labeled 15.1, above). To preserve equality, if we take the log of the left-hand side, we also have to take the log of the right-hand side:

$$\begin{aligned} \log y &= \log\left(\alpha \cdot e^{\beta_1 t}\right) \\ &= \log \alpha + \beta t\,. \end{aligned}$$

The second equation says that the log of y is a linear function of the time variable, t, with intercept $\beta_0 = \log \alpha$ and slope β_1.

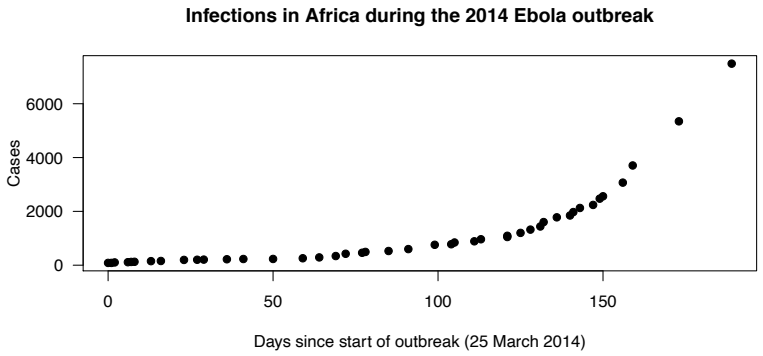

Figure 15.1: Cases of Ebola over time in West Africa, 2014. Compiled from CDC reports by Francis Smart, as described here.

Thus to fit the exponential growth model for any response variable y, we need to follow two steps:

(1) Define a new variable $z = \log y$ by taking the logarithm of the original response variable.
(2) Fit a linear model for the transformed variable z versus the original predictor, using ordinary least squares.

Figure 15.2 shows the result of following these two steps for the Ebola data. The left panel shows the straight-line fit on the log scale:

$$\log \text{Cases} = 4.54 + 0.021 \cdot \text{Days}\,.$$

The right panel shows the corresponding exponential-growth curve on the original scale:

$$\text{Cases} = 93.5 \cdot e^{0.021 \cdot \text{Days}}\,.$$

The leading constant is calculated from the intercept on the log scale: $93.5 \approx e^{4.54}$. From Figure 15.2, we can see that the exponential-growth model fits adequately, although imperfectly: the rate of growth seems to be accelerating at the right of the picture, and the upward trajectory is visibly nonlinear on the log scale. (Remember: all models are wrong, but some models are useful.)

An exponential model with a negative slope β_1 on the log scale is called an exponential decay model. Exponential decay is a good model for, among other things, the decay of a radioactive isotope.

Interpreting the coefficient in an exponential model. To interpret the coefficient in an exponential growth model, we will use it to calculate the doubling time—that is, how many time steps it takes for the response variable (here, Ebola cases) to double.

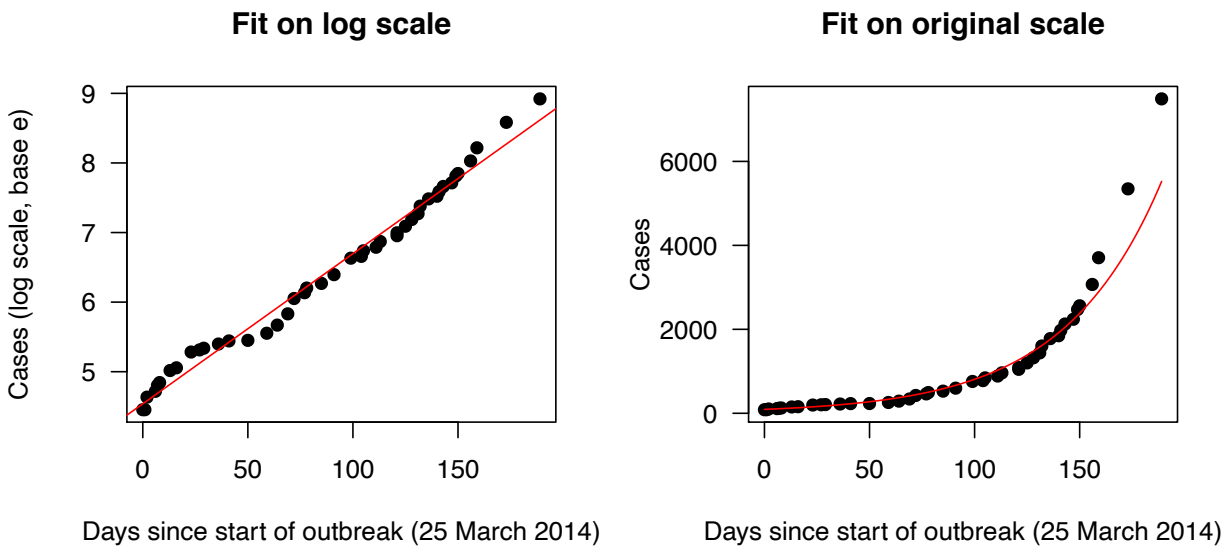

Figure 15.2: An exponential-growth model fit to the Ebola data by ordinary least squares, where the y variable is shown on the log scale (left) and on the original scale (right).

In terms of our estimated model, the number of cases doubles between days t_1 and t_2 whenever

$$\frac{\alpha e^{\beta_1 t_2}}{\alpha e^{\beta_1 t_1}} = 2\,,$$

so that the number of cases on day t_2 (in the numerator) is precisely twice the number of cases on day t_1, in the denominator. If we simplify this equation using the basic rules of algebra for exponentials, we find that the number of days that have elapsed between t_1 and t_2 is

$$t_2 - t_1 = \frac{\log 2}{\beta_1}\,.$$

This is our doubling time. For Ebola in West Africa, the number of cases doubled roughly every

$$\frac{\log 2}{0.021} \approx 32$$

days during the spring and early summer of 2014.

In an exponential decay model (where $\beta_1 < 0$), a similar calculation would tell you the half life, not the doubling time.[1]

[1] Instead, solve the equation

$$\frac{\alpha e^{\beta_1 t_2}}{\alpha e^{\beta_1 t_1}} = 1/2$$

for the difference $t_2 - t_1$.

Power laws: double log transformations

In some cases, it may be best to take the log of both the predictor and the response, and to work on this doubly transformed scale. For example, in the upper left panel of Figure 15.3, we see a scatter plot of brain weight (in grams) versus body weight (in kilos) for 62 different mammalian species, ranging from the lesser short-tailed shrew (weight: 10 grams) to the African elephant (weight: 6000+ kilos). You can see that most species are scrunched up in a small box at the lower left of the plot. This happens because the observations span many orders of magnitude, and most are small in absolute terms.

But if we take the log of both body weight and brain weight, as in the top-right panel of Figure 15.3, the picture changes considerably. Notice that, in each of the top two panels, the red box encloses the same set of points. On the right, however, the double log transformation has stretched the box out in both dimensions, allowing us to see the large number of data points that, on the left, were all trying to occupy the same space. Meanwhile, the two points outside the box (the African and Asian elephants) have been forced to cede some real estate to the rest of Mammalia.

This emphasizes the taking the log is an "unsquishing" operator. To see this explicitly, look at the histograms in the second and third row of panels in Figure 15.3. Whenever the histogram of a variable looks highly skewed right, as on the left, a log transformation is worth considering. It will yield a much more nicely spread-out distribution of points, as on the right.

Power laws. It turns out that when we take the log of both variables, we are actually fitting a *power law* for the relationship between y and x. The equation of a power law is

$$y = \alpha \cdot x^{\beta_1}$$

for some choices of α and β. This is a very common model for data sets that span many orders of magnitude (like the body/brain weight data). To see the connection with the double log transformation, simply take the logarithm of both sides of the power law:

$$\begin{aligned} \log y &= \log\left(\alpha \cdot x^{\beta_1}\right) \\ &= \log \alpha + \log x^{\beta_1} \\ &= \log \alpha + \beta_1 \log x\,. \end{aligned}$$

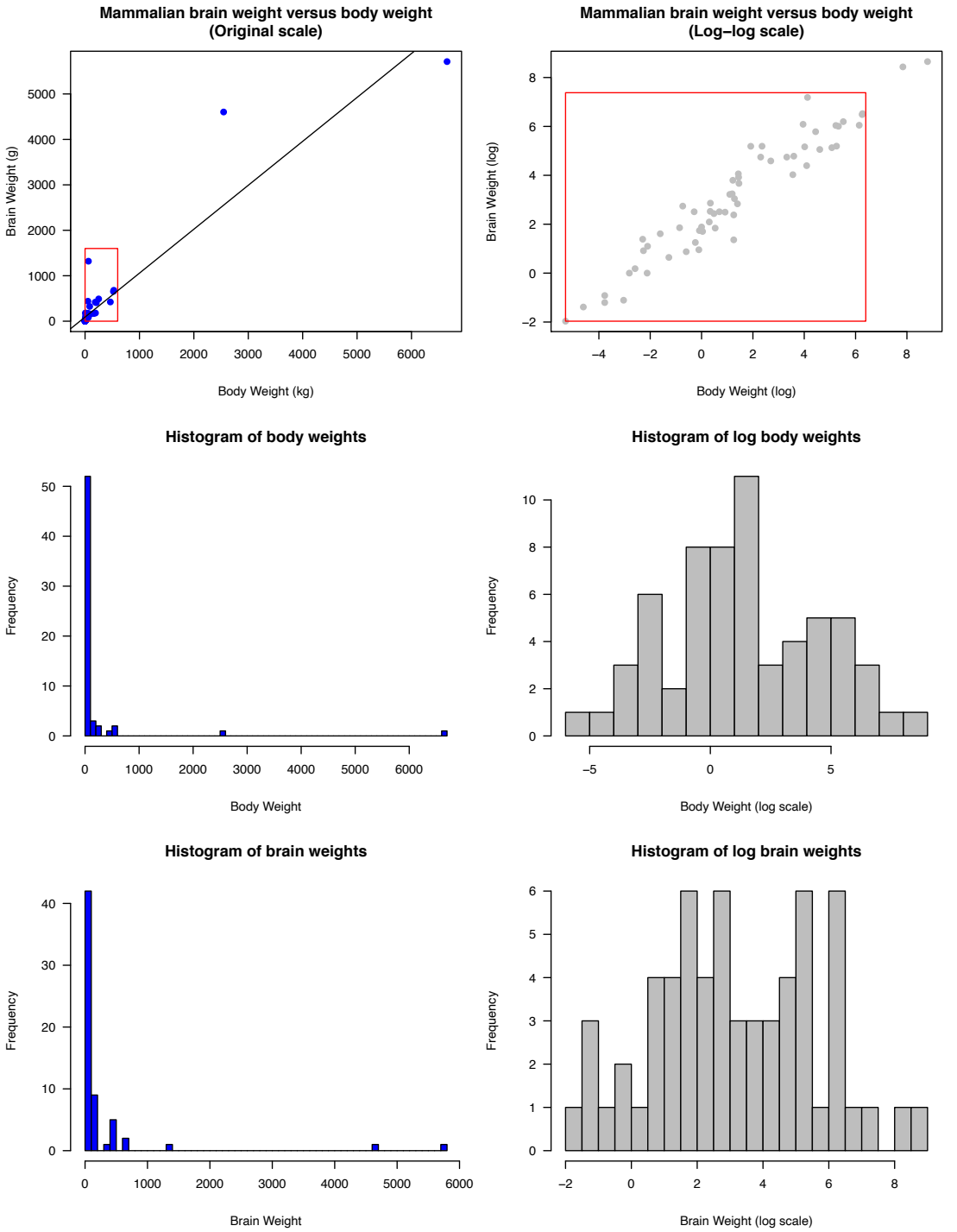

Figure 15.3: Brain weight versus body weight for 62 mammalian species, both on the original scale and the log scale. Notice how the log transformation "unsquishes" the points.

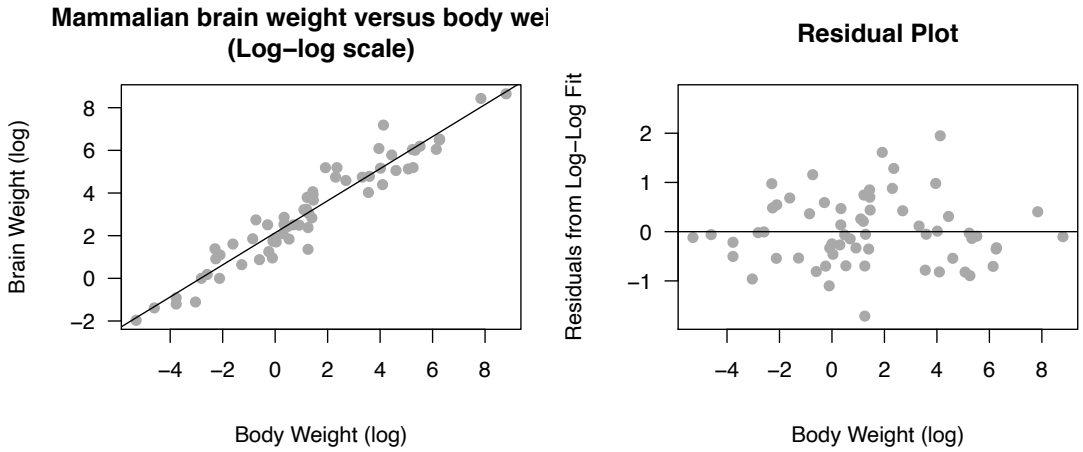

Figure 15.4: A straight-line fit to the mammalian brain weight data after a double log transformation.

Therefore, if y and x follow a power law, then $\log y$ and $\log x$ follow a linear relationship with intercept $\log \alpha$ and slope β_1. This implies that we can fit the parameters of the power law by applying the double log transformation and using ordinary least squares. For our mammalian brain weight data, applying this recipe yields the fitted equation

$$\log \text{brain} = 2.13 + 0.75 \cdot \log \text{body}\,,$$

or expressed as a power law on the original scale,

$$\text{brain} = 8.4 \cdot \text{body}^{0.75}\,.$$

The residuals in a power-law model. As we've just seen, we can fit power laws using ordinary least squares after a log transformation of both the predictor and response. In introducing this idea, we ignored the residuals and focused only on the part of the model that describes the systematic relationship between y and x. If we keep track of these residuals a bit more carefully, we see that the model we're fitting for the ith response variable is this:

$$\log y_i = \log \alpha + \beta_1 \log x_i + e_i\,, \tag{15.2}$$

where e_i is the amount by which the fitted line misses $\log y_i$. We suppressed these residuals before the lighten the algebra, but now we'll pay them a bit more attention.

Equation 15.2 says that the residuals affect the model in an additive way on the log scale. But if we exponentiate both sides,

we find that they affect the model in a multiplicative way on the original scale:

$$\begin{aligned} \exp(\log y_i) &= \exp(\log \alpha) \cdot \exp(\beta_1 \log x) \exp(e_i) \\ y_i &= \alpha x^{\beta_1} \exp(e_i) \,. \end{aligned}$$

Therefore, in a power low, the exponentiated residuals describe the percentage error made by the model on the original scale. Let's work through the calculations for two examples:

- If $e_i = 0.2$ on the log–log scale, then the actual response is $\exp(0.2) \approx 1.22$ times the value predicted by the model. That is, our model underestimates this particular y_i by 22%.
- If $e_i = -0.1$ on the log–log scale, then the actual response is $\exp(-0.1) \approx 0.9$ times the value predicted by the model. That is, our model overestimates this particular y_i by 10%.

The key thing to realize here is that the *absolute* magnitude of the error will therefore depend on whether the y variable itself is large or small. This kind of multiplicative error structure makes perfect sense for our body–brain weight data: a 10% error for a lesser short-trailed shrew will have us off by a gram or two, while a 10% error for an elephant will have us off by 60 kilos or more. Bigger critters mean bigger errors—but only in an absolute sense, and not if we measure error relative to body weight.

Interpreting the slope under a double log transformation. To correctly interpret the slope β_1 under a double log transformation, we need a little bit of calculus. The power law that we want to fit is of the form $y = \alpha x^{\beta_1}$. If we take the derivative of this expression, we get

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \beta_1 \alpha x^{\beta_1 - 1} \,.$$

We can rewrite this as

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}x} &= \frac{\beta_1 \alpha x^{\beta_1}}{x} \\ &= \beta_1 \frac{y}{x} \,. \end{aligned}$$

If we solve this expression for β_1, we get

$$\beta_1 = \frac{\mathrm{d}y/y}{\mathrm{d}x/x} \,. \tag{15.3}$$

Since the dy in the derivative means "change in y", the numerator is the rate at which the y variable changes, as a fraction of its value. Similarly, since dx means "change in x", the denominator is the rate at which the x variable changes, as a fraction of its value.

Putting this all together, we find that β_1 measures the ratio of percentage change in y to percentage change in x. In our the mammalian brain-weight data, the least-squares estimate of the slope on a log-log scale was $\widehat{\beta}_1 = 0.75$. This means that, among mammals, a 100% change (i.e. a doubling) in body weight is associated with a 75% expected change in brain weight. The bigger you are, it would seem, the smaller your brain gets—at least relatively speaking.

The coefficient β_1 in a power law is often called an *elasticity* parameter, especially in economics, where it is used to quantify the responsiveness of consumer demand to changes in the price of a good or service. The underlying model for consumer behavior that's often postulated is that

$$Q = \alpha P^{\beta_1},$$

where Q is the quantity demanded by consumers, P is the price, and $\beta_1 < 0$. Economists would call β_1 the price elasticity of demand,[2] which may be a familiar concept from a microeconomics course.

[2] They actually define elasticity as the ratio in Equation 15.3, but as we've seen, this is mathematically equivalent to the regression coefficient you get when you fit the x–y relationship using a power law.