# Supplement: Interpreting Parameters after Transformation

36-401, Fall 2015, Section B

After unscheduled class discussion, 29 September 2015

## 1 Transformed Predictor

The model becomes

$$Y = \beta_0 + \beta_1 f(X) + \epsilon \tag{1}$$

for some invertible, nonlinear function $f$, with $\epsilon$ still IID Gaussian. The usual interpretations apply, but now are all in terms of $f(x)$, not $x$:

1. We can never find coefficients $\gamma_0, \gamma_1$ where

   $$\beta_0 + \beta_1 f(x) = \gamma_0 + \gamma_1 x \tag{2}$$

   for all $x$. That is to say, applying a nonlinear transformation to the predictor doesn't just amount to making some adjustment to the slope and intercept.

2. $\beta_0 = \mathbb{E}\left[Y|f(X) = 0\right]$. This is (usually) not $\mathbb{E}\left[Y|X = 0\right]$.

   (a) $\beta_0$ is still the intercept when $f(x)$ goes on the horizontal axis.

   (b) Instead, $\mathbb{E}\left[Y|X = 0\right] = \beta_0 + \beta_1 f^{-1}(0)$.

3. $\beta_1$ is the slope in units of $Y$ per units of $f(X)$. That is, it's the difference in the expected response for a difference in $f(x)$ of 1, not for a difference in $x$ of 1.

   (a) A difference of 1 in $x$ predicts a difference of $\beta_1(f(x) - f(x-1))$ in $\mathbb{E}\left[Y\right]$ if $x$ decreases by 1, and a difference of $\beta_1(f(x+1) - f(x))$ if $x$ increases by 1. (These are generally not the same.) So even the response to increases and decreases isn't necessarily of the same size.

   (b) Very small differences in $x$, of size $h$, predict very small differences in $\mathbb{E}\left[Y\right]$, of size $h\beta_1 \frac{df}{dx}(x)$. So there is a slope at each point, but it changes. (That's what makes $f$ non-linear.)

4. $\sigma^2$ is still the variance of the Gaussian noise around the regression curve, but now that curve really is curved and not a straight line.

(a) A plot of $y_i$ against $x_i$ should not be a straight line.

(b) A plot of $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 f(x_i))$ should still be a flat line around 0.

We estimate the model by transforming the data, going from $(x_1, y_1), \ldots (x_n, y_n)$ to $(f(x_1), y_1), \ldots (f(x_n), y_n)$, and then running a regression of $y_i$ on $f(x_i)$.

### 1.1 Special case: Log transformation of the predictor

Suppose we select $f = \log$. Our model then is

$$Y = \beta_0 + \beta_1 \log X + \epsilon \tag{3}$$

In this setting,

- $\log X = 0$ means $X = 1$, so $\beta_0 = \mathbb{E}[Y|X = 1]$.
- A $k$ unit change in $\log x$ means multiplying $x$ by $e^k$:

  $$k + \log x = \log e^k + \log x = \log xe^k \tag{4}$$

  Hence, $\beta_1$ is the expected difference in $Y$ for an $e$-fold change in $X$.
- The slope of $\mathbb{E}[Y]$ with respect to $X$ decreases in $x$:

  $$\frac{d\mathbb{E}[Y|X = x]}{dx} = \frac{\beta_1}{x} \tag{5}$$

## 2 Transforming the response

Again, we select an invertible, non-linear function $g$, and transform the response variable:

$$g(Y) = \beta_0 + \beta_1 X + \epsilon \tag{6}$$

All of the parameters retain their old interpretations in terms of $g(Y)$. None of them have their old interpretations in terms of $Y$. This is because the model for $Y$ is now

$$Y = g^{-1}(\beta_0 + \beta_1 X + \epsilon) \tag{7}$$

1. We can never find coefficients $\gamma_0, \gamma_1$ where

   $$\beta_0 + \beta_1 x = g(\gamma_0 + \gamma_1 x) \tag{8}$$

   for all $x$. That is to say, applying a nonlinear transformation to the response doesn't just amount to making some adjustment to the slope and intercept.

2. $\beta_0 = \mathbb{E}[g(Y)|X = 0]$. Note that since $g$ is not a linear function, neither is $g^{-1}$, and so $\mathbb{E}[Y|X = 0] \neq g^{-1}(\beta_0)$.

3. More generally, $\mathbb{E}[Y|X = x] \neq g^{-1}(\beta_0 + \beta_1 x)$.

   (a) Since we're assuming $\epsilon$ is Gaussian centered at 0, the median value of $\epsilon = 0$. Therefore, according to the model,

$$\mathbb{P}\left(g(Y) \leq \beta_0 + \beta_1 x \mid X = x\right) = 0.5 \tag{9}$$

   Since $g$ is invertible, it is therefore also true that

$$\mathbb{P}\left(Y \leq g^{-1}(\beta_0 + \beta_1 x) \mid X = x\right) = 0.5 \tag{10}$$

   and the transformation can be simply undone for the conditional median, but not the conditional mean.

   (b) In particular, $g^{-1}(\beta_0)$ is the conditional median of $Y$ when $X = 0$.

4. $\beta_1$ is the difference in the mean of $g(Y)$ predicted by a 1 unit change in $X$.

   (a) There is generally no simple interpretation of $\beta_1$ for the original $Y$.

   (b) By the argument above, increasing $x$ by $h$ predicts that the conditional *median* will change by $g^{-1}(\beta_0 + \beta_1 x + \beta_1 h) - g^{-1}(\beta_0 + \beta_1 x)$. This, generally speaking, does not simplify.

   (c) When the change $h$ is very small, the change to the conditional median will tend towards $h\beta_1 \left.\frac{dg^{-1}(u)}{du}\right|_{u=\beta_0+\beta_1 x}$.

5. $\sigma^2$ is the variance of the Gaussian noise around the regression line for $g(Y)$.

   (a) Because $g^{-1}$ is a nonlinear transformation, the noise around the regression curve for $Y$, $g^{-1}(\beta_0+\beta_1 x)$, will not (in general) be Gaussian.

   (b) In fact, the noise around that curve will generally not have mean zero, or constant variance, or even just be an additive perturbation around the curve. (See the example for the log transformation below.)

   (c) A plot of $Y$ against $X$ will not show a straight line, though a plot of $g(Y)$ against $X$ should.

   (d) Residuals for $Y$, calculated as $y_i - g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$ need not be centered at 0, or have constant variance, etc., etc.

   (e) Residuals for the transformed response, calculated as $g(y_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$, should show all the usual properties.

## 2.1 Special case: log transformation of the response

If we select $g = \log$, the model becomes

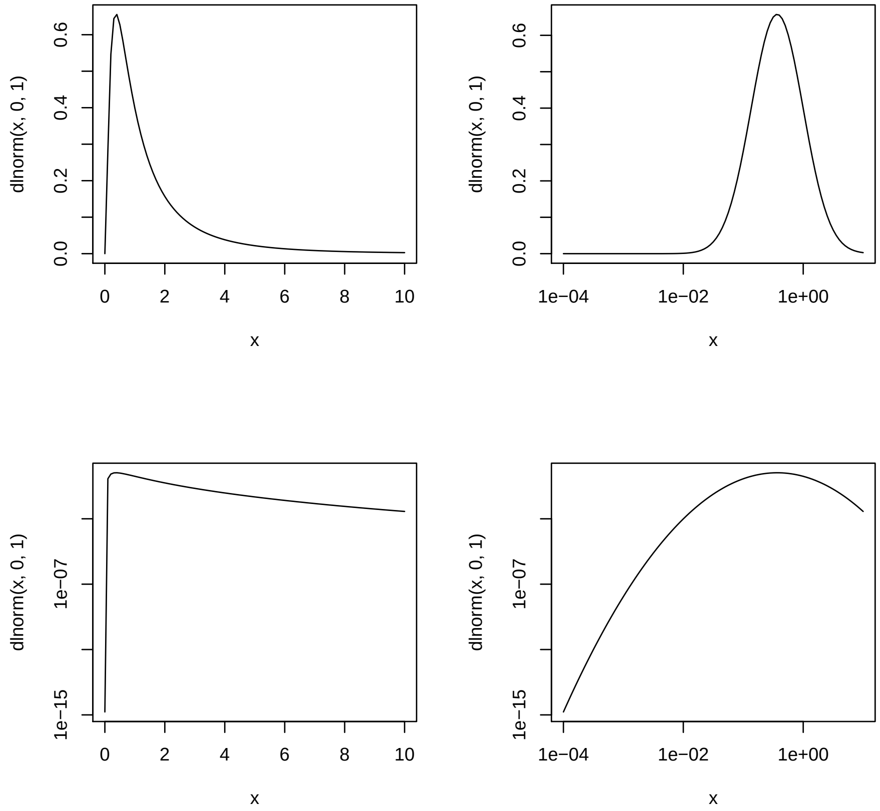
$$\log Y = \beta_0 + \beta_1 X + \epsilon \tag{11}$$

and the interpretation simplifies slightly, especially on taking the inverse transformation:

$$Y = e^{\beta_0 + \beta_1 X + \epsilon} = e^{\beta_0} e^{\beta_1 x} e^{\epsilon} \tag{12}$$

1. $e^{\beta_0}$ is the median value of $Y$ when $X = 0$. It is common to abbreviate it as a single number, say $y_0$.
2. A one-unit increase in $x$ predicts that $Y$ should be larger by a *factor* of $e^{\beta_1}$. That is, additive, equal-size changes to $x$ lead to multiplicative changes in $Y$.
3. The slope of $Y$ with respect to $x$ is $\beta_1 e^{\beta_1 x}$, so, again, there is no one answer to "what gets added to $Y$ when $x$ changes a little?"
4. Because $\epsilon \sim N(0, \sigma^2)$, $e^\epsilon$ is not Gaussian (no matter what mean and variance we might try). Rather, we say that $e^\epsilon$ is **log-normal** or **log-gaussian**, because it's log is normal or Gaussian[1] (R functions: `dlnorm`, `plnorm`, `qlnorm`, `rlnorm`). This is written $e^\epsilon \sim LN(0, \sigma^2)$, i.e., the log-normal is parameterized by the mean and variance of its log.
   (a) $e^\epsilon \geq 0$, so the $LN$ distribution is supported on the positive numbers, not (like the $N$) the whole number line.
   (b) It further follows that $\mathbb{E}\left[e^\epsilon\right] > 0$.
   (c) By the argument above, the median of $e^\epsilon = 1$.
   (d) Because making $\epsilon < 0$ can only decrease $e^\epsilon$ a little below 1 (at worst to 0), but making $\epsilon > 0$ can increase $e^\epsilon$ by a lot (up to $\infty$), the distribution is skewed to the right.
   (e) By directly using the transformation-of-variables formula (which you remember from your probability class), the probability density function of an $LN(\mu, \sigma^2)$ distribution is
   $$\frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(\mu - \log x)^2}{\sigma^2}} \tag{13}$$
   (Figure 1). By integration, then, one can show that the expectation of this distribution is $e^{\mu + \sigma^2/2}$, and its variance is $e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$.
   (f) Specifically, $\mathbb{E}\left[e^\epsilon\right] = \exp \sigma^2/2$.
   (g) Similarly, $\mathrm{Var}\left[e^\epsilon\right] = e^{\sigma^2}(e^{\sigma^2} - 1)$.
   (h) The noise $e^\epsilon$ *multiplies* the deterministic function $e^{\beta_0 + \beta_1 x}$, it does not add to it. Therefore we have
   $$\begin{aligned}
   \mathbb{E}\left[Y|X=x\right] &= \mathbb{E}\left[e^{\beta_0+\beta_1 x + \epsilon} \mid X = x\right] && (14)\\
   &= e^{\beta_0} e^{\beta_1 x} \mathbb{E}\left[e^\epsilon \mid X = x\right] && (15)\\
   &= e^{\beta_0} e^{\beta_1 x} \mathbb{E}\left[e^\epsilon\right] && (16)\\
   &= e^{\sigma^2/2} e^{\beta_0} e^{\beta_1 x} && (17)\\
   \mathrm{Var}\left[Y|X=x\right] &= \mathrm{Var}\left[e^{\beta_0+\beta_1 x} e^\epsilon \mid X = x\right] && (18)\\
   &= e^{2(\beta_0+\beta_1 x)} \mathrm{Var}\left[e^\epsilon \mid X = x\right] && (19)\\
   &= e^{\sigma^2}(e^{\sigma^2} - 1) e^{2(\beta_0+\beta_1 x)} && (20)
   \end{aligned}$$

[1] Some people prefer to call $e^\epsilon$ "anti-log-normal", which has a kind of logic to it, but they're very much a minority.



```
par(mfrow=c(2,2)) # Set up 2x2 plotting grid
curve(dlnorm(x,0,1),from=1e-4,to=10)
curve(dlnorm(x,0,1),from=1e-4,to=10, log="x")
curve(dlnorm(x,0,1),from=1e-4,to=10, log="y")
curve(dlnorm(x,0,1),from=1e-4,to=10, log="xy")
par(mfrow=c(1,1)) # Restore usual plot playout for later
```

FIGURE 1: *Probability density function of the standard log-normal distribution, $LN(0,1)$. Top left: ordinary (linear) scale on both axes; top right: log scale on horizontal axis; bottom left: log scale on vertical axis; bottom right: log scale on both axes.*

# 3 Transforming both predictor and response

The model is

$$g(Y) = \beta_0 + \beta_1 f(X) + \epsilon \qquad (21)$$

All the considerations of both the previous sections apply.

1. If this model applies, no linear model also applies.
2. $\beta_0 = \mathbb{E}\left[g(Y)|f(X) = 0\right]$.
3. $\beta_1$ is the slope of the curve of $g(Y)$ against $f(X)$.
   (a) There is generally no simple way to express this in terms of the original variables.
   (b) There is also generally no simple way to write the slope of the curve of $Y$ on $X$.
4. $\sigma^2$ is the variance of the Gaussian noise around the line of $g(Y)$ against $f(X)$. The distribution of $Y$ around its curve against $f(X)$, let alone against $X$, is generally not even additive.
5. The function $g^{-1}(\beta_0 + \beta_1 f(x))$ continues to give the conditional median of $Y$.

## 3.1 Special case: log of both predictor and response

The model is

$$\log Y = \beta_0 + \beta_1 \log X + \epsilon \qquad (22)$$

Undo the log on both sides:

$$Y = e^{\beta_0} X^{\beta_1} e^{\epsilon} \qquad (23)$$

Because this says that $Y$ is some power of $X$, up to noise, this sort of model is often called a **power law**.

Abbreviate $e^{\beta_0}$ by $y_0$. Then, in the power law model with log-normal noise,

1. $y_0$ is the median value of $Y$ when $X = 1$.
2. $\beta_1$ is the slope of $\log Y$ against $\log X$. It is also power to which we raise $X$ to get the systematic part of $Y$.
   (a) In the jargon, one says that "Y **scales like** $X^{\beta_1}$".
   (b) Ignore the noise temporarily, so we have a deterministic relationship $y = y_0 x^{\beta_1}$. A small difference in $x$, say $dx$, means a fractional difference of $dx/x$. The ratio of the fractional difference in $y$, $dy/y$ to the

fractional difference in $x$, sometimes called the **elasticity** of $y$ with respect to $x$, is $(dy/y)/(dx/x) = (dy/dx)/(y/x)$, is[2]

$$\frac{\frac{dy}{dx}}{\frac{y}{x}} = \frac{y_0\beta_1 x^{\beta_1 - 1}}{\frac{y_0 x^{\beta_1}}{x}} = \beta_1 \tag{24}$$

3. $\mathbb{E}\left[Y|X=x\right] \neq y_0 x^{\beta_1}$.

   (a) $y_0 x^{\beta_1}$ is the conditional median, however.

   (b) By parallel reasoning to what we went through above with the log-normal,

$$\begin{aligned} \mathbb{E}\left[Y|X=x\right] &= y_0 x^{\beta_1} e^{\sigma^2/2} & (25) \\ \mathrm{Var}\left[Y|X=x\right] &= y_0^2 x^{2\beta_1} e^{\sigma^2}(e^{\sigma^2}-1) & (26) \end{aligned}$$

Notice that this is a *different* statistical model from

$$Y = y_0 X^{\beta_1} + \epsilon \tag{27}$$

which would lead to $\mathbb{E}\left[Y|X=x\right] = y_0 X^{\beta_1}$, $\mathrm{Var}\left[Y|X=x\right] = \sigma^2$. (It is, unfortunately, often unclear whether people mean a power law with multiplicative, log-normal noise or with additive, normal noise.) I will leave it as an exercise to check how the interpretations change.

[2] If you are appalled by expressions like $dy/y$, you have good mathematical taste, and are invited to reach this conclusion through a proper proof using limits. (It can be done.)