Chapter 19

Estimating the dose-response function

"Dose-response"

Students and former students often tell me that they found evidence for dose-response. I know what they mean because that's the phrase I have also been taught to use, many years ago. It is, however, a poorly stated idea.

Let E be a continuous exposure variable, and let D be the disease variable or some function of it such as mean, probability, and log-odds. Let $D=f(E)$ — with causality in mind, not prediction. Then the function I just defined is a dose-response function. When specified, it tells us how the value of D "responds" to the "dose" of E, where "responds" denotes cause-and-effect relation and "dose" substitutes for "value". Notice that the dose-response function is defined even when E has a null effect on D. In that case the function is flat: $f(E)$=constant. The value of D is not influenced by the value of E.

"Evidence for dose-response" is therefore a meaningless claim; there is always *some* dose-response function relating D to E. What users of the phrase try to say is that they found evidence for a special type of a relation: a monotonic dose-response function. They want to say that the larger the value of E, the larger the value of D (monotonically increasing), or that the larger the value of E, the smaller the value of D (monotonically decreasing). I will skip the formal definitions of these terms, as well as the distinction between a monotonic function and a strictly monotonic function.

Whether monotonic or not, a dose-response function provides estimates of the effect of E on D for any pair of values of E, say, e_1 and e_2. If the function is known, then the effect of the causal contrast $E=e_2$ versus $E=e_1$ is the contrast between $f(e_2)$ and $f(e_1)$, usually quantified as $f(e_2)-f(e_1)$, $\exp(f(e_2)-f(e_1))$, or $f(e_2)/f(e_1)$. Knowing the dose-response function is, therefore, knowing all the effects of E on D when E is continuous. No small matter.

How do we know what the dose-response function is?

Relax, I am not about to reveal any great news. We don't know. We never know, other than in the scientific sense of "knowing". Everything worth knowing in science is no more than a conjecture (even when the whole of science votes "we know" and label any criticism as denial, bias, ignorance, and the like). So rather than asking "how do we know?", let's ask "how do we explore?", or "how do we estimate?"

Method 1: Display the data

This might work, rarely. If D is a continuous variable, and you don't need to condition on any variable to estimate the effect, you may display the data points {E, D} in a graph and look for a pattern. That is, you may try to fit a smoother through the cloud of data, thereby approximating the dose-response function. Beyond that rare example, the method fails. What will you learn from displaying the values of a binary D against the values of E? Not much. What values of D will you display, when D is the log-odds of the disease, and you have to condition on variables X, Y, and Z

to remove confounding bias? There are no values to display, unless you specify the dose-response function… Circularity.

Method 2: Linear, until proven otherwise

If I had to come up with a single example of a poor statistical paradigm, "linear, until proven otherwise" might be my choice. Here is the method:

Fit and solve the following equation (or some variation of it):

$$D = \beta_0 + \beta_1 E + \boldsymbol{\beta V}$$

where **βV** is notation for the covariates on which you need to condition and their coefficients. Then, test a null hypothesis, $\beta_1=0$. If rejected by a magical p-value, conclude that you have evidence that $\beta_1\neq 0$. But you do NOT have evidence for a linear function. You have evidence for a non-zero coefficient — *if you assume a linear function*. Let's repeat: *if you assume a linear function*, you have evidence that the function is not flat. But who says that the dose-response function is linear to begin with? By analogy (from logic): suppose you *assume* that the sun is colder than we think, and then show that *if that were the case*, life on earth is still possible. Is this evidence that the sun is colder than we think?

What happens if you did not succeed in rejecting the null? Do you have evidence for $f(E)$=constant, a flat dose-response function (null effect of E on D)? No. You have evidence for nothing. The lack of evidence against the null is not evidence for the null. [This might be a good time to visit, or revisit, my home page and read the dialogues on null hypothesis testing.]

The story does not end, however, even when the null is rejected. In that case you are told to fit another model, a quadratic function,

$$D = \beta_0 + \beta_1 E + \beta_2 E^2 + \boldsymbol{\beta V}$$

and test another null: $\beta_2=0$. Again, you are told to draw the following (erroneous) inference:

If the null is rejected, you have evidence for a quadratic dose-response function. (No, you don't. *If you assume a quadratic dose-response function*, you have evidence that $\beta_2\neq 0$. But who says that a quadratic function is indeed the true dose-response function to begin with? Can't the function be some higher order polynomial?)

If the null is not rejected, you have evidence for a linear dose-response function. (No, you don't. The lack of evidence against $\beta_2=0$ is not evidence for $\beta_2=0$. It is evidence for nothing. The function may still be anything: linear, quadratic, cubic, higher order polynomial. You have gathered no new knowledge whatsoever.) [Again, look up "A non-significant dialogue" on my home page.]

Statisticians are usually competent in logical inference, so what has caused them to take here an erroneous route? The explanation may be found in a paradigm called "simplicity". Linear function is the simplest dose-response function, and a quadratic function is next in line. That's true, but scientific discovery is not about providing a simple description of causality. It is about searching for a *true* description of causal reality. In this case, simplicity is both simplistic and

wrong. You may choose a linear function only when a thorough inquiry, as described later, leads to that choice. Linearity may be the *conclusion* at end of the road; it should never be the *assumption* at the beginning.

Method 3: A step function

Imagine a dose-response function that is composed of a series of horizontal lines connected by vertical lines. That's a step function. The steps may go only up, only down, or both up and down in any order. We have seen this function before, many times. Let's review the method by which it is fit.

First, categorize the exposure into k successive categories, say, k=4. In each of the four categories, fit a horizontal line, $f(E)$=constant, where the constant is D (in our notation): mean, log-odds, probability, etc. For example, the constant may be the mean of a continuous D in each of the four categories of E, or the log-odds (D=1) for a binary D. In regression, these constants are easily found by creating k-1 dummy variables and regressing D on those variables (and anything else that should be conditioned on).

Table 1 shows the coding of dummy variables, where E is LDL cholesterol and the four categories are approximately quartiles of the LDL distribution. The lowest quartile serves as the reference category.

Table 1.

LDL (mg/dL) Approximate quartiles	LDL2	LDL3	LDL4
70-110	0	0	0
110-135	1	0	0
135-160	0	1	0
160-240	0	0	1

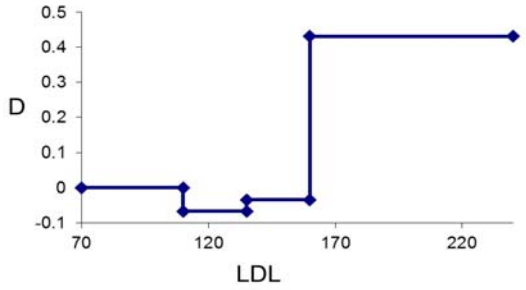
Regressing the log hazard of stroke[a] (D in our notation) on the dummy variables, we find the following solution:

$$D = -0.07\,LDL2 - 0.03\,LDL3 + 0.43\,LDL4$$

And the graph is displayed below.

[a] Not exactly. In Cox regression, we don't actually predict the log hazard since we ignore a time-dependent intercept. But that omission only moves the function up or down. The *shape* of the function, $f(LDL2, LDL3, LDL4)$, does not depend on the unspecified intercept.

Figure 1.



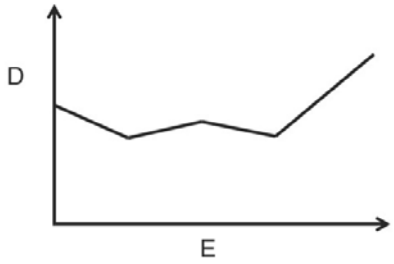
As you can see in the graph, the step function explicitly assumes that E has a null effect on D within each category (say, no effect for 230 vs.165) and a sizeable effect for similar values on opposite sides of each cutoff point (say, 162 vs. 158. It is therefore a joke. No one would assume that a step function is a correct description of the true dose-response function. What is often done next is "a test for linear trend", a quadruple joke.[b] First, the test depends on arbitrary coding of the distance between categories. Second, it is not a test of monotonicity across adjacent categories. Third, it is not a test of trend across the range of a continuous E. Fourth, the procedure rests on the shaky foundation of null testing and p-values.

I prefer another approach. Simply try to draw a smoother through the steps. This is your rough guess of the dose-response function, and it is no more than a semi-quantitative guess, a good start. What's next?

Method 4: Linear spline function

Look again at Figure 1. Would it not be better if you could fit a sequence of contiguous lines that are not constrained to be horizontal? I mean something like the following graph:

Figure 2.



[b] You can find this nonsense in my first-author papers too, early in my career. Advice for the novice: don't assume that what epidemiologists and statisticians around you are doing is necessarily correct. Many of them are just duplicating uncritically what they had been taught.

This type of a dose-response function allows you to estimate the effect not only between two categories of E, but also within each category. In addition, it is easier to draw a smoother through such a graph than through a step function.

Figure 2 is a linear spline. Using the same example of LDL and stroke, I will explain the technique of linear spline regression,[c] and then, why it works.

First, create three "spline variables" for the four categories of LDL (k-1), analogous to three dummy variables (Table 2). Notice that unlike dummy variables, spline variables are continuous, not binary.

Table 2.

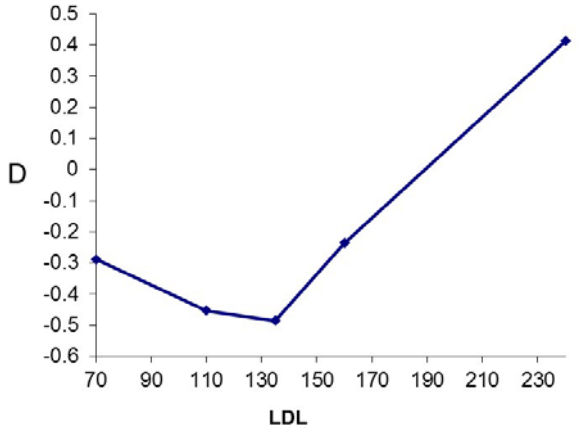
LDL (mg/dL)	S2	S3	S4
70–110	0	0	0
110–135	LDL-110	0	0
135–160	LDL-110	LDL-135	0
160–240	LDL-110	LDL-135	LDL-160

Next, regress the log hazard of stroke (D in our notation) on the spline variables *and* LDL. We find the solution

$$D = -0.0041\ LDL + 0.0028\ S2 + 0.0113\ S3 - 0.0019\ S4$$

the graph of which is displayed below.

Figure 3.



[c] Many students confuse "linear spline regression" with "linear regression". The former is *not* the well-known least-squares regression. Rather, linear splines can be fit within any regression model we have discussed (logistic, Poisson, Cox, *and* linear), depending on the nature of D.

So how does it work? Why do the spline coding and spline regression create a function that is composed of contiguous straight lines?

It is not that difficult to understand, once we see the actual function in each category of LDL.

Quartile 1: D = –0.0041 LDL (notice that all spline variables take the value of zero)
Quartile 2: D = –0.0041 LDL + 0.0028 S2 (notice that S3=0 and S4=0)
Quartile 3: D = –0.0041 LDL + 0.0028 S2 + 0.0113 S3 (S4=0)
Quartile 4: D = –0.0041 LDL + 0.0028 S2 + 0.0113 S3 – 0.0019 S4

Recall that the spline variables were created from LDL (Table 2):

S2 = LDL–110
S3 = LDL–135
S4 = LDL–160

Replacing the spline variables with the expressions above, we see the following functions:

Quartile 1: D = –0.0041 LDL
Quartile 2: D = –0.0041 LDL + 0.0028 (LDL–110)
Quartile 3: D = –0.0041 LDL + 0.0028 (LDL–110) + 0.0113 (LDL–135)
Quartile 4: D = –0.0041 LDL + 0.0028 (LDL–110) + 0.0113 (LDL–135) – 0.0019 (LDL–160)

First, notice that the function D=*f*(LDL) is *linear in each quartile*. For example, simplifying the function in Quartile 2, we get:

D = –0.0041 LDL + 0.0028 (LDL–110) = (–0.0041 + 0.0028) LDL – 0.0028x110
= – 0.308 – 0.0013 LDL

Try to simplify the function in Quartiles 3 and 4 and get a concise linear expression ($\beta_0+\beta_1$LDL).

Second, notice that at the cutoff points between adjacent quartiles, *the function to the left and the function to the right predict the same value* of D. Take, for example, the cutoff between the first quartile and the second quartile (LDL=110).

The function to the left (Quartile 1) predicts: D = –0.0041x110 = –0.451
The function to the right (Quartile 2) predicts: D = –0.0041x110 + 0.0028 (110–110) = –0.451

Which means that *the function is contiguous at* LDL=110.

That is also the case at the cutoff between the second quartile and the third (LDL=135), and at the cutoff between the third quartile and the fourth (LDL=160). (Check it.) That's why the cutoff points are formally called "knots". The line to the left of a knot and the line to the right of a knot meet (are "tied") at the knot. And that's why the spline coding and regression always generate something like Figure 2, a sequence of contiguous straight lines.

Method 5: Quadratic spline function

Linear spline is better than a step function, but we can do even better. How about allowing for some curvature (a quadratic function) in each quartile of LDL, and forcing not only contiguity at the knots but also smooth continuity? So smooth that no one can tell from the graph where the knots are located?

As you might have guessed, we would still need spline variables, but this time -- their squared version. Also needed is LDL-squared. Here is the quadratic spline model and its solution:

$$D = 0.0791\ LDL - 0.0004\ LDL^2 + 0.0009\ S2^2 - 0.0004\ S3^2 - 0.00003\ S4^2$$

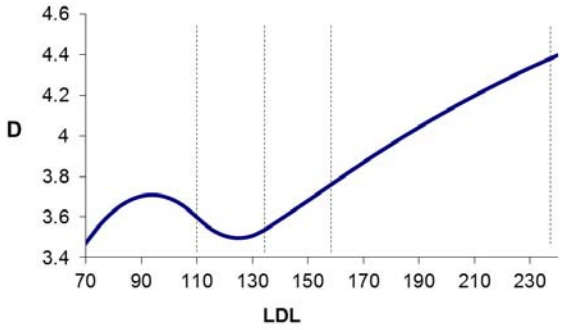
Replacing the spline variables, we get

$$D = 0.0791\ LDL - 0.0004\ LDL^2 + 0.0009\ (LDL-110)^2 - 0.0004\ (LDL-135)^2 - 0.00003(LDL-160)^2$$

If you simplify the function in each quartile, you will see that you get four (different) quadratic functions. As before, it is easy to check that the functions are contiguous at the knots. Lastly, the continuity is smooth because, at each knot, the first derivative of the function to the left is identical to the first derivative of the function to the right.

Figure 4 displays the quadratic spline function. The vertical lines indicate the knots (110, 135, 160). Had they not been shown, you would not have been able to tell where they are located.

Figure 4.



What can we learn from Figure 4?

Many of us will dislike the "inconsistency" in the graph. We tend to like monotonicity and dismiss almost any other dose-response pattern as "implausible", "noisy data", and the like. Well, these are the results and we ought to interpret them one way or another. Here is my take:[d]

I am not particularly troubled by the up-down-up pattern because the "hill" is low and the "valley" is shallow. If this is a good approximation of the dose-response function, I see tiny to small effects in the range of 70 to 150 or so. In this range, the most extreme fluctuation we

[d] Pretending that we don't need to condition on any variable to estimate the effect…

observe (on the Y-axis) is about 0.2. Since D is log-hazard, a difference of 0.2 between two log hazards translates to a hazard ratio of about 1.2 (or 0.82), nothing to get excited about. By contrast, I see nearly linear shape in the upper quartile and some meaningful effect sizes in that region versus others. For instance, the contrast between 230 and 130 translates to a hazard ratio of about 2.2. So, perhaps the effect is somewhere between null and very small up to 150 or so and fairly linear thereafter, reaching as high as two-fold hazard for some distant contrasts.

What I wrote above sounds to me like science (i.e., interesting conjectural knowledge), and a lot more scientific than "a statistically significant linear trend across four quartiles ($p<0.001$)".

A few summary points

- That we allow a quadratic function does not imply that we enforce much curvature. Look, for example, at the shape of the graph in the upper quartile (Figure 4). It is close to a straight line. Although the function is quadratic, it does not have much curvature in that uphill region. The downhill region is farther to the right, well into non-existing values of LDL.

- A quadratic function has a minimum or a maximum, which means that the graph can change direction only once between two adjacent knots. Some people choose cubic spline -- allowing a cubic function between successive knots -- which means that the graph can change direction *twice* in any category (say, up-down-up). I think it's overdoing the method. If the categories are not too wide and the data are not sparse, I do not expect to miss a change of direction by using quadratic spline.

- The choice of the number (and location) of the knots is arbitrary. I think that three knots are a good choice whenever the data are not sparse. Four may be used with abundant data, and two -- when you are worried about sparse data. Oftentimes, the graph is not very sensitive to some variation in the location of the knots.

- It's a good idea to give some indication of how much data you have in each category. For example, I could have displayed the number of strokes in each quartile. The amount of data might also guide the choice of the knots (i.e., their locations).

- Do not assume that quadratic spline regression is some kind of panacea. Remember that we supply assumptions and constraints (number and location of the knots, quadratic function, continuity at the knots, maximum likelihood estimation), which in turn allows the data to deliver results (coefficients). The scientific road is always a two-way road between us and data. Data never speak for themselves; they respond to whatever we impose in our models.

- Spline regression is actually a family of methods. There are variations which I did not discuss here, such as restricted quadratic spline, penalized spline, and "a moving window". My viewpoint on these methods is as follows: Whenever a different smoothing method gives you vastly different results from the method I explained here, it is time to start worrying about the data, not about choosing the "right" method.