Chapter 9

Estimating the Arithmetic Mean Difference

Sex group and blood pressure

The arithmetic mean difference was not my preferred measure of effect (chapter 3), but for several related reasons I decided to give it a place of honor in the second part of the book: First, it is easier to explain key principles of estimation on the additive scale (difference) than on the multiplicative scale (ratio). Second, we often estimate the mean difference by linear regression—the historical foundation of all regression models. Third, many of the principles of linear regression hold in other statistical models (by which we estimate other measures of effect.) I have not abandoned, however, my commitment to ratio measures of effect and suggest that you do the same. Try to learn from this chapter and the next one about regression and estimation in general, but compute the geometric mean ratio in causal inquiry (chapter 11).

In chapter 3, you may recall, we estimated the arithmetic mean difference in FEV_1 between smokers and former smokers. Here, we will first consider an example of sex group (causal variable) and systolic blood pressure (effect). Assuming that one's sex and one's blood pressure have no common cause, we may set aside fears of confounding paths (chapter 6). And if prior knowledge of blood pressure did not affect the chances of getting into the sample, we may also assume the absence of selection bias (chapter 7). On these assumptions, some measures of the marginal association between sex and systolic blood pressure (SBP) are measures of the effect of sex (chapter 3). Table 9–1 shows relevant variables and several observations in the data file that we'll use.

Table 9–1. The first six observations in a data file (N=1,000 people)

Observation	SEX (0=female; 1=male)	SBP* (mmHg)	AGE (years)
1	0	161.0	73
2	0	145.5	77
3	0	147.0	60
4	1	111.5	81
5	0	102.5	59
6	1	98.0	72

* Systolic blood pressure

The means and the mean difference

In this sample of 1,000 people, 45 to 84 years old, mean SBP was 126.8 and 124.5 mmHg in men and women, respectively. Figure 9–1 schematically illustrates several values of SBP in each sex and shows the two means (open circles). I also drew two lines: a dashed line that passes through the means and vertical line that corresponds to the mean difference.

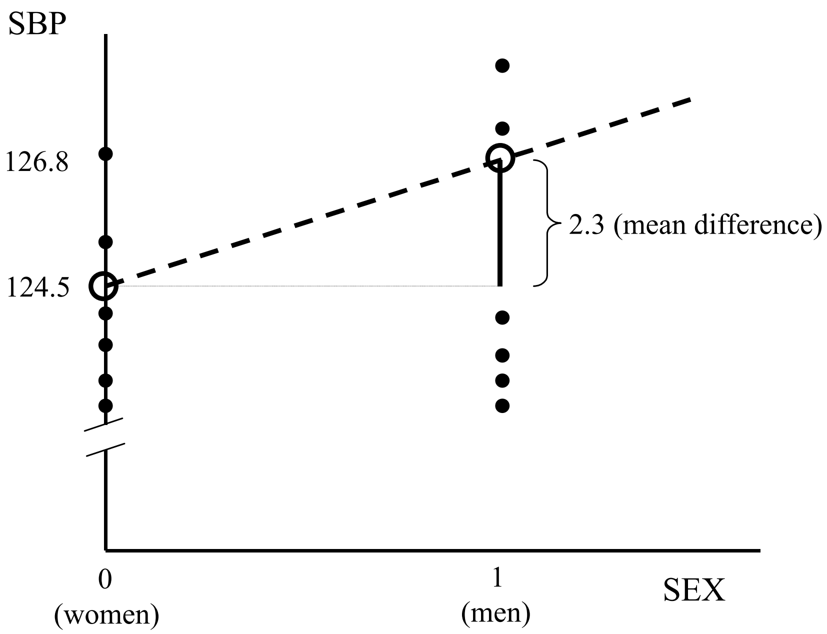


Figure 9–1. Illustration of data points, mean values, and the mean difference

As you know, every straight line can be described by a unique equation, $Y = \beta_0 + \beta_1 X$, where β_0 is the intercept (the value of Y when X=0) and β_1 is the slope. We define the slope as $\Delta Y/\Delta X$ (the "rise" divided by the "run"), or alternatively, as the change in Y when X increases by 1 unit. Every point (x, y) on the line fulfills the equality $y = \beta_0 + \beta_1 x$.

Notice that the dashed line in Figure 9–1 contains only two meaningful points: (0, 124.5) and (1, 126.8). All other points do not correspond to any reality (because the variable SEX takes only two values: 0 and 1). We may still write, however, the line's equation, $SBP = \beta_0 + \beta_1$ SEX, and find out the values and meaning of β_0 and β_1.

By definition, the intercept (β_0) should be the value of SBP when SEX=0. But we already know that number and its meaning from Figure 9–1: it is 124.5 mmHg—the mean value of SBP in women. To find out the meaning of the slope (β_1), identify the "rise" of SBP when the value of SEX increases by 1 unit, from 0 to 1. That rise, shown as the vertical line, is also known. It is the mean difference (2.3 mmHg): mean SBP in men minus mean SBP in women. To sum up, the dashed line in Figure 9–1 corresponds to the equation "SBP = 124.5 + 2.3 SEX".

Linear regression and the mean difference

In Figure 9–1, I conveniently drew a line to connect the two means, and thereby set its intercept to be the mean of SBP in women and its slope to be the arithmetic mean difference. But that is not the only possible line that may connect a column of data points in women with a column of data points in men. I could have connected the median values in each sex, or the 25th percentiles, or any other pair of values. In fact, I could have chosen from an infinite number of lines. What is so special, then, about connecting the means? Why choose a line whose intercept is the arithmetic mean of SBP in women and whose slope is the mean difference?

Linear regression offers an answer to this question by starting from a neutral viewpoint, with no a priori preference for any pair of values. It tells you to look for a line that is "as close as possible" to the various data points—a line whose intercept and slope are determined by the observations. In the method of linear regression (or more precisely, *ordinary least-squares regression*), you should select the "winning" line for the blood pressure data according to the following theoretical algorithm:

1. Draw a candidate straight line
2. Calculate the *vertical* distance (call it "e") between each blood pressure value and the candidate line. (Figure 9–2 shows several examples.)
3. Square the vertical distance: e^2
4. Sum the squares: $\Sigma\, e^2$
5. Among *all* possible lines, choose the one for which you get the *smallest* sum of squares. (Rest assured that no two lines would meet that condition.)

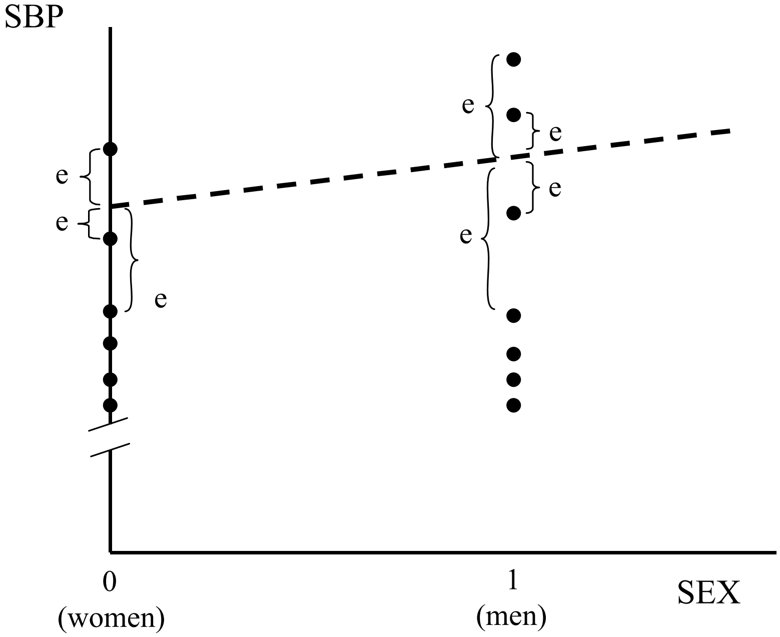


Figure 9–2. A few examples of the vertical distance (e) from a candidate line

The line that generates the smallest sum of squares is called the *linear regression line*. And it happens to be the one that I drew in Figure 9–1: a line whose intercept is the mean of SBP in women and whose slope is the mean difference. This is no one-time coincidence. For every similar example, the mean difference will win the contest, proving to be the slope of the line for which the sum of squares is the smallest. The mathematical explanation is simple: The sum of squares about the arithmetic mean is smaller than the sum of squares about any alternative number.

Linear regression is therefore a method by which we can estimate the mean difference. Of course, you don't really need it to compute the marginal (crude) association between a binary exposure and a continuous effect, but it will prove essential in more complex tasks of estimation—when we'll need to condition on confounders, for example (chapter 6).

More on linear regression

Linear regression and other regression models are exploited for causal inquiry, but their underlying principle is *prediction*, not causation. These equations predict the value of one variable from the value of another (or from the values of other variables), regardless of which are the causes and which is the effect—if at all. If we are told, for instance, that John belongs to our sample, we may predict his systolic blood pressure ("guess" the number) by entering SEX=1 into the equation "SBP = 124.5 + 2.3 SEX". Of course, John and all of his fellow men share a single predicted blood pressure (126.8 mmHg), and so do all of the women (124.5 mmHg). But if there were other variables next to SEX, we would have found a more diversified prediction.

You can tell from Figure 9–1 that for almost every member of the sample the measured blood pressure differs from our model-based prediction (which is the sex-specific mean). That difference is the vertical distance (e) from the regression line. For this reason the mathematical relation between *measured* SBP and SEX takes the following form: "SBP = 124.5 + 2.3 SEX + e". And if you wish to write precise statistical notation, add the subscript "*i*" to indicate the *i*-th person: $SBP_i = 124.5 + 2.3\ SEX_i + e_i$. John's measured value, for instance, is the sum of his predicted value (124.5 + 2.3 x 1 = 126.8) and the vertical distance (e) between that prediction and his measured value. (If measured SBP is smaller than predicted SBP, we add the negative of the distance.)

Textbooks in statistics specify several assumptions about "e", which I will not discuss here. This term goes by several names, neither of which is particularly good: "error term", "disturbance", "random disturbance ", and "random noise". Although I have no better name to offer, I suggest that "e" is the expression of indeterministic causation (chapter 1)—whenever the model claims to estimate an effect. Male sex generates a propensity toward some value of blood pressure, but the actual blood pressure of John has emerged as probabilistic realization of that propensity.

In the absence of randomization, deterministic statisticians usually attribute "e" to sampling-related randomness, though they never explain what was "sampled" from what (chapter 4). Sometimes, "e" is attributed to "random measurement error", assuming that John and his fellow men share the same true blood pressure (?). And in some minds, this term is supposed to represent the combined effect of unspecified determinants of blood pressure, which somehow add up to a random component. Statistics textbooks list mathematical requirements from "e", without which the model does not rest on a solid foundation. They also explain how to check whether some of the requirements are met.

SAS PROC GLM

Let's re-examine our example of sex and systolic blood pressure from the very beginning. Our goal is to estimate, by linear regression, the mean difference in SBP for the causal contrast between male sex and female sex. We already know that the slope (β_1) of the regression line "$SBP = \beta_0 + \beta_1\ SEX$" is equal to the mean difference, but we don't know the value of the slope, yet. At the moment, both β_1 and β_0 are unknown *coefficients*. The task at hand is, therefore, to solve the equation "$SBP = \beta_0 + \beta_1$ SEX + e": to find the values of β_0 and β_1 for which the sum of squares (Σe^2) is minimal.

Statistical theory has found formulae for β_0 and β_1, which you can find in many textbooks. SAS, the statistical software I will use throughout, offers several procedures to fit a linear regression line—that is, to find its coefficients. Of these, **PROC GLM** is widely used. (PROC is short for procedure; GLM is the acronym for "general linear models".)

SAS code

```
PROC GLM;
 MODEL sbp = sex/SOLUTION CLPARM;
 run;
```

Below the **PROC GLM** statement, you find a "model statement" that specifies the regression variables. Analogous to the equation "$SBP = \beta_0 + \beta_1$ SEX", systolic blood pressure is written to the left of the equality sign and the sex variable is written to the right. In the language of regression, SBP is called the dependent variable or the response variable, whereas SEX carries at least four names: independent variable, explanatory variable, regressor, and predictor. Note that in correct terminology we "regress the dependent variable on the independent variable" and not the other way around. Again, the math of the model does not rest on any causal assumption.

To the right of the slash, I added two key words: **SOLUTION** requests the solution of the regression equation ($SBP = \beta_0 + \beta_1$ SEX + e), namely, the values of the coefficients (of the line with the smallest the sum of squares…) **CLPARM** requests confidence limits for the regression coefficients (parameters.)

Selected SAS printout

SAS printout contains many numbers, some are more useful than others. Here and elsewhere, I selected those pieces that serve my emphasis and pedagogical preference. Test statistics and *p*-values do not show up for reasons that were explained in chapter 8.

```
                              The GLM Procedure

Dependent Variable: sbp       SYSTOLIC BLOOD PRESSURE (mmHg)

                                                 Sum of
        Source                      DF          Squares

        Model                        1        1304.4140
        Error                      998      454303.9420
        Corrected Total            999      455608.3560

                                     Standard
 Parameter          Estimate            Error       95% Confidence Limits

 Intercept       124.5202991
 sex               2.2889114       1.35216360      -0.3644985    4.9423213
```

As you can see, the method of linear regression indeed identified the same line that I drew in Figure 9–1: SBP = 124.5 + 2.3 SEX. The sum of squares (Σ e^2) about this line is shown at the top of the printout in the row titled "Error". No other line would have generated a sum of squares smaller than 454,303.942. If you have taken a course in linear regression, you probably know the meaning of DF and other kinds of sum of squares. (If you don't know, it doesn't matter.)

For our purpose the key number on the printout is, of course, the coefficient of SEX— the slope of the regression line. This is the mean difference in SBP between men and women. Again, in this sample the men's average is 2.3 mmHg higher than the women's average. Assuming that the estimator is unbiased, we may use the standard error (1.4) to compute a confidence interval or a confidence limit difference (chapter 8). The 95% confidence limits are already provided.

Age and blood pressure

Suppose that instead of estimating the effect of sex on systolic blood pressure, we wish to estimate the effect of age. Unlike sex, however, age is a continuous variable so we have to compute the mean difference for many causal contrasts (chapter 2): ages 50 and 52; 50 and 63; 58.2 and 62.9; and many other pairs. Obviously, it is impossible to replicate the method we have just used—for every conceivable contrast. In some ages the number of observations is small; blood pressure was not measured, and cannot be measured, in every possible age; and the number of pairs is infinite.

Linear regression offers several solutions, the simplest of which ask you to make a non-trivial assumption: For any specified age difference (call it Δ), assume that the mean difference is constant, regardless of the ages you contrast. For example, if you specify Δ=1 year, causal contrasts such as [50, 51], [51.5, 52.5], and [62.9, 63.9] should produce the same mean difference. And for Δ=2.5 years, contrasts such as [45, 47.5], [57.5, 60], and [60.1, 62.6] should also share a single mean difference. In the language of causation, you are asked to assume that the effect of "Δ years of aging" on mean systolic blood pressure is constant (on the additive scale) whatever the starting age may be.

In Figure 9–3, I plotted several hypothetical data points that satisfy that restrictive assumption. I chose a mean difference of +2 mm Hg per Δ=1 year, and five successive contrasts: [50, 51]; [51, 52]; [52, 53]; [53, 54]; [54, 55].

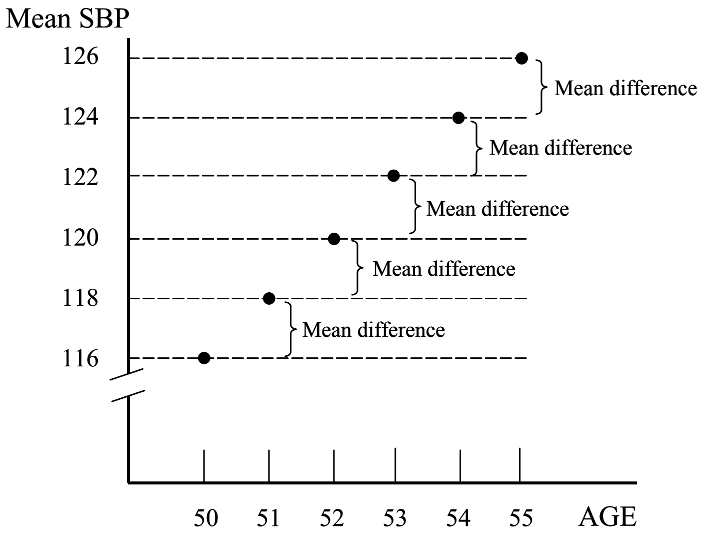


Figure 9–3. An example of a constant mean difference (+2 mmHg) per 1 year of aging

As the graph shows, on the assumption of a constant mean difference, mean SBP must be a *linear function* of AGE: Mean SBP = $\beta_0 + \beta_1$ AGE

The intercept of this line is not particularly helpful (mean SBP when AGE=0), but the slope, if estimated, should serve us well in causal inquiry. By definition, β_1 is the mean difference in SBP per 1 year of aging. To compute the (constant) mean difference for any Δ years of aging (such as 2.5), just multiply β_1 x Δ. Table 9–2 shows a general proof of this rule, using the notation "k" and "k+Δ" for the contrasted pair.

Table 9–2. Computing the effect of Δ years of aging, on the restrictive assumption of a constant mean difference

Causal assignments	Mean SBP = $\beta_0 + \beta_1$ AGE
AGE = k+Δ	Mean SBP = $\beta_0 + \beta_1$ (k+Δ)
AGE = k	Mean SBP = $\beta_0 + \beta_1$ k
Effect of Δ years of aging (mean difference in SBP)	$\boldsymbol{\beta_1 \Delta}$

Linear regression of SBP on AGE

If you are willing to make the assumption of a constant mean difference per Δ years of aging, the task at hand is reduced to finding a straight line that will pass through many mean values of SBP. Again, linear regression offers a method for choosing that line: Solve the equation "SBP = $\beta_0 + \beta_1$ AGE + e"; find the coefficients for which the sum of squared "error" is the smallest.

Figure 9–3 illustrates several data points and some of the vertical distances that should be squared and summed up. On several assumptions about "e", statistical theory reassures us, again, that the "winning" straight line will pass through the (estimated) means of SBP, connecting an infinite number of them. The product "β_1 x Δ" will tell us the mean difference for any specified Δ of age.

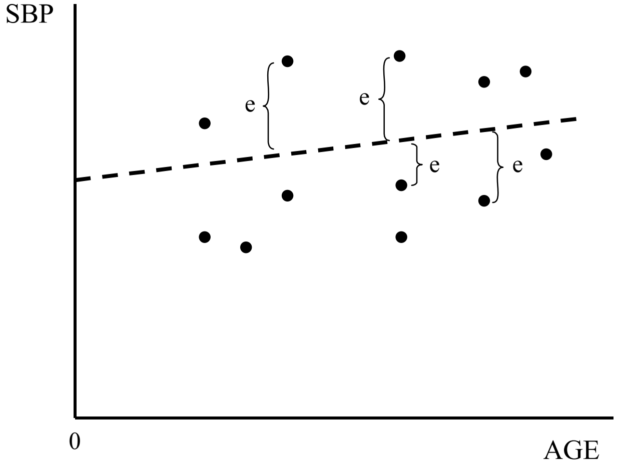


Figure 9–3. Illustration of linear regression of SBP on AGE

SAS code

```
PROC GLM;
 MODEL sbp = age/SOLUTION CLPARM;
 run;
```

As you see, the code is identical to the one I have used to regress blood pressure on sex. I have just substituted the continuous variable AGE for the binary variable SEX.

Selected SAS printout

```
                               The GLM Procedure

Dependent Variable: sbp       SYSTOLIC BLOOD PRESSURE (mmHg)

                                      Sum of
       Source                  DF          Squares

       Model                    1       58081.6546

       Error                  998      397526.7014

       Corrected Total        999      455608.3560
```

```
                                   Standard
Parameter          Estimate        Error           95% Confidence Limits

Intercept       80.53075063
age              0.73842716        0.06115135     0.61842719    0.85842714
```

The regression line of SBP on AGE has an intercept of 80.5 and slope of 0.7 (Figure 9–4). No other straight line would have generated a sum of squared "error" smaller than 397,526.7014.

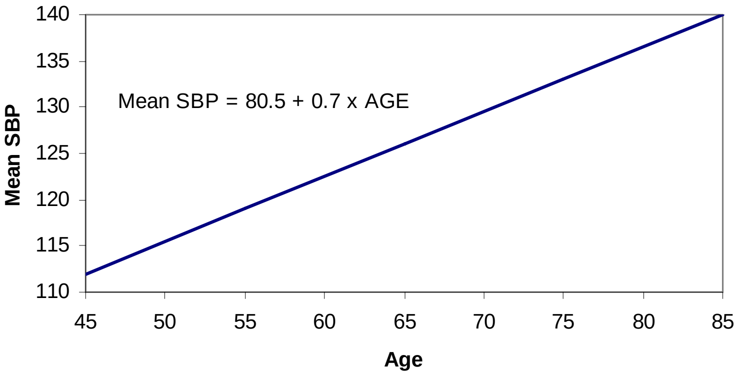


Figure 9–4. The linear regression line: SBP regressed on AGE

With this line at hand we can estimate the mean difference for any Δ years of interest. For example, the estimated difference in blood pressure for the contrast between the ages of 79.4 and 67.3 (Δ=12.1) is β_1 x Δ = 0.7 x 12.1= 8.5 mmHg.

The printout shows the 95% CI for the slope—for the mean difference in SBP when Δ=1. The 95% CI for any other Δ may be computed according to the following formula:

$$(\beta_1 \text{ x } \Delta) \pm 1.96 \text{ x SE } (\beta_1 \text{ x } \Delta)$$

A rule of arithmetic for standard errors tells us that SE (β_1 x Δ) = SE (β_1) x Δ. Therefore, the 95% CI may also be written as

$$(\beta_1 \text{ x } \Delta) \pm 1.96 \text{ x SE } (\beta_1) \text{ x } \Delta$$

For example, when Δ=12.1 years, the 95% CI for the mean difference in SBP is

0.7 x 12.1 ± 1.96 x 0.06 x 12.1= [7.1, 9.9]

Beyond a straight line

Whenever we fit a regression model to estimate the effect of a continuous exposure (E) on a dependent variable (Y), we do not come with empty hands, just asking the model to hand us the estimates. We must first specify a function that connects Y to E (Y=f(E), in notation), and inevitably force the dependent variable to change in a pre-specified manner. Such a function is called the *dose-response function*, because it is supposed to tell us how the value of the response variable changes as the "dose" (value) of the exposure changes. For example, by writing the equation "Mean SBP = $\beta_0 + \beta_1$ AGE", we impose a linear function on the relation of mean systolic blood pressure with age. That constraint, as you recall, has resulted from our assumption about a constant mean difference per Δ years of aging. *We* supplied a causal assumption.

But how do we know a priori that the mean difference should be constant? Or alternatively, how do we know that a straight line captures the true dose-response function for the effect of age on systolic blood pressure?

The simple answer is that we usually don't know. In rare instances, we may be able to display the data and convince ourselves visually that the observations scatter around a straight line. But in most cases we don't have that luxury, either because it is difficult to identify a pattern in a cloud of data points, or because there are no data points to display on a 2-dimensional graph (when the model contains several regressors, for example.) Nonetheless, you will find many advocates for the linear function: Some invoke the principle of simplicity, arguing that scientists should always prefer a simple theory to a complex one. And is there anything simpler than "Mean SBP = $\beta_0 + \beta_1$ AGE"? Others tell us to learn from experience: in retrospect, many causal relations seem to comply with a linear function. Others, yet, say that a straight line is much more plausible than many alternative lines, especially for a causal relation.

If you subscribe to Popper's philosophy of science (chapter 4) and cherish the pursuit of Truth, none of these arguments hold any merit. Your scientific duty is to discover the true dose-response relation, from which you may estimate the effect of various causal contrasts. To that end, you should mercilessly interrogate the data: specify alternative functions, examine the resulting graphs, and decide which function to accept as a good approximation for the unknown Truth. A priori preference for a straight line has no place in this inquiry and could lead you astray. Moreover, the first question to be answered is "what do the data tell us about the dose-response function?"—regardless of whether their story seems "plausible". Of course, we may later modify the story (smooth a bumpy graph, for example), or remain undecided between two graphs, or even decide against coherent inference from the data.

As you might have guessed, alternatives to a linear function are not assumption-free, either. They just replace one set of assumptions with another set, removing some of the constraints at the cost of imposing others. To illustrate two commonly used alternative functions, let's return to the example of age and systolic blood pressure.

A “step” function

If there were enough observations for some discrete ages, we could have computed the mean value of SBP for those ages, display the sequence of means, and perhaps visually learn something about the dose-response function. This method might occasionally prove helpful, but at least two drawbacks lurk in the background: we ignore some of the data (ages for which we don't have multiple observations), and we run the risk of "noisy estimates" (means that are based on too few observations).

This idea, however, has set the foundation for another method, which has been adopted by many researchers. Rather than computing the mean of SBP at discrete ages, compute it for several contiguous age groups. Figure 9–5 shows an example: I computed the mean SBP in four successive 10-year age groups (45-54, 55-64, 65-74, 75-84) and displayed each number as a horizontal line. Then, I added connecting vertical lines to create a continuous graph.

By adding one assumption, we can turn the graph in Figure 9–5 into a dose-response function. Let's assume that the mean in a 10-year interval also estimates the mean in *each nested age*—in each causal assignment within the interval. The value of 116.5 mmHg, for instance, estimates the mean SBP in every age within the first interval [45, 54] whereas the value 125.5 mmHg estimates the mean in every age within the second interval [55, 64]. On this assumption, the "step" function is Figure 9–5 is a primitive dose-response function from which we can compute several mean differences, though not all.

We cannot estimate the effect of age on systolic blood pressure for pairs of causal assignments that belong to the same interval, because they share the same mean (zero mean difference.) But pairs of causal assignments that reside in different age intervals will generate meaningful estimates: we just have to subtract the mean of SBP in one interval from the mean in another. In Figure 9–5, I show three estimates (9.0, 14.2, 21.9), using the youngest group as the reference causal assignment. Three other mean differences may be derived from the first three by subtraction—for a total of six. For example, the estimated mean difference between a causal assignment in the third age group and its counterpart in the second is 14.2–9.0 = 5.2 mm Hg (or 130.7–125.5).

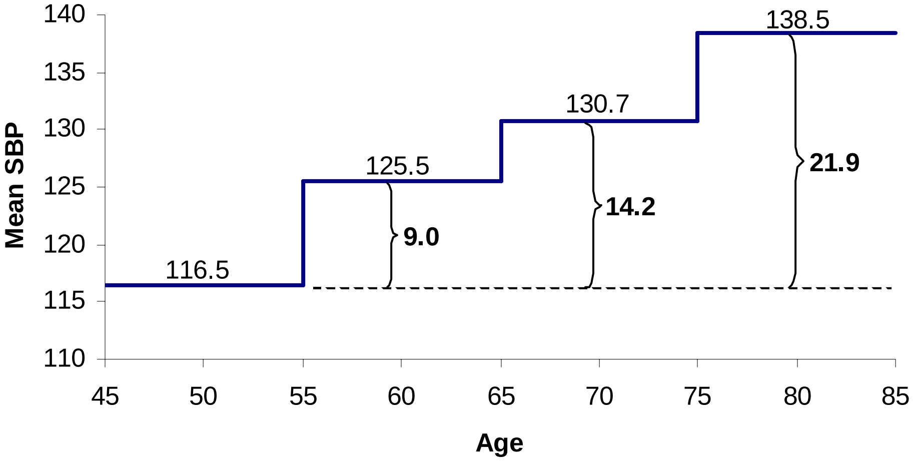


Figure 9–5. Mean SBP by four age groups

How do the assumptions of the step function differ from those of a straight line?

As you can tell from the graph, we abandoned the assumption of a constant mean difference per Δ years of aging. On the other hand, we imposed two new assumptions:

zero mean difference within each 10-year interval and an abrupt change of the mean at the junction points.

Figure 9–5 looks as implausible to me as it looks to you. Is there any reason for Nature to create causal relations that follow a step-like dose-response function? Is there any reason why the mean difference should be zero for the causal assignments [52, 54] and 9.0 mmHg for the nearby pair [54, 56]? The answer to both questions is "No". Again, we have imposed the step function on the data, just as we previously imposed a linear function (which may be similarly criticized for being simplistic and naïve.) But we did gain something. We now have the possibility of imagining a smooth line that would pass through the steps, and possibly guess the shape of the underlying dose-response function. Fitting that imaginary line does not follow any algorithm and may occasionally be as subjective as the answers to a Rorscharch inkblot test. Nonetheless, we got another view of the data and moved one step away from a world that contains nothing but straight lines.

Fitting the "step" function by linear regression

Just as we didn't need linear regression to estimate the difference in mean SBP between men and women, we don't really need it to compute the differences between four age groups. The method will become essential, however, when confounding or effect modification start playing a role. For the time being, let's find out how linear regression could create the dose-response function in Figure 9–5.

Our task is to write a regression equation whose coefficients will estimate the means and the mean differences (Figure 9–5.) On first impression, it is not at all clear that a single equation could describe a graph that is composed of four horizontal lines. Since a horizontal line corresponds to "Y=constant", it seems that we must write four equations:

If $45 \leq \text{age} < 55$, Mean SBP = A
If $55 \leq \text{age} < 65$, Mean SBP = B
If $65 \leq \text{age} < 75$, Mean SBP = C
If $75 \leq \text{age} < 85$, Mean SBP = D

One method of reducing these equations into one requires a preliminary step called "dummy coding". In this step we replace the four categories of age with three binary "dummy variables" as shown in Table 9–3.

Table 9–3. Replacing the four categories of age with three "dummy variables"

Age category	AGE2	AGE3	AGE4
45-54	0	0	0
55-64	1	0	0
65-74	0	1	0
75-84	0	0	1

The wisdom behind the conversion is not apparent yet, but notice first that the new variables preserve the original data, because the joint values of [AGE2, AGE3, AGE4] identify each age category: The oldest group, for example, is uniquely identified by [AGE2=0, AGE3=0, AGE4=1] whereas the youngest group is uniquely identified by [AGE2=0, AGE3=0, AGE4=0]. For reasons that will become clear shortly, the group that takes a zero value on all dummy variables is called the "reference category." I chose the youngest group, but any other choice is permissible.

Second, it turns out that equation 9–1, below, corresponds to a step function:

Mean SBP = $\beta_1 + \beta_2$ AGE2 + β_3 AGE3 + β_4 AGE4 (Equation 9–1)

To convince ourselves, let's derive the mean SBP for each age group by entering the appropriate values of the three dummy variables.

If $45 \le$ age < 55: **Mean SBP =** $\beta_1 + \beta_2 \text{ x } 0 + \beta_3 \text{ x } 0 + \beta_4 \text{ x } 0$ **=** $\boldsymbol{\beta_1}$
If $55 \le$ age < 65, **Mean SBP =** $\beta_1 + \beta_2 \text{ x } \mathbf{1} + \beta_3 \text{ x } 0 + \beta_4 \text{ x } 0$ **=** $\boldsymbol{\beta_1 + \beta_2}$
If $65 \le$ age < 75, **Mean SBP =** $\beta_1 + \beta_2 \text{ x } 0 + \beta_3 \text{ x } \mathbf{1} + \beta_4 \text{ x } 0$ **=** $\boldsymbol{\beta_1 + \beta_3}$
If $75 \le$ age < 85, **Mean SBP =** $\beta_1 + \beta_2 \text{ x } 0 + \beta_3 \text{ x } 0 + \beta_4 \text{ x } \mathbf{1}$ **=** $\boldsymbol{\beta_1 + \beta_4}$

Because the expressions on the right hand side are all constants, equation 9–1 indeed corresponds to four horizontal lines, one line per age group. The intercept (β_1) is the value of the dependent variable when all regressors take the value of zero, namely: mean SBP in the reference interval (ages 45 to 54). The three slopes ($\beta_2, \beta_3, \beta_4$) estimate the mean difference between any causal assignment in the respective age interval and any causal assignment in the reference. Table 9–4 shows one example of the explicit proof:

Table 9–4. Deriving the effect of a causal contrast between age 60 (second interval) and age 50 (reference interval) from equation 9–1

Causal assignment	Values of dummy variables	Mean SBP = $\beta_1 + \beta_2$ AGE2 + β_3 AGE3 + β_4 AGE4

AGE = 60	AGE2=1
	AGE3=0
	AGE4=0
		Mean SBP = $\beta_1 + \beta_2 \text{ x } 1 + \beta_3 \text{ x } 0 + \beta_4 \text{ x } 0$

AGE = 50	AGE2=0
	AGE3=0
	AGE4=0
		Mean SBP = $\beta_1 + \beta_2 \text{ x } 0 + \beta_3 \text{ x } 0 + \beta_4 \text{ x } 0$

Effect (mean difference)		$\boldsymbol{\beta_2}$

To fit the regression model by SAS, we should create the dummy variables in a data step. Then, we add them to the "model statement", replacing the original age variable.

SAS code

```
DATA one;
IF     age<55 THEN DO;   age2=0; age3=0; age4=0; END;
IF 55<=age<65 THEN DO;   age2=1; age3=0; age4=0; END;
IF 65<=age<75 THEN DO;   age2=0; age3=1; age4=0; END;
IF 75<=age    THEN DO;   age2=0; age3=0; age4=1; END;
```

```
PROC GLM;
 MODEL sbp = age2 age3 age4/SOLUTION;
 run;
```

Selected SAS printout

```
                              The GLM Procedure

Dependent Variable: sbp       SYSTOLIC BLOOD PRESSURE (mmHg)

                                         Sum of
Source                       DF         Squares

Model                         3      55825.2486

Error                       996     399783.1074

Corrected Total             999     455608.3560



                                              Standard
       Parameter         Estimate                Error

       Intercept      116.5308642
       age2             8.9982267           1.64269566
       age3            14.2303298           1.65425898
       age4            21.9465794           2.06320473
```

The regression equation is therefore,

Mean SBP = 116.5 + 9.0 x AGE2 + 14.2 x AGE3 + 21.9 x AGE4

And the coefficients are identical to the numbers in Figure 9–5. Again, 116.5 mmHg is the mean SBP in the first age interval, whereas 9.0, 14.2, and 21.9 are the mean differences between the next three intervals and the first interval.

It is not entirely clear whether the step function supports a straight line. If the differences between adjacent means (the "risers") were identical, then a straight line would have fit perfectly. But the three risers are 9.0; 5.2 (=14.2–9.0); and 7.7 (=21.9–14.2)—not identical and not even monotonically increasing or monotonically decreasing. On the other hand, these numbers have followed an arbitrary choice of the intervals and it is possible that a different choice (five-year intervals, for example) would have generated a different pattern. So, how do you decide which intervals to choose? How do you decide how to categorize a continuous exposure to explore the dose-response function?

I have no simple answer to offer. We should naturally prefer many small intervals because a sequence of small steps gets us closer to the image of a smooth line. But the unavoidable cost is fewer observations per interval, which means imprecision of the means and a bouncy graph. Unfortunately, no rule can strike a balance between the preference and the cost. It is another example of the bias-variance tension (chapter 8), and another reminder that no algorithm can take us to the Truth. Many researchers use the percentile distribution of the exposure to categorize the sample into quartiles or quintiles (equal size groups); and if the sample is small, tertiles may be the limit. Others choose the cutoff points according to prior assumptions about "clinically important values", or simply for reasons of simplicity, as I have done here. The idea of statistical efficiency sometimes support an even splitting of the sample size, though at the end, all choices share an element of arbitrariness. If you try a few options and still draw the same inference about the dose-response function, that's a good sign.

More on "dummy coding"

Dummy coding followed by linear regression is more than a two-step method to explore the dose-response function. This method helps us to compute the mean differences between the *k* values of any categorical variable ($k \geq 3$). Again, neither dummy coding nor linear regression is needed to compute marginal associations, but both will prove essential when we'll need to account for confounders and effect modifiers.

To illustrate the two steps of the method, consider the three categories of smoking status (never smoking, former smoking, and current smoking) and the same postulated effect: systolic blood pressure. If you want to estimate the mean difference in SBP between pairs of smoking categories, choose one category as the reference (say, never smokers) and create two dummy variables (Table 9–5.)

Table 9–5. Replacing the three categories of smoking status with two "dummy variables"

Smoking Status	VAR1	VAR2
Never smoker	0	0
Former smoker	1	0
Current smoker	0	1

Then, fit the model "Mean SBP = $\beta_0 + \beta_1$ VAR1 + β_2 VAR2". This model provides all that we need: three means and three mean differences:

Mean SBP (never smokers) = $\beta_0 + \beta_1 \text{ x } 0 + \beta_2 \text{ x } 0 = \boldsymbol{\beta_0}$
Mean SBP (former smokers) = $\beta_0 + \beta_1 \text{ x } \mathbf{1} + \beta_2 \text{ x } 0 = \boldsymbol{\beta_0 + \beta_1}$
Mean SBP (current smokers) = $\beta_0 + \beta_1 \text{ x } 0 + \beta_2 \text{ x } \mathbf{1} = \boldsymbol{\beta_0 + \beta_2}$

Evidently, β_1 and β_2 are mean differences: β_1 —between former smokers and never smokers; β_2—between current smokers and never smokers. The mean difference between current smokers and former smokers can be easily computed: $(\beta_0 + \beta_2) - (\beta_0 + \beta_1) = \beta_2 - \beta_1$

Dummy variables always take the values 0 and 1, and their number is always *k*–1: one fewer than the *k* categories of the variable they replace. A dummy variable may be given any name, but it's helpful to check which category is identified by the value of 1, and name the variable after that category. For example, we may give VAR1 the name FORMER because the value of 1 identifies the former smokers. Likewise, we may give VAR2 the name CURRENT because the value of 1 identifies the current smokers (Table 9–6).

Table 9–6. Replacing the three categories of smoking status with two "dummy variables"

Smoking Status	FORMER	CURRENT
Never smoker	0	0
Former smoker	1	0
Current smoker	0	1

$$\text{Mean SBP} = \beta_0 + \beta_1 \text{ FORMER} + \beta_2 \text{ CURRENT}$$

Naming the dummy variables in this way should help us to interpret the coefficients quickly: β_1 in front of FORMER is the mean difference between former smokers and the reference category, whereas β_2 in front of CURRENT is the mean difference between current smokers and the reference.

If you took a course in statistics, you might have learned about *analysis of variance* (ANOVA) as a method to compute and compare three means or more. To set the record straight, you should know that what is called one-way ANOVA is equivalent to linear regression on dummy variables. So what is the difference between the two approaches?

ANOVA puts the emphasis on the means themselves rather than on the mean differences, and on statistical hypothesis testing of the global null "all of the means are equal". As we realized earlier, however, the means are also available from a comparable regression model with dummy variables: one mean shows up on the printout (the intercept) and the others can be computed easily by adding the coefficient of the appropriate dummy variable. Even the global null hypothesis of ANOVA is tested in linear regression, but I deleted the test statistic and the *p*-value for two reasons: First, my arguments against the use of p-values hold here as well (chapter 8). Second, the "overall null" should be of little interest because its rejection endorses the statement "at least two means are not equal", which entails the possibility that just two (unspecified) means are not equal. What do we learn from the last statement? Not much, if anything at all. In ANOVA you can also directly test null hypotheses about the equality of any pair of means, but the merit of these tests is just as questionable as the merit of any *p*-value. In summary, unless you have residual sympathy for null hypothesis testing, or have unique interest in the means themselves, I suggest that you think about ANOVA as a special case of the linear regression model and archive the term in your mind. (By the way, two-way ANOVA is also not much more than linear regression.)

A quadratic function

Less restrictive than a linear function but perhaps more restrictive than a step function—is a quadratic function. Its rationale is simple. Rather than forcing a straight line, allow the dose-response line to express some curvature. Not any curved shape, of course, but the kind of structured curvature of a quadratic equation ($Y = a + bX + cX^2$).

In our example: Mean SBP = $\beta_0 + \beta_1$ AGE + β_2 AGE^2

To fit this regression model in SAS, you have two options:

1) Create a new variable "AGESQUARE=AGE*AGE" in a data step, and add it to the "model statement".
2) Add the term "AGE * AGE" directly to the "model statement", as shown below.

Regardless of the method you choose, the variable AGE should be retained in the model. Whenever a "high-order term", such as AGE^2, enters the model, all lower-order terms must be included as well.

SAS code

```
PROC GLM;
 MODEL sbp = age age*age/SOLUTION;
 run;
```

Selected SAS printout

```
                                The GLM Procedure

Dependent Variable: sbp   SYSTOLIC BLOOD PRESSURE (mmHg)

                                          Sum of
Source                       DF          Squares

Model                         2       58179.9051

Error                       997      397428.4509

Corrected Total             999      455608.3560



       Parameter          Estimate

       Intercept       68.95768659
       age              1.12145239
       age*age         -0.00308108
```

The linear regression equation is therefore,

$$\text{Mean SBP} = 69.0 + 1.1 \text{ x AGE} - 0.003 \text{ x AGE}^2$$

You might be wondering why I used the term "linear regression" for such a quadratic equation. The explanation is simple: "linear" refers to the regression coefficients, not to the variables. The coefficients, β_1, β_2, and β_3, form a linear combination; we do not use terms such as the square of β_1 or β_1 x β_2.

As you may know, the graph of a quadratic function ($Y = a + bX + cX^2$) is a parabola, which has a maximum value ("dome" shape), or a minimum value ("inverted dome"), depending on the signs of the constants *b* and *c*. Since the coefficient of AGE is positive and that of AGE^2 is negative, the function we found should resemble a dome, having a maximum value of SBP. Figure 9–6 displays the graph of that quadratic function for the age range of the sample. As you can tell, the graph barely differs from a straight line and no parabola is in sight. Why do we see only minimal curvature? Where is the dome?

Figure 9–7 reveals the answers. When the function is displayed over the non-existing age range of 30 to 330, we see a parabola with a maximum value around the "age" of 180. (Calculus tells us that the maximum will be reached at "AGE"= $-\beta_1/2\beta_2$.) It also becomes clear why the graph in Figure 9–7 resembles a straight line. In that segment of the function, there is very little curvature. To sum up, the quadratic function largely agrees with the linear function over the age range of the sample.

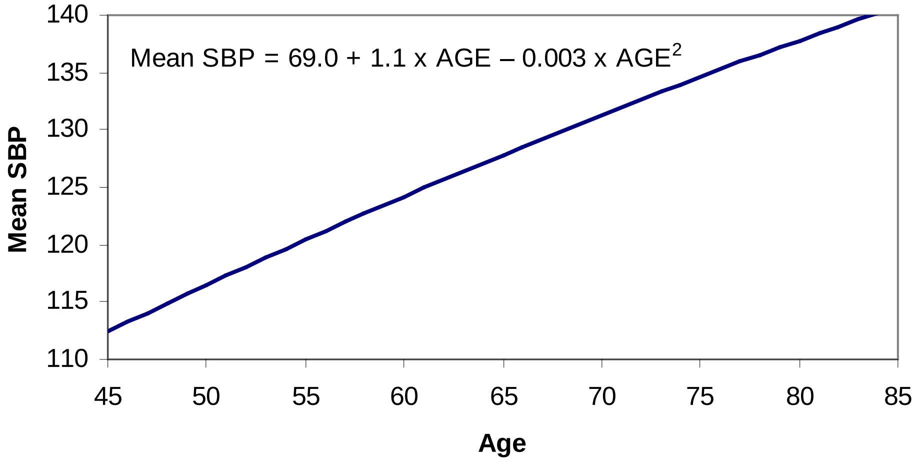


Figure 9–6. Mean SBP as a quadratic function of AGE (ages 45-84)

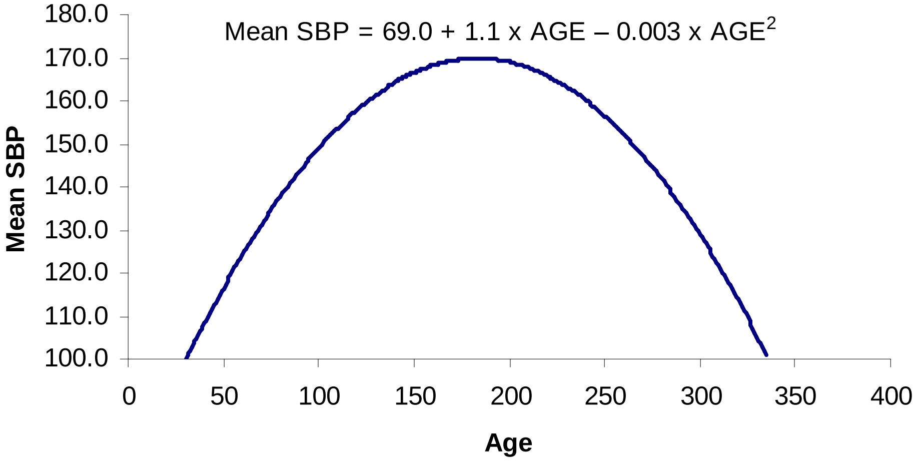


Figure 9–7. A wide view of mean SBP as a quadratic function of AGE ("ages" 30-330)

How do we compute estimates of the mean difference from a quadratic dose-response function?

The coefficients of AGE and AGE^2 are not interpretable individually, but the method is not different from the method we have used to compute the mean difference from a linear function. Table 9–7 shows an example for the causal contrast [50, 60].

Table 9–7. Computing the effect of a causal contrast between age 60 and age 50 from the quadratic function

Causal assignment	Values of age variables	Mean SBP = 69.0 + 1.1 x AGE – 0.003 x AGE^2
AGE = 60	AGE = 60 AGE^2 = 3600	Mean SBP = 69.0 + 1.1 x 60 – 0.003 x 3600 = 124.2
AGE = 50	AGE = 50 AGE^2 = 2500	Mean SBP = 69.0 + 1.1 x 50 – 0.003 x 2500 = 116.5
Effect (mean difference)		7.7 mmHg

Notice that the estimate (7.7 mmHg) is similar to the estimate from the straight line (0.74 x 10 =7.4), which is not surprising. We have already realized that in the sample's range of ages, the graph of the quadratic function is not that different from a straight line. The estimate from the step function (Figure 9–5) is a little larger (9.0 mmHg).

*

Quadratic functions are typically fit to explore the dose-response function, as we have done here, but the model hides a deeper secret—a special kind of effect modification. To detect that property of the function, let's compute again the effect of 10 years of aging, but this time we'll try a different causal contrast: AGE=70 versus AGE=60 (Table 9–8).

Table 9–8. Computing the effect of a causal contrast between age 70 and age 60 from the quadratic function

Causal assignment	Values of age variables	Mean SBP = 69.0 + 1.1 x AGE – 0.003 x AGE^2
AGE = 70	AGE = 70 AGE^2 = 4900	Mean SBP = 69.0 + 1.1 x 70 – 0.003 x 4900 = 131.3
AGE = 60	AGE = 60 AGE^2 = 3600	Mean SBP = 69.0 + 1.1 x 60 – 0.003 x 3600 = 124.2
Effect (mean difference)		7.1 mmHg

The effect of 10 years of aging starting at age 60 (Table 9–8) is different from the effect of 10 years of aging starting at age 50 (Table 9–7). In fact, for each pair of ages that differ by 10 years, you will find a unique estimate.

Unlike a linear function, the estimated mean difference from a quadratic function is not a constant coefficient anymore. In the language of causation, the model assumes that the effect of Δ years of aging on mean systolic blood pressure *varies by age*, which is nothing but a claim of effect-modification. The effect of the exposure is modified by the exposure itself! Furthermore, calculus-based math, as well as the convexity of the graph, reveals two interesting derivations from our quadratic function: First, the effect of aging on systolic blood pressure is attenuated with aging. Second, that effect is attenuated in a monotonic fashion, precisely as a linear function of age. The first derivation may be tolerated as "plausible", but the second reveals the restrictive face of a quadratic function. Not only does age modify the effect of age, but it also does so in a strictly monotonic fashion with a constant degree of attenuation. You might agree that the last derivation does not sound much more plausible than the assumption of a constant mean difference of the linear function or the two assumptions of the step function. But, again, had we chosen to obey the vague psychological idea of "plausibility", scientific inquiry would not have taken us very far. (Just recall how implausible Einstein's ideas were to the human mind.)

Where do we go from here?

Linear function, step function, and quadratic function are not the only methods to explore the dose-response function. In chapter 21, you will find another approach that allows the data to express greater flexibility of the dose-response line: quadratic spline regression. To some extent, that method is a hybrid of the step function and the quadratic function: First, we decide on cutoff points for the exposure distribution, just as we do in a step function. Then, we fit a quadratic function in each interval, but ensure that the end of one segment smoothly merges with the beginning of the next segment, so the resulting line is continuous.

Methods to discover the dose-response relation are not restricted to linear regression; they are helpful in other regression models as well (logistic, Poisson, Cox.)

Deconfounding by linear regression

The marginal association between systolic blood pressure and age, as estimated by the "crude" mean difference, might contain not only the effect of age but also the effect of confounding paths. For example, on the naïve assumptions of the causal diagram below (Figure 9–9), the various mean differences we have computed so far do not estimate the effect of age. They are biased. Or more precisely (chapter 8): the estimators from which they emerged embed confounding bias due to sex group.

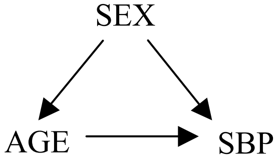


Figure 9–9. A causal diagram relating age, sex, and systolic blood pressure

Before finding out how linear regression could deconfound the mean difference, let's spend another moment on Figure 9–9. I had assumed that sex group is antecedent to age and not vice versa, which is my background conjecture about the determinants of life expectancy. But this claim may be challenged. After some thinking about surrogate variables (chapter 2), I could have also rationalized an arrow in the opposite direction, turning SEX into an intermediary variable between AGE and SBP. If, for example, AGE were a surrogate for the efficiency of metabolic pathways and SEX were a surrogate for the levels of sex hormones (which are products of such pathways), I could have proposed that AGE→SEX, and thereby eliminate the need to deconfound. In short, causal diagrams and the analytical route they dictate are as good as our underlying theories.

Assuming that Figure 9–9 correctly describes causal reality, we should deconfound the marginal association between age and systolic blood pressure by conditioning on sex. That is: we should stratify on sex, compute the mean difference in men and women, and calculate a (weighted) average of the two mean differences (chapter 6). Regression models perform the very same task, albeit behind the scenes—and much more. Not only do they allow us to condition the association on one binary variable such as sex, but they also offer simultaneous conditioning on several variables, including continuous variables on which stratification is not possible.

Regression-based conditioning is achieved by simply adding the confounders as "covariates" to the right hand side of the model, thereby creating a multivariable model. (Many researchers refer to these models as *multivariate* regression, but the correct adjective is *multivariable* or *multiple*. "Multivariate" denotes a model with several *dependent* variables.)

In our example, where only one confounder is proposed, we should fit the following equation: $\text{SBP} = \beta_0 + \beta_1\ \text{AGE} + \beta_2\ \text{SEX} + e$

And look for a solution, for the values of the coefficients.

If we use the method of linear regression to solve the equation, the coefficient of AGE will estimate the conditional mean difference per 1 year of aging, acting like a weighted average of the sex-specific mean differences. Most people will call that number the "sex-adjusted" mean difference, but the term *conditional* is far more accurate. The word "adjusted" delivers the promise of something better than the "crude", yet we never know that one estimate is better than another, because the causal diagram we drew might be wrong. C*onditional* is always a true claim about reality; *adjusted* is not.

How do we solve the regression equation?

When a linear regression model contains two regressors or more, it is no longer possible to display pairs of data points in a 2-dimentional graph, fit a candidate line, and calculate vertical distances. At most we may display triplets of data points (AGE, SEX, SBP) in a 3-dimentional graph, fit a candidate *surface*, and calculate the vertical distances of observed SBP values from the surface. And if there are three regressors, no graphical display is possible anymore. Nonetheless, the method to solve the equation is identical in all cases, regardless of the number of variables. Instead of drawing a candidate line or a candidate surface, we simply have to propose candidate *values of the coefficients*.

In our example (SBP = $\beta_0 + \beta_1$ AGE + β_2 SEX + e), you would solve the equation according to the following theoretical steps:

1. Propose candidate *values* of the coefficients (β_0, β_1, β_2).
2. Use the candidate coefficients to compute the predicted value of SBP for each person: SBP (predicted) = $\beta_0 + \beta_1$ AGE + β_2 SEX
3. Calculate the difference, e, between the observed SBP of each person and the predicted value: e = SBP (observed) – SBP (predicted)
4. Square the difference: e^2
5. Sum the squares: Σe^2
6. Among *all* possible values of β_0, β_1, and β_2, choose the set of values for which you get the smallest sum of squares.

SAS code

```
PROC GLM;
 MODEL sbp = age sex/SOLUTION CLPARM;
 run;
```

Selected SAS printout

```
                                 The GLM Procedure

Dependent Variable: sbp   SYSTOLIC BLOOD PRESSURE (mmHg)

                                        Sum of
Source                      DF         Squares

Model                        2      58795.1134

Error                      997     396813.2426

Corrected Total            999     455608.3560


                                           Standard
       Parameter         Estimate             Error   95% Confidence Limits

       Intercept      79.82561708
       age             0.73522359        0.06117390   0.61517923   0.85526796
       sex             1.69409655
```

The conditional mean difference per 1 year of aging (0.735 mmHg before rounding) barely differs from the marginal estimate we computed earlier (0.738 mmHg), implying that confounding by sex did not exist. This little surprise, which is not uncommon in causal inquiry, gives us the opportunity to entertain a few explanations and to understand better the constraints of scientific uncertainty. First, there is no guarantee that a true theory will always be corroborated in the empirical world—including a true theory about confounding paths. Second, both 0.735 and 0.738 are single estimates from different estimators whose expected values remain unknown (chapter 8). We tend to forget—perhaps wish to forget—that remarkable similarity of two point estimates does not necessarily imply remarkable similarity of the corresponding expected values. (Remember that each *estimate* is no more than what the word means.) Third, properties of the sample at hand inevitably affect the results of conditioning (chapter 7). In this sample, the marginal association between age and sex was weak, in part because of sampling procedures, and therefore conditioning on sex was not expected to have made a striking difference, if at all. Fourth, as we'll see in the next chapter, it is possible that both estimates have originated in biased estimators so their near-perfect agreement might be irrelevant.

Just as we conditioned the linear association between age and blood pressure on sex, we may also condition other functions by which we explored the dose-response relation. Here are the models we would fit to deconfound (SAS code and printout omitted):

Step function: $\text{Mean SBP} = \beta_1 + \beta_2\,\text{AGE2} + \beta_3\,\text{AGE3} + \beta_4\,\text{AGE4} + \beta_5\,\text{SEX}$

Quadratic function: $\text{Mean SBP} = \beta_1 + \beta_2\,\text{AGE} + \beta_3\,\text{AGE}^2 + \beta_4\,\text{SEX}$

Finally, you might wonder why I didn't comment on the coefficient of SEX in the multivariable model (1.7 mmHg), and even deleted its standard error. Well, this coefficient is also a conditional mean difference: the mean difference in SBP between men and women—after conditioning on age. But according to the causal diagram I drew (Figure 9–9), we should *not* condition on age if we wish to estimate the effect of sex group on systolic blood pressure. AGE is an intermediary variable on a path from SEX to SBP, rather than a confounder! More on this topic in the next section.

To make matters worse, notice that we cannot confidently predict the effect of conditioning just from knowledge of some qualitative properties of the sample. Even though AGE and SEX were weakly associated in our sample and the coefficient of AGE barely changed after adding SEX to the model, conditioning on AGE *did* change the coefficient of SEX: from 2.3 in the model "$\text{Mean SBP} = \beta_0 + \beta_1\,\text{SEX}$" (the first printout in this chapter) to 1.7 in the last printout. Part of the explanation has to do with a strong association between age and blood pressure in the sample.

Beware of biased coefficients!

"Not all regression coefficients are created equal." –Miguel Hernán

Some researchers assume that each coefficient in a multivariable model is estimating the effect of the respective variable after "adjusting for all other covariates." They would say, for instance, that when we regress systolic blood pressure on SEX and WEIGHT, the coefficient of SEX estimates the sex effect whereas the coefficient of WEIGHT estimates the weight effect, each "adjusted for the other." Those who avoid using the words cause and effect might cautiously write "estimates the independent association of each variable"—and still have cause-and-effect in mind.

This assumption may be false in many regression models. The theory of causal diagrams has taught us that conditioning (or what is called "adjustment") is not a symmetrical process in causal inquiry (chapter 6). Some conditional estimates serve to deconfound, whereas others could be worse than the marginal estimates, because confounding is not a reciprocal idea. To illustrate the pitfall, let's compare the estimated effect of sex group on systolic blood pressure from two linear regression modes: a model with SEX alone (the marginal association we've already seen) and a model that includes SEX, AGE, and WEIGHT (measured in pounds).

Model 1: $\text{Mean SBP} = \beta_0 + \beta_1\ \text{SEX}$

Model 2: $\text{Mean SBP} = \beta_0 + \beta_1\ \text{SEX} + \beta_2\ \text{AGE} + \beta_3\ \text{WEIGHT}$

Selected SAS printout from the two models is shown side by side:

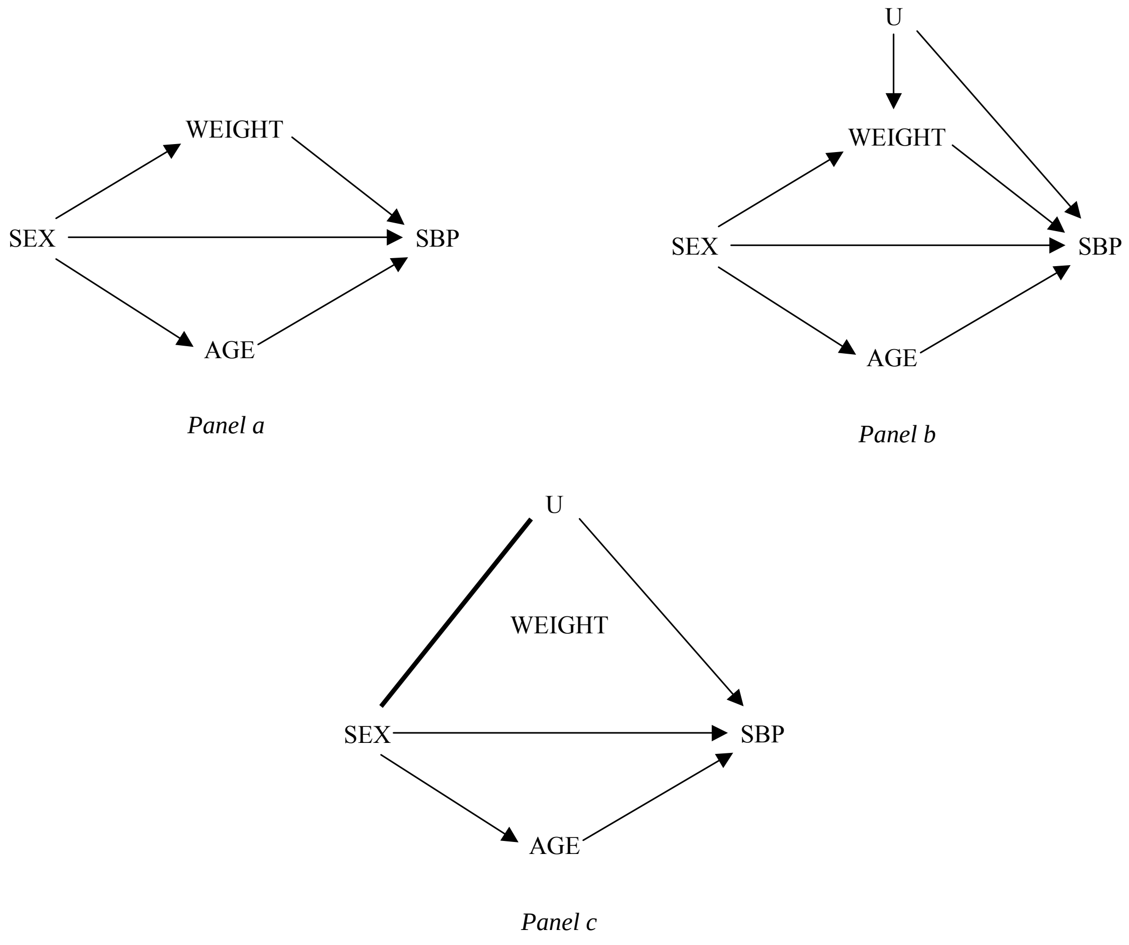
```
Dependent Variable: sbp  SYSTOLIC BLOOD PRESSURE (mmHg)

               Model 1                                   Model 2

                             Sum of                                     Sum of
Source              DF      Squares          Source              DF     Squares

Model                1    1304.4140          Model                3   76710.7659
Error              998  454303.9420          Error              996  378897.5901

Corrected Total 999     455608.3560          Corrected Total 999   455608.3560


Parameter         Estimate                   Parameter         Estimate

Intercept      124.5202991                   Intercept      56.34767205
sex              2.2889114                   sex            -0.85095005
                                             age             0.79548598
                                             weight          0.11963538
```

By statistical criteria, every statistician will prefer model 2 because "it fits the data better". The sum of the squared "error", $\Sigma\ e^2$, is much smaller in model 2, which means that the model should predict the value of systolic blood pressure much better. Nonetheless, you will shortly see that bias may be lurking behind the negative coefficient of SEX (–0.85). That a regression model might do a better job in predicting John's blood pressure does not endow all of its coefficients with the title "measure of effect". Try to keep in mind the asymmetrical relation between causal inquiry and statistical prediction, which so many seem to forget: to estimate a causal parameter, we often seek help from a prediction model, but not every prediction model, however good it may be, delivers unbiased estimators of causal parameters. Please read the last sentence again and be sure to share it occasionally with your fellow statisticians.

Figure 9–10 (*panel a*) shows a diamond-shaped causal diagram that connects the four variables of interest. In addition to arrows from SEX to AGE and from SEX to SBP, which we had assumed earlier, I drew a path from SEX to SBP via the variable WEIGHT. According to this simplistic diagram, the marginal association between sex group and systolic blood pressure (model 1) is not confounded. The expected value of the estimator behind the coefficient of SEX should be equal to the causal parameter—to the net effect of sex on blood pressure via the three causal pathways. No conditioning is needed.

Could conditioning (model 2, for example) cause any harm?

As you might recall from chapter 6, it certainly can—if we happen to condition on colliders. Suppose that weight and blood pressure share a common cause, U, such as a genotype or a hormone (Figure 9–10, *panel b*). On this assumption, WEIGHT is a collider on the path SEX→WEIGHT←U→SBP, and conditioning on it will open a confounding path via the colliding variables (*panel c*). As a result, the "adjusted" mean difference between men and women (model 2) will not deconfound anything. On the contrary: it will contain the confounding effect of the path we have opened.



Panel a

Panel b

Panel c

Figure 9–10. A causal diagram relating sex, age, weight, and systolic blood pressure (*panel a*); a common cause, U, of weight and systolic blood pressure was added (*panel b*); causal and confounding paths after conditioning on weight (*panel c*).

At this point, some researchers might be inclined to say that adjustment for weight and age (model 2) has "explained" the effect of sex group on systolic blood pressure, or perhaps has estimated the direct effect, SEX→SBP, alone. Unfortunately, neither is necessarily true. First, if U does exist (*panel b*), we have just realized that the coefficient of SEX in the second model contains an artificial component, which we have created by conditioning on a collider. Whatever the difference between the two coefficients of SEX may tell us, if anything useful at all, it is not only an "explanation" of causal pathways. Second, even if U is absent, other assumptions must be invoked when we try to separate a direct effect from indirect effects by adjusting for intermediary variables.

The proportion difference: a special kind of a mean difference:

Suppose that the effect of interest is not a continuous variable such as SBP, but a binary variable called HTN (hypertension status): 1=hypertension; 0=normotension. If we fit a linear regression model, regressing hypertension status on sex group, the model will predict the mean of HTN, and the coefficient of SEX will estimate the mean difference in HTN between men and women:

Mean HTN = $\beta_1 + \beta_2$ SEX

But what is the meaning of that mean?

After thinking for a moment, you would probably realize that "mean HTN" is simply a proportion: the proportion of people with hypertension. (The mean of a binary "0, 1" variable turns out to be the proportion of "ones", because the sum of "zeros" and "ones" divided by the number of observations is simply the proportion of "ones".) Moreover, if the proportion in question may also be called "probability", we seem to have found a model that estimates the *probability difference*, a measure of effect. It is called the *linear probability model*:

$$\Pr(\text{HTN}=1) = \beta_1 + \beta_2 \text{ SEX}$$

Life is never simple, though. As you may recall, the precise specification of linear regression includes notation for the *i*-th person and an "error term". Unfortunately, when the dependent variable is binary, rather than continuous, some of the assumptions about the behavior of the error term do not hold, and therefore, it is not strictly valid to fit such a model. Another problem arises because the expression on the right ($\beta_1 + \beta_2$ SEX) can predict probability values that do not exist (greater than 1, for example). Nonetheless, knowledgeable authors have reassured us that the coefficients of the linear probability model are still unbiased and that the consequences of violating some of the statistical assumptions should not disrupt our sleep. Just remember to not give too much weight to the standard errors because they are wrong.

To illustrate how we would estimate the (marginal) probability difference between men and women, I created a binary variable called HTN by dichotomizing systolic blood pressure at 140 mmHg. Every blood pressure value greater than 140 qualified for HTN=1; otherwise HTN=0. In our sample of 1,000 people, 22.9% of men and 21.1% of women met that criterion of hypertension, a difference of 1.8 percentage points against men. Keep these numbers in mind as we fit the linear probability model (below).

SAS code

```
PROC GLM;
 MODEL htn = sex/SOLUTION;
 run;
```

Selected SAS printout

```
                              The GLM Procedure

Dependent Variable: htn   HYPERTENSION STATUS

                                            Sum of
Source                        DF           Squares

Model                          1         0.0787513
Error                        998       172.0802487
Corrected Total              999       172.1590000


       Parameter         Estimate

       Intercept     0.2115384615
       sex           0.0177848467
```

The regression equation of this linear probability model is therefore:

Pr (HTN=1) = 0.211 + 0.018 SEX

For women, Pr (HTN=1) = 0.211 + 0.018 x 0 = 0.211, identical to the actual proportion of hypertensives among women (21.1%).
â
For men, Pr (HTN=1) = 0.211 + 0.018 x 1 = 0.229, identical to the actual proportion of hypertensives among men (22.9%).

As is always the case with linear regression, the coefficient of SEX estimates the mean difference in the dependent variable between men and women. Here, that coefficient is the probability difference of having hypertension. It is 0.018—identical to the percentage difference in the sample (22.9% – 21.1% = 1.8 percentage points).

Of course, you don't really need the linear probability model to compute marginal associations, but the model becomes essential when you have to deconfound. Consider, for instance, the following task: Estimate the probability difference of hypertension in ascending age groups (reference: the youngest) after conditioning on sex group. The simple code below, with three dummy variables for age, provides the requested estimates.

SAS code

```
PROC GLM;
 MODEL htn = age2 age3 age4 sex/SOLUTION;
 run;
```

Selected SAS printout

```
                          The GLM Procedure

Dependent Variable: htn    HYPERTENSION STATUS

                                          Sum of
Source                      DF           Squares

Model                        4        13.2435394
Error                      995       158.9154606
Corrected Total            999       172.1590000


      Parameter          Estimate

      Intercept      0.0963533943
      age2           0.0867445207
      age3           0.1956133640
      age4           0.3446042887
      sex            0.0047652453
```

The regression equation:

Pr (HTN=1) = 0.096 + 0.087 x AGE2 + 0.196 x AGE3 + 0.345 x AGE4 + 0.005 x SEX

Compare this equation to the regression of systolic blood pressure on age group (the step function.) To interpret the coefficients here, you just have to substitute the words "probability of hypertension" for "mean systolic blood pressure". For example, the coefficient of AGE4 (0.345) is the probability difference of hypertension for the contrast between the oldest group and the youngest group, after conditioning on sex.

Before leaving this chapter behind, let's reflect for a moment on the conversion of systolic blood pressure to the binary variable HTN. Was it wise to do so? May we transform a continuous dependent variable (presumed effect) into a categorical variable, using one or more cutoff points of its distribution?

We may do whatever we want, of course, but in my view this common practice is mistaken. The effect of sex group on the binary variable we created (HTN) is solely due to its effect on systolic blood pressure; there is no other mechanism by which sex group could affect hypertension (as defined here). Why, then, model an artificial surrogate for the real effect if we can model the real effect? Moreover, by replacing a continuous dependent variable with some categorical version of it, we are making two mistakes: First, we unnecessarily increase the standard errors of the estimators. Second, we throw away detailed data (blood pressure values) for no good reason. (Do not draw analogy to categorization of a continuous exposure, which serves our interest in the dose-response function.)

We will return to the linear probability model toward the end of the next chapter.