Chapter 10

Estimating the Modified Mean Difference

Sex group: an effect modifier?

At the end of the last chapter we tried to estimate the effect of age on systolic blood pressure, assuming that the marginal association between the two variables contained the confounding effect of sex (Figure 9–9). To deconfound, we regressed SBP on AGE and SEX, and found the following solution:

$$\text{Mean SBP} = 79.8 + 0.7 \times \text{AGE} + 1.7 \times \text{SEX}$$

As you know by now, the coefficient of age in this model (0.735 mmHg before rounding) is a conditional mean difference: a weighted average of the sex-specific coefficients. Although regression-based conditioning, unlike simple stratification, does not reveal the sex-specific estimates, we can find them by regressing SBP on AGE in each sex group, using the following code:

SAS code

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

Selected printout from the two regression models is shown below, side-by-side.

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

	WOMEN		MEN	
Number of observations	**468**		**532**	
Source	DF	Sum of Squares	DF	Sum of Squares
Model	1	47940.7502	1	15473.0107
Error	466	204860.3069	530	186029.8742
Corrected Total	467	252801.0572	531	201502.8849
Parameter	Estimate	Standard Error	Estimate	Standard Error
Intercept	64.22615646		94.88851349	
age	**0.99183335**	0.09497783	**0.51819630**	0.07804773

The simple average of the sex-specific coefficients of age is (0.992+0.518)/2=0.755, a little larger than the weighted average from the regression of systolic blood pressure on age and sex (0.735). Since the weighted average of 0.735 resides closer to the estimate in men (0.518) than to the estimate in women (0.992), the men in the sample must have pulled the average toward their end more strongly than have the women. To understand the reason for the inequity, compare the two standard errors (0.078 versus 0.095) behind the sex-specific coefficients: the men's estimate has emerged from an estimator that has a smaller standard error (a tighter sampling distribution), in part because men outnumbered women in the sample. Their estimate rightly carried more weight.

The equation "Mean SBP = 79.8 + 0.7 x AGE + 1.7 x SEX" is called a *main effect model*, a name that serves to distinguish it from an *interaction model*, which we'll encounter in the next section. Again, the math of that model assumes that the sex-specific estimates of the mean difference have originated from a *single parameter*, and if so, the best estimate of that number must be a weighted average of the two estimates. From the viewpoint of cause-and-effect (chapter 4), we assume homogeneity of the causal parameter across the strata of sex, and request the model to compute a single estimate for both men and women.

To realize the last, key point, I explicitly derived the estimated effect of 1 year of aging in each sex group from the main effect model (Table 10–1). As you see, it is the same estimate: 0.7.

Table 10–1. Deriving the effect of 1 year of aging on systolic blood pressure in each sex group from the main effect model

	Mean SBP = 79.8 + 0.7 x AGE + 1.7 x SEX
Causal assignments (women)	
AGE=K+1 and SEX=0	Mean SBP = 79.8 + 0.7 x (K+1) + 1.7 x 0
AGE=K and SEX=0	Mean SBP = 79.8 + 0.7 x K + 1.7 x 0
Effect of aging in women (mean difference in SBP)	0.7
Causal assignments (men)	
AGE=K+1 and SEX=1	Mean SBP = 79.8 + 0.7 x (K+1) + 1.7 x 1
AGE=K and SEX=1	Mean SBP = 79.8 + 0.7 x K + 1.7 x 1
Effect of aging in men (mean difference in SBP)	0.7

Regardless of whether sex group played the role of a confounder, the models we have fit so far—both of marginal associations and of conditional associations—are misleading if sex happens to modify the effect of age on systolic blood pressure. If this is the case, we have been following the wrong path: the estimators are all biased because the causal parameter takes *two* values, not one (chapter 4). Indeed, the sex-specific estimates cast some doubt on the wisdom behind our models: the estimated effect of age on systolic blood pressure in women (a mean difference of 0.99 mmHg per 1 year of aging) is about twice the estimated effect in men (0.52 mmHg). Perhaps the two numbers estimate *two* values of the causal parameter (or, if you prefer, *two* causal parameters).

Stratified regression, as shown in the last printout, could help us to explore effect modification, but the method quickly fails when a little complexity is added; for example,

when the effect modifier contains many strata or happens to be a continuous variable itself. It would, therefore, be helpful to find a way to estimate heterogeneous effects from a single regression model. In the next section, we will study the general method in its simplest form—with one binary exposure (E) and one binary effect modifier (M). Later, we'll return to the example of age, a continuous variable, and sex.

The regression viewpoint of effect-modification (interaction)

In chapter 5, we explored both interaction and effect modification in the context of 2x2 tables and realized that the two ideas are tightly linked mathematically: one can be derived from the other by simple arithmetic. In this section we will discover the mathematical expression of these ideas when effects are estimated by a linear regression model. Since our measure of effect is the mean difference, interaction and effect modification will necessarily be computed on an additive scale, but much of the theory applies to ratio measures of effect and to other kinds of regression models (logistic, Poisson, Cox).

Let E be a binary exposure (1,0), such as alcohol drinking status (drinking or abstaining), and let M be another binary variable (1,0), such as smoking status (smoking or not smoking). Let Y be their effect—a continuous dependent variable such as systolic blood pressure. As we saw earlier, if we wish to condition the association of E with Y on M (having deconfounding in mind), we will typically regress Y on M and E simultaneously:

$$\text{Mean } Y = \beta_0 + \beta_1 M + \beta_2 E \qquad \text{(Equation 10–1)}$$

If conditioning on M has followed the deconfounding rules of a causal diagram, the coefficient of E (β_2) estimates the effect of E on Y (assuming no other confounders). Specifically, β_2 is the mean difference in Y for the causal contrast between E=1 and E=0.

If, however, we are entertaining the idea that M is an effect modifier, rather than a confounder, we should add to the main effect model another independent variable called an "interaction term":

$$\text{Mean } Y = \beta_0 + \beta_1 M + \beta_2 E + \beta_3 (M \times E) \qquad \text{(Equation 10–2)}$$

In this *interaction model*, Y is regressed on three variables: on M, on E, and on a new variable which is the product of M and E. Sometimes the analytical software requires you to create that product variable in a data step (Z=M*E); other times, you may explicitly fit the model with a product term (type in "M*E"). Note that regardless of any technical specification, MxE is a variable, just like M and E. Table 10–1 shows all possible values of these three variables:

Table 10–1. Possible values of M, E, and M x E

M	E	M x E
0	0	0 x 0 = 0
0	1	0 x 1 = 0
1	0	1 x 0 = 0
1	1	1 x 1 = 1

In our minds the multiplication of two variables might resonate with the causal idea of interaction, but intuition is not good enough. How do we know that such a model allows us to examine an interaction between E and M? Moreover, if you prefer to talk about effect modification by M (as I do), how does this model allow the effect of E to vary according to the values of M (or vice versa)?

Let's examine first the idea of interaction, which alludes to "the joint effect of M and E". What does our model say about that effect?

To find out we should contrast the pair of causal assignments "M=1 and E=1" with the pair "M=0 and E=0". Specifically, we should calculate the mean of Y for each pair of causal assignment (by entering the values of M and E into the regression equation), and then, compute the mean difference (Table 10–2).

Table 10–2. The joint effect of E and M from an interaction model

Causal assignments	Mean $Y = \beta_0 + \beta_1 M + \beta_2 E + \beta_3 (M \times E)$
M=1 and E=1	Mean $Y = \beta_0 + \beta_1 \times 1 + \beta_2 \times 1 + \beta_3 \times (1 \times 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$
M=0 and E=0	Mean $Y = \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 + \beta_3 \times (0 \times 0) = \beta_0$
Joint Effect (mean difference in Y)	$\beta_1 + \beta_2 + \beta_3$

Our "interaction model" has led us to conclude that the joint effect of M and E (the mean difference) is $\beta_1 + \beta_2 + \beta_3$, but how do we interpret each coefficient? What, for example, does β_1 mean in that model?

It is not too difficult to show that β_1 estimates the effect of M (the causal contrast between M=1 and M=0) <u>when E=0</u>. Similarly and reciprocally, β_2 estimates the effect of E (the causal contrast between E=1 and E=0) <u>when M=0</u>. Tables10–3 and 10–4, below, show the formal proofs.

Table 10–3. The effect of M (when E=0) from an interaction model

Causal assignments	Mean $Y = \beta_0 + \beta_1 M + \beta_2 E + \beta_3 (M \times E)$
M=1 (and E=0)	Mean $Y = \beta_0 + \beta_1 \times 1 + \beta_2 \times 0 + \beta_3 \times (1 \times 0) = \beta_0 + \beta_1$
M=0 (and E=0)	Mean $Y = \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 + \beta_3 \times (0 \times 0) = \beta_0$
Effect of M (mean difference in Y)	β_1

Table 10–4. The effect of E (when M=0) from an interaction model

Causal assignments	Mean $Y = \beta_0 + \beta_1 M + \beta_2 E + \beta_3 (M \times E)$
E=1 (and M=0)	Mean $Y = \beta_0 + \beta_1 \times 0 + \beta_2 \times 1 + \beta_3 \times (0 \times 1) = \beta_0 + \beta_2$
E=0 (and M=0)	Mean $Y = \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 + \beta_3 \times (0 \times 0) = \beta_0$
Effect of E (mean difference in Y)	β_2

These effects of M and E (β_1, β_2, respectively) are often called "separate effects" or "independent effects", terms that appeal to our intuition when the zero value of each binary variable makes up a "natural" reference. For instance, if M is smoking status (smoking, non-smoking) and E is alcohol-drinking status (drinking, non-drinking), we will naturally define the separate effect of drinking on blood pressure as its effect in the absence of smoking. Nonetheless, as soon as we replace smoking status with a variable

such as sex group, the term "separate effect" loses its rhetorical force. What should be called the separate effect of alcohol drinking when M is sex: its effect in men or its effect in women? Perhaps neither.

Setting that difficulty aside, you might be inclined to state an expectation about the joint effect of M and E. Intuitively, you probably expect it to be the sum of their so-called separate effects—that is, $\beta_1 + \beta_2$. Our regression model tells us, however, that the joint effect should be $\beta_1 + \beta_2 + \boldsymbol{\beta_3}$. The model allows the joint effect to be greater or smaller than the sum of the separate effects, depending on whether β_3 were positive or negative. If $\beta_3 > 0$, we will claim synergistic interaction (more than we expected from summing the "separate effects"), and if $\beta_3 < 0$ we will claim antagonistic interaction (less than we expected). Theoretically, β_3 could be precisely zero, but that never happens in regression; at most its magnitude might be trivially small. (If your mind is wandering right now in the direction of null hypothesis testing of $\beta_3=0$, please read chapter 8 again.)

*

So far we exclusively examined the model in the language of interaction. One line of algebra, however, will radically change our perspective from interaction to effect-modification, and eliminate the need to discuss deterministic ideas such as "separate effects", synergistic interaction, and antagonistic interaction (sometimes called "sub-additivity".)

Notice that our "interaction model" $\text{Mean } Y = \beta_0 + \beta_1 M + \beta_2 E + \beta_3 (M \times E)$

may also be written as: $\text{Mean } Y = \beta_0 + \beta_1 M + (\boldsymbol{\beta_2 + \beta_3 M}) E$

Examining the second form, we should realize an interesting property of the model. The effect of E, the multiplier in front of E, is not just a constant coefficient anymore as we were used to see in a main effect model. The "coefficient" has turned into a *function* of M. It is "$\beta_2 + \beta_3 M$". No longer can we provide a single answer to the question "What is the effect of E on Y?" because there are two answers:

When M=0, the effect of E on Y (the mean difference) is $\beta_2 + \beta_3 \times 0 = \beta_2$ (as we saw).
When M=1, the effect of E on Y (the mean difference) is $\beta_2 + \beta_3 \times 1 = \beta_2 + \beta_3$

Read out loud the last two sentences and you will find yourself claiming effect modification by M. The effect of E on Y—the mean difference in Y—varies according to the values of M. It is not homogenous across the strata of M.

When examined earlier from the viewpoint of interaction, β_3 estimated the departure from the sum of the so-called separate effects. It was the difference between our model-based joint effect ($\beta_1 + \beta_2 + \beta_3$) and the "expected" joined effect ($\beta_1 + \beta_2$). In the language of effect modification, β_3 plays a different role, estimating the departure from homogeneity of effect. It is the difference between the effect of E when M=1 (which is $\beta_2 + \beta_3$) and the effect of E when M=0 (which is β_2). β_3 tells us how much larger, or how much smaller, is one effect than the other.

Notice that when the coefficient of E (β_2) is positive, the following may be stated:

If $\beta_3 > 0$ (described earlier as synergistic interaction), the effect of E on Y when M=1 is greater than its effect when M=0 (because $\beta_2 + \beta_3 > \beta_2$)

If $\beta_3 < 0$ (described earlier as antagonistic interaction), the effect of E on Y when M=1 is smaller than its effect when M=0 (because $\beta_2 + \beta_3 < \beta_2$.)

In chapter 5 we also discovered that effect modification is a reciprocal idea because the labels "exposure" and "effect modifier" were arbitrary and could have been switched. Since equation 10–2 is symmetrical with respect to E and M, the reciprocal property holds, as shown below:

Instead of expressing the model as $\text{Mean } Y = \beta_0 + \beta_1 M + (\boldsymbol{\beta_2 + \beta_3 M}) E$

We can also express it as $\text{Mean } Y = \beta_0 + \beta_2 E + (\boldsymbol{\beta_1 + \beta_3 E}) M$

So the effect of M on Y has turned into a function of E. Variable E modifies the effect of M on Y.

To conclude this section, let's turn back to equation 10–1 (the main effect model) and compare it to equation 10–2 (the interaction model). In equation 10–1, we fit a model with no interaction term because we had deconfounding in mind and had assumed that M does not modify the effect of E on Y. But notice that equation 10–1 can also be written in the format of equation 10–2 provided that we force the coefficient of M x E to be zero (β_3=0): we can rewrite the main effect model

$$\text{Mean } Y = \beta_0 + \beta_1 M + \beta_2 E$$

as a pseudo-interaction model

$$\begin{aligned}\text{Mean } Y &= \beta_0 + \beta_1 M + \beta_2 E + \mathbf{0}\,(M \times E) \\ &= \beta_0 + \beta_1 M + (\boldsymbol{\beta_2 + 0 \times M})\, E.\end{aligned}$$

But that pseudo-interaction model implies no effect modification by M! For both M=1 and M=0, the estimated effect of E on Y is the same: $\beta_2 + \mathbf{0} \times M = \beta_2$.

To sum up, the main effect model treats M as a confounder and denies the possibility of effect modification by M, implicitly forcing a coefficient of zero for the interaction term, M x E. Indeed, as we realized in chapters 5 and 6, effect modification and confounding never converge. Once we decide that a variable modifies the effect of another variable, we can no longer treat either variable as a confounder of the other. And vice versa: once we decide that a variable confounds the causal relation of interest, we can no longer treat it as a modifier of that relation. (Our decision, of course, may be wrong.) Finally, in regression as in tabular analysis the distance between interaction and effect modification is just one line of algebra, but your choice between the two ideas is not arbitrary; it is a choice between the two trails of causal inquiry. For a deterministic scientist, causes interact with each other and join hands to complete a sufficient cause of the effect, whereas for an indeterministic scientist, causal variables modify each other's effect by altering the background causal propensity (chapters 3-5).

Effect-modification by sex group

To illustrate the math of the previous section, I dichotomized age at 65, and created a binary exposure variable AGEBIN (age, binary) that takes the values 0 and 1 for the "young" and the "old", respectively. Then, I fit the following interaction model:

Mean SBP = $\beta_0 + \beta_1$ SEX + β_2 AGEBIN + β_3 (SEX x AGEBIN) (Equation 10–3)

SAS code

```
PROC GLM;
 MODEL sbp = sex agebin sex*agebin/SOLUTION;
 run;
```

Selected SAS printout

```
                            The GLM Procedure

Dependent Variable: sbp   SYSTOLIC BLOOD PRESSURE (mmHg)

                                   Sum of
Source                      DF        Squares

Model                        3     45179.8130

Error                      996    410428.5430

Corrected Total            999    455608.3560


                                    Standard
      Parameter          Estimate         Error

      Intercept       117.9548495
      sex               5.4051505      1.65884687
      agebin           18.1812452      1.95358849
      sex*agebin      -10.2718486      2.63937319
```

The regression equation is therefore:

Mean SBP = 118.0 + 5.4 x SEX + 18.2 x AGEBIN – 10.3 x (SEX x AGEBIN)

Or alternatively,

Mean SBP = 118.0 + 5.4 x SEX + $\underbrace{(18.2 - 10.3 \text{ x SEX})}_{\text{The age effect}}$ x AGEBIN

From the viewpoint of effect modification, the model forces heterogeneity of the age effect by sex group (Tables 10–5 and 10–6).

Table 10–5. Effect of AGEBIN in women

Causal assignments	Mean SBP = 118.0 + 5.4 SEX + 18.2 AGEBIN – 10.3 (SEX x AGEBIN)
AGEBIN=1 (and SEX=0)	Mean SBP = 118.0 + 5.4 x 0 + 18.2 x 1 – 10.3 (0 x 1)
AGEBIN=0 (and SEX=0)	Mean SBP = 118.0 + 5.4 x 0 + 18.2 x 0 – 10.3 (0 x 0)
Effect of AGEBIN (mean difference in Y)	18.2

Table 10–6. Effect of AGEBIN in men

Causal assignments	Mean SBP = 118.0 + 5.4 SEX + 18.2 AGEBIN – 10.3 (SEX x AGEBIN)
AGEBIN=1 (and SEX=1)	Mean SBP = 118.0 + 5.4 x 1 + 18.2 x 1 – 10.3 (1 x 1)
AGEBIN=0 (and SEX=1)	Mean SBP = 118.0 + 5.4 x 1 + 18.2 x 0 – 10.3 (1 x 0)
Effect of AGEBIN (mean difference in Y)	18.2 – 10.3 = 7.9

In women, the mean difference in systolic blood pressure between "old" and "young" is 18.2 mmHg, whereas in men that difference is 10.3 mmHg *smaller*: 18.2–10.3 = 7.9 mmHg. Recalling the reciprocal property of effect modification, we may also state heterogeneity of the sex effect by age:

$$\text{Mean SBP} = 118.0 + 18.2 \times \text{AGEBIN} + \underbrace{(5.4 - 10.3 \times \text{AGEBIN})}_{\text{The sex effect}} \times \text{SEX}$$

In "young" people (AGEBIN=0) the mean difference between men and women is 5.4 mmHg (men's blood pressure is higher), whereas in "old" people (AGEBIN=1), it is 5.4 – 10.3 = –4.9 mmHg (women's blood pressure is higher). Notice that the sex effect not only changes quantitatively with aging, but it also changes direction, a phenomenon called *qualitative* effect modification.

Focusing again on the age effect, we find one standard error on the printout (1.95 for the coefficient of AGEBIN in women); the other may be computed with additional programming code and some effort. We can get, however, both standard errors by specifying the interaction model differently in SAS. The alternative code (shown below) also saves us the trouble of summing two coefficients to estimate the effect in men. What we'll find on the printout is precisely what we need: two sex-specific mean differences and their standard errors.

Alternative SAS code

```
PROC GLM;
 CLASS sex;
 MODEL sbp = sex agebin(sex)/SOLUTION;
 run;
```

Selected SAS printout

The GLM Procedure

Class Level Information

Class	Levels	Values
sex	2	0 1

Dependent Variable: sbp SYSTOLIC BLOOD PRESSURE (mmHg)

Source	DF	Sum of Squares
Model	3	45179.8130
Error	996	410428.5430
Corrected Total	999	455608.3560

Parameter		Estimate	Standard Error
Intercept		123.3600000	
sex	0	-5.4051505	
sex	1	0.0000000	
agebin(sex)	0	18.1812452	1.95358849
agebin(sex)	1	7.9093966	1.77476275

Notwithstanding some technical differences (a different intercept on the printout, which may be ignored, and a different kind of parameter estimates), this model is identical to the previous one. Both models display the same distributions of the sum of squares and both correspond to equation 10–3. What we don't find on the last printout is the coefficient of "SEX x AGEBIN", the estimated heterogeneity of effect, but we can easily compute it: 7.9 (the effect in men) – 18.2 (the effect in women) = –10.3 mmHg.

Finally, notice that there is no point in trying to express our results in the language of interaction and expected joined effects, because there is no natural reference for the "independent effect of aging" and perhaps not even for the "independent effect of sex group". As I mentioned before, the deterministic idea of interaction sounds appealing when smoking interacts with drinking, but not when age group interacts with sex group.

Since I dichotomized age only for pedagogical reasons, it's time to return to our original models. We may explore effect modification by sex group for any of the three dose-response functions of age and blood pressure: linear, quadratic and step function. In each case, all that we have to do is to create product terms: to "multiply" each age variable by SEX. The next three sections show the math and its application to the blood pressure data.

Sex group modifies the age effect: the linear function

For AGE, a continuous variable, the interaction model is almost identical. We just substitute AGE for AGEBIN.

$$\text{Mean SBP} = \beta_1 + \beta_2\ \text{AGE} + \beta_3\ \text{SEX} + \beta_4\ \text{AGE x SEX}$$

$$= \beta_1 + \beta_3\ \text{SEX} + (\boldsymbol{\beta_2 + \beta_4\ \text{SEX}})\ \text{AGE}$$

As the last expression shows, the mean difference per 1 year of aging is a function of sex group: ($\beta_2 + \beta_4$ SEX). Explicitly, it is β_2 in women and $\beta_2 + \beta_4$ in men.

SAS code

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

Selected SAS printout

```
                              The GLM Procedure

Dependent Variable: sbp   SYSTOLIC BLOOD PRESSURE (mmHg)

       Parameter         Estimate

       Intercept      64.22615646
       age             0.99183335 (β2)
       sex            30.66235703
       age*sex        -0.47363705 (β4)
```

Notice that the results are identical to those of stratified regression (the first printout in this chapter): 0.99 mmHg per 1 year of aging in women and 0.52 mmHg (0.99–0.47) in men. Again, alternative SAS code (below) will provide the sex-specific estimates directly, as well as the two standard errors.

SAS code

```
PROC GLM;
 CLASS sex;
 MODEL sbp = sex age(sex)/SOLUTION;
 run;
```

Selected SAS printout

The GLM Procedure

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

                                          Standard
     Parameter           Estimate             Error

     Intercept        94.88851349
     sex       0     -30.66235703
     sex       1       0.00000000
     age(sex)  0       0.99183335        0.08973957
     age(sex)  1       0.51819630        0.08252862
```

The printout shows the sex-specific estimates and the standard errors. No computation is needed. Ignore the other estimates.

Sex group modifies the age effect: the "step" function

When a step function connects age to systolic blood pressure, the interaction model should contain the set of dummy variables of age, the sex variable, and as many product terms as there are dummy variables (because each age variable should be multiplied by SEX).

$$\text{Mean SBP} = \beta_1 + \beta_2\,\text{AGE2} + \beta_3\,\text{AGE3} + \beta_4\,\text{AGE4} + \beta_5\,\text{SEX} + \beta_6\,\text{AGE2 x SEX} + \beta_7\,\text{AGE3 x SEX} + \beta_8\,\text{AGE4 x SEX}$$

$$= \beta_1 + \beta_5\,\text{SEX} + (\boldsymbol{\beta_2 + \beta_6\,\text{SEX}})\,\text{AGE2} + (\boldsymbol{\beta_3 + \beta_7\,\text{SEX}})\,\text{AGE3} + (\boldsymbol{\beta_4 + \beta_8\,\text{SEX}})\,\text{AGE4}$$

The expressions in bold print show the effect of each age variable on blood pressure, which varies by sex. Specifically:

- "$\beta_2 + \beta_6$ SEX" is the mean difference in systolic blood pressure between the second and first age groups. Explicitly, the mean difference is β_2 in women and $\beta_2 + \beta_6$ in men.
- "$\beta_3 + \beta_7$ SEX" is the mean difference in systolic blood pressure between the third and first age groups. Explicitly, the mean difference is β_3 in women and $\beta_3 + \beta_7$ in men.
- "$\beta_4 + \beta_8$ SEX" is the mean difference in systolic blood pressure between the fourth and first age groups. Explicitly the mean difference is β_4 in women and $\beta_4 + \beta_8$ in men.

SAS code

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

Selected SAS printout

```
                          The GLM Procedure

Dependent Variable: sbp   SYSTOLIC BLOOD PRESSURE (mmHg)

       Parameter         Estimate

       Intercept      113.2562500
       age2            10.1070594 (β2)
       age3            19.8755682 (β3)
       age4            28.4810381 (β4)
       sex              6.4693598
       age2*sex        -2.0900220 (β6)
       age3*sex       -10.4904184 (β7)
       age4*sex       -12.3282695 (β8)
```

Again, to get all of the sex-specific estimates and all of the standard errors directly, we could use the alternative code below:

SAS code

```
PROC GLM;
 CLASS sex;
 MODEL sbp = sex age2(sex) age3(sex) age4(sex) /SOLUTION;
 run;
```

Selected SAS printout

```
                                          Standard
     Parameter           Estimate             Error

     Intercept        119.7256098
     sex       0       -6.4693598
     sex       1        0.0000000
     age2(sex) 0       10.1070594        2.30756135
     age2(sex) 1        8.0170373        2.30809831
     age3(sex) 0       19.8755682        2.46496412
     age3(sex) 1        9.3851497        2.21851578
     age4(sex) 0       28.4810381        3.03124552
     age4(sex) 1       16.1527686        2.78699217
```

The printout shows six sex-specific estimates (bold print) and six standard errors. (Ignore again the intercept and the coefficient of SEX.)

Sex group modifies the age effect: the quadratic function

Since the quadratic model contains two age variables (AGE and AGE^2), both should be multiplied by SEX.

Mean SBP = $\beta_1 + \beta_2$ AGE + β_3 AGE^2 + β_4 SEX + β_5 AGE x SEX + β_6 AGE^2 x SEX

Here, however, we cannot compute the age effect by grouping terms (for reasons that require some knowledge of calculus.) The safest method would be to enter the values of the causal assignments of interest, compute two predicted means, and then subtract one predicted mean from the other to get the estimated mean difference. Since the quadratic function already contains the idea of effect modification by the exposure itself (chapter 9), the mean difference will vary by *both* age and sex.

SAS code

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

Selected SAS printout

```
                            The GLM Procedure

Dependent Variable: sbp   SYSTOLIC BLOOD PRESSURE (mmHg)

      Parameter           Estimate

      Intercept        64.20942108
      age               0.99238737 (β2)
      age*age          -0.00000446 (β3)
      sex               8.20730762 (β4)
      age*sex           0.26943363 (β5)
      age*age*sex      -0.00597460 (β6)
```

We will stop at this level of complexity, which is complicated enough, and not worry about the standard errors or about using an alternative code with a "class statement".

A single interaction model (versus stratification on the modifier)

You may wonder why not simply use stratified regression (sex-specific models) instead of complex models that contain interaction terms. Indeed, the results of stratified regression would have been identical in all of the examples above. The two methods, however, might generate different results when the model contains other covariates. If we had to condition on GENOTYPE, for example, while estimating the sex-specific mean difference for the age effect, the following two SAS codes might produce *different* estimates:

Stratified regression:

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

Interaction model:

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

So which method should you choose if you have to account for confounders while estimating heterogeneity of effects: stratum-specific models or a single model that contains interaction term(s)?

You will find both approaches in the literature. I prefer an interaction model to stratified regression (for a reason that has nothing to do with statistical testing of a null hypothesis about the coefficients of interaction terms.) When we search for modification of the age effect by sex while conditioning on genotype, the estimated effect of age in each sex group behaves like a weighted average across the strata of the genotype. If the genotype distribution in women differs from its distribution in men, the sex-specific estimates (of the age effect) from stratified regression will be based on different sets of weights. In contrast, sex-specific estimates from an interaction model will rely on the genotype distribution in the *entire* sample—on the same set of weights. (This paragraph may require a second reading.)

Obviously, stratified regression is not possible when the postulated modifier is a continuous variable. For example, if we reverse our interest in sex and age, wishing to estimate the effect of sex in the "strata" of age, we cannot stratify on age (unless we categorize the variable). Nonetheless, the various interaction models we have already fit also provide estimates of the sex effect for any specified age—because effect modification is a reciprocal property. To illustrate the computation, let's look again at the interaction model between AGE (the continuous variable) and SEX.

SAS code

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

Selected SAS printout

```
                              The GLM Procedure

Dependent Variable: sbp   SYSTOLIC BLOOD PRESSURE (mmHg)

       Parameter         Estimate

       Intercept      64.22615646
       age             0.99183335
       sex            30.66235703
       age*sex        -0.47363705
```

The regression equation is therefore:

Mean SBP = 64.2 + 0.99 x AGE + 30.7 x SEX – 0.47 x (AGE x SEX)
= 64.2 + 0.99 x AGE + **(30.7 – 0.47 x AGE)** SEX

The expression in bold print is the effect of sex, which is a function of age. For instance, at age 50, the estimated mean difference in systolic blood pressure ("men minus women") is 7.2 mmHg (30.7–0.47x50), whereas at age 60 it is only 2.5 mmHg (30.7–0.47x60).

You may similarly reorganize other interaction models to compute the effect of sex on blood pressure. In each case, you just have to place the variable SEX outside a parenthetical expression by which it will be multiplied. That multiplier in front of SEX will be a function of age, estimating the mean difference in blood pressure between men and women for any specified age.

The modified probability difference

At the end of chapter 9, we used linear regression to estimate probability differences of hypertension. The very same model (linear probability) may also be used to explore effect modification of the probability difference.

SAS code

```
PROC GLM;
 MODEL htn = age sex age*sex/SOLUTION;
```

Selected SAS printout

```
                              The GLM Procedure

Dependent Variable: htn  HYPERTENSION STATUS

       Parameter         Estimate

       Intercept     -.6251373904
       age           0.0137632442
       sex           0.3115557943
       age*sex       -.0049497992
```

The regression equation is

Pr (HTN=1) = –0.625 + 0.014 x AGE + 0.312 x SEX – 0.005 x (AGE x SEX)
= –0.625 + 0.014 x AGE + **(0.312 – 0.005 x AGE)** SEX

And the expression in bold print is the estimated probability difference between men and women—forced to be a function of age. For example, at age 50, the probability difference of hypertension is 0.062 (=0.312–0.005x50), or 6.2 percentage points higher for men, whereas at age 60 that difference is 1.2 percentage points. By solving the equation "0.312–0.005xAGE=0" we can even estimate the age at which sex equity is reached. According to this model the sex effect is zero at age 62.4. Beyond that age, the probability of hypertension is actually *lower* in men than in women and the absolute difference gradually increases.

If you wish to estimate the probability difference per 1 year of aging as a function of sex, regroup the terms, and place AGE (rather than SEX) outside a parenthetical expression:

Pr (HTN=1) = –0.625 + 0.014 x AGE + 0.312 x SEX – 0.005 x (AGE x SEX)
= –0.625 + 0.312 x SEX + **(0.014 – 0.005 x SEX)** AGE

Then, the probability difference per 1 year of aging is 0.014 in women (1.4 percentage points) and 0.009 in men (0.9 percentage point). The model indicates that the effect of aging on hypertension is larger in women than in men, similar to the effect of aging on mean systolic blood pressure (see the first printout in this chapter). Notice that we have been exploring effect modification on the additive scale, because our measure of effect was the probability difference, not the probability ratio.

A word about ANCOVA

ANCOVA (analysis of covariance) is an extension of ANOVA (analysis of variance), which was mentioned in the previous chapter. What's the difference? In ANCOVA, covariates are added to the model to deconfound or to explore effect modification, so the main feature of the output is something called "adjusted means": predicted means of the dependent variable at some values of the covariates (usually at their means). But just like ANOVA, the story is no more than linear regression of the dependent variable on dummy variables (and covariates).

You don't really need ANCOVA models to estimate the conditional mean difference or the modified mean difference. All of these models can be expressed in the language of linear regression.