13. Multicausality – analysis approaches

Concepts and methods for analyzing epidemiologic data involving more than two variables; control of confounding through stratified analysis and mathematical modeling.

Multivariable analysis

In the preceding two chapters we delved into the kinds of issues and situations that arise in a multivariable context. We introduced the additive and multiplicative models for joint effects of multiple exposure variables and employed stratified analysis to examine the effects of one variable whle controlling for the values of others. In this chapter we consider analytic approaches for examining the relationships between an outcome and multiple explanatory variables. The latter may consist of a study factor and potential confounders, a study factor and potential modifiers, or several exposures all of which are of interest.

Confounding:

To restate briefly, confounding is a situation where a factor or combination of factors other than the study factor is responsible for at least part of the association we observe for the association between the study factor and the outcome. If we do not control for confounding, then we may misattribute an effect to the study factor when the association really reflects the effect of another variable. In a situation of confounding, the crude data give us the wrong picture of the relationship between the exposure and outcome. Other factors may be exaggerating the strength of the relationship or obscuring some or all of it. To see the correct picture, we need to take into account the effects of other factors.

In a law enforcement analogy, the exposure is the suspect in a bank robbery and the other factors are known offenders with whom he associates. We need to establish the suspect's guilt. The suspect may be completely innocent, may have had some role in the crime, or may have had a greater role than at first appears. In order to determine the suspect's guilt, we need to examine the total picture of the actions of all of the individuals. In this analogy, confounding would occur if we charge the suspect with a crime he did not commit or with a role in the crime greater or smaller than accords with his actions. For example, it would be confounding to charge the suspect with bank robbery if he was just passing by and one of the robbers called him in. Confounding would also occur if we charged the suspect as an accomplice when in fact he was the principal organizer of the robbery.

The most common method of deciding whether or not confounding exists is to compare the crude (uncontrolled) results with the controlled results. If these two sets of results are meaningfully different, if they send a different "message" or suggest a different conclusion about the association under study, then confounding is present; the crude results are "confounded". The conclusion about the presence of confounding, however, is secondary to our main purpose, which is to obtain a valid estimate of the existence and strength of association between the exposure of interest and the disease outcome. When

we determine that confounding is present, we either present the stratum-specific findings or compute an adjusted measure of association (e.g., a standardized rate ratio) that controls for the effects for the confounding variables.

Effect modification

Effect modification is a situation where neither the crude and nor the adjusted measure provides an adequate picture of the relationship under study. The picture is **not wrong**, but it could nevertheless mislead. We may not have fulfilled our responsibility to present the full picture. Effect modification means that there are important differences between groups (or at different levels of some modifying variable) in the relationship between the exposure and the disease on our scale of measurement. Where effect modification is present, then the relationship between exposure and disease is not susceptible to being stated in such a simple formulation as "D and E are associated with a relative risk of about 2". Rather, an answer to the question "what is the relative risk for D given E?", must be "It depends." For example, any discussion of the heart disease risks for women talomg oral contraceptives would be seriously incomplete if it did not explain that the situation is quite different for women who smoke cigarettes or not, especially at ages above 35 years.

Where effect modification is present, the summary measure is an average of disparate components, so that the summary is too uninformative by itself. If Carlos is 90 cm tall, Shizue is 120 cm tall, and Rhonda is 150 cm tall, it may be useful to know that their average height is 120 cm, but probably not a good idea to buy three medium size (120 cm) school uniforms.

Analytic approaches

There are two primary approaches to analyzing data involving more than two variables: stratified analysis and modeling. We have already encountered both. In stratified analysis we divide the observations into one group for each level or combination of levels of the control variables. We analyze the association between the study factor and outcome separately within each group. In this way we may be able to observe the association involving the study factor without interference from the stratification variables.

Comparison of the crude measure of association to the stratum-specific measures or their weighted average will disclose whether the crude measure of association is confounded. Examination of the data within the individual strata will reveal if the measure of association varies so greatly that a summary measure by itself may mislead. For a fuller exploration we can stratify by each of the covariables and by various combinations of them. Stratified analysis gives us a full picture that we can examine in detail.

At some point, however, detail becomes an obstacle instead of an advantage. Modeling is a strategy for submerging the detail and focusing on relationships. Viewing the data through the framework of the model we gain analytic power and convenience. Rather than confuse ourselves and our audience by presenting a plethora of tables, we employ the elegant simiplicity of the model and its parameters, through which we can estimate the measures of association we seek. If we have chosen our model well and evaluated its suitability, we can obtain an optimal analysis of the data. But just as a pilot can fly

very far on instruments but needs to see the runway when landing, a modeling analysis should be supplemented with some stratified analyses. On the other hand, in a stratified analysis, computation of summaries across strata generally involves at least an implicit model framework.

Whichever approaches we use, there's no escaping the fact that how we proceed and how we interpret the results we observe depend on our conceptual model of the relationships among outcome, exposure, and stratification variables. If nothing is known about the factors under study, we may have to proceed in a completely empirical manner. But if there is some knowledge, it will serve as a guide. For example, suppose we see an association involving our study factor and outcome, but when we control for another factor the association disappears. Whether we conclude "confounding" and dismiss the crude association as an artifact or not depends upon whether or not we think of the stratification variable as a "real" cause of the outcome rather than the study factor. If the stratification variable is an intermediate factor in the causal pathway between the study factor and the outcome, then the situation is not one of confounding even though it can be numerically identical.

Stratified analysis — interpretation

Stratified analysis is conceptually simple. It involves disaggregating a dataset into subgroups defined by one or more factors that we want to control. For example, in studying the effect of reserpine use on breast cancer risk, we could stratify by obesity. Analyses within each strata can then be regarded as unconfounded by that risk factor, to the degree that the strata are sufficiently narrow. (If the strata are broad, e.g., "body mass index of 2.2 through 3.2" or "blood pressure greater than 95 mmHg", we may have "residual confounding" due to heterogeneity of the stratification variable within one or more strata.)

We have already encountered stratified analyses, notably in the chapters on confounding and effect modification. In this chapter we will gain a more indepth understanding of stratified analysis and how it relates to other concepts we have learned. We will also see when and how to obtain an overall summary measure that takes account of the stratification.

Example

Suppose that four case-control studies have investigated a possible association between reserpine and breast cancer (a question that arose in the 1970s) and that each controlled for obesity by dividing the data into two strata. The table below shows the crude and stratum-specific odds ratios from these four (hypothetical) studies. How would we describe the results of each study?

Association between reserpine and breast cancer controlling for body weight (odds ratios)
Hypothetical data

Study	Obese	Nonobese	Summary (adjusted)	Total (crude)
A	2.0	2.2	2.1	4.0
B	4.0	2.2	3.1	3.0
C	2.0	2.2	2.1	2.0
D	4.0	2.2	3.1	1.5

In study A, we see that the OR within each body weight category is about 2.0, whereas the crude OR is 4.0. Study A, therefore, illustrates a situation of **confounding**: the crude measure of association lies outside the range of the stratum-specific measures. The crude OR is meaningfully different than the adjusted OR and no other method of adjustment would change that, since any weighted average of the stratum-specific OR's would have to lie between 2.0 and 2.2.

In studies B and C, on the other hand, the crude OR could equal (or nearly equal) a weighted average of the stratum-specific measures (as is in fact the case for the adjusted OR's shown), because it (nearly) lies within the range of those measures. Therefore, confounding is not a feature of the data in either of these studies. In study B, if the numbers of participants in each stratum are large enough for us to regard the difference between the stratum-specific OR's as meaningful (not simply due to "noise"), then the difference indicates effect modification of the OR. It was important for the study to report the stratum-specific OR's and not rely completely on the crude or adjusted measures.

If the strata were large enough and the OR's were regarded as reasonably free of bias, we might wonder whether in some way obesity could potentiate the effect of reserpine (at least on the odds ratio scale). If the relationship is judged to be causal and these OR's the best estimates of the strength of relationship, then the stronger OR for obese patients suggests that they especially should avoid taking reserpine if they cannot lose weight (the usual criterion for "public health interaction" and "individual risk management interaction" are departure from the additive model of expected joint effect. However, if the observed association is "supra-multiplicative" [stronger than that expected from multiplicative model], it will also be "supra-additive" [stronger than expected form an additive model]). In study C, on the other hand, the slight difference between the two strata, even if not attributable to random variation, is insufficient to warrant attention. Any weighted average of the two stratum-specific measures would be a satisfactory summary.

Study D illustrates both confounding and effect modification, since the crude OR lies outside the range of the stratum-specific ORs and therefore could not equal any weighted average of the two. At the same time, the stratum-specific ORs appear to be importantly different (assuming adequate stratum sizes). It would not be sufficient to provide only a summary measure (on the OR scale).

Summarizing the relationships

Often we are interested in obtaining an overall assessment of the role of the study factor, controlling for other risk factors. The usefulness of an overall measure of association will obviously differ in these four studies. In studies A and C, a single overall measure could adequately summarize the OR's in the two strata so that it would not be essential to present them as well. In studies B and D, however, we clearly need to present the stratum-specific OR's, though for some purposes a summary measure may also be useful.

The most convenient overall estimate, if it is not confounded, is the measure based on the aggregate data, the crude estimate. The stratified analysis in study C above indicates no confounding by obesity. If that is the only variable we need to control for, then we can use the crude OR to summarize the relationship.

In both study A and study D, however, confounding is present. Relying on the crude OR as the summary of the stratified results will clearly mislead. Therefore, we require a summary measure that "adjusts for" obesity. The summary measure we derive is a weighted average of the stratum-specific measures. The summary measures we encountered in the chapter on standardization (the SMR and the SRR) are examples of such summary measures.

Relationship between stratified analysis and models for joint effects

The additive and multiplicative models introduced earlier express the joint incidence or effect of two (or more) factors in terms of the separate incidence or effect of each. The multiplicative model, for example, expresses the joint RR as:

$$RR_{11} = RR_{10} \times RR_{01}$$

and the joint risk (or rate) as:

$$R_{11} = \frac{R_{10} \times R_{01}}{R_{00}}$$

where the first and second subscripts indicate presence (1) or absence (0) of the first and second factors, respectively. It turns out that if the data fit this model, then in a stratified analysis controlling for either factor the stratum-specific RR's for the other factor will be equal to each other.

To see this, simply divide both sides of the second form of the model by R_{01}:

$$\frac{R_{11}}{R_{01}} = \frac{R_{10} \times R_{01}}{R_{00} \times R_{01}} = \frac{R_{10}}{R_{00}}$$

Let's examine the term on the left and the term on the right. In both of these terms, the first factor is present in the numerator rate but absent from the denomator rate. Thus, each of these terms is a rate ratio for the effect of the first factor.

$$\begin{array}{c} \text{RR for 1st factor} \\ \text{(2nd factor present)} \end{array} = \begin{array}{c} \text{RR for 1st factor} \\ \text{(2nd factor absent)} \end{array}$$

Meanwhile, the second factor is present in both numerator and denominator rates on the left, and absent from both rates on the right. Since each rate requires a number of cases and a person or person-time denominator, then each RR must come from a 2 x 2 table containing exposed cases, unexposed cases, exposed noncases or person-time, and unexposed noncases or person-time.

Thus, these two RR's correspond to a stratified analysis that controls for the second factor as present vs. absent. Their equality means that the RR for the outcome with respect to the first factor is the same in both strata of the second factor. Had we originally divided by RR_{01}, instead of RR_{10}, we would have found that the RR for the second factor is the same in both strata of the first factor.

To see the relationship with some familiar numbers, here is a portion of the Mann et al. data presented earlier:

Incidence of myocardial infarction (MI) in oral contraceptive (OC) users age 40-44 years, per 100,000 women-years

Cigarettes/day	OC*	$\overline{OC}$*	RR**	AR***
0-14	47 (R_{01})	12 (R_{00})	4	35
15 +	246 (R_{11})	61 (R_{10})	4	185

* Rate per 100,000 women-years

** RR=relative risk (rate ratio)

*** AR=attributable risk (rate difference, absolute difference)

We saw in the chapter on effect modification that the full table conformed quite closely to a multiplicative model. If we look back at the table we see that the RR's for the first two rows (3) were the same and those for the second two rows (4, shown above) were the same.

Suppose we let the four rates in the table be represented by R_{00}, R_{10}, R_{01},and R_{11}, with the first subscript denoting smoking and the second denoting OC. Then we can write:

$$R_{11} = \frac{R_{10} \times R_{01}}{R_{00}}$$

and

$$246 \approx \frac{61 \times 47}{12}$$

The above equality is only approximate, but then the rate ratios weren't exactly the same (3.92 versus 4.03). Therefore, the statement that the RR is the same in all strata is equivalent to saying that the data conform to a multiplicative model.

We could equally well have demonstrated this fact by using the OR (try it!). Had we instead used the rate or risk difference as the parameter of interest, we would find (by subtraction, rather than division) that equality of the stratum-specific difference measures is equivalent to having the data conform to an additive model (try this, too!).

$$R_{11} = R_{10} + R_{01} - R_{00}$$

$$R_{11} - R_{01} = R_{10} + R_{01} - R_{00} - R_{01} = R_{10} - R_{00}$$

This relationship between the multiplicative and additive models on the one hand and stratified analysis on the other is fundamentally trivial, but also fundamental, so it is worth a little more time.

Stratified analysis as "tables" or "columns"

A stratified analysis involving a dichotomous outcome, a dichotomous exposure, and a dichotomous stratification variable involves two 2 x 2 tables, each with two columns of cases and noncases (or person-time). If we look at the data as columns, rather than as tables, we can almost "see" the multiplicative or additive model structure in the stratification. For example, here are two 2 x 2 tables created with hypothetical numbers that produce rates similar to those in the Mann et al. data above and presented in the form of our earlier stratified analyses.

Hypothetical data on incidence of myocardial infarction (MI) in oral contraceptive (OC) users per 100,000 women-years, controlling for smoking (after Mann et al.)

Cigarettes /day	15+	15+	0-14	0-14
OC use	OC	$\overline{\text{OC}}$	OC	$\overline{\text{OC}}$
CHD	49	11	19	8
Women-years*	20	18	40	66
Rate**	245	61	48	12
	R_{11}	R_{10}	R_{01}	R_{00}

* (in thousands)

** per 100,000 (some differ slightly from Mann et al.'s)

The lefthand 2 × 2 table shows the relationship between OC and CHD among women who smoke 15+ cigarettes/day; the righthand table shows the relationship among women who smoke less than 15 cigarettes/day. **Equivalently**, the four columns show the number of cases, women-years of risk, and CHD rate in, from left to right:

15+ cigarette/day OC users	(R_{11}, = 49/20,000 = 245/100,000wy)
15+ cigarette/day OC nonusers	(R_{10}, = 11/18,000 = 61/100,000wy)
0-14 cigarette/day OC users	(R_{01}, = 19/40,000 = 48/100,000wy)
0-14 cigarette/day OC nonusers	(R_{00}, = 8/66,000 = 12/100,000wy)

Similarly, all of the relevant RR estimates can be obtained by forming ratios of the appropriate rates, e.g.:

Rate ratios

Both factors (versus neither)	$RR_{11} = R_{11} / R_{00} = 245/12 = 20$
Smoking (1st factor) acting alone	$RR_{10} = R_{10} / R_{00} = 61/12 = 5$
Smoking (1st factor) in presence of OC (2nd factor)	$RR_{S\|O} = R_{11} / R_{01} = 245/48 = 5$
OC (2nd factor) acting alone	$RR_{01} = R_{01} / R_{00} = 48/12 = 4$
OC (2nd factor) in presence of smoking (1st factor)	$RR_{O\|S} = R_{11} / R_{10} = 245/61 = 4$

So the multiplicative model for joint effects, introduced in the chapter on effect modification, is equivalent to stratified analyses in which the ratio measure is the same in all strata. The same can be shown for the additive model and the difference measure, though not with these data since they do not fit an additive model.

"Homogeneity" and "heterogeneity" vs. "synergy" or "antagonism"

In the terminology used when discussing summary measures of association, stratum-specific measures are said to be "homogeneous" when they are the same and "heterogeneous" when they are meaningfully different. Obviously, a summary measure works best in a situation where the measure being summarized is homogenous across strata. In the usual case, for a ratio measure of effect, homogeneity across strata is equivalent to rates, odds, or ratios that conform to a multiplicative model of joint effects. In the case of difference (absolute) measures, homogeneity is equivalent to an additive model of joint effects. "Effect modification" (or "effect measure modification", in Greenland and Rothman's new terminology) signifies heterogeneity for that measure.

Typically, epidemiologic analyses of risk factors employ ratio measures of effect. On the ratio scale, summary measures from stratified analysis (and as we will soon see, from mathematical models) are derived on the premise of homogeneity of effects across strata, equivalent to a multiplicative model of expected joint effects, and also generally inconsistent with an additive model. So the term "effect modification" is most commonly applied to situations where the ratio measure of effect is heterogeneous across strata – even if it should happen (admittedly as the exception) that the data do conform to an additive model! In contrast, "synergism" from a public health perspective is now generally regarded as an observed effect greater than expected from an **additive** model. So when there is "effect modification of the relative risk" there is generally "interaction from a public health perspective".

Such inconsistency is undoubtedly an indication that these concepts were designed by mortals, rather than by a higher power, and also underlines the point that "effect modification" is relative to the scale of measurement or expected model for joint effects. We can hope that as the discipline evolves, a new synthesis will develop that will avoid this "schizophrenic" approach. In the meantime, perhaps the following summary table will help.

Homogeneity, heterogeneity, and effect modification in relation to additive and multiplicative models

	Public health impact perspective	Summary measure perspective
1. Data conform to an additive model (**homogeneity of the difference measure** across strata)	**No interaction** (no synergism)	**No effect modification (of difference measure)**, summary difference measure is adequate **Effect modification (of ratio measure)**, summary ratio measure is not adequate
2. Joint effect exceeds expectation under an **additive** model ("supra-additive" – may or may not equal or exceed multiplicative model)	**Public health interaction (synergistic effect)**	**Effect modification** (of difference measure, perhaps also ratio measure), summary difference measure is not adequate (perhaps also summary ratio measure)
3. Data conform to expectation under a multiplicative model (**homogeneity of ratio measure** across strata)	**Public health interaction (synergistic effect)**	**No effect modification (of ratio measure)**, summary ratio measure is adequate
4. Joint effect exceeds expectation under a **multiplicative** model ("supra-multiplicative")	**Public health interaction (synergistic effect)**	**Effect modification (of difference and ratio measures)**, summary difference and ratio measures are not adequate

Types of overall summary measures

When the crude and stratum-specific measures are all similar, then the crude measure serves as a fully satisfactory summary measure. When there is meaningful heterogeneity, then we will need to present the stratum specific measures themselves. There remains the situation where the stratum-specific measures are sufficiently homogenous that a summary measure of some kind is of interest but, due to confounding, the crude measure cannot serve this roll. In such cases the crude measure is outside the range of the stratum-specific measures or so far from the middle of the range that it would be a

misleading summary. These circumstances call for an adjusted measure, generally some form of weighted average of the stratum-specific measures.

Suppose that all of the stratum-specific measures are close together (i.e., homogeneous), so that we are inclined to regard all of them as estimates of the same population parameter (the "true" measure of association) plus or minus some distortion from sampling variability (if we want to quantify the compatibility of the data with this supposition, we can employ a statistical test, such as the Breslow-Day homogeneity chi-square, to assess the expected range of chance variability). If there is a "true" underlying value, how can we best estimate it? Obviously some sort of weighted average is called for, but what kind?

If there is only one "true" measure of association and each of the strata provides an estimate of that true measure, then we will want to pay more attention to strata that provide "better" (i.e., more precise) estimates. So the averaging procedure we employ should give more weight to the estimates from such strata. We can meet this objective by using as weights the estimated precision of each stratum-specific estimate. Such a weighted average provides the best estimate of the "true" measure of association, under the assumptions on which we have been proceeding. (Rothman refers to summary estimates derived in this way as "directly pooled" estimates. However, the term "pooled" is sometimes used to refer to the crude total over a set of strata or studies.)

[Note: the calculation of summary measures of association as explained below is NOT a required part of EPID 168. The only things from this discussion of summary measures that EPID 168 students are expected to know concern: (1) summary measures are typically weighted averages; (2) if the crude measure of association falls comfortably within the range of the stratum-specific measures, then it is not confounded and may serve as a summary measure; (3) if the crude measure is outside the range of the stratum-specific measures, then confounding is present and the crude measure is <u>not</u> an adjusted measure of association must be used to summarize the relationship; (4) if stratum-specific measures are meaningfully different from each other, then any summary measure (crude or adjusted) provides an incomplete picture of the relationship, so the investigator should report the stratum-specific results and take that heterogeneity into account in interpreting a summary measure. The following discussion is provided for the more advanced or adventurous. Others may wish to come back to this section during or after their next course in epidemiologic methods.]

Precision-based weighted summary measure estimates – *optional topic*

The **im**precision of an estimate can be defined as the width of the confidence interval around it. Since we are used to estimating 95% confidence intervals by adding and subtracting 1.96 times the standard error of the estimate, the total width is 2 x 1.96 × standard error. Since all of these width's will include the 2 x 1.96, all of the variability in precision is contained in the standard errors. The **smaller** the standard error, the greater the degree of precision, so weights consisting of the receiprocals of the standard errors will accomplish precision-weighting. In fact, the weights used are the squares of these reciprocals and are called "inverse variance weights".

Difference measure – the CID

The variance of the CID is an easy one to derive, since the CID is simply a difference of two proportions. When there are at least 5 "successes", the variance of a proportion (p) can be estimated simply as p(1-p)/n, where n is the size of the sample. The variance of a sum or difference of two independent random variables is the sum of their variances. So the variance (square of the standard error) of the CID is:

$$\text{var(CID)} = \text{var}(CI_1) + \text{var}(CI_0)$$

$$[\text{s.e.(CID)}]^2 = \frac{CI_1\,(1\text{-}CI_1)}{n_1} + \frac{CI_0\,(1\text{-}CI_0)}{n_0}$$

Using the notation from our 2 × 2 tables, where "a" represents exposed cases and "b" represents unexposed cases, we can write this formula as:

$$[\text{s.e.(CID)}]^2 = \frac{a/n_1\,(c/n_1)}{n_1} + \frac{b/n_0\,(d/n_0)}{n_0}$$

$$[\text{s.e.(CID)}]^2 = \frac{ac}{n_1^3} + \frac{bd}{n_0^3} = \frac{n_0^3 ac + n_1^3 bd}{n_1^3\, n_0^3}$$

whose reciprocal (and the stratum-specific weight) is:

$$w = \frac{1}{[\text{s.e.(CID)}]^2} = \frac{n_1^3 + n_0^3}{n_0^3 ac + n_1^3 bd}$$

This value is computed for each stratum and used as the weight for the CID for that stratum. For two strata (indicated by subscripts 1 and 2):

$$\text{Summary CID} = \frac{w_1\,CID_1 + w_2\,CID_2}{w_1 + w_2}$$

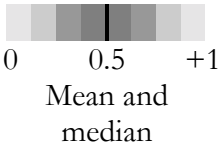
Since we have just derived the variances of the stratum-specific CID estimates and since the variance of the summary CID estimate is simply their sum, the variance of this summary CID estimate is simply $1/w_1 + 1/w_2$, and a 95% confidence interval for the summary CID estimate is:

$$\text{95\% CI for (summary) CID} = \text{CID} \pm 1.96 \sqrt{1/w_1 + 1/w_2}$$

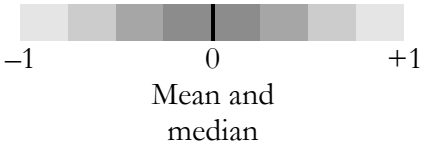
Ratio measures

A uniform random variable that is a proportion has a symmetric distribution, since its possible values lie between 0 and 1, and the mean of the distribution (0.5) is the same as its median. Similarly, the distribution of the CID, based on the difference in two uniform random proportions, is symmetric, since it lies between -1 and 1 and has its mean and median at its null value, 0.

Distribution of a proportion:



Distribution of a difference of two proportions:

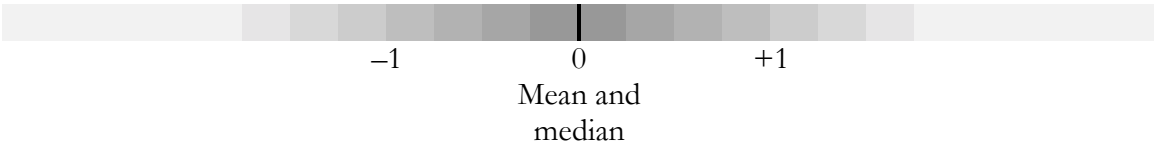


Because of this symmetry, variance estimates based on an approximate normal distribution could be used. Ratio measures, however, do not have symmetric distributions. The CIR (a ratio of two proportions) and the OR (a ratio of odds, which are in turn ratios of two non-independent proportions) both have a lower limit of 0, a median (and null value) at 1.0, and no upper limit.

Distribution of CIR, IDR, OR

This asymmetry makes the use of a normal approximation more problematic. However, the logarithm of a ratio measure **does** have a symmetric distribution, so that the normal approximation can be used.

Distribution of ln(CIR), ln(IDR), ln(OR):



Therefore, variances for the CIR, IDR, and OR are estimated using a logarithmic transformation.

Ratio measures – CIR:

The natural logarithm of the CIR is:

$$\ln(CIR) = \ln\left[\frac{CI_1}{CI_0}\right] = \ln(CI_1) - \ln(CI_0)$$

If each stratum-specific CI is an independent random proportion, then the variance of the logarithm of the estimate of the stratum-specific CIR is the sum of the variances of the logarithms of the estimates of the statum-specific CI's.

$$\mathrm{Var}(\ln(CIR)) = \mathrm{Var}(\ln(CI_1)) + \mathrm{Var}(\ln(CI_0))$$

The variance of these logarithms is obtained using a Taylor's series approximation as (Kleinbaum, Kupper, and Morgenstern; Rothman and Greenland):

$$\text{Var}(\ln(\text{CIR})) \approx \frac{c}{an_1} + \frac{d}{bn_0} = \frac{bcn_0 + adn_1}{abn_1n_0}$$

so that the stratum-specific weights are:

$$w = \frac{1}{\text{Var}(\ln(\text{CIR}))} = \frac{abn_1n_0}{adn_1 + bcn_0}$$

For two strata, then, the precision-weighted summary **ln(CIR)** is:

$$\text{Summary } (\ln(\text{CIR})) = \frac{w_1 \ln(\text{CIR}_1) + w_2 \ln(\text{CIR}_2)}{w_1 + w_2}$$

In order to obtain the summary estimate for the CIR, the summary ln(CIR) must now be converted to the natural scale by exponentiation:

$$\text{Summary CIR} = \exp(\text{summary } \ln(\text{CIR}))$$

Again, we can use the w_i to obtain the variance of the overall CIR estimate, though again a transformation of scale will be needed. The variance of the summary ln(CIR) estimate is simply $1/w_1 + 1/w_2$, so the 95% confidence interval is:

$$95\% \text{ confidence interval for } \ln(\text{CIR}) = \ln(\text{CIR}) \pm 1.96\sqrt{1/w_1 + 1/w_2}$$

$$95\% \text{ confidence interval for CIR} = \exp[\ln(\text{CIR}) \pm 1.96\sqrt{1/w_1 + 1/w_2}]$$

Ratio measures – OR:

An approximate variance estimate for the ln(OR) in the ith stratum is:

$$\text{Var}(\ln(\text{OR})) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

so that the weight for the ith stratum is:

$$w_i = \frac{1}{\left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)}$$

(Notice that a small number in any cell makes the variance large and, therefore, the weight small.) The overall ln(OR) is then estimated as:

$$\ln(OR) = \frac{w_1 OR_1 + w_2 OR_2}{w_1 + w_2}$$

and the overall OR as:

$$OR = \exp(\ln(OR))$$

The variance of the ln(OR) is $1/\Sigma w_i$ and can be used to obtain a 95% confidence interval for the ln(OR), which can then be exponentiated to obtain a confidence interval for the OR, as for the CIR.

Mantel-Haenszel summary measures:

Nathan Mantel and William Haenszel, in their classic 1959 paper, introduced a summary OR that is particularly easy to calculate:

$$OR_{MH} = \frac{\Sigma[a_i d_i / n_i]}{\Sigma[b_i c_i / n_i]}$$

Rothman shows that the OR_{MH} is a weighted average, with stratum-specific weights of $b_i c_i / n_i$. These weights are also precision-based, since they are inversely proportional to the variance of the logarithm of the stratum-specific OR's. The difference between these weights and the ones in the previous formula is that for the OR_{MH} the weights are based on variances that apply on the assumption that the OR's are 1.0, whereas the previous weights did not require that assumption. However, the two summary measures produce similar results and are essentially equivalent when the stratum-specific OR's are not far from 1.0. An advantage of the OR_{MH} is that it can be used with sparse data including an "occasional" zero cell (see Rothman).

Formulas for the these and other summary measures of association (IDD, IDR), confidence intervals, and overall tests of statistical significance can be found in the textbooks by

Kleinbaum, Kupper, and Morgenstern; Hennekins and Buring; Schlesselman; and Rothman. The Rothman text includes discussion of maximum likelihood methods of estimating summary measures.

Although the discussion here has emphasized the usefulness of summary measures in analyses where there is little heterogeneity across strata, at times an investigator may wish to present a summary measure even when substantial heterogeneity is present. Standardized (rather than adjusted) measures are used in these situations (see Rothman and/or Kleinbaum, Kupper, and Morgenstern).

[Note: Time to tune back in if you skipped through the section on weighting schemes for summary measures of association. On the other hand, if you are already familiar with mathematical models you may wish to skim or skip this section.]

Matched designs

As we saw in the chapter on confounding, when the study design uses matching, it may be necessary to control for the matching variables in the analysis. In a follow-up study, analyzing the data without taking account of matching may not yield the most precise estimates, but the estimates will not be biased. A case-control study with matched controls, however, can yield biased estimates if the matching is not allowed for in the analysis. Thus, the matching variables should always be controlled in analyzing matched case-control data. If the result is no different from that in the unmatched analysis, then the unmatched analysis can be used, for simplicity.

The most straightforward way to control for matching variables is through stratified analysis, as presented above. If matching was by category (i.e., frequency matching, e.g., by sex and age group) was employed, then the analysis procedure is a stratified analysis controlling for those variables. If individual matching (e.g., pair matching, matched triples, etc.) was employed, then each pair or "n-tuple" is treated as a strata.

Suppose that the data from a case-control study using pair matching are as shown in the following table.

Pair	Case	Control	Type
6	n	n	A
9	n	n	A
10	n	n	A
1	Y	n	B
2	Y	n	B

5	Y	n	B
3	n	Y	C
8	n	Y	C
4	Y	Y	D
7	Y	Y	D

If each pair is a stratum, then the stratified analysis of the above data consists 10 tables, each with one case and one control. There will be 3 tables like table A, 3 like table B, 2 like table C, and 2 like table D.

	Exp	Unexp	Exp	Unexp	Exp	Unexp	Exp	Unexp
Case	0	1	1	0	0	1	1	0
Control	0	1	0	10	1	0	1	0
Type	A		B		C		D	

Although we cannot compute any stratum-specific measures of association, we can compute a Mantel-Haenszel summary odds ratio using the formula:

$$OR_{MH} = \frac{\Sigma[a_i d_i / n_i]}{\Sigma[b_i c_i / n_i]}$$

where a_i, b_i, c_i, d_i are the cells in table i, and n_i is the number of participants in table i. This general formula becomes much simpler for pair-matched data, because all of the n_i are 2 and many of the terms disappear due to zero cells. When we remove these terms and multiply numerator and denominator by 2 (n_i), we are left with (a) a one ($a_i d_i$) in the numerator for each table where the control is exposed and the case is not (table type B); and (b) a one ($b_i c_i$) in the denominator for each table where the case is exposed and the control is not (table type C). For the above data:

$$OR_{MH} = \frac{1 + 1 + 1}{1 + 1} = \frac{3}{2} = 1.5$$

So the formula becomes simply OR=B/C, where B is the number of discordant pairs in which the case is exposed and C is the number of pairs in which the control is exposed. Note that the concordant pairs (types A and D) have no effect on the OR.

Mathematical models

Earlier in this chapter we showed that when the RR is the same in all strata of a stratified analysis, then data conform to a multiplicative model, and vice-versa. We also stated that for difference measures, equality of the stratum-specific difference measures is equivalent to having the data conform to an additive model. In fact, these simple models can serve as a jumping off point for understanding mathematical models used to control confounding.

Returning to the topic of breast cancer in relation to obesity and/or reserpine use, suppose that the following table shows data from a cohort study. (Note that this is hypothetical - reserpine was at one time suspected of being related to breast cancer risk, but that evidence has since been discounted.)

Ten-year risk of breast cancer, by obesity and use of reserpine (hypothetical data)

Risk factors	Numeric (illustrative)	Algebraic
None (background risk)	.01	R_{00}
Obesity only	.03	R_{10}
Reserpine only	.02	R_{01}
Both reserpine and obesity	.04	R_{11}

Thus:

R_{00} indicates background risk (no reserpine, non-obese)

R_{10} indicates risk for obesity (without reserpine)

R_{01} indicates risk for reserpine (without obesity)

R_{11} indicates risk both reserpine and obesity

In this example, the joint risk conforms to an additive model:

$RD_{11} = RD_{10} + RD_{01}$ (Risk differences are additive)

$$R_{11} - R_{00} = (R_{10} - R_{00}) + (R_{01} - R_{00})$$

$$(.04 - .01) = (.03 - .01) + (.02 - .01)$$

$$0.03 = 0.02 + 0.01$$

or, equivalently:

$$R_{11} = R_{10} + R_{01} - R_{00}$$

$$0.04 = 0.03 + 0.02 - 0.01$$

We can also express the various risks in terms of the baseline risk and the "effect" of the risk factors:

$R_{10} = R_{00} + RD_{10}$ (.03 = .01 + .02) (Obesity "effect")

$R_{01} = R_{00} + RD_{01}$ (.02 = .01 + .01) (Reserpine "effect")

$R_{11} = R_{00} + RD_{01} + RD_{10}$ (.04 = .01 + .02 + .01) (Both)

Note that the word "effect" is used here by convention and for convenience, rather than to suggest causality.

Another way we might think about these various risk equations is to try to put them all into a single equation with "switches" for which effects are "turned on". The baseline risk R_{00} is always present, so we require only two "switches", one for the obesity effect and one for the reserpine effect:

Risk	=	R_{00}	+	Obesity effect		Obesity "switch"	+	Reserpine effect		Reserpine "switch"
Risk	=	R_{00}	+	RD_{10}	×		+	0.01	×	
Risk	=	0.01	+	0.02	×		+	0.01	×	

When a "switch" is on (=1) then the 0.02 (obesity effect) or 0.01 (reserpine effect) comes into play, making the Risk from the model larger.

Risk	=	R_{00}	+	Obesity effect		Obesity "switch"		Reserpine effect		Reserpine "switch"	
Risk	=	0.01	+	0.02	×	0	+	0.01	×	0	= 0.1
Risk	=	0.01	+	0.02	×	1	+	0.01	×	0	= 0.03
Risk	=	0.01	+	0.02	×	0	+	0.01	×	1	= 0.02
Risk	=	0.01	+	0.02	×	1	+	0.01	×	1	= 0.04

We now have a "model" that we can use to compute the risk for any combination of the two risk factors. Although this example is trivial, as well as contrived, the model structure is the same as in multiple linear regression. To see our model in a more sophisticated form, we have merely to replace the "switches" by indicator variables that can take the value of 0 or 1.

Linear models:

If we let:

B = 1 if the woman is obese and 0 if she is not

E = 1 if the woman uses reserpine and 0 if she does not

then our model becomes:

$$R(B,E) = R_{00} + (RD_{10})B + (RD_{01})E$$

Substituting values from the table:

$$R(B,E) = .01 + (0.02)B + (0.01)E$$

Our two dichotomous variables (B=1 or 0, E=1 or 0) yield four possible combinations of reserpine use and obesity, just as did our switches model. We now have a professional-looking linear model for breast cancer risk in terms of baseline risk, presence or absence of each of two dichotomous risk factors, and the risk difference (or increase in risk) attributable to each factor. The risk differences (0.02, 0.01) are called "coefficients" and are often represented by the Greek letter β; the baseline risk is often represented by the Greek letter α.

You may well wonder what is the value of the above machinations, since we have no more information from our model than we had in our table of risks (i.e., in our stratified analysis). The accomplishment lies in the ability to estimate risk differences for each factor, controlling for the other(s), by estimating the coefficients in the model. The power of modeling is the ability to use the study data to estimate model coefficients by using a statistical technique known as regression analysis. The estimated coefficients yield epidemiologic measures that are adjusted for the effects of the other variables in the model.

We can make our model more complex and professional-looking by adding a third variable and introducing additional notation:

$$\text{Risk} = \Pr(D=1 \mid X_1, X_2, X_3) = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Here, we express risk as the probability that the disease variable equals 1 (as opposed to 0) based on the values of X_1, X_2, X_3. Each β represents the risk difference, or increase in risk, for the corresponding factor (X). The estimate of each β would be based on the observed risk difference across each stratum of the other variables.

This model, of course, is just like the one we developed, except that to make it more impressive, α's and β's are used instead of RD's, X's are used instead of more familiar letters, and a third term has been added. For example, if X_1 is obesity, X_2 reserpine, and X_3 parity (also coded as a dichotomous variable, e.g., nulliparous vs. parous) then the coefficient for X_1 will be a weighted average of the risk difference for obesity use among the four subgroups defined by the other two risk factors:

1. no reserpine-nulliparous women
2. no reserpine-parous women
3. reserpine-nulliparous women
4. reserpine-parous women.

Therefore, each coefficient (risk difference) will be adjusted for the effects of the other variables in the model, more or less as if we had computed an adjusted overall measure in a stratified analysis.

Just as in stratified analysis, the suitability of the coefficient as an adjusted risk difference depends on whether the risk difference for reserpine is essentially the same across the four groups. The model is designed to handle random variability in the risk differences, but not biological (or sociological, artefactual, etc.) reality. So as with any summary measure, the suitability of the linear regression coefficient (i.e., the estimate of the overall risk difference) can be compromised by meaningful heterogeneity of the risk difference across strata of the other variables (i.e., on the extent of statistical interaction or effect modification of the risk difference).

If necessary, the model can accomodate some heterogeneity with the help of an "interaction" term to represent the "difference in risk differences". Interaction terms are usually created as a product of the

two (or more) factors that "interact", since such a term is zero if either of the factors is absent and one only when both are present. For the price of one more Greek letter (γ , gamma) we can write the model:

$$\text{Risk} = \Pr(D=1 \mid X_1,X_2,X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 X_1 X_2$$

provides for the effect of X_1 to depend upon whether X_2 is present or absent (as well as for the effect of X_2 to depend upon whether X_1 is present or absent). But if we incorporate interaction terms for all possible pairs, triplets, . . ., of variables, we will find ourselves right back where we started from – a fully-stratified analysis and no summary measure to use.

The linear model we have just seen has many attractive features, not unimportantly its simplicity and the ease with which statistical estimation of its coefficients can be carried out. Moreover, although we have developed and illustrated the model using only dichotomous, or "binary" variables, the model can readily accomodate count and continuous variables, and with some caution, ordinal variables. (For a nondichotomous variable, the coefficient is the risk difference for a one-unit increase in the variable.)

But linear models also have several drawbacks. First, of course, the data may not conform to an additive model, perhaps to an extent beyond which a single interaction term will suffice to "fit" the data. Second, it is possible to obtain estimates of coefficients that will result in "risks" that are less than zero or greater than one. The linear model in the homework assignment will do that for certain combinations of risk factors, though this is more of a technical objection. Third, linear regression estimates risk differences, but epidemiologists are usually interested in estimating ratio measures of association.

Logistic models:

More widely used in epidemiologic analysis is the logistic model (also referred to as the multiple logistic model or the logit analysis model). In our linear model, above, we chose to model risk as a linear function of two risk factors. In the logistic model, we model the "logit" as a linear function of the risk factors:

$$\text{Logit}(D=1 \mid X_1,X_2,X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

The logit is the natural logarithm of the odds, ln(odds) or ln[p/(l-p)]. It may seem a bit farfetched to work with the logit, rather than risk, but recall our explanation for the use of a logarithmic transformation in order to estimate the variance of a ratio measure.

Whereas risk ranges from 0 to 1, a confining situation for mathematicians, the logit has no bounds. Whereas the risk ratio and the OR have their null value (1.0) way to one side of the range of possible values (zero to infinity), the log(OR) has an unlimited range, with its null value (zero) right in the middle (i.e., it has a symmetrical distribution). We generally use Naperian or "natural" logarithms (base e), abbreviated as ln.

Moreover, the logistic model, we will see, corresponds to a multiplicative model, which we saw earlier is the model that is implied by stratified analysis based on the OR or the risk ratio. Furthermore, the coefficients that we estimate using logistic regression can be converted into OR's, so that we now have a ratio measure of association.

It is easy to discover what the logistic coefficients are. Since the logit is the logarithm of the odds, then the difference of two logits is the logarithm of an OR (because subtraction of logs corresponds to division of their arguments – see the appendix to the chapter on Measures of Frequency and Extent).

Suppose that X_3 is a dichotomous (0-1) variable indicating absence (0) or presence (1) of an exposure. First write the model with the exposure "present" (X_3=1), and underneath write the model with the exposure "absent" (X_3=0).

$$\begin{aligned} &\text{logit}(D=1\,|\,X_1,X_2,X_3=1) = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3 \quad (X_3=1, \text{ present}) \\ -\;&\text{logit}(D=1\,|\,X_1,X_2,X_3=0) = \alpha + \beta_1X_1 + \beta_2X_2 + 0 \quad (X_3=0, \text{ absent}) \\ &\text{———————————————————————————} \end{aligned}$$

When we subtract the second model from the first, all the terms on the right are removed except the coefficient for X_3. On the left, we have the (rather messy) difference of the two logits, one for X_3 present and the other for X_3 absent:

$$\text{logit}(D=1\,|\,X_1,X_2,X_3=1) - \text{logit}(D=1\,|\,X_1,X_2,X_3=0) = \beta_3$$

Spelling out the logits:

$$\ln(\text{odds}(D=1\,|\,X_1,X_2,X_3=1)) - \ln(\text{odds}(D=1\,|\,X_1,X_2,X_3=0)) = \beta_3$$

and, since a difference of logarithms is the logarithm of a ratio:

$$\ln\left[\frac{\text{odds}(D=1\,|\,X_1,X_2,X_3=1)}{\text{odds}(D=1\,|\,X_1,X_2,X_3=0)}\right] = \beta_3$$

A ratio of odds is simply an OR, in this case, the OR for the disease with respect to the exposure represented by X_3:

$$\ln[\,OR\,] = \beta_3$$

$$\exp(\ln[\,OR\,]) = \exp(\beta_3)$$

$$OR = \exp(\beta_3)$$

β_3 is the difference of the logits, hence the log of the OR for the exposure represented by X_3. Therefore $\exp(\beta_3)$ is the OR for a one-unit change in X_3.

Note: $\exp(\beta_1)$ means the anti-logarithm: e, the base for Naperian logarithms, raised to the β_1 power. Since the coefficients are on the logarithmic scale, to see the result on the OR scale, we needed to take the anti-logarithm. For example, a logistic model coefficient of 0.7 corresponds to an OR of about 2.0 for a dichotomous variable or 2.0 for a one-unit increase in a measurement variable.

So the coefficient of a dichotomous explanatory variable is the log of the OR of the outcome with respect to that explanatory variable, controlling for the other terms included in the model. The constant term (α) in a model with only dichotomous risk factor variables is the baseline logit (log odds) for the outcome – the log of the disease odds for a person who has none of the risk factors ($\ln[\Pr(CI_0/(1-CI_0)]$).

For a nondichotomous risk factor, we can compare the odds at two different levels. For example, if age is expressed by a continuous variable X_1 for the number of years, then $\exp(\beta_1)$ gives the OR per year of age and $\exp(10\ \beta_1)$ gives the OR per decade of age.

The logistic model can also be written in terms of risk (i.e., probability) by taking anti-logs (exponents) and employing some algebra. The tranformation is left as an optional exercise for those of you who are interested. The result is:

$$\Pr(D=1 \mid X_1,X_2,X_3) = \frac{1}{1 + \exp(-\alpha - \beta_1X_1 - \beta_2X_2 - \beta_3X_3)}$$

or, if we let L = logit = $\alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$

$$\Pr(D=1 \mid X_1,X_2,X_3) = \frac{1}{1 + \exp(-L)}$$

From the risk formulation we can readily see that the logistic function must range between zero and one, a desirable property for modeling risk. When L (the logit) is "infinitely negative", then exp(-L) is "infinitely large" and the probability estimate is zero. When L is "infinitely large", then exp(-L) is also "infinitely small" and the probability estimate is one. When L is zero, then exp(-L) is 1, and the probability estimate is one-half.

Key epidemiologic assumptions in the logistic model

1. the log odds of disease are linearly related to each of the risk factors (X variables), or equivalently, the disease odds are exponentially related to each of the risk factors, or equivalently, the disease risk is related to each of the risk factors by the logistic (sigmoidal) curve;

2. the joint effects of the risk factors are multiplicative on disease odds (e.g., if a one-unit increase in X_1 alone multiplies incidence odds two-fold and a one-unit increase in X_2 alone multiplies incidence odds three-fold, then a simultaneous one-unit increase in both X_1 and X_2 multiplies incidence odds six-fold) (Greenland, *AJPH*, 1989; Rothman, Modern epidemiology).

In addition, to estimate the coefficients using regression procedures, it must be assumed that the subjects are a random sample of independent observations from the population about which inferences are to be drawn (Harrell, Lee, and Pollock, 1988).

Thus the logistic model corresponds to the multiplicative model for the stratified analysis we considered above. The true OR is assumed constant across all strata. As with the linear model, it is the assumption of homogeneity that permits us to estimate coefficients that are simple to interpret.

We can relax the assumption by including product terms, as illustrated above for the linear model. But then the coefficients are more difficult to interpret. In addition, carried too far that tactic will return us toward a fully-stratified situation and will exhaust our sample size, computer resources, and imagination.

Though we have illustrated both of these models with dichotomous (zero-one) variables, they can readily accomodate continuous variables. Again, the model structure is based on an assumption – that the relationship of the dependent variable (risk, for the linear model, or the logit, for the logistic model) with the independent variable is linear.

For some relationships, this assumption is readily tenable, e.g., CHD risk and number of cigarettes smoked. For others, e.g., mortality risk and body weight, the relationship is U-shaped, so that a simple linear or logistic model will not be suitable (more complex forms of the linear and logistic models are available for U-shaped variables through such techniques as the incorporation of squares of variable values).

Other limitations of the logistic model are that ORs are not the preferred epidemiologic measure of association, and where the outcome is not rare, the proximity of the OR to the risk ratio does not hold. Also, the model cannot provide what the study cannot. Although the logistic model in the above form can be used with case control data, estimates of risk require follow-up data. Mathematics can substitute for data only to a point.

Other regression models *[Optional for EPID 168]*

Two other mathematical model forms that epidemiologists commonly use to control for confounding and to obtain adjusted measures of effects are the proportional hazards and Poisson models.

For an outcome with an extended risk period, especially an outcome that is not rare, it is frequently desirable to use an analysis approach, such as incidence density or survivorship, that takes into account time to the occurrence of the event. The proportional hazards model, developed by David R. Cox, is a widely-used mathematical model for analyzing epidemiologic data where "time to occurrence" is important. The "hazard" (conventionally represented by the Greek letter lambda, λ) is essentially the same concept as instantaneous incidence density.

For three independent variables, the proportional hazards model can be written:

$$\log[ID(t|X_1,X_2,X_3)] = \log[ID_0(t)] + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$$

(i.e., the natural log of incidence density as a function of time is the sum of the log of a background or underlying incidence density plus an increment for each predictor variable).

The model can also be formulated in terms of survivorship:

$$S(t|X_1,X_2,X_3) = [S_0(t)]\exp(\beta_1X_1 + \beta_2X_2 + \beta_3X_3)$$

where S(t) is the probability that the event has not occurred by time t.

The coefficient of a dichotomous predictor is the logarithm of the incidence density ratio [ln(IDR)] for that predictor:

$$\log[ID(t|X_1,X_2,X_3=1)] = \log[ID_0(t)] + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 \quad (X_3 \text{ present})$$

$$-\ \log[ID(t|X_1,X_2,X_3=0)] = \log[ID_0(t)] + \beta_1X_1 + \beta_2X_2 + 0 \quad (X_3 \text{ absent})$$

$$\log[IDR(t)] = \beta_3$$

$$IDR(t) = \exp(\beta_3)$$

In addition to the assumptions required for the logistic model, the Cox proportional hazards model requires that the hazard ratio (the IDR) be constant over time, though more complex survivorship models employing "time-dependent covariates" relax this assumption.

The Poisson model is similar to the logistic model and the proportional hazards model in that the three involve a logarithmic transformation of the risk function (i.e., odds, hazard) being estimated and have a linear combination (i.e., an expression of the form:

$a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots)$ on the right-hand side. The Poisson model is of particular interest when outcomes are very rare.

Key points [EPID 168 students please tune back in here.]

Some guiding principles for multivariable analysis are:

1. Keep in mind that our principal objectives are to describe and interpret the data at hand, using informed judgment, insight, and substantive knowledge as well as technique.
2. Stratified analysis is a very powerful approach. Although it does not hold when we try to analyze many variables simultaneously, we can control for two or three at a time, using different subsets, and let judgment help to fill the gaps. It is always possible that an observed association that is not eliminated when we control for smoking, cholesterol, blood pressure, and Type A behavior pattern individually could still be due to some combined effect of all of these. But how likely is it, especially if we have controlled for each pair of these risk factors and still found the association?
3. In carrying out a stratified analysis for a variable or a combination of variables, we are asking the question "is that combination of variables responsible for the observed result?" The question must be a reasonable one for us to ask. If a few principal risk factors individually do not account for an observed finding, the probability that some combination of them would do so appears less likely. [But no one has demonstrated that proposition empirically.]
4. Mathematical modeling is a very powerful approach to data analysis. But in all cases, a key question is whether the form of the model is appropriate for the data, and the underlying relationships, at hand. Using an inappropriate model can produce biased results. There are statistical techniques for assessing the statistical appropriateness of the models employed ("ask your statistician").

(It is recommended (see Greenland, *AJPH*, 1989; 79(3):340-349 and Vanderbroucke JP: Should we abandon statistical modeling altogether? *Am J Epidemiol* 1987; 126:10-13) that before embarking on modeling exercises that cannot be directly validated against the results of stratified analyses, one should first perform parallel analyses with the same variables in order to validate model choices and results against the stratified data.)

Expectations for EPID 168

- Know the relationship between the multiplicative model and stratified analysis, and (only) basic concepts of linear regression models and logistic regression models. Expectations for your understanding of mathematical modeling are modest:
- Know advantages and disadvantages of modeling (compared to, for example, stratified analysis), as presented in the chapter on confounding.

- Know the epidemiologic meaning of the coefficient of an exposure term in a linear regression model and how the linear regression model relates to stratified analysis and the additive model discussed in the Effect Modification chapter.
- Know the epidemiologic meaning of the coefficient of an exposure term in a logistic model and how that model relates to stratified analysis and the multiplicative model.
- Know the epidemiologic meaning of the coefficient of an exposure term in a proportional hazards model and that that model is used for analyses in terms of incidence density [survivorship]
- For all three models, the coefficients in a model with several variables are all "adjusted" for the effects of the other variables in the model.

Bibliography

Rothman, Modern epidemiology, pp. 285-295; Schlesselman, Case-control studies, pp. 227-234.

Harrell, Frank E., Jr.; Kerry L. Lee, Barbara G. Pollock. Regression models in clinical studies. *JNCI* 1988; 80(15):1198-1202.

Godfrey, Katherine. Simple linear regression in medical research. N Engl J Med 1985;313:1629-36.

Silberberg, Jonathan A. Estimating the benefits of cholesterol lowering: are risk factors for coronary heart disease multiplicative. *J Clin Epidemiol* 1990; 43(9):875-879.

J. Paul Leigh. Assessing the importance of an independent variable in multiple regression: is stepwise unwise? *J Clin Epidemiol* 1988; 41:669-678.

Vandenbroucke, Jan P. Should we abandon statistical modeling altogether? *Am J Epidemiol* 1987; 126:10-13.

Greenland, Sander. Modeling and variable selection in epidemiologic analysis. *Am J Public Health* 1989; 79:340-349 (Advanced)

Kleinbaum, Kupper, and Morgenstern. *Epidemiologic research: Principles and quantitative methods*. Chapters 16-17.

Breslow and Day. *Statistical methods in cancer research. I. The analysis of case-control studies.* Chapters 3-7 (Primarily Chapter 3).

Wilcosky, Timothy C. and Lloyd E. Chambless. A comparison of direct adjustment and regression adjustment of epidemiologic measures. *J Chron Dis* 1985; 38:849-356.

See also: Flanders, W. Dana; and Philip H. Rhodes. Large sample confidence intervals for regression standardized risks, risk ratios, and risk differences. *J Chron Dis* 1987; 40(7):697-704. [includes SAS program]

Deubner, David C., William E. Wilkinson, Michael J. Helms, et al. Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia. *Am J Epidemiol* 1980;112:135-143, 1980.

Weinstein, Milton C.; Pamela G. Coxson, Lawrence W. Williams, Theodore M. Pass, et al. Forecasting coronary heart disease incidence, mortality, and cost: the Coronary Heart Disease Policy Model. *Am J Pubic Health* 1987; 77:1417-1426.

McGee, Daniel; Dwayne Reed, Katsuhika Yano. The results of logistic analyses when the variables are highly correlated: an empirical example using diet and CHD incidence. *J Chron Dis* 1984; 37:713-719.

Breslow and Storer. General relative risk functions for case-control studies. *Am J Epidemiol* 1985;

Szklo, Moyses; F. Javier Nieto. Epidemiology: beyond the basics. Gaithersburg MD, Aspen, 2000. Chapter 7 has an excellent presentation of stratified analysis and mathematical modeling at a basic and understandable level.