CLINICAL EPIDEMIOLOGY AND POPULATION HEALTH

Key Points – Multivariable Modeling

Correlation
A simple way to assess the relationship between two continuous variables is to look for correlation. Plot the values of the two variables on a graph. The *Pearson Correlation Coefficient* (R) is a statistical method for quantifying the correlation between the two variables. R comes out as a value between -1 and +1. The sign (+) or (-) indicates positive or negative slope. The closer that R is to 1 (either sign) the stronger the correlation.

Why multivariable analysis?
Multivariable analysis, like stratified analysis, allows us to isolate the relationship between two variables, holding all other variables constant. In stratified analysis, we stratify by a variable that we suspect might be a confounder (e.g. age), and look at the exposure-outcome relationship for different age strata. If the stratum-specific RRs differ from the crude RR, then the exposure-outcome relationship is confounded by age. But, what if we are interested in considering age, sex, and hair color (and maybe other variables) as possible confounders of the exposure-outcome relationship? Stratified analyses become unwieldy, so we instead turn to multivariable analysis, which can be thought of as multiple, simultaneous stratified analyses.

Fundamentally, regression models fit a mathematical equation to predict the value of an outcome (y) based on values of predictor variables (x_1, x_2, x_3…). We fit the best equation we can to the data we have and then can use it to predict the outcome of y for new values of x we might encounter.

Three main types of multivariable analysis → depends on the outcome variable:

Type of regression	Outcome variable
Linear	Continuous (interval)
Logistic	Dichotomous (yes/no)
Proportional hazards	Length of time until outcome

<u>Linear Regression</u>
Let's say we want to predict blood pressure for all people, but we can only study a sample of 200.

First, start with our data. Examine the association with 1 predictor: age (x axis) and BP (y axis). We could fit the best straight line we could [y = intercept + βx] to these data and we'd use this to predict the BP for others in the population. BUT…our data points do not fall exactly on the line. *We have residual error in our prediction that we'd like to reduce.* So, the equation is actually y = intercept + βx + residual error.

Then, imagine adding BMI to the model by creating a third axis for the graph. If BMI is a useful predictor (in addition to age), the data points will be closer to the line (less residual error) – a more accurate prediction.

y = intercept + $\beta_1 x_1 + \beta_2 x_2$ + residual error (where x_1 is age and x_2 is BMI).

Each β is called a parameter estimate. In linear regression, for a given change in x_1 we multiply it by β_1 to get the change in y, *accounting for all of the other variables in the model for this individual*. In this example the units of β are (mm Hg of BP/year of age).

One can then do this with many variables. This is difficult to visualize but conceptually one is simply adding more and more axes to the graph in multidimensional space in order to enable us to fit the observed data as closely as possible to an equation. These are very powerful techniques used commonly in medical research, but they do have many embedded assumptions that warrant consideration when assessing their validity.

Note:

- We can include categorical variables as predictors (e.g. male or female gender) by assigning them values like 0 and 1.
- One of the underlying assumptions is that the relationship between the predictor of interest and the outcome is a *straight line*. This might not always be true! There are other assumptions as well that we won't focus on here.

Logistic Regression

Think about fitting a straight line to a dichotomous outcome (like developing cancer or not) -- essentially a bunch of 1's and 0's. Fitting a straight line won't work.

Instead, we fit an equation to predict the odds of an individual having the outcome. (Recall that odds are the ratio of the probability of an event happening over it not happening, or: p/1-p). And, rather than a linear association, we assume a logarithmic relationship.

$$\frac{p}{1-p} = e^{\text{intercept} + \beta 1x1 + \beta 2x2 + \text{residual error}}$$

then, if we take the natural log of both sides, the equation is that for a line (similar to linear regression), we've just made the outcome the ln(odds):

$$\ln(p/1-p) = \text{intercept} + \beta_1 x_1 + \beta_2 x_2 + \text{residual error}$$

(This looks scary, but breathe deeply…). All this means is:

- As in linear regression, we use our data to create the best equation (with the smallest residual error) and that gives us the parameter estimates (β) that go with each predictor.
- Then, if we are given new values for x_1 and x_2, we can calculate the odds of an individual getting the outcome.
- Note that in logistic regression, β has no units, but e^{β} is the odds ratio associated with each predictor variable, in predicting the odds of the outcome.
- The predictor variables may be categorical (e.g., e^{β_1} could be the odds of developing lung cancer in males vs. females) or an interval variable (e.g., e^{β_2} could be the odds of getting lung cancer for each additional kilo of body fat).
- As always, if an odds ratio is 1.0 there is no effect, OR > 1 indicates a positive association (higher likelihood of outcome), OR <1 indicates a negative association (lower likelihood of the outcome).
- And, if you have several odds ratios from independent predictors, they can be multiplied to provide an overall estimate of the odds of the outcome.

Proportional Hazards Models (aka. Cox Regression, Survival Analysis)

Analogous to logistic regression, but used when the outcome is *time to* a dichotomous event (like death, or relapse), which allows you to include subjects with varying length of follow-up time (i.e., censoring). The outcome is an estimate of relative risk called a hazard ratio, which is the probability of the event (e.g. dying) during a particular time interval, given that a subject has survived until that time. Hazard ratios are not as intuitive to interpret – in the literature they are often referred to it as a "rate." The math is complex, but it yields coefficients for predictor variables that are exponentiated to hazard ratios (analogous to taking e^{β} to get the odds ratio for logistic regression). Hazard ratios can be interpreted almost exactly like odds ratios in logistic regression (1.0 means no effect, > 1 a positive association, <1 a negative association with the outcome). Just like the models above, proportional hazard models can include several predictor variables at once to control for the effects of all of the other variables.

Final notes

Multivariable techniques are complex, powerful, and very frequently used. We all need to be savvy consumers of results of these models. Here are some issues to keep in mind:

- The intent of a model can be primarily explanatory or predictive. In the former, we seek information about potential *causes* of an outcome. In a predictive model, we just want to use available data to most accurately predict an outcome variable for a new individual—we may be less concerned about whether the predictors are causes or associations.
- As in all statistical tests, we are making inferences from a sample. If we have observed too few individuals in our sample we will be less confident about the conclusions we draw. We need enough outcome observations (at least 10-20 per predictor variable included) to make the model valid.
- We use statistical tests that are similar to those we have studied previously to determine the confidence interval around a parameter estimate β derived from a multivariable model, or can test (at a given level of alpha error) the hypothesis that a predictor is associated with the outcome of interest.
- Choosing which variables to include in a multivariable model is complex, and a bit of an art. Generally, we want to include those that:
    - we know from other research to be important
    - add to the ability of the model to explain or predict the outcome
    - whose inclusion changes the parameter estimates of the main predictor(s) of interest substantially (a common rule of thumb is more than 10%), since this suggests that the additional variable is a confounder of the exposure-outcome relationship.
- Effect modification won't be apparent from a regression model unless you look for it. The simplest way is to stratify the data on the potential effect modifier, run the same model on each stratum, and see if the effect estimates for predictor variables of interest differ within strata. For example, if you are looking for effect modification by age, you would separately examine the relationship between BMI and blood pressure among those under 65 years and among those 65 and older. If the β coefficient for BMI is at least 10% different in the older vs. younger individuals, effect modification exists, and you would report the stratum specific results. There are other methods for addressing "interactions" between variables that are beyond the level of this course.