BST 210
Applied Regression Analysis

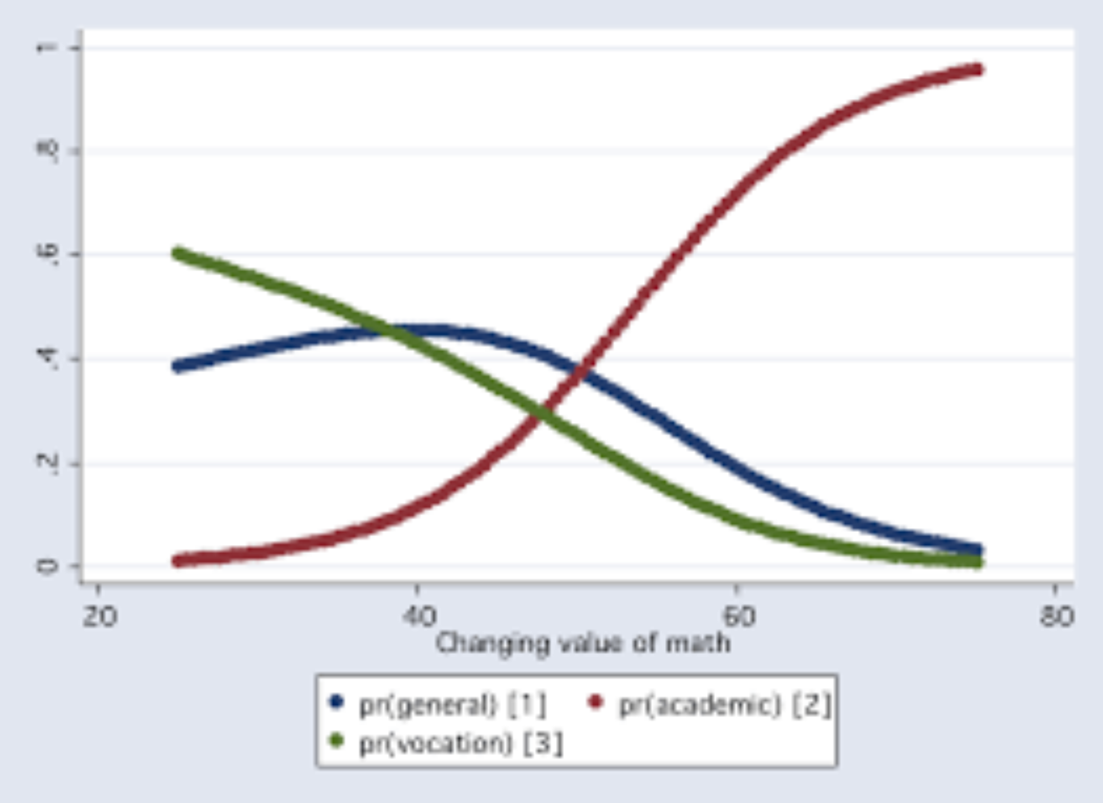
Lecture 18

Plan for Today

- Finish example of Ordinal Logistic Regression (> 2 ordered levels)

- Next: Logistic regression for Case-Control Studies
  - Retrospective design
  - Conditional logistic regression for matched design

Recall: Multinomial and Ordinal Regression



Recall Example: Hyponatremia

- Data not collected on many (488 out of 14k+ runners); no elite runners
- *Self reported*: gender, run time, body mass index, weight gain, water consumption, urination freq, and more
- Useful to have clinically meaningful cutoff for sodium variable (≤135 mmol/l)
- Have normal, moderate, severe sodium level categories to work with

Recall Ordinal Logistic Regression

- Multinomial logistic regression makes no assumptions about the possible ordering of the categories of the outcome variable – the outcome is treated as a nominal variable (and it doesn't even matter which group is chosen to be reference)

- In the hyponatremia example, the outcome is in fact ordinal

- Here we could use *ordinal logistic regression*

Ordinal Logistic Regression

- Suppose that an ordinal outcome *Y* has *c* ordered categories (lowest to highest) labeled as **j** = 1, 2, ... *c*

- The *proportional odds ordinal logistic regression model* is based on the *probability of success* P(*Y* ≥ *j*) *(or cumulative logit)*:

$$\mathbf{logit[P(}\boldsymbol{Y \geq j}\mathbf{)] = log}\ \frac{\mathbf{P(}\boldsymbol{Y \geq j}\mathbf{)}}{\mathbf{P(}\boldsymbol{Y < j}\mathbf{)}}$$

$$= \boldsymbol{\alpha_j + \beta_1 x_1 + \ \ldots + \beta_p x_p}$$

where *j* = 2, ... *C*

- The β coefficients do not differ across outcome categories (but the intercept terms do) – much easier to interpret than multinomial!

- And…we once again have an *odds ratio*!

Ordinal Logistic Regression

- Suppose we have two subjects A and B who differ by 1 unit for the variable x_i, with all other covariates taking the same value

- Then $\exp(\beta_i)$ is just the *odds ratio* that $Y \geq j$ versus $Y < j$ for subject A versus subject B

- $\exp(\beta_i)$ is assumed to be the *same* for each possible value of j

- Estimated probabilities do sum to 1 here

Ordinal Logistic Regression

- This is called the *proportional odds* assumption

- If this assumption is valid, there are many fewer parameters to be estimated than in the multinomial logistic regression model

- These betas are also seeming easier to interpret (OR!)

- If c = 2, again the model reduces to (ordinary/binary) logistic regression

Ordinal versus Multinomial

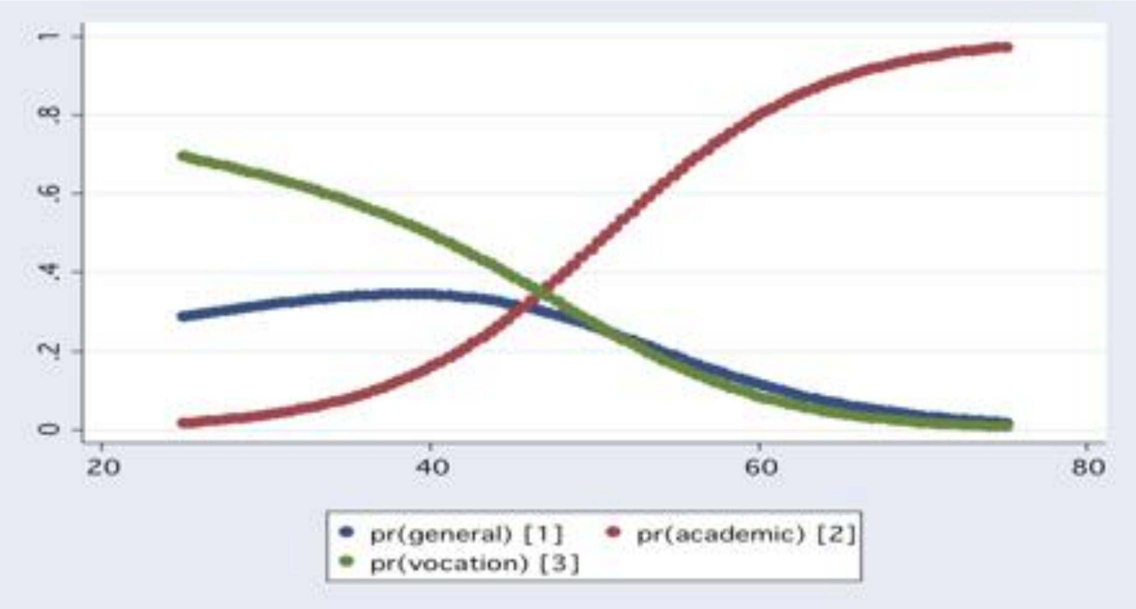
- The ordinal logistic regression model has more assumptions that the multinomial model (ie proportional odds), which may be hard to satisfy (but can be tested)

- Different software packages may have different tests (often approximate tests) of the proportional odds assumption

- This assumption is worth checking, because if the ordinal model is appropriate, model interpretation is simpler (only 1 set of beta coefficients!)—otherwise we use multinomial model.

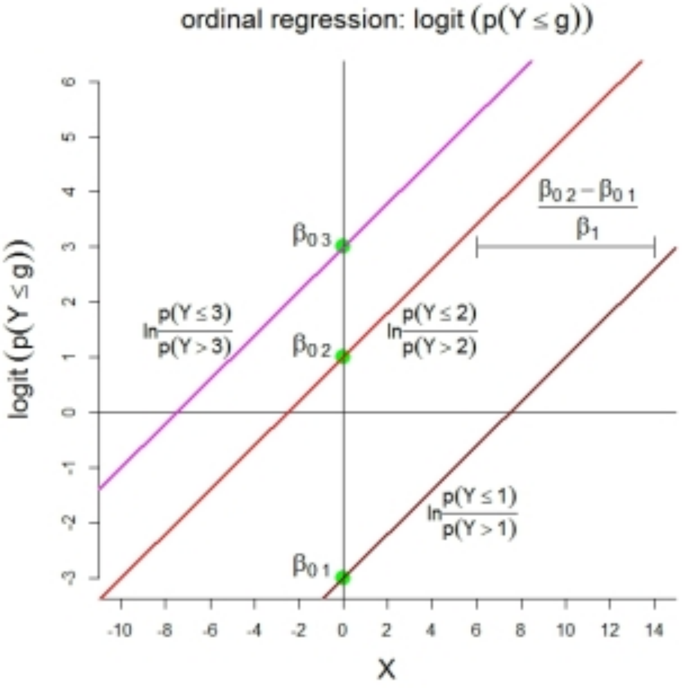
→ *Think Parsimony when possible.*

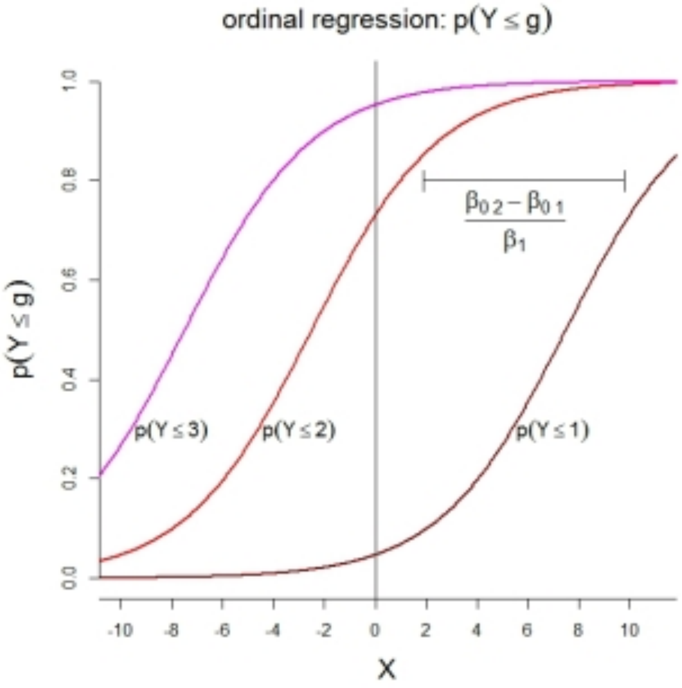
Recall: Multinomial logistic regression



Not concerned about proportional odds! These odds do what they wish!

Ordinal (proportional odds) logistic regression





Ordinal (proportional odds) logistic regression

Let’s return to our example in Stata: (see Example output for Lectures 16 and 18)

Next

- Logistic regression for Case-Control Studies

  - Retrospective design
  - Conditional logistic regression for matched design

Case-Control versus Cohort Studies

Prospective Study:

- watches for outcomes, such as the development of a disease, during the study period and relates this to other factors such as suspected risk or protection factor(s)
- usually involves taking a cohort of subjects and watching them over a long period
- outcome of interest should be common; otherwise, the number of outcomes observed will be too small to be statistically meaningful (indistinguishable from those that may have arisen by chance)
- all efforts should be made to avoid sources of bias such as the loss of individuals to follow up during the study
- Prospective studies usually have fewer potential sources of bias and confounding than retrospective studies

Case-Control versus Cohort Studies

Restrospective Study:

- looks backwards and examines exposures to suspected risk or protection factors in relation to an outcome that is established at the start of the study

- Many valuable case-control studies, such as Lane and Claypon's 1926 investigation of risk factors for breast cancer, were retrospective investigations

- Confounding and bias are more common in retrospective studies than in prospective studies; special care must be taken to try to avoid (retrospective studies are thus sometimes criticized)

- In retrospective studies the odds ratio provides an estimate of relative risk (sampling fraction issue).

Case-Control versus Cohort Studies

- Cohort studies are usually but not exclusively, prospective

- Case-Control studies are usually but not exclusively, retrospective

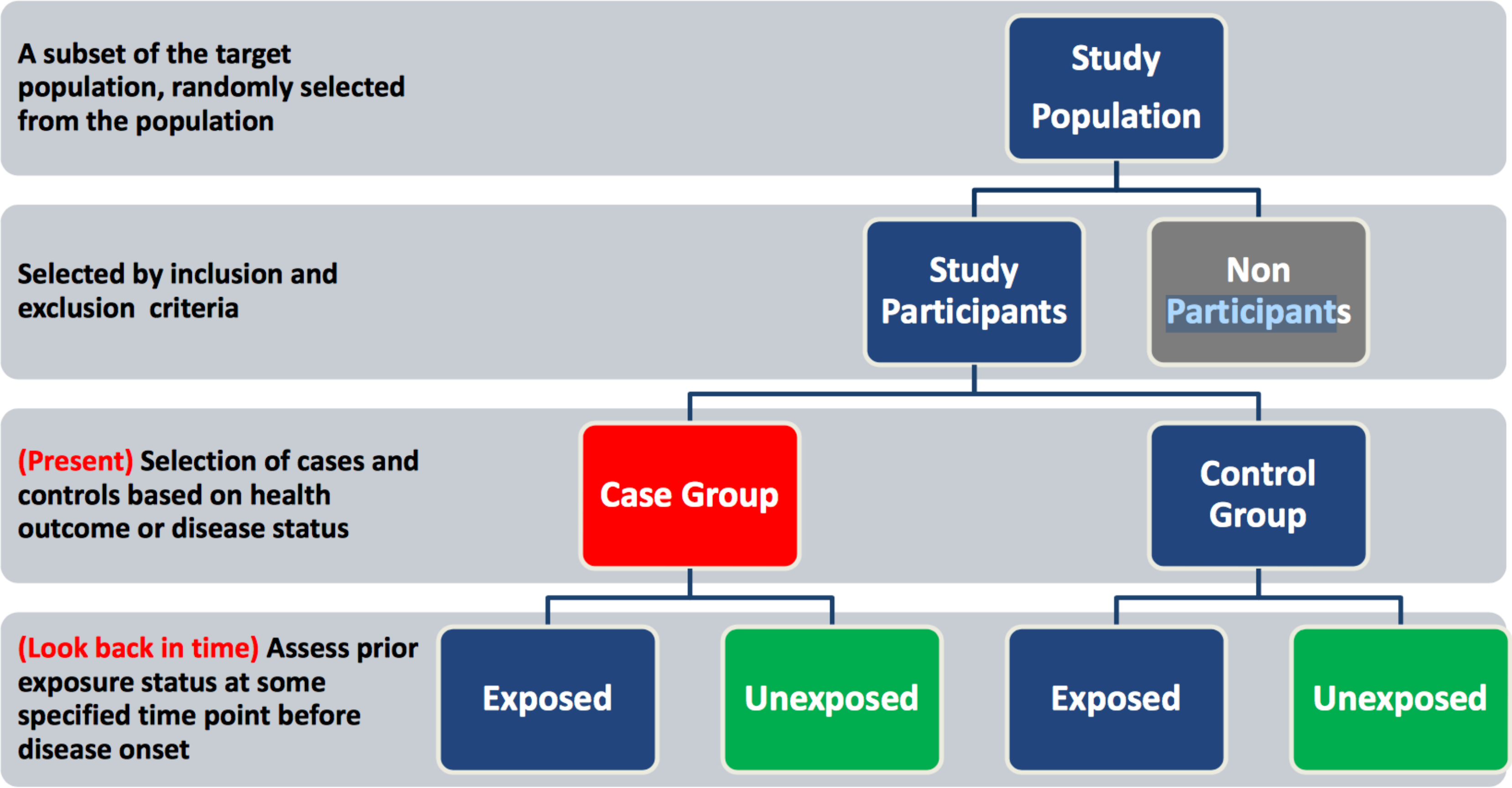
Cohort Studies

- **outcome is measured after exposure (we are deciding on exposure, then waiting for outcomes)**
- yields true incidence rates, relative risks, odds ratios (no *sampling fraction* to worry about)
- may uncover unanticipated associations with outcome
- best for common outcomes
- expensive
- requires large numbers
- takes a long time to complete
- prone to attrition bias or bias of change in methods over time

Case-Control Studies

- **outcome is measured <u>before</u> exposure (we are deciding on outcome, then looking back for exposures– 2x2 table)**
- controls are selected on the basis of not having the outcome
- often conducted before a cohort or an experimental study to identify the possible etiology of the disease.
- must incorporate *sampling fraction* in calculation of incidence rates, relative risks -- not possible to estimate incidence of disease unless study is population based and all cases in a defined population are obtained (odds ratios are always valid in cohort or case-control)
- good for rare outcomes
- relatively inexpensive
- smaller numbers required
- quicker to complete
- prone to selection bias and recall/retrospective bias

Case-Control Study Design



A subset of the target population, randomly selected from the population

Selected by inclusion and exclusion criteria

(Present) Selection of cases and controls based on health outcome or disease status

(Look back in time) Assess prior exposure status at some specified time point before disease onset

Recall: Multiple Logistic Regression Model

- We know that in logistic regression, to predict a binary outcome *Y* with covariates $X_1, \ldots, X_{p,}$ we use the model:

$$\text{logit}(p) = \log[p / (1 - p)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- Here we assume that the relationship between logit(*p*) and the covariates $x_1, \ldots, x_p$ is linear

- Usually we are thinking that the probability *p* is the probability of the event occurring *prospectively*

- What if the events of interest actually happened in the *past*? How does this change our statistical approach?

Case-Control Studies

- We then work under the assumptions of a case-control design

- Suppose we have risk factors x_1, …, x_K in a case-control study with the underlying logistic regression model given by:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki}.$$

- The complicating issue in case-control studies is that the *sampling fraction of cases* might be different (probably higher) than the *sampling fraction of controls*

Case-Control Studies

How do we address this?

- Let τ_1 = proportion of cases sampled in the case-control study, and τ_0 = proportion of controls sampled (different from τ_1, probably lower)

- Assume that the sampling fractions of cases and controls are independent of other risk factors

- We run a logistic regression model for this data with sampling fractions of τ_1 and τ_0

Case-Control Studies

- The true logistic regression model can then be shown to be:

$$\ln[p_i / (1 - p_i)] = \alpha + \ln(\tau_1 / \tau_0) + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki}$$

- We can validly estimate the odds ratio for the *k*th risk factor by $\exp(\beta_k)$

- However, we cannot estimate absolute probabilities of disease (the p_i), since the sampling fractions τ_1 and τ_0 are usually not known in a case-control study

Example: Passive Smoking

- Case-control study examining the association between passive smoking and risk of (any) cancer

- 509 cancer cases and 489 controls with a similar distribution of age and gender were enrolled

- Passive smoking was defined as cigarette smoking by a spouse of ≥ 1 cigarette per day for ≥ 6 months

- One possible confounding variable is active smoking by the subjects themselves

Example: Passive Smoking

Non-Smokers		**Passive Smoker**		
		Yes	No	Total
Case Control Status	Case	120	111	231
	Control	80	155	235
	Total	200	266	466

Example: Passive Smoking

Smokers		**Passive Smoker**		
		Yes	No	Total
Case Control Status	Case	161	117	278
	Control	130	124	254
	Total	291	241	532