

PS 150C/350C
Final Exam
Department of Political Science
Stanford University
Spring 2007

This is an open-book exam. Answer all questions.

1. Let \mathbf{x}_i be a vector of variables on observations $i = 1, \dots, n$, z_i be a continuous variable, y_i be a binary dependent variable, and Φ be the normal CDF. In the model

$$\Pr(y_i = 1 | \mathbf{x}_i, z_i) = \Phi(\mathbf{x}_i \boldsymbol{\beta} + \gamma_1 z_i + \gamma_2 z_i^2)$$

what is the partial effect of z_i on the response probability? What is an estimate of that partial effect? How would you obtain the standard error of the estimated partial effect?

2. Download the Marinov [data](#) (in R dpt format, read it with the `dget` command) on interstate trade sanctions and answer the following questions. Key variables in the data set are
 - `statea`: name of sender
 - `stateb`: name of target
 - `startyear`: length of spell, start of current year
 - `endyear`: length of spell, end of current year
 - `ended`: did sanctions end in current year
 - `usdummy`: is the United States the sender (time invariant)
 - `demtarg`: is the target a democracy (time invariant)
 - `align`: is there an alliance between the two countries (time invariant)
 - `multilat`: is the sanction episode part of a multilateral sanction (time invariant)
 - `gdppctarg`: GDP per capita of the target country (time varying)

- (a) Show the Kaplan-Meier estimates of the survivor functions for sanctions for when the United States is the sender, versus when the United States is not the sender. Comment briefly on what these graphs reveal.
- (b) Show the Kaplan-Meier estimates of the survivor functions for sanctions for when the target country is a democracy, versus when the target country is not a democracy. Comment briefly on what these graphs reveal.
- (c) Show the Kaplan-Meier estimates of the survivor functions for sanctions for when the target country is allied with the sender country, versus when there is no such alliance. Comment briefly on what these graphs reveal.
- (d) In a brief research note (no more than three pages of text), summarize the findings of the analysis. Estimate a Cox proportional hazards model for these data, using all the available predictors. What does the baseline hazard look like, at least qualitatively? What are the effects of various predictors (or combinations of predictors) on risk or expected survival time? Use whatever graphs and tables you need to convey your findings.

Computing hint: `(Surv(startyear, endyear, ended))` will get you started, at least in R. Note that the data are organized by sanction, with one row in the data set for each year of a sanction. A sanction ends when `ended=1` or is otherwise right-censored. You'll also see the `startyear` and `endyear` variables steadily incrementing over the course of a sanction.

3. The file [wagepan.dta](#) contains data (Stata format) from 545 men who worked every year from 1980 to 1987. Consider the wage equation

$$\log(\text{wage}_{it}) = \theta_t + \beta_1 \text{educ}_i + \beta_2 \text{black}_i + \beta_3 \text{hispan}_i + \beta_4 \text{exper}_{it} + \beta_5 \text{exper}_{it}^2 + \beta_6 \text{married}_{it} + \beta_7 \text{union}_{it} + c_i + u_{it}$$

Notice that education is time-invariant. In the data set, the variable `lwage` is the log of the wages, while the other variables have obvious names (`educ`, `black`, `hisp`, etc). The variable `nr` is a unique identifier for each subject. `exper` is the experience variable, measured as years that a person has been in the labor market. Summarize the models you fit in the following questions in a publication-quality table.

- (a) How much of the variation in log wages is cross-sectional, and how much is longitudinal?
- (b) Estimate this model by pooled OLS (i.e., ignoring the unit-specific term c_i). Are the OLS standard errors reliable, even if c_i is uncorrelated with all explanatory variables? Explain. Compute appropriate standard errors.
- (c) Estimate the wage equation treating the c_i as “random effects”. Compare your estimates with the pooled OLS estimates.
- (d) Now estimate the equation by fixed effects. Why is exper_{it} redundant in the model even though it changes over time? What happens to the marriage and union wage premiums as compared with their corresponding random effects estimates?
- (e) What does a Hausman test say about the plausibility of the random effects assumption?
- (f) Estimate an education effect specific to each year. How well does this model fit the data compared to the model with a time-invariant education effect? Has the return to education increased over time? Offer a substantive interpretation of the returns to education.

4. Suppose that we have the unobserved effects model

$$y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma} + h_i + u_{it}$$

where the \mathbf{x}_{it} are time-varying, the \mathbf{z}_i are time-constant, $E(u_{it}|\mathbf{x}_{it}, \mathbf{z}_i, h_i) = 0$, $t = 1, \dots, T$ and $E(h_i|\mathbf{x}_i, \mathbf{z}_i) = 0$. Let $\sigma_h^2 = \text{var}(h_i)$ and $\sigma_u^2 = \text{var}(u_{it})$. If we estimate $\boldsymbol{\beta}$ by fixed effects, we are estimating the equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}$$

where $c_i = \alpha + \mathbf{z}_i\boldsymbol{\gamma} + h_i$.

- (a) Find $\sigma_c^2 = \text{var}(c_i)$. Show that σ_c^2 is at least as large as σ_h^2 and usually strictly larger.
- (b) Explain why estimation of the model by fixed effects will lead to a larger estimated variance of the unobserved effect than if we estimate the model by random effects.

5. The data frame `bacteria` in the R package MASS contains data from a drug trial administered to children with a history of otitis media (infection of the middle ear). Fifty children participated in 5 regular screenings for the presence of *H. influenzae* (bacteria responsible for a wide range of diseases, including some nasties like acute bacterial meningitis, but more run of the mill stuff like conjunctivitis, sinusitis, and ear infections such as otitis media). Children were randomly assigned to one of four conditions at the initial screening in a crossed experimental design: (a) a drug or a placebo; (b) encouragement to take their assigned medication, or no such encouragement. The data frame has 220 observations; not all children were monitored at all 5 times points. The variables are

- `y`: presence or absence of *H. influenzae*, a factor with levels `n` and `y`;
- `ap`: active drug or placebo, a factor with levels `a` and `p`;
- `hilo`: hi/low encouragement to comply with treatment, a factor with levels `hi` and `lo`;
- `week`: numeric, the week of the screening;
- `ID`: subject ID, a factor;
- `trt`: a factor with levels `placebo`, `drug`, `drug+`, a re-coding of `ap` and `hilo`

We are interested in the efficacy of the treatments, taking into account the fact that we don't expect to see any response to treatment until the second screening (since treatments were assigned at the initial screening).

- (a) Provide some kind of assessment as to how much of the variation in these data is “between”, and how much is “within”.
- (b) Estimate a logit model for these data with `y` as the (binary) response, and the treatment variables as the only predictors. Briefly comment on what you find.
- (c) Estimate the unconditional transition probabilities for `y`, week by week. That is, at each week (other than week 0) what is the probability of testing positive for *H. influenzae* given the status of

a child's test in the previous week. Hint: this is nothing other than computing 2-by-2 tables, but you have to set them up the right way (or rather, set up the lagged variables the right way).

- (d) A plausible rival hypothesis is *maturation*: that over time, the children's immune systems would attack the bacteria, and absent treatment, children would tend to test negative for *H. influenzae*. What does your answer to the previous question suggest about this possibility?
- (e) Augment your logit model with some kind of control for this maturation hypothesis. Does this alter your conclusion about the effects of the treatments? What specification of time dependence is best supported by the data: a linear time trend (on the log-odds scale), or fixed effects for each time period, something else, or nothing?
- (f) The data available for analysis have no individual level covariates, but it is plausible that unobserved subject-level heterogeneity is a factor in these data. Comment on this possibility, keeping in mind that we have an experiment.
- (g) The R package `glmmML` has a function with the same name for fitting GLMs with random intercepts ("random effects"). The argument `cluster` in the `glmmML` function specifies the grouping variable for the random effects (i.e., in this case, ID). Use this function to augment your preferred logit model from the previous steps.
- (h) Why is random effects appropriate here? How do your results change, if at all, via the introduction of the random effects?
- (i) Provide a short statement as to the efficacy of the treatments deployed in this case. In a graph or two, show how the predicted probability of having *H. influenzae* changes over time, as a function of the two overlapping treatments. Augment your graphs with confidence intervals etc.

END OF EXAM