Lecture 5. Nonlinear regression models

Economics 8379
George Washington University

Instructor: Prof. Ben Williams

Binary choice

If Y_i is binary then $E(Y_i \mid X_i = x) = Pr(Y_i = 1 \mid X = x)$

- the CEF is likely not linear
- but OLS provides the best linear approximation to the CEF, $Pr(Y_i = 1 \mid X_i)$

Binary choice

- Suppose D_i is a randomly assigned binary treatment variable.
  - let β^{OLS} denote the OLS estimand from a regression of Y_i on D_i
  - then

$$\beta^{OLS} = E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0) = ATE$$

Binary choice

- Suppose that $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i \mid X_i$
  - if the model is fully saturated in X_i,

$$\beta^{OLS} = \sum_x \delta_x w_x$$

where

- w_x are weights proportional to $P(x)(1 - P(x))Pr(X_i = x)$
- $\delta_x = E(Y_1 - Y_0 \mid X_i = x)$

AP's reasons to avoid probit/logit

- "regression gives us what we need with or without the probit distributional assumptions"
- "if the CEF has a causal interpretation, it seems fair to say that regression has a causal interpretation as well, because it still provides the MMSE approximation to the CEF"
- "...while a nonlinear model may fit the CEF ... more closely than a linear model, when it comes to marginal effects, this probably matters little."
- too many decisions to make along the way, while OLS is standardized
- life gets more complicated with IV and panel data

Latent index model

- Let $Y_i^* = \beta' X_i + \varepsilon_i$ denote a latent index and suppose that we observe $Y_i = \mathbf{1}(Y_i^* \geq 0)$.

Latent index model

- Let $Y_i^* = \beta' X_i + \varepsilon_i$ denote a latent index and suppose that we observe $Y_i = \mathbf{1}(Y_i^* \geq 0)$.
  - back to generic notation where X_i can include a "treatment" and "controls"

Latent index model

- Let $Y_i^* = \beta' X_i + \varepsilon_i$ denote a latent index and suppose that we observe $Y_i = \mathbf{1}(Y_i^* \geq 0)$.
  - back to generic notation where X_i can include a “treatment” and “controls”
- If ε_i and X_i are independent then

$$Pr(Y_i = 1 \mid X_i) = F_{\varepsilon_i}(\beta' X_i)$$

Latent index model

- Let $Y_i^* = \beta' X_i + \varepsilon_i$ denote a latent index and suppose that we observe $Y_i = \mathbf{1}(Y_i^* \geq 0)$.
  - back to generic notation where X_i can include a "treatment" and "controls"
- If ε_i and X_i are independent then

$$Pr(Y_i = 1 \mid X_i) = F_{\varepsilon_i}(\beta' X_i)$$

  - if F_{ε_i} is the standard normal cdf this is the *probit* model
  - if $F_{\varepsilon_i}(x) = \frac{\exp(x)}{1+\exp(x)}$ this is the *logit* model
  - if $F_{\varepsilon_i}(x) = x\mathbf{1}(0 \leq x \leq 1)$ this is the linear probability model

Latent index model

- The latent index may have a structural interpretation (random utility, shadow price, etc.).
- In the structural interpretation it often does not make sense to restrict the standard deviation of ε_i.
  - Assume that $\varepsilon_i \mid X_i \sim N(0, \sigma_\varepsilon^2)$
  - Then

$$\begin{aligned} Y_i &= \mathbf{1}(\beta' X_i + \varepsilon_i \geq 0) \\ &= \mathbf{1}(\frac{\beta}{\sigma_\varepsilon}' X_i + \frac{\varepsilon_i}{\sigma_\varepsilon} \geq 0) \end{aligned}$$

  - Thus $Pr(Y_i = 1 \mid X_i) = \Phi\left(\frac{\beta'}{\sigma_\varepsilon} X_i\right)$

Latent index model

- The latent index may have a structural interpretation (random utility, shadow price, etc.).
- In the structural interpretation it often does not make sense to restrict the standard deviation of ε_i.
  - Assume that $\varepsilon_i \mid X_i \sim N(0, \sigma_\varepsilon^2)$
  - Then

$$
\begin{aligned}
Y_i &= \mathbf{1}(\beta' X_i + \varepsilon_i \geq 0) \\
&= \mathbf{1}(\frac{\beta}{\sigma_\varepsilon}' X_i + \frac{\varepsilon_i}{\sigma_\varepsilon} \geq 0)
\end{aligned}
$$

  - Thus $Pr(Y_i = 1 \mid X_i) = \Phi\left(\frac{\beta'}{\sigma_\varepsilon} X_i\right)$
  - so we can't separate β from σ_ε

Marginal effects

- marginal effect of a continuous regressor:
  $\frac{\partial}{\partial x_k} Pr(Y_i = 1 \mid X_i = x) = \beta_k f_{\varepsilon_i}(\beta' x)$
- the partial effect of a discrete regressor
  - Suppose $X_i = (D_i, \tilde{X}_i)$.
  - We estimate the partial effect of D_i as a difference:
    $F_{\varepsilon_i}(\beta_0 + \beta_1 + \beta_2' \tilde{x}) - F_{\varepsilon_i}(\beta_0 + \beta_2' \tilde{x})$
- marginal effects at the mean: $\beta_k f_{\varepsilon_i}(\beta' \bar{X})$
- average marginal effect: $\beta_k E\left(f_{\varepsilon_i}(\beta' X_i)\right)$
- `margins` command in Stata

Estimation

- estimation is via maximum likelihood:

$$\hat{\beta} = max_{\beta} \sum_{i=1} Y_i ln\left(F_{\varepsilon}(\beta' X_i)\right) + (1 - Y_i) ln\left(1 - F_{\varepsilon}(\beta' X_i)\right)$$

- in small samples or high dimensional models you might experience convergence problems:

Estimation

- estimation is via maximum likelihood:

$$\hat{\beta} = max_{\beta} \sum_{i=1} Y_i ln\left(F_{\varepsilon}(\beta' X_i)\right) + (1 - Y_i) ln\left(1 - F_{\varepsilon}(\beta' X_i)\right)$$

- in small samples or high dimensional models you might experience convergence problems:
  - the MLE does not exist if there is a β such that $\beta' X_i \geq 0$ for all $i : Y_i = 0$ and $\beta' X_i \leq 0$ for all $i : Y_i = 1$
  - it is not clear whether Stata is able to catch all cases of this
  - if the "overlap" is small and there are many regressors then Stata's algorithm my have difficulty converging
  - problems with approximating probit cdf when probabilites are close to 0/1 (outliers)

probit/logit versus OLS

- causal effects in the latent index model:
  - independence between ε_i and (D_i, X_i) implies CIA
  - then

$$\begin{aligned}\delta_x &= E(Y_{1i} - Y_{0i} \mid X_i = x) \\ &= F_{\varepsilon_i}(\beta_0 + \beta_1 + \beta_2' x) - F_{\varepsilon_i}(\beta_0 + \beta_2' x)\end{aligned}$$

  - nonlinearity induces heterogeneous effects
  - if the model is not fully saturated in X_i, the nonlinearity can make problems even worse

probit/logit versus OLS

- causal effects in the latent index model:
  - independence between ε_i and (D_i, X_i) implies CIA
  - then

$$\begin{aligned}\delta_x &= E(Y_{1i} - Y_{0i} \mid X_i = x) \\ &= F_{\varepsilon_i}(\beta_0 + \beta_1 + \beta_2' x) - F_{\varepsilon_i}(\beta_0 + \beta_2' x)\end{aligned}$$

  - nonlinearity induces heterogeneous effects
  - if the model is not fully saturated in X_i, the nonlinearity can make problems even worse

- misspecification is a valid concern
  - suppose ε_i is heteroskedastic
  - one solution to this problem is a semiparametric model (average derivative methods or maximum score methods)

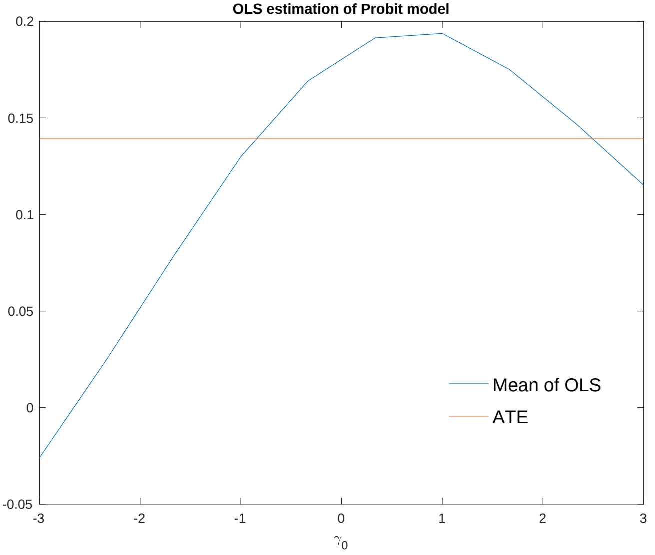
Illustration of OLS bias

- I simulated the following model:

$$
\begin{aligned}
X_i &\sim N(0,1) \\
D_i &= \mathbf{1}(\gamma_0 + X_i \geq v_i), \quad v_i \sim N(0,1) \\
Y_i &= \mathbf{1}(0.5D_i + X_i \geq u_i), \quad u_i \sim N(0,1)
\end{aligned}
$$

- The ATE is $E\left(\Phi(.5 + X_i) - \Phi(X_i)\right) \approx 0.14$
- I simulate the model for a grid of values of γ_0 between -3 and 3 for $n = 1000$ observations.

Illustration of OLS bias



Structural models

- Over the next few lectures I want to introduce you to structural estimation methods.
- Today I begin by familiarizing you with some nonlinear models which are commonly used.
- We will also take about maximum likelihood because this gives us practice in moving from an economic model to an econometric specification.
- Next class we will discuss other estimation methods.

Maximum likelihood

- You've seen theoretical conditions for maximum likelihood estimation before. See Cameron and Trivedi for a review.
- Suppose we observe a vector of outcomes Y_i and covariates X_i.
- Our model fully specifies, up to a parameter vector β, the distribution of Y_i conditional on X_i via a density $f_{Y|X}(Y_i \mid X_i; \beta)$.

Maximum likelihood

- With iid data, the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} f_{Y|X}(Y_i \mid X_i; \beta)$$

- Let $\mathcal{L}(\beta) = log(L(\beta)) = \sum_{i=1}^{n} \log(f_{Y|X}(Y_i \mid X_i; \beta))$. Then

$$\hat{\beta}_{MLE} = arg \max_{\beta} \mathcal{L}(\beta)$$

Properties of MLE

- $\hat{\beta}_{MLE} \rightarrow_p \beta$ and $\sqrt{n}(\hat{\beta}_{MLE} - \beta) \rightarrow_d N(0, \mathcal{I}^{-1})$ where
- $\mathcal{I} = plim_{n\rightarrow\infty} \frac{1}{N} \frac{\partial\mathcal{L}(\beta)}{\partial\beta} \frac{\partial\mathcal{L}(\beta)}{\partial\beta'}$ (Fisher information matrix)

Properties of MLE

- $\hat{\beta}_{MLE} \rightarrow_p \beta$ and $\sqrt{n}(\hat{\beta}_{MLE} - \beta) \rightarrow_d N(0, \mathcal{I}^{-1})$ where
- $\mathcal{I} = plim_{n \rightarrow \infty} \frac{1}{N} \frac{\partial \mathcal{L}(\beta)}{\partial \beta} \frac{\partial \mathcal{L}(\beta)}{\partial \beta'}$ (Fisher information matrix)
- $E\left(\frac{\partial \mathcal{L}(\beta)}{\partial \beta} \frac{\partial \mathcal{L}(\beta)}{\partial \beta'}\right) = -E\left(\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta \partial \beta'}\right)$ (information matrix equality)

Properties of QMLE

- Suppose $f_{Y|X}(Y_i \mid X_i; \beta)$ is not the correct density.
  - $\hat{\beta}_{MLE} \rightarrow_p \beta^*$, pseudo-true value that maximizes $plim_{n\rightarrow\infty} \frac{1}{n}\mathcal{L}(\beta)$

Properties of QMLE

- Suppose $f_{Y|X}(Y_i \mid X_i; \beta)$ is not the correct density.
    - $\hat{\beta}_{MLE} \to_p \beta^*$, pseudo-true value that maximizes $plim_{n\to\infty} \frac{1}{n}\mathcal{L}(\beta)$
    - $\sqrt{n}(\hat{\beta}_{MLE} - \beta^*) \to_d N(0, A^{-1}BA^{-1})$ where
    - $B = plim_{n\to\infty} \frac{1}{N} \frac{\partial \mathcal{L}(\beta)}{\partial \beta} \frac{\partial \mathcal{L}(\beta)}{\partial \beta'}$ and $A = plim_{n\to\infty} \frac{1}{N} \frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta \partial \beta'}$

Properties of (Q)MLE

- Under correct specification, $A^{-1}BA^{-1} = B^{-1} = \mathcal{I}^{-1}$.
- Example:
  - OLS is equivalent to MLE assuming homoskedastic normal errors
  - If errors are heteroskedastic, we can use a sandwich formula that accounts for heteroskedasticity (Eicker-Huber-White standard errors)
  - In this case, the pseudo-true value is β.
  - The "robust" option for a probit does the same thing, but the pseudo-true value is *not* β

Nonlinear least squares

- The nonlinear least squares (NLS) estimator is an alternative to MLE.
  - less efficient than MLE
  - but relies on weaker distributional assumptions

Nonlinear least squares

- The nonlinear least squares (NLS) estimator is an alternative to MLE.
  - less efficient than MLE
  - but relies on weaker distributional assumptions
- Suppose $Y_i = g(X_i, \beta) + u_i$ and $E(u_i \mid X_i) = 0$.
- Then $\hat{\beta}_{NLS}$ minimizes

$$\sum_{i=1}^{n} (Y_i - g(X_i, \beta))^2$$

Nonlinear least squares

- Sandwich variance matrix:
  - $\hat{\beta}_{NLS} \to_p \beta$ and $\sqrt{n}(\hat{\beta}_{NLS} - \beta) \to_d N(0, A^{-1}BA^{-1})$
  - where $A = plim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial g(X_i,\beta)}{\partial \beta} \frac{\partial g(X_i,\beta)}{\partial \beta'}$ and $B = plim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} E(u_i u_j \mid X) \frac{\partial g(X_i,\beta)}{\partial \beta} \frac{\partial g(X_i,\beta)}{\partial \beta'}$

Other estimators

- Variations on NLS (e.g., FGNLS)
- GMM (more on this in a few classes)
- simulation-based versions of these

Random utility model for multinomial outcomes

We can start with a very general random utility model.

- Individual (or household, firm, etc.) i has a choice among m alternatives.
- For $j = 1, \ldots, m$, utility for choice j is $U_{ij} = V_{ij} + \varepsilon_{ij}$ where V_{ij} will be a function of observables and ε_{ij} is unobservable.
- Then the probability that i chooses j (conditional on observables) is:

$$\begin{aligned} p_{ij} &:= Pr\left(U_{ij} = \max_{k=1,\ldots,m} U_{ik}\right) \\ &= Pr\left(\varepsilon_{ik} - \varepsilon_{ij} \leq V_{ij} - V_{ik} \text{ for all } k \neq j\right) \end{aligned}$$

Random utility model for multinomial outcomes

- The log likelihood is then $\sum_{i=1}^{n} \sum_{j=1}^{m} \log(p_{ij}) y_{ij}$ where y_{ij} is equal to 1 if observation i chose option j and 0 otherwise.
- There are then two choices to make:
  - how to specify $V_{i1}, \ldots, V_{im}$
  - how to specify the joint distribution of $\varepsilon_{i1}, \ldots, \varepsilon_{im}$

Multinomial logit model

- Logit models are derived from the assumption that $\varepsilon_{i1}, \ldots, \varepsilon_{im}$ are independent with identical type 1 extreme value distributions
  - sometimes called the Gumbel distribution, sometimes abbreviated EV1, this distribution has cdf $F(x) = e^{-e^{-x}}$
- Under this assumption,

$$
\begin{aligned}
p_{ij} &= \frac{\exp(V_{ij})}{\sum_{k=1}^{m} \exp(V_{ij})} \\
&= \frac{\exp(V_{ij} - V_{i1})}{1 + \sum_{k=2}^{m} \exp(V_{ij} - V_{i1})}
\end{aligned}
$$

Multinomial logit model

Independence of irrelevant alternatives (IIA)

- Notice that for two choices $j \neq k$,

$$\frac{p_{ij}}{p_{ik}} = \frac{\exp(V_{ij})}{\exp(V_{ik})}$$

- The relative probability of the two options is not affected by other options at all!
- “red bus-blue bus” problem

Multinomial logit model

Specifying V_{ij}

- What makes utility of one choice higher than utility of another?
  - choice-specific characteristics, including price
  - preferences, which vary with individual characteristics
- A general model that includes both: $V_{ij} = \beta' x_{ij} + \gamma_j' w_i$
  - x_{ij} are choice-specific characteristics, which may also vary with the individual
  - w_i is an individual characteristic and γ_j reflects how this characteristic influence utility of choice j
  - note that we must normalize $\gamma_1 = 0$

Multinomial logit model

Specifying V_{ij}

- note that $V_{ij} - V_{i1} = \beta'(x_{ij} - x_{i1}) + (\gamma_j - \gamma_1)' w_i$
- we can always add the same constant to γ_j and γ_1 and the likelihood does not change
- so we must normalize $\gamma_1 = 0$

Multinomial logit model

Marginal effects

- For the choice-specific variables:

$$\frac{\partial p_{ij}}{\partial x_{ij}} = p_{ij}(1 - p_{ij})\beta$$
$$\frac{\partial p_{ij}}{\partial x_{ik}} = -p_{ij}p_{ik}\beta, k \neq j$$

- For regressors that don't vary with choice:

$$\frac{\partial p_{ij}}{\partial w_i} = p_{ij}\left(\gamma_j - \sum_{k=1}^{m} \gamma_k p_{ik}\right)$$

Multinomial logit model

Log odds ratio interpretation

- If $V_{ij} = \gamma_j' w_i$ then

$$\log\left(\frac{p_{ij}}{p_{ik}}\right) = (\gamma_j - \gamma_k)' w_i$$

- Since $\gamma_1 = 0$, coefficient estimates $\hat{\gamma}_j$ can then be interpreted as the increase in the log odds ratio of choice j relative to choice 1 due to a one unit increase in w_i.

Alternatively, we can simulate the model to answer different policy counterfactuals.

Multinomial logit model

In Stata

- `mlogit`
  - data structure – each row is an individual and the `depvar` is a categorical variable
  - syntax – `mlogit depvar indepvars, baseoutcome(value)` where `value` is the value for the dependent variable indicating the choice where we impose the normalization
  - model – only works for $V_{ij} = \gamma_j' w_i$
- `asclogit`
  - data structure – each row is an individual, choice pair and the `depvar` is a dummy variable
  - syntax – `asclogit depvar indepvars, case(id) alternatives(choice) basealternative(value)` where `value` is the value for the dependent variable indicating the choice where we impose the normalization.
  - model – works for $V_{ij} = \beta' x_{ij} + \gamma_j' w_i$
  - w_i are specified using `casevars` option

Multinomial logit model

More on IIA

- Suppose $V_{ij} = \alpha \cdot price_j + \beta' x_{ij} + \gamma_j' w_i$.
- Then the cross price elasticity $\left(\frac{\partial p_{ij}}{\partial price_k} \frac{price_k}{p_{ij}}\right)$ is equal to

$$\alpha price_k p_{ik}.$$

- It is the same for all j!!
- One solution is to model the correlations between ε_{ij} and ε_{ik} explicitly (see multinomial probit next class).
- Two more solutions will be previewed.

Nested logit model

- In some cases we can group choices together – red bus and blue bus are both buses.
- The nested logit models the probability of choosing option k which is part of group j by

$$p_{jk} = p_j \times p_{k|j}$$

- For the nested logit with $V_{jk} = \alpha' z_j + \beta' x_{jk}$ for J groups where group j has K_j choices:

$$p_{jk} = \frac{\exp(\alpha' z_j + \rho_j I_j)}{\sum_{m=1}^{J} \exp(\alpha' z_j + \rho_j I_j)} \frac{\exp\left((\beta_j/\rho_j)' x_{jk}\right)}{\sum_{l=1}^{K_j} \exp\left((\beta_j/\rho_j)' x_{jl}\right)}$$

Random coefficients logit model

- We can generalize the utility model to include an individual-specific coefficient that is treated as a "random effect" $V_{ij} = \beta_i' x_{ij}$.
- In this model,

$$p_{ij} = \int p_{ij}(\beta_i) f_{\beta_i}(\beta_i) d\beta_i$$

where $p_{ij}(\beta) = \frac{\exp(\beta' x_{ij})}{\sum_{k=1}^{m} \exp(\beta' x_{ik})}$

- Typically f_{β_i} is specified as a normal distribution with a mean and variance to be estimated.

Kleven et al. (2013)

- A model for country choice (of European football players).
  - The multinomial choice model: for player i in time t playing in country n yields utility:

$$U^i_{nt} = \alpha log(1 - \tau^i_{nt}) + \alpha log(w^i_{nt}) + home^i_n + x^i_t \beta_n + \gamma_n + \nu^i_{nt}$$

Kleven et al. (2013)

- A model for country choice (of European football players).
  - The multinomial choice model: for player i in time t playing in country n yields utility:

$$U^i_{nt} = \alpha log(1 - \tau^i_{nt}) + \alpha log(w^i_{nt}) + home^i_n + x^i_t \beta_n + \gamma_n + \nu^i_{nt}$$

  - multinomial logit
  - can you map the notation here to the general notation for the multinomial logit above in the slides?
  - various specifications to account for not observing w^i_{nt}
  - probability that i chooses n in year t is $P^i_{nt} = Pr(U^i_{nt} \geq U^i_{mt} \quad \forall m)$

Kleven et al. (2013)

- Tax elasticities:
  - they compare estimates of

$$\varepsilon^n_{domestic} = \frac{d \log(\sum_{i \in I_n} P^i_{nt})}{d \log(1 - \tau_{nd})} = \alpha(1 - \bar{P}^d_n)$$

and

$$\varepsilon^n_{foreign} = \frac{d \log(\sum_{i \in I^C_n} P^i_{nt})}{d \log(1 - \tau_{nf})} = \alpha(1 - \bar{P}^f_n)$$

  - these formulas show how restrictive the multinomial logit can be

Christensen and Kiefer

A job search model

- Suppose job offers are distributed according to a density $f(w)$.
- There is a reservation wage w_r such that each worker i accepts offer w_i if $w_i \geq w_r$.
- The distribution of accepted offers is

$$g(w) = \frac{f(w)}{\int_{w_r}^{\infty} f(w)dw} \mathbf{1}(w \geq w_r)$$

Christensen and Kiefer

Taking the model to data

- Suppose $f(w)$ and w_r are parameterized by a vector θ.
- We observe a sample of wages for workers (who are assumed to have accepted a wage offer).
- So we observe $(w_1, \ldots, w_n)$, an iid sample from $g(w)$.
- Thus, the likelihood is

$$L(\theta) = \prod_{i=1}^{n} \frac{f(w_i;\theta)}{\int_{w_r(\theta)}^{\infty} f(w;\theta)dw} \mathbf{1}(w_i \geq w_r(\theta))$$

Christensen and Kiefer

Taking the model to data

- One option is to take $f(w) = \gamma \exp(-\gamma(w - c))$.
- It turns out that g does not end up depending on c so we can take $\theta = (\gamma, w_r)$ and

$$L(\theta) = \gamma^n \exp\left(-\gamma \sum_{i=1}^{n}(w_i - w_r)\right) \mathbf{1}(min(w_i) \geq w_r)$$

- This likelihood function has some weird properties (regardless of how $f(w)$ is parameterized; assumption (iv) in Prop 5.5 in CT; see the paper for details) so the authors assume wages are observed with error.

Christensen and Kiefer

Model with measurement error

- It is assumed that we observe $w_i^e = w_i m_i$ where w_i is iid from $g(w)$.
- They maintain the shifted exponential assumption, $f(w) = \gamma \exp(-\gamma(w - c))$.
- The measurement error, m_i is assumed to have density $h(m_i)$ with support on $[0, \infty)$.
- Note then that for any x,

$$Pr(w_i^e \leq x) = \int_0^\infty Pr(w_i \leq \frac{x}{m_i} \mid m_i)h(m_i)dm_i$$

Christensen and Kiefer

Model with measurement error

- This can be written as

$$Pr(w_i^e \leq x) = \int_0^{x/w_r} (1 - \exp(-\gamma(x/m_i - w_r)))\, h(m_i) dm_i$$

- To derive the likelihood function we need the density of w_i^e, which will be denoted $f_e(x)$.

$$\begin{aligned} f_e(x) &= \frac{d}{dx} Pr(w_i^e \leq x) \\ &= \gamma \exp(\gamma w_r) \int_0^{x/w_r} \frac{1}{m} h(m) \exp(-\gamma x/m) dm \end{aligned}$$

- This is derived assuming certain properties of h.

Christensen and Kiefer

Specifying the distribution of measurement error

- First, the density h must satisfy some properties for f_e to take the form on the previous slide.
- Second, we want to use a flexible family of distributions as we do not know much about what the distribution should look like.
- Further, we want the resulting expression for f_e to be tractable.

Christensen and Kiefer

The resulting f_e:

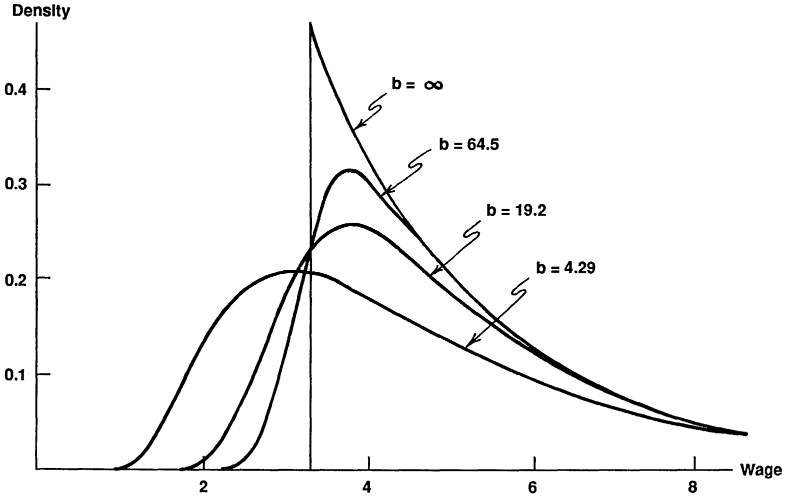


FIG. 1.—Observed wage densities

Likelihood function

- Moving from a model, written in equations, to the appropriate likelihood function?
- Can be difficult if your model isn't a textbook case.
- Here I will provide some examples.

Censoring

- Let Y^* denote the outcome of interest.
- (Right-) censoring occurs when we observe $Y = Y^*$ if $Y^* \leq C$ and we observe $Y = C$ for the individuals with $Y^* > C$.
- We will consider both the case where C varies across individuals and the case where it is a constant.

Truncation

- Let Y^* denote the outcome of interest.
- Truncation occurs when we observe $Y = Y^*$ if $Y^* \leq C$ and we don't observe the individuals with $Y^* > C$ at all (as in the Christensen and Kiefer model).
- Again, C may or may not vary across individuals.

Censoring

- Consider a sample of durations $(y_1, \ldots, y_n)$ and covariates $(x_1, \ldots, x_n)$.
- Suppose the conditional density for Y^* is given by $f(y \mid x, \theta)$.
- If $y_i = y_i^*$ for all i then the likelihood is simply $\mathcal{L}(\theta) = \sum_{i=1}^{n} \log(f(y_i \mid x_i, \theta))$.
- What if some observations are censored?

Censoring

- Non-random censoring:
  - The likelihood should be the distribution of what we observe.
  - Here we observe both Y_i and $D_i = \mathbf{1}(Y_i^* \leq C)$.

Censoring

- Non-random censoring:
  - The likelihood should be the distribution of what we observe.
  - Here we observe both Y_i and $D_i = \mathbf{1}(Y_i^* \leq C)$.
  - If $d_i = 1$ then $Pr(Y_i = y_i, D_i = d_i \mid X_i) = Pr(Y_i^* = y_i, Y_i^* \leq C \mid X_i) = Pr(Y_i^* = y_i \mid X_i)$.
  - If $d_i = 0$ then $y_i = C$ and $Pr(Y_i = y_i, D_i = d_i \mid X_i) = Pr(Y_i^* > C \mid X_i)$.

Censoring

- Non-random censoring:
  - The likelihood should be the distribution of what we observe.
  - Here we observe both Y_i and $D_i = \mathbf{1}(Y_i^* \leq C)$.
  - If $d_i = 1$ then $Pr(Y_i = y_i, D_i = d_i \mid X_i) = Pr(Y_i^* = y_i, Y_i^* \leq C \mid X_i) = Pr(Y_i^* = y_i \mid X_i)$.
  - If $d_i = 0$ then $y_i = C$ and $Pr(Y_i = y_i, D_i = d_i \mid X_i) = Pr(Y_i^* > C \mid X_i)$.
  - So the log-likelihood is given by

$$\sum_{i=1}^{n} D_i \ln(f(y_i \mid x_i, \theta)) + (1 - D_i) \ln \left(\int_C^{\infty} f(y \mid x_i, \theta) dy \right)$$

Censoring

- Random censoring:
  - Suppose censoring times are random, C_i, with distribution $f_{C|X}(c \mid x, \theta)$.
  - Assume that Y_i^* and C_i are independent conditional on X_i.

Censoring

- Random censoring:
  - Suppose censoring times are random, C_i, with distribution $f_{C|X}(c \mid x, \theta)$.
  - Assume that Y_i^* and C_i are independent conditional on X_i.
  - If $d_i = 1$ then

$$\begin{aligned} Pr(Y_i = y_i, D_i = d_i \mid X_i) &= Pr(Y_i^* = y_i, Y_i^* \leq C_i \mid X_i) \\ &= Pr(Y_i^* = y_i, C_i \geq y_i \mid X_i) \\ &= f(y_i \mid x_i, \theta) \int_{y_i}^{\infty} f_C(y \mid x_i, \theta) dy \end{aligned}$$

Censoring

- Random censoring:
  - Suppose censoring times are random, C_i, with distribution $f_{C|X}(c \mid x, \theta)$.
  - Assume that Y_i^* and C_i are independent conditional on X_i.
  - If $d_i = 1$ then

$$\begin{aligned} Pr(Y_i = y_i, D_i = d_i \mid X_i) &= Pr(Y_i^* = y_i, Y_i^* \leq C_i \mid X_i) \\ &= Pr(Y_i^* = y_i, C_i \geq y_i \mid X_i) \\ &= f(y_i \mid x_i, \theta) \int_{y_i}^{\infty} f_C(y \mid x_i, \theta) dy \end{aligned}$$

  - If $d_i = 0$ then $y_i = C_i$ and

$$\begin{aligned} Pr(Y_i = y_i, D_i = d_i \mid X_i) &= Pr(C_i = y_i, Y_i^* > C_i \mid X_i) \\ &= f_C(y_i \mid x_i, \theta) \int_{y_i}^{\infty} f(y \mid x_i, \theta) dy \end{aligned}$$

Truncation

- non-random censoring:
  - The likelihood should be the distribution of what we observe, *conditional on being observed*.
  - This is usually implicit.

Truncation

- non-random censoring:
  - The likelihood should be the distribution of what we observe, *conditional on being observed*.
  - This is usually implicit.
  - Here,

$$
\begin{aligned}
Pr(Y_i = y_i \mid D_i = 1, x_i) &= Pr(Y_i^* = y_i \mid Y_i^* \leq C, x_i) \\
&= \frac{Pr(Y_i^* = y_i, Y_i^* \leq C \mid x_i)}{Pr(Y_i^* \leq C \mid x_i)} \\
&= \frac{f(y_i \mid x_i, \theta)}{\int_{-\infty}^{C} f(y \mid x_i, \theta) dy}
\end{aligned}
$$

Truncation

- non-random censoring:
  - The likelihood should be the distribution of what we observe, *conditional on being observed*.
  - This is usually implicit.
  - Here,

$$
\begin{aligned}
Pr(Y_i = y_i \mid D_i = 1, x_i) &= Pr(Y_i^* = y_i \mid Y_i^* \leq C, x_i) \\
&= \frac{Pr(Y_i^* = y_i, Y_i^* \leq C \mid x_i)}{Pr(Y_i^* \leq C \mid x_i)} \\
&= \frac{f(y_i \mid x_i, \theta)}{\int_{-\infty}^{C} f(y \mid x_i, \theta) dy}
\end{aligned}
$$

  - So the log-likelihood is

$$
\sum_{i=1}^{n} \log(f(y_i \mid x_i, \theta)) - \log\left(\int_{-\infty}^{C} f(y \mid x_i, \theta) dy\right)
$$

Truncation

- random censoring:
  - Now we get

$$
\begin{aligned}
Pr(Y_i = y_i \mid D_i = 1, x_i) &= Pr(Y_i^* = y_i \mid Y_i^* \leq C_i, x_i) \\
&= \frac{Pr(Y_i = y_i, Y_i^* \leq C_i \mid x_i)}{Pr(Y_i^* \leq C_i \mid x_i)} \\
&= \frac{f(y_i \mid x_i, \theta)(1 - F_C(y_i \mid x_i, \theta))}{\int_{-\infty}^{\infty} f(y \mid x_i, \theta)(1 - F_C(y \mid x_i, \theta))dy}
\end{aligned}
$$

Identification

- An important part of structural modeling: determining model identification
  - Just because we can write down a likelihood function does not mean the model is identified.
  - Consider the random censoring model:
    - suppose we assume instead that $D_i = \mathbf{1}(Y_i^* \leq C_i + c_0)$
    - we can add a constant to c_0 and shift the density of C_i by the same constant without changing the likelihood function
    - so the model is not identified!
  - We will give some more examples of this next week.