Truncated, Discrete-Continuous and Stratified Responses

Truncated versus Censored Data

Models for Discrete-Continuous Responses

- Tobit Model
- Two-Part Model
- Sample Selection Model

Exogenous / Endogenous Stratification

Truncated versus Censored Data

- **Truncated data**: a subsample associated to particular responses is missing from the data (data is missing on (Y, X))

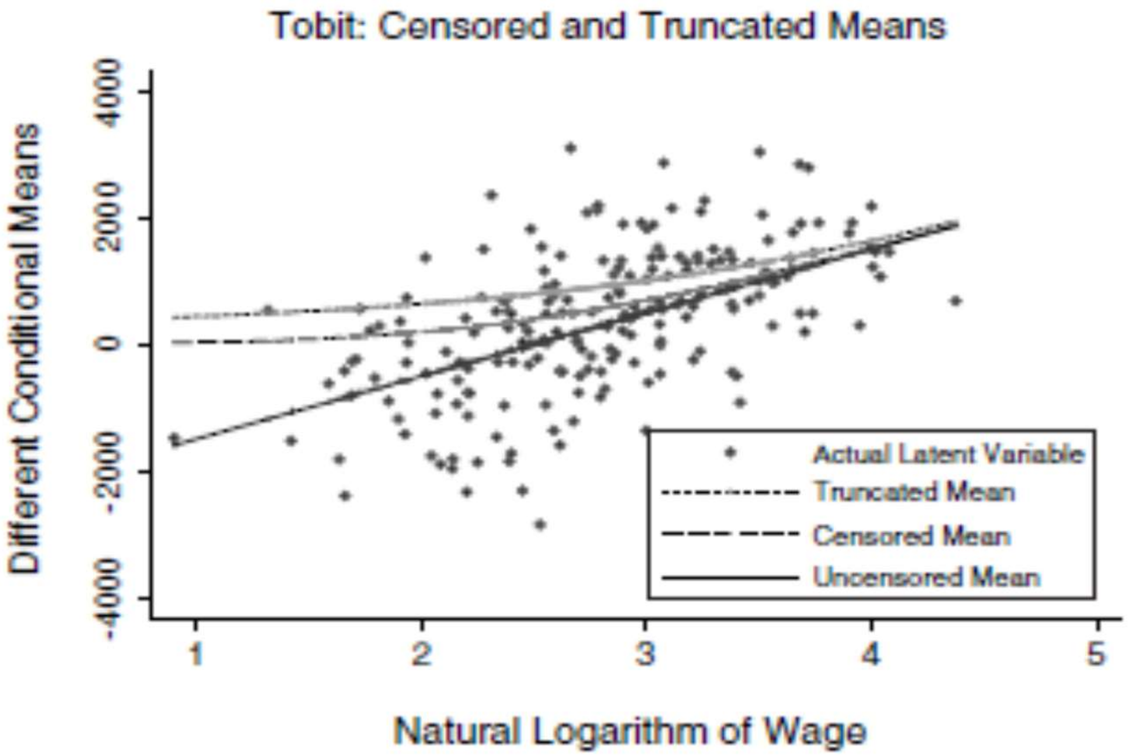
- **Censored data**: for a subsample associated to particular responses only X is observed. Y assumes a particular fixed value, say L, reflecting the fact that is not observed. For these observations we have (L, X)

Example: expenditures on alcohol at household level

- Truncated data: only households with positive expenditures are observed
- Censored data: all households are included in the sample, but zero expenditures (tend to present an accumulation) may correspond to families that would present negative expenditures. For this subsample covariates are observed, but the dependent variable is not, only the 0 is recorded

Truncated versus Censored Data

- Figure CT, p. 531



Truncated versus Censored Data

- Figure Greene (), p. 835

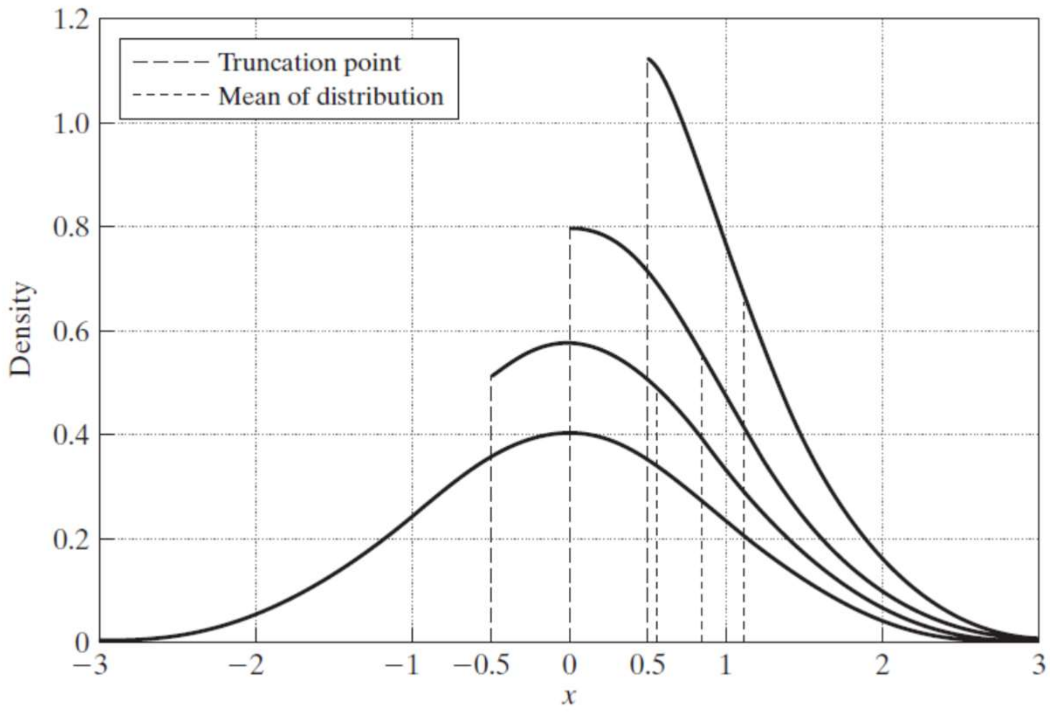


FIGURE 19.1 Truncated Normal Distributions.

Truncated versus Censored Data

Formal framework:

Assume that the aim is modelling a latent variable Y^*, but in fact the observed variable is Y

- Truncated data: (Y, X) are observed if $Y^* > \mathrm{L}$

$$Y = Y^* \; if \; Y^* > L$$

described by $f(y|x, y > L)$

- Censored data: X is available for all sample but Y^* is only observed if $Y^* > \mathrm{L}$

$$Y = \begin{cases} Y^* \; if \; Y^* > L \\ L \; if \; Y^* \leq L \end{cases}$$

described by $\begin{cases} f(y|x, y > L) \; if \quad y > L \\ F(L|x) \; if \quad\; y \leq L \end{cases}$, where the second line adds information on X relative to the truncated case

Truncated versus Censored Data

Example for truncated data – zero truncated Poisson:

- $Pr(Y_i = y|x_i, y_i > 0) = \frac{Pr(Y_i=y|x_i)}{Pr(Y_i>0|x_i)} = \frac{Pr(Y_i=y|x_i)}{1-Pr(Y_i=0|x_i)} = \frac{e^{-\lambda_i}{\lambda_i}^y}{y!(1-e^{-\lambda_i})}$

where $\lambda_i = E(Y|X) = exp(x'\beta)$

Stata
tpoisson Y X_1 ... X_k, ll(0)

- ML estimation
  - QML no longer available
  - It is possible to compute probabilities and expected values

Conditional on X	Conditional on X and Y>0
$E(Y\|X) = \lambda_i$	$E(Y\|X, Y > 0) = \frac{\lambda_i}{(1 - e^{-\lambda_i})}$
$Pr(Y = y\|X) = \frac{e^{-\lambda_i}{\lambda_i}^y}{y!}$	$Pr(Y = y\|X, Y > 0) = \frac{e^{-\lambda_i}{\lambda_i}^y}{y!\,(1 - e^{-\lambda_i})}$

Truncated versus Censored Data

Consider the file "CameronTrivedi2010-ch18-health.dta" used in illustration 1. Keep year=1 and compare the mean of mdu for all data and for the positive values

```
. drop if year!=1
(14,548 observations deleted)

. sum mdu

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         mdu |       5638    2.877971    4.332918          0         69


. sum mdu if mdu>0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         mdu |      3,909    4.150934    4.668514          1         69
```

Note the increase in the mean

Truncated versus Censored Data

Using all the data: ML Poisson

```
. poisson mdu lcoins ndisease female age lfam child
…
Poisson regression                              Number of obs     =      5,638
                                                LR chi2(6)        =    2296.50
                                                Prob > chi2       =     0.0000
Log likelihood = -17173.058                     Pseudo R2         =     0.0627
------------------------------------------------------------------------------
         mdu |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |  -.0682563   .0038214   -17.86   0.000    -.0757462   -.0607665
    ndisease |     .03538   .0010286    34.40   0.000     .0333639     .037396
      female |   .1276798   .0164078     7.78   0.000     .0955211    .1598386
         age |   .0041167   .0007841     5.25   0.000     .0025798    .0056536
        lfam |   -.141855   .0154716    -9.17   0.000    -.1721789   -.1115312
       child |    .056673   .0278998     2.03   0.042     .0019903    .1113557
       _cons |   .7463066   .0385653    19.35   0.000       .67072    .8218932
------------------------------------------------------------------------------
```

Truncated versus Censored Data

Droping 0's:

```
. poisson mdu lcoins ndisease female age lfam child if mdu>0
…
Poisson regression                              Number of obs     =      3,909
                                                LR chi2(6)        =     902.89
                                                Prob > chi2       =     0.0000
Log likelihood = -11927.131                     Pseudo R2         =     0.0365
------------------------------------------------------------------------------
         mdu |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |  -.0305899   .0038326    -7.98   0.000    -.0381017   -.0230781
    ndisease |   .0248502   .0010552    23.55   0.000     .0227821    .0269184
      female |   .0490165   .0164567     2.98   0.003     .0167619    .0812711
         age |   .0029178   .0007854     3.72   0.000     .0013786    .0044571
        lfam |  -.0876082   .0161344    -5.43   0.000    -.1192312   -.0559853
       child |   .0546617   .0282745     1.93   0.053    -.0007554    .1100787
       _cons |   1.149892   .0397799    28.91   0.000     1.071925    1.227859
------------------------------------------------------------------------------
```

Truncated versus Censored Data

Accounting for truncation at 0:

```
. drop if mdu==0
(1,729 observations deleted)

. tpoisson mdu lcoins ndisease female age lfam child, ll(0)

(…)

Truncated Poisson regression                    Number of obs     =      3,909
Truncation point: 0                             LR chi2(6)        =     961.47
                                                Prob > chi2       =     0.0000
Log likelihood = -11835.785                     Pseudo R2         =     0.0390

------------------------------------------------------------------------------
         mdu |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lcoins |  -.0331493   .0039933    -8.30   0.000     -.040976   -.0253226
    ndisease |   .0261598   .0010833    24.15   0.000     .0240366    .0282829
      female |   .0540894   .0171947     3.15   0.002     .0203885    .0877903
         age |   .0030495   .0008127     3.75   0.000     .0014567    .0046423
        lfam |  -.0936962   .0167247    -5.60   0.000     -.126476   -.0609163
       child |   .0583544    .029489     1.98   0.048     .0005571    .1161517
       _cons |    1.11516   .0414084    26.93   0.000     1.034001    1.196319
------------------------------------------------------------------------------
```

Models for Discrete-Continuous Responses

Motivation:

- Describes cases where the dependent variable has both discrete and continuous values; typically:
  - Discrete value: for many individuals, $Y_i = 0$
  - Continuous component: for the remaining individuals, Y_i may take on some positive value

- Examples:
  - Expenditures on durable goods, alcohol,,...
  - Work hours

Models for Discrete-Continuous Responses

Alternative models:

- Tobit model: a single model explains all values
- Two-part model: uses two independent models for explaining separately the zeros and the positive values
- Sample selection model: uses two different models, but interdependent, for explaining the zeros and the positive values

Model specification:

- Latent model: $Y_i^* = x_i'\beta + u_i$
- Instead of Y_i^*, it is observed:

$$Y_i = \begin{cases} 0 & \text{if } Y_i^* \leq 0 \\ Y_i^* & \text{if } Y_i^* > 0 \end{cases}$$

- Assumption: $u_i \sim N(0, \sigma^2)$
  - $Pr(Y_i = 0|x_i) = Pr(Y_i^* \leq 0|x_i) = Pr(x_i'\beta + u_i \leq 0|x_i) = Pr(u_i \leq -x_i'\beta|x_i) = Pr\left(\frac{u_i}{\sigma} \leq -\frac{x_i'\beta}{\sigma}\middle| x_i\right) = \Phi\left(-\frac{x_i'\beta}{\sigma}\right) = 1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)$
  - Hence: $f(y_i|x_i) = \begin{cases} 1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right) & \text{if } \mathrm{Y} = 0 \\ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}} & \text{if } Y > 0 \end{cases}$

Models for Discrete-Continuous Responses
Tobit Model

Estimation:

- Method: ML

Stata
tobit $YX_1 \dots X_k$, ll(0)

- Parameters to be estimated: β and σ

- Log-likelihood function:

$$LL = \sum \left\{ (1-d_i) log \left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right) \right] + d_i log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}} \right] \right\}$$

where $d_i = \begin{cases} 0 \text{ if } Y_i = 0 \\ 1 \text{ if } Y_i > 0 \end{cases}$

Adding and subtracting $d_i log\left[\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]$, it is clear that the LL function combines that of a binary model (first line) and a truncated model (second line):

$$LL =$$

$$= \sum\left\{(1-d_i)log\left[1-\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] + d_i log\left[\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]\right.$$

$$\left. + d_i log\left[\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}}\right] - d_i log\left[\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]\right\}$$

Models for Discrete-Continuous Responses

Tobit Model

Quantities of interest:

- Conditional mean given that y_i is positive:

$$E(Y_i|x_i, Y_i > 0) = x_i'\beta + \sigma\lambda\left(\frac{x_i'\beta}{\sigma}\right)$$

where $\lambda\left(\frac{x_i'\beta}{\sigma}\right) = \frac{\phi\left(\frac{x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta}{\sigma}\right)}$ is the Mills ratio

- Probability of observing positive values for y_i:

$$\Pr(Y_i > 0|x_i) = \Phi\left(\frac{x_i'\beta}{\sigma}\right)$$

- Overall conditional mean:

$$E(Y_i|x_i) = Pr(Y_i = 0|x_i)E(Y_i|x_i, Y_i = 0) + \Pr(Y_i > 0|x_i)E(Y_i|x_i, Y_i > 0)$$
$$= \Phi\left(\frac{x_i'\beta}{\sigma}\right)x_i'\beta + \sigma\phi\left(\frac{x_i'\beta}{\sigma}\right)$$

Partial effects:

- $\Delta X_j = 1 \Longrightarrow$
  - $\Delta E(Y_i|x_i, Y_i > 0) = \beta_j \left\{1 - \lambda\left(\frac{x_i'\beta}{\sigma}\right)\left[\frac{x'_i\beta}{\sigma} + \lambda\left(\frac{x_i'\beta}{\sigma}\right)\right]\right\}$
  - $\Delta Pr(Y_i > 0|x_i) = \frac{\beta_j}{\sigma}\phi\left(\frac{x_i'\beta}{\sigma}\right)$
  - $\Delta E(Y_i|x_i\ ) = \beta_j\Phi\left(\frac{x_i'\beta}{\sigma}\right)$
- The three effects have the same sign

Model specification:

- First part – binary regression model:

$$Pr(d_i = 1|x_i) = G_1(x_i'\beta)$$

  - $d_i = \begin{cases} 0 \text{ se } Y_i = 0 \\ 1 \text{ se } Y_i > 0 \end{cases}$

- Second part – exponential or fractional regression model

$$E(Y_i|x_i, d_i = 1) = G_2(x_i'\theta)$$

- Overall conditional mean:

$$E(Y_i|x_i\ ) \\ = Pr(Y_i = 0|x_i)E(Y_i|x_i, Y_i = 0) + \Pr(Y_i > 0|x_i)E(Y_i|x_i, Y_i > 0) \\ = G_1(x_i'\beta)G_2(x_i'\theta)$$

Estimation:

- Each part of the model is estimated separately:
  - In each part, use the standard methods for the type of data being analyzed
  - In the first part of the model, use the full sample
  - In the second part of the model, use the subsample for which $Y_i > 0$
  - One may use different explanatory variables in each part of the model

Partial effects:

- $\Delta Pr(d_i = 1|x_i) = \Delta Pr(Y_i > 0|x_i)$
- $\Delta E(Y_i|x_i, d_i = 1) = \Delta E(Y_i|x_i, Y_i > 0)$
- $\Delta E(Y_i|x_i) = \Delta Pr(d_i = 1|x_i) E(Y_i|x_i, d_i = 1) + Pr(d_i = 1|x_i) \Delta E(Y_i|x_i, d_i = 1)$

Latent variable:

- Y_{2i}^*: main variable
- Y_{1i}^*: variable that determines whether Y_{2i}^* is observed or not

Two equations:

- Participation equation (*e.g.* to work or not):

$$Y_{1i} = \begin{cases} 0 \text{ if } Y_{1i}^* \leq 0 \\ 1 \text{ if } Y_{1i}^* > 0 \end{cases}$$

- Outcome equation (*e.g.* how much to work):

$$Y_{2i} = \begin{cases} - & \text{if } Y_{1i}^* \leq 0 \\ Y_{2i}^* & \text{if } Y_{1i}^* > 0 \end{cases}$$

Latent linear models:

$$\begin{cases} Y_{1i}^* = x_{1i}'\beta_1 + u_{1i} \\ Y_{2i}^* = x_{2i}'\beta_2 + u_{2i} \end{cases}$$

Assumptions:

- The error terms of the two equations are assumed to be correlated, having a bivariate normal distribution:

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim N\left\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right\}$$

- Only when $\sigma_{12} = 0$ the two equations will be independent (the selection mechanism is exogenous or ignorable):
  - The second equation may be estimated by OLS using only the observed data

Quantities of interest:

- Conditional mean of the main latent variable:

$$E(Y_{2i}^{*}|x_i) = x_{2i}'\beta_2$$

- Conditional mean of the main observed dependent variable:

$$E(Y_{2i}|x_i, Y_{1i} = 1) = x_{2i}'\beta_2 + \sigma_{12}\lambda(x_{1i}'\beta_1)$$

- Probability of observing positive values:

$$Pr(Y_{1i}^{*} > 0|x_i) = Pr(Y_{1i} = 1|x_i) = \Phi(x_{1i}'\beta_1)$$

Parameters to be estimated: $\beta, \sigma_{12}, \sigma_2$

Estimation methods:

- ML
- Heckman's two-step method

ML:

- Based on the following log-likelihood function:

$$LL = \sum \{(1 - d_i)\Pr(Y_{1i} = 0|x_{1i}) + d_i[f(Y_{1i} = 1|Y_{2i}) + f(Y_{2i})]\}$$

Stata
heckman Y_2 X_1 ... X_k, select(Y_1 X_1 ... X_k)

Heckman's two-step method:

- Based on $E(Y_{2i}|x_i, Y_{1i} = 1) = x'_{2i}\beta_2 + \sigma_{12}\lambda(x'_{1i}\beta_1)$
- First step: estimate the probit model $Pr(Y_{1i} = 1|x_i) = \Phi(x'_{1i}\beta_1)$ and get $\lambda\left(x'_{1i}\hat{\beta}_1\right) = \frac{\phi(x'_{1i}\hat{\beta}_1)}{\Phi(x'_{1i}\hat{\beta}_1)}$
- Second step: regress Y_{2i} on x_{2i} and $\lambda\left(x'_{1i}\hat{\beta}_1\right)$ using only individuals fully observed and OLS, and correct the variances
- t test for $\mathrm{H}_0: \sigma_{12} = 0$ (exogenous selection mechanism)
- If the same regressors are used in both steps, multicolinearity may arise; to avoid it, it is usual to **exclude from $\boldsymbol{x_{2i}}$ some of the variables included in $\boldsymbol{x_{1i}}$**

Stata
heckman $Y_2\ X_1\ \ldots\ X_k$, twostep select($Y_1\ X_1\ \ldots\ X_k$)

Endogenous Stratification
Exogenous / Endogenous Sampling

- Endogenous stratification: the probability of observing a sampling unit depends on the response variable
  - Choice-based sampling (binary case): endogenous sampling for $Y \in \{0,1\}$
  - Motivation: promoting efficiency gains by inflating the proportion of rare cases
    - Example: assume Y=1 for those travelling from city A to B by plane

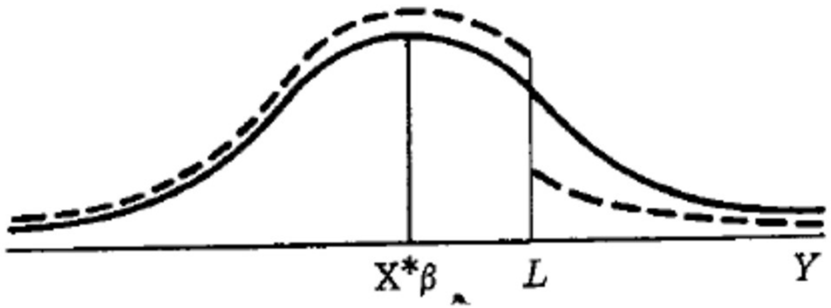
Mode	Proportion in the population	Proportion in the sample
Y=1	Q=0.1	H=0.5
Y=0	1-Q=0.9	1-H=0.5

- Exogenous stratification: the probability of observing a sampling unit depends on one or more explanatory variables

Endogenous Stratification

Exogenous / Endogenous Sampling

- Illustration of stratification for a normal distributed variable, Hausman and Wise (1981)



Endogenous Stratification

Exogenous / Endogenous Sampling

- Under RS: joint density of y and x

$$f(y,x) = f(y|x,\theta)f(x)$$

Analysis is undertaken conditional on x, based on $f(y|x,\theta)$

- Consider S strata, s=1, 2, …S, defined on $\mathscr{Y}_s$ and/or $\mathscr{X}_s$ with
  - Proportion in the sample: H_s
  - Proportion in the population: Q_s

$$Q_s = \int_{\mathscr{X}_s}\int_{\mathscr{Y}_s} f(y|x,\theta)f(x)\,dydx$$

- Sample: one observes (y,x,s) according to

$$h(y,x,s) = \frac{H_s}{Q_s} f(y|x,\theta)f(x)$$

Endogenous Stratification

Exogenous / Endogenous Sampling

- The sampling density of x

$$h(x) = \sum_S \int_{\mathscr{Y}_s} \frac{H_s}{Q_s} f(y|x,\theta) f(x)\, dy$$

$$= f(x) \sum_S \int_{\mathscr{Y}_s} \frac{H_s}{Q_s} f(y|x,\theta)\, dy$$

Is a function of θ: x no longuer exogenous

- With exogenous stratification - simplification:

$$h(x) = f(x) \sum_S \frac{H_s}{Q_s} \int_{\mathscr{Y}} f(y|x,\theta)\, dy = f(x) \sum_S \frac{H_s}{Q_s}$$

with $Q_s = \int_{\mathscr{X}_s} \int_{\mathscr{Y}} f(y|x,\theta) f(x)\, dy dx = \int_{\mathscr{X}_s} f(x) dx$ not a function of θ.

The analysis may be udertaken conditional on x

Dealling with choice-based sampling

- If Pr(y|x) is specified as logit, with a constant term, the standard ML estimator of slope parameters is still consistent. The corresponding standard deviations are correct
- For other models: with a known proportion Q, the likelihoood function is re-weighted

$$LL = \sum_{i=1}^{N} \left\{ y_i \frac{Q_i}{H_i} \ln[G(x_i'\beta)] + (1 - y_i) \frac{1 - Q_i}{1 - H_i} \ln[1 - G(x_i'\beta)] \right\}$$