Regression in Complex Surveys

- Learn about relationships between variables
- Give more accurate estimates of population means and totals

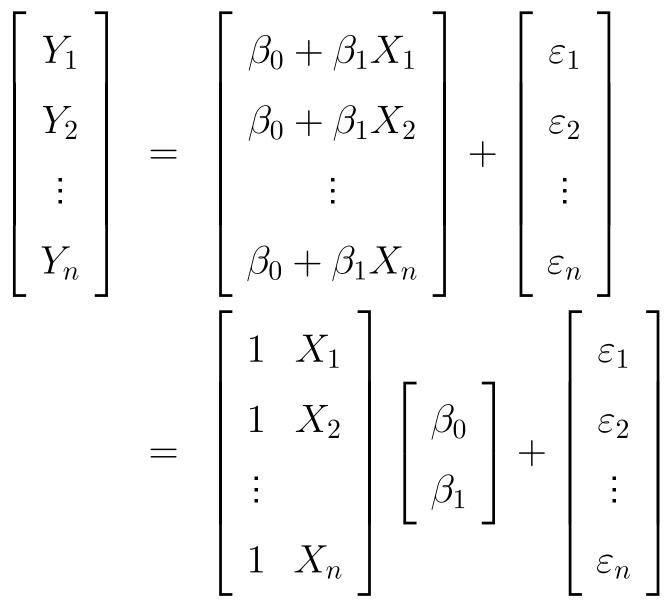
Review Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Y_i is a random variable for the response
- x_i is an explanatory variable
- β_0 and β_1 are unknown parameters
- ϵ_i 's are the deviations of the response variable about the line described by the model
- $E[\epsilon_i] = 0$, or $E[Y_i|x_i] = \beta_0 + \beta_1 x_i$
- $V[\epsilon_i] = \sigma^2$

- $Cov[\epsilon_i, \epsilon_j] = 0$ for $i \neq j$
- Often, conditionally on the x_i 's, ϵ_i 's are independent and identically distributed from a normal distribution with mean 0 and variance σ^2

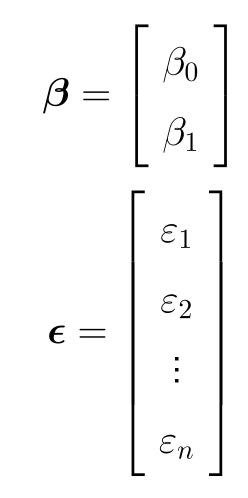
The SLR model in Matrix Form



where

- $\bullet~\mathbf{X}$ is called the design matrix
- $oldsymbol{eta}$ is the vector of parameters
- ϵ is the error vector
- ullet \mathbf{Y} is the response vector

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots \\ 1 & X_n \end{bmatrix}$$



$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

SLM in Matrix Form

$$\mathbf{Y} = \mathbf{X} \boldsymbol{eta} + \boldsymbol{\epsilon}.$$

The ordinary Least Squares (OLS) estimates

• Measure
$$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Minimize Q to find estimates b_0 and b_1 for β_0 and β_1
- Normal equations

$$\beta_0 n + \beta_1 \sum x_i = \sum y_i$$
$$\beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i$$

$$b_1 = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$
$$b_0 = \frac{\sum y_i - b_1 \sum x_i}{n}$$

• b_0 and b_1 are the best linear unbiased estimates

Inferences

$$b_1 \sim N(\beta_1; \sigma^2(b_1))$$
where $\sigma^2(b_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$

$$\hat{\sigma}^2(b_1) = s^2(b_1) = \frac{\mathsf{MSE}}{\sum (x_i - \bar{x})^2}$$

$$\frac{b_1 - \beta_1}{s(b_1)} \sim t(n-2)$$
Confidence interval $b_1 \pm t(1 - \frac{\alpha}{2}, n-2) s(b_1)$

Confidence Interval $b_1 \pm t(1 - \frac{\pi}{2}, n - 2)s(b_1)$

$$b_0 \sim N(\beta_0; \sigma^2(b_0))$$

where $\sigma^2(b_0) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$
 $\hat{\sigma}^2(b_0) = s^2(b_0) = \frac{\text{MSE} \sum x_i^2}{n \sum (x_i - \bar{x})^2}$
 $\frac{b_0 - \beta_0}{s(b_0)} \sim t(n-2)$

Confidence Interval $b_0 \pm t(1 - \frac{\alpha}{2}, n - 2)s(b_0)$

Regression in Complex Survey

- Estimating quantities from a finite population
- The finite population quantities of interest for regression are the least squares coefficients for the population, B_0 and B_1 , that minimize

$$\sum_{i=1}^{N} (y_i - B_0 - B_1 x_i)^2$$

over the entire population

- Observations may have different probabilities of selection, π_i . If the probability of selection is related to the response variable y_i , then an analysis that does not account for the different probabilities of selection may lead to biases in the estimated regression parameters.
- Nonrespondents, who may be thought of as having zero probability of selection, can distort the relationship
- Stratification may also need to be taken into account

Normal equations

$$B_0 N + B_1 \sum_{i=1}^N x_i = \sum_{i=1}^N y_i$$
$$B_0 \sum_{i=1}^N x_i + B_1 \sum_{i=1}^N x_i^2 = \sum_{i=1}^N x_i y_i$$

Estimates of B_1 and B_0

$$\hat{B}_{1} = \frac{\sum_{i=1}^{N} x_{i}y_{i} - (\sum_{i=1}^{N} x_{i})(\sum_{i=1}^{N} y_{i})/N}{\sum_{i=1}^{N} x_{i}^{2} - (\sum_{i=1}^{N} x_{i})^{2}/N}$$

$$= \frac{t_{xy} - t_{x}t_{y}/N}{t_{x^{2}} - (t_{x})^{2}/N}$$

$$= \frac{\sum_{i\in S} w_{i}x_{i}y_{i} - (\sum_{i\in S} w_{i}x_{i})(\sum_{i\in S} w_{i}y_{i})/\sum_{i\in S} w_{i}}{\sum_{i\in S} w_{i}x_{i}^{2} - (\sum_{i\in S} w_{i}x_{i})^{2}/\sum_{i\in S} w_{i}}$$

$$\hat{B}_0 = \frac{\sum_{i=1}^N y_i - \hat{B}_1 \sum_{i=1}^N x_i}{N}$$

$$= \frac{\frac{t_y - \hat{B}_1 t_x}{N}}{\sum_{i \in S} w_i y_i - \hat{B}_1 \sum_{i \in S} w_i x_i}{\sum_{i \in S} w_i}$$

Standard Errors

• An approximate 100(1-lpha)% confidence interval for B_1 is

$$\hat{B}_1 \pm t_{\alpha/2} \sqrt{\hat{V}(\hat{B}_1)}$$

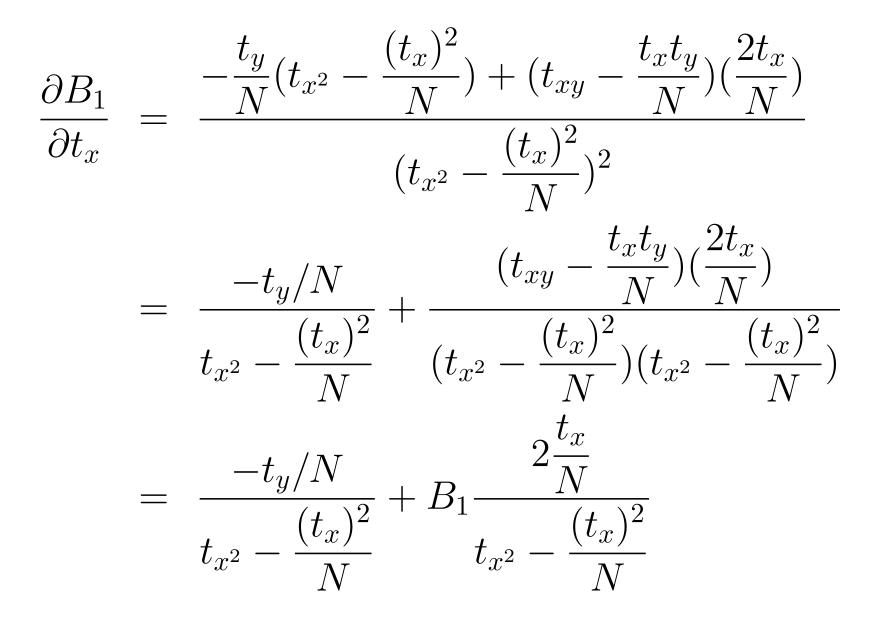
- For linearization, jackknife, Or BRR in a stratified multistage sample, we would use (number of sampled psu's) - (number of strata) as the degrees of freedom
- Random group method of estimating the variance, the appropriate degrees of freedom would be (number of groups)-1

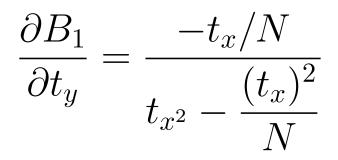
Standard Errors Using Linearization for \hat{B}_1

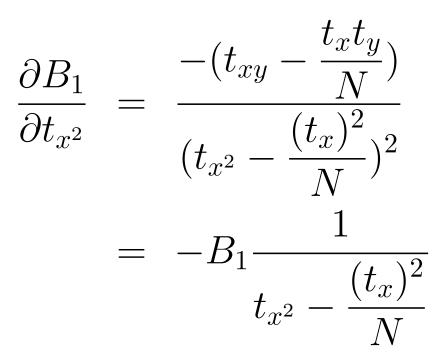
- B_1 is a function of four population totals t_{xy}, t_x, t_y , and t_{x^2} .
- Using Taylor expansion,

$$V(\hat{B}_{1}) \approx V\{\frac{\partial B_{1}}{\partial t_{xy}}(\hat{t}_{xy} - t_{xy}) + \frac{\partial B_{1}}{\partial t_{x}}(\hat{t}_{x} - t_{x}) + \frac{\partial B_{1}}{\partial t_{y}}(\hat{t}_{y} - t_{y}) + \frac{\partial B_{1}}{\partial t_{x^{2}}}(\hat{t}_{x^{2}} - t_{x^{2}})\}$$

$$\frac{\partial B_1}{\partial t_{xy}} = \frac{1}{t_{x^2} - \frac{(t_x)^2}{N}}$$







$$\begin{split} &V(\hat{B}_{1}) \\ &= V\left[[t_{x^{2}} - \frac{(t_{x})^{2}}{N}]^{-1} \{ \hat{t}_{xy} - \hat{t}_{y} \frac{t_{x}}{N} - B_{0} \hat{t}_{x} + B_{0} t_{x} - B_{1} \hat{t}_{x^{2}} + B_{1} \frac{\hat{t}_{x} t_{x}}{N} \} \right] \\ &= V\left[[t_{x^{2}} - \frac{(t_{x})^{2}}{N}]^{-1} \sum_{i \in S} w_{i} (y_{i} - B_{0} - B_{1} x_{i}) (x_{i} - \frac{t_{x}}{N}) \right] \\ \\ &\mathsf{Define} \end{split}$$

$$q_i = (y_i - \hat{B}_0 - \hat{B}_1 x_i)(x_i - \hat{\bar{x}})$$

where $\hat{\bar{x}} = \hat{t}_x / \hat{N}$.

$$\hat{V}(\sum_{i \in S} w_i q_i)$$

$$\frac{\hat{V}(\sum_{i \in S} w_i q_i)}{\left[\sum_{i \in S} w_i x_i^2 - \left(\sum_{i \in S} w_i x_i\right)^2 / \sum_{i \in S} w_i\right]^2}$$

Consider simple random sampling

$$\hat{V}(\sum_{i \in S} w_i q_i) = \hat{V}(\hat{t}_q) = \frac{N^2 s_q^2}{n}$$

$$s_q^2 = \frac{\sum_{i \in S} (x_i - \bar{x}_S)^2 (y_i - \hat{B}_0 - \hat{B}_1 x_i)^2}{n - 1}$$

$$\hat{V}_L(\hat{B}_1) = \frac{n \sum_{i \in S} (x_i - \bar{x}_S)^2 (y_i - \hat{B}_0 - \hat{B}_1 x_i)^2}{(n - 1) [\sum_{i \in S} (x_i - \bar{x}_S)^2]^2}$$

$$\hat{V}_M(\hat{\beta}_1) = \frac{\sum_{i \in S} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{(n - 2) \sum_{i \in S} (x_i - \bar{x}_S)^2}$$

Design-based v.s Model-based variance estimator

- Design based estimator of the variance \hat{V}_L comes from the selection probabilities of the design
- V_M comes from the average squared deviation over all possible realizations of the model

• For $\hat{B}_1 \pm t_{\alpha/2} \sqrt{\hat{V}_L(\hat{B}_1)}$, the confidence level is $\sum u(S)P(S)$, where the sum is over all possible samples S that can be selected using the sampling design, P(S) is the probability that sample S is selected, u(S) = 1, if the confidence interval constructed from sample S contains the population character-

istic B_1 and u(S) = 0 otherwise.

- In an SRS, The design-based confidence level is the proportion of possible samples that result in a confidence interval that includes B_1 , from the set of all SRS's of size n from the finite population of fixed values $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$.
- For the model-based confidence interval $\beta_1 \pm t_{\alpha/2} \sqrt{\hat{V}_M(\hat{\beta}_1)}$, the confidence level is the expected proportion of confidence intervals that will include β_1 , from the set of all samples that could be generated from the model

Standard errors using jackknife

• Stratified multistage cluster sample, the jackknife can be applied separately in each stratum at the first stage of sampling, with one psu deleted at a time Suppose there are H strata, from each stratum h, n_h psu's are sampled. w_i 's are the original weight. Define a new weight variable:

$$w_{i(hj)} = \begin{cases} w_i, & \text{unit } i \text{ is not in stratum } h, \\ 0, & \text{unit } i \text{ is in psu } j \text{ of stratum } h, \\ \frac{n_h}{n_h - 1} w_i, & \text{unit } i \text{ is in stratum } h \text{ but not in psu } j. \end{cases}$$

Then use the weights $w_{i(hj)}$ to calculate $\hat{B}_{1(hj)}$, the jackknife estimator is defined as follows:

$$\hat{V}_{JK}(\hat{B}_1) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (\hat{B}_{1(hj)} - \hat{B}_1)^2.$$

Example: Consider the two samples of size 200 from the 3,000 criminal. For SRS, $w_i = 3000/200$, so $w_{i(j)} = 200w_i/199 = 3000/199$ for $i \neq j$. For the unequal probability sample, $w_i = 1/\pi_i$, so $w_{i(j)} = 200w_i/199$ for $i \neq j$.

| | estimates | variance | variance |
|----------------|-----------|----------------|-----------------|
| SRS | 3.0453 | $V_{L} = .048$ | $V_{JK} = .050$ |
| Unequal sample | 3.055 | $V_{L} = .346$ | $V_{JK} = .461$ |

• The jackknife estimated variance is larger than the linearization variance, as often occurs in practice.

Multiple Linear Regression

$$Y_i = X_{i,1}B_0 + B_1X_{i,2} + B_2X_{i,3} + \dots + B_{p-1}X_{i,p} + \epsilon_i,$$

where

- Multiple–More than one predictor variable
- Y_i is the response variable
- $X_{i,1}, X_{i,2}, \cdots X_{i,p}$ are the p explanatory variables for cases i = 1 to N.

Let

$$\mathbf{x}_{i}^{T} = \begin{bmatrix} x_{i1}, x_{i2}, \cdots, x_{ip} \end{bmatrix}$$
$$\mathbf{B} = \begin{bmatrix} B_{0} \\ B_{1} \\ \vdots \\ B_{p-1} \end{bmatrix}$$

$$\mathbf{X}_{U} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} & X_{1,p} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} & X_{2,p} \\ \vdots & \vdots & & & \\ X_{N,1} & X_{N,2} & \cdots & X_{N,p-1} & X_{N,p} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{1}^{T} \\ \mathbf{x}_{2}^{T} \\ \vdots \\ \mathbf{x}_{N}^{T} \end{bmatrix}$$
$$\mathbf{Y}_{N\times 1} = \mathbf{X}_{U(N\times p)} \mathbf{B}_{p\times 1} + \boldsymbol{\epsilon}_{N\times 1}.$$

• Normal equation

$$\mathbf{X}_U^T \mathbf{X}_U \mathbf{B} = \mathbf{X}_U^T \mathbf{Y}_U$$
$$\mathbf{B} = (\mathbf{X}_U^T \mathbf{X}_U)^{-1} \mathbf{X}_U^T \mathbf{Y}_U$$

•
$$\mathbf{X}_U^T \mathbf{X}_{U(j,k)} = \sum_{i=1}^N x_{ij} x_{ik};$$

 $\mathbf{X}_U^T \mathbf{y}_{U(k)} = \sum_{i=1}^N x_{ik} y_i.$

- Estimate the matrices $\mathbf{X}_U^T \mathbf{X}_U$ and $\mathbf{X}_U^T \mathbf{y}_U$ using weights
- Let \mathbf{X}_S be the matrix of explanatory values for the sample, \mathbf{y}_S be the response vector of sample observations, and let \mathbf{W}_S be a diagonal matrix of the sample weights w_i

•
$$\mathbf{X}_{S}^{T}\mathbf{W}_{S}\mathbf{X}_{S(j,k)} = \sum_{i \in S} w_{i}x_{ij}x_{ik}$$
, which estimates $\sum_{i=1}^{N} x_{ij}x_{ik}$;
 $\mathbf{X}_{S}^{T}\mathbf{W}_{S}\mathbf{y}_{S(k)} = \sum_{i \in S} w_{i}x_{ik}y_{i}$, which estimates the population
total $\sum_{i=1}^{N} x_{ik}y_{i}$

•
$$\hat{\mathbf{B}} = (\mathbf{X}_S^T \mathbf{W}_S \mathbf{X}_S)^{-1} \mathbf{X}_S^T \mathbf{W}_S \mathbf{y}_S$$

• Let $\mathbf{q}_i = \mathbf{x}_i(y_i - \mathbf{x}_i^T \hat{\mathbf{B}})$, using linearization,

$$\hat{V}(\hat{\mathbf{B}}) = (\mathbf{X}_{S}^{T}\mathbf{W}_{S}\mathbf{X}_{S})^{-1}\hat{V}(\sum_{i\in S}w_{i}\mathbf{q}_{i})(\mathbf{X}_{S}^{T}\mathbf{W}_{S}\mathbf{X}_{S})^{-1}$$

• CI:
$$\hat{B}_k \pm t \sqrt{\hat{V}(\hat{B}_k)}$$

Notes:

- Sampling weighted least squares are different from weighted least squares
- The weighted least squares estimate minimizes $\sum (y_i \mathbf{x}_i^T \boldsymbol{\beta})^2 / \sigma_i^2$, and gives observations with smaller variance more weight in determining the regression equation
- Sampling weighted least square: weights come from the sampling design, not from an assumed covariance structure
- Sampling weighted least squares is not maximum likelihood