Regression Methods Improve Precision if Auxiliary Variable x is available

Linear Regression thru Origin

Ratio Estimator (7)

Estimators in Text

Model

$y_i = \beta x_i + \varepsilon_i$

Standard Linear Regression

Regression Estimator (8)

Estimators in Text

Model

$y_i = \alpha + \beta x_i + \varepsilon_i$

$$\hat{\mu}_r = r\mu_x = [\frac{\bar{y}}{\bar{x}}]\mu_x$$

Regression thru origin with errors increasing with x. Discussed in class and text.

$$\hat{\tau}_r = N\hat{\mu}_r$$

$$\hat{\mu}_L = a + b\mu_x$$

$$\hat{\mu}_L = \bar{y} + b(\mu_x - \bar{x})$$

a and b standard least squares estimators of intercept and slope

$$\hat{\tau}_L = N\hat{\mu}_L$$

$$\hat{\mu} = \bar{y}$$

$$\hat{\tau} = N\bar{y}$$

If we ignore the x's then we lose precision!!.

Linear Regression thru Origin

Ratio Estimator (Ch 7)

Model $y_i = \beta x_i + \varepsilon_i$

Important Results

$$\hat{\mu}_r = r\mu_x = [\frac{\bar{y}}{\bar{x}}]\mu_x$$

$$\hat{\mu}_r = \bar{y}[\frac{\mu_x}{\bar{x}}]$$ This shows the motivation for the estimator

$$\hat{\tau}_r = N\hat{\mu}_r$$

$$\hat{V}ar(\hat{\mu}_r) = [\frac{N-n}{N}]\frac{s_r^2}{n}$$

$$\hat{V}ar(\hat{\tau}_r) = N^2\hat{V}ar(\hat{\mu}_r)$$

$$s_r^2 = \frac{1}{n-1}\sum_1^n (y_i - rx_i)^2$$

$$\hat{\mu}_r \pm t_{n-1}(\alpha/2)\sqrt{\hat{V}ar(\hat{\mu}_r)}$$

Linear Regression thru Origin

Ratio Estimator (Ch 7)

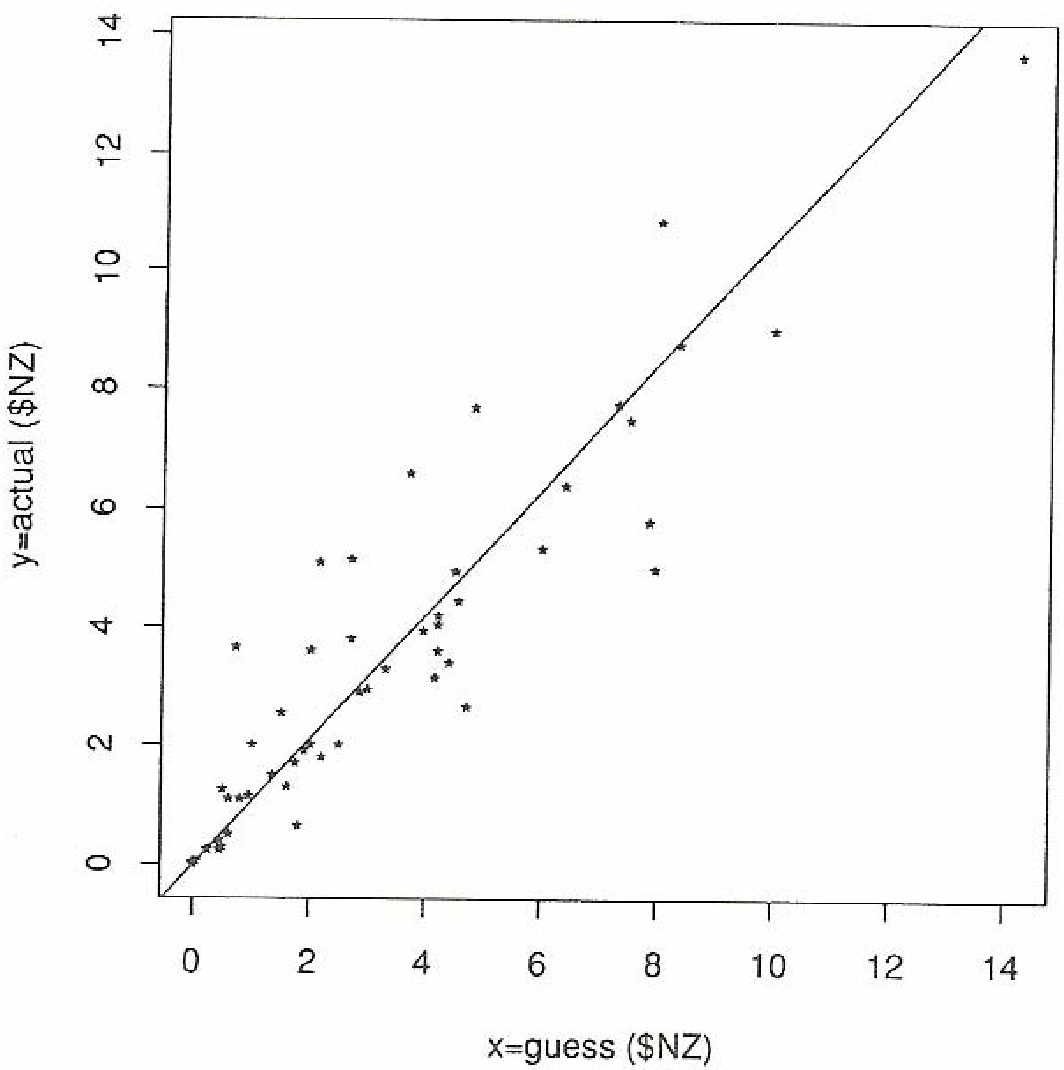
Model $y_i = \beta x_i + \varepsilon_i$

Key Assumption

$$\hat{\mu}_r = r\mu_x = [\frac{\bar{y}}{\bar{x}}]\mu_x$$

Key assumption is that μ_x is known for the whole popn!

P70-71 Ratio Estimator Example (Notice Regression through the Origin is appropriate from Plot! This is the Plot for the Whole Population of 53 pts.



The Sample Data

	1	2	3	4	5	6	7	8	9	10
x	8.35	1.50	10.00	0.60	7.50	7.95	0.95	4.40	1.00	0.50
y	8.75	2.55	9.00	1.10	7.50	5.00	1.15	3.40	2.00	1.25

Comparison of Ratio Estimator vs. Mean of y Variable.

Based on the sample of 10 (y,x) values on P70 text

y and x

$$\hat{\mu}_r = r\mu_x = [\frac{\bar{y}}{\bar{x}}]\mu_x = \frac{4.17}{4.275}(3.33) = 3.25$$

$$SE(\hat{\mu}_r) = 0.3354$$

$$\hat{\mu}_r \pm t_{n-1}(\alpha / 2)\sqrt{\hat{V}ar(\hat{\mu}_r)}$$

$$3.25 \pm 2.262 x 0.3354$$

$$3.25 \pm 0.76$$

$$(2.49, 4.01)$$

y only

$$\hat{\mu} = \bar{y} = \$4.17$$

$$SE(\hat{\mu}) = 0.9061$$

We lose precision!!.(0.9061 vs. 0.3354)

Several Ratio Estimators based on Least Squares Criteria Section 7.7 Text B L U E's.

$$y_i = \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim (0, v_i \sigma^2)$$

$$\hat{\beta} = \frac{\sum_1^n (x_i y_i / v_i)}{\sum_1^n (x_i^2 / v_i)}$$

$$\hat{\beta} = \frac{\sum_1^n (x_i y_i)}{\sum_1^n (x_i^2)} \qquad v_i = 1$$

$$\hat{\beta} = \frac{\sum_1^n (y_i)}{\sum_1^n (x_i)} \qquad v_i = x_i$$

$$\hat{\beta} = \frac{\sum_1^n (y_i / x_i)}{n} \qquad v_i = x_i^2$$

Constant Variance Ratio Estimator

$$y_i = \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim (0, v_i \sigma^2)$$

$$\hat{\beta} = \frac{\sum_1^n (x_i y_i / v_i)}{\sum_1^n (x_i^2 / v_i)}$$

$$\hat{\beta} = \frac{\sum_1^n (x_i y_i)}{\sum_1^n (x_i^2)} \qquad v_i = 1$$

Usual Ratio Estimator- Variance Proportional to x.

$$y_i = \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim (0, v_i \sigma^2)$$

$$\hat{\beta} = \frac{\sum_1^n (x_i y_i / v_i)}{\sum_1^n (x_i^2 / v_i)}$$

$$\hat{\beta} = \frac{\sum_1^n (y_i)}{\sum_1^n (x_i)} \qquad v_i = x_i$$

Another Ratio Estimator-Variance Proportional to x^2

$$y_i = \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim (0, v_i \sigma^2)$$

$$\hat{\beta} = \frac{\sum_1^n (x_i y_i / v_i)}{\sum_1^n (x_i^2 / v_i)}$$

$$\hat{\beta} = \frac{\sum_1^n (y_i / x_i)}{n} \qquad v_i = x_i^2$$

When is the Ratio Estimator a Good One?

- x variable is available for whole popn.
- y and x are linearly related with
- no intercept necessary in model
- The usual ratio estimator assumes the variance structure is such that variance in y's increases as x increases.

Good Example of the Ratio Estimator

- y-Yield of crop, x –Size of farm
- y small then x small so no intercept needed (Regression thru the origin example)
- y is linearly related to x.
- All the x’s on the frame available from a prior Survey

Good Example of the Ratio Estimator

- y-amount overpaid on medical claim, x – amount paid on a medical claim
- y should increase approx linearly as x increases and past analyses have shown this to be a good model.
- All the x’s on the frame available from Medicaid Records

Standard Linear Regression

Regression Estimator Popn Mean

Estimators in Text

Model

$y_i = \alpha + \beta x_i + \varepsilon_i$

$$\hat{\mu}_L = a + b\mu_x$$

$$\hat{\mu}_L = \bar{y} + b(\mu_x - \bar{x})$$

Least Squares Estimates

$$a = \bar{y} - b\bar{x} \qquad b = \frac{\sum_{1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{1}^{n}(x_i - \bar{x})^2}$$

Linear Regression

Important Var Results

$$\hat{V}ar(\hat{\mu}_L) = [\frac{N-n}{N}]\frac{s_e^2}{n}$$

$$s_e^2 = \frac{1}{n-2}\sum_1^n (y_i - a - bx_i)^2$$

$$\hat{\mu}_L \pm t_{n-2}(\alpha/2)\sqrt{\hat{V}ar(\hat{\mu}_L)}$$

Standard Linear Regression

Regression Estimator Popn Total

Estimators in Text

Model

$y_i = \alpha + \beta x_i + \varepsilon_i$

$$\hat{\mu}_L = \bar{y} + b(\mu_x - \bar{x})$$

$$\hat{\tau}_L = N\hat{\mu}_L$$

$$\mathrm{Var}(\hat{\tau}_L) = N^2 \mathrm{Var}(\hat{\mu}_L)$$

P 91 Text Example

	1	2	3	4
y	1410	1690	1680	1850
x	50	100	150	200

Standard Linear Regression
Example Calculations

$$\hat{\mu}_L = \bar{y} + b(\mu_x - \bar{x}) = 1657.5 + 2.62(100 - 125)$$

$$\hat{\mu}_L = 1592$$

$$\hat{\tau}_L = N\hat{\mu}_L \quad = \quad 100 \text{ x } 1592 = 159{,}200$$

$$\text{Var}(\hat{\tau}_L) = 16{,}884{,}000$$

$$\text{SE}(\hat{\tau}_L) = 4{,}109$$

$$\hat{\mu} = 1657.5$$

$$\hat{\tau} = N\hat{\mu} \quad = \quad 100 \text{ x } 1657.5 = 165{,}750$$

$$SE(\hat{\tau}) = 8{,}939$$

Notice the much larger SE if one ignores the information in the x's.

Conclusions

- Use Regression Estimator when x information available on frame, linear relationship of y and x, need for an intercept
- Generalize to multiple linear regression and also nonlinear regression models in finite population sampling

Conclusions
Double sampling Later

μ_x unknown then can use double sampling and get generalised estimators where $\hat{\mu}_x$ comes from a preliminary sample. we will come back to this later. Chapter 14.1

$$\hat{\mu}_r = \overline{y}(\hat{\mu}_x / \overline{x})$$

$$\hat{\mu}_L = \overline{y} + b(\hat{\mu}_x - \overline{x})$$