Ⅰ ランダムに欠損する場合

(1) 問題設定

・真のモデル

$$(\text{賃金})_i = 1000 + 2(IQ)_i + \varepsilon_i \quad (\varepsilon_i \sim N(0, 15^2))$$

$$(\text{就業})_i \sim Ber(0.5)$$

・超人的な視点の完全データ（真のモデルに基づいている）

| i | 賃金 | IQ | 就業 |
|---|---|---|---|
| 1 | 1100 | 110 | 1 |
| 2 | 980 | 95 | 1 |
| 3 | 1050 | 105 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ |

(2) 分析

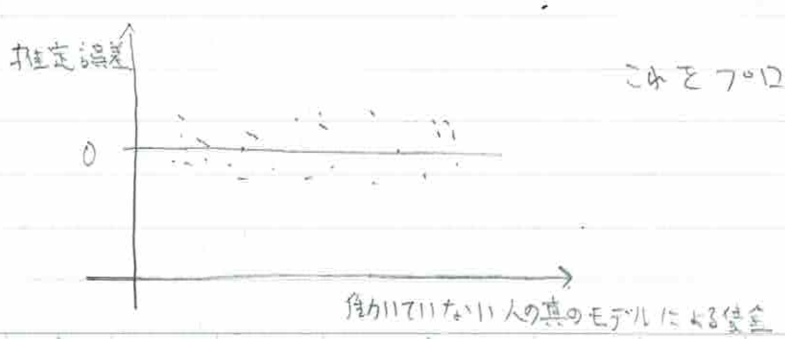
観察される現実のデータ

| i | 〃 | 〃 | 〃 |
|---|---|---|---|
| 1 | 1100 | 110 | 1 |
| 2 | 980 | 95 | 1 |
| 3 | NA | 105 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ |

⇩ $(\text{就業})_i = 1$ のデータに限定して OLS wage on IQ

$(\text{賃金})_i = b_0 + b_1 (IQ_i)$ を得る。

これをもとに、$S_i = 0$ となる人の$(\text{賃金})_i$を推定。




これをプロット

## ② 系統的に欠損する場合

### (1) 問題設定

• 真のモデル　・$(\text{賃金})_i = 1000 + 2(IQ_i) + \varepsilon_i$

・$(\text{就業})_i = \begin{cases} 1 & \text{if } -100 + (IQ_i) + u_i > 0 \\ 0 & \text{if } -100 + (IQ_i) + u_i \leq 0 \end{cases}$

ただし $\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim N\left(0, \begin{bmatrix} \sigma_1^2 & \rho_{21} \\ \rho_{12} & \sigma_2^2 \end{bmatrix}\right)$

• 超人的視点のデータ

| $i$ | 賃金 (W) | IQ (iq) | 就業 |
|---|---|---|---|
| 1 | 1100 | 105 | 1 |
| 2 | (980) | 90 | 0 |
| 3 | 1700 | 120 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ |

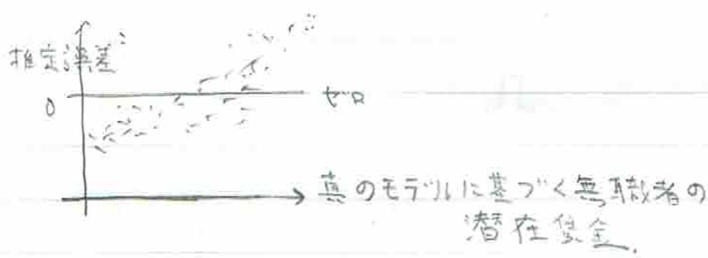
### (2) サンプルセレクションバイアス

$s_i = 1$ となるものに限定したデータについて、

| $i$ | W | iq | s |
|---|---|---|---|
| 1 | 1100 | 105 | 1 |
| 3 | 1700 | 120 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ |

OLS wage on iq. $\mathbb{E}[\hat{\beta}(\text{Observed data})] \neq \beta$.

さらに、プロットもする。




### (3). Heckman's two-step Sample Selection Correction.

(I). 全てのデータに対して、Probit Reg $S_i$ on IQ.

↓

$S_i = 1$ について Inv Mill Ratio $\lambda_i = \dfrac{\phi(\gamma_0 + \gamma_1 IQ)}{\Phi(\gamma_0 + \gamma_1 IQ)}$ を計算

$S_i = 0$ ：　$\lambda_i = \dfrac{-\phi(\gamma_0 + \gamma_1 IQ)}{1 - \Phi(\gamma_0 + \gamma_1 IQ)}$ を計算

(II). $S_i = 1$となるものに限定したデータについて、OLS wage on iq, $\lambda$.