Linear Probability Models

In Binary Variable Case \( y = \begin{cases} 1 & \text{by \( \text{OLS} \)} \\ 0 & \text{estimate \( \beta \)'s.} \end{cases} \)

\( \hat{y} = E[y=1|x] = p(x) \) may not hold the conditions of probability.

In Logit/Probit Models:

\( y = \begin{cases} 1 & \text{and estimation of \( \beta \)'s is by Maximum} \\ 0 & \text{Likelihood Estimation method} \end{cases} \)

\( p(x) = G(x \hat{\beta}) \) where \( G \) is the cumulative distribution function.

Partial effects:

\[ \frac{\partial}{\partial \hat{\beta}_j} G(x \hat{\beta}) = g(x \hat{\beta}) x_j \]

where \( g \) is the probability density function.

- Partial Effects at Average: Replace \( x_i \)'s by \( \bar{x}_j \)'s
- Average Partial Effects: Replace \( x_i \)'s & take the average

\[ \frac{1}{n} \sum_{i=1}^{n} G(x_i \hat{\beta}) \]

Likelihood Ratio Test

Given \( Y = X \beta + u \) \( k \)-Variables, \( k+1 \) parameters

Testing

\[ H_0: \beta_1 = \ldots = \beta_k = 0 \quad \Rightarrow \quad Y = \beta_0 + u \]

\[ H_A: \text{At least one differs} \quad Y = x \beta + u \]

In OLS case:

Test statistic is:

\[ F = \frac{(SSR_R - SSR_{UR})/k}{SSR_{UR} / (n-k-1)} \]

\[ F = \frac{(R_{UR}^2 - R_{UR}^2 Y_k)}{(1 - R_{UR}^2)/(n-k-1)} \]

Reject \( H_0 \) if \( F \geq F_{k,(n-k-1),\alpha} \)
Linear Probability Models

Likelihood function:
\[ L = \log L = \sum_{i=1}^{n} \log g(x_i \beta) \]

Test statistic \[ LR = 2 \left( L_{UR} - L_R \right) \sim \chi^2 \text{ of \# of par. in } H_0, d \]

Example 1: (Handout 8)
Model: Unrestricted
\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_7 X_6 + u \quad L_{UR} = -401.76 \]

Model 0: Restricted
\[ Y = \beta_0 + u \quad L_{R_0} = -514.87 \]

\( H_0: \beta_1 = \beta_2 = \ldots = \beta_7 = 0 \)
\( H_A: \text{not} \)

Test stat \[ LR = 2 \left( -401.76 + 514.87 \right) = 226.216 \]
\[ > \chi^2_{7, 0.05} = 14. \]
Reject \( H_0 \). That is, all explanatory variables are significant

Model 1: Restricted: Impact of having children
\( H_0: \beta_6 = \beta_7 = 0 \quad L_{R_1} = -432.78 \)

\[ LR = 2 \left( -401.76 + 432.78 \right) = 62.022 > \chi^2_{2, 0.05} \]
Reject \( H_0 \).
Take: Age = 30, \( x_1 = x_{\bar{1}} \), \( x_2 = 5 \text{yr} \), \( x_3 = 0 \)

Consider the distribution of \( x_5 \) (\# of children-age < 6)

<table>
<thead>
<tr>
<th>( x_5 )</th>
<th>freq. ( \hat{p}(x)_{\text{OLS}} )</th>
<th>( \hat{p}(x)_{\text{Logit}} )</th>
<th>( \hat{p}(x)_{\text{infl}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>606</td>
<td>0.7</td>
<td>0.75</td>
</tr>
<tr>
<td>1</td>
<td>118</td>
<td>0.42</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>0.16</td>
<td>0.15</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>-0.1</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>753</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Constant marginal effects

\[ \text{Diminishing magnitudes of P.E.} \]

Goodness of Fit Measure

\( R^2_{\text{OLS}} = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\sum \hat{u}_i^2}{\sum (y_i - \bar{y})^2} \)

\( R^2_{\text{pseudo}} = 1 - \frac{\log L(\text{estimated model})}{\log L(\text{constant probability})} \)

Constant probability \( \Rightarrow H_0: \beta_1 = \ldots = \beta_k = 0 \)

If \( R^2_{\text{pseudo}} = 0 \) no explanatory power
\( R^2_{\text{pseudo}} = 1 \) Perfect Fit !!! PROBLEM !!!

\[ 0 \leq R^2_{\text{pseudo}} < 1 \]

To assess the \( R^2_{\text{pseudo}} \): Calculate "Percent Correctly Predicted"
Define $\hat{Y}_i = \begin{cases} 1 & \text{if } \sigma(x^T \beta) > 0.5 \\ 0 & \text{otherwise} \end{cases}$

We count the cases where $\hat{Y}_i$ matches with $Y_i$.

If $Y_i = 0$ and $\hat{Y}_i = 0$ correct estimate
If $Y_i = 1$ and $\hat{Y}_i = 1$ " "
$Y_i = 1$ and $\hat{Y}_i = 0$ incorrect "
$Y_i = 0$ and $\hat{Y}_i = 1$ " "

We define a measure depending on the counts of correct estimates:

Let $n_{ij} =$ no. of times $Y_i = i$ predicted and $Y_j = j$ observed; $i, j = 0, 1$.

<table>
<thead>
<tr>
<th>Predicted $Y_i$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$n_{00}$</td>
<td>$n_{10}$</td>
</tr>
<tr>
<td>1</td>
<td>$n_{10}$</td>
<td>$n_{11}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observed $Y_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Overall percent correctly predicted

$$cP = \frac{n_{00} + n_{11}}{n}$$

% of zero predicted given observed is zero is

$$cP_0 = \frac{n_{00}}{n_0}$$

% of one predicted given observed is one is

$$cP_1 = \frac{n_{11}}{n_1}$$
Measure $\tilde{c}_p = c_p + c_{po}$

In case of constant probability model, $\tilde{c}_p = 1$
In case of probability model, $\tilde{c}_p > 1$

If $\tilde{c}_p > 1$, then $y = x\beta$ predicts better.

Example 1 (48)

<table>
<thead>
<tr>
<th>Observed</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>207</td>
<td>81</td>
</tr>
<tr>
<td>118</td>
<td>347</td>
<td></td>
</tr>
<tr>
<td>325</td>
<td>428</td>
<td>753</td>
</tr>
</tbody>
</table>

Take $\hat{y}_i = 1$ if $\hat{p}(x) \geq 0.5$

$\tilde{c}_p = \frac{207 + 347}{753} = 0.7357$

% of $\hat{y}_i$ having 1 as above = 0.5634

$C_{po} = 0.6369$  $C_{p1} = 0.8107$

$\tilde{c}_p = 1.4437$

When $y$ is continuous, dominated by many zeros and having wide range in its non-zero values, we may have estimates of $y$, $\hat{y}_i < 0$.

Define $y^* = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$

Re-define $y = \max(0, y^*) \implies$ eliminate $y < 0$

$y^*$ satisfies CLM assumption.
Tobit Model

When the percentage is significant and the range of non-zero Y values is big, we use Tobit Model.

For \( Y \in \mathbb{R} \), the classical LM yields negative estimations.

Let \( Y = x\beta + u \), \( X \): \( k \)-variables, \( k+1 \) parameters

Define a latent variable \( y^* \) such as
\[
y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u
\]

\[
y = \max(y^*, 0)
\]

Probability at zero:
\[
P(Y=0 | x) = P(Y^* < 0 | x) = P(u < - \frac{x\beta}{\sigma})
\]

\[
= \Phi(- \frac{x\beta}{\sigma}) = 1 - \Phi(\frac{x\beta}{\sigma})
\]

Interpretation of \( \beta_j \)’s

\[
\hat{y} = E[Y|x] = P[Y > 0 | x] \cdot E[Y | Y > 0, x]
\]

\[
E[Y|x] = \Phi(\frac{x\beta}{\sigma}) \cdot E[Y | Y > 0, x]
\]

\[
\frac{\partial E[Y|x]}{\partial x_j} = \frac{\partial}{\partial x_j} P(Y > 0 | x) E[Y | Y > 0, x] + P(y > 0 | x) \frac{\partial}{\partial x_j} E[Y | Y > 0, x]
\]

\[
= \frac{\partial}{\partial x_j} P(Y > 0 | x) E[Y | Y > 0, x]
\]

\[
= \frac{\partial}{\partial x_j} \Phi(\frac{x\beta}{\sigma}) E[Y | Y > 0, x]
\]

\[
= \frac{\beta_j}{\sigma} \Phi(\frac{x\beta}{\sigma}) \quad \text{is the partial effect of } x_j
\]
Adjustment Factor to compare with OLS estimators
\[
\hat{\beta}_i - \rho \left( \frac{\hat{\varepsilon}_i}{\hat{\sigma}} \right) \quad \text{PE at Average}
\]
\[
\frac{1}{n} \sum_{i=1}^{n} \rho \left( \frac{\hat{\varepsilon}_i}{\hat{\sigma}} \right) \quad \text{Average PE}
\]

Predicted \( \hat{Y}_i \):
\[
\hat{Y}_i = \hat{\Omega} \left( \frac{\hat{\varepsilon}_i}{\hat{\sigma}} \right) \cdot x_i \hat{\beta} + \hat{\varepsilon}_i
\]

Hypothesis Testing:
CLH \rightarrow t - statistic for single parameters also for T-stat
For more than one parameters
Ho: \( \beta_1 = \ldots = \beta_q = 0 \)
we use LR test statistic \( LR = 2(\log L_{LR} - \log L_{UE}) \approx \chi^2_q \)

Example 2: Handout 8
\( Y = \) pension benefit
percent zero = 18%

Model I:
\[
t_w = \frac{144.09}{102.08} = 1.41 < 1.96 \quad \text{Fail to Reject Ho}
\]
\[
t_u = \frac{308.15}{69.85} = 4.409 > 1.96 \quad \text{Reject Ho}
\]

Model II:
\[
t_w = 1.606 < 1.96 \quad \text{Fail to Reject}
\]
\[
t_u = 3.78 > 1.96 \quad \text{Reject Ho}
\]

In both models, gender has significant impact on pension whereas race does not.
\[
t_u = \frac{439.05}{62.49} > 1.96 \quad \text{Reject Ho}
\]

with LR-test \( 2 \left( -3648.55 + 3672.96 \right) \not\sim \chi^2_1 \), so Reject Ho.
\( \hat{Y}_i = \Phi \left( \frac{X_i \hat{\beta}}{\hat{\sigma}} \right) x_i \hat{\beta} + \hat{\sigma} \phi \left( \frac{x_i \hat{\beta}}{\hat{\sigma}} \right) \)
\( \hat{\sigma} = 677.24 \)

**WM**

\( X_\beta = -1.252.5 + 5.20(10) - 4.64(35) + 36.02(10) + 93.21(16) + 144.05 + 308.15 = 940.80 \)

\( \hat{Y}_m = \Phi \left( \frac{940.80}{677.24} \right) \cdot 940.80 + 677.24 \phi \left( \frac{940.80}{677.24} \right) \)

\( \hat{Y}_m = 966.40 \)

**NW-F**

\( X_\beta = 488.66 \quad \text{(Last two terms vanish)} \)

\( \hat{Y}_{NW-F} = \Phi \left( \frac{488.66}{677.24} \right) \cdot 488.66 + 677.24 \phi \left( \frac{488.66}{677.24} \right) \)

\( = 582.10 \)

\( \hat{Y}_m - \hat{Y}_{NW-F} = \$384.30 \)

**Example 3:**

**Adjustment Factor**

\( APE : \frac{1}{n} \sum \Phi \left( \frac{X_{\beta}}{\hat{\sigma}} \right) = 0.589 \)

\( \hat{\beta}_{21} = 80.65 \)

\( 80.65 \times 0.589 = 47.50 \text{ hrs} \) to compare with DLS \( \hat{\beta}_{21LS} = 28.76 \text{ hrs} \)

**PEA :**

\( \Phi \left( \frac{X_{\beta}}{\hat{\sigma}} \right) = 0.645 \)

\( \hat{\beta}_{21} = 80.65 \)

\( 80.65 \times 0.645 = 52 \text{ hrs} \) \( \hat{\beta}_{21LS} = 28.76 \text{ hrs} \)
Poisson Regression

Let $Y$ be a random variable with Poisson ($\lambda$)

$$P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!} \quad y = 0, 1, 2, \ldots$$

Ex: No. of claims per year, no. of accidents per month

$E[Y] = \text{Var}[Y] = \lambda$

Distribution of time between Poisson arrivals is Exponential ($\lambda$)

$$f(t) = \frac{1}{\lambda} e^{-\lambda t} \quad t > 0$$

Regression Model

$Y_i$: count variable / unit $y_i = 0, 1, 2, \ldots$

$E[y_{i|x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

$\% \Delta E[y_{i|x}] \approx (100 \beta_j) \Delta x_j$

$\lambda = \text{mean} \Rightarrow \lambda = E[y_{i|x}] = e^{x^T \beta}$

Then

$$P(Y = h | x) = \frac{(e^{x^T \beta})^h}{h!} \quad h = 0, 1, 2, \ldots$$

Interpretation of $\beta_j$'s

$$\frac{\partial}{\partial x_j} E[y_{i|x}] = \frac{\partial}{\partial x_j} \lambda = e^{x^T \beta} \beta_j$$

Remark: $\text{Var}[y_{i|x}] \neq E[y_{i|x}] \Rightarrow$ to handle this we use LSE of

$$\text{Var}[y_{i|x}] = \sigma^2 E[y_{i|x}]$$

If $\sigma^2 = 1$ Poisson Assumption is fulfilled

$\sigma^2 < 1$ underdispersion

$\sigma^2 > 1$ overdispersion