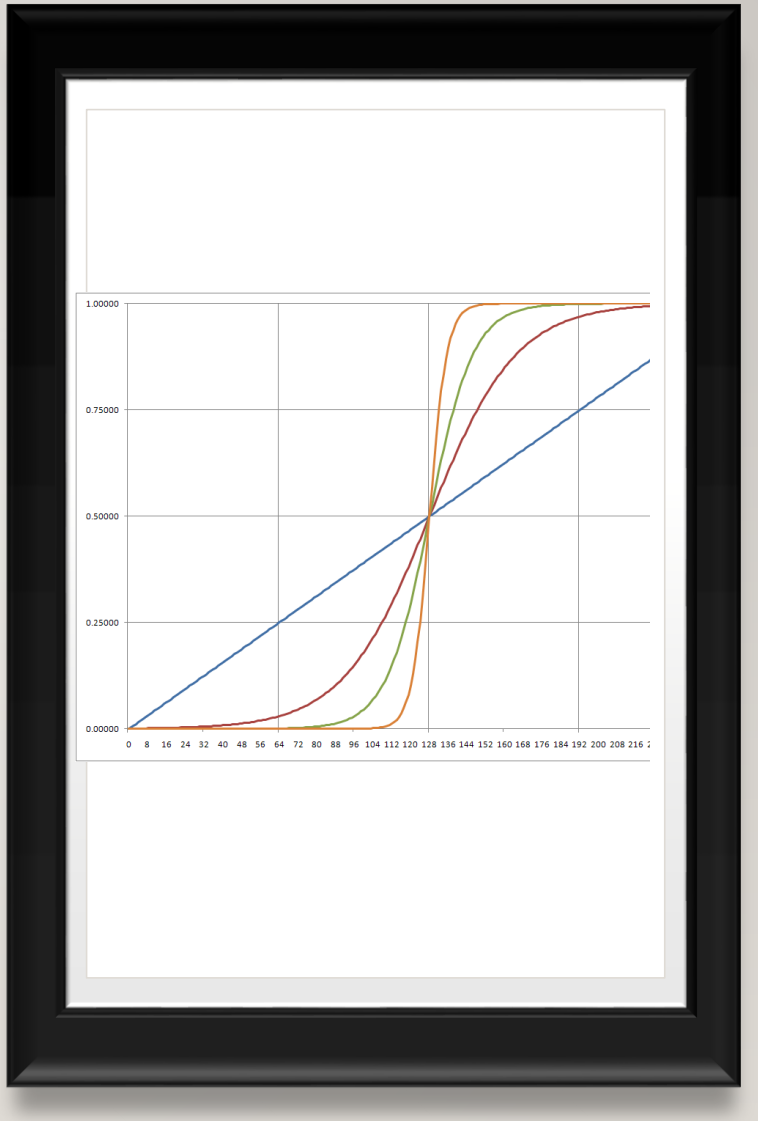
Latent Variables and Discrete Choice Models

(Part 2)

อ.ดร. ภารวี มณีจักร



เนื้อหากระบวนวิชา

1. นิยามของตัวแปรตามที่มีค่าจำกัด
2. กรณีที่ตัวแปรตามเป็นไปได้ 2 ค่า (Binary Responses)
    - แบบจำลองเชิงเส้น (Linear Probability Models)
    - แบบจำลองโลจิท (Logit Models)
    - แบบจำลองโพรบิต (Probit Models)
3. การประมาณแบบจำลองด้วย MLE
4. Odd ratio และ Marginal Effect และการแปลผล
5. กรณีที่ตัวแปรตามเป็นไปได้มากกว่า 2 ค่า (Multiple Responses)
    - Multinomial Logit Model
    - Nested Logit Model
6. Ordered Logit and Probit Models

ทำไมเราต้องใช้ Binary regression?

- เนื่องจากในหลายๆการศึกษาหรืองานวิจัย ตัวแปรตามหรือ Y ในบางครั้งมีลักษณะที่ไม่ใช่ค่าที่เป็นตัวเลขสุ่มหรือตัวเลขทั่วไป เช่น การศึกษาปัจจัยที่ส่งผลต่อการตัดสินใจซื้อ ซึ่งกรณีนี้ Y คือ ซื้อ และ ไม่ซื้อ แค่ 2 ทางเลือก
- ดังนั้นลักษณะข้อมูล Y ไม่มีลักษณะแบบต่อเนื่อง หรือ ลักษณะแบบแจกแจงปกติ (distributed normally)
- ดังนั้นตัวแปร Y จึงมีลักษณะเป็นข้อมูลเชิงคุณภาพ ดังนั้น เราต้องแปลงข้อมูลนี้ให้อยู่ในรูปของตัวเลข (ตัวแปร dummy) เช่น ซื้อ (Y=1) ไม่ซื้อ (Y=0) เป็นต้น

ตัวอย่างการลงข้อมูล

Y

X

EVAC	PETS	MOBLHOME	TENURE	EDUC
0	1	0	16	16
0	1	0	26	12
0	1	1	11	13
1	1	1	1	10
1	0	0	5	12
0	0	0	34	12
0	0	0	3	14
0	1	0	3	16
0	1	0	10	12
0	0	0	2	18
0	0	0	2	12
0	1	0	25	16
1	1	1	20	12

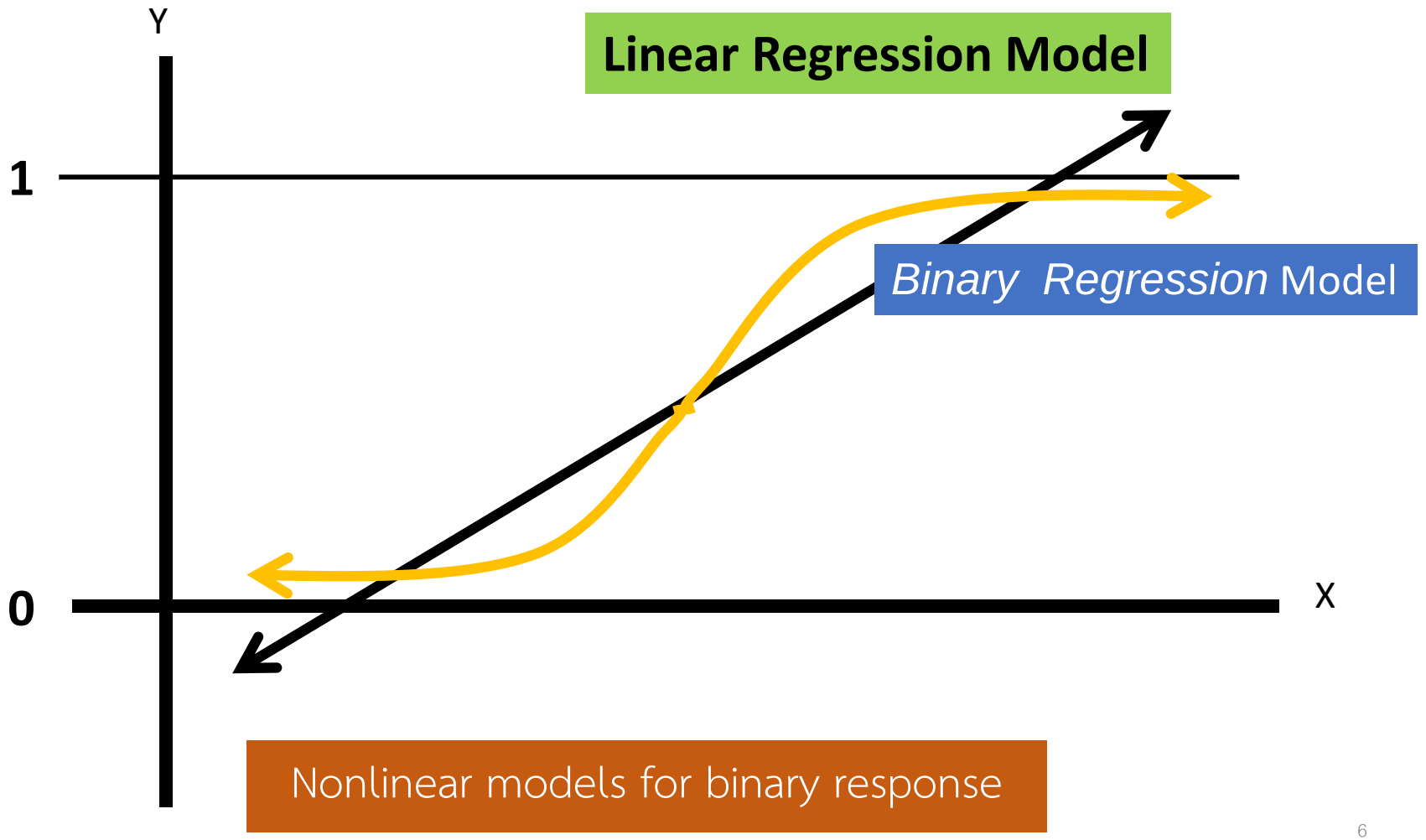
ทำไมเราไม่ใช้แบบจำลอง Linear regression ?

- ตามปกติจะใช้สมการถดถอยเชิงเส้น และประมาณค่าพารามิเตอร์ในแบบจำลองโดยใช้ OLS
- แต่การประมาณแบบจำลองที่ y*= (0,1)

$$y^* = x\beta + e$$

- จะทำให้เกิดปัญหาดังต่อไปนี้
  - ค่าความแปรปรวนของ error ไม่นิ่ง -> เกิดปัญหา Heteroskedasticity
  - Error มีค่าเฉลี่ยไม่เท่ากับ 0 และไม่มีการแจกแจงแบบปกติ
  - ค่า Y ที่พยากรณ์จะไม่มีค่าอยู่ระหว่าง 0 ถึง 1 ซึ่งไม่ตรงกับความเป็นจริง
  - Guass Markov assumption ไม่เป็นจริงในหลายๆข้อ ทำให้การประมาณด้วย OLS ได้ผลที่ Bias

Comparing Linear Regression and Binary Regression Models



Logit and Probit models for Binary choice Model

- ปัจจุบันแบบจำลอง Binary choice regression มี 2 แบบจำลอง คือ
- 1) Logit regression

  2) Probit regression

$$P(y=1|\mathbf{x}) = G(\beta_0+\beta_1 x_1+\ldots+\beta_k x_k) = G(\mathbf{x}\boldsymbol{\beta})$$

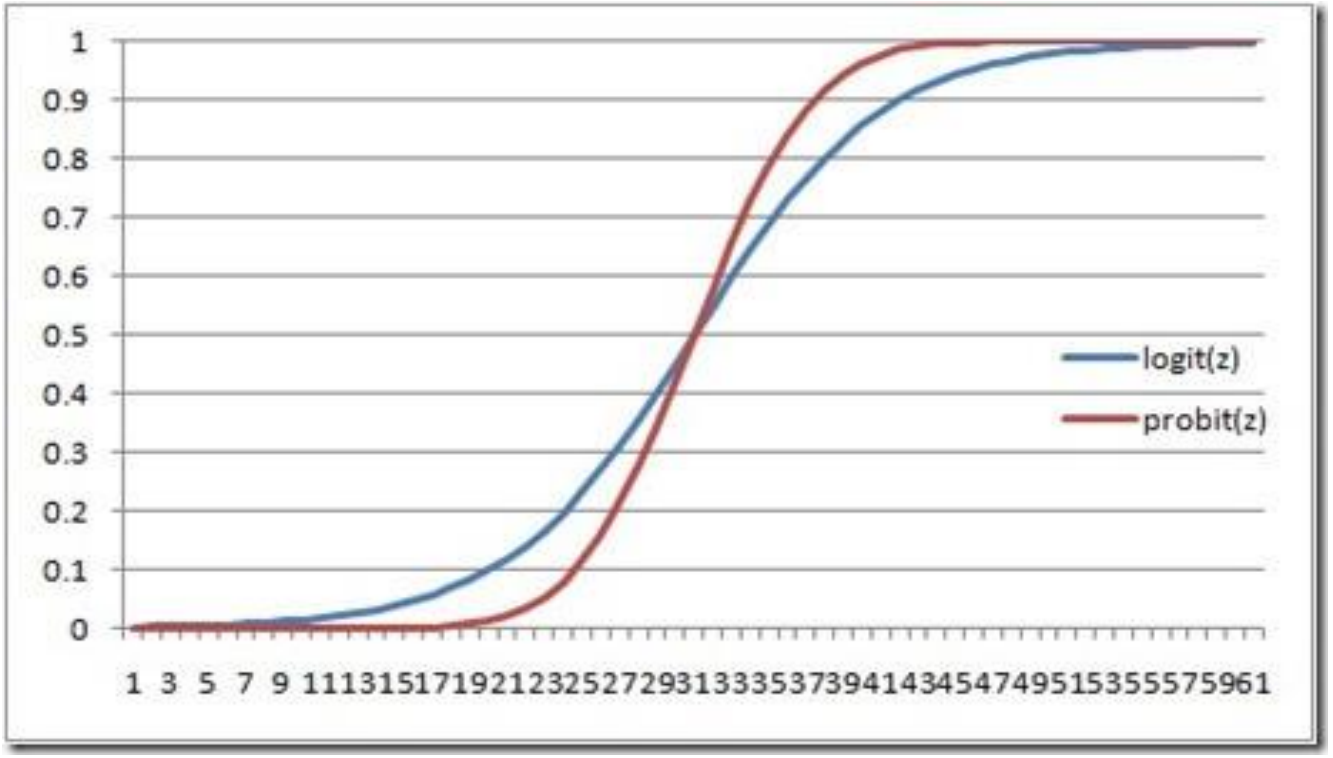
ความน่าจะเป็นที่ Y=1

ความน่าจะเป็นสะสม (cumulative distribution function : CDF) ซึ่งเป็นฟังก์ชันที่ทำให้สมการอยู่ในช่วง 0-1 นั่นเองทำให้เราสามารถประมาณค่า Yที่อยู่ในช่วง 0-1

$0 < G(z) < 1$

Logit and Probit distribution

- Cumulative Distribution Function : CDF ของ Logit และ Probit



Logit and Probit Models for Binary Response

- Cumulative Distribution Function : CDF ของ Logit และ Probit

Probit: $G(z) = \Phi(z) \equiv \int_{-\infty}^{z} \phi(v)dv$ (normal distribution)

Logit: $G(z) = \Lambda(z) = \exp(z)/\left[1 + \exp(z)\right]$ (logistic function)

whereas $z = x\beta$

- ดังนั้นสมการแบบจำลอง Logit และ Probit

$$y^* = x\beta + e \quad \text{and} \quad y = 1\left[y^* > 0\right]$$

$$\Rightarrow P(y = 1|\mathbf{x}) = P(y^* > 0|\mathbf{x})$$

$$= P(e > -x\beta) = 1 - G(-x\beta) = G(x\beta)$$

เพิ่มเติม:

logistic และ normal distribution ทำให้เราสมารถสร้างสมการ probability ได้ ซึ่งเป็นทางเดียวที่ทำให้ Y อยู่ในช่วง 0-1

ถ้า β X =0, ดังนั้น p = .50

- ยิ่งค่า β X สูงขึ้น, p จะเข้าใกล้ 1 มากขึ้น
- ยิ่งค่า β X ต่ำขึ้น, p จะเข้าใกล้ 0 มากขึ้น

การประมาณ

- เนื่องจาก ข้อสมมุติของ Gauss Markov ไม่เป็นจริงในหลายๆข้อ ดังนั้นการประมาณโดย OLS จะผิดดั้งนั้นในการประมาณแบบจำลอง Probit และ Logit จึงต้องใช้การประมาณแบบอื่นซึ่งก็คือ Maximum Likelihood Estimation (MLE)
- โดยที่ likelihood function (L) คือค่าฟังก์ชั่นความน่าจะเป็นที่ค่าตัวแปร Y จะเข้าใกล้ค่าจริงหรือเหมือนกับตัวแปร Y ที่เป็นกลุ่มตัวอย่างของเรา (p_1, p_2, ..., p_n), เช่น ถ้า คนที่แรกตอบ $Y_1 = 0$, คนที่สองตอบ $Y_2 = 1$ ดังนั้น p_1 *คือความน่าจะเป็นที่คนแรกตอบ* $Y_1 = 0$ และ p_2, *คือ*คนที่สองตอบ $Y_2 = 1$ นั่นเอง ซึ่งปกติเราจะเก็บข้อมูล N ตัวอย่าง ดังนั้น function likelihood คือ

$$L = \text{Prob}\,(p_1 * p_2 * * * p_n)$$

- ดังนั้นการประมาณที่ดีที่สุดเพื่อหาค่า β คือการหาค่า β ที่ทำให้ค่า L สูงที่สุดนั่นเอง เพราะค่าที่ L สูงสุดคือค่าความน่าจะเป็นที่เราจะพยาการณ์ค่า Y ถูกต้องที่สุด

- Maximum likelihood estimation ของแบบจำลอง Logit และ Probit

$$f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) = [G(\mathbf{x}_i\boldsymbol{\beta})]^{y_i}\,[1 - G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y_i}$$

ความน่าจะเป็นของคนที่ตอบ y_i =0 โดยที่การตอบนั่นขึ้นอยู่กับปัจจัย x ต่างๆในแบบจำลอง

ความน่าจะเป็นของคนที่ตอบ y_i =1 โดยที่การตอบนั่นขึ้นอยู่กับปัจจัย x ต่างๆในแบบจำลอง

- ดังนั้นเราสามารถสร้างสมการ log-likelihood ได้ดังนี้

$$\log L(\boldsymbol{\beta}) = \log\left(\prod_{i=1}^{n} f(y_i|\mathbf{x}_i;\boldsymbol{\beta})\right) = \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i;\boldsymbol{\beta})$$

$$\widehat{\boldsymbol{\beta}} = \max \sum_{i=1}^{n} \log L_i(\boldsymbol{\beta})$$

Maximum likelihood estimates

การประมาณ

$$\widehat{\boldsymbol{\beta}} = \max \sum_{i=1}^{n} \log L_i(\boldsymbol{\beta})$$

$$\log L(\beta) = \log \left(\prod_{i=1}^{n} f(y_i|\mathbf{x}_i;\beta) \right) = \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i;\beta)$$

FOC

$$\frac{\partial \log L(\beta)}{\partial \beta} = 0$$

SOC

$$\frac{\partial \log L(\beta)}{\partial \beta} < 0$$

ตัวอย่างผล (STATA)

$bvap_i$ is black voting age population

$$black\ elected_i = \beta_0 + \beta_1 bvap_i + e_i$$

```
. probit black bvap

Iteration 0:   log likelihood = -735.15352
Iteration 1:   log likelihood = -292.89815
Iteration 2:   log likelihood = -221.90782
Iteration 3:   log likelihood = -202.46671
Iteration 4:   log likelihood = -198.94506
Iteration 5:   log likelihood = -198.78048
Iteration 6:   log likelihood = -198.78004

Probit estimates                                  Number of obs   =       1507
                                                  LR chi2(1)      =    1072.75
                                                  Prob > chi2     =     0.0000
Log likelihood = -198.78004                       Pseudo R2       =     0.7296

------------------------------------------------------------------------------
       black |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bvap |    0.092316   .5446756    16.95   0.000     0.081641    0.102992
       _cons |   -0.047147   0.027917   -16.89   0.000    -0.052619   -0.041676
------------------------------------------------------------------------------
```

Maximizing the log-likelihood function!

Coefficients are significant

Interpretation (การแปลผล)

- โดยปกติการแปลผลสมการ linear regression คือ ถ้าเราจะแปลผลจากค่า $\hat{\beta}$ ซึ่งก็คือถ้า x เปลี่ยนไป 1 หน่วย Y จะเปลี่ยนแปลงไปเท่ากับ $\hat{\beta}$.

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$$\frac{\partial y}{\partial x} = \widehat{\beta}$$

- แต่ในแบบจำลอง probit หรือ logit

$$y_i = \mathrm{G}(\beta_0 + \beta_1 x_i)$$

$$\frac{\partial y}{\partial x} = \widehat{\beta}\, G(x\beta)$$

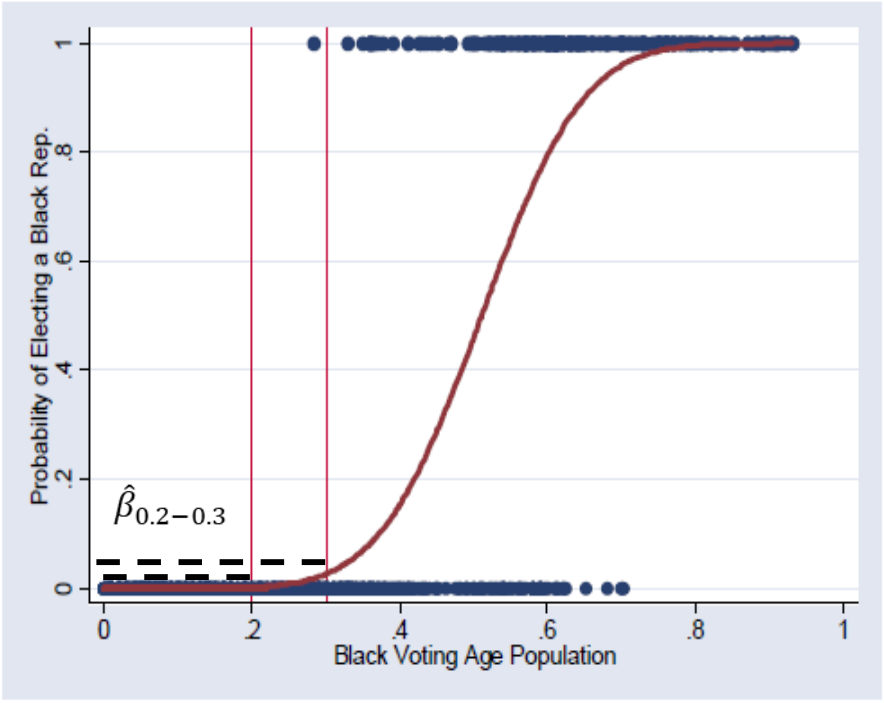
- เราจะแปลผลว่า ถ้า x เปลี่ยนไป 1 หน่วย Pr(Y=1) จะเปลี่ยนแปลงไปเท่ากับ $\hat{\beta}$.

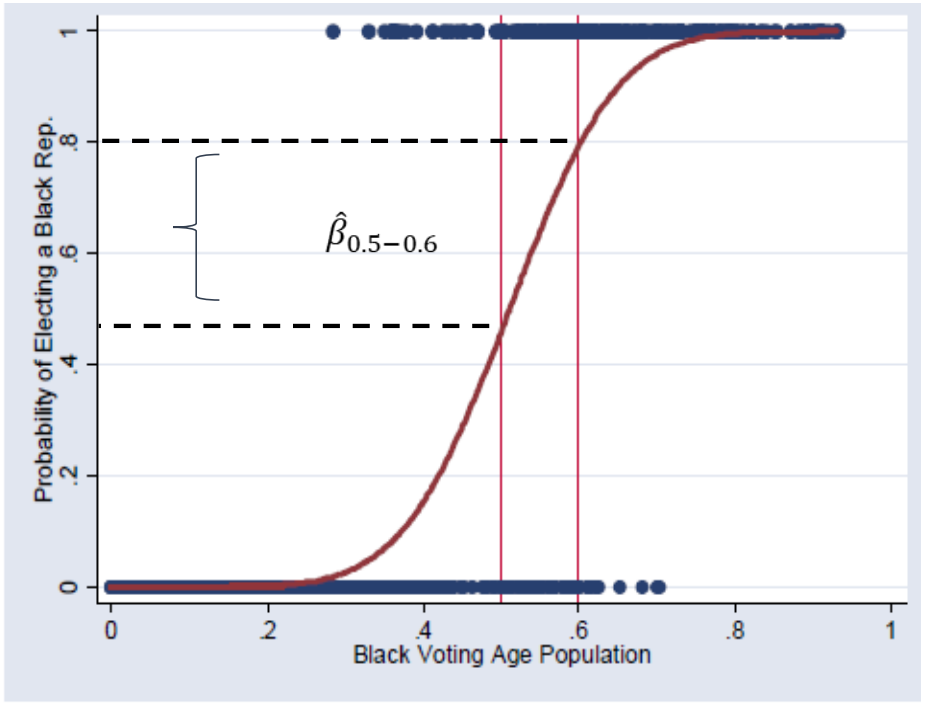
แต่อย่างไรก็ตามการแปลผลแบบนี้ยังมีข้อบกพร่องอยู่ เนื่องจากเป็นการแปลผลที่คงที่ตลอดไป กล่าวคือ ถ้า x เปลี่ยนไป 1 หน่วย Pr(Y=1) จะเปลี่ยนแปลงไปเท่ากับ $\hat{\beta}$ ตลอด ซึ่งไม่จริงสำหรับ แบบจำลอง logit และ probit ซึ่งทั้งสองแบบจำลองนี้มีการแปลผลที่จะไม่คงที่

ปัญหาของการแปลผลแบบเดิม

ตัวอย่างเช่น การเพิ่มขึ้นของสัดส่วน BVAP จาก 0.2 ไป 0.3 (20%→30%) มีผลทำให้ความน่าจะเป็นที่คนดำจะได้รับเลือกตั้งเพิ่มขึ้นไม่มาก (Black Elected)

แต่ การเพิ่มขึ้นของสัดส่วน BVAP จาก 0.5 ไป 0.6 (50%→60%) มีผลทำให้ความน่าจะเป็นที่คนดำจะได้รับเลือกตั้งเพิ่มขึ้นอย่างมาก





$\hat{\beta}_{0.2-0.3}$ น้อยกว่า $\hat{\beta}_{0.5-0.6}$

เราจะแก้ไขปัญหานี้อย่างไร: Marginal Effects

- ดังนั้นค่า $\hat{\beta}$ ที่ของแบบจำลอง **Probit** และ **Logit** นั้นมีลักษณะเป็น **partial effects** ดังนั้นเราจะแปลอย่างไรดี ถ้า $\hat{\beta}$ ไม่คงที่แบบนี้ คำตอบคือ เราต้องใช้ผลเฉลี่ยของ $\hat{\beta}$ ในแต่ละช่วงนั่นเอง ซึ่งสามารถทำได้ 2 วิธีด้วยกัน

- **Partial effects at the average:**

$$\widehat{PEA}_j = g(\bar{x}\hat{\beta})\hat{\beta}_j$$

แปลงตัวแปร x ให้อยู่ในค่าเฉลี่ย $\bar{x}$ (ในกรณีจะมีปัญหาเมื่อตัวแปร x มีค่าเป็น 1 และ 0 เพราะค่าเฉลี่ยจะเป็นอยู่ระหว่าง 0-1 เสมอซึ่งค่อนข้างจะไม่สะท้อนค่าที่แท้จริงของตัวแปร)

- **Average partial effects:**

$$\widehat{APE}_j = n^{-1}\sum_{i=1}^{n} g(\mathbf{x}_i\hat{\beta})\hat{\beta}_j$$

ในกรณีนี้คือหาค่าเฉลี่ยของ partial effects ทั้งหมดเลยนั่นเอง (แบบนี้ดีกว่าและได้รับความนิยมมากกว่า)

- เราจะเรียกวิธีการเหล่านี้ว่าการทำ Marginal effect นั้นเอง เพื่อที่จะสามารถแปลผลแบบจำลอง Logit และ Probit ได้

Marginal effect (continuous data)

$$\frac{\partial G(x\beta)}{\partial x}$$

Multivariate test

- Hypothesis testing (สำหรับ maximum likelihood estimation)
  - t-tests and confidence intervals ก็สามารถได้ในกรณี univariate test
  - แต่ในกรณีของ multivariate hypotheses test จะใช้วิธีอื่นๆ เช่น
    - Lagrange multiplier or score test
    - Wald test
    - Likelihood ratio test (คล้ายกับการทำ F-test)

$$LR = 2(\log L_{ur} - \log L_r) \sim \chi^2_q$$

Chi-square distribution with q degrees of freedom

$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ → Lr

$H_1 : H_0$ is not true → Lur

- Goodness-of-fit measures for Logit and Probit models
  - Percent correctly predicted

$$\tilde{y}_i = \begin{cases} 1 & \text{if } G(\mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq .5 \\ 0 & \text{otherwise} \end{cases}$$

Individual i's outcome is predicted as one if the probability for this event is larger than .5, then percentage of correctly predicted y = 1 and y = 0 is counted

  - Pseudo R-squared

$$\tilde{R}^2 = 1 - \log L_{ur} / \log L_0$$

Compare maximized log-likelihood of the model with that of a model that only contains a constant (and no explanatory variables)

  - Correlation based measures

$$Corr(y_i, \tilde{y}_i), \;\; Corr(y_i, G(\mathbf{x}_i\hat{\boldsymbol{\beta}}))$$

Look at correlation (or squared correlation) between predictions or predicted prob. and true values

เนื่องจากเราใช้เครื่องมือ Maximum Likelihood ดังนั้นเราต้องทำการทดสอบสถิติที่ใช้ Likelihood

การทดสอบสมมติฐาน

ผู้วิจัยสามารถทดสอบแบบจำลองโลจิทและโพรบิตได้ว่าแบบจำลองที่ประมาณได้นั้นสามารถอธิบายพฤติกรรมได้อย่างมีนัยสำคัญหรือไม่ โดยพิจารณาจากค่า Likelihood ratio statistic ซึ่งมีค่าเป็นสองเท่าของส่วนต่างของค่า log-likelihoods ดังนี้

$$LR = 2(L_{ur} - L_r)$$

เมื่อกำหนดให้ $\mathbf{L_{ur}}$ คือค่า log-likelihood เมื่อแบบจำลองไม่มีข้อจำกัด (Unrestricted Model) และ $\mathbf{L_r}$ คือค่า log-likelihood เมื่อแบบจำลองมีข้อจำกัด (Restricted Model) เนื่องจาก $\mathbf{L_{ur} - L_r}$ เสมอ ดังนั้นค่า Likelihood ratio statistic จึงจะมีค่าเป็นบวกเสมอ

LR มีการกระจายแบบไคสแควร์ (Chi-square distribution) ภายใต้ข้อสมมติฐานหลัก $\mathbf{H_0}$ ที่ว่าแบบจำลองไม่สามารถอธิบายพฤติกรรมของตัวแปรตามได้ ดังนั้นเมื่อต้องการทดสอบว่าแบบจำลองสามารถอธิบายพฤติกรรมของตัวแปรตามได้หรือไม่ ผู้วิจัยจึงต้องพิจารณาจากค่า LR ถ้าสมการมีตัวแปรอิสระ q ตัว (ไม่นับค่าคงที่) ค่า LR จะมีการกระจายแบบไคสแควร์ที่มี degree of freedom (Df) เท่ากับ q ถ้าค่าสถิติ LR มีนัยสำคัญ ณ ระดับ 5 % แสดงว่าผู้วิจัยสามารถปฏิเสธข้อสมมติฐานหลัก ($\mathbf{H_0}$) และยอมรับว่าแบบจำลองสามารถอธิบายพฤติกรรมของตัวแปรตามได้ โดยปรกติแล้วโปรแกรมคอมพิวเตอร์ที่เกี่ยวข้องกับเศรษฐมิติจะคำนวณค่า LR และ p-value ทุกครั้งเมื่อมีการประมาณสมการโลจิทและโพรบิต

การแปรผลของค่าที่ประมาณได้จากแบบจำลองโลจิทและโพรบิต

โปรแกรมคอมพิวเตอร์จะคำนวณค่าสัมประสิทธ์ ค่า standard error ของสัมประสิทธ์ และค่าของฟังก์ชัน Log-likelihood ค่าสัมประสิทธ์จะเป็นค่าบอกทิศทางของผลกระทบของตัวแปรอิสระนั้นๆ (X) ที่มีต่อโอกาสในการเกิดเหตุการณ์ (Y) ถ้าค่าสัมประสิทธ์เป็นบวกหมายความว่าตัวแปรอิสระนั้นทำให้โอกาสที่จะเกิดเหตุการณ์ ดังกล่าวเพิ่มขึ้น ตรงกันข้ามถ้าค่าสัมประสิทธ์เป็นลบหมายความว่าตัวแปรอิสระนั้นทำให้โอกาสที่จะเกิดเหตุการณ์ดังกล่าวลดลง ค่า standard error ของสัมประสิทธ์ จะเป็นตัวกำหนดค่านัยสำคัญทางสถิติของตัวแปร $\mathbf{X_j}$ แต่ละตัว ซึ่งระดับนัยสำคัญจะบอกได้ว่าผู้วิจัยจะสามารถปฏิเสธสมมติฐานหลัก ($\mathbf{H_0}$) ที่ว่าค่าสัมประสิทธ์ของ $\mathbf{X_j}$ เท่ากับศูนย์ $(\mathrm{H_0}: \beta_j = 0)$ ได้หรือไม่

ความสามารถของแบบจำลองที่ประมาณได้ในการที่จะพยากรณ์ค่าความน่าจะเป็นของกลุ่มตัวอย่างหรือที่เรียกว่า Goodness-of-fit นั้นคือเปอร์เซ็นต์ที่แบบจำลองจะพยากรณ์ได้ถูกต้องสามารถคำนวณได้ดังนี้ หากพบว่าค่าพยากรณ์ $G(\hat{\beta}_0 + X_i\hat{\beta}) > 0.5$ ค่าประมาณโอกาสในการเกิดเหตุการณ์ Y=1 และหากพบว่าค่าพยากรณ์ $G(\hat{\beta}_0 + X_i\hat{\beta}) \leq 0.5$ ค่าประมาณโอกาสในการเกิดเหตุการณ์ Y=0 เปอร์เซ็นต์ที่แบบจำลองพยากรณ์เท่ากับค่าจริงจะหมายถึงแบบจำลองพยากรณ์ถูกต้อง นั่นคือพยากรณ์ว่า Y=0 เมื่อไม่มีเหตุการณ์ Y และพยากรณ์ว่า Y=1 เมื่อมีเหตุการณ์ Y เกิดขึ้น แต่การคิดเปอร์เซ็นต์รวมก็อาจจะนำไปสู่ข้อสรุปที่ไม่ถูกต้องได้ เช่น หากมีตัวอย่างทั้งหมด 100 คน โดยกลุ่มตัวอย่างที่มี Y=1 มีจำนวน 10 คน และโดยกลุ่มตัวอย่างที่มี Y=0 มีจำนวน 90 คน ถึงแม้ว่าแบบจำลองจะพยากรณ์ Y=0 สำหรับกลุ่มตัวอย่างทั้ง 100 คน เปอร์เซ็นต์ ความแม่นยำของแบบจำลองเท่ากับ 90 % ทั้ง ๆ ที่แบบจำลองพยากรณ์กลุ่มตัวอย่างที่มีค่า Y=1 ไม่ถูกต้องเลย ดังนั้นการแสดงความแม่นยำจึงควรแยกกันระหว่างสองผลลัพธ์ดังแสดงในตารางต่อไปนี้

ตารางที่ 8.2 ตัวอย่างการรายงานความแม่นยำของตัวอย่างแบบจำลอง

	ค่าพยากรณ์ $(\hat{Y}_i = 0)$	ค่าพยากรณ์ $(\hat{Y}_i = 1)$	รวม
ค่าจริง $(Y_i = 0)$	90	0	90
ค่าจริง $(Y_i = 1)$	10	0	10
รวม	100	0	100

McFadden (1974) ได้เสนอให้มีการใช้ Pseudo R-squared เพื่อพิจารณาแบบจำลอง โดยค่า Pseudo R-squared จะมีความหมายคล้ายกับค่า R-squared ในการประมาณสมการถดถอยเชิงเส้น (Linear Regression) คำนวณได้จาก $1 - \mathbf{L_{ur}}/\mathbf{L_0}$ กำหนดให้ $\mathbf{L_{ur}}$ คือค่า log-likelihood เมื่อแบบจำลองไม่มีข้อจำกัด (Unrestricted Model) และ $\mathbf{L_0}$ คือค่า log-likelihood เมื่อแบบจำลองมีเพียงค่าคงที่ (Intercept) เพียงอย่างเดียว ถ้าแบบจำลองไม่สามารถอธิบายเหตุการณ์ได้ $\mathbf{L_{ur}}/\mathbf{L_0} = \mathbf{1}$ ทำให้ Pseudo R-squared เท่ากับ 0 ตรงกันข้ามถ้า $\mathbf{L_{ur}} = \mathbf{0}$ ค่า Pseudo R-squared เท่ากับ 1 หมายความว่าแบบจำลองสามารถอธิบายเหตุการณ์ได้ดีมาก*

สมมติให้ต้องการประมาณแหล่งผลิตรถยนต์ (Y=1 เมื่อเป็นรถยนต์ต่างประเทศ,Y=0 เมื่อเป็นรถยนต์ผลิตในประเทศ) จากลักษณะการใช้พลังงานต่อระยะทาง (mpg) น้ำหนักรถยนต์ (weight) และอัตราเกียร์ (gear_ratio)

---

*ศูนย์เป็นค่าสูงสุดของ Log-likelihood ratio แต่ในความเป็นจริงค่า $\mathbf{L_{ur}}$ไม่สามารถเป็นศูนย์ได้ในแบบจำลองโลจิทหรือโพรบิต

ตัวอย่างการแสดงผลของการประมาณแบบจำลองโลจิท

```
. logit foreign mpg weight gear_ratio

Iteration 0:   log likelihood = -42.400729
Iteration 1:   log likelihood =  -21.25145
Iteration 2:   log likelihood = -16.956941
Iteration 3:   log likelihood = -15.692303
Iteration 4:   log likelihood = -15.495945
Iteration 5:   log likelihood = -15.488658
Iteration 6:   log likelihood = -15.488644

Logit estimates                                   Number of obs   =         69
                                                  LR chi2(3)      =      53.82
                                                  Prob > chi2     =     0.0000
Log likelihood = -15.488644                       Pseudo R2       =     0.6347

------------------------------------------------------------------------------
     foreign |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -.2524188   .1167192    -2.16   0.031    -.4811843   -.0236533
      weight |  -.0033825   .0013854    -2.44   0.015    -.0060978   -.0006671
  gear_ratio |   5.543769   1.856432     2.99   0.003     1.905229    9.182309
       _cons |   -3.43858   7.283729    -0.47   0.637    -17.71443    10.83727
------------------------------------------------------------------------------
```

ตัวอย่างจากคู่มือ STATA 8.0

ตัวอย่างการแสดงผลของการประมาณแบบจำลองโพรบิต

```
. probit foreign mpg weight gear_ratio

Iteration 0:   log likelihood = -42.400729
Iteration 1:   log likelihood = -20.721735
Iteration 2:   log likelihood =  -16.38795
Iteration 3:   log likelihood = -15.284765
Iteration 4:   log likelihood = -15.148649
Iteration 5:   log likelihood = -15.145575
Iteration 6:   log likelihood = -15.145573

Probit estimates                                  Number of obs   =         69
                                                  LR chi2(3)      =      54.51
                                                  Prob > chi2     =     0.0000
Log likelihood = -15.145573                       Pseudo R2       =     0.6428

------------------------------------------------------------------------------
     foreign |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -.1487816   .0694506    -2.14   0.032    -.2849022    -.012661
      weight |  -.0019479   .0007826    -2.49   0.013    -.0034817    -.000414
  gear_ratio |   3.279957   1.054985     3.11   0.002     1.212224     5.34769
       _cons |  -2.182765   4.259567    -0.51   0.608    -10.53136    6.165832
------------------------------------------------------------------------------

note: 1 failure and 0 successes completely determined.
```

ตัวอย่างและการเปรียบเทียบผลการประมาณจากแบบจำลองต่างๆ

ตัวอย่าง: Married women's labor force participation

TABLE 17.1 LPM, Logit, and Probit Estimates of Labor Force Participation

Dependent Variable: *inlf*			
Independent Variables	LPM (OLS)	Logit (MLE)	Probit (MLE)
nwifeinc	−.0034 (.0015)	−.021 (.008)	−.012 (.005)
educ	.038 (.007)	.221 (.043)	.131 (.025)
exper	.039 (.006)	.206 (.032)	.123 (.019)
$exper^2$	−.00060 (.00019)	−.0032 (.0010)	−.0019 (.0006)
age	−.016 (.002)	−.088 (.015)	−.053 (.008)
kidslt6	−.262 (.032)	−1.443 (.204)	−.868 (.119)
kidsge6	.013 (.014)	.060 (.075)	.036 (.043)
constant	.586 (.152)	.425 (.860)	.270 (.509)
Percentage correctly predicted	73.4	73.6	73.4
Log-likelihood value	—	−401.77	−401.30
Pseudo *R*-squared	.264	.220	.221

จะเห็นว่าผลการประมาณแตกต่างกันอย่างมากและเราจะเลือกแบบจำลองใดหรือเชื่อแบบจำลองใด?

LPM = linear Probability model or linear regression model

Logit and Probit Models in R

- **ใช้ Package ‘mfx’ ในการหา Marginal Effects และ Odds Ratios**
- **ตัวอย่างข้อมูลที่ใช้**

ศึกษาผลของปัจจัยดังต่อไปนี้ที่มีต่อการตัดสินใจเข้าศึกษาต่อในระดับบัณฑิตศึกษา ประกอบด้วย

admit	คือ ตัวแปรตามมีลักษณะเป็น Binary responses โดย 1 = เข้าร่วม และ 0 = ไม่เข้าร่วม
gre	คือ คะแนนสอบความถนัดทั่วไป (General Test Score)
gpa	คือ ผลการเรียนเฉลี่ยสะสม (Grade Point Average)
rank	คือ ชื่อเสียงของมหาวิทยาลัย โดยมีค่าเรียงลำดับระหว่าง 1 ถึง 4 โดย 1 หมายถึง มีชื่อเสียงมากที่สุด และ 4 หมายถึง มีชื่อเสียงน้อยที่สุด

- Install package และเรียก Library (mfx)
- นำเข้าข้อมูลโดย

```
data <-read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
```

- กำหนดให้แบบจำลอง คือ *admit = gpa + gre + rank*
- ในการประมาณ Logit model ใช้ฟังก์ชัน glm (generalized linear model)

```
m1=glm(admit~gpa+gre+rank,family=binomial(link = "logit"))
summary(m1)
```

```
Call:
glm(formula = admit ~ gpa + gre + rank, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5802  -0.8848  -0.6382   1.1575   2.1732

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.449548   1.132846  -3.045  0.00233 **
gpa          0.777014   0.327484   2.373  0.01766 *
gre          0.002294   0.001092   2.101  0.03564 *
rank        -0.560031   0.127137  -4.405 1.06e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 459.44  on 396  degrees of freedom
AIC: 467.44
```

แสดงการแจกแจงของ residuals สำหรับ individual cases

ทุกตัวแปรมีนัยสำคัญทางสถิติ ณ ระดับนัยสำคัญแตกต่างกันไป

คำถามคือ เราจะแปลผลจากตารางนี้เลยหรือไม่ ?

- คำตอบ คือ ***ไม่*** เนื่องจากค่าเบต้าที่ประมาณได้มีค่าไม่คงที่ในแต่ละช่วง ดังนั้น ต้องใช้ผลเฉลี่ยของเบต้าในแต่ละช่วง
- แนวทางแก้ไข 2 วิธี

1. Odd ratio

```
logitor(formula=admit~gpa+gre+rank, data=data)
```

2. Marginal effect

```
logitmfx(formula=admit~gpa+gre+rank, data=data)
```

Odds ratio

- **Odds ratio (OR) หรือ Exp(B) หรือ e^B**

การแปลผล

- **ถ้าค่า Odds ratio = 1** แสดงว่าการเปลี่ยนแปลงของ X ไม่มีผลต่อ y
- **ถ้าค่า Odds ratio >1** แสดงว่าเมื่อ X เพิ่มขึ้น ทำให้โอกาสของการเกิดเหตุการณ์เพิ่มขึ้น (ค่าสัมประสิทธิ์ถดถอย จะมีค่าเป็นบวก)

แปลผลได้ว่า เมื่อ x เปลี่ยนไป 1 หน่วย โอกาสที่จะเกิดเหตุการณ์ที่สนใจ จะเพิ่มขึ้น ..... เท่า เมื่อเทียบกับค่าเดิมของ x

- **ถ้าค่า Odds ratio <1** แสดงว่าเมื่อ X เพิ่มขึ้น ทำให้โอกาสของการเกิดเหตุการณ์ลดลง (ค่าสัมประสิทธิ์ถดถอย จะมีค่าเป็นลบ)

แปลผลได้ว่า เมื่อ X เปลี่ยนไป 1 หน่วย โอกาสที่จะเกิดเหตุการณ์ที่สนใจ จะลดลง ..... เท่า เมื่อเทียบกับค่าเดิมของ x

```
Call:
logitor(formula = admit ~ gpa + gre + rank, data = data)

Odds Ratio:
     OddsRatio Std. Err.       z     P>|z|
gpa  2.1749672 0.7122667  2.3727   0.01766 *
gre  1.0022966 0.0010943  2.1010   0.03564 *
rank 0.5711911 0.0726195 -4.4049 1.058e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

Gpa = 2.17 ซึ่ง > 1 หมายถึง เมื่อ x เปลี่ยนไป 1 หน่วย ค่า odds หรือโอกาสที่จะเกิดเหตุการณ์ จะเพิ่มขึ้น 2.17 เท่าเมื่อเทียบกับค่าเดิมของ x
นั่นคือ เมื่อ gpa เพิ่มขึ้น 1 หน่วย โอกาสที่จะเข้าศึกษาต่อเพิ่มขึ้น 2.17 เท่า

rank = 0.57 ซึ่ง < 1 หมายถึง เมื่อ x เปลี่ยนไป 1 หน่วย ค่า odds หรือโอกาสที่จะเกิดเหตุการณ์ จะลดลงไป 0.57 เท่าเมื่อเทียบกับค่าเดิมของ x
นั่นคือ เมื่ออันดับมหาลัยเพิ่มสูงขึ้นอีก 1 ระดับ โอกาสที่จะเข้าศึกษาต่อจะลดลงไป 0.57 เท่าจากระดับเดิม

Marginal effect

ทุกตัวแปรมีนัยสำคัญทางสถิติ ณ ระดับนัยสำคัญแตกต่างกันไป

```
Call:
logitmfx(formula = admit ~ gpa + gre + rank, data = data)

Marginal Effects:
            dF/dx   Std. Err.       z     P>|z|
gpa    0.16240171  0.06811125  2.3844   0.01711 *
gre    0.00047945  0.00022724  2.1099   0.03486 *
rank -0.11705079  0.02609555 -4.4855 7.275e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Gpa = 0.1624 หมายถึง ทุกๆ การเปลี่ยนแปลง (เพิ่มขึ้น) ของ gpa 1 หน่วย ทำให้ความน่าจะเป็นที่จะเข้าศึกษาต่อในระดับบัณฑิตศึกษาเพิ่มขึ้น 16.24%

Gre = 0.0004 หมายถึง ทุกๆ การเปลี่ยนแปลง (เพิ่มขึ้น) ของ gre 1 หน่วย ทำให้ความน่าจะเป็นที่จะเข้าศึกษาต่อในระดับบัณฑิตศึกษาเพิ่มขึ้น 0.04%

rank = -0.117 หมายถึง ทุกๆ การเปลี่ยนแปลงอันดับของมหาลัยขึ้นอีก 1 หน่วย ทำให้ความน่าจะเป็นที่จะเข้าศึกษาต่อในระดับบัณฑิตศึกษาลดลง 11.7%

- ในการประมาณ Probit model ใช้ฟังก์ชัน glm เช่นกัน

```
m1=glm(admit~gpa+gre+rank,family=binomial(link = “probit"))
```

Results of Probit model

```
Call:
glm(formula = admit ~ gpa + gre + rank, family = binomial(link = "probit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5626  -0.8920  -0.6403   1.1631   2.2097

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.0915037  0.6718360  -3.113  0.00185 **
gpa          0.4643598  0.1950263   2.381  0.01727 *
gre          0.0013982  0.0006487   2.156  0.03112 *
rank        -0.3317117  0.0745524  -4.449 8.61e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 459.48  on 396  degrees of freedom
AIC: 467.48
```

Marginal effect

```
Call:
probitmfx(formula = admit ~ gpa + gre + rank, data = data)

Marginal Effects:
            dF/dx   Std. Err.        z     P>|z|
gpa    0.16183104  0.06784070  2.3855   0.01706 *
gre    0.00048729  0.00022568  2.1592   0.03084 *
rank -0.11560270  0.02581064 -4.4789 7.504e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

Let’s Practice