Lecture 6: Binary Dependent Variable

Introduction to Econometrics,Spring 2025

Zhaopeng Qu

Business School,Nanjing University

April 04 2025

Review of the last lecture

Nonlinear Regression Functions

- How to extend linear OLS model to be nonlinear? Two categories based on which is nonlinear?

1. **Nonlinear in Xs**(the previous lecture)
    - **Polynomials,Logarithms and Interactions**
    - The multiple regression framework can be extended to handle regression functions that are nonlinear in one or more X.
    - the difference from a standard multiple OLS regression is *how to explain estimating coefficients*.

- So far the dependent variable (Y) has been continuous:
    - testscore
    - average hourly earnings
    - GDP growth rate
- What if the outcome variables(Y) is **discrete** or **limited**.

Nonlinear Regression Functions

2. **Nonlinear in β or Nonlinear in Y**

- Discrete(or Categorical) dependent variables
    - employment status: full-time,part-time,or none
    - ways to commute to work:by bus, car or walking
    - occupation(or sector) choices
    - demand for products: buy A, B or C
- Linear function is not a good prediction function. Need a certain function which parameters enter nonlinearly.
- **OLS** is not our first choice to estimate the model but the **Maximum Likelihood Estimation(MLE)** with the cost of pre-assumption about the known distribution families.
- Interpreting the results more difficult for the nonlinearity.

Discrete and Limited Dependent Variable Models

- **Discrete Models**:

- **Limited Dependent Variable**
  - **Censored data(删截)**: The information on the dependent variable of some observations is lost,but not data on the regressors.(Only some Xs are missing)
  - **Truncated data(断尾)**: Both dependent variable and independent. variables of some observations are missing for some reasons.(Both X and Y are missing for Y)
  - **Sample selection(样本选择)**: The sample are not randomly selected but based in part on values taken by a dependent variable.(Both X and Y are missing for Z)
- **Binary outcomes models** and **Multinomial choice models** are covered here.

Binary Outcome Models

Binary Outcome Models

- **Binary outcomes**
  - Y= get into college, or not; X = parental income.
  - Y= person smokes, or not; X = cigarette tax rate, income.
  - Y= mortgage application is accepted, or not; X = race, income, house characteristics, marital status
- Binary outcomes models:
  - **Logit Probability Model(LPM)**
  - **Logit model**
  - **Probit model**

The Linear Probability Model(LPM)

The Conditional Expectation

- If a outcome variable Y is **binary**, thus

$$Y = \begin{cases} 1 \text{ if } D = 1 \\ 0 \text{ if } D = 0 \end{cases}$$

- The expectation of Y is

$$E[Y] = 1 \times Pr(Y = 1) + 0 \times Pr(Y = 0) = Pr(Y = 1)$$

  which is the probability of $Y = 1$.

- Then we can extend it to the **conditional expectation** of Y equals to the the probability of $Y = 1$ conditional on Xs,thus

$$E[Y|X_{1i}, ..., X_{ki}] = Pr(Y = 1|X_{1i}, ..., X_{ki})$$

Multiple OLS Regression

- Suppose our regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

- Based on **Assumption 1**, thus

$$E[u_i | X_{1i}, ..., X_{ki}] = 0$$

- Then

$$E[Y | X_{1i}, ..., X_{ki}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$$

The Linear Probability Model

- The **conditional expectation** equals the probability that $Y_i = 1$ conditional on $X_{1i}, ..., X_{ki}$

$$E[Y|X_{1i}, ..., X_{ki}] = Pr(Y = 1|X_{1i}, ..., X_{ki})$$
$$= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$$

- Now a **Linear Probability Model** can be defined as following

$$Pr(Y = 1|X_{1i}, ..., X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$$

The Linear Probability Model

- The model does not change essentially.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

- The different part is the interpretation the coefficient.Now the **population coefficient** β_j

$$\frac{\partial Pr(Y_i = 1 | X_{1i}, ..., X_{ki})}{\partial X_j} = \beta_j$$

- β_j can be explained as **the change in the probability** that $Y = 1$ associated with **a unit change in** X_j

LPM and Multiple OLS

- Almost all of the tools of Multiple OLS regression can carry over to the LPM model.
  - **Assumptions** are the same as for general multiple regression model.
  - The coefficients can be also estimated by **OLS**.
  - Both **t-statistic** and **F-statistic** can be constructed as before.
  - The errors of the LPM are **always heteroskedastic**, so it is essential that **heteroskedasticity-robust s.e.** be used for inference.
  - One difference is that both original R^2 and adjusted-R^2 are not meaningful statistics now.

An Example: Mortgage Applications

- Most individuals who want to buy a house apply for a mortgage at a bank. And not all mortgage applications are approved.
- **Question**: What determines whether an application is approved or denied?
- **Boston HMDA data**: a data set on mortgage applications collected by the Federal Reserve Bank in Boston.

Variable	Description	Mean	SD
deny	= 1 if application is denied	0.120	0.325
pi_ratio	monthly loan payments / monthly income	0.331	0.107
black	= 1 if applicant is black	0.142	0.350

- Our linear probability model is

$$Pr(Y = 1|X_{1i}, X_{2i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

An Example: Mortgage Applications

- *Does the payment to income ratio affect whether or not a mortgage application is denied?*

$$\widehat{deny} = -0.080 + 0.604\ P/I\ ratio$$
$$(0.032)(0.098)$$

- The estimated OLS coefficient on the payment to income ratio

$$\hat{\beta}_1 = 0.604$$

- The estimated coefficient is significantly different from 0 at a 1% significance level as the t-statistic is over 6.

An Example: Mortgage Applications

- How should we interpret $\hat{\beta}_1$?
  - An original one: *payments/monthly income ratio increase 1(100%)*,then *probability being denied* will also increase **0.6(60%)**.
  - Another more reasonable one: payments/monthly income ratio increase **10%(0.1)**,then probability being denied will also increase **6%(0.06)**.
- **Question**: Does the effect matter? Or the magnitude of the effect is economically large enough.
- **Answer**: An option is comparing with the **mean** of dependent variable.
  - Here $deny\ rate = 0.12$ means that the deny ratio will increase $0.06/0.12 \times 100\% = 50\%$ if P/I Ratio increases 10%.

An Example: Mortgage Applications

- *What is the effect of race on the probability of denial*,**holding constant the P/I ratio?**
- **the differences between *black* applicants and *white* applicants.**

$$\widehat{deny} = \underset{(0.029)}{-0.091} + \underset{(0.089)}{0.559}\ P/I\ ratio + \underset{(0.025)}{0.177} black$$

- **The coefficient on black, 0.177, indicates that an African American applicant has a 17.7% higher probability of having a mortgage application denied than a white applicant, holding constant their payment-to-income ratio.**
- **This coefficient is significant at the 1% level (the t-statistic is 7.11).**

LPM Assumptions Similar to an OLS Regression

- Assumptions are the same as for general multiple regression model:
  1.
  2.
  3.
  4.
- Advantages of the linear probability model:
  - Easy to estimate and inference
  - Coefficient estimates are easy to interpret
  - Very useful under some circumstances like using IV.

LPM's Weakness: Heteroskedasticity

- The conditional variance of the error term u_i is always heteroskedasticity.

$$\operatorname{Var}\left(u_i \mid X_{1i}, \cdots, X_{ki}\right) \neq \sigma_u^2$$

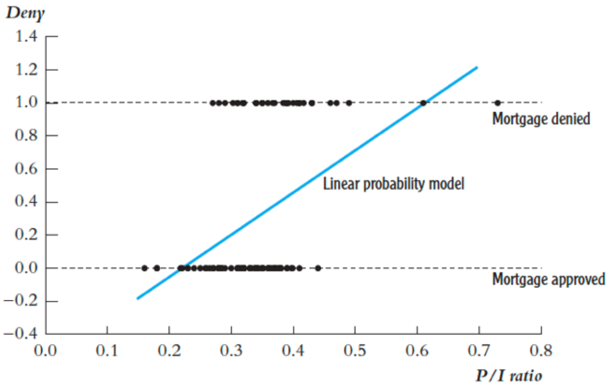
- Always use **heteroskedasticity robust standard errors** when estimating a linear probability model!

LPM's Weakness: Predicted values

- More serious problem: **the predicted probability can be below 0 or above 1!**

FIGURE 11.1 Scatterplot of Mortgage Application Denial and the Payment-to-Income Ratio

Mortgage applicants with a high ratio of debt payments to income (*P/I ratio*) are more likely to have their application denied (*deny* = 1 if denied, *deny* = 0 if approved). The linear probability model uses a straight line to model the probability of denial, conditional on the *P/I ratio*.



Nonlinear Probability Models

Introduction

- **Intuition**: Probabilities must be bounded between 0 and 1.
- To address this limitation, we consider a **general** probability model:

$$
\begin{aligned}
Pr(Y_i = 1 | X_1, ... X_k) &= G(Z) \\
&= G(\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i})
\end{aligned}
$$

  where $Z = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i}$

- The function $G(\cdot)$ must satisfy two essential conditions:
  - $0 \leq G(Z) \leq 1$
  - Monotonicity and continuity
- The central challenge is identifying an appropriate function $G(Z)$ that constrains predicted probabilities to the interval $(0, 1)$.
  - The **cumulative distribution function(c.d.f)**

Math Review: The cumulative distribution function(c.d.f)

- The cumulative distribution function (c.d.f) of a random variable X at a given value x is defined as the probability that X is smaller than x

$$F_X(x) = \Pr(X \leq x)$$

- Assume that the probability mass function or probability distribution function is $f_X(x)$, then the c.d.f is

$$F_X(x) = \begin{cases} \sum_{\substack{t \in \mathcal{X} \\ t \leq x}} f_X(t) \text{ if X is discrete} \\ \int_{-\infty}^{x} f_X(t) \mathrm{d}t \text{ if X is continuous} \end{cases}$$

- More importantly,the c.d.f satisfies
    - $0 \leq F_X(x) \leq 1$
    - monotonicity and continuity

Logit and Probit functions

- Two common nonlinear functions
  1. **Probit Model**

$$G(Z) = \Phi(Z) = \int_{-\infty}^{Z} \phi(Z)dZ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{t^2}{2}} dt$$

     which is the **standard normal** cumulative distribution function
  2. **Logit Model**

$$G(Z) = \frac{1}{1+e^{-Z}} = \frac{e^Z}{1+e^Z}$$

     which is the **logistic** cumulative distribution function.
  - where

$$Z = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$$

- Several reasons why these two are chosen:
  - good shapes, thus the predictions make more senses.
  - relatively easy to use and interprete them.

Probit Model

- Probit regression models the probability that $Y = 1$

$$Pr(Y_i = 1|X_1, ...X_k) = \Phi(\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i})$$

  - where $\Phi(Z)$ is the **standard normal** c.d.f, then we have

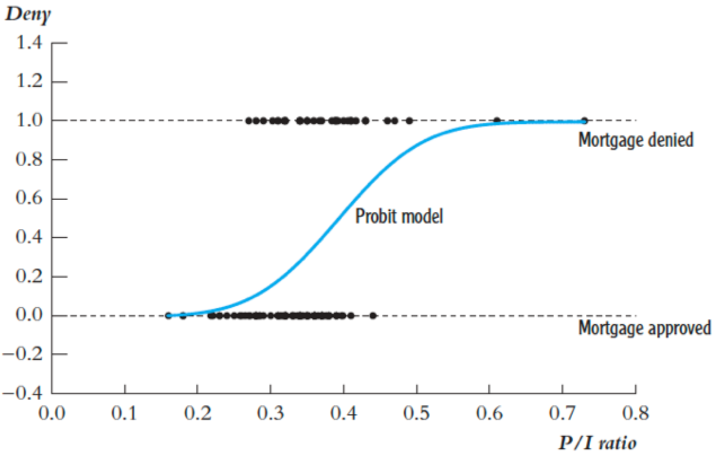
$$0 \leq \Phi(Z) \leq 1$$

- Then it make sure that the **predicted probabilities** of the probit model are between 0 and 1.

Probit Model: Shape and Prediction Value

FIGURE 11.2 Probit Model of the Probability of Denial, Given *P/I Ratio*

The probit model uses the cumulative normal distribution function to model the probability of denial given the payment-to-income ratio or, more generally, to model $\Pr(Y = 1 | X)$. Unlike the linear probability model, the probit conditional probabilities are always between 0 and 1.



Probit Model: Explaination to the Coefficient

- How should we interpret $\hat{\beta}_1$?
  - Recall $Z = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i}$
  - The coefficient β_j is **the change in the** Z-value rather than the probability arising from a unit change in X_j, holding constant other X_is.
- The effect on the predicted probability of a change in a regressor should be computed by the **general formula in the nonlinear regression model**(*Key concept 8.3*)
  1. computing the predicted probability for the initial value of the regressors,
  2. computing the predicted probability for the new or changed value of the regressors,
  3. taking their difference.

Probit Model: Explaination to the Coefficient

The Expected Change on Y of a Change in X_1 in the Nonlinear Regression Model (8.3)

KEY CONCEPT 8.1

The expected change in Y, ΔY, associated with the change in X_1, ΔX_1, holding $X_2, \ldots, X_k$ constant, is the difference between the value of the population regression function before and after changing X_1, holding $X_2, \ldots, X_k$ constant. That is, the expected change in Y is the difference:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k). \qquad (8.4)$$

The estimator of this unknown population difference is the difference between the predicted values for these two cases. Let $\hat{f}(X_1, X_2, \ldots, X_k)$ be the predicted value of Y based on the estimator $\hat{f}$ of the population regression function. Then the predicted change in Y is

$$\Delta \hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k). \qquad (8.5)$$

The Predicted Probability: one regressor

- Suppose the probit population regression model with only one regressors, X_1

$$Pr(Y = 1|X_1) = \Phi(Z) = \Phi(\beta_0 + \beta_1 X_1)$$

- Suppose the estimate result is $\hat{\beta}_0 = -2$ and $\hat{\beta}_1 = 3$,which means

$$Z = -2 + 3X_1$$

- **Question**: *how to compute the probability change of X_1 with a change from 0.4 to 0.5?*

The Predicted Probability: one regressor

- The probability that $Y=1$ when $X_1=0.4$, then $z=-2+3\times 0.4=-0.8$, then the predicted probability is

$$Pr(Y=1|X_1=0.4)=Pr(z\leq -0.8)=\Phi(-0.8)$$

- Likewise the probability that $Y=1$ when $X_1=0.5$, then $z=-2+3\times 0.5=-0.5$,the predicted probability is

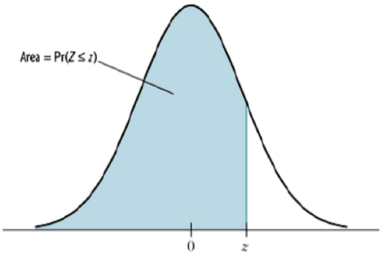
$$Pr(Y=1|X_1=0.5)=Pr(z\leq -0.5)=\Phi(-0.5)$$

- Then the difference is

$$Pr(Y=1|X_1=0.5)-Pr(Y=1|X_1=0.4)=$$
$$\Phi(-.5)-\Phi(-.8)=0.3085-0.2119=0.097$$

The Predicted Probability: one regressor

TABLE 1 The Cumulative Standard Normal Distribution Function, $\Phi(z) = \Pr(Z$ ″ $z)$



z	Second Decimal Value of z									
	0	1	2	3	4	5	6	7	8	9
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121

Example: Mortgage Applications

- **The probit model:**

$$Pr(Y = 1|X_1) = \Phi(Z) = \Phi(\beta_0 + \beta_1 X_1)$$

- **Question**: *Does the payment to income ratio affect whether or not a mortgage application is denied?*

$$\widehat{Pr(deny = 1|P/I\ ratio)} = \Phi(\underset{(0.16)}{-2.19} + \underset{(0.47)}{2.97} P/I\ ratio)$$

- **Answer**: Yes, the payment to income ratio affects whether or not a mortgage application is denied.

Example: Mortgage Applications

- **Question**: *What is the change in the predicted probability* that an application will be denied if **P/I ratio** increases from 0.3 to 0.4?
- The probability of denial when $P/I\ ratio = 0.3$

$$\Phi(-2.19 + 2.97 \times 0.3) = \Phi(-1.3) = 0.097$$

- The probability of denial when $P/I\ ratio = 0.4$

$$\Phi(-2.19 + 2.97 \times 0.4) = \Phi(-1.0) = 0.159$$

- **Answer**: The estimated change in the probability of denial is $0.159 - 0.097 = 0.062$, which means that the P/I ratio increase from *from 0.3 to 0.4*, the denial probability increase 6.2%.

Effect of a Change in X: When X is continuous

- the P/I ratio increase from
  - *0.3 to 0.4*, denial probability increase 6.2%.
  - *0.4 to 0.5*, denial probability increase 9.7%.
- **Marginal Effects** for X_j

$$\frac{\partial Pr(Y=1|X_1,...X_k)}{\partial X_j} = \phi(\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i}) \times \beta_j$$

- Where $\phi(\cdot)$ is the **probability distribution function(p.d.f)** of the standard normal c.d.f.
- Hence, the effect of a change in X depends on the starting value of X and other Xs like other nonlinear functions.

Effect of a Change in X: Marginal Effects

- Then the **Marginal Effects** varies with the point of evaluation
  - **Marginal Effect at a Representative Value** (MER):ME at $X = X^*$ (at representative values of the regressors)
  - **Marginal Effect at Mean** (MEM): ME at $X = \bar{X}$(at the sample mean of the regressors)
  - **Average Marginal Effect** (AME): average of ME at each $X = X_i$ (at sample values and then average)
- The most common one is MEM while the other two are not meaningless.

Example: Mortgage Applications

- The **Marginal Effect**

$$\frac{\partial Pr(deny=1|P/I\ ratio)}{\partial P/I\ ratio} = \phi(-2.19+2.97P/I\ ratio)\times 2.97$$

- Then **Marginal Effect at Mean** (MEM):(at the sample mean of the regressors: $P/I\ ratio_{mean}=0.331$

$$\left.\frac{\partial Pr(deny=1|P/I\ ratio)}{\partial P/I\ ratio}\right|_{at\ mean} = \phi(-2.19+2.97\times 0.331)\times 2.97$$
$$= \phi(-1.21)\times 2.97$$

- **The the effect of** $P/I\ ratio$ **change** 10%**(0.1) on the probability of deny is** $3.36\%(0.0336)$

Effect of a Change in X: When X is discrete

- If X_j is a *discrete* variable, then we should not rely on calculus in evaluating the effect on the response probability.
- Assume X_2 is a dummy variable, then partial effect of X_2 changing from 0 to 1:

$$G(\beta_0 + \beta_1 X_{1,i} + \beta_2 \times 1 + \ldots + \beta_k X_{k,i}) - G(\beta_0 + \beta_1 X_{1,i} + \beta_2 \times 0 + \ldots + \beta_k X_{k,i})$$

Example: Race in Mortgage Applications

- Mortgage denial (deny) and the payment-income ratio (P/I ratio) and race

$$\widehat{Pr(deny = 1|P/I\ ratio)} = \Phi(-2.26 + 2.74 P/I\ ratio + 0.71 black)$$
$$(0.16) \qquad (0.44) \qquad (0.083)$$

- **Question**: What is the effect of race on the probability of denial, holding constant the P/I ratio?

- The probability of denial when $black = 0$,thus whites(non-blacks) is

$$\Phi(-2.26 + 2.74 \times 0.3 + 0.71 \times 0) = \Phi(-1.43) = 0.075$$

- The probability of denial when $black = 1$,thus blacks is

$$\Phi(-2.26 + 2.74 \times 0.3 + 0.71 \times 1) = \Phi(-0.73) = 0.233$$

- **Answer**: The difference between whites and blacks at $P/Iratio = 0.3$ is $0.233 - 0.075 = 0.158$, which means probability of denial for blacks is 15.8% higher than that for whites.

Logit Model

Logistic Function

- Using the standard **logistic** cumulative distribution function

$$
\begin{aligned}
Pr(Y_i = 1|Z) &= \frac{1}{1+e^{-Z}} \\
&= \frac{e^Z}{1+e^Z}
\end{aligned}
$$

- As in the Probit model

$$Z = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i}$$

- Since $F(z) = Pr(Z \leq z)$ we have that the predicted probabilities of the logit model are also between 0 and 1.

Logit Model: Predicted Probabilities

- Suppose we have only one regressor X and $Z = -2 + 3X_1$
- We want to know the probability that $Y = 1$ when $X_1 = 0.4$
- Then

$$Z = -2 + 3 \times 0.4 = -0.8$$

- So the probability is

$$Pr(Y = 1 | X_1 = 0.4) =$$

Logit Model: Predicted Probabilities

- Suppose we have only one regressor X and $Z = -2 + 3X_1$
- We want to know the probability that $Y = 1$ when $X_1 = 0.4$
- Then

$$Z = -2 + 3 \times 0.4 = -0.8$$

- So the probability is

$$Pr(Y = 1|X_1 = 0.4) = Pr(Z \leq -0.8)$$

Logit Model: Predicted Probabilities

- Suppose we have only one regressor X and $Z = -2 + 3X_1$
- We want to know the probability that $Y = 1$ when $X_1 = 0.4$
- Then

$$Z = -2 + 3 \times 0.4 = -0.8$$

- So the probability is

$$\begin{aligned} Pr(Y = 1|X_1 = 0.4) = Pr(Z \leq -0.8) \\ = F(-0.8) \end{aligned}$$

Logit Model: Predicted Probabilities

- Suppose we have only one regressor X and $Z = -2 + 3X_1$
- We want to know the probability that $Y = 1$ when $X_1 = 0.4$
- Then

$$Z = -2 + 3 \times 0.4 = -0.8$$

- So the probability is

$$\begin{aligned} Pr(Y = 1|X_1 = 0.4) = Pr(Z \leq -0.8) \\ =& F(-0.8) \\ =& \frac{1}{1 + e^{-0.8}} \end{aligned}$$

Logit Model: Predicted Probabilities

- Suppose we have only one regressor X and $Z = -2 + 3X_1$
- We want to know the probability that $Y = 1$ when $X_1 = 0.4$
- Then

$$Z = -2 + 3 \times 0.4 = -0.8$$

- So the probability is

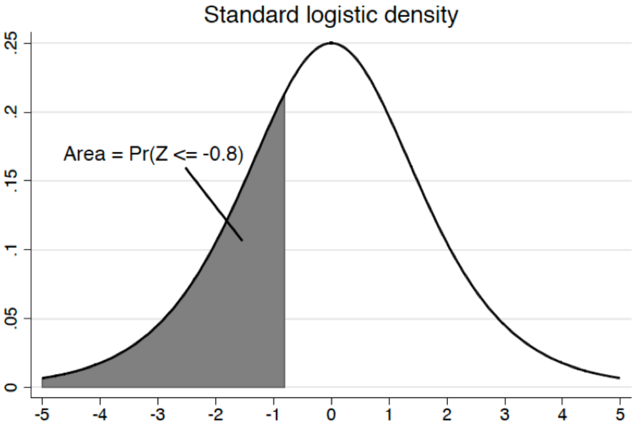
$$\begin{aligned} Pr(Y = 1 | X_1 = 0.4) = Pr(Z \leq -0.8) \\ =& F(-0.8) \\ =& \frac{1}{1 + e^{-0.8}} \\ =& 0.31 \end{aligned}$$

Logit Model: Predicted Probabilities

- $Pr(Y = 1) = Pr(Z \leq -0.8) = \frac{1}{1+e^{-0.8}} = 0.31$

Logit Model: Predicted Probabilities

- $Pr(Y = 1) = Pr(Z \leq -0.8) = \frac{1}{1+e^{-0.8}} = 0.31$



Logit Model: Explaination to the Coefficient

- How should we interpret $\hat{\beta}_1$?
- Similar to the Probit model,$Z = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i}$
  - The coefficient β_j can not be explained directly.
  - **the change in the Z-value** rather than the probability arising from a unit change in X_j, holding constant other X_i.
- However, Logit can be different from the Probit model in some way
  - The **odds ratio**

Logit Model: the Odds Ratio

- Let p is the conditional probability of $Y = 1$,then

$$p = Pr(Y_i = 1|Z) = \frac{e^Z}{1 + e^Z}$$

- Then $1 - p$ is the probability of $Y = 0$

$$1 - p = Pr(Y_i = 0|Z) = 1 - \frac{e^Z}{1 + e^Z} = \frac{1}{1 + e^Z}$$

- Then the **ratio** of probability of $Y = 1$ to the probability of $Y = 0$ is

$$\frac{p}{1 - p} = \frac{Pr(Y_i = 1|Z)}{Pr(Y_i = 0|Z)} = e^z$$

- the $\frac{p}{1-p}$ is called as **Odds Ratio** .

Logit Model: the Odds Ratio

- Then the logit model can be expressed as

$$ln(\frac{p}{1-p}) = Z = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i}$$

- Therefore $100 \times \hat{\beta}_j$ can be expressed that the **percentage change in odds ratio** arising from 1 unit change in X_j.

Example: Mortgage Applications

- **Logit Model: Mortgage denial (deny) and the payment-to-income ratio (P/I ratio)**

$$Pr(deny \widehat{= 1|P}/I\ ratio) = F(-4.03 + 5.88P/I\ ratio)$$
$$\qquad\qquad\qquad (0.359) \qquad (1.000)$$

- **If** $P/I\ ratio$ **increases** 10%**(0.1), then odds ratio of deny to accept will be increased** 58.8%**.**

Marginal Effect in logit model

- Then **Marginal Effect at Mean** (MEM):(at the sample mean of the regressors: $P/I\ ratio_{mean} = 0.331$

$$\frac{\partial Pr(deny = 1|P/I\ ratio)}{\partial P/I\ ratio}_{at\ mean} = f(-2.19 + 2.97 \times 0.331) \times 2.97$$
$$= f(-1.21) \times 2.97$$
$$= 0.526$$

- **The the effect of** $P/I\ ratio$ **change** 10%**(0.1) on the probability of deny is** $5.26\%(0.0526)$

Example: Mortgage Applications on Race

- **Logit Model: Mortgage denial (deny) and the payment-to-income ratio (P/I ratio) and race**

$$Pr(deny \widehat{= 1|P}/I\ ratio) = F(\underset{(0.35)}{-4.13} + \underset{(0.96)}{5.37} P/I\ ratio + \underset{(0.15)}{1.27} black)$$

Example: Mortgage Applications on Race

- **The predicted denial probability of a *white* applicant with $P/I\ ratio = 0.3$ is**

$$\frac{1}{1+e^{-(-4.13+5.37\times0.3+1.27\times0)}} = 0.074$$

- **The predicted denial probability of a *black* applicant with $P/I\ ratio = 0.3$ is**

$$\frac{1}{1+e^{-(-4.13+5.37\times0.3+1.27\times1)}} = 0.222$$

- **the difference is**

$$0.222 - 0.074 = 0.148 = 14.8\%$$

which indicates that the probability of denial for blacks is 14.8% higher than that for whites when $P/Iratio = 0.3$.

Maximum Likelihood Estimation(MLE) to Probit and Logit

Estimation and Inference in Probit and Logit Models

- How do we estimate $\beta_0, \beta_1, ..., \beta_k$?
- And how to get the sampling distribution of these estimators? $\sigma_{\hat{\beta}_j}$
- Logit and Probit models are nonlinear in the coefficients $\beta_0, \beta_1, ..., \beta_k$
  - These models cannot be estimated directly by OLS, but require **Nonlinear Least Squares (NLS)**.
  - In practice, **Maximum Likelihood Estimation** (MLE) is the most common method for estimating logit and probit models.

Review: Maximum Likelihood Estimation

- The **likelihood function** is a *joint probability distribution* of the data, treated as a function of the unknown coefficients.
- It describes the **probability** of the data we observed or the sample from the population, given the unknown coefficients.
- The **maximum likelihood estimator** (MLE) are the estimate values of the unknown coefficients that maximize the likelihood function.
- **MLE's logic**: the most likely function is the function to have produce the data we observed.

Review: Maximum Likelihood Estimation

- Random Variables Y_i have n observations, thus $Y_1, Y_2, Y_3, ...Y_n$ have a **joint density function** denoted

$$f_\theta(Y_1, Y_2, ..., Y_n) = f(Y_1, Y_2, ..., Y_n|\theta)$$

  - where θ is an unknown parameter.
- Given observed values $Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n$,the likelihood of θ is the function

$$likelihood(\theta) = f(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n|\theta) = f(\theta; y_1, ..., y_n)$$

  - which can be considered as a function of θ.
- Then the **Maximum Likelihood Estimation** to θ is a solution to the question

$$arg \max_{\hat{\theta}} f(\theta; Y_1 = y_1, ..., Y_n = y_n))$$

Maximum Likelihood Estimation of a Binary Variable

- Suppose we flip a coin which is yields heads ($Y = 1$) and tails ($Y = 0$). We want to estimate the probability p of heads($Y = 1$).
- Therefore, let $Y_i = 1(heads)$ be a **binary** variable that indicates whether or not a heads is observed.

$$Y_i = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p \end{cases}$$

- Then **the probability mass function** for a single observation is a Bernoulli distribution

$$Pr(Y_i) = \begin{cases} p & \text{when } Y_i = 1 \\ 1 - p & \text{when } Y_i = 0 \end{cases}$$

- Which can be transform into a **probability density function** as

$$Pr(Y_i = y) = Pr(Y_i = 1)^y (1 - Pr(Y_i = 1))^{1-y} = p^y (1 - p)^{1-y}$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 1: *write down the likelihood function, the joint probability distribution of the data*

- Since $Y_1, ..., Y_n$ are **i.i.d**,the joint probability distribution of the observations, thus the Likelihood function is the **product** of the individual distributions

$$f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n) = Pr(Y_1 = y_1, ..., Y_n = y_n)$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 1: *write down the likelihood function, the joint probability distribution of the data*

- Since $Y_1, ..., Y_n$ are **i.i.d**,the joint probability distribution of the observations, thus the Likelihood function is the **product** of the individual distributions

$$\begin{aligned} f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n) &= Pr(Y_1 = y_1, ..., Y_n = y_n) \\ &= Pr(Y_1 = y_1) \times ... \times Pr(Y_n = y_n) \end{aligned}$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 1: *write down the likelihood function, the joint probability distribution of the data*

- Since $Y_1, ..., Y_n$ are **i.i.d**,the joint probability distribution of the observations, thus the Likelihood function is the **product** of the individual distributions

$$\begin{aligned} f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n) &= Pr(Y_1 = y_1, ..., Y_n = y_n) \\ &= Pr(Y_1 = y_1) \times ... \times Pr(Y_n = y_n) \\ &= p^{y_1}(1-p)^{1-y_1} \times ... \times p^{y_n}(1-p)^{1-y_n} \end{aligned}$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 1: *write down the likelihood function, the joint probability distribution of the data*

- Since $Y_1, ..., Y_n$ are **i.i.d**,the joint probability distribution of the observations, thus the Likelihood function is the **product** of the individual distributions

$$
\begin{aligned}
f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n) &= Pr(Y_1 = y_1, ..., Y_n = y_n) \\
&= Pr(Y_1 = y_1) \times ... \times Pr(Y_n = y_n) \\
&= p^{y_1}(1-p)^{1-y_1} \times ... \times p^{y_n}(1-p)^{1-y_n} \\
&= p^{(y_1+y_2+...+y_n)}(1-p)^{n-(y_1+y_2+...+y_n)}
\end{aligned}
$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 1: *write down the likelihood function, the joint probability distribution of the data*

- Since $Y_1, ..., Y_n$ are **i.i.d**,the joint probability distribution of the observations, thus the Likelihood function is the **product** of the individual distributions

$$
\begin{aligned}
f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n) &= Pr(Y_1 = y_1, ..., Y_n = y_n) \\
&= Pr(Y_1 = y_1) \times ... \times Pr(Y_n = y_n) \\
&= p^{y_1}(1-p)^{1-y_1} \times ... \times p^{y_n}(1-p)^{1-y_n} \\
&= p^{(y_1+y_2+...+y_n)}(1-p)^{n-(y_1+y_2+...+y_n)} \\
&= p^{\sum y_i}(1-p)^{n-\sum y_i}
\end{aligned}
$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 2: *Write down the maximization problem*

- More easier to maximize the **logarithm** of the likelihood function

$$
\begin{aligned}
ln(f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n)) &= ln\left(\prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}\right) \\
&= \left(\sum y_i\right) ln(p) + \left(n - \sum y_i\right) ln(1-p)
\end{aligned}
$$

- Since the logarithm is a **strictly increasing** function, maximizing the likelihood or the log likelihood will give the same estimator.
- Then the **maximization** problem is

$$
arg \max_{\hat{p}} ln(f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n))
$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 3: *Maximize the likelihood function*

- **F.O.C**: taking the derivative and setting it to zero.

Maximum Likelihood Estimation of a Binary Variable

MLE Step 3: *Maximize the likelihood function*

- **F.O.C: taking the derivative and setting it to zero.**

$$\Rightarrow \frac{d}{dp}\left[\left(\sum y_i\right) ln(p) + \left(n - \sum y_i\right) ln(1-p)\right] = 0$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 3: *Maximize the likelihood function*

- **F.O.C: taking the derivative and setting it to zero.**

$$\Rightarrow \frac{d}{dp}\left[\left(\sum y_i\right)ln(p) + \left(n - \sum y_i\right)ln(1-p)\right] = 0$$

$$\Rightarrow \frac{\sum y_i}{p} - \frac{n - \sum y_i}{1-p} = 0$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 3: *Maximize the likelihood function*

- **F.O.C: taking the derivative and setting it to zero.**

$$\Rightarrow \frac{d}{dp}\left[\left(\sum y_i\right)ln(p) + \left(n - \sum y_i\right)ln(1-p)\right] = 0$$

$$\Rightarrow \frac{\sum y_i}{p} - \frac{n - \sum y_i}{1-p} = 0$$

$$\Rightarrow \sum y_i(1-p) = (n - \sum y_i)p$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 3: *Maximize the likelihood function*

- **F.O.C**: taking the derivative and setting it to zero.

$$\Rightarrow \frac{d}{dp}\left[\left(\sum y_i\right) ln(p) + \left(n - \sum y_i\right) ln(1-p)\right] = 0$$

$$\Rightarrow \frac{\sum y_i}{p} - \frac{n - \sum y_i}{1-p} = 0$$

$$\Rightarrow \sum y_i(1-p) = (n - \sum y_i)p$$

$$\Rightarrow \sum y_i - p\sum y_i = np - p\sum y_i$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 3: *Maximize the likelihood function*

- **F.O.C: taking the derivative and setting it to zero.**

$$
\begin{aligned}
&\Rightarrow \frac{d}{dp}\left[\left(\sum y_i\right) ln(p) + \left(n - \sum y_i\right) ln(1-p)\right] = 0 \\
&\Rightarrow \frac{\sum y_i}{p} - \frac{n - \sum y_i}{1-p} = 0 \\
&\Rightarrow \sum y_i (1-p) = (n - \sum y_i) p \\
&\Rightarrow \sum y_i - p \sum y_i = np - p \sum y_i \\
&\Rightarrow p = \frac{1}{n} \sum y_i
\end{aligned}
$$

Maximum Likelihood Estimation of a Binary Variable

MLE Step 3: *Maximize the likelihood function*

- **F.O.C**: taking the derivative and setting it to zero.

$$
\begin{aligned}
&\Rightarrow \frac{d}{dp}\left[\left(\sum y_i\right) ln(p) + \left(n - \sum y_i\right) ln(1-p)\right] = 0 \\
&\Rightarrow \frac{\sum y_i}{p} - \frac{n - \sum y_i}{1-p} = 0 \\
&\Rightarrow \sum y_i (1-p) = (n - \sum y_i) p \\
&\Rightarrow \sum y_i - p \sum y_i = np - p \sum y_i \\
&\Rightarrow p = \frac{1}{n} \sum y_i
\end{aligned}
$$

- Then the MLE estimator for a binary variable, p, is $\hat{p}_{MLE} = \frac{1}{n}\sum y_i = \overline{Y}$

MLE of the Probit Model

- **Assume our probit model is**

$$P(Y_i = 1|X_i) = \Phi(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}) = p_i$$

- **Step 1: write down the likelihood function**

$$f_{probit}(\beta_0, ..., \beta_k; Y_1, ..., Y_n|X_{1i}, ..., X_{ki}, i = 1, ..., n) = Pr(Y_1 = y_1, .., Y_n = y_n)$$

MLE of the Probit Model

- Assume our probit model is

$$P(Y_i = 1|X_i) = \Phi(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}) = p_i$$

- **Step 1**: write down the likelihood function

$$\begin{aligned} f_{probit}(\beta_0, ..., \beta_k; Y_1, ..., Y_n|X_{1i}, ..., X_{ki}, i = 1, ..., n) &= Pr(Y_1 = y_1, .., Y_n = y_n) \\ &= Pr(Y_1 = y_1) \times ... \times Pr(Y_n = y_n) \end{aligned}$$

MLE of the Probit Model

- Assume our probit model is

$$P(Y_i = 1|X_i) = \Phi(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}) = p_i$$

- **Step 1**: write down the likelihood function

$$\begin{aligned} f_{probit}(\beta_0, ..., \beta_k; Y_1, ..., Y_n | X_{1i}, ..., X_{ki}, i = 1, ..., n) &= Pr(Y_1 = y_1, .., Y_n = y_n) \\ &= Pr(Y_1 = y_1) \times ... \times Pr(Y_n = y_n) \\ &= p^{y_1}(1-p)^{1-y_1} \times ... \times p^{y_n}(1-p)^{1-y_n} \end{aligned}$$

MLE of the Probit Model

- Assume our probit model is

$$P(Y_i = 1|X_i) = \Phi(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}) = p_i$$

- **Step 1**: write down the likelihood function

$$\begin{aligned} f_{probit}(\beta_0, ..., \beta_k; Y_1, ..., Y_n|X_{1i}, ..., X_{ki}, i = 1, ..., n) &= Pr(Y_1 = y_1, .., Y_n = y_n) \\ &= Pr(Y_1 = y_1) \times ... \times Pr(Y_n = y_n) \\ &= p^{y_1}(1-p)^{1-y_1} \times ... \times p^{y_n}(1-p)^{1-y_n} \end{aligned}$$

$$= \left[\Phi(\beta_0 + \beta_1 X_{11} + ... + \beta_k X_{k1})^{y_1}(1 - \Phi(\beta_0 + \beta_1 X_{11} + ... + \beta_k X_{k1}))^{1-y_1}\right] \times$$

$$... \times \left[\Phi(\beta_0 + \beta_1 X_{1n} + ... + \beta_k X_{kn})^{y_n}(1 - \Phi(\beta_0 + \beta_1 X_{1n} + ... + \beta_k X_{kn}))^{1-y_n}\right]$$

MLE of the Probit Model

- **Step 2**: Maximize the log likelihood function

$$ln(f_{probit}(\beta_0, ..., \beta_k; Y_1, ..., Y_n | X_{1i}, ..., X_{ki}, i = 1, ..., n))$$

MLE of the Probit Model

- **Step 2**: Maximize the log likelihood function

$$
\begin{aligned}
& ln(f_{probit}(\beta_0, ..., \beta_k; Y_1, ..., Y_n | X_{1i}, ..., X_{ki}, i = 1, ..., n)) \\
&= \sum_i^n y_i \times ln[\Phi(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})] \\
&+ \sum_i^n (1 - y_i) \times ln[1 - \Phi(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})]
\end{aligned}
$$

MLE of the Probit Model

- **Step 2**: Maximize the log likelihood function

$$
\begin{aligned}
& ln(f_{probit}(\beta_0, ..., \beta_k; Y_1, ..., Y_n | X_{1i}, ..., X_{ki}, i = 1, ..., n)) \\
&= \sum_{i}^{n} y_i \times ln[\Phi(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})] \\
&+ \sum_{i}^{n} (1 - y_i) \times ln[1 - \Phi(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})]
\end{aligned}
$$

- **Then the maximization problem is**

$$
arg \max_{\hat\beta_0, \hat\beta_1, ..., \hat\beta_k} ln(f_{probit}(\beta_0, \beta_1, ..., \beta_k; Y_1 = y_1, ..., Y_n = y_n | X_{1i}, ..., X_{ki}, i = 1, ..., n))
$$

MLE of the Logit Model

- **Step 1** write down the likelihood function

$$Pr(Y_1 = y_1, ..., Y_n = y_n) = p^{y_1}(1-p)^{1-y_1} \times ... \times p^{y_n}(1-p)^{1-y_n}$$

- Similar to the Probit model but with a different function for p_i

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})}}$$

MLE of the Logit Model

- **Step 2**: Maximize the log likelihood function

$$ln(f_{logit}(\beta_0, ..., \beta_k; Y_1, ..., Y_n | X_{1i}, ..., X_{ki}, i = 1, ..., n))$$

MLE of the Logit Model

- **Step 2**: Maximize the log likelihood function

$$
\begin{aligned}
ln(f_{logit}(\beta_0, ..., \beta_k; Y_1, ..., Y_n | X_{1i}, ..., X_{ki}, i = 1, ..., n)) \\
= \sum y_i \times ln\left(\frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})}}\right) \\
+ \sum (1 - y_i) \times ln\left(\frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})}}\right)
\end{aligned}
$$

MLE of the Logit Model

- **Step 2**: Maximize the log likelihood function

$$
\begin{aligned}
&ln(f_{logit}(\beta_0, ..., \beta_k; Y_1, ..., Y_n | X_{1i}, ..., X_{ki}, i = 1, ..., n)) \\
&= \sum y_i \times ln\left(\frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})}}\right) \\
&+ \sum (1 - y_i) \times ln\left(\frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})}}\right)
\end{aligned}
$$

- Then the maximization problem is

$$
\underset{\hat{\beta}_0, \hat{\beta}_1, .., \hat{\beta}_k}{arg\max}\, ln(f_{logit}(\beta_0, ..., \beta_k; Y_1 = y_1, ..., Y_n = y_n | X_{1i}, ..., X_{ki}, i = 1, ..., n))
$$

Computation of MLE Estimators

- In most cases the computation of maximum likelihood estimators is not easy to obtain since the first order conditions **do not have closed form solutions necessarily**.
- We can still obtain the values of estimators using **numerical algorithm** with iterative methods.
- One of common methods is **Newton-Raphson Method** based on low order *Taylor series expansions*.

Measures of Fit

- R^2 is a poor measure of fit for the linear probability model. This is also true for probit and logit regression.
- Two measures of fit for models with binary dependent variables

1. *fraction correctly predicted*
    - If $Y_i = 1$ and the predicted probability exceeds 50% or if $Y_i = 0$ and the predicted probability is less than 50%, then Y_i is said to be correctly predicted.

Measures of Fit

2. **The pseudo-R2**
   - The $pseudo - R^2$ compares the value of the likelihood of the estimated model to the value of the likelihood when none of the Xs are included as regressors.

$$pseudo - R^2 = 1 - \frac{ln(f^{max}_{probit})}{ln(f^{max}_{bernoulli})}$$

   - f^{max}_{probit} is the value of the maximized probit likelihood (which includes the X's)
   - $f^{max}_{bernoulli}$ is the value of the maximized Bernoulli likelihood (the probit model excluding all the X's).

Statistical inference based on the MLE

- It can be prove that under very general conditions,the MLE estimator is **unbiased**,**consistent**, **asymptotic normally distributed** in large samples. See the Appendix for MLE in OLS regression.
- Because the MLE is normally distributed in large samples, statistical inference about the probit and logit coefficients based on the MLE proceeds in the **same way** as inference about the linear regression function coefficients based on the OLS estimator.
- That is, hypothesis tests are performed using the **t-statistic**(or **z-statistic**) and confidence intervals are also formed using the **normal distribution**.
  - For example, the **95% confidence intervals** are formed as 1.96 standard errors.

Statistical inference based on the MLE

- Testing of **joint hypotheses** on multiple coefficients are very similar to the **F-statistic** which is discussed in multiple OLS model.
- The **likelihood ratio test** is based on comparing the log likelihood values of the unrestricted and the restricted model. The test statistic is

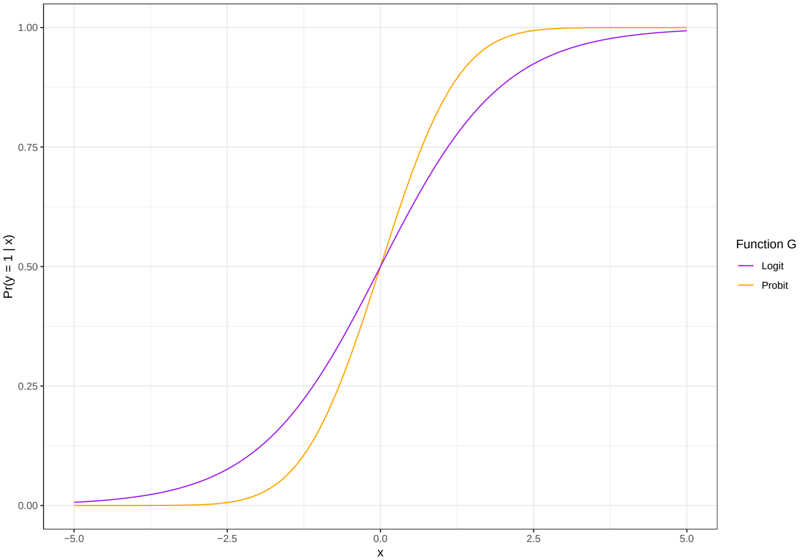
$$LR = 2(logL_{ur} - LogL_r) \sim \chi^2_q$$

  - where $logL_{ur}$ is the log likelihood of the unrestricted model, $logL_r$ is the log likelihood of the restricted model, and q is the number of restrictions being tested.
- Because the MLE maximizes the log-likelihood function, dropping variables generally leads to a **smaller—or at least no larger—log-likelihood**.
- The question is whether the fall in the log-likelihood is **large enough** to conclude that the dropped variables are important.
  - Therefore, the likelihood ratio test statistic is always non-negative.

Comparing the LPM,Probit and Logit

- All three models: *linear probability, probit, and logit* are just approximations to the unknown population regression function $E(Y|X) = Pr(Y = 1|X)$.
  - LPM is easiest to use and to interpret, but it cannot capture the nonlinear nature of the true population regression function.
  - Probit and logit regressions model this nonlinearity in the probabilities, but their regression coefficients are more difficult to interpret.
- So which should you use in practice?
  - *There is no one right answer*, and different researchers use different models.
  - *Probit and logit regressions frequently produce similar results*.

Logit v.s. Probit



Comparing the LPM,Probit and Logit

- The marginal effects and predicted probabilities are much more similar across models.
- Coefficients can be compared across models, using the following rough conversion factors (Amemiya 1981)

$$\hat{\beta}_{logit} \simeq 4\hat{\beta}_{ols}$$
$$\hat{\beta}_{probit} \simeq 2.5\hat{\beta}_{ols}$$
$$\hat{\beta}_{logit} \simeq 1.6\hat{\beta}_{probit}$$

Example: Mortgage Applications(short regression)

Dependent variable: *deny* = 1 if mortgage application is denied, = 0 if accepted			
regression model	LPM	Probit	Logit
black	0.177***	0.71***	1.27***
	(0.025)	(0.083)	(0.15)
P/I ratio	0.559***	2.74***	5.37***
	(0.089)	(0.44)	(0.96)
constant	-0.091***	-2.26***	-4.13***
	(0.029)	(0.16)	(0.35)
difference Pr(*deny*=1) between black and white applicant when *P/I ratio*=0.3	17.7%	15.8%	14.8%

Multinomial Regression Models

Introduction

- The multinomial regression model is an extension of the binary dependent variable model to allow for more than two categories of the dependent variable, shuch as the choice of occupation, transportation mode, etc.
  - **Occupation choice**: self-employed, government employee, private sector employee, etc.
  - **Major Choices**: Economics, Statistics, Computer Science, etc.
  - **Transportation mode**: car, bus, bike, subway, etc.
  - **Demand for goods**: Coke, Pepsi, Sprite, etc.
- One important feature: the outcomes cannot be ordered in any natural way.

Multinomial Regression Models

- There are m **mutually-exclusive** alternatives:
  - Y_i takes value j if the outcome is alternative j , $j = 1, ..., m$, where $m \geq 2$.

$$Y_i = \begin{cases} 1 & \text{if the outcome is A} \\ 2 & \text{if the outcome is B} \\ \vdots & \vdots \\ m & \text{if the outcome is M} \end{cases}$$

- The respondents face those m alternatives and can only choose one among them.
- **Question**: Can we use an OLS regression to model this situation? Like

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + u_i$$

- **Answer**: No, because the dependent variable's values lack a meaningful order.

Multinomial Regression Models

- Naturally we can use a binary choice model(LPM, probit, logit) to model the situation by grouping all categories into two major ones.
- Suppose the i individual's choice is J,then we can turn the Y_{ij} into a binary variable.

$$Y_{ij} = \begin{cases} 1 & \text{if the outcome is J} \\ 0 & \text{if the outcome is not J} \end{cases}$$

  - $Y_{ij} = 1$ if alternative J is chosen and Y_{ij} = 0 for all non-chosen alternatives for any individual i.
- Though **Binary choice models** could potentially be used,this is not ideal.
  - We can not compare the coefficients across different alternatives directly.
  - Those alternatives are mutually exclusive.

Multinomial Regression Models

- **However, if Y_i takes value j if the outcome is alternative j , $j = 1, ..., m$, then the probability** that the outcome is alternative j can be modeled as

$$P(Y_i = J|X_{1i}, ..., X_{ki}) = f_{ij}(\beta : X) = p_{ij}$$

- Then the p.d.f of individual i' choice among alternatives j is

$$f_i(\beta : X) = p_{i1}^{y_{i1}} \times p_{i2}^{y_{i2}} \times \ldots \times p_{im}^{y_{im}} = \prod_{j=1}^{m} p_{ij}^{y_{ij}}$$

- Using MLE estimation to maximize the log-likelihood function to solve the parameters β.

$$\ln L(\cdot) = \ln \left(\prod_{i=1}^{N} f_{ij} \left(\beta : X\right) \right) = \ln \left(\prod_{i=1}^{N} \prod_{j=1}^{m} p_j^{y_j} \right) = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln p_{ij}$$

Multinomial Logit Model

- The functional form is the key to solve the multinomial regression model.Likewise, there are two functional forms for the multinomial models:
  - **Logit** and **Probit**
- The **multinomial logit model** or **M-logit** is the most common form of multinomial regression model.
- As dealing with the categorized independent variables in linear regression models, we still need a reference category, **the base category**, to compare with alternatives.
  - Which is the necessary condition for the identification of the model.

Multinomial Logit Model

- The base category will not be included in the model, as we avoid the **dummy variable trap** in linear regression with **categorized independent variables**.
- Assume the reference category is $J = 1$, and let $\beta^1 = 0$, then the probability that the outcome is alternative j can be expressed as following:

Multinomial Logit Model

- The base category will not be included in the model, as we avoid the **dummy variable trap** in linear regression with **categorized independent variables**.
- Assume the reference category is $J = 1$, and let $\beta^1 = 0$, then the probability that the outcome is alternative j can be expressed as following:

$$P(Y_i = J|X_{1i}, ..., X_{ki}) = \begin{cases} \frac{1}{1+\sum_{j=2}^{M} exp(X'\beta^j)} & \textbf{if J=1} \\ \frac{exp(X'\beta^J)}{1+\sum_{j=2}^{M} exp(X'\beta^j)} & \textbf{if J=2,3,...,M} \end{cases}$$

Multinomial Logit Model: Coefficients Interpretation

- Then *the probability that the outcome is alternative* J can be expressed as following under the distributional assumption of the error term:(Skip the derivation, you can prove it by yourself.)

$$p_{ij} = P(Y_i = J | X_{1i}, ..., X_{ki}) = \frac{exp(X'\beta^J)}{\Sigma_{j=1}^{M} exp(X'\beta^j)}$$

- **Question**: How to interpret the coefficients?
- **Answer**:the parameters of the multinomial logit model are difficult to interpret. Neither the sign nor the magnitude of the parameter has an direct intuitive meaning.

Multinomial Logit Model: Marginal Effects

- The **marginal probability effects** of the multinomial logit model for a change of X_k for choice J can be calculated as follows:

$$MPE_{ijk} = \frac{\partial p_{ij}}{\partial X_{ik}} = p_{ij}\left(\beta_{jk} - \sum_{j=1}^{M} p_{ij}\beta_{jk}\right)$$

- Then **the average marginal probability effects (AMPE)** for a change of X_k can be calculated as follows:

$$\widehat{AMPE}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\widehat{MPE}_{ijk} \quad (j = 1, \ldots, M)$$

- **the marginal probability effects at mean(MPEM)** for a change of X_k can be calculated as

$$\widehat{MPEM}_{jk} = \bar{p}_{ij}\left(\hat{\beta}_{jk} - \sum_{j=1}^{M}\bar{p}_{ij}\right) \quad (j = 1, \ldots, M)$$

Multinomial Logit Model: odds/risk ratio

- Recall the **odds ratio** in the binary choice model, thus the **ratio of probability** of $Y = 1$ to the probability of $Y = 0$ is

$$\frac{p}{1-p} = \frac{Pr(Y_i = 1|Z)}{Pr(Y_i = 0|Z)} = e^z$$

- Then the **odds ratio of the multinomial logit model** is the ratio of the probability of choosing alternative J to the probability of choosing the base category 1 is

$$\frac{p_j}{p_1} = \frac{Pr(Y_i = j|Z)}{Pr(Y_i = 1|Z)} = \frac{\frac{exp(X'\beta^J)}{1+\sum_{j=2}^{M} exp(X'\beta^j)}}{\frac{1}{1+\sum_{j=2}^{M} exp(X'\beta^j)}} = exp(X'\beta^J)$$

- Therefore $100 \times \hat{\beta}_k$ can be expressed that **the percentage change in odds ratio for choice J relative to the base category** 1 arising from a unit change in X_k.

Multinomial Logit Model: Strong Assumption

- Likewise,the odds between two alternatives j and k is

$$\frac{p_j}{p_k} = \frac{Pr(Y_i = j|Z)}{Pr(Y_i = k|Z)} = \frac{\frac{exp(X'\beta^j)}{1+\sum_{j=2}^{M} exp(X'\beta^j)}}{\frac{exp(X'\beta^k)}{1+\sum_{j=2}^{M} exp(X'\beta^j)}} = exp(X'(\beta^j - \beta^k))$$

- Then the log odds ratio is

$$log(\frac{p_j}{p_k}) = X'(\beta^j - \beta^k)$$

- It **only** depends on the corresponding two probabilities (but not those of other alternatives). This is known as **independence of irrelevant alternatives (IIA)**.
- Essentially, the IIA assumption requires that all the alternatives are **independent of each other**.

Independence of Irrelevant Alternatives (IIA)

- The IIA assumption is a strong assumption, which is not always satisfied in practice.
  - Example: **Transportation Mode Choice**: suppose a person chooses between car, subway, and bus
  - Under IIA, the ratio of probabilities between any two choices (e.g., car vs subway) should not change if a third option (bus) is added or removed.
  - However, in reality, if bus service is removed, many bus riders might switch to subway rather than car, violating IIA
  - This is because subway and bus are **closer substitutes** than car and bus.
- Therefore, we have more flexible models to relax the IIA assumption as **nested logit model** and **mixed logit model**.(You may learn them in some advanced courses in your future study.)

Wrap-up

- Multinomial Probit Model
  - The multinomial probit model is a generalization of the probit model to the case of more than two outcomes.
  - The model assumes that the error terms are normally distributed.
  - The model is more flexible than the multinomial logit model, but it is computationally more demanding.
- Extension for relaxing the IIA assumption:
  - **Nested Logit Model**
  - **Mixed Logit Model**
  - **Conditional Logit Model**
- Another extension: **Ordered Probit or Logit models**
  - The ordered probit or logit models are used when the dependent variable is ordinal.

A Lastest Application: Jia,Lan and Miquel(2021)

Parental background and Entrepreneurship in China

- Ruixue Jia(贾瑞雪), Xiaohuan Lan(兰小欢) and Gerard Padrói Miquel, “Doing Business in China: Parental background and government intervention determine who owns business”,The Journal of Development Economics,Volume 151, June 2021.
- **Main Question**:
  1. the parental determinants of entrepreneurship in China.
  2. how the parental determinants of entrepreneurship vary with government intervention in the economy.

Jia,Lan and Miquel(2021): Data

1. Individual-level data:
    - China General Social Survey (GCSS) 2006,2008,2010,2012,2013
    - 31 provinces, 22801 urban respondents.
2. Province-level data:
    - China Statistic Yearbooks.

Jia,Lan and Miquel(2021): Main Variables

- **Independent Variables**: **cadre parents** and **entrepreneur parents**
  - **cadre parents**: *“does a parent work in government or in a public organization affiliated with the government?”*
  - **entrepreneur parents**: *business owner + self-employed*
- **Dependent Variables**: whether the respondent is
  - **business owner**: all owners of incorporated businesses, who must pay corporation tax and follow corporation law.
  - **self-employment**: owners of non-incorporated small businesses.
  - **goverment employee**: work in government or in a public organization affiliated with the government.

Parental Background and Doing Business

- **Goal**: examine the difference in the probability of being in different occupations between those with *entrepreneur parents, cadre parents and others*.
- Linear Probability Model:

$$Pr(Y=1|X) = \beta_1 \textbf{CardreParent}_i + \beta_2 \textbf{EntreParent}_i + \gamma X_i + Prov_p \times Year_t + u_{ipt}$$

  - Y_i is a dummy indicating the respondent's occupation,all the other occupations grouped together in the reference group.
  - X_i are individual-level characteristics such as gender,age, marital status, college education or not, and minority status.
  - $Prov_p \times Year_t$ are the province-by-year fixed effects.

Empirical Results: LPM

Table 3A

Parent background and child occupations: OLS estimates.

	(1)	(2)	(3)	(4)	(5)	(6)
	Government worker (0/1, mean = 0.217)		Business owner (0/1, mean = 0.022)		Self-employed (0/1, mean = 0.107)	
Cadre Parent	0.144***	0.115***	0.006**	0.003	−0.009*	−0.011**
	(0.009)	(0.009)	(0.003)	(0.003)	(0.005)	(0.005)
Entrepreneur Parent	−0.006	−0.006	0.016***	0.014**	0.063***	0.057***
	(0.012)	(0.011)	(0.006)	(0.006)	(0.013)	(0.013)
Province FE*Year FE	Y	Y	Y	Y	Y	Y
Individual Characteristics		Y		Y		Y
Observations	22,801	22,801	22,801	22,801	22,801	22,801
R-squared	0.057	0.139	0.015	0.022	0.039	0.067

- **Cadre Parents** increase the probability of being government workers(11.5%).
- **Entrepreneur Parents** do not.

Empirical Results: LPM

Table 3A
Parent background and child occupations: OLS estimates.

	(1)	(2)	(3)	(4)	(5)	(6)
	Government worker (0/1, mean = 0.217)		Business owner (0/1, mean = 0.022)		Self-employed (0/1, mean = 0.107)	
Cadre Parent	0.144***	0.115***	0.006**	0.003	−0.009*	−0.011**
	(0.009)	(0.009)	(0.003)	(0.003)	(0.005)	(0.005)
Entrepreneur Parent	−0.006	−0.006	0.016***	0.014**	0.063***	0.057***
	(0.012)	(0.011)	(0.006)	(0.006)	(0.013)	(0.013)
Province FE*Year FE	Y	Y	Y	Y	Y	Y
Individual Characteristics		Y		Y		Y
Observations	22,801	22,801	22,801	22,801	22,801	22,801
R-squared	0.057	0.139	0.015	0.022	0.039	0.067

- **Entrepreneur Parents** increase the probability of being business owner(1.6%).
- **Cadre Parents** also increase the probability of being business owner(0.6%). However, the effect will go away when controlling individual characteristics.

Empirical Results: LPM

Table 3A
Parent background and child occupations: OLS estimates.

	(1)	(2)	(3)	(4)	(5)	(6)
	Government worker (0/1, mean = 0.217)		Business owner (0/1, mean = 0.022)		Self-employed (0/1, mean = 0.107)	
Cadre Parent	0.144***	0.115***	0.006**	0.003	−0.009*	−0.011**
	(0.009)	(0.009)	(0.003)	(0.003)	(0.005)	(0.005)
Entrepreneur Parent	−0.006	−0.006	0.016***	0.014**	0.063***	0.057***
	(0.012)	(0.011)	(0.006)	(0.006)	(0.013)	(0.013)
Province FE*Year FE	Y	Y	Y	Y	Y	Y
Individual Characteristics		Y		Y		Y
Observations	22,801	22,801	22,801	22,801	22,801	22,801
R-squared	0.057	0.139	0.015	0.022	0.039	0.067

- **Entrepreneur Parents** increase the probability of being self-employed(6%).
- **Cadre Parents** *decrease* the probability of self-employment(1.1%).

Empirical Results: Multinomial Logit

Table 3B
Relative risk ratios in diff. Occupations by parental background –multinomial logit estimates.

	(1)	(2)	(3)
Reference group: being a firm employee			
Work in government			
Cadre Parents	2.327***	2.056***	2.043***
	(0.122)	(0.120)	(0.104)
Entrepreneur Parents	1.116	1.047	1.047
	(0.095)	(0.093)	(0.092)
Being a business owner			
Cadre Parents	1.656***	1.406***	1.413***
	(0.187)	(0.162)	(0.167)
Entrepreneur Parents	2.225***	1.912***	1.750***
	(0.379)	(0.315)	(0.300)
Being self-employed			
Cadre Parents	1.120*	1.058	1.080
	(0.075)	(0.070)	(0.073)
Entrepreneur Parents	1.921***	1.763***	1.579***
	(0.186)	(0.169)	(0.015)
Individual Characteristics		Y	Y
Province FE*Year FE			Y
Observations	22,801	22,801	22,801

Notes: In Table 3A, the comparison group is all other occupations. In Table 3B, the reference group is being a firm employee. Individual characteristics include: age, gender, marital status, ethnic minority status and college education. Standard errors are clustered at the province-year level. Significance level: $*p < 0.1$, $**p < 0.05$, $***p < 0.01$.

Work in government

- **Cadre Parents** increase the odds of being a government relative to be a firm employee by **over 2 times significantly**.
- **Entrepreneur Parents** increase the odds of being a government relative to be a firm employee but **the effect is not significant**.

Empirical Results: Multinomial Logit

Table 3B
Relative risk ratios in diff. Occupations by parental background –multinomial logit estimates.

	(1)	(2)	(3)
Reference group: being a firm employee			
Work in government			
Cadre Parents	2.327***	2.056***	2.043***
	(0.122)	(0.120)	(0.104)
Entrepreneur Parents	1.116	1.047	1.047
	(0.095)	(0.093)	(0.092)
Being a business owner			
Cadre Parents	1.656***	1.406***	1.413***
	(0.187)	(0.162)	(0.167)
Entrepreneur Parents	2.225***	1.912***	1.750***
	(0.379)	(0.315)	(0.300)
Being self-employed			
Cadre Parents	1.120*	1.058	1.080
	(0.075)	(0.070)	(0.073)
Entrepreneur Parents	1.921***	1.763***	1.579***
	(0.186)	(0.169)	(0.015)
Individual Characteristics		Y	Y
Province FE*Year FE			Y
Observations	22,801	22,801	22,801

Notes: In Table 3A, the comparison group is all other occupations. In Table 3B, the reference group is being a firm employee. Individual characteristics include: age, gender, marital status, ethnic minority status and college education. Standard errors are clustered at the province-year level. Significance level: $*p < 0.1$, $**p < 0.05$, $***p < 0.01$.

Being a business owner

- **Cadre Parents** increase the odds of being a business owner relative to be a firm employee by **over 1.4 times significantly**.
- **Entrepreneur Parents** increase the odds of being a business owner relative to be a firm employee by **over 1.7 times significantly**.

Empirical Results: Multinomial Logit

Table 3B
Relative risk ratios in diff. Occupations by parental background –multinomial logit estimates.

	(1)	(2)	(3)
Reference group: being a firm employee			
Work in government			
Cadre Parents	2.327***	2.056***	2.043***
	(0.122)	(0.120)	(0.104)
Entrepreneur Parents	1.116	1.047	1.047
	(0.095)	(0.093)	(0.092)
Being a business owner			
Cadre Parents	1.656***	1.406***	1.413***
	(0.187)	(0.162)	(0.167)
Entrepreneur Parents	2.225***	1.912***	1.750***
	(0.379)	(0.315)	(0.300)
Being self-employed			
Cadre Parents	1.120*	1.058	1.080
	(0.075)	(0.070)	(0.073)
Entrepreneur Parents	1.921***	1.763***	1.579***
	(0.186)	(0.169)	(0.015)
Individual Characteristics		Y	Y
Province FE*Year FE			Y
Observations	22,801	22,801	22,801

Notes: In Table 3A, the comparison group is all other occupations. In Table 3B, the reference group is being a firm employee. Individual characteristics include: age, gender, marital status, ethnic minority status and college education. Standard errors are clustered at the province-year level. Significance level: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Being self-employed

- **Cadre Parents** don't increase the odds of being self-employed relative to be a firm employee.
- **Entrepreneur Parents** increase the odds of self-employed relative to be a firm employee by **over 1.6 times significantly**.

Summary of LPM and Multinomial Logit

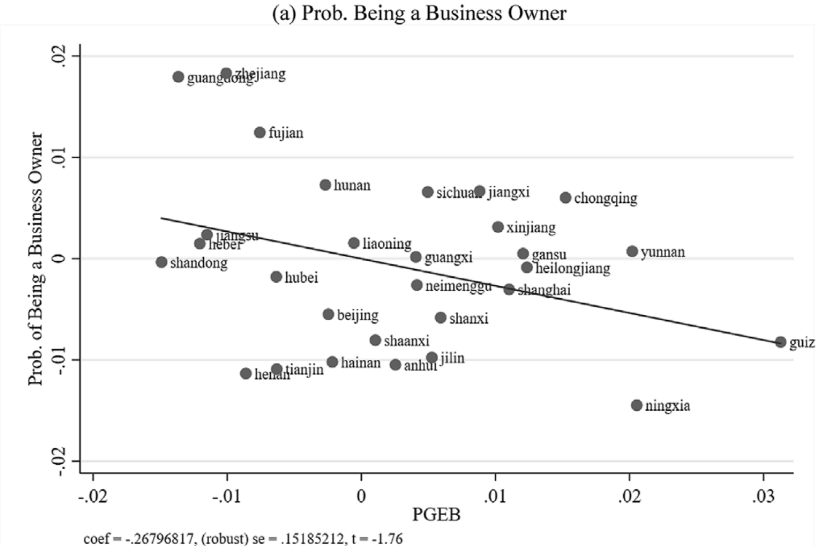
Parents	Model	Government	Business Owner	Self-employ
Cadre	LPM	$\uparrow$	$\uparrow$	$\downarrow$
Cadre	MLogit	$\uparrow$	$\uparrow$	—
Entrepreneur	LPM	—	$\uparrow$	$\uparrow$
Entrepreneur	MLogit	—	$\uparrow$	$\uparrow$

- The LPM and MLogit models provide very similar results.

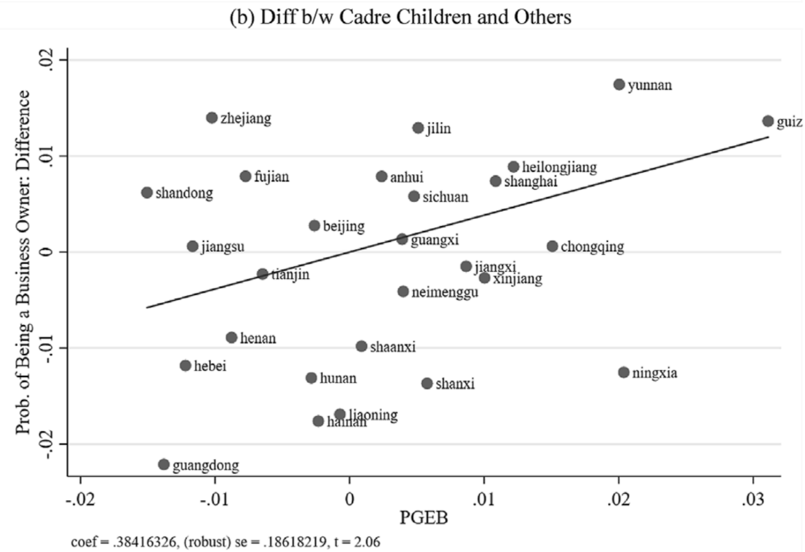
Parental Background and Local Economic Context

- **Measurement**: *Provincial Government Expenditure on Business-related activities(PGEB)* as a measure of the role of government on the private business environment.
  - Expenditure on Business-related activities: Infrastructure and MCF (Manufacturing/Commerce/Finance).
- Robustness:
  - weakly correlated with GDP
  - negatively correlated marketization index.
  - relatively smaller share of private sector.

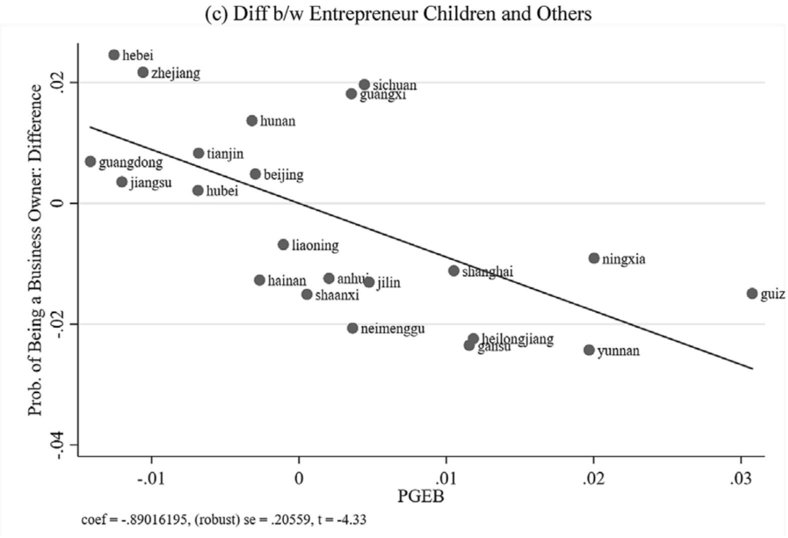
Descriptive patterns: cross-provinces



Descriptive patterns: cross-provinces



Descriptive patterns: cross-provinces



Parental Background and Local Economic Context

- **Question**: Whether the association between parental occupation and business ownership varies with the level of government intervention in the business environment?.
- **Linear Probability Model: Interacted with PGEB**

$$
\begin{aligned}
Pr(Y=1|X) = \beta_1 CardreParent_i &+ \beta_2 CardreParent_i \times PGEB_{pt} \\
+\beta_3 EntreParents_i &+ \beta_4 EntreParents_i \times PGEB_{pt} \\
+\gamma X_i + \gamma X_i \times PGEB_{pt} &+ Prov_p \times Year_t + u_{ipt}
\end{aligned}
$$

- **Question**: Which parameter is our interest? and how to interpret it?
- **Answer**: β_2 and β_4 are the coefficients of the interaction terms between parental occupation and PGEB.
- **Thinking 1**: *Why there is no PGEB term in the model?*

Empirical Results: LPM+Interactions

Table 4
The impact of cadre Parent × PGEB in determining business ownership.

	(1)	(2)	(3)	(4)	(5)	(6)
	Y = business owner (mean = 0.022)					
Cadre Parent * PGEB (sd)	0.004*	0.004*	0.005**			0.007**
	(0.002)	(0.002)	(0.002)			(0.003)
Cadre Parent	0.006**	0.003	0.003	0.003	0.003	0.003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Entrepreneur Parent * PGEB (sd)	−0.008*	−0.008**	−0.008*			−0.006
	(0.004)	(0.004)	(0.004)			(0.008)
Entrepreneur Parent	0.016***	0.014**	0.014**	0.014**	0.013**	0.014**
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Cadre Parent * GDP Per Capita (sd)				−0.001		−0.001
				(0.002)		(0.002)
Entre. Parent * GDP Per Capita (sd)				−0.006		−0.006
				(0.005)		(0.004)
Cadre Parent * Other Expend (sd)					0.003	−0.002
					(0.003)	(0.004)
Entrepreneur Parent * Other Expend (sd)					−0.007	−0.003
					(0.005)	(0.010)
Province FE*Year FE	Y	Y	Y	Y	Y	Y
Individual Characteristics		Y	Y	Y	Y	Y
PGEB *Individual Characteristics			Y	Y	Y	Y
Observations	22,801	22,801	22,801	22,801	22,801	22,801
R-squared	0.015	0.023	0.023	0.023	0.023	0.023

Empirical Results: LPM+Interactions

Table 4
The impact of cadre Parent × PGEB in determining business ownership.

	(1)	(2)	(3)	(4)	(5)	(6)
	Y = business owner (mean = 0.022)					
Cadre Parent * PGEB (sd)	0.004*	0.004*	0.005**			0.007**
	(0.002)	(0.002)	(0.002)			(0.003)
Cadre Parent	0.006**	0.003	0.003	0.003	0.003	0.003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Entrepreneur Parent * PGEB (sd)	−0.008*	−0.008**	−0.008*			−0.006
	(0.004)	(0.004)	(0.004)			(0.008)
Entrepreneur Parent	0.016***	0.014**	0.014**	0.014**	0.013**	0.014**
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Cadre Parent * GDP Per Capita (sd)				−0.001		−0.001
				(0.002)		(0.002)
Entre. Parent * GDP Per Capita (sd)				−0.006		−0.006
				(0.005)		(0.004)
Cadre Parent * Other Expend (sd)					0.003	−0.002
					(0.003)	(0.004)
Entrepreneur Parent * Other Expend (sd)					−0.007	−0.003
					(0.005)	(0.010)
Province FE*Year FE	Y	Y	Y	Y	Y	Y
Individual Characteristics		Y	Y	Y	Y	Y
PGEB *Individual Characteristics			Y	Y	Y	Y
Observations	22,801	22,801	22,801	22,801	22,801	22,801
R-squared	0.015	0.023	0.023	0.023	0.023	0.023

Empirical Results: LPM+Interactions

Table 4
The impact of cadre Parent × PGEB in determining business ownership.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Y = business owner (mean = 0.022)						Y = self-employed (mean = 0.107)
Cadre Parent * PGEB (sd)	0.004*	0.004*	0.005**			0.007**	0.002
	(0.002)	(0.002)	(0.002)			(0.003)	(0.007)
Cadre Parent	0.006**	0.003	0.003	0.003	0.003	0.003	−0.011**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)
Entrepreneur Parent * PGEB (sd)	−0.008*	−0.008**	−0.008*			−0.006	0.017
	(0.004)	(0.004)	(0.004)			(0.008)	(0.011)
Entrepreneur Parent	0.016***	0.014**	0.014**	0.014**	0.013**	0.014**	0.057***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.012)
Cadre Parent * GDP Per Capita (sd)				−0.001		−0.001	
				(0.002)		(0.002)	
Entre. Parent * GDP Per Capita (sd)				−0.006		−0.006	
				(0.005)		(0.004)	
Cadre Parent * Other Expend (sd)					0.003	−0.002	
					(0.003)	(0.004)	
Entrepreneur Parent * Other Expend (sd)					−0.007	−0.003	
					(0.005)	(0.010)	
Province FE*Year FE	Y	Y	Y	Y	Y	Y	Y
Individual Characteristics		Y	Y	Y	Y	Y	Y
PGEB *Individual Characteristics			Y	Y	Y	Y	Y
Observations	22,801	22,801	22,801	22,801	22,801	22,801	22,801
R-squared	0.015	0.023	0.023	0.023	0.023	0.023	0.068

Notes: This table shows that the advantage in becoming a business owner (1) increases with PGEB for those with cadre parents and (2) decreases with PGEB for those with entrepreneur parents. Individual characteristics include: age, gender, marital status, ethnic minority status, and college education. Standard errors are clustered at the province-year level. Significance level: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

Jia,Lan and Miquel(2021): Main Findings

1. **Is there intergenerational transmission of entrepreneurship in China?**
   - Yes, and the magnitude is similar to findings elsewhere.
2. **Do children of government officials have a higher likelihood of becoming entrepreneurs?**
   - Yes, in particular they have a high likelihood of owning incorporated businesses.
3. **Do parental determinants depend on the role of government?**
   - Yes. the larger is government involvement in business-related spending, the larger the business-ownership propensity of children of government officials, and the smaller the propensity of children of entrepreneurs.

Wrap Up

Summary

- The key assumptions of these models are similar to those of OLS regression.
  - If it suffers OVB or other potential endogenous bias, then the coefficient estimates are biased and inconsistent even we use the MLE to estimate the parameters rather than OLS.
- Although Probit and Logit offer some advantages in model specifications over LPM, LPM is more intuitive and easier to interpret.
  - This is particularly useful when we want to deal with the endogeneity problem.
- When the dependent variable is binary,even multinomial, the LPM remains a good starting point for empirical analysis.

Appendix 1

Appendix 1: MLE in Simple Linear Regression

- Suppse the simple linear regression model is

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

- Now we have two estimation approaches:
  - OLS
  - MLE
- Recall the Simple OLS estimator is

$$\hat{\beta}_{1,OLS} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- How to get the MLE estimator of β_0 and β_1?

MLE Estimation of Simple Linear Regression

- We maintain the same three assumptions as in OLS:
  - **MLE Assumption 1**: X_{1i} is **exogenous**, thus $E(u_i|X_{1i}) = 0$
  - **MLE Assumption 2**: u_i is **independently** distributed.
  - **MLE Assumption 3**: Large outliers are unlikely.
- Additionally, MLE requires two more assumptions:
  - **MLE Assumption 4**: u_i is **normally** distributed, thus

$$u_i \sim N(0, \sigma^2)$$

  - **MLE Assumption 5**: u_i is **homoskedastic**, thus

$$Var(u_i|X_{1i}) = \sigma^2$$

MLE Estimation of Simple Linear Regression

- **Step 1: Write down the likelihood function**

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(Y_i | X_i, \beta_0, \beta_1, \sigma^2)$$

- **where**

$$f(Y_i | X_i, \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right)$$

- **Step 2: Maximize the log likelihood function**

$$ln(L(\beta_0, \beta_1, \sigma^2)) = \sum_{i=1}^{n} ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right)\right)$$

MLE Estimation of Simple Linear Regression

- **First order conditions(FOC):**
  - **For** β_0: $\frac{\partial \ln L}{\partial \beta_0} = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0$
  - **For** β_1: $\frac{\partial \ln L}{\partial \beta_1} = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i = 0$
  - **For** σ^2: $\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 = 0$
- **MLE Solutions:**
  - $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
  - $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

MLS vs OLS

- For slope and intercept:

$$\hat{\beta}_0^{MLE} = \hat{\beta}_0^{OLS}$$
$$\hat{\beta}_1^{MLE} = \hat{\beta}_1^{OLS}$$

- Therefore, the MLE estimator is **identical to the OLS estimator**.
- However, for variance:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$
$$\hat{\sigma}^2_{OLS} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

- When n is small, $\hat{\sigma}^2_{MLE} < \hat{\sigma}^2_{OLS}$ (MLE **underestimates** the variance)
- When n is large, $\hat{\sigma}^2_{MLE}$ and $\hat{\sigma}^2_{OLS}$ are very close to each other
- Note: While we focus more on the properties of coefficient estimates (β), the

MLS vs MLE

- Under the assumption of **normality** and **homoskedasticity**, OLS and MLE give identical point estimates for β_0 and β_1.
- OLS estimator is **BLUE (Best Linear Unbiased Estimator)**
- MLE provides theoretical justification for OLS under normality
- When error distribution is non-normal, MLE may differ from OLS
- This relationship extends to multiple regression: $Y = X\beta + u$
- If these assumptions fail, more specialized estimation methods may be needed.

Appendix 2: Newton-Raphson Method

Math Review: Taylor Expressions

- Recall **Taylor series** of a function $f(x)$ at a certain value of x,thus x_0,

Math Review: Taylor Expressions

- Recall **Taylor series** of a function $f(x)$ at a certain value of x,thus x_0,

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0)$$

Math Review: Taylor Expressions

- Recall **Taylor series** of a function $f(x)$ at a certain value of x,thus x_0,

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2$$

Math Review: Taylor Expressions

- Recall **Taylor series** of a function $f(x)$ at a certain value of x,thus x_0,

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

Math Review: Taylor Expressions

- Recall **Taylor series** of a function $f(x)$ at a certain value of x,thus x_0,

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

- Then we can have the **Taylor expression** of $f(x)$ at first and second orders

Math Review: Taylor Expressions

- Recall **Taylor series** of a function $f(x)$ at a certain value of x,thus x_0,

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \dots \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

- Then we can have the **Taylor expression** of $f(x)$ at first and second orders

$$f(x) \simeq f(x_0) + f'(x_0)(x - x_0)$$

Math Review: Taylor Expressions

- Recall **Taylor series** of a function $f(x)$ at a certain value of x,thus x_0,

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \dots \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

- Then we can have the **Taylor expression** of $f(x)$ at first and second orders

$$f(x) \simeq f(x_0) + f'(x_0)(x - x_0)$$

$$f(x) \simeq f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2$$

Newton-Raphson Method

- **Objective**: find the solution of x to a equation: $f(x) = 0$
- An alternative way: find some x make

$$f(x_0) + f'(x_0)(x - x_0) = 0$$

- Here the x_0 is **some initial value** that we guess, which is close to the desired solution. And then we obtain a **better** approximation x_1, based on

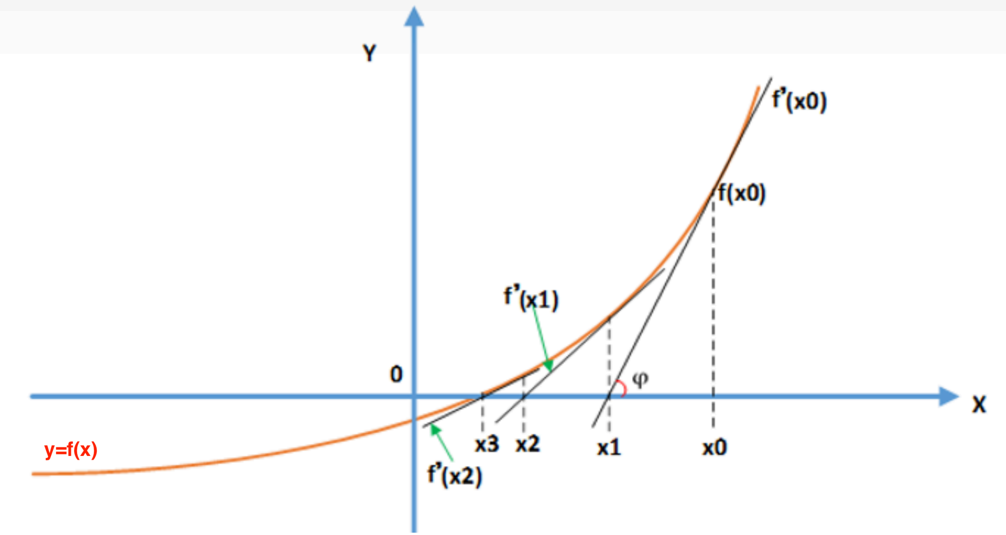
$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

- We do not stop repeating this procedure until

$$f(x_j) = 0$$

where the x_j is the solution to the function.

Newton-Raphson Method



Newton-Raphson Method

- **Objective**: find the solution of x to the **F.O.C** equation: $f'(x) = 0$
- Then we need the Taylor expression of $f(x)$ at second order

$$f(x) \simeq f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2$$

- F.O.C for $f'(x) = 0$

Newton-Raphson Method

- **Objective**: find the solution of x to the **F.O.C** equation: $f'(x) = 0$
- Then we need the Taylor expression of $f(x)$ at second order

$$f(x) \simeq f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2$$

- F.O.C for $f'(x) = 0$

$$\frac{d}{d(x - x_0)}\left[f\left(x_0\right) + f'\left(x_0\right)\left(x - x_0\right) + \frac{1}{2}f''\left(x_0\right)\left(x - x_0\right)^2\right] = 0$$

Newton-Raphson Method

- **Objective**: find the solution of x to the **F.O.C** equation: $f'(x) = 0$
- Then we need the Taylor expression of $f(x)$ at second order

$$f(x) \simeq f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2$$

- F.O.C for $f'(x) = 0$

$$\frac{d}{d(x - x_0)}\left[f\left(x_0\right) + f'\left(x_0\right)\left(x - x_0\right) + \frac{1}{2}f''\left(x_0\right)\left(x - x_0\right)^2\right] = 0$$

$$\Rightarrow f'\left(x_0\right) + f''\left(x_0\right)\left(x - x_0\right) = 0$$

Newton-Raphson Method

- **Objective**: find the solution of x to the **F.O.C** equation: $f'(x) = 0$
- Then we need the Taylor expression of $f(x)$ at second order

$$f(x) \simeq f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2$$

- F.O.C for $f'(x) = 0$

$$\frac{d}{d(x-x_0)}\left[f\left(x_0\right) + f'\left(x_0\right)\left(x - x_0\right) + \frac{1}{2}f''\left(x_0\right)\left(x - x_0\right)^2\right] = 0$$

$$\Rightarrow f'\left(x_0\right) + f''\left(x_0\right)\left(x - x_0\right) = 0$$

$$\Rightarrow x = x_0 - \frac{f'\left(x_0\right)}{f''\left(x_0\right)}$$

Newton-Raphson Method

- **Objective**: find the solution of x to the **F.O.C** equation: $f'(x) = 0$
- Then we need the Taylor expression of $f(x)$ at second order

$$f(x) \simeq f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2$$

- F.O.C for $f'(x) = 0$

$$\frac{d}{d(x - x_0)}\left[f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2\right] = 0$$

$$\Rightarrow f'(x_0) + f''(x_0)(x - x_0) = 0$$

$$\Rightarrow x = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

- Repeating this procedure until $f'(x_j) = 0$ where the x_j is the solution to the function.

Computation of MLE estimators

- For simplicity, assume only one parameter θ, the maximum likelihood function is $L(\theta_{MLE})$.
- Then the F.O.C for the problem of maximization is as following

$$\frac{\partial L(\theta_{MLE})}{\partial \theta} = 0$$

- A initial guess of the parameter value, which denotes as θ_0. Then the MLE estimator,$\hat{\theta}_1$ can be calculated by

$$\hat{\theta}_1 \simeq \theta_0 - \left[\frac{\partial^2 L(\theta_0)}{\partial \theta^2}\right]^{-1} \frac{\partial L(\theta_0)}{\partial \theta}$$

- We do not stop repeating this procedure until

$$\frac{\partial L(\hat{\theta}_{MLE,j})}{\partial \theta} = 0$$

,where the $\hat{\theta}_{MLE,j}$ is the solution to the function.