# Lecture 6: Limited Dependent Variable

*Introduction to Econometrics,Spring 2023*

Zhaopeng Qu

Business School,Nanjing University

April 05 2023

1. Review of the last lecture

2. The Linear Probability Model(LPM)

3. Nonlinear Probability Models

4. Logit Model

5. Estimation and Inference in Probit and Logit Model

6. Sample Selection Model

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

# 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 $\beta$** 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...
- Linear function is not a good prediction function. Need a certain function which parameters enter nonlinearly, such as **logistic** function.
- 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:
    - Binary outcomes: (LPM,logit and probit)
    - Multinomial outcomes: Multiple responses or choices without orders (multi-logit and multi-probit)
    - Ordered outcomes: Ordered Response Models(ordered probit and logit)
    - Count outcomes: The outcomes is a nonnegative integer or a count (possion model)
    - Duration data(spell lengths or transitions): Duration model or hazard model
- Limited Dependent Variable
    - Censored data: The information on the dependent variable of some observations is lost,but not data on the regressors.
    - Truncated data: Both dependent variable and independent variables of some observations are missing for some reasons.
    - Sample selection: The sample are not randomly selected but based in part on values taken by a dependent variable.
- Only **Binary outcomes models** are covered here.

# 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 $X_{1i}, ..., X_{ki}$

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

- 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** $\beta_j$

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

- $\beta_j$ 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. Not all mortgage applications are approved.
- Question: *What determines whether or not a mortgage 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.(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,then probability being denied will also increase 0.6*
  - 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 large enough.
- **Answer**: compare with the mean value of dependent variable. Here *deny rate* $= 0.12$ means that the deny ratio will increase $0.06/0.12 \times 100\% = 50\%$ if PI 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: 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: Heteroskedasticity

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

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

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

# LPM: Predicted values

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

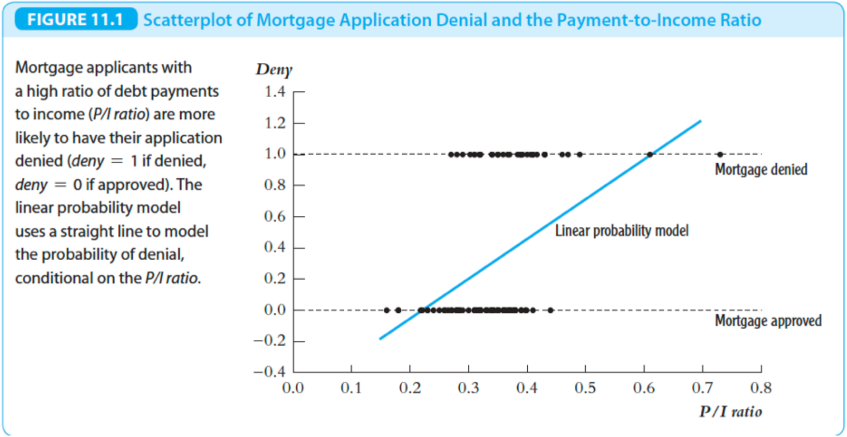


FIGURE 11.1 Scatterplot of Mortgage Application Denial and the Payment-to-Income Ratio

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 should not be less than 0 or greater than 1
- To address this problem,consider a *nonlinear* probability models

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

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

- And the function have to satisfy the conditions:
  - $0 \leq G(Z) \leq 1$
  - monotonicity and continuity
- The key is whether we could find a proper function $G(x)$ which can limit the prediction value less than 1 and greater than 0.
  - 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 $x$

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

- If $X$ is a discrete r.v. with with possible outcomes$\mathcal{X}$ and the probability mass function is $f_X(x)$, then the c.d.f is the

$$F_X(x) = \sum_{\substack{t \in \mathcal{X} \\ t \leq x}} f_X(t)$$

- If $X$ is a discrete and the probability mass function is $f_X(x)$, then the p.d.f is the

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \mathrm{d}t$$

- 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

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

- 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: Prediction Value

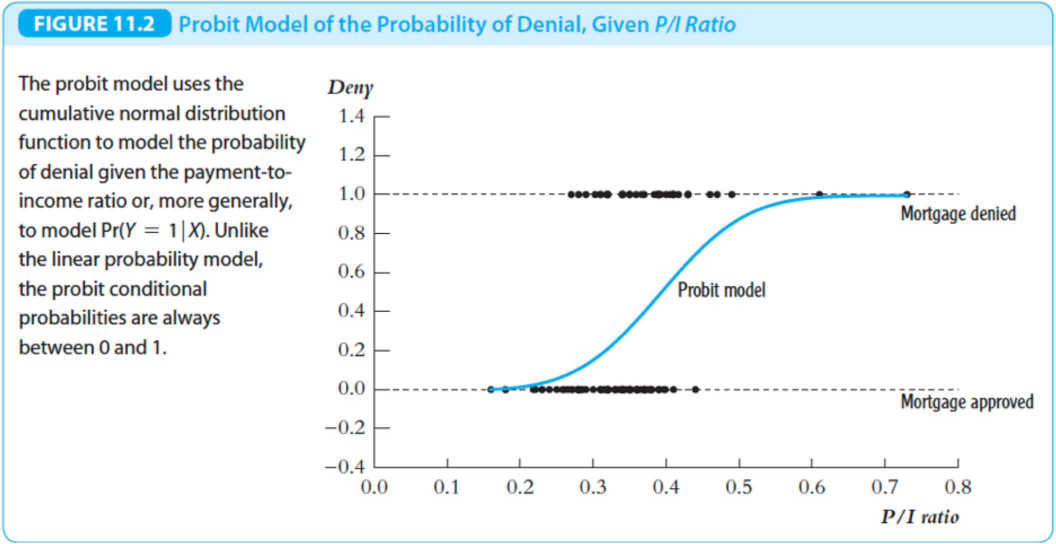


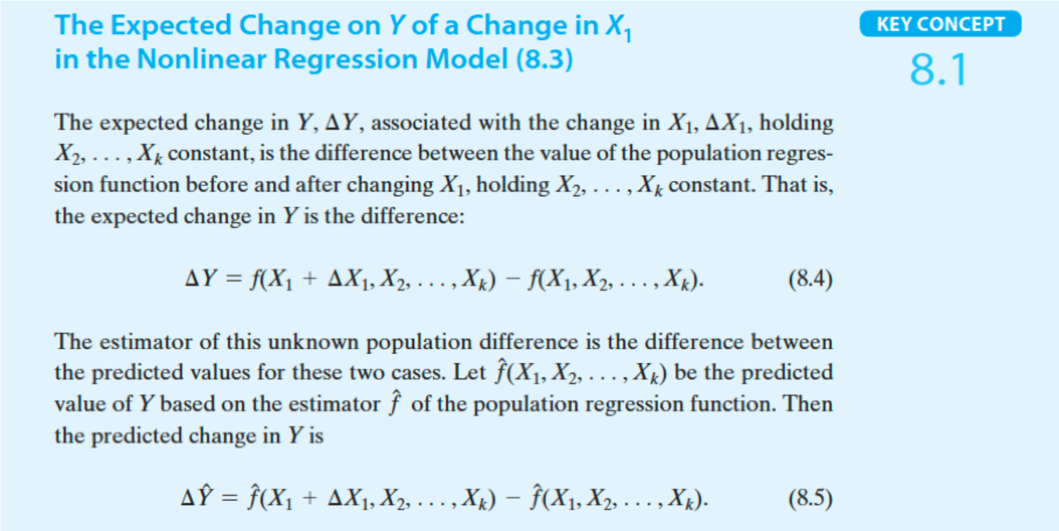
**FIGURE 11.2** Probit Model of the Probability of Denial, Given *P/I Ratio*

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 $\beta_j$ is **the change in the $Z$-value** rather than the probability arising from a unit change in $X_j$, holding constant other $X_i$.
- 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 $Y$, $\Delta Y$, associated with the change in $X_1$, $\Delta X_1$, 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

$$Z = -2 + 3X_1$$

- 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

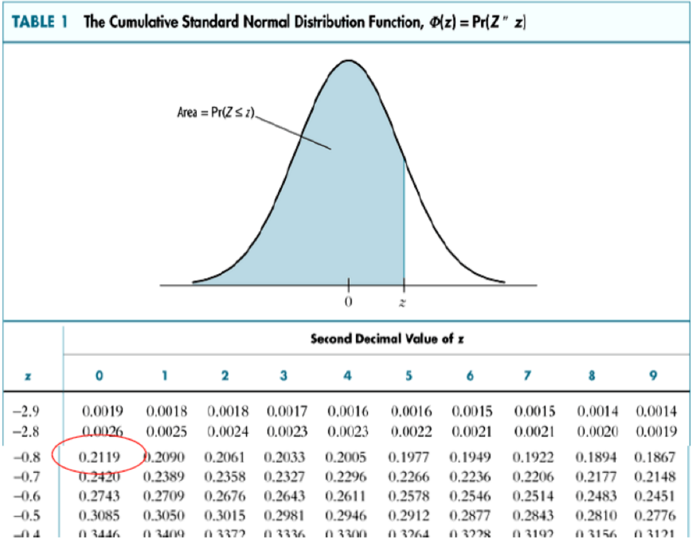


**TABLE 1** **The Cumulative Standard Normal Distribution Function, $\Phi(z)$ = Pr(Z ″ z)**

Area = Pr(Z ≤ z)
0
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:

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

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

$$\begin{array}{rl} Pr\left(deny \widehat{= 1|P}/I\ ratio\right) = \Phi\Big(& -2.19 \ + \ 2.97 P/I\ ratio\Big) \\ & (0.16) \qquad (0.47) \end{array}$$

# Example: Mortgage Applications

- *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 $P/I\ ratio = 0.4$

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

- 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 continous

- **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)

# 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 $3.36\%(0.0336)$

# Discrete Explanatory Variable

- 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

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

$$(0.16) \qquad (0.44) \qquad (0.083)$$

- 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

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

- so 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

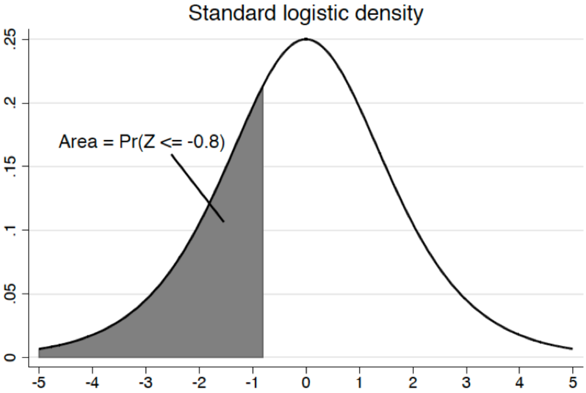
$$Z = -2 + 3 \times 0.4 = -0.8$$

- So the probability

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

# Logit Model: Predicted Probabilities

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



Standard logistic density

# 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 $\beta_j$ 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 $X_i$.
- Different from the Probit model
  - The **odds ratio**

# Logit Model: the Odds Ratio

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

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

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

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

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

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

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

# Logit Model: the Odds Ratio

- Then

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

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

# Example: Mortgage Applications

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

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

- 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:

  $P/I\ ratio_{mean} = 0.331$

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

- 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

$$\frac{1}{1+e^{-(-4.13+5.37\times 0.3+1.27\times 0)}} = 0.074$$

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

$$\frac{1}{1+e^{-(-4.13+5.37\times 0.3+1.27\times 1)}} = 0.222$$

- 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 $P/Iratio = 0.3$.

# Estimation and Inference in Probit and Logit Model

# Estimation and Inference in Probit and Logit Model

- How to estimate $\beta_0, \beta_1, ..., \beta_k$?
- What is the sampling distribution of the estimators?
- Logit and Probit models are nonlinear in the coefficients $\beta_0, \beta_1, ..., \beta_k$
    - These models can NOT be estimated directly by OLS, but by Nonlinear Least Squares(NLS).
    - In practice,the most common method used to estimate logit and probit models is **Maximum Likelihood Estimation** (MLE).

# Review: Maximum Likelihood Estimation

- The **likelihood function** is a *joint probability distribution* of the data, treated as a function of the unknown coefficients.
- The **maximum likelihood estimator** (MLE) are the estimate values of the 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 $Y_1, Y_2, Y_3, ...Y_n$ have a joint density function denoted

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

  - where $\theta$ is an unknown parameter.

- Given observed values $Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n$,the likelihood of $\theta$ 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 $\theta$.

- Then the **Maximum Likelihood Estimation** to $\theta$ 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.
- 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

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

# Maximum Likelihood Estimation of a Binary Variable

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

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

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

# Maximum Likelihood Estimation

**MLE Step 2**: *Write down the maximization problem*

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

$$ln(f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n)) = \left(\sum y_i\right) ln(p) + \left(n - \sum y_i\right) ln(1 - p)$$

- 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

**MLE Step 3**: *Maximize the likelihood function*

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

$$
\begin{aligned}
\frac{d}{dp} ln(f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n)) &= 0 \\
\frac{d}{dp}\left[\left(\sum y_i\right) ln(p) + \left(n - \sum y_i\right) ln(1-p)\right] &= 0 \\
\frac{\sum y_i}{p} - \frac{n - \sum y_i}{1-p} &= 0 \\
p(n - \sum y_i) &= \sum y_i(1-p)
\end{aligned}
$$

- Solving the equation for $p$ yields the MLE estimator; that is, $\hat{p}_{MLE}$ satisfies

$$
\hat{p}_{MLE} = \frac{1}{n}\sum y_i = \overline{Y}
$$

- You can prove that $\hat{p}_{MLE}$ is an unbiased, consistent and normally distributed estimator of $p$.

# 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

$$\begin{aligned} f_{probit}(\beta_0, ..., \beta_k; Y_1, ..., Y_n | X_{1i}, ..., X_{ki}, i = 1, ..., n) = & \\ Pr(Y_1 = y_1, .., Y_n = y_n) &= Pr(Y_1 = y_1) \times ... \times Pr(Y_n = y_n) \\ &= p^{y_1}(1-p)^{1-y_1} \times ... \times p^{y_n}(1-p)^{1-y_n} \\ = \left[\Phi(\beta_0 + \beta_1 X_{11} + ... + \beta_k X_{k1})^{y_1}(1 - \Phi(\beta_0 + \beta_1 X_{11} + ... + \beta_k X_{k1}))^{1-y_1}\right] \times & \\ ... \times \left[\Phi(\beta_0 + \beta_1 X_{1n} + ... + \beta_k X_{kn})^{y_n}(1 - \Phi(\beta_0 + \beta_1 X_{1n} + ... + \beta_k X_{kn}))^{1-y_n}\right] & \end{aligned}$$

# MLE of the Probit Model

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

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

- Then the maximization problem is

$$\underset{\hat{\beta}_0, \hat{\beta}_1, .., \hat{\beta}_k}{arg\max}\, ln(f_{probit}(\beta_0, \beta_1, ..., \beta_k; Y_1 = y_1, ..., Y_n = y_n|X_{1i}, ..., X_{ki}, i = 1, ..., n))$$

# 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

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

- 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 from solutions necessarily.
- We can still obtain the values of estimators using **numerical algorithm** with iterative methods.
- One of common methods is **Gradient Method** based on low order *Taylor series expansions*.

# 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

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

# Newton-Raphson Method

- Our 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 $x_0$ we guess, which is close to the desired solution. And then we obtain a **better** approximation $x_1$, based on

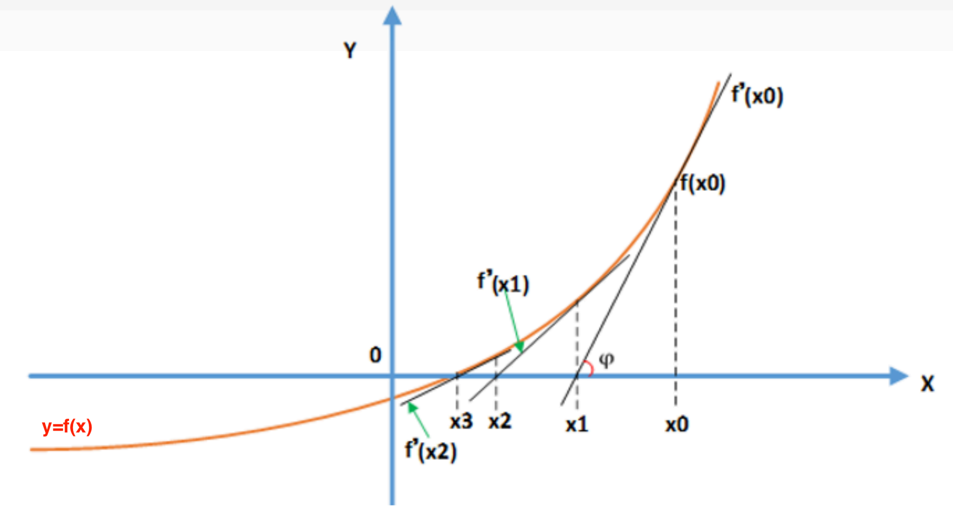
$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

- We do not stop repeating this procedure until

$$f(x_j) = 0$$

, here the $x_j$ is the solution to the function.

# Newton-Raphson Method



# Newton-Raphson Method

- Our objective: find the solution of $x$ to a 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 $f'(x) = 0$

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

- repeating this procedure until

$$f'(x_j) = 0$$

, here the $x_j$ is the solution to the function.

# Computation of MLE estimators

- For simplicity, assume only one parameter $\theta$, 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 $\theta_0$. Then the MLE estimator can be calculated by

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

- We do not stop repeating this procedure until

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

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

# 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.
- 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** and **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**, it is based on comparing the log likelihood values of the unrestricted and the restricted model. The test statistic is

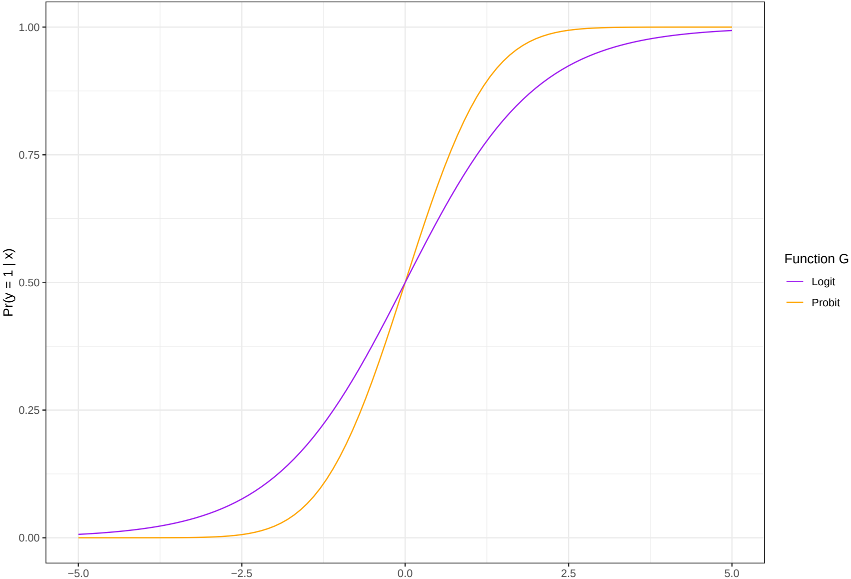
$$LR = 2\left(logL_{ur} - LogL_r\right) \sim \chi^2_q$$

- where q is the number of restrictions being tested.

# 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

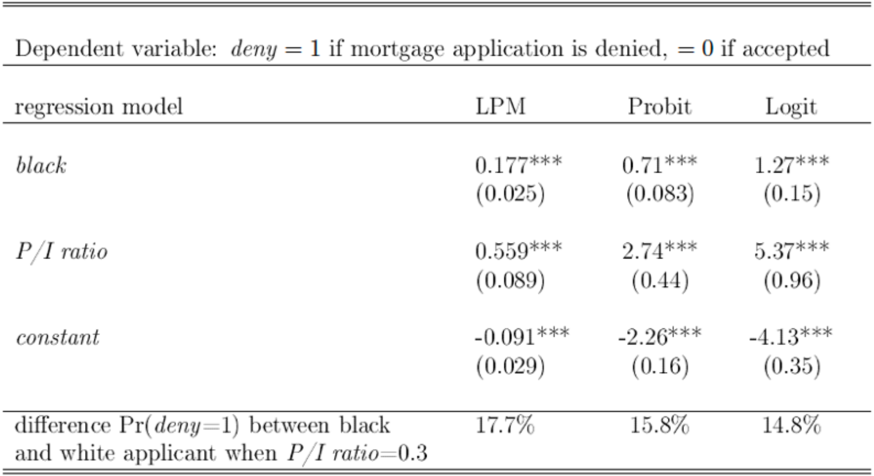


# Comparing the LPM,Probit and Logit

- 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***<br>(0.025) | 0.71***<br>(0.083) | 1.27***<br>(0.15) |
| *P/I ratio* | 0.559***<br>(0.089) | 2.74***<br>(0.44) | 5.37***<br>(0.96) |
| *constant* | -0.091***<br>(0.029) | -2.26***<br>(0.16) | -4.13***<br>(0.35) |
| difference Pr(*deny*=1) between black and white applicant when *P/I ratio*=0.3 | 17.7% | 15.8% | 14.8% |

# Sample Selection Model

# More Extensions: Limited Dependent Variables families

- Multinomial outcomes: No order, such as (multinomial-logit,probit)
- Ordered outcomes: Ordered Response Models(order probit and logit)
- Count outcomes: The outcomes is a nonnegative integer or a count. (possion model)
- Limited Dependent Variable(Censored, Tobit and Selection Models)
- Time: (Duration Model)

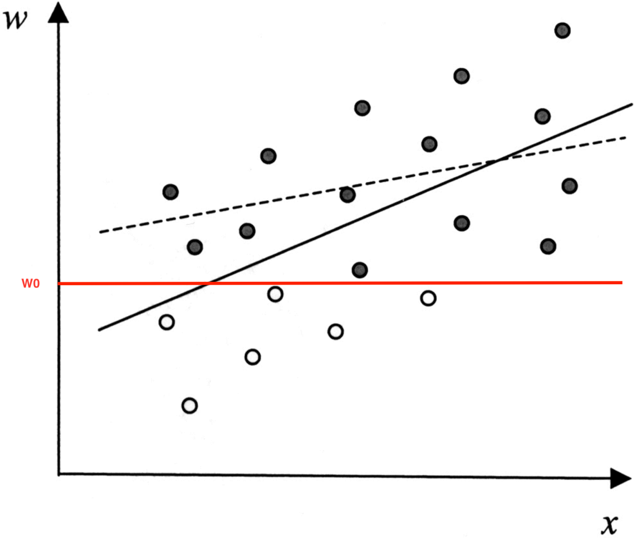
# Introduction to Sample Selection Model

- The classical example: wage determination of working women

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

- $Y_i$ is logwage
- $X_i$ is schooling years
- The sample selection problem arises in that the sample consists only of women who choose to work.
    - If the selection into working and not working for women is random,then OK.
    - But in reality, working women probably smarter, more career-oriented, more ambitious which can not observed or measured in the data.

# Wage determination of working women



# Intro to Heckman Sample Selection Model

- A **two-equation behavioral model**

1. *selection equation*

$$Z_i^* = W_i'\gamma + e_i$$

where $Z_i$ is a latent variable which indicates the propensity of working for a married woman

- and the error term $e_i$ satisfies

$$E[e_i|W_i] = 0$$

- Then $Z_i$ is a dummy variable to represent whether a woman to work or not,thus

$$Z_i = \begin{cases} 1 \; if \; Z^* > 0 \\ 0 \; if \; Z^* \leq 0 \end{cases}$$

# Heckman Sample Selection Model

2. *outcome equation*

$$Y_i^* = X_i'\beta + u_i$$

- where the outcome($Y_i$) can be observed only when $Z_i$=1 or $Z_i^* > 0$

$$Y_i^* = \begin{cases} Y_i \; if \; Z_i = 1 \\ 0 \; or \; missing \; if \; Z_i = 0 \end{cases}$$

- The error term $u_i$ satisfies $E\left[u_i \middle| X_i\right] = 0$

# Heckman Sample Selection Model

- The conditional expectation of wages on $X_i$ is

$$E[Y_i^*|X_i] = X_i'\beta$$

- The conditional expectation of wages on $X_i$ is *only for women who work*$(Z^* > 0)$

$$\begin{aligned} E[Y_i^*|X_i, Z_i^* > 0] &= E[Y_i|X_i, Z_i^* > 0] \\ &= E[X_i'\beta + u_i|X_i, Z_i^* > 0] \\ &= X_i'\beta + E[u_i|Z_i^* > 0] \\ &= X_i'\beta + E[u_i|e_i > -W_i'\gamma] \end{aligned}$$

# Heckman Sample Selection Model

- If $u_i$ and $e_i$ is independent, then $E[u_i|e_i > -W_i'\gamma] = 0$, then

$$E[Y_i^*|X_i, Z_i^* > 0] = E[Y_i^*|X_i] = X_i'\beta$$

,which means that using sample-selected data does not matter to the estimation of $\beta$

- But in reality, unobservables in the two equations, thus $u_i$ and $e_i$, are likely to be **correlated**
  - eg. innate ability
- Instead assume that $u_i$ and $e_i$ are **jointly normal distributed**, which means that

$$\begin{pmatrix} u_i \\ e_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{eu} \\ \sigma_{ue} & \sigma_e^2 \end{pmatrix}\right)$$

# Math Review: Two Normal Distributed R.V.s

## Two Normal Distributed R.V.s

For any two normal variables $(n_0, n_1)$ with zero mean, we can write $n_1 = \alpha_0 n_0 + \eta$, where $\eta \sim N(0, \sigma_\eta)$ and $E(\eta | n_0) = 0$.Then we have

$$\alpha_0 = \frac{Cov(n_0, n_1)}{Var(n_0)}$$

or

$$E(n_1 \mid n_0) = \frac{Cov(n_0, n_1)}{Var(n_0)} n_0$$

Then

$$n_1 = E(n_1 \mid n_0) + \eta = \frac{Cov(n_0, n_1)}{Var(n_0)} n_0 + \eta$$

# Heckman Sample Selection Model

- For two normal variables $u_i$ and $e_i$ with zero mean, we have

$$\alpha_0 = \frac{Cov(u_i, e_i)}{Var(e_i)} = \frac{\sigma_{ue}}{\sigma_e^2}$$

- Then

$$u_i = \alpha_0 e_i + \eta = \frac{\sigma_{ue}}{\sigma_e^2} e_i + \eta$$

where $\eta \sim N\left(0, \sigma_\eta\right)$ and $E\left(\eta | e_i\right) = 0$

# Heckman Sample Selection Model

- Then the conditional expectation of $u_i$

$$\begin{aligned} E[u_i|e_i > -W_i'\gamma] &= E[\frac{\sigma_{ue}}{\sigma_e^2}e_i + \eta|e_i > -W_i'\gamma] \\ &= \frac{\sigma_{ue}}{\sigma_e^2}E[e_i|e_i > -W_i'\gamma] + E[\eta|e_i > -W_i'\gamma] \\ &= \frac{\sigma_{ue}}{\sigma_e^2}E[e_i|e_i > -W_i'\gamma] \end{aligned}$$

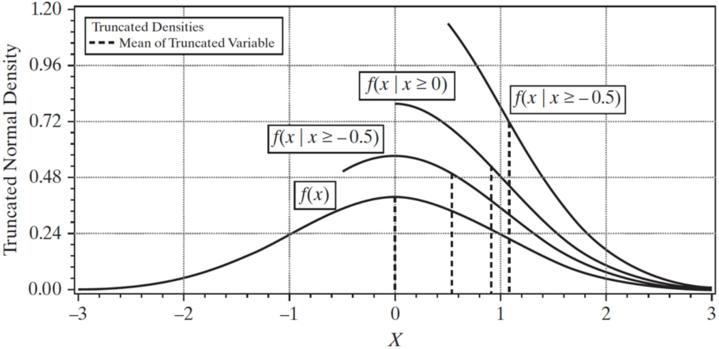
# Math Review: Truncated Density Function

## Truncated Density Function

If a continuous random variable $X$ has p.d.f. $f(x)$ and c.d.f. $F(x)$ and $a$ is a constant, then the conditional density function

$$f(x|x > a) = \begin{cases} \frac{f(x)}{1-F(a)} & \mathit{if}\ x > a \\ 0 & \mathit{if}\ x \leq a \end{cases}$$

# Math Review: Truncated Density Function



- It amounts merely to **scaling** the density so that it integrates to one over the range above a.

# Math Review: Truncated Density Function

## Truncated Density Function

The proof follows from the definition of a conditional probability is

$$Pr(A|B) = \frac{Pr(AB)}{Pr(B)}$$

then,

$$\begin{aligned} F(x|X > c) &= \frac{Pr(X < x, X > c)}{Pr(X > c)} = \frac{Pr(c < X < x)}{1 - F(c)} \\ &= \frac{F(x) - F(c)}{1 - F(c)} \end{aligned}$$

then,

$$f(x|x > c) = \frac{d}{dx} F(x|X > c) = \frac{\frac{d}{dx}[F(x)] - 0}{1 - F(c)} = \frac{f(x)}{1 - F(c)}$$

# Standard Normal Truncated Density Function

- If X is distributed as standard normal, thus $X \sim N\left(0,1\right)$, then the p.d.f and c.d.f are as follow

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

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

- And $c$ is a scalar, then we can get the Truncated Density Function of an R.V. distributed in Standard Normal

$$f\left(x \mid x > c\right) = \frac{\phi\left(x\right)}{1 - \Phi\left(c\right)}$$

- The Expectation of in a standard normal truncated p.d.f

$$E(x|x > c) = \frac{f(c)}{1 - \Phi\left(c\right)} \equiv \lambda\left(c\right)$$

where $\lambda\left(c\right)$ is called by Inverse Mills Ratio.

## Proof

$$
\begin{aligned}
E(x|x>c) &= \int_c^{+\infty} xf(x|x>c)dx = \int_c^{+\infty} x\frac{\phi\left(x\right)}{1-\Phi\left(c\right)}dx \\
&= \frac{1}{1-\Phi\left(c\right)}\int_c^{+\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\
&= \frac{1}{1-\Phi\left(c\right)}\int_c^{+\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}d(\frac{x^2}{2}) \\
&= \frac{1}{1-\Phi\left(c\right)}\int_{\frac{c^2}{2}}^{+\infty} \frac{1}{\sqrt{2\pi}}e^{-t}d(t) \\
&= \frac{1}{1-\Phi\left(c\right)}\times\frac{1}{\sqrt{2\pi}} - e^{-t}\ |_{\frac{c^2}{2}}^{+\infty} \\
&= \frac{1}{1-\Phi\left(c\right)}\times\frac{1}{\sqrt{2\pi}}e^{-\frac{c^2}{2}} = \frac{f(c)}{1-\Phi\left(c\right)}
\end{aligned}
$$

# Heckman Sample Selection Model

- Then the conditional expectation of $u_i$

$$\begin{aligned}
E[u_i|e_i > -W_i'\gamma] &= \frac{\sigma_{ue}}{\sigma_e^2} E[e_i|e_i > -W_i'\gamma] \\
&= \frac{\sigma_{ue}}{\sigma_e} E[\frac{e_i}{\sigma_e}|\frac{e_i}{\sigma_e} > \frac{-W_i'\gamma}{\sigma_e}] \\
&= \frac{\sigma_{ue}}{\sigma_e} \frac{\phi(-W_i'\gamma/\sigma_e)}{1-\Phi(-W_i'\gamma/\sigma_e)} \\
&= \frac{\sigma_{ue}}{\sigma_e} \frac{\phi(W_i'\gamma/\sigma_e)}{\Phi(W_i'\gamma/\sigma_e)} \\
&= \sigma_\lambda \lambda(W_i'\gamma)
\end{aligned}$$

# Heckman Sample Selection Model

- Then the conditional expectation of wages on $X_i$ is only for women who work($Z^* > 0$)

$$E[Y_i^*|X_i, Z_i^* > 0] = E[Y_i|X_i, Z_i = 1] = X_i'\beta + \sigma_\lambda \lambda(W_i'\gamma)$$

- It means that if we could include $\lambda(W_i'\gamma)$ as **an additional regressor** into the outcome equation, thus we run

$$Y_i = X_i'\beta + \sigma_\lambda \lambda(W_i'\gamma) + u_i$$

  then we can obtain the **unbiased** and **consistent** estimate $\beta$ using a self-selected sample.
- The coefficient before $\lambda(\cdot)$ can be testing significance to indicate whether the term should be included in the regression, in other words, *whether the selection should be corrected.*

# Heckit Model Estimation

1. Estimate selection equation using **all observations**,thus

$$Z_i = W_i'\gamma + e_I$$

- obtain estimates of parameters $\hat{\gamma}$
- computer the Inverse Mills Ratio(IMR) $\dfrac{\phi(W_i'\gamma)}{\Phi(W_i'\gamma)} = \hat{\lambda}(W_i'\gamma)$

2. Estimate the outcome equation using **only the selected observations**.

$$Y_i = X_i'\beta + \sigma_\lambda\hat{\lambda}(W_i'\gamma) + u_i$$

- **Note**: standard error is not right, have to be adjusted because we use $\hat{\lambda}(W_i'\gamma)$ instead of $\lambda(W_i'\gamma)$ in the estimation.

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

# Jia,Lan and Miquel(2021)

- 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.
- Interaction:
  - Provincial Government Expenditure on Business-related activities(PGEB) as a measure of the role of government on the private business environment.

# 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:

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

  - $Y_i$ 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

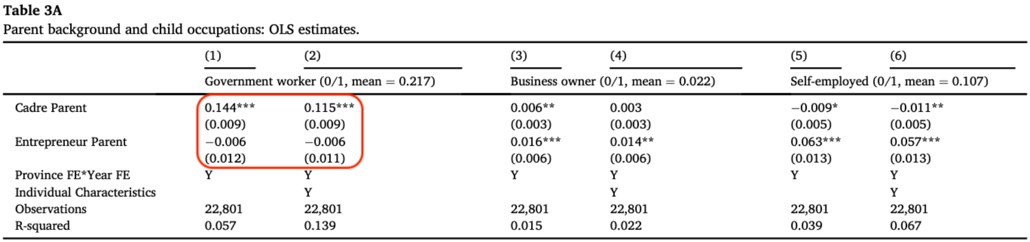


**Table 3A**
Parent background and child occupations: OLS estimates.

| | (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

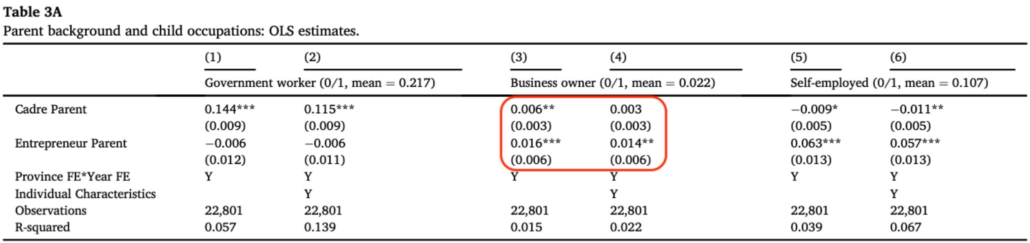


**Table 3A**
Parent background and child occupations: OLS estimates.

| | (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

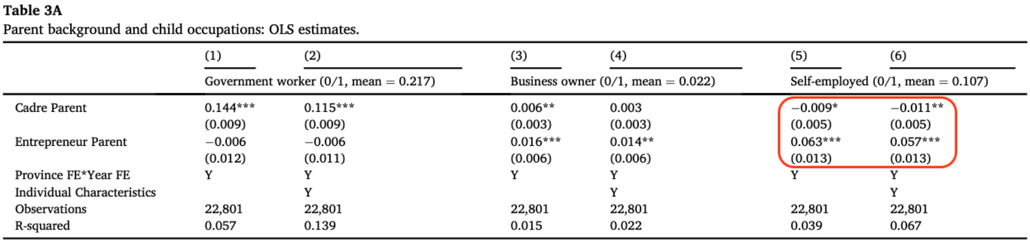


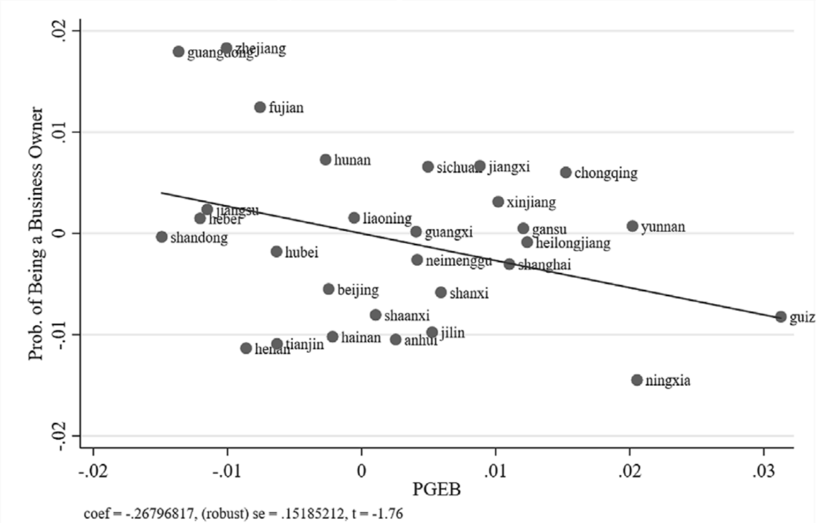
**Table 3A**
Parent background and child occupations: OLS estimates.

| | (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 business owner(6%).
- *Cadre Parents* **decrease** the probability of self-employment(1.1%).

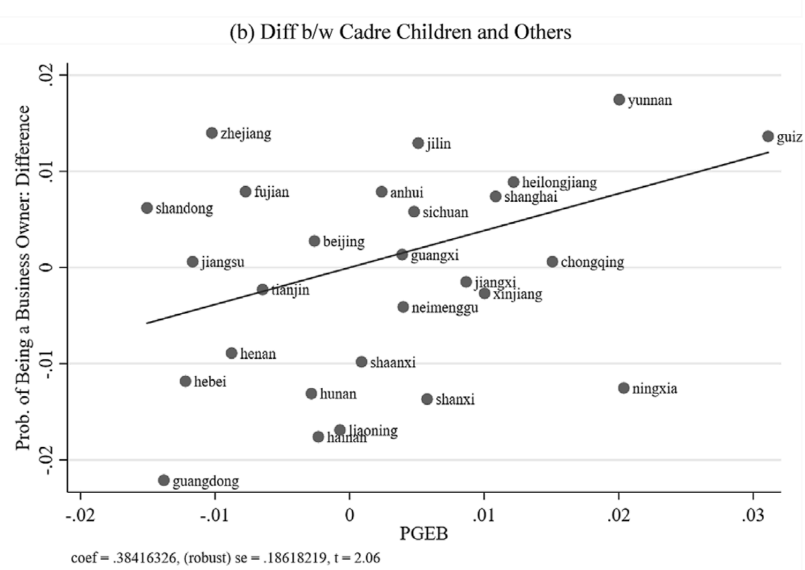
# Descriptive patterns: Cross-provinces

(a) Prob. Being a Business Owner



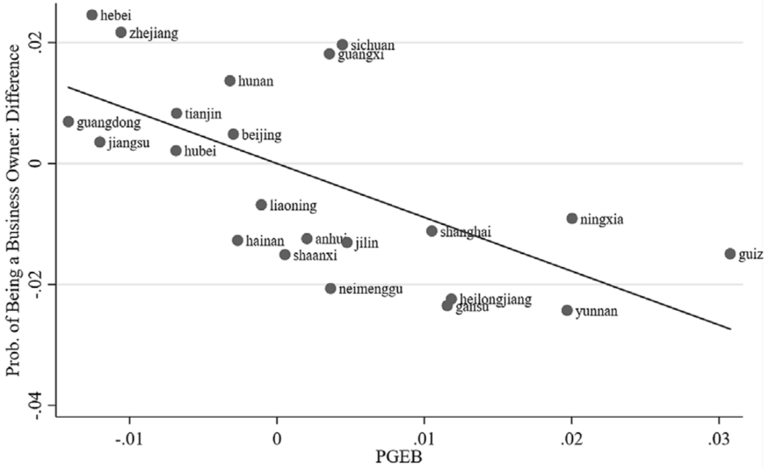
coef = -.26796817, (robust) se = .15185212, t = -1.76

# Descriptive patterns



# Descriptive patterns

(c) Diff b/w Entrepreneur Children and Others



coef = -.89016195, (robust) se = .20559, t = -4.33

# 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**

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

# Empirical Results: LPM+Interactions

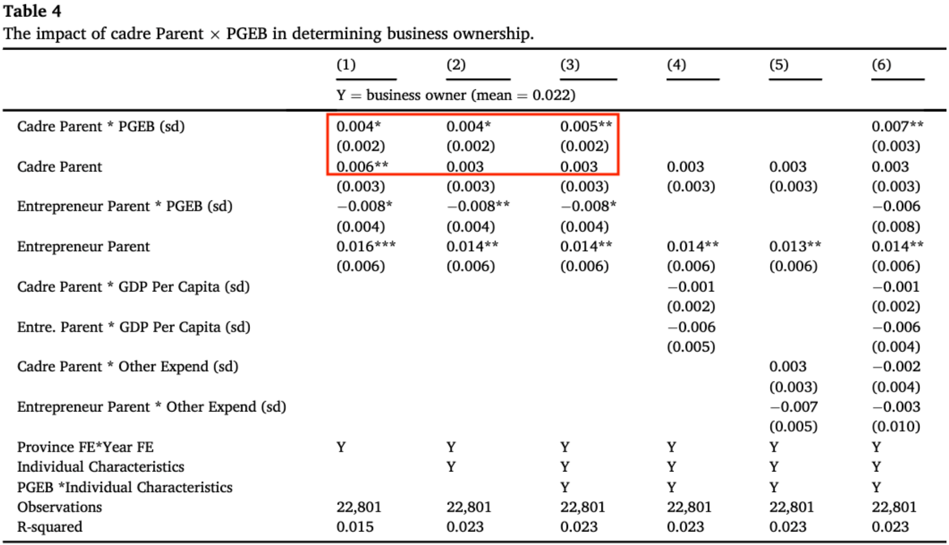


**Table 4**

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

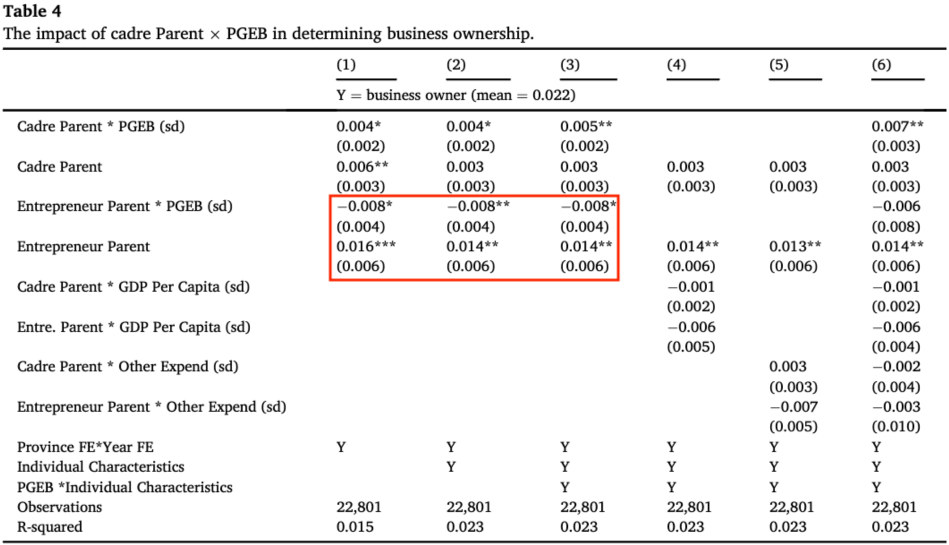


**Table 4**
The impact of cadre Parent × PGEB in determining business ownership.

| | (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

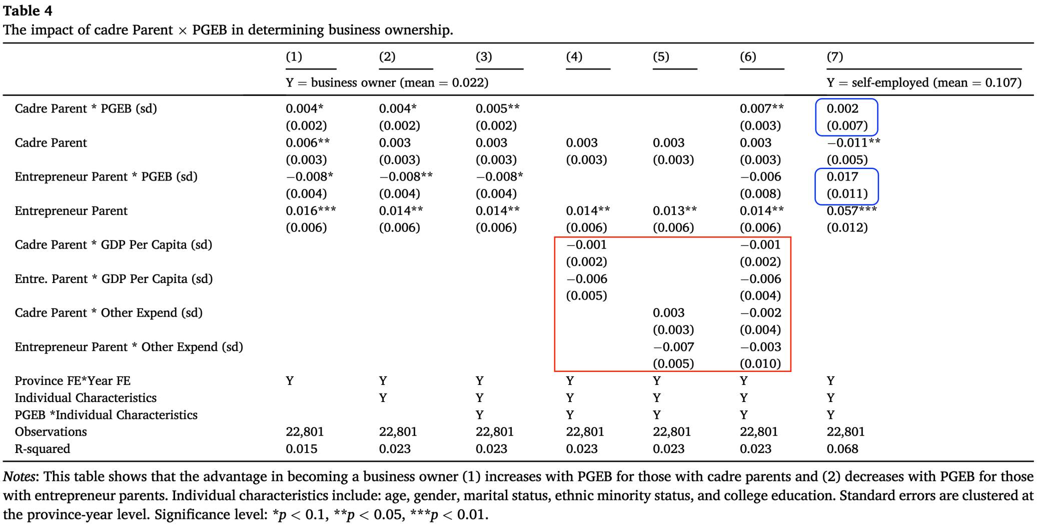


**Table 4**

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?
   - 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.