大数据时代的管理决策 (2024 年春)

Lecture 3: OLS Regression(II):Nonlinear Model

Zhaopeng Qu

Nanjing University Business School

March 15 2024

1. Review of previous lecture
2. Nonlinear Regression Functions:
3. Nonlinear in Xs
4. Polynomials in X
5. Logarithms
6. Interactions Between Independent Variables
7. A Lastest and Smart Application: Jia and Ku(2019)
8. Summary

Review of previous lecture

OLS Regression and Hypothsis Testing

- **Hypothesis Testing** is a formal procedure in statistics for testing assumptions regarding a population parameter.
- **Hypothesis Testing** in OLS regressions
  - single coefficient: the **t-statistic**
  - two or more coefficents: **the F-statistic**
- The key component in obtaining the t-statistic is the **standard error**(S.E.), which is the estimation of Standard Deviation of estimated coefficients($\hat{\beta}$).

OLS Regression and Hypothsis Testing

- Assumption 4: The error terms are **homoskedastic**

$$Var(u_i \mid X_i) = \sigma_u^2$$

  Then $\hat{\beta}^{OLS}$ is the **Best Linear Unbiased Estimator(BLUE)**.

- However,in most cases it is **NOT** a plausible assumption.
- Homoskedasticity is a special case of heteroskedasticity, these heteroskedasticity robust formulas are also valid if the error terms are homoskedastic.
- Using the hypothesis testing and confidence interval in OLS regression, we could make a more reliable judgment about the relationship between the treatment and the outcomes.

Nonlinear Regression Functions:

Introduction

- Recall the assumption of Linear Regression Model

Linear Regression Model

The observations, (Y_i, X_i) come from a random sample(i.i.d) and satisfy the linear regression equation,

$$Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i} + u_i$$

- Everything what we have learned so far is under this assumption of linearity. But this linear approximation is not always a good one.

Introduction: Recall the whole picture

- A general formula for a population regression model may be

$$Y_i = f(X_{1,i}, X_{2,i}, ..., X_{k,i}) + u_i$$

- **Parametric methods**: assume that the function form(families) is known, we just need to assure(estimate) some unknown parameters in the function.
  - **Linear**
  - **Nonlinear**
- **Nonparametric methods**: assume that the function form is unknown or unnecessary to known.

Nonlinear Regression Functions

- How to extend linear OLS model to be nonlinear?

1. **Nonlinear in Xs**(the lecture now)
   - **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*.
2. **Nonlinear in β** or **Nonlinear in Y**(the next lecture)
   - **Discrete Dependent Variables** or **Limited Dependent Variables**.
   - Linear function in Xs is not a good prediciton function or Y.
   - Need a function which parameters enter nonlinearly, such as logisitic or negative exponential functions.
   - Then the parameters can not obtained by OLS estimation any more but *Nonlinear Least Squres* or Maximum Likelyhood Estimation.

Marginal Effect of X in Nonlinear Regression

- If our regression model is linear: $Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i} + u_i$
  - Then the **marginal effect** of X, thus *the effect of Y on a change in X_j by 1 (unit)* is **constant** and equals β_j:

$$\beta_j = \frac{\partial Y_i}{\partial X_{ji}}$$

- But if a relation between Y and X is **nonlinear**, thus

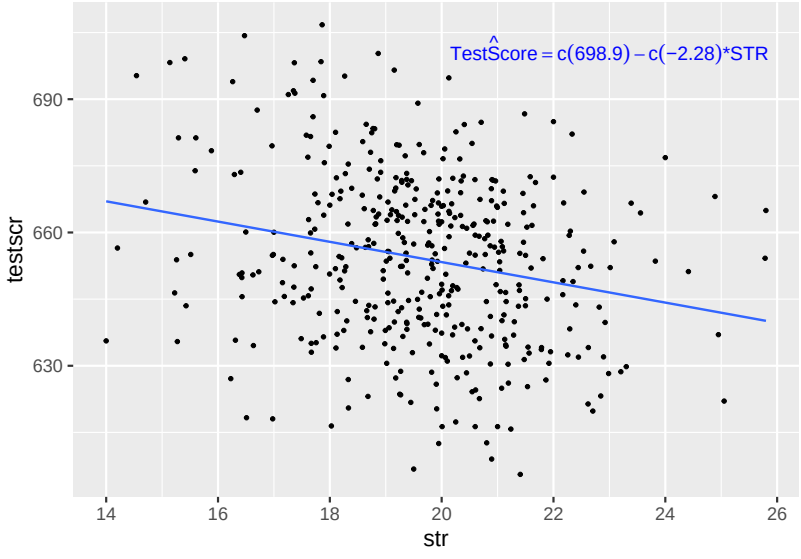
$Y_i = f(X_{1,i}, X_{2,i}, ..., X_{k,i}) + u_i$

  - Then the marginal effect of X is not constant, but depends on the value of Xs(including X_i itself or/and other X_js) because

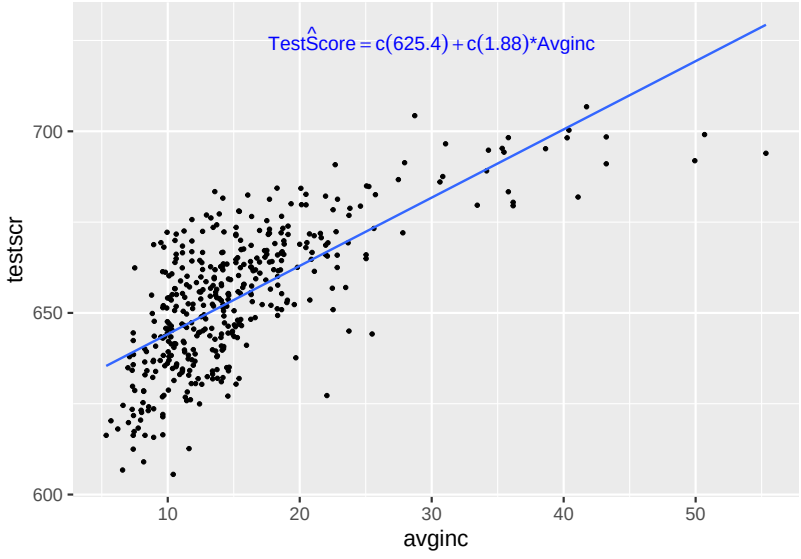
$$\frac{\partial Y_i}{\partial X_{ji}} = \frac{\partial f(X_{1,i}, X_{2,i}, ..., X_{k,i})}{\partial X_{ji}}$$

Nonlinear in Xs

The TestScore – STR relation looks linear (maybe)



But the TestScore – Income relation looks nonlinear

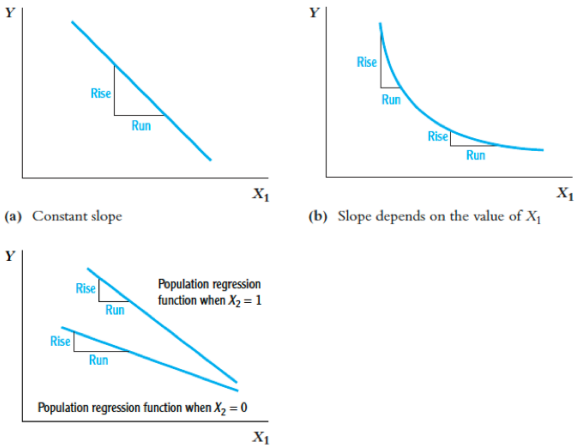


Three Complementary Approaches:

1. **Polynomials in X**
   - The population regression function is approximated by a quadratic, cubic, or higher-degree polynomial.
2. **Logarithmic transformations**
   - Y and/or X is transformed by taking its logarithm
   - this gives a *percentages* interpretation that makes sense in many applications
3. **Interactions**
   - the effect X on Y depends on the value of another independent variable
   - very often used in the analysis of hetergenous effects, some time used as analysis(channel).

Population Regression Functions with Different Slopes

FIGURE 8.1 Population Regression Functions with Different Slopes



(a) Constant slope

(b) Slope depends on the value of X_1

(c) Slope depends on the value of X_2

In Figure 8.1a, the population regression function has a constant slope. In Figure 8.1b, the slope of the population regression function depends on the value of X_1. In Figure 8.1c, the slope of the population regression function

The Effect of a Change in X in Nonlinear Functions

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:

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

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

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

Polynomials in X

Example: the TestScore-Income relation

- If a straight line is NOT an adequate description of the relationship between district income and test scores, what is?
- Two options
  - Quadratic specification:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + u_i$$

  - Cubic specification:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + \beta_3 (Income_i)^3 + u_i$$

- How to estimate these models?
  - We can see **quadratic** and **cubic** terms as two independent variables.
  - Then the model turns into a **special form** of a multiple OLS regression model.

Estimation of the quadratic specification in R

```
#>
#> Call:
#>    felm(formula = testscr ~ avginc + I(avginc^2), data = ca)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -44.416  -9.048   0.440   8.348  31.639
#>
#> Coefficients:
#>              Estimate Robust s.e t value Pr(>|t|)
#> (Intercept) 607.30174    2.90175 209.288   <2e-16 ***
#> avginc        3.85100    0.26809  14.364   <2e-16 ***
#> I(avginc^2)  -0.04231    0.00478  -8.851   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 12.72 on 417 degrees of freedom
#> Multiple R-squared(full model): 0.5562   Adjusted R-squared: 0.554
#> Multiple R-squared(proj model): 0.5562   Adjusted R-squared: 0.554
#> F-statistic(full model, *iid*):261.3 on 2 and 417 DF, p-value: < 2.2e-16
```

Estimation of the cubic specification in R

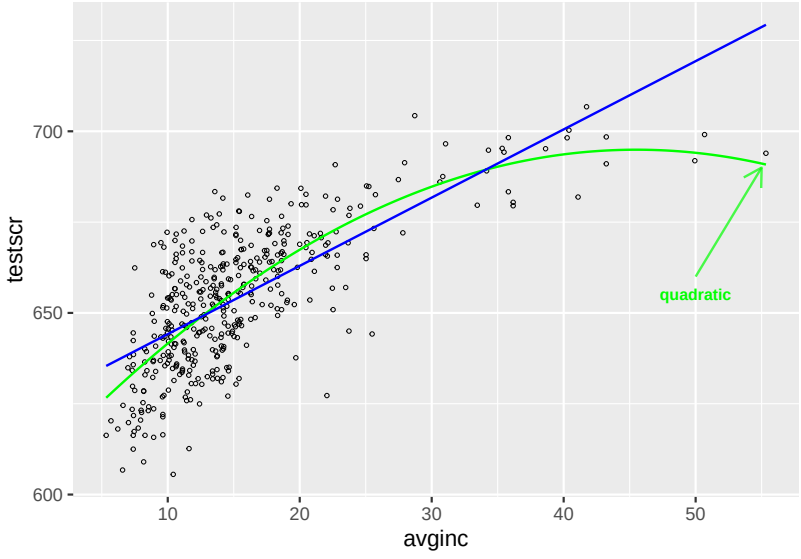
```
#>
#> Call:
#>    felm(formula = testscr ~ avginc + I(avginc^2) + I(avginc3), data = ca)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -44.28  -9.21   0.20   8.32  31.16
#>
#> Coefficients:
#>               Estimate Robust s.e t value Pr(>|t|)
#> (Intercept)  6.001e+02  5.102e+00 117.615  < 2e-16 ***
#> avginc       5.019e+00  7.074e-01   7.095 5.61e-12 ***
#> I(avginc^2) -9.581e-02  2.895e-02  -3.309  0.00102 **
#> I(avginc3)   6.855e-04  3.471e-04   1.975  0.04892 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 12.71 on 416 degrees of freedom
#> Multiple R-squared(full model): 0.5584   Adjusted R-squared: 0.5552
#> Multiple R-squared(proj model): 0.5584   Adjusted R-squared: 0.5552
```

TestScore and Income: OLS Regression Results

Table 1

	Dependent Variable: Test Score		
	(1)	(2)	(3)
avginc	1.879***	3.851***	5.019***
	(0.113)	(0.267)	(0.704)
I(avginc^2)		−0.042***	−0.096***
		(0.005)	(0.029)
I(avginc^3)			0.001**
			(0.0003)
Constant	625.384***	607.302***	600.079***
	(1.863)	(2.891)	(5.078)
Observations	420	420	420
Adjusted R^2	0.506	0.554	0.555
Residual Std. Error	13.387	12.724	12.707
F Statistic	430.830***	261.278***	175.352***

Figure: Linear and Quadratic Regression



Quadratic vs Linear

- **Question**: Is the quadratic model better than the linear model?
- We can test the null hypothesis that the regression function is linear against the alternative hypothesis that it is quadratic:

$$H_0 : \beta_2 = 0 \; and \; H_1 : \beta_2 \neq 0$$

- the t-statistic

$$t = \frac{(\hat{\beta}_2 - 0)}{SE(\hat{\beta}_2)} = \frac{-0.0423}{0.0048} = -8.81$$

- Since $8.81 > 2.58$, we reject the null hypothesis (the linear model) at a 1% significance level.
- Based on the F-test, we can also reject the null hypothesis

$$F - statistic_{q=2,d=417} = 261.3, p - value \cong 0.00$$

Interpreting the estimated quadratic regression

- What is the **marginal effect** of X on Y in a quadratic regression function.
- The regression model now is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

- The marginal effect of X on Y

$$\frac{\partial Y_i}{\partial X_i} = \beta_1 + 2\beta_2 X_i$$

- It means that the **marginal effect** of X on Y depends on the specific value of X_i

Interpreting the estimated quadratic regression

- The estimated regression function with a quadratic term of income is

$$\widehat{TestScore}_i = \underset{(2.90)}{607.3} + \underset{(0.27)}{3.85} \times income_i - \underset{(0.0048)}{0.0423} \times income_i^2.$$

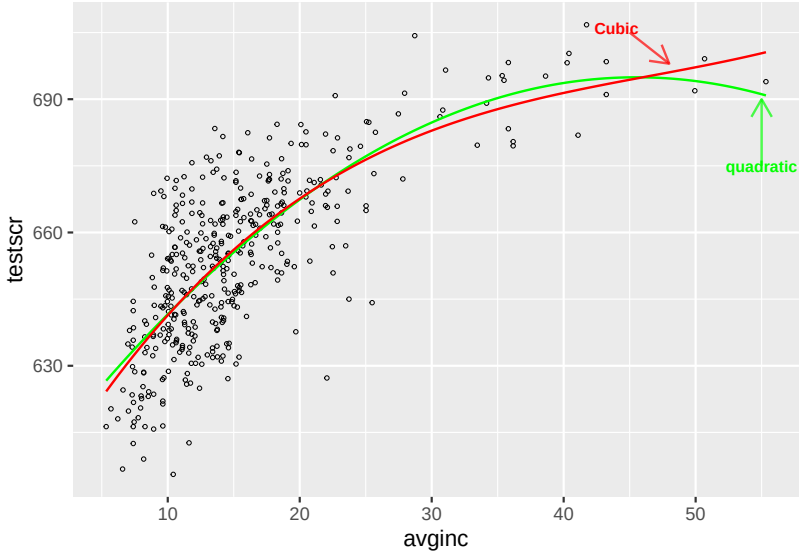
- Suppose *the effect of an $1000 increase on average income on test scores*
- A group: from $10,000 per capita to $11,000 per capita:

$$\begin{aligned}\Delta TestScore &= 607.3 + 3.85 \times 11 - 0.0423 \times (11)^2 \\ &\quad - [607.3 + 3.85 \times 10 - 0.0423 \times (10)^2] \\ &= 2.96\end{aligned}$$

- B group: from $40,000 per capita to $41,000 per capita:

$$\begin{aligned}\Delta TestScore &= 607.3 + 3.85 \times 41 - 0.0423 \times (41)^2 \\ &\quad - [607.3 + 3.85 \times 40 - 0.0423 \times (40)^2] \\ &= 0.42\end{aligned}$$

Figure: Cubic and Quadratic Regression



Quadratic vs Cubic

- **Question**: Is the cubic model better than the quadratic model?
- **Answer**: testing the null hypothesis that the regression function is *quadratic* against the alternative hypothesis that it is *cubic*:

$$H_0 : \beta_3 = 0 \; and \; H_1 : \beta_3 \neq 0$$

- the t-statistic

$$t = \frac{(\hat{\beta}_3 - 0)}{SE(\hat{\beta}_3)} = \frac{-0.001}{0.0003} = -3.33$$

- Since $3.33 > 2.58$, we reject the null hypothesis (the linear model) at a 1% significance level.
- the F-test also reject

$$F - statistic_{q=3,d=416} = 175.35, p - value \cong 0.00$$

Interpreting the estimated cubic regression function

- **The regression model now is**

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$$

- **The marginal effect of X on Y**

$$\frac{\partial Y_i}{\partial X_i} = \beta_1 + 2\beta_2 X_i + 3\beta_3 X_i^2$$

Interpreting the estimated regression function

- **The estimated cubic model is**

$$\widehat{TestScore}_i = \underset{(5.83)}{600.1} + \underset{(0.86)}{5.02} \times income - \underset{(0.03)}{0.96} \times income^2 - \underset{(0.00047)}{0.00069} \times income^3.$$

- **A group**: **from $10,000 per capita to $11,000 per capita:**

$$\begin{aligned}\Delta TestScore &= 600.079 + 5.019 \times 11 - 0.96 \times (11)^2 + 0.001 \times (11)^3 \\ &\quad - [600.079 + 5.019 \times 10 - 0.96 \times (10)^2 + 0.001 \times (10)^3]\end{aligned}$$

- **B group**: **from $40,000 per capita to $41,000 per capita:**

$$\begin{aligned}\Delta TestScore &= 600.079 + 5.019 \times 41 - 0.96 \times (41)^2 + 0.001 \times (41)^3 \\ &\quad - [600.079 + 5.019 \times 40 - 0.96 \times (40)^2 + 0.001 \times (40)^3]\end{aligned}$$

Polynomials in X Regression Function

- Approximate the population regression function by a polynomial:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2 ... + \beta_r X_i^r + u_i$$

- This is just the multiple linear regression model – except that the regressors are **powers of X!**
- Estimation, hypothesis testing, etc. proceeds as in the multiple regression model using OLS.
- Although, the coefficients are difficult to interpret, the regression function itself is interpretable.

Testing the population regression function is linear

- If the population regression function is linear, then the higher-degree terms should not enter the population regression function.
- To perform hypothesis test

$$H_0 : \beta_2 = 0, \beta_3 = 0, ..., \beta_r = 0 \; and \; H_1 : at \; least \; one \; \beta_j \neq 0$$

- Because H_0 is a **joint null hypothesis** with $q = r - 1$ restrictions on the coefficients, it can be tested using the F-statistic.

Which degree polynomial should I use?

- How many powers of X should be included in a polynomial regression?
- The answer balances a **trade-off** between flexibility and statistical precision. (many ML or non-parametric or semi-parametric methods work on this)
  - Increasing the degree r introduces more flexibility into the regression function and allows it to match more shapes; a polynomial of degree r can have up to r - 1 bends (that is, inflection points) in its graph.
  - But increasing r means adding more regressors, which can reduce the precision of the estimated coefficients.

Which degree polynomial should I use?

- A practical way: asking whether the coefficients in the regression associated with the largest values of r are **zero**. If so, then these terms can be dropped from the regression.
- This procedure, which is called *sequential hypothesis testing*
  1. Pick a maximum value of r and estimate the polynomial regression for that r.
  2. Use the t-statistic to test whether the coefficient on X^r,β_r is **ZERO**.
  3. If reject, then the degree is r; if not then test the whether the coefficient on X^{r-1},β_{r-1} is **ZERO**.
  4. continue this procedure until the coefficient on the highest power in your polynomial is statistically significant.

Which degree polynomial should I use?

- The **initial degree** r of the polynomial is still missing.
- In many applications involving economic data, the nonlinear functions are smooth, that is, they do not have sharp jumps, or "spikes."
- If so, then it is appropriate to choose a small maximum degree for the polynomial, such as 2, 3, or 4.

Which degree polynomial should I use?

- There are also several formal testing to determine the degree.
  - The F-statistic approach
  - The Akaike Information Criterion(AIC)
  - The Bayes Information Criterion(BIC)
- We will introduce them later on.

Wrap Up

- The nonlinear functions in Polynomials in Xs are just a special form of Multiple OLS Regression.
- If the true relationship between X and Y is nonlinear in polynomials in Xs, then a fully linear regression is misspecified – the functional form is wrong.
- The estimator of the effect on Y of X is biased(a special case of OVB).
- Estimation, hypothesis testing, etc. proceeds as in the multiple regression model using OLS, which can also help us to tell the degrees of polynomial functions.
- The big difference is how to explained the estimate coefficients and make the predicted change in Y with a change in Xs.

Logarithms

Logarithmic functions of Y and/or X

- Another way to specify a nonlinear regression model is to use the natural logarithm of Y and/or X.
- $Ln(x)$ = the natural logarithm of x is the inverse function of the exponential function e^x, here $e = 2.71828$.

$$x = ln(e^x)$$

Review of the Basic Logarithmic functions

- If X and a are variables, then we have

$$
\begin{aligned}
ln(1/x) &= -ln(x) \\
ln(ax) &= ln(a) + ln(x) \\
ln(x/a) &= ln(x) - ln(a) \\
ln(x^a) &= aln(x)
\end{aligned}
$$

Logarithms and percentages

- **Because**

$$ln(x + \Delta x) - ln(x) = ln\left(\frac{x + \Delta x}{x}\right)$$
$$\cong \frac{\Delta x}{x} \; (when \; \frac{\Delta x}{x} \; is \; very \; small)$$

- **For example:**

$$ln(1 + 0.01) = ln(101) - ln(100) = 0.00995 \cong 0.01$$

- **Thus,logarithmic transforms permit modeling relations in percentage terms (like elasticities), rather than linearly.**

The three log regression specifications:

Case	Population regression function
I.linear-log	$Y_i = \beta_0 + \beta_1 ln(X_i) + u_i$
II.log-linear	$ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$
III.log-log	$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + u_i$

- The interpretation of the slope coefficient **differs** in each case.
- The interpretation is found by applying the general "before and after" rule: "figure out the change in Y for a given change in X."(Key Concept 8.1 in S.W.pp301)

I. Linear-log population regression function

- **Regression Model:**

$$Y_i = \beta_0 + \beta_1 ln(X_i) + u_i$$

- **Change X ΔX:**

$$\begin{aligned}\Delta Y &= [\beta_0 + \beta_1 ln(X + \Delta X)] - [\beta_0 + \beta_1 ln(X)] \\ &= \beta_1 [ln(X + \Delta X) - ln(X)] \\ &\cong \beta_1 \frac{\Delta X}{X}\end{aligned}$$

- **Note** $100 \times \frac{\Delta X}{X} = percentage\ change\ in\ X$**, and**

$$\beta_1 \cong \frac{\Delta Y}{\frac{\Delta X}{X}}$$

- **Interpretation of** β_1**:** *a 1 percent increase in X (multiplying X by 1.01 or* $100 \times \frac{\Delta X}{X}$*) is associated with a* $0.01\beta_1$ *or* $\frac{\beta_1}{100}$ *change in Y.*

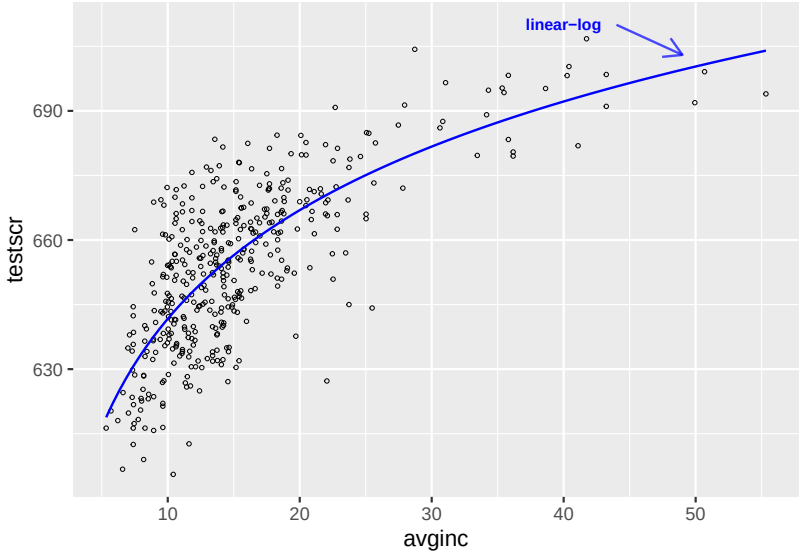
Example: the TestScore – log(Income) relation

- The OLS regression of ln(Income) on Testscore yields

$$\widehat{TestScore} = 557.8 + 36.42 \times ln(Income)$$
$$(3.8) \quad (1.4)$$

- Interpretation of β_1: **a 1% increase in Income is associated with an increase in TestScore of 0.3642 points on the test.**

Test scores: linear-log function



Case II. Log-linear population regression function

- **Regression model:**

$$ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$$

- **Change X:**

$$ln(\Delta Y + Y) - ln(Y) = [\beta_0 + \beta_1(X + \Delta X)] - [\beta_0 + \beta_1 X]$$

$$ln(1 + \frac{\Delta Y}{Y}) = \beta_1 \Delta X$$

$$\Rightarrow \frac{\Delta Y}{Y} \cong \beta_1 \Delta X$$

- **So** $100\frac{\Delta Y}{Y} = percentage\ change\ in\ Y$ **and**

$$\beta_1 = \frac{\frac{\Delta Y}{Y}}{\Delta X}$$

- **Then** *a change in X by one unit is associated with a* $\beta_1 \times 100$ *percent change in Y.*

Mincer Earning Function: log-linear functions

- Example: Age(working experience) and Earnings
- The OLS regression of age on earnings yields

$$\widehat{ln(Earnings)} = 2.811 + 0.0096 Age$$
$$(0.018) \quad (0.0004)$$

- According to this regression, when one more year old, earnings are predicted to increase by $100 \times 0.0096 = 0.96\%$

Case III. Log-log population regression function

- **the regression model is**

$$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + u_i$$

- **Change X:**

$$ln(\Delta Y + Y) - ln(Y) = [\beta_0 + \beta_1 ln(X + \Delta X)] - [\beta_0 + \beta_1 ln(X)]$$
$$ln(1 + \frac{\Delta Y}{Y}) = \beta_1 ln(1 + \frac{\Delta X}{X})$$
$$\Rightarrow \frac{\Delta Y}{Y} \cong \beta_1 \frac{\Delta X}{X}$$

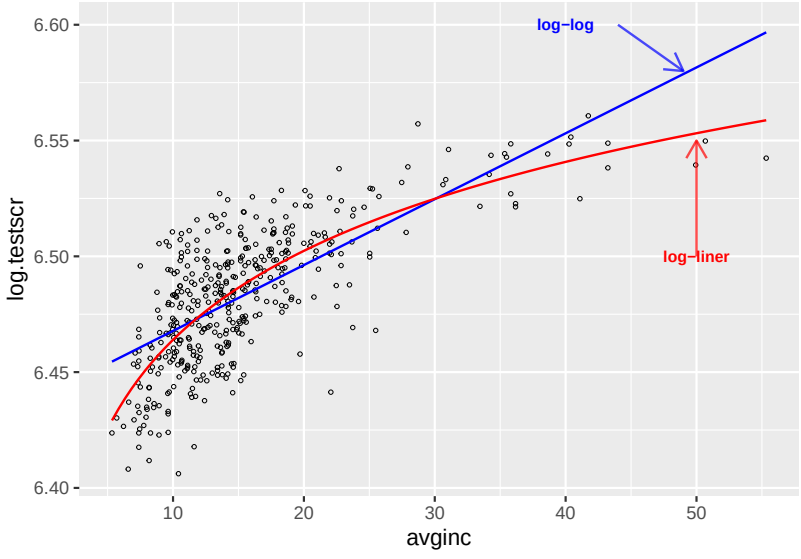
- **Now** $100\frac{\Delta Y}{Y} = percentage\ change\ in\ Y$ **and** $100\frac{\Delta X}{X} = percentage\ change\ in\ X$
- **Therefore** *a 1% change in X by one unit is associated with a* β_1*% change in Y*,**thus** β_1 **has the interpretation of an elasticity.**

Test scores and income: log-log specifications

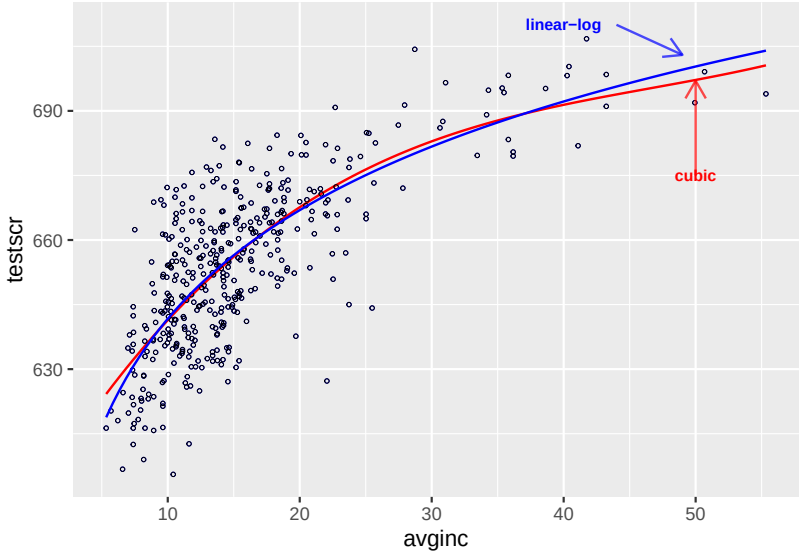
$$\widehat{ln(TestScore)} = 6.336 + 0.055 \times ln(Income)$$
$$(0.006) \quad (0.002)$$

- **A 1% increase in Income is associated with an increase of 0.055% in TestScore.**

Test scores: The log-linear and log-log functions



Test scores: The linear-log and cubic functions



Logarithmic and cubic functions

Table 3

	Dependent Variable: Test Score		
	testscr	log.testscr	testscr
	(1)	(2)	(3)
loginc	36.420***	0.055***	
		(0.002)	
avginc			5.019***
			(0.704)
I(avginc^2)			−0.096***
			(0.029)
I(avginc^3)			0.001**
			(0.0003)
Constant	557.832***	6.336***	600.079***
	(5.078)	(0.006)	(5.078)
Observations	420	420	420
Adjusted R^2	0.561	0.557	0.555

Choice of specification should be guided

- The two estimated regression functions are quite similar. So how to choose?
- The general rules:
  - By **economic logic or theories**(which interpretation makes the most sense in your application?).
  - There are several formal tests, while seldom used in reality. Actually t-test and F-test are enough.
  - Plotting predicted values and use $\overline{R^2}$ or SER can help to make further judgment.

Summary

- We can add polynomial terms of any significant variables to a model and to perform a single and joint test of significance. If the additional quadratics are significant, they can be added to the model.
- We can also change the variables values into logarithms to capture the nonlinear relationships.
- In reality, it can be difficult to pinpoint the precise reason for functional form misspecification.
- Fortunately, using **logarithms** of certain variables and adding **quadratic** or **cubic** functions are **sufficient** for detecting many(almost) important nonlinear relationships in Xs in economics.

Interactions Between Independent Variables

Introduction

- The product of two variables is called an **interaction term**.
- Try to answer *how the effect on Y of a change in an independent variable depends on the value of another independent variable.*
- Consider three cases:
  1. Interactions between **two binary** variables.
  2. Interactions between **a binary and a continuous variable**.
  3. Interactions between **two continuous variables**.

Interactions Between Two Binary Variables

- Assume we would like to study the earnings of worker in the labor market
- The population linear regression of Y_i is

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

- Dependent Variable: **log earnings**(Y_i,where $Y_i = ln(Earnings)$)
- Independent Variables: two binary variables
  - $D_{1i} = 1$ if the person graduate from college
  - $D_{2i} = 1$ if the worker's gender is female
- So β_1 is the effect on log earnings of having a college degree, *holding gender constant*, and β_2 is the effect of being female, *holding schooling constant*.

Interactions Between Two Binary Variables

- The effect of having a college degree in this specification, holding constant gender, is the **same** for men and women. No reason that this must be so.
- the effect on Y_i of D_{1i}, holding D_{2i} constant, could depend on the value of D_{2i}
- there could be an interaction between having a college degree and gender so that the value in the job market of a degree is different for men and women.
- The new regression model of Y_i is

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

- The new regressor, the product $D_{1i} \times D_{2i}$, is called an **interaction term** or an interacted regressor,

Interactions Between Two Binary Variables:

- **The regression model of Y_i now is**

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

- **Then the *conditional expectation of Yi for* $D_{1i} = 0$, given a certain value of D_{2i},d_2**

$$E(Y_i | D_{1i} = 0, D_{2i} = d_2) = \beta_0 + \beta_1 \times 0 + \beta_2 d_2 + \beta_3 (0 \times d_2) = \beta_0 + \beta_2 d_2$$

- **Then the *conditional expectation of Yi for* $D_{1i} = 1$, given a certain value of D_{2i},d_2**

$$\begin{aligned} E(Y_i | D_{1i} = 1, D_{2i} = d_2) &= \beta_0 + \beta_1 \times 1 + \beta_2 d_2 + \beta_3 (1 \times d_2) \\ &= \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2 \end{aligned}$$

Interactions Between Two Binary Variables:

- The effect of this change is the difference of expected values,which is

$$E(Y_i|D_{1i} = 1, D_{2i} = d_2) - E(Y_i|D_{1i} = 0, D_{2i} = d_2) = \beta_1 + \beta_3 d_2$$

- In the binary variable interaction specification, the effect of acquiring a college degree (a unit change in D_{1i}) depends on the person's gender.
  - If the person is male,thus $D_{2i} = d_2 = 0$,then the effect is β_1
  - If the person is female,thus $D_{2i} = d_2 = 1$,then the effect is $\beta_1 + \beta_3$
- So the coefficient β_3 is **the difference in the effect of acquiring a college degree for women versus men.**

Application: the STR and the English learners

- Let $HiSTR_i$ be a binary variable for STR
  - $HiSTR_i = 1$ if the $STR > 20$
  - $HiSTR_i = 0$ otherwise
- Let $HiEL_i$ be a binary variable for the share of English learners
  - $HiEL_i = 1$ if the $el_{pct} > 10 percent$
  - $HiEL_i = 0$ otherwise

Application: the STR and the English learners

- the OLS regression result is

$$\widehat{TestScore} = \underset{(1.4)}{664.1} - \underset{(1.9)}{1.9} HiSTR - \underset{(2.3)}{18.2} HiEL - \underset{(3.1)}{3.5}(HiSTR \times HiEL)$$

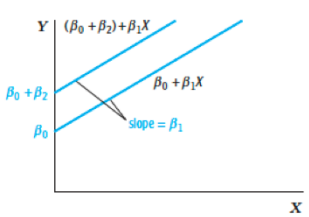
- The value of β_3 here(3.5) means that performance gap in test scores between large class($STR > 20$) and small class($STR \leq 20$) varies between the "higher-share-immigrant" class and the "lower-share immigrants" class.
- More precisely,the gap of test scores is positively related with the "higher-share-immigrant" class though insignificantly.

Interactions: a Continuous and a Binary Variable

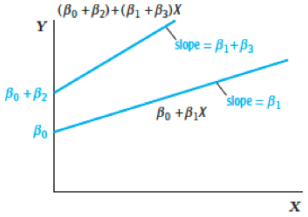
- **Binary Variable**: eg, whether the worker has a college degree (D_i)
- **Continuous Variable**: eg, the individual's years of work experience (X_i)
- In this case, we can have three specifications:
  1. No interaction
  $$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$
  2. a interaction and only one independent variable
  $$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (D_i \times X_i) + u_i$$
  3. Interaction and two independent variables
  $$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (D_i \times X_i) + u_i$$

A Continuous and a Binary Variable: Three Cases

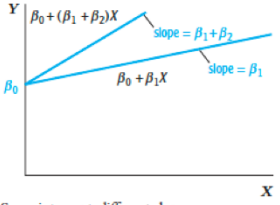
FIGURE 8.8 Regression Functions Using Binary and Continuous Variables



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

Interactions of binary variables and continuous variables can produce three different population regression functions: (a) $\beta_0 + \beta_1 X + \beta_2 D$ allows for different intercepts but has the same slope, (b) $\beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D)$ allows

A Continuous and a Binary Variable: Specifications

- All three specifications are just different versions of the multiple regression model.
- Different specifications are based on different assumptions of the relationships of X on Y depending on D.
- The **Model 3** is preferred, because it allows for both different intercepts and different slops.

Application: the STR and the English learners

- $HiEL_i$ is still a binary variable for English learner
- The estimated interaction regression

$$\widehat{TestScore} = 682.2 - 0.97STR + 5.6HiEL - 1.28(STR \times HiEL)$$
$$(11.9) \quad (0.59) \quad (19.5) \qquad\qquad (0.97)$$
$$\overline{R^2} = 0.305$$

- For districts with a low fraction of English learners,the estimated regression line is $682.2 - 0.97STR_i$
- For districts with a high fraction of English learners,the estimated regression line is $682.2 + 5.6 - 0.97STR_i - 1.28STR_i = 687.8 - 2.25STR_i$
- The difference between these two effects, 1.28 points, is the coefficient on the interaction term.

Application: the STR and the English learners

- The value of β_3 here(-1.28) means that *the effect of class size on test scores varies between the "higher-share-immigrant" class and the "lower-share immigrants or more native" class.*
- More precisely,negatively related with the "higher-share-immigrant" class though insignificantly.

Hypotheses Testing

1. High fraction is the same as low fraction, thus the two line are in fact the same

- computing the **F-statistic** testing the joint hypothesis

$$\beta_2 = \beta_3 = 0$$

- This F-statistic is 89.9, which is significant at the 1% level.

2. The effects between two groups is the same,thus two lines have the same slope

- testing whether the coefficient on the interaction term is zero, which can be tested by using a **t-statistic**
- This t-statistic is -1.32, which is insignificant at the 10% level.

Hypotheses Testing

3. the lines have the same intercept

- Testing that the population coefficient on $HiEL$ is zero,which can be tested by using a **t-statistic**.
- This t-statistic is 0.29, which is insignificant even at the 10% level.
- The reason is that the regressors, $HiEL$ and $STR * HiEL$, are highly correlated. Then large standard errors on the individual coefficients.
- Even though it is impossible to tell which of the coefficients is nonzero, there is strong evidence against the hypothesis that both are zero.

Interactions Between Two Continuous Variables

- Now suppose that both independent variables (X_{1i} and X_{2i}) are continuous.
  - X_{1i} is his or her years of work experience
  - X_{2i} is the number of years he or she went to school.
- there might be an interaction between these two variables so that the effect on wages of an additional year of experience depends on the number of years of education.
- the population regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

Interactions Between Two Continuous Variables

- Thus the effect on Y of a change in X_1, holding X_2 constant, is

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$$

- A similar calculation shows that the effect on Y of a change ΔX_1 in X_2, holding X_1 constant, is

$$\frac{\Delta Y}{\Delta X_2} = \beta_2 + \beta_3 X_1$$

- That is, if X_1 changes by ΔX_1 and X_2 changes by ΔX_2, then the expected change in Y

$$\Delta Y = (\beta_1 + \beta_3 X_2)\Delta X_1 + (\beta_2 + \beta_3 X_1)\Delta X_2 + \beta_3 \Delta X_1 \Delta X_2$$

Application: the STR and the English learners

- **The estimated interaction regression**

$$\widehat{ln(TestScore)} = 686.3 - 1.12STR - 0.67PctEL + 0.0012(STR \times PctEL)$$
$$(11.8) \quad (0.059) \quad (0.037) \qquad (0.019)$$

- **The value of β_3 here means** *how the effect of class size on test scores varies along with the share of English learners in the class*.
- **More precisely,** *increase 1 unit of the share of English learners* **make the effect of class size on test scores increase extra 0.0012 scores.**

Application: the STR and the English learners

- when the percentage of English learners is at the **median**($PctEL = 8.85$), the slope of the line relating test scores and the STR is

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2 = -1.12 + 0.0012 \times 8.85 = -1.11$$

- when the percentage of English learners is at the **75th percentile**($PctEL = 23.0$), the slope of the line relating test scores and the STR is

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2 = -1.12 + 0.0012 \times 23.0 = -1.09$$

- The difference between these estimated effects is not statistically significant.Because?
  - The t-statistic testing whether the coefficient on the interaction term is zero $t = 0.0012/0/019 = 0.06$

Application: STR and Test Scores in a Summary

- Although these nonlinear specifications extend our knowledge about the relationship between STR and Testscore, it must be augmented with control variables such as **economic background** to avoid OVB bias.
- Two measures of the economic background of the students:
  1. the percentage of students eligible for a subsidized lunch
  2. the logarithm of average district income.

Application: STR and Test Scores in a Summary

- Then three specific questions about test scores and the student–teacher ratio.
  1. After controlling for differences in economic characteristics, does the effect on test scores of STR depend on the fraction of English learners?
  2. Does this effect depend on the value of the student–teacher ratio(STR)?
  3. Most important, after taking economic factors and nonlinearities into account,what is the estimated effect on test scores of reducing the student–teacher ratio by **2** students per teacher?

	score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
str	−1.00***	−0.73**	−0.97	−0.53	64.34**	83.70**	65.29**
	(0.27)	(0.26)	(0.59)	(0.34)	(24.86)	(28.50)	(25.26)
I(str^2)					−3.42**	−4.38**	−3.47**
					(1.25)	(1.44)	(1.27)
I(str^3)					0.06**	0.07**	0.06**
					(0.02)	(0.02)	(0.02)
str:HiEL			−1.28	−0.58		−123.28*	
			(0.97)	(0.50)		(50.21)	
I(str^2):HiEL						6.12*	
						(2.54)	
I(str^3):HiEL						−0.10*	
						(0.04)	
english	−0.12***	−0.18***					−0.17***
	(0.03)	(0.03)					(0.03)
HiEL			5.64	5.50	−5.47***	816.08*	
			(19.51)	(9.80)	(1.03)	(327.67)	
lunch	−0.55***	−0.40***		−0.41***	−0.42***	−0.42***	−0.40***
	(0.02)	(0.03)		(0.03)	(0.03)	(0.03)	(0.03)
log(income)		11.57***		12.12***	11.75***	11.80***	11.51***
		(1.82)		(1.80)	(1.77)	(1.78)	(1.81)
Constant	700.15***	658.55***	682.25***	653.67***	252.05	122.35	244.81
	(5.57)	(8.64)	(11.87)	(9.87)	(163.63)	(185.52)	(165.72)
N	420	420	420	420	420	420	420
Adjusted R^2	0.77	0.79	0.31	0.79	0.80	0.80	0.80

*p < .05; **p < .01; ***p < .001

Robust S.E. are shown in the parentheses

	score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
str	−1.00***	−0.73**	−0.97	−0.53	64.34**	83.70**	65.29**
	(0.27)	(0.26)	(0.59)	(0.34)	(24.86)	(28.50)	(25.26)
I(str^2)					−3.42**	−4.38**	−3.47**
					(1.25)	(1.44)	(1.27)
I(str^3)					0.06**	0.07**	0.06**
					(0.02)	(0.02)	(0.02)
str:HiEL			−1.28	−0.58		−123.28*	
			(0.97)	(0.50)		(50.21)	
I(str^2):HiEL						6.12*	
						(2.54)	
I(str^3):HiEL						−0.10*	
						(0.04)	
english	−0.12***	−0.18***					−0.17***
	(0.03)	(0.03)					(0.03)
HiEL			5.64	5.50	−5.47***	816.08*	
			(19.51)	(9.80)	(1.03)	(327.67)	
lunch	−0.55***	−0.40***		−0.41***	−0.42***	−0.42***	−0.40***
	(0.02)	(0.03)		(0.03)	(0.03)	(0.03)	(0.03)
log(income)		11.57***		12.12***	11.75***	11.80***	11.51***
		(1.82)		(1.80)	(1.77)	(1.78)	(1.81)
Constant	700.15***	658.55***	682.25***	653.67***	252.05	122.35	244.81
	(5.57)	(8.64)	(11.87)	(9.87)	(163.63)	(185.52)	(165.72)
N	420	420	420	420	420	420	420
Adjusted R^2	0.77	0.79	0.31	0.79	0.80	0.80	0.80

*p < .05; **p < .01; ***p < .001

Robust S.E. are shown in the parentheses

	score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
str	−1.00***	−0.73**	−0.97	−0.53	64.34**	83.70**	65.29**
	(0.27)	(0.26)	(0.59)	(0.34)	(24.86)	(28.50)	(25.26)
I(str^2)					−3.42**	−4.38**	−3.47**
					(1.25)	(1.44)	(1.27)
I(str^3)					0.06**	0.07**	0.06**
					(0.02)	(0.02)	(0.02)
str:HiEL			−1.28	−0.58		−123.28*	
			(0.97)	(0.50)		(50.21)	
I(str^2):HiEL						6.12*	
						(2.54)	
I(str^3):HiEL						−0.10*	
						(0.04)	
english	−0.12***	−0.18***					−0.17***
	(0.03)	(0.03)					(0.03)
HiEL			5.64	5.50	−5.47***	816.08*	
			(19.51)	(9.80)	(1.03)	(327.67)	
lunch	−0.55***	−0.40***		−0.41***	−0.42***	−0.42***	−0.40***
	(0.02)	(0.03)		(0.03)	(0.03)	(0.03)	(0.03)
log(income)		11.57***		12.12***	11.75***	11.80***	11.51***
		(1.82)		(1.80)	(1.77)	(1.78)	(1.81)
Constant	700.15***	658.55***	682.25***	653.67***	252.05	122.35	244.81
	(5.57)	(8.64)	(11.87)	(9.87)	(163.63)	(185.52)	(165.72)
N	420	420	420	420	420	420	420
Adjusted R^2	0.77	0.79	0.31	0.79	0.80	0.80	0.80

*p < .05; **p < .01; ***p < .001

Robust S.E. are shown in the parentheses

	score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
str	−1.00***	−0.73**	−0.97	−0.53	64.34**	83.70**	65.29**
	(0.27)	(0.26)	(0.59)	(0.34)	(24.86)	(28.50)	(25.26)
I(str^2)					−3.42**	−4.38**	−3.47**
					(1.25)	(1.44)	(1.27)
I(str^3)					0.06**	0.07**	0.06**
					(0.02)	(0.02)	(0.02)
str:HiEL			−1.28	−0.58		−123.28*	
			(0.97)	(0.50)		(50.21)	
I(str^2):HiEL						6.12*	
						(2.54)	
I(str^3):HiEL						−0.10*	
						(0.04)	
english	−0.12***	−0.18***					−0.17***
	(0.03)	(0.03)					(0.03)
HiEL			5.64	5.50	−5.47***	816.08*	
			(19.51)	(9.80)	(1.03)	(327.67)	
lunch	−0.55***	−0.40***		−0.41***	−0.42***	−0.42***	−0.40***
	(0.02)	(0.03)		(0.03)	(0.03)	(0.03)	(0.03)
log(income)		11.57***		12.12***	11.75***	11.80***	11.51***
		(1.82)		(1.80)	(1.77)	(1.78)	(1.81)
Constant	700.15***	658.55***	682.25***	653.67***	252.05	122.35	244.81
	(5.57)	(8.64)	(11.87)	(9.87)	(163.63)	(185.52)	(165.72)
N	420	420	420	420	420	420	420
Adjusted R^2	0.77	0.79	0.31	0.79	0.80	0.80	0.80

*p < .05; **p < .01; ***p < .001

Robust S.E. are shown in the parentheses

	score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
str	−1.00***	−0.73**	−0.97	−0.53	64.34**	83.70**	65.29**
	(0.27)	(0.26)	(0.59)	(0.34)	(24.86)	(28.50)	(25.26)
I(str^2)					−3.42**	−4.38**	−3.47**
					(1.25)	(1.44)	(1.27)
I(str^3)					0.06**	0.07**	0.06**
					(0.02)	(0.02)	(0.02)
str:HiEL			−1.28	−0.58		−123.28*	
			(0.97)	(0.50)		(50.21)	
I(str^2):HiEL						6.12*	
						(2.54)	
I(str^3):HiEL						−0.10*	
						(0.04)	
english	−0.12***	−0.18***					−0.17***
	(0.03)	(0.03)					(0.03)
HiEL			5.64	5.50	−5.47***	816.08*	
			(19.51)	(9.80)	(1.03)	(327.67)	
lunch	−0.55***	−0.40***		−0.41***	−0.42***	−0.42***	−0.40***
	(0.02)	(0.03)		(0.03)	(0.03)	(0.03)	(0.03)
log(income)		11.57***		12.12***	11.75***	11.80***	11.51***
		(1.82)		(1.80)	(1.77)	(1.78)	(1.81)
Constant	700.15***	658.55***	682.25***	653.67***	252.05	122.35	244.81
	(5.57)	(8.64)	(11.87)	(9.87)	(163.63)	(185.52)	(165.72)
N	420	420	420	420	420	420	420
Adjusted R^2	0.77	0.79	0.31	0.79	0.80	0.80	0.80

*p < .05; **p < .01; ***p < .001

Robust S.E. are shown in the parentheses

	score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
str	−1.00***	−0.73**	−0.97	−0.53	64.34**	83.70**	65.29**
	(0.27)	(0.26)	(0.59)	(0.34)	(24.86)	(28.50)	(25.26)
I(str^2)					−3.42**	−4.38**	−3.47**
					(1.25)	(1.44)	(1.27)
I(str^3)					0.06**	0.07**	0.06**
					(0.02)	(0.02)	(0.02)
str:HiEL			−1.28	−0.58		−123.28*	
			(0.97)	(0.50)		(50.21)	
I(str^2):HiEL						6.12*	
						(2.54)	
I(str^3):HiEL						−0.10*	
						(0.04)	
english	−0.12***	−0.18***					−0.17***
	(0.03)	(0.03)					(0.03)
HiEL			5.64	5.50	−5.47***	816.08*	
			(19.51)	(9.80)	(1.03)	(327.67)	
lunch	−0.55***	−0.40***		−0.41***	−0.42***	−0.42***	−0.40***
	(0.02)	(0.03)		(0.03)	(0.03)	(0.03)	(0.03)
log(income)		11.57***		12.12***	11.75***	11.80***	11.51***
		(1.82)		(1.80)	(1.77)	(1.78)	(1.81)
Constant	700.15***	658.55***	682.25***	653.67***	252.05	122.35	244.81
	(5.57)	(8.64)	(11.87)	(9.87)	(163.63)	(185.52)	(165.72)
N	420	420	420	420	420	420	420
Adjusted R^2	0.77	0.79	0.31	0.79	0.80	0.80	0.80

*p < .05; **p < .01; ***p < .001

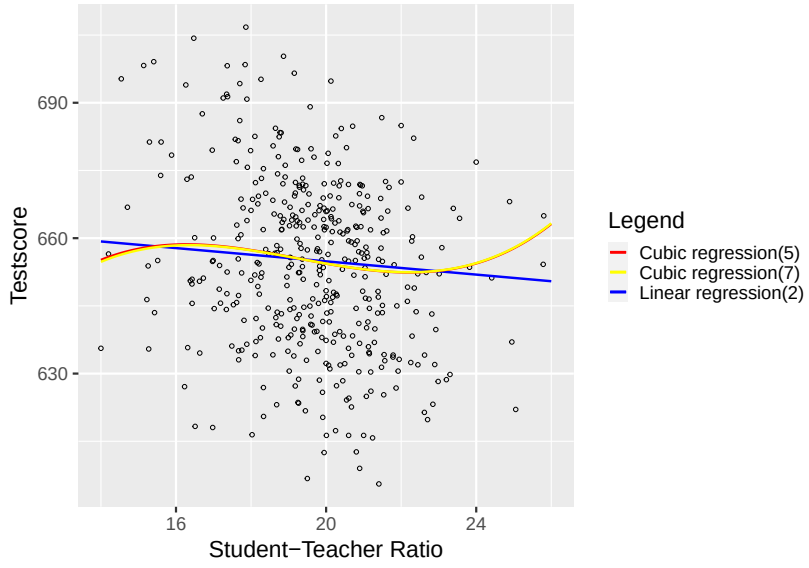
Robust S.E. are shown in the parentheses

	score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
str	−1.00***	−0.73**	−0.97	−0.53	64.34**	83.70**	65.29**
	(0.27)	(0.26)	(0.59)	(0.34)	(24.86)	(28.50)	(25.26)
I(str^2)					−3.42**	−4.38**	−3.47**
					(1.25)	(1.44)	(1.27)
I(str^3)					0.06**	0.07**	0.06**
					(0.02)	(0.02)	(0.02)
str:HiEL			−1.28	−0.58		−123.28*	
			(0.97)	(0.50)		(50.21)	
I(str^2):HiEL						6.12*	
						(2.54)	
I(str^3):HiEL						−0.10*	
						(0.04)	
english	−0.12***	−0.18***					−0.17***
	(0.03)	(0.03)					(0.03)
HiEL			5.64	5.50	−5.47***	816.08*	
			(19.51)	(9.80)	(1.03)	(327.67)	
lunch	−0.55***	−0.40***		−0.41***	−0.42***	−0.42***	−0.40***
	(0.02)	(0.03)		(0.03)	(0.03)	(0.03)	(0.03)
log(income)		11.57***		12.12***	11.75***	11.80***	11.51***
		(1.82)		(1.80)	(1.77)	(1.78)	(1.81)
Constant	700.15***	658.55***	682.25***	653.67***	252.05	122.35	244.81
	(5.57)	(8.64)	(11.87)	(9.87)	(163.63)	(185.52)	(165.72)
N	420	420	420	420	420	420	420
Adjusted R^2	0.77	0.79	0.31	0.79	0.80	0.80	0.80

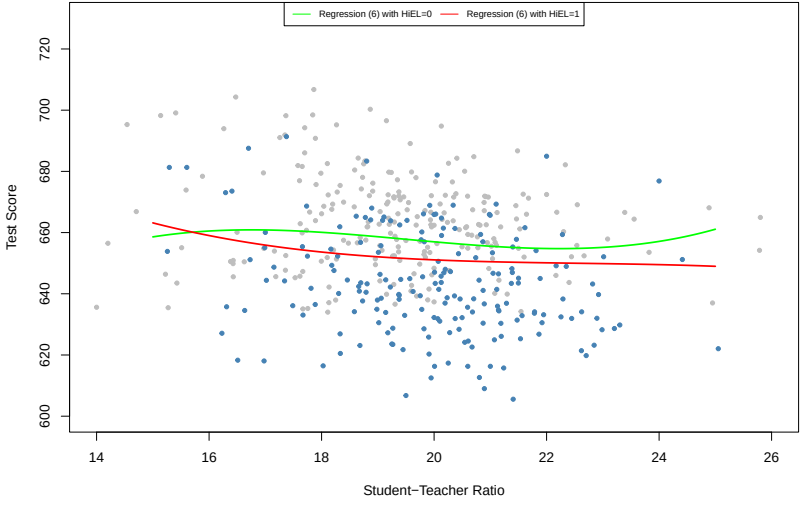
*p < .05; **p < .01; ***p < .001

Robust S.E. are shown in the parentheses

Three Regressions on graph



Interaction on graph



A Lastest and Smart Application: Jia and Ku(2019)

Jia and Ku(2019)

- Ruixue Jia and Hyejin Ku, "Is China's Pollution the Culprit for the Choking of South Korea?Evidence from the Asian Dust",The Economic Journal.
- **Main Question**: Whether the air pollution spillover from China to South Korea and affect the health of South Koreans?

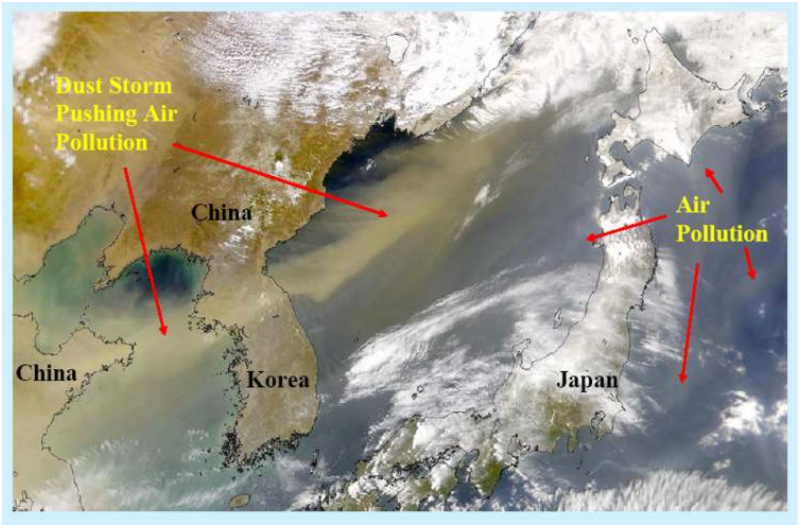
Empirical Strategy

- A naive strategy:
  - Dependent variable: **Deaths in South Korea**(respiratory and cardiovascular mortality)
  - Independt variable: **Chinese pollution**(Air Quality Index)
- Because the observed or measured air quality (i.e., pollution concentration) in Seoul or Tokyo increases in periods when China is more polluted does not mean that the pollution must have **originated from China**.

Jia and Ku(2019): Asian Dust as a carrier of pollutants

- **Asian Dust** (also yellow dust, yellow sand, yellow wind or China dust storms) is a **meteorological phenomenon** which affects much of East Asia year round but especially during the spring months.
  - The dust originates in China, the deserts of Mongolia, and Kazakhstan where high-speed surface winds and intense dust storms kick up dense clouds of fine, dry soil particles.
  - These clouds are then carried eastward by prevailing winds and pass over China, North and South Korea, and Japan, as well as parts of the Russian Far East.
  - In recent decades,Asian dust brings with it China's man-made pollution as well as its by-products.

Jia and Ku(2019): Asian Dust



Jia and Ku(2019): Asian Dust

1. A clear directional aspect in that the wind which transport Chinese pollutants to Korea but not vice versa.
2. Exogenous to South Korea's local activities. And wind patterns and topography generate rich spatial and temporal variation in the incidence.
3. The occurrence of Asian dust is monitored and recorded station by station in South Korea.(because of its visual salience)

Econometric Method: OLS with an interaction term

- **Dependent variable**: *Deaths in South Korea*(respiratory and cardiovascular mortality of South Koreans)
- **Treatment variable**: *Chinese pollution*(Air Quality Index in China)
- **Interaction Variable**: **Asian dust**(the number of Asian dust days in South Korea)
- **Control Variables**: Time, Regions, Weather,Local Economic Conditions

Jia and Ku(2019): Estimation Strategy

- **The impact of *Chinese pollution* on district-level *mortality* that operates via Asian dust**

$$\begin{aligned} Mortality_{ijk} = \beta_0 &+ \beta_1 AsianDust_{ijk} + \beta_2 ChinesePollution_{jk} \\ &+ \beta_3 AsianDust_{ijk} \times ChinesePollution_{jk} \\ &+ \delta_1 X_{ijk} + u_{ijk} \end{aligned}$$

- **Main coefficient of interest is β_3, which measures the effect of Chinese pollution in year j and month k on mortality in district i of South Korea.**

Jia and Ku(2019): the result of interaction terms

Table 2: The Impact of Dust*China's Pollution on Mortality Rates in South Korea

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Baseline							Placebo Tests	
	Mortality rates: Respiratory and Cardiovascular							Cancers	Accidents
Mean Dependent Var.	12.23							16.30	4.21
#Dust*China's Mean AQI				0.038**	0.033*	0.043**	0.040**	-0.008	0.007
				(0.016)	(0.018)	(0.018)	(0.019)	(0.020)	(0.009)
#Dust	0.076***		0.039	-0.251*	-0.214	-0.313**	0.072	0.525*	-0.102
	(0.025)		(0.032)	(0.131)	(0.142)	(0.149)	(0.240)	(0.275)	(0.119)
China's Mean AQI		0.265***	0.193*	0.117	0.138	0.202*	0.200*	-0.080	0.005
		(0.081)	(0.104)	(0.105)	(0.107)	(0.110)	(0.111)	(0.121)	(0.060)
District FE*Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Province FE*Month FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Weather (cubic polynomial)					Y	Y	Y	Y	Y
Local Prod (export, energy)						Y	Y	Y	Y
Local Prod*Dust							Y	Y	Y
Observations	29,464	29,464	29,464	29,464	28,952	28,024	28,024	28,024	28,024
R-squared	0.695	0.695	0.695	0.696	0.703	0.717	0.717	0.718	0.473

Jia and Ku(2019): the result of interaction terms

Table 2: The Impact of Dust*China's Pollution on Mortality Rates in South Korea

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Baseline							Placebo Tests	
	Mortality rates: Respiratory and Cardiovascular							Cancers	Accidents
Mean Dependent Var.	12.23							16.30	4.21
#Dust*China's Mean AQI				0.038**	0.033*	0.043**	0.040**	-0.008	0.007
				(0.016)	(0.018)	(0.018)	(0.019)	(0.020)	(0.009)
#Dust	0.076***		0.039	-0.251*	-0.214	-0.313**	0.072	0.525*	-0.102
	(0.025)		(0.032)	(0.131)	(0.142)	(0.149)	(0.240)	(0.275)	(0.119)
China's Mean AQI		0.265***	0.193*	0.117	0.138	0.202*	0.200*	-0.080	0.005
		(0.081)	(0.104)	(0.105)	(0.107)	(0.110)	(0.111)	(0.121)	(0.060)
District FE*Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Province FE*Month FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Weather (cubic polynomial)					Y	Y	Y	Y	Y
Local Prod (export, energy)						Y	Y	Y	Y
Local Prod*Dust							Y	Y	Y
Observations	29,464	29,464	29,464	29,464	28,952	28,024	28,024	28,024	28,024
R-squared	0.695	0.695	0.695	0.696	0.703	0.717	0.717	0.718	0.473

Jia and Ku(2019): the result of interaction terms

Table 2: The Impact of Dust*China's Pollution on Mortality Rates in South Korea

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Baseline							Placebo Tests	
	Mortality rates: Respiratory and Cardiovascular							Cancers	Accidents
Mean Dependent Var.	12.23							16.30	4.21
#Dust*China's Mean AQI				0.038**	0.033*	0.043**	0.040**	-0.008	0.007
				(0.016)	(0.018)	(0.018)	(0.019)	(0.020)	(0.009)
#Dust	0.076***		0.039	-0.251*	-0.214	-0.313**	0.072	0.525*	-0.102
	(0.025)		(0.032)	(0.131)	(0.142)	(0.149)	(0.240)	(0.275)	(0.119)
China's Mean AQI		0.265***	0.193*	0.117	0.138	0.202*	0.200*	-0.080	0.005
		(0.081)	(0.104)	(0.105)	(0.107)	(0.110)	(0.111)	(0.121)	(0.060)
District FE*Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Province FE*Month FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Weather (cubic polynomial)					Y	Y	Y	Y	Y
Local Prod (export, energy)						Y	Y	Y	Y
Local Prod*Dust							Y	Y	Y
Observations	29,464	29,464	29,464	29,464	28,952	28,024	28,024	28,024	28,024
R-squared	0.695	0.695	0.695	0.696	0.703	0.717	0.717	0.718	0.473

Jia and Ku(2019): the result of interaction terms

Table 2: The Impact of Dust*China's Pollution on Mortality Rates in South Korea

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Baseline							Placebo Tests	
	Mortality rates: Respiratory and Cardiovascular							Cancers	Accidents
Mean Dependent Var.	12.23							16.30	4.21
#Dust*China's Mean AQI				0.038**	0.033*	0.043**	0.040**	-0.008	0.007
				(0.016)	(0.018)	(0.018)	(0.019)	(0.020)	(0.009)
#Dust	0.076***		0.039	-0.251*	-0.214	-0.313**	0.072	0.525*	-0.102
	(0.025)		(0.032)	(0.131)	(0.142)	(0.149)	(0.240)	(0.275)	(0.119)
China's Mean AQI		0.265***	0.193*	0.117	0.138	0.202*	0.200*	-0.080	0.005
		(0.081)	(0.104)	(0.105)	(0.107)	(0.110)	(0.111)	(0.121)	(0.060)
District FE*Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Province FE*Month FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Weather (cubic polynomial)					Y	Y	Y	Y	Y
Local Prod (export, energy)						Y	Y	Y	Y
Local Prod*Dust							Y	Y	Y
Observations	29,464	29,464	29,464	29,464	28,952	28,024	28,024	28,024	28,024
R-squared	0.695	0.695	0.695	0.696	0.703	0.717	0.717	0.718	0.473

Jia and Ku(2019): Placebo Test

Table 2: The Impact of Dust*China's Pollution on Mortality Rates in South Korea

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				Baseline				Placebo Tests	
	Mortality rates: Respiratory and Cardiovascular							Cancers	Accidents
Mean Dependent Var.	12.23							16.30	4.21
#Dust*China's Mean AQI				0.038**	0.033*	0.043**	0.040**	-0.008	0.007
				(0.016)	(0.018)	(0.018)	(0.019)	(0.020)	(0.009)
#Dust	0.076***		0.039	-0.251*	-0.214	-0.313**	0.072	0.525*	-0.102
	(0.025)		(0.032)	(0.131)	(0.142)	(0.149)	(0.240)	(0.275)	(0.119)
China's Mean AQI		0.265***	0.193*	0.117	0.138	0.202*	0.200*	-0.080	0.005
		(0.081)	(0.104)	(0.105)	(0.107)	(0.110)	(0.111)	(0.121)	(0.060)
District FE*Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Province FE*Month FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Weather (cubic polynomial)					Y	Y	Y	Y	Y
Local Prod (export, energy)						Y	Y	Y	Y
Local Prod*Dust							Y	Y	Y
Observations	29,464	29,464	29,464	29,464	28,952	28,024	28,024	28,024	28,024
R-squared	0.695	0.695	0.695	0.696	0.703	0.717	0.717	0.718	0.473

Summary

Wrap up

- We extend our multiple ols model form linear to nonlinear in Xs(the independent variables)
  - 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 between a standard multiple OLS regression and a nonlinear OLS regression model in Xs is how to explain estimating coefficients.
- All are very useful and common tools with OLS regressions. You had better understand it very clear.

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
- 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 Dependent Variables

- 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
- **Binary outcomes models** is 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 $X_{1i}, ..., X_{ki}$**

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

- **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** 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. 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.(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**: An option is comparing 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.**

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

- **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.

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

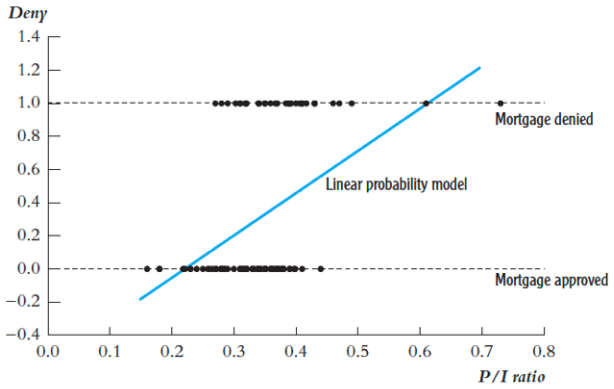
- Always use **heteroskedasticity robust standard errors** when estimating a linear probability model!

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

- 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

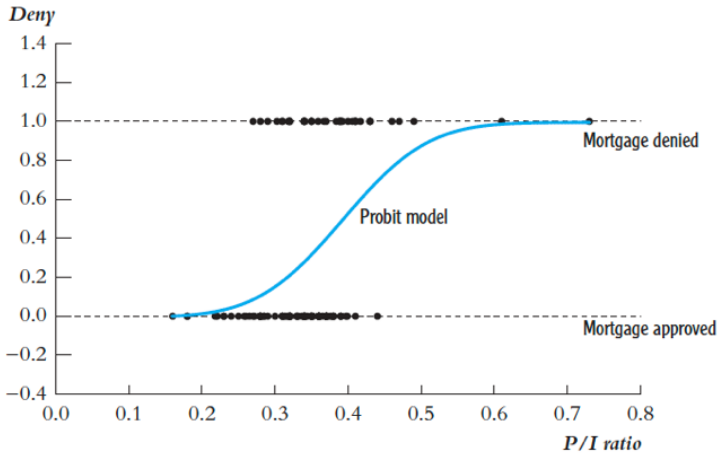
$$0 \leq \Phi(Z) \leq 1$$

- 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 $\Pr(Y = 1 \mid X)$. 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:

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

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

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

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**

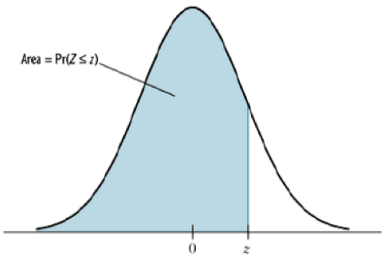
$$Pr(Y = 1|X_1 = 0.5) = Pr(z \leq -0.5) = \Phi(-0.5)$$

- **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, $\Phi(z) = \Pr(Z \leq z)$



	Second Decimal Value of z									
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:**

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

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

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

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 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**

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

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

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

- **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:

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

Example: Race in Mortgage Applications

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

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

- **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

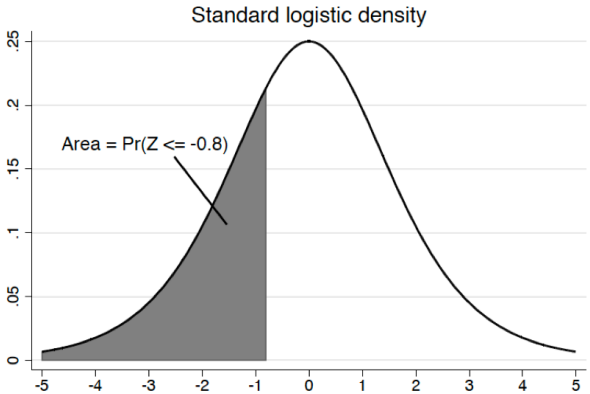
$$Z = -2 + 3 \times 0.4 = -0.8$$

- 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

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



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.
- Different from the Probit model
  - The **odds ratio**

Logit Model: the Odds Ratio

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

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

- 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**

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

- **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)**

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

- **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:** $P/I\ ratio_{mean} = 0.331$

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

- **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**

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

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 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.
- 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 $Y_1, ..., Y_n$ 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

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

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

Maximum Likelihood Estimation

MLE Step 3: *Maximize the likelihood function*

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

Maximum Likelihood Estimation

MLE Step 3: *Maximize the likelihood function*

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

$$\frac{d}{dp} ln(f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n)) = 0$$

Maximum Likelihood Estimation

MLE Step 3: *Maximize the likelihood function*

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

$$\frac{d}{dp} ln(f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n)) = 0$$

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

Maximum Likelihood Estimation

MLE Step 3: *Maximize the likelihood function*

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

$$\frac{d}{dp} ln(f_{bernouilli}(p; Y_1 = y_1, ..., Y_n = y_n)) = 0$$

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

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

Maximum Likelihood Estimation

MLE Step 3: *Maximize the likelihood function*

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

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

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

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

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**

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

MLE of the Probit Model

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

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

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**

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

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

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

MLE of the Logit Model

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

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

- **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 **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**

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

- **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**

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

- **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

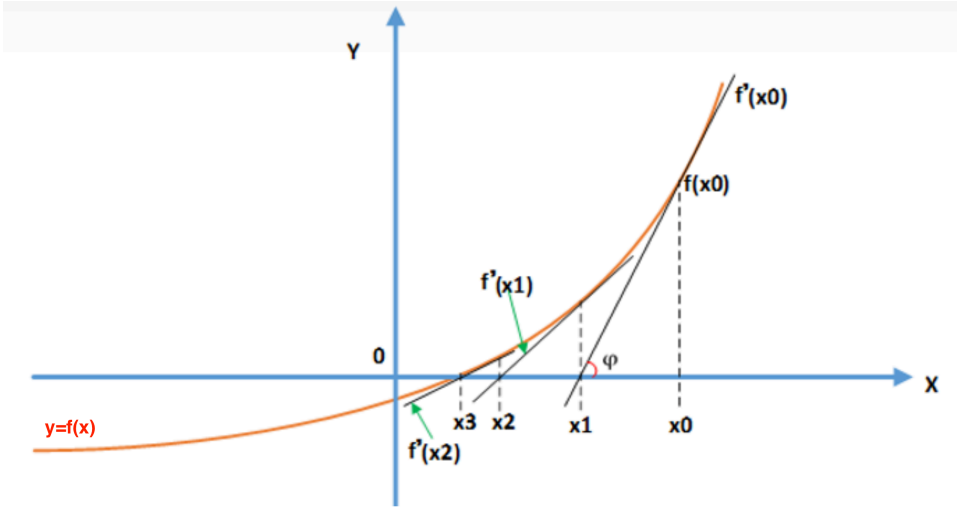
$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

- 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 $f'(x) = 0$

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 $f'(x) = 0$**

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

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 $f'(x) = 0$

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

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

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 $f'(x) = 0$

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

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

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

- repeating this procedure until

$$f'(x_j) = 0$$

, here 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 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

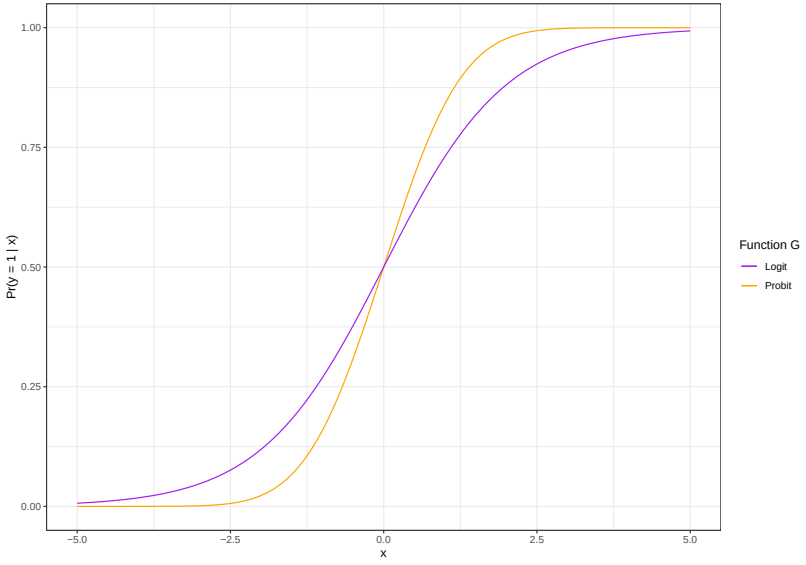
$$LR = 2(logL_{ur} - LogL_r) \sim \chi^2_q$$

- 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***	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%

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:

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

  - 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.

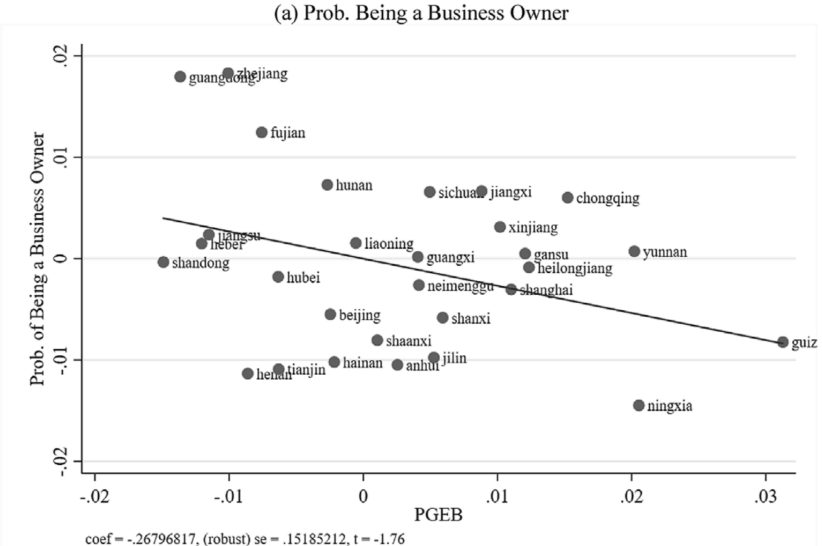
Empirical Results: LPM

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

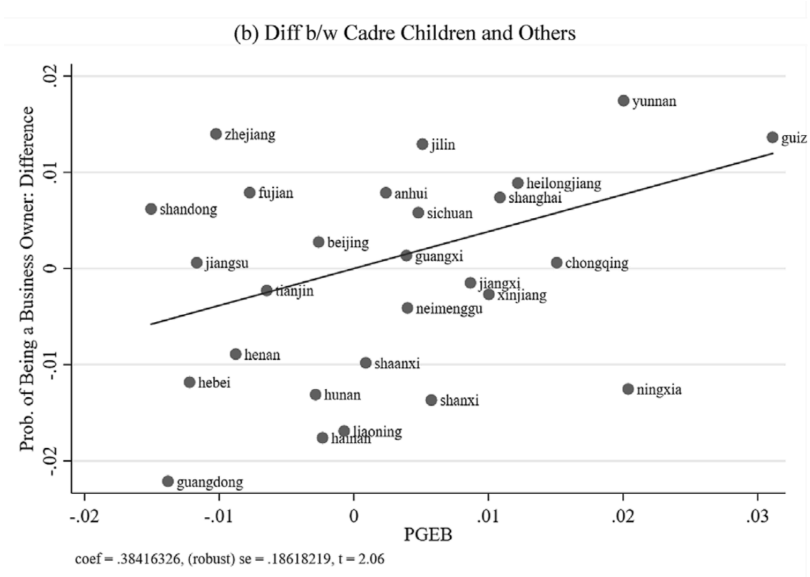
	(1)	(2)	(3)	(4)	(5)	(6)
	Government worker (0/1, mean = 0.217)		Business owner (0/1, mean = 0.022)		Self-employed (0/1, mean = 0.107)	
Cadre Parent	0.144***	0.115***	0.006**	0.003	−0.009*	−0.011**
	(0.009)	(0.009)	(0.003)	(0.003)	(0.005)	(0.005)
Entrepreneur Parent	−0.006	−0.006	0.016***	0.014**	0.063***	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 business owner**(6%).
- *Cadre Parents* **decrease** the probability of self-employment(1.1%).

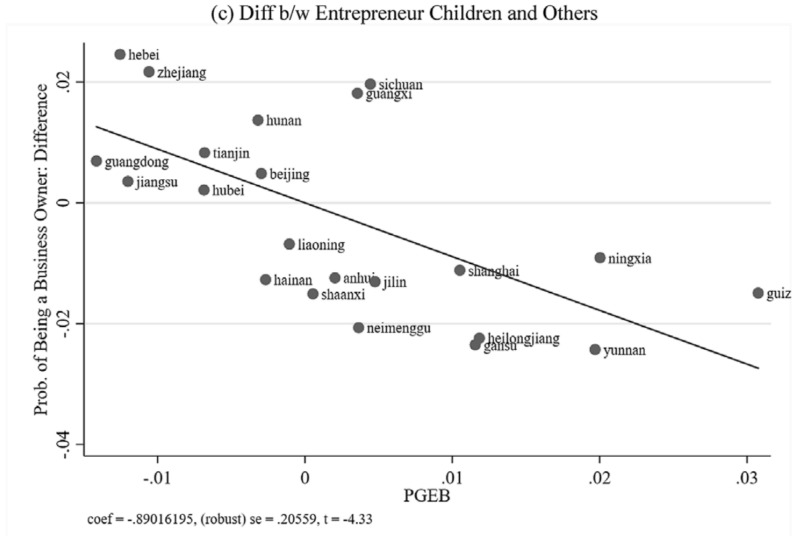
Descriptive patterns: Cross-provinces



Descriptive patterns



Descriptive patterns



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.