期末

2025年6月11日 15:01

序号7：

Regression with a Binary Dependent Variable

In the linear probability model, the predicted value of Y is interpreted as the predicted probability that Y = 1, and β1 is the change in that predicted probability for a unit change in X.

When Y *is binary, the linear regression model*

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

is called the ***linear probability model*** because

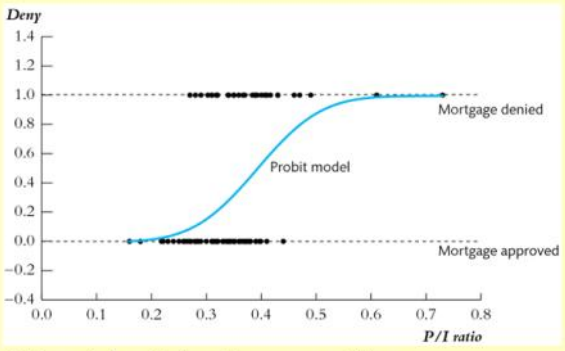
$$\Pr(Y = 1|X) = \beta_0 + \beta_1 X$$

Disadvantages:

– A LPM says that the change in the predicted probability for a given change in X is the same for all values of X, but that doesn’ t make sense. Think about the HMDA example…

– Also, LPM predicted probabilities can be 1!

These disadvantages can be solved by using a nonlinear probability model: **probit and logit regression**



- The probit model satisfies these conditions:
  - I. $\Pr(Y=1|X)$ to be increasing in X for $\beta_1 > 0$, and
  - II. $0 \le \Pr(Y=1|X) \le 1$ for all X

Probit regression models the probability that $Y = 1$ using the cumulative standard normal distribution function, $\Phi(z)$, evaluated at $z = \beta_0 + \beta_1 X$. The probit regression model is,

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

where Φ is the cumulative normal distribution function

– $z = \beta_0 + \beta_1 X$ is the “z-value” or “z-index” of the probit model.

Example: Suppose $\beta_0 = -2$, $\beta_1 = 3$, $X = .4$, so

$$\Pr(Y=1|X=.4) = \Phi(-2+3\times.4) = \Phi(-0.8)$$

$\Pr(Y = 1|X = .4)$ = area under the standard normal density to left of $z = -.8$, which is…

– $\beta_0 + \beta_1 X = z$-value

– $\hat{\beta}_0 + \hat{\beta}_1 X$ is the predicted z-value, given X

– β_1 is the change in the z-value for a unit change in X

在线性概率模型（LPM）中，模型假设当自变量X变化一个单位时，预测的发生概率的变化是固定的，不论X的初始值是多少。换句话说，这个模型认为概率-变化关系是线性的。然而，这样的假设在实际情况中并不合理。

以HMDA（房产抵押贷款数据集）为例，假设我们在考虑借款人收入比（P/I比例）对贷款被拒绝的影响。当P/I比例很低时，增加一点比例可能会显著提高被拒绝的概率；但当P/I比例已经非常高时，增加比例可能几乎不会再改变被拒绝的概率。这意味着概率的变化需要依据X的不同值而调整。

因此，“线性概率模型假设每单位X变化导致的概率变化是恒定的不合理”——在实际中，概率的变化通常是非线性的，更合理的模型应该允许这种关系随X的不同水平而变化，比如logit和probit模型。

Probit regression with multiple regressors

$$\Pr(Y = 1|X_1, X_2) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

- Φ is the cumulative normal distribution function.
- $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ is the "z-value" or "z-index" of the probit model.
- β_1 is the effect on the z-score of a unit change in X_1, holding constant X_2 (when a causal interpretation is justified)

Logit regression models the probability of $Y = 1$, given X, as the cumulative standard *logistic* distribution function, evaluated at $z = \beta_0 + \beta_1 X$:

$$\Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$$

where F is the cumulative logistic distribution function:

$$F(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Because logit and probit use different probability functions, the coefficients (β's) are different in logit and probit.

In practice, logit and probit are very similar – since empirical results typically don't hinge on the logit/probit choice, both tend to be used in practice.

Probit estimation by nonlinear least squares

Recall OLS: $\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$

- The result is the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

  Nonlinear least squares extends the idea of OLS to models in which the parameters enter nonlinearly:

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - \Phi(b_0 + b_1 X_i)]^2$$

How to solve this minimization problem?

- Calculus doesn't give an explicit formula for the NLLS estimators.
- Instead, the minimization problem is solved *numerically* using the computer (specialized minimization algorithms)
- Nonlinear least squares isn't actually used in practice. A more efficient estimator (smaller variance) is…

The Maximum Likelihood Estimator of the Coefficients in the Probit Model

The ***likelihood function*** is the **conditional density** of $Y_1,\ldots,Y_n$ given $X_1,\ldots,X_n$, treated as a function of the unknown parameters β_0 and β_1.

- The maximum likelihood estimator (MLE) is the value of (β_0, β_1) that maximize the likelihood function.
- The MLE is the value of (β_0, β_1) that best describe the full distribution of the data.
- In large samples, the MLE is:
  - Consistent
  - Normally distributed
  - Efficient (has the smallest variance of all consistent estimators)

Probit和Logic模型使用MLE来估计参数:

The derivation starts with the density of Y_1, given X_1:

$$\Pr(Y_1=1|X_1)=\Phi(\beta_0+\beta_1X_1)$$

$$\Pr(Y_1=0|X_1)=1-\Phi(\beta_0+\beta_1X_1)$$

so

$$\Pr(Y_1=y_1|X_1)=\Phi(\beta_0+\beta_1X_1)^{y_1}[1-\Phi(\beta_0+\beta_1X_1)]^{1-y_1}$$

The probit likelihood function is the joint density of $Y_1,\ldots,Y_n$ given $X_1,\ldots,X_n$, treated as a function of β_0, β_1:

$$f(\beta_0,\beta_1;Y_1,\ldots,Y_n|X_1,\ldots,X_n)$$
$$=\{\Phi(\beta_0+\beta_1X_1)^{Y_1}[1-\Phi(\beta_0+\beta_1X_1)]^{1-Y_1}\}\times$$
$$\ldots\times\{\Phi(\beta_0+\beta_1X_n)^{Y_n}[1-\Phi(\beta_0+\beta_1X_n)]^{1-Y_n}\}$$

The Probit Likelihood with one X (2 of 3)

$$f(\beta_0,\beta_1;Y_1,\ldots,Y_n|X_1,\ldots,X_n)$$
$$=\{\Phi(\beta_0+\beta_1X_1)^{Y_1}[1-\Phi(\beta_0+\beta_1X_1)]^{1-Y_1}\}\times$$
$$\ldots\times\{\Phi(\beta_0+\beta_1X_n)^{Y_n}[1-\Phi(\beta_0+\beta_1X_n)]^{1-Y_n}\}$$

- $\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}$ maximize this likelihood function.
- But we can't solve for the maximum explicitly! So the MLE must be maximized using numerical methods
- As in the case of no X, in large samples:
  - $\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}$ are consistent
  - $\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}$ are normally distributed
  - $\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}$ are asymptotically efficient – among all estimators (assuming the probit model is the correct model)

The only difference between probit and logit is the functional form used for the probability: Φ is replaced by the cumulative logistic function.

The R^2 and $\overline{R}^2$ don't make sense here (*why*?). So, two other specialized measures are used:

1. The ***fraction correctly predicted*** = fraction of Y's for which the predicted probability is >50% when $Y_i = 1$, or is <50% when $Y_i = 0$.
2. The ***pseudo-R^2*** measures the improvement in the value of the log likelihood, relative to having no X's (see SW App.

这个段落主要介绍了在用Logit和Probit模型进行模型拟合后，评估模型表现的两种指标：正确预测比例（fraction correctly predicted）和伪R-squared（pseudo-R^2）。

1. **正确预测比例**（Fraction correctly predicted）：这是衡量模型预测准确性的一个指标。具体定义为：
   - 对于所有观测值，当模型预测的概率大于50%时，模型会预测Y=1；
   - 当预测的概率小于50%时，模型会预测Y=0。然后，将所有实际Y=1且模型预测概率>50%的样本数与所有Y=1样本的总数相除，得到真正被正确预测为“1”的比例。此外，将Y=0且预测概率<50%的样本数除以所有Y=0的样本总数，得到被正确预测为“0”的比例。两个比例相加得到模型的总体正确预测比例。
2. **伪R-squared（pseudo-R^2）**：这是用来衡量Logit或Probit模型拟合优度的指标，反映模型在对数似然函数上的改进程度。它是通过比较模型的对数似然值（log likelihood）与没有任何自变量（即仅有截距的模型）时的对数似然值，来评估模型的改进效果。伪R^2的定义类似于线性回归中的R^2，但它适用于非线性模型（如Logit和Probit），而不像线性模型那样直接。

值得注意的是，虽然伪R^2不能完全等同于线性模型中的R^2，但在大样本条件下，它可以作为一种衡量模型拟合好坏的有用指标。

$Y_i = 0$.

2. The ***pseudo-R^2*** measures the improvement in the value of the log likelihood, relative to having no X's (see SW App. 11.2). The pseudo-R^2 simplifies to the R^2 in the linear model with normally distributed errors.

像线性模型那样直接。

值得注意的是，虽然伪R^2不能完全等同于线性模型中的R^2，但在大样本条件下，它可以作为一种衡量模型拟合好坏的有用指标。

这些评估指标帮助研究人员衡量Logit和Probit模型在分类任务中的性能，并理解模型在预测事件发生概率方面的效果。

序号8：Instrumental Variables Regressions

IV Regression: Why?

Three important threats to internal validity are:

• Omitted variable bias from a variable that is correlated with X but is unobserved (so cannot be included in the regression) and for which there are inadequate control variables;

• Simultaneous causality bias (X causes Y, Y causes X);

• Errors-in-variables bias (X is measured with error) All three problems result in E(u|X) ≠ 0.

• Instrumental variables regression can eliminate bias when E(u|X) ≠ 0– using an instrumental variable (IV), Z.

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

• The goal is an estimate of the causal effect β1. However, X is correlated with the error term, and we cannot solve the problem simply by including control variables.

• Instrumental variables (IV) regression breaks X into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate β1.

• This is done using an instrumental variable, Zi, which is correlated with Xi but uncorrelated with ui.

An **endogenous variable** is one that is correlated with u

An **exogenous variable** is one that is uncorrelated with u In IV regression, we focus on the case that X is endogenous and there is an instrument, Z, which is exogenous.

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

For an instrumental variable (an "***instrument***") Z to be valid, it must satisfy two conditions:

1. ***Instrument relevance***: $\text{corr}(Z_i, X_i) \neq 0$
2. ***Instrument exogeneity***: $\text{corr}(Z_i, u_i) = 0$

The IV estimator with one *X* and one *Z* (1 of 7)

Explanation #1: Two Stage Least Squares (TSLS)

As it sounds, TSLS has two stages – two regressions:

(1) Isolate the part of X that is uncorrelated with u by regressing X on Z using OLS:

$$X_i = \pi_0 + \pi_1 Z_i + v_i \qquad (1)$$

- Because Z_i is uncorrelated with u_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i. We don't know π_0 or π_1 but we have estimated them, so…
- Compute the predicted values of X_i, where $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, $i = 1,\ldots,n$.

(2) Replace X_i by $\hat{X}_i$ in the regression of interest: regress Y on $\hat{X}_i$ using OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i \qquad (2)$$

- **Because $\hat{X}_i$ is uncorrelated with u_i, the first least squares assumption holds for regression (2).** (This requires n to be large so that π_0 and π_1 are precisely estimated.)
- Thus, in large samples, β_1 can be estimated by OLS using regression (2)
- The resulting estimator is called the *Two Stage Least Squares* (*TSLS*) estimator, $\hat{\beta}_1^{TSLS}$.

Summary

Suppose Z_i, satisfies the two conditions for a valid instrument:

1. ***Instrument relevance***: $\text{corr}(Z_i, X_i) \neq 0$
2. ***Instrument exogeneity***: $\text{corr}(Z_i, u_i) = 0$

Two-stage least squares:

Stage 1: Regress X_i on Z_i (including an intercept), obtain the predicted values $\hat{X}_i$

Stage 2: Regress Y_i on $\hat{X}_i$ (including an intercept); the coefficient on $\hat{X}_i$ is the TSLS estimator, $\hat{\beta}_1^{TSLS}$.

$\hat{\beta}_1^{TSLS}$ is a consistent estimator of β_1.

TSLS in the supply-demand example (1 of 2)

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

Let Z = rainfall in dairy-producing regions.

Is Z a valid instrument?

TSLS in the supply-demand example (1 of 2)

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

Let Z = rainfall in dairy-producing regions.

Is Z a valid instrument?

(1) Relevant? $\text{corr}(rain_i, \ln(P_i^{butter})) \neq 0$?

Plausibly: insufficient rainfall means less grazing means less butter means higher prices

(2) Exogenous? $\text{corr}(rain_i, u_i) = 0$?

Plausibly: whether it rains in dairy-producing regions shouldn't affect demand for butter

Summary of IV Regression with a Single X and Z

- A valid instrument Z must satisfy two conditions:
  1. *relevance*: $\text{corr}(Z_i, X_i) \neq 0$
  2. *exogeneity*: $\text{corr}(Z_i, u_i) = 0$
- TSLS proceeds by first regressing X on Z to get $\hat{X}$, then regressing Y on $\hat{X}$
- The key idea is that the first stage isolates part of the variation in X that is uncorrelated with u
- If the instrument is valid, then the large-sample sampling distribution of the TSLS estimator is normal, so inference proceeds as usual

- So far we have considered IV regression with a single endogenous regressor (X) and a single instrument (Z).
- We need to extend this to:
  - multiple endogenous regressors $(X_1, \ldots, X_k)$
  - multiple included exogenous variables $(W_1, \ldots, W_r)$ or control variables

In general, a parameter is said to be identified if different values of the parameter produce different distributions of the data.

In IV regression, whether the coefficients are identified depends on the relation between the number of instruments (m) and the number of endogenous regressors (k)

The coefficients $\beta_1,\ldots,\beta_k$ are said to be:

- ***exactly identified*** if $m = k$.

  There are just enough instruments to estimate $\beta_1,\ldots,\beta_k$.

- ***overidentified*** if $m > k$.

  There are more than enough instruments to estimate $\beta_1,\ldots,\beta_k$. If so, you can test whether the instruments are valid (a test of the "overidentifying restrictions") – we'll return to this later

- ***underidentified*** if $m<k$.

  There are too few instruments to estimate $\beta_1,\ldots,\beta_k$. If so, you need to get more instruments!

$$Y_i=\beta_0+\beta_1X_{1i}+\ \ldots\ +\beta_kX_{ki}+\beta_{k+1}W_{1i}+\ldots+\beta_{k+r}W_{ri}+u_i$$

- Y_i is the ***dependent variable***
- $X_{1i},\ldots,X_{ki}$ are the ***endogenous regressors*** (potentially correlated with u_i)
- $W_{1i},\ldots,W_{ri}$ are the ***included exogenous regressors*** (uncorrelated with u_i) or ***control variables*** (included so that Z_i is uncorrelated with u_i, once the *W*'s are included)
- $\beta_0,\beta_1,\ldots,\beta_{k+r}$ are the unknown regression coefficients
- $Z_{1i},\ldots,Z_{mi}$ are the *m* ***instrumental variables*** (the ***excluded exogenous variables***)
- The coefficients are ***overidentified*** if $m > k$; ***exactly identified*** if $m = k$; and ***underidentified*** if $m < k$.

TSLS with a Single Endogenous Regressor

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2W_{1i}+\ \ldots\ +\beta_{1+r}W_{ri}+u_i$$

- *m* instruments: $Z_{1i},\ldots,Z_m$
- First stage
  - Regress X_1 on *all* the exogenous regressors: regress X_1 on $W_1,\ldots,W_r$, $Z_1,\ldots,Z_m$, and an intercept, by OLS
  - Compute predicted values $\hat{X}_{1i}$, $i=1,\ldots,n$
- Second stage
  - Regress Y on $\hat{X}_{1i}, W_1,\ldots,W_r$, and an intercept, by OLS
  - The coefficients from this second stage regression are the TSLS estimators, but *SE*s are wrong
- To get correct *SE*s, do this in a single step in your regression software

The IV Regression Assumptions

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

1. $E(u_i | W_{1i}, \ldots, W_{ri}) = 0$
   - #1 says "the exogenous regressors are exogenous."
2. $(Y_i, X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi})$ are i.i.d.
   - #2 is not new
3. The *X*'s, *W*'s, *Z*'s, and *Y* have nonzero, finite 4th moments
   - #3 is not new
4. The instruments $(Z_{1i}, \ldots, Z_{mi})$ are valid.
   - We have discussed this

- Under 1–4, TSLS and its *t*-statistic are normally distributed
- The critical requirement is that the instruments be valid

- Technically, the condition for W's being effective control variables is that the conditional mean of u_i does not depend on Z_i, given W_i:

$$E(u_i | W_i, Z_i) = E(u_i | W_i)$$

- Thus an alternative to IV regression assumption #1 is that conditional mean independence holds:

$$E(u_i | W_i, Z_i) = E(u_i | W_i)$$

  This is the IV version of the conditional mean independence assumption in Chapter 6.

- *Here is the key idea*: in many applications you need to include control variables (*W*'s) so that *Z* is plausibly exogenous (uncorrelated with *u*).

Example #1: Effect of Studying on Grades (5 of 6)

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad \text{corr}(Z_i, u_i) < 0,$$

Why might *Z* be correlated with *u*?

- Here's a hypothetical possibility: the student's sex. Suppose:
  - Roommates are randomly assigned – except always men with men and women with women.
  - Women get better grades than men, holding constant hour spent studying
  - Men are more likely to bring a video game than women
  - Then corr $(Z_i, u_i) < 0$ (males are more likely to have a [male] roommate who brings a video game – but males also tend to have lower grades, holding constant the amount of studying).
- Because corr $(Z_i, u_i) < 0$ the IV (roommate brings video game) isn't valid.
  - This is the IV version of OV bias.
  - The solution to OV bias is to control for (or include) the OV – in this case, sex.

- This logic leads you to include *W* = student's sex as a control variable in the IV regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

- The TSLS estimate reported above is from a regression that included gender as a *W* variable – along with other variables such as individual *i*'s major.
- The conditional mean independence condition for an exogenous instrument is, $E(u_i|Z_i,W_i) = E(u_i|W_i)$.

First stage regression:

$$X_i = \pi_0 + \pi_1 Z_{1i} + \ldots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \ldots + \pi_{m+k} W_{ki} + u_i$$

- The instruments are relevant if at least one of $\pi_1,\ldots,\pi_m$ are nonzero.
- The instruments are said to be ***weak*** if all the $\pi_1,\ldots,\pi_m$ are either zero or nearly zero.
- ***Weak instruments*** explain very little of the variation in X, beyond that explained by the *W*'s

What are the consequences of weak instruments?

If instruments are weak, the sampling distribution of TSLS and its *t*-statistic are not (at all) normal, even with n large.

Consider the simplest case of 1 *X*, 1 Z, no control variables:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$X_i = \pi_0 + \pi_1 Z_i + u_i$$

- The IV estimator is $\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$
- If $\text{cov}(X, Z)$ is zero or small, then s_{XZ} will be small: With weak instruments, the denominator is nearly zero.
- If so, the sampling distribution of $\hat{\beta}_1^{TSLS}$ (and its *t*-statistic) is not well approximated by its large-n normal approximation...

Measuring the Strength of Instruments in Practice: The First-Stage *F*-statistic

The first stage regression (one X):

- Regress X on $Z_1,..,Z_m,W_1,\ldots,W_k$.
- Totally irrelevant instruments ↔ *all* the coefficients on $Z_1,\ldots,Z_m$ are zero.
- The ***first-stage F-statistic*** tests the hypothesis that $Z_1,\ldots,Z_m$ do not enter the first stage regression.
- Weak instruments imply a small first stage F-statistic.

- Compute the first-stage *F*-statistic.

  Rule-of-thumb: If the first stage F-statistic is less than 10, then the set of instruments is weak.

Which F-statistic? Use the heteroskedasticity-robust F, for the usual reasons.

Checking Assumption #2: Instrument Exogeneity

Consider the simplest case:

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

- Suppose there are two valid instruments: Z_{1i}, Z_{2i}
- Then you could compute two separate TSLS estimates.
- Intuitively, if these 2 TSLS estimates are very different from each other, then something must be wrong: one or the other (or both) of the instruments must be invalid.
- The *J*-test of overidentifying restrictions makes this comparison in a statistically precise way.
- This can only be done if #*Z*'s > #*X*'s (overidentified).

Suppose # instruments = m > # *X*'s = k (overidentified)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ \ldots\ + \beta_k X_{ki} + \beta_{k+r} W_{1i} + \ \ldots\ + \beta_{k+r} W_{ri} + u_i$$

The *J*-test is the Anderson-Rubin test, using the TSLS estimator instead of the hypothesized value $\beta_{1,0}$. The recipe:

1. First estimate the equation of interest using TSLS and all m instruments; compute the predicted values $\hat{Y}_i$, using the *actual* *X*'s (not the $\hat{X}$'s used to estimate the second stage)
2. Compute the residuals $\hat{u}_i = Y_i - \hat{Y}_i$
3. Regress $\hat{u}_i$ against $Z_{1i}, \ldots, Z_{mi}, W_{1i}, \ldots, W_{ri}$
4. Compute the *F*-statistic testing the hypothesis that the coefficients on $Z_{1i}, \ldots, Z_{mi}$ are all zero;
5. The ***J-statistic*** is $J = mF$

Checking Instrument Validity: Summary (1 of 2)

This summary considers the case of a single X. The two requirements for valid instruments are:

1. *Relevance*

- At least one instrument must enter the population counterpart of the first stage regression.
- If instruments are weak, then the TSLS estimator is biased and the *t*-statistic has a non-normal distribution
- To check for weak instruments with a single included endogenous regressor, check the first-stage F
  - If $F > 10$, instruments are strong – use TSLS
  - If $F < 10$, weak instruments – take some action.

2. *Exogeneity*

- ***All*** the instruments must be uncorrelated with the error term: $\text{corr}(Z_{1i},u_i) = 0,\ldots,\text{corr}(Z_{mi},u_i) = 0$
- We can partially test for exogeneity: if $m > 1$, we can test the null hypothesis that all the instruments are exogenous, against the alternative that as many as $m - 1$ are endogenous (correlated with u)
- The test is the *J*-test, which is constructed using the TSLS residuals.
- If the *J*-test rejects, then at least some of your instruments are endogenous – so you must make a difficult decision and jettison some (or all) of your instruments.

第4章 补充内容

刷题感悟：

1. SER和R^2的公式要记熟。
2. 关注教材里的Key Concept.
3. 这一章和下一章主要考计算。要会算β、置信区间、t-value、R^2等
4. 其余章节主要关注应用，需要把案例好好看一遍。刷实际应用类型的题。

知识点：

拟合曲线过样本均值。可以代入一个求另一个。

SER的单位和Y一致。R^2无量纲（unit-free）。

KEY CONCEPT 4.1

Terminology for the Linear Regression Model with a Single Regressor

The linear regression model is

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

where

the subscript i runs over observations, $i = 1, \ldots, n$;
Y_i is the *dependent variable*, the *regressand*, or simply the *left-hand variable*;
X_i is the *independent variable*, the *regressor*, or simply the *right-hand variable*;
$\beta_0 + \beta_1 X$ is the *population regression line* or the *population regression function*;
β_0 is the *intercept* of the population regression line;
β_1 is the *slope* of the population regression line; and
u_i is the *error term*.

The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \quad (4.5)$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \quad (4.6)$$

The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n \quad (4.7)$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n. \quad (4.8)$$

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual ($\hat{u}_i$) are computed from a sample of n observations of X_i and Y_i, $i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

KEY CONCEPT 4.3

The Least Squares Assumptions for Causal Inference

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, \ldots, n,$$

where β_1 is the causal effect on Y of X, and:

1. The error term u_i has conditional mean 0 given X_i: $E(u_i|X_i) = 0$;
2. $(X_i, Y_i), i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) draws from their joint distribution; and
3. Large outliers are unlikely: X_i and Y_i have nonzero finite fourth moments.

KEY CONCEPT 4.4

Large-Sample Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

If the least squares assumptions in Key Concept 4.3 hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a jointly normal sampling distribution. The large-sample normal distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, where the variance of this distribution, $\sigma^2_{\hat{\beta}_1}$, is

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}. \tag{4.19}$$

The large-sample normal distribution of $\hat{\beta}_0$ is $N(\beta_0, \sigma^2_{\hat{\beta}_0})$, where

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ where } H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)}\right]X_i. \tag{4.20}$$