# Econometrics Summary

YiFan Li

June 1, 2020

## Contents

# 1 Linear Regression with One Regressor

## 1.1 The Model

The linear regression model with a single regressor:

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

- $X$ is the independent variable or regressor
- $Y$ is the dependent variable
- $B_0$ = intercept
- $B_1$ = slope
- $u_i$ = the regression error

## 1.2 The Ordinary Least Squares (OLS) estimator

Solve the minimization problem

$$\min \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

The resulting OLS estimators are

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum\limits_{i=1}^{n} (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

The residuals is

$$\hat{u}_i = Y_i - \hat{Y}_i$$

## 1.3 Algebraic Facts

- $\sum \hat{u}_i = 0$
- $\frac{1}{n} \sum \hat{y}_i = \bar{Y}$
- $\sum \hat{u}_i X_i = 0$
- $TSS = ESS + SSR$

## 1.4 Measures of Fit

### 1.4.1 The regression R-squared

Explained sum of squares (ESS) = $\sum\limits_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$

Total sum of squares (TSS) = $\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$

Sum of squared residuals (SSR)= $\sum_{i=1}^{n} \hat{u}_i^2$

$R^2$ measures the fraction of the variance of $Y$ that is explained by $X$.

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

It is always satisfied that $0 \leq R^2 \leq 1$.

### 1.4.2 The Standard Error os the Regression (SER)

The SER measures the spread of the observations around the regression line.

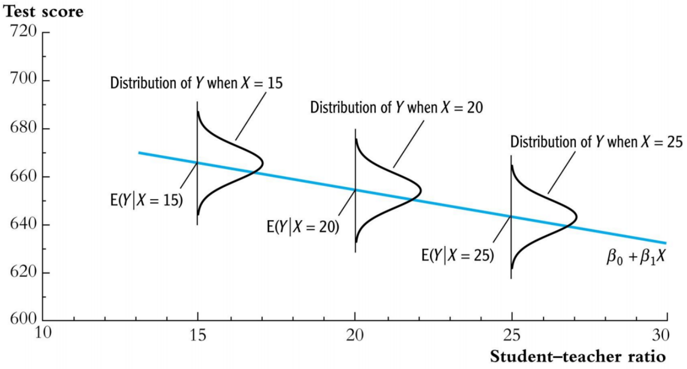
$$\text{SER} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

Division by $n-2$ is a "degrees of freedom" correction.

## 1.5 The Least Squares Assumptions

1. $E(u_i|X_i = x) = 0$. This implies that $\hat{\beta}_1$ is unbiased.
2. All $(X_i, Y_i)$ are i.i.d.
3. Large outliers in X and/or Y are rare.

Explanation for #1: It means that the conditional distribution of $u_i$ given $X_i$ has mean zero. That is, the "other factors" contained in $u_i$ should be uncorrelated with $X_i$. See the following image.



With the three assumptions, it can be derived that

- $E\left(\hat{\beta}_1\right) = \beta_1$
- $\operatorname{var}\left(\hat{\beta}_1\right) = \frac{1}{n} \times \frac{\operatorname{var}\left[\left(X_i - \mu_x\right) u_i\right]}{\sigma_X^4} \propto \frac{1}{n}$

Therefore:

- When $n$ is large, $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\operatorname{var}\left(\hat{\beta}_1\right)}}$~$N(0,1)$ because of the CLT
- The larger the variance of X, the smaller the variance of $\hat{\beta}_1$
- $\hat{\beta}_1$ is unbiased
- $\hat{\beta}_1$ is consistent, i.e. $\hat{\beta}_1 \xrightarrow{p} \beta_1$

# 2 Regression Hypothesis Tests and Confidence Intervals

Standard error of $\hat{\beta}_1$: $SE(\hat{\beta}_1)$

The standard error will be reported in such form:

$$\begin{aligned} \widehat{\text{TestScore}} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times \text{STR}, R^2 = 0.051, SER = 18.6 \end{aligned}$$

Here, 10.4 is the standard error of $\beta_0$, and 0.52 is the standard error of $\beta_1$.

## 2.1 Hypotheses Testing

t-statistic: $t = \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error}}$.

For testing $\beta_1$, $t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$, where $\beta_{1,0}$ is the hypothesized value which is provided arbitrarily.

p-value for two-sided tests = $2\Phi(-|t|)$

Reject at 5% significance level if $|t| > 1.96$, or if $p$-value $< 5\%$.

For 1% significance level, use 2.576 instead of 1.96.

## 2.2 Confidence Intervals

95% confidence interval for $\beta_1$ = $\left[\hat{\beta}_1 - 1.96\, SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96\, SE(\hat{\beta}_1)\right]$

The 99% confidence interval replaces 1.96 with 2.576.

The same goes for $\beta_0$.

## 2.3 Regression when X is Binary

When a regressor is binary, having only 0 or 1 as its value, it is called “dummy” variable.

In such case:

- $\beta_0$ = mean of Y when X = 0
- $\beta_0 + \beta_1$ = mean of Y when X = 1
- $\beta_1$ is no longer a "slope", instead it is the difference in group means.
- t-statistics, confidence Intervals, $SE(\hat{\beta_1})$ all have the usual interpretation

### 2.4 Heteroskedasticity and Homoskedasticity

If var $(u|X = x)$ is constant (the variance of $u$ does not depend on X), then $u$ is homoskedastic, otherwise it is heteroskedastic.

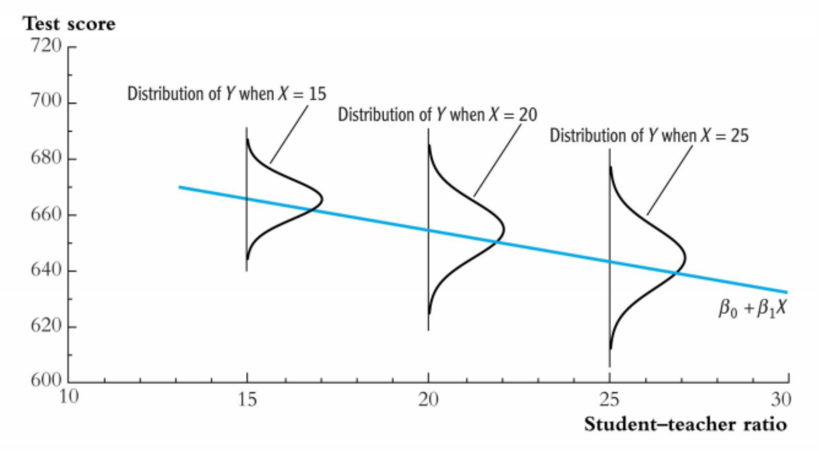


Figure 1: Example of Heteroskedasticity

If the error is homoskedastic and $u$ is distributed $N(0, \sigma^2)$, OLS estimators has the lowest variance among all linear consistent estimators.

var$(\hat{\beta}_1)$ simplifies to $\dfrac{\sigma_u^2}{n\sigma_X^2}$.

If the error is heteroskedastic and you assume that it is homoskedastic, the computed standard error is smaller, therefore is no robust. In STATA, always use robust regression, which treats the error as heteroskedastic.

# 3 Multiple Regression

### 3.1 Omitted Variable Bias

If an omitted factor "$Z$":

- is a determinant of $Y$, i.e. $Z$ is part of $u$
- correlated with the regressor $X$, i.e. Corr$(Z, X) \neq 0$

then $Z$ will result in omitted variable bias, i.e. the OLS estimator does not approach the true value

The omitted variable bias formula:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}$$

where $\rho_{Xu} = \text{Corr}(X, u)$

## 3.2 Multiple Regression Model

The case of two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ i = 1, \ldots, n$$

Meaning of the coefficients:

- $\beta_1 = \dfrac{\Delta Y}{\Delta X_1}$, holding $X_2$ constant
- $\beta_2 = \dfrac{\Delta Y}{\Delta X_2}$, holding $X_1$ constant
- $\beta_0 =$predicted value of $Y$ when $X_1 = X_2 = 0$

The two regressors OLS solves

$$\min_{\beta_0, \beta_1, \beta_2} \sum_{i=1}^{n} \left[Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})\right]^2$$

This can be generalized to $k$ regressors.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i,\ i = 1, \ldots, n$$

The derivations is

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

## 3.3 Measures of Fit for Multiple Regression

$Y_i = \hat{Y}_i + \hat{u}_i$

$\text{SER} = \sqrt{\dfrac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2}$

$\text{RMSE} = \sqrt{\dfrac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2}$

$R^2 = \dfrac{\text{ESS}}{\text{TSS}} = 1 - \dfrac{\text{SSR}}{\text{TSS}}$, where ESS=$\sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}})^2$, SSR=$\sum_{i=1}^{n} \hat{u}_i^2$, TSS=$\sum_{i=1}^{n} (Y_i - \bar{Y})^2$

$R^2$ will always increase when another regressor is added, unless its estimated coefficient is exactly zero. This can be fixed with:

Adjusted $R^2$: $\bar{R}^2 = 1 - \left(\dfrac{n-1}{n-k-1}\right) \dfrac{\text{SSR}}{\text{TSS}}$

### 3.4 The Least Squares Assumptions for Multiple Regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i,\ i = 1, \ldots, n$$

- The conditional distribution of $u$ given $X$ has mean zero, that is, $E(u|X_1 = x_1, \ldots, X_k = x_k) = 0$
- $(X_{1i}, \cdot, X_{ki}, Y_i)$ are i.i.d.
- Large outliers are rare, i.e. $Y$ have limited fourth moments: $E(X_{ki}^4) < \infty, E(Y_i^4) < \infty$
- There is no perfect multicollinearity

#### 3.4.1 Assumption #1

If an omitted variable:

- is a part of $u$
- is correlated with an included X

, then this condition fails and leads to omitted variable bias.

The solution is to include the omitted variable in the regression.

#### 3.4.2 Assumption #4

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

### 3.5 The Sampling Distribution of the OLS Estimator

- $E(\hat{\beta}_1) = \beta_1$, var$(\hat{\beta}_1)$ is inversely proportional to $n$
- The exact distribution of $\hat{\beta}_1$ is complicated, but for large $n$,
  - $\hat{\beta}_1$ is consistent: $\hat{\beta}_1 \xrightarrow{p} \beta_1$
  - $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately $N(0, 1)$ (CLT)

, the same goes for $\hat{\beta}_2, \ldots, \hat{\beta}_k$

### 3.6 Multicollinearity

Perfect multicollinearity must be removed otherwise the regression cannot be done.

Imperfect multicollinearity occurs when two or more regressors are highly correlated. It will result in large standard errors for one or more of the OLS coefficients.

## 4 Hypothesis Tests and Confidence Intervals in Multiple Regression

### 4.1 Single Coefficient

Hypotheses on $\beta_1$ can be tested using the usual t-statistic $t = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$, and confidence intervals are constructed as $\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)$. The same goes for $\beta_2, \ldots, \beta_k$

## 4.2 Tests for Joint Hypotheses

A joint hypothesis is: $H_0 : \beta_1 = 0$ and $\beta_2 = 0$ v.s. $H_1$ : either $\beta_1 \neq 0$ or $\beta_2 \neq 0$ or both.

The above example considered $q = 2$ coefficients, we call it 2 "restrictions". It can also be extended to any $q$, where $H_0$ is $\beta_1 = 0$ and . . . and $\beta_q = 0$

In this case, we should use the F-statistic instead of the t-statistic.

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

where $t_1$ and $t_2$ are one-variable t-statistics and $\hat{\rho}_{t_1,t_2}$ estimates the correlation between $t_1$ and $t_2$.

In large samples, the F-statistic is distributed as $\chi_q^2/q$ (or $F_{q,\infty}$), whose 5% critical value is:

| q | 5% critical value |
|---|---|
| 1 | 3.84 |
| 2 | 3.00 |
| 3 | 2.60 |
| 4 | 2.37 |
| 5 | 2.21 |

The corresponding p-value is the tail probability of the $F_{q,\infty}$ distribution beyond the F-statistic actually computed.

Simple formula for the homoskedastic-only F-statistic:

$$F = \frac{\left(R^2_{\text{unrestricted}} - R^2_{\text{restricted}}\right)/q}{\left(1 - R^2_{\text{unrestricted}}\right)/(n - k_{\text{unrestricted}} - 1)}$$

where:

- $R^2_{\text{restricted}}$ = the $R^2$ for the restricted regression
- $R^2_{\text{unrestricted}}$ = the $R^2$ for the unrestricted regression
- $q$ = the number of restrictions under the null
- $k_{\text{unrestricted}}$ = the number of regressors in the unrestricted regression

where the restricted regression is

$$Y_i = \beta_0 + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

which is removing the two regressor involved in the null hypothesis. The unrestricted regression is just the original regression.

The formula is equivalent to

$$F = \frac{(SSR_{\text{restricted}} - SSR_{\text{unrestricted}})/q}{SSR_{\text{unrestricted}}/(n - k_{\text{unrestricted}} - 1)}$$

## 4.3 Testing Single Restrictions on Multiple Coefficients

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

Consider the hypothesis $H_0 : \beta_1 = \beta_2$ v.s. $H_1 : \beta_1 \neq \beta_2$.

### 4.3.1 Method 1: Rearrange the regression

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\ &= \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + u_i \end{aligned}$$

Let

$$\begin{aligned} \gamma_1 &= \beta_1 - \beta_2 \\ W_i &= X_{1i} + X_{2i} \end{aligned}$$

, the original regression is now $Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$

Now the hypothesis is $H_0 : \gamma_1 = 0$ v.s. $H_1 : \gamma \neq 0$.

### 4.3.2 Method 2: Perform the test directly

Use STATA.

## 4.4 Confidence Sets for Multiple Coefficients

Let $F(\beta_{1,0}, \beta_{2,0})$ be the F-statistic testing the hypothesis that $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$, the 95% confidence set = $\{\beta_{1,0}, \beta_{2,0} : F(\beta_{1,0}, \beta_{2,0}) < 3.00\}$, where 3.00 is the 5% critical value of the $F_{2,\infty}$ distribution.

## 4.5 Model specification

- Specify a “base” or “benchmark” model
- Specify a range of plausible alternative models, which include additional variables
- Compare the results for $\beta_1$, $\beta_{\text{new-var}}$, etc., with judgement (instead of a mechanical recipe).

Do not just try to maximize $R^2$! A higher $R^2$ or $\bar{R}^2$ means that the regressors explain the variation in Y, but it does not mean that you have an unbiased estimator of a casual effect $\beta_1$, which is in fact our final purpose.

A table of regression results should at least include:

- estimated regression coefficients
- standard errors
- measures of fit
- number of observations (sample size)
- relevant F-statistics

# 5 Nonlinear Regression Functions

## 5.1 General Ideas

Model:

$$Y_i = f\left(X_{1i}, X_{2i}, \ldots, X_{ki}\right) + u_i,\ i = 1, \ldots, n$$

Assumptions:

- $E(u_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$, which implies that $f$ is the conditional expectation of Y given the X’s.
- $(X_{1i}, \ldots, X_{ki}, Y_i)$ are i.i.d.

- Big outliers are rare
- No perfect multicollinearity

## 5.2 Single Independent Variable

Two complementary approaches: polynomials in X and logarithmic transformations.

### 5.2.1 Ploynomials in X

Model:

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

Estimation, hypothesis testing, etc. proceeds as in the multiple regression model using OLS, by simply regarding $X_i^r$ as different variables. However, the coefficients have complicated interpretations.

Choice of degree $r$ should be determined by plotting the data, looking at t-tests and F-tests, etc.

### 5.2.2 Logarithmic functions of Y and/or X

Three specifications:

| Case | Regression Function |
|---|---|
| linear-log | $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$ |
| log-linear | $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$ |
| log-log | $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$ |

Because of $\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$:

For small $\Delta X$ in linear-log case, $\beta_1 \approx \frac{\Delta Y}{\Delta X / X}$. Therefore a 1% increase in X (multiply X by 1.01, regardless the current value of X) is associated with a $0.01\beta_1$ change in Y.

For small $\Delta X$ in log-linear case, $\beta_1 \approx \frac{\Delta Y / Y}{X}$. Therefore a change in X by one unit ($\Delta X = 1$) is associated with a $100\beta_1$% change in Y ($\Delta Y = \beta_1 Y$).

For small $\Delta X$ in log-log case, $\beta_1 \approx \frac{\Delta Y / Y}{\Delta X / X}$. Therefore a 1% change in X is associated with a $\beta_1$% change in Y.

**Important! Interpret the result of regression with the above-mentioned "percentage" way, instead of computing the logarithmic/exponential number!**

Hypotheses tests, confidence intervals, etc. are implemented as usual.

## 5.3 Interactions Between Independent Variables

Generally, $\frac{\Delta Y}{\Delta X_1}$ might depend on $X_2$.

### 5.3.1 Two Binary Variables

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

where $D_{1i}, D_{2i}$ are binary.

To allow the effect of changing $D_1$ to depend on $D_2$, change it to:

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

The interpretation of $\beta_3$ is the increment to the effect of $D_1$, when $D_2 = 1$. If $D_2 = 0$, $\beta_3$ will not effect the predictor.

### 5.3.2 Continuous and Binary Variables

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

where $D_i$ is binary and $X$ is continuous.

To allow the effect of changing $X$ to depend on $D$, change it to:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i \cdot X_i) + u_i$$

It can be derived that $\dfrac{\Delta Y}{\Delta X} = \beta_2 + \beta_3 D$. The interpretation of $\beta_3$ is the increment to the effect of $X$, when $D = 1$. If $D = 0$, $\beta_3$ will not effect the predictor.

### 5.3.3 Two Continuous Variables

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

To allow the effect of changing $X_1$ to depend on $X_2$, change it to:

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

It can be derived that $\dfrac{\Delta Y}{\Delta X} = \beta_1 + \beta_3 X_2$. The interpretation of $\beta_3$ is the increment to the effect of $X_1$ from a unit change in $X_2$.

## 5.4 Other Nonlinear Functions

Negative Exponential Growth:

$$Y_i = \beta_0 - \alpha e^{-\beta_1 X_i} + u_i$$

Let $\alpha = \beta_0 e^{\beta_2}$, it becomes

$$Y_i = \beta_0 \left[ 1 - e^{-\beta_1 (X_i - \beta_2)} \right] + u_i$$

The nonlinear least squares is

$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} \left\{ Y_i - \beta_0 \left[ 1 - e^{-\beta_1 (X_i - \beta_2)} \right] \right\}^2$$

### 5.5 Application of Nonlinear Effects in Studies

Question: Are there nonlinear effects of $X$ on $Y$?

1. Estimate linear and nonlinear functions of $X$, holding other regressors
2. See whether adding the nonlinear terms makes an "real-word" importance
3. Test whether the nonlinear terms are statistically significant

Question: Are there nonlinear interactions between $X_1$ and $X_2$?

- Estimate linear and nonlinear functions of $X_1$, interacted with $X_2$
- Note that in the nonlinear case, add interactions with all the terms $(X_1, X_1^2, X_1^3, \ldots)$

## 6 Assessing Studies Based on Multiple Regression

The framework:

- Internal validity: A study's statistical inferences about casual effects are valid for the population being studied
- External validity: A study's statistical inferences can be generalized to other populations and settings

We focus on internal validity, which has two components:

- The estimator of the causal effect should be unbiased and consistent
- Hypothesis tests should have the desired significance level, and the standard errors are computed correctly

Threats that might cause the OLS estimator biased:

- Omitted variable bias
- Mis-specification of the functional form
- Measurement error
- Sample selection bias
- Simultaneous causality bias

All of these imply that $E(u_i|X_{1i}, \ldots, X_{ki}) \neq 0$.

### 6.1 Omitted Variable Bias

Omitted variable bias arises if an omitted variable is both:

- a determinant of Y
- correlated with at least one included regressor

Potential solutions:

- If the variable can be measured, include it as an additional regressor

- Else use instrumental variables regression (introduced later)
- Run a randomized controlled experiment

## 6.2 Mis-specification of the functional form

This happens when the true population regression function is nonlinear, but the estimated regression is linear.

Potential solutions:

- Continuous dependent variable: use the appropriate nonlinear specification
- Discrete dependent variable: use probit or logit analysis (introduced later)

## 6.3 Errors-in-variables Bias

This happens when the data is measured with error. The error term is typically correlated with the regressor, so that $\hat{\beta}_1$ is biased.

Potential solutions:

- Obtain better data
- Develop a specific model of the measurement error process (not pursued)
- Instrumental variables regression (introduced later)

## 6.4 Sample selection bias

This happens when the sample selection process

- influences the availability of data
- process is related to the dependent variable

For example, those with lower education level are less likely to present on the job market, therefore they do not have income at all, while the sample selection process

Potential solutions:

- Collect the sample in a way that avoids sample selection
- Randomized controlled experiment
- Estimate that model using Heckman two-step estimation

## 6.5 Simultaneous Causality Bias

This happens when $Y$ causes $X$.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ X_i &= \gamma_0 + \gamma_1 Y_i + v_i \end{aligned}$$

Large $u_i$ means large $Y_i$, which implies large $X_i$, therefore corr$(X_i, u_i) \neq 0$

Potential solutions:

- Randomized controlled experiment
- Develop and estimate a complete model of both directions of causality (extremely difficult)

- Use instrumental variables regression, which can estimate the effect of X on Y, ignoring effect of Y on X

# 7 Instrumental Variable Regression

Instrumental variables regression eliminate bias when $E(u|X) \neq 0$, using an instrumental variable, $Z$.

Terminology:

- Endogenous: a variable that is correlated with $u$
- Exogenous: a variable that is uncorrelated with $u$

For an instrumental variable $Z$ to be valid, it must satisfy two conditions:

- Relevance: corr$(Z_i, X_i) \neq 0$
- Exogeneity: corr$(Z_i, u_i) = 0$

## 7.1 IV Regression with One Regressor and One Instrument

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

If Z is relevant and exogenous, we can use the two stage least squares estimator (TSLS).

Stage 1: regress $X$ on $Z$ using OLS

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

Compute $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i,\ i = 1, \ldots, n$. As $Z$ is uncorrelated with $u$, $\hat{X}$ is uncorrelated with $u$ and the least square assumptions hold.

Stage 2: regress $Y$ on $\hat{X}_i$ using OLS.

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$$

Then $\beta_0$ and $\beta_1$ can be estimated by OLS, whose result is called the Two Stage Least Squares (TSLS) estimator, $\hat{\beta}^{\text{TSLS}}$

It can be shown that $\hat{\beta}_1^{\text{TSLS}} = \dfrac{s_{YZ}}{s_{XZ}}$, where $s$ means the sample covariance. It is also a consistent estimator.

## 7.2 Inference using TSLS

It can be derived that $\hat{\beta}_1^{\text{TSLS}}$ is approximately distributed $N(\beta_1, \sigma^2_{\hat{\beta}_1^{\text{TSLS}}})$, where $\sigma^2_{\hat{\beta}_1^{\text{TSLS}}} = \dfrac{1}{n} \dfrac{\text{var}\left[(Z_i - \mu_Z)u_i\right]}{\left[\text{cov}(Z_i, X_i)\right]^2}$

Therefore the statistical inference proceeds in the usual way ($\mu \pm 1.96\sigma$), based on large samples. Note that the standard error here is not the OLS standard error in the second stage. Compute the specific TSLS standard error instead.

## 7.3 The General IV Regression Model

The model should be extended to:

- Multiple endogenous regressors, $X_1, \ldots, X_k$
- Multiple exogenous regressors, $W_1, \ldots, W_r$
- Multiple instrumental variables, $Z_1, \ldots, Z_m$. More instrumental variables can produce larger $R^2$ in the first stage, therefore you have more variation in $\hat{X}$

For IV regression to be possible, there must be as many as instruments as endogenous variables. The coefficients $\beta$ are said to be:

- exactly identified, if $m = k$
- over-identified, if $m > k$. You can test whether the instruments are valid.
- under-identified, if $m < k$. It is needed to get more instruments.

The model:

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

- $Y$ is the dependent variable
- $X$ are the endogenous regressors
- $W$ are the included exogenous regressors
- $\beta$ are the unknown regression coefficients
- $Z$ are the $m$ instrumental variables

### 7.3.1 TSLS with a Single Endogenous Regressor

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

with $m$ instruments: $Z_1, \ldots, Z_m$.

Stages:

1. Regress $X_1$ on all exogenous regressors, $W_1, \ldots, W_r, Z_1, \ldots, Z_m$ by OLS. Compute predicted $\hat{X}_1$
2. Regress $Y$ on $\hat{X}_1, W_1, \ldots, W_r$ by OLS

The resulting coefficients from the second stage are the TSLS estimators.

## 7.4 IV Regression Assumptions

- $E(u_i|W_{1i}, \ldots, W_{ri}) = 0$, that is, "the exogenous regressors are exogenous"
- All variables are i.i.d.
- All variables have nonzero, finite fourth moments
- The instruments are valid

Under such assumptions, the TSLS and its t-statistic are normally distributed.

## 7.5 Checking Instrument Validity

### 7.5.1 Checking Relevance

We focus on a single included endogenous regressor:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

First stage regression:

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

The instruments are relevant if at least one of $\pi$ is nonzero. They are weak if all $\pi$ are either zero or nearly zero. Use first-stage F-statistic to test the hypothesis that $\pi_1, \ldots, \pi_m$ are all zero. If the first-stage F-statistic is less than 10 (NOT the typical critical value for F-statistic), then the set of instruments is weak.

### 7.5.2 Checking Exogeneity

Could be done only when over-identifying.

Suppose that $m > k$ in

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

Steps:

1. Estimate the equation using all $m$ instruments to obtain the TSLS's. Compute the predicted $\hat{Y}_i$ using the **actual** $X$'s.
2. Compute the residuals $\hat{u}_i = Y_i - \hat{Y}_i$
3. Regress $\hat{u}_i$ against all $Z$ and $W$
4. Compute the F-statistic testing the hypotheses that the coefficients on $Z$ are all zero
5. The J-statistic is $J = mF$

Under the null hypotheses that all the instruments are exogenous, $J$ has a $\chi^2$ distribution with $m - k$ degrees of freedom and it will be small. If at least one of the instrument is endogenous, $J$ will be larger than the $\chi^2$ critical value. That is, small J is preferred, which interprets that your IV is valid.

# 8 Regression with a Binary Dependent Variable

Regression when $Y$ is binary.

## 8.1 The Linear Probability Model

$$Y_i = \beta_1 X_i + u_i$$

The predicted value $\hat{Y}$ is the predicted probability that $Y_i = 1$, given $X$. $\beta_1$ is the change in probability that $Y - 1$ for a given $\Delta x$.

All interpretations, hypotheses tests, confidence intervals are the same as for multiple regression.

The disadvantage is that the predicted probability can be not in $[0, 1]$

## 8.2 Probit and Logit Regression

Probit and logit regression can ensure that the predicted $\Pr(Y = 1|X) \in [0, 1]$.

### 8.2.1 Probit Regression

Model:

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

where $\Phi$ is the cumulative normal distribution function. $z = \beta_0 + \beta_1 X$ is the z-value of the probit model.

Example: For $\beta_0 = -2, \beta_1 = 3, X = 0.4$, $\Pr(Y = 1|X = 0.4) = \Phi(-2 + 3 \times 0.4) = \Phi(-0.8)$, which is the area under the PDF of the normal distribution to the left of $-0.8$, i.e. $0.212$.

Model with multiple regressors:

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

$\beta_1$ is the effect on the z-score of a unit change in $X_1$, holding $X_2$ constant.

### 8.2.2 Logit Regression

Model:

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

where $F$ is the cumulative logistic distribution function:

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

Practically, logit and probit models are very similar.

## 8.3 Estimation and Inference in Probit and Logit Models

Calculus does not give explicit solution for OLS estimator for probit models, instead, we should use a maximum likelihood estimator.

The likelihood function is the conditional density of $Y$ given $X$'s treated as a function of the unknown parameters $\beta$'s. The MLE (maximum likelihood estimator) is the value of $\beta$'s that maximize the likelihood function.

In large samples, the MLE is consistent, normally distributed and efficient.

### 8.3.1 Introduction to MLE – Linear Model as an Example

The MLE for the simple regression model $Y_i = \beta_0 + \beta_1 X_1 + u_i$ can be derived as follows.

As we assume i.i.d. normal errors, the distribution of the regression residual $u$ is $N(0, \sigma)$, where $\sigma$ represents the standard error of $u$.

Therefore, the conditional pdf of $Y$ for each $x$, $p(y|X = x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\beta_0-\beta_1 x)^2}{2\sigma^2}}$

Given the data $X_i, Y_i$, the overall probability is

$$\prod_{i=1}^{n} p(Y_i|X_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y_i-\beta_0-\beta_1 X_i)^2}{2\sigma^2}}$$

The maximum likelihood estimator is the estimator $\hat{\beta_0}$ and $\hat{\beta_1}$ that can maximize this function. Take logarithmic on this formula,

$$\begin{aligned} f(\beta_0, \beta_1) &= \log \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y_i-\beta_0-\beta_1 X_i)^2}{2\sigma^2}} \\ &= -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 \end{aligned}$$

Take partial derivative on $f(\beta_0, \beta_1)$,

$$\begin{aligned} \frac{\partial}{\partial\beta_0}(\beta_0, \beta_1) &= \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i) \\ \frac{\partial}{\partial\beta_1}(\beta_0, \beta_1) &= \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)X_i \end{aligned}$$

Let the two partial derivative be 0, it can be derived that

$$\begin{aligned} \hat{\beta_0} &= \bar{Y} - \hat{\beta_1}\bar{X} \\ \hat{\beta_1} &= \frac{\sum\limits_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2} \end{aligned}$$

### 8.3.2 The Probit Likelihood with One X

The density of $Y_1$ given $X_1$ is:

$$\begin{aligned} \Pr(Y_1 = 1|X_1) &= \Phi(\beta_0 + \beta_1 X_1) \\ \Pr(Y_1 = 0|X_1) &= 1 - \Phi(\beta_0 + \beta_1 X_1) \end{aligned}$$

Therefore,

$$\Pr(Y_1 = y_1|X_1) = \Phi(\beta_0 + \beta_1 X_1)^{y_1}\left[1 - \Phi(\beta_0 + \beta_1 X_1)\right]^{1-y_1}$$

The probit likelihood function, which is the joint density of conditional $Y_1, \cdots, Y_n$, is:

$$f(\beta_0, \beta_1; Y_1, \cdots, Y_n|X_1, \cdots, X_n) = \Pr(Y_1|X_1) \times \Pr(Y_2|X_2) \times \cdots \times \Pr(Y_n|X_n)$$

There is not explicit solution for this.

Testing, confidence intervals proceeds as usual:

- Hypotheses testing via t-statistic

- Confidence interval as ± 1.96 SE
- Joint hypotheses with F-statistic

For no X, see slide “7_Binary Regression” Page 29. For multiple X’s, see textbook Page 419.

#### 8.3.3 The Logit Likelihood with One X

The only difference is that the functional form $\Phi$ is replaced by the cumulative logistic function.

### 8.4 Measures of Fit for Logit and Probit

The $R^2$ don’t make much sense here. Two other specialized measures are used instead.

1. The fraction correctly predicted.
    - Advantage: easy to understand
    - Disadvantage: does not reflect the quality of the prediction (51% is the same as 99%)
2. The pseudo-$R^2$, $1-\frac{\ln L(f^{\max}_{\text{probit}})}{\ln L(f^{\max}_{\text{Bernoullit}})}$, which measures the improvement in the value of the log likelihood, relative to having no X’s.
   $f^{\max}_{\text{probit}}$ is the value of the maximized probit likelihood function (see 8.3.2). $f^{\max}_{\text{Bernoulli}}$ is the value of the maximized probit likelihood function **of the model excluding all X’s**.

## 9 Regression with Panel Data

A panel dataset contains observations on multiple entities, where each entity is observed at two or more points in time.

Notation for panel data is a double subscript. $i$ is the entity and $n$ is the number of entity. $t$ is the time period and $T$ is the number of time periods.

With one regressor, the data is: $(X_{it}, Y_{it})$, $i = 1, \ldots, n$; $t = 1, \ldots, T$. With $k$ regressors, the data is: $(X_{1it}, X_{2it}, \ldots, X_{kit}, Y_{it})$, $i = 1, \ldots, n$; $t = 1, \ldots, T$

Another term for panel data is longitudinal data. If the panel contains no missing observations, it’s called a balanced panel.

### 9.1 Panel Data with Two Time Periods

Consider the panel data model

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$$

where $Z_i$ is a factor that does not change over time, and is not observed. Its omission could cause omitted variable bias in the previous setup, but can be eliminated using $T = 2$ years.

Subtract the equation of $t = t_2$ from that of $t = t_1$ eliminates the item $\beta_2 Z_i$, resulting in

$$Y_{it_1} - Y_{it_2} = \beta_1(X_{it_1} - X_{it_2}) + (u_{it_1} - u_{it_2})$$

Therefore, $\beta_1$ can be correctly estimated without the omitted variable bias.

## 9.2 Entity Fixed Effects Regression - the Derivation

If you have more than 2 time periods, it can be rewritten in the fixed effects form.

Suppose we have $n = 3$ and the original model is

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_i,\ i = 1, \ldots, n; t = 1, \ldots, T$$

Population regression for $i = 1$ is:

$$\begin{aligned} Y_{1t} &= \beta_0 + \beta_1 X_{1t} + \beta_2 Z_1 + u_{1t} \\ &= (\beta_0 + \beta_2 Z_1) + \beta_1 X_{1t} + u_{1t} \end{aligned}$$

Let $\alpha_1 = \beta_0 + \beta_2 Z_1$, which does not change over time, the model is

$$Y_{1t} = \alpha_1 + \beta_1 X_{1t} + u_{1t}$$

The intercept is unique for state 1, but the slope is the same in all the states.

For i=2, The model is $Y_{2t} = \alpha_2 + \beta_1 X_{2t} + u_{2t}$ For i=3, The model is $Y_{3t} = \alpha_3 + \beta_1 X_{3t} + u_{3t}$

In the generalized form, that is

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it},\ i = 1, \ldots, n;\ t = 1, \ldots, T$$

For different $i$, the slope is the same but the intercept is different. The shifts in the intercept can be represented using binary regressors.

$$Y_{it} = \beta_0 + \gamma_1 D_i^1 + \gamma_2 D_i^2 + \beta_1 X_{it} + u_{it}$$

where $D_i^1 = 1$ if $i = 1$, 0 otherwise; $D_i^2 = 1$ if $i = 2$, 0 otherwise. Note that there is no $D_i^3$ because its intercept is just $\beta_0$. This is the “n-1 binary regressor form”.

## 9.3 Estimate the Entity Fixed Effects Regression - the Summary

There are three methods:

- “Changes” specification without an intercept (only works for $T = 2$), see Section 9.1
- “n-1 binary regressors” OLS regression
- “Entity-demeaned” OLS regression

### 9.3.1 “n-1 Binary Regressors” OLS Regression

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D_i^2 + \ldots + \gamma_n D_i^n + u_{it}$$

where $D_i^2 = 1$ for $i = 2$, 0 otherwise, etc.

Estimate with OLS and inference as usual.

### 9.3.2 “Entity-demeaned” OLS Regression

The fixed effects regression model is

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$$

Average this over the T time periods:

$$\bar{Y_i} = \frac{1}{T}\sum_{t=1}^{T} Y_{it} = \alpha_i + \beta_1 \frac{1}{T}\sum_{t=1}^{T} X_{it} + \frac{1}{T}\sum_{t=1}^{T} u_{it}$$

It can be converted into such form:

$$\tilde{Y_{it}} = \beta_1 \tilde{X_{it}} + \tilde{u}_{it}$$

where $\tilde{Y_{it}} = Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ and $\tilde{X_{it}} = X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it}$

Estimate this formula using OLS, and do inference/testing as usual.

## 9.4 Regression with Time Fixed Effects

An omitted variable might vary over time but not across states, which can be denoted by $S_t$, the resulting regression model is:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

If there is only time-dependent omitted variable ($S$), but not state-dependent omitted variable ($Z$), it’s called time fixed effects only. The model is:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_3 S_t + u_{it}$$

It can be also converted into two forms:

- “T-1 binary regressor” formulation
- “Year-demeaned” formulation

### 9.4.1 “T-1 binary regressor” OLS estimation

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B_t^2 + \ldots + \delta_T B_t^T + u_{it}$$

where $B_t^2$ = 1 when $t = 2$, 0 otherwise, etc.

Estimate with OLS as usual.

### 9.4.2 "Year-demeaned" OLS estimation

The original model is

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}$$

It can be converted to

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it}$$

by letting $\tilde{Y}_{it} = Y_{it} - \frac{1}{N}\sum_{i=1}^{N} Y_{it}$ and $\tilde{X}_{it} = X_{it} - \frac{1}{N}\sum_{i=1}^{N} X_{it}$.

Estimate with OLS as usual.

## 9.5 Regression with Both Entity and Time Fixed Effects

The original model is:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

The techniques for entity fixed effects and time fixed effects are decoupled so they can be combined. You can apply the "demeaned" technique for both effects.

Or apply the "X-1 binary regressor":

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D_i^2 + \ldots + \gamma_n D_i^n + \delta_2 B_t^2 + \ldots + \delta_T B_t^T + u_{it}$$

where $\beta_0 = \alpha_1 + \lambda_1, \gamma_i = \alpha_i - \alpha_1, \delta_t = \lambda_t - \lambda_1$.

It is also plausible to use entity demeaning & T-1 time indicators, or vice versa.

## 9.6 Assumptions for Fixed Effects Regression

For a single X:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it},\ i = 1, \ldots, n;\ t = 1, \ldots, T$$

Assumptions:

1. $E(u_{it}|X_{i1}, \ldots, X_{iT}, \alpha_i) = 0$
2. $(X_{i1}, \ldots, X_{iT}, Y_{i1}, \ldots, Y_{iT})$ are i.i.d.
3. $(X_{it}, Y_{it})$ have finite fourth moments
4. There is no perfect multicollinearity (for multiple $X$'s)
5. corr$(u_{it}, u_{is}|X_{it}, X_{is}, \alpha_i) = 0$ for $t \neq s$

Assumption 5 is new while others are just extensions of the normal least square assumptions.

### 9.6.1 Assumption #1

$u_{it}$ has mean zero, given the state fixed effect and the entire history of the X's for that state. This means there are no omitted lagged effects imposed by X, and there is no feedback from $u$ to future X.

### 9.6.2 Assumption #2

Entities should be sampled randomly, then data for the entities are collected over time. Note that this does not require observations to be i.i.d. over time

### 9.6.3 Assumption #5

The error terms are uncorrelated over time within a state.

It might not hold sometimes. The analogy is hetroskedasticity. In that case, we still use the OLS estimation and the estimator will still be unbiased, but the standard error should be computed with "Heteroskedasticity and autocorrelation-consistent standard errors".

## 9.7 Standard Errors for Fixed Effects Regression

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it}$$

OLS fixed effects estimator:

$$\beta_1 = \frac{\sum\limits_{i=1}^{n}\sum\limits_{t=1}^{T} \tilde{X}_{it}\tilde{Y}_{it}}{\sum\limits_{i=1}^{n}\sum\limits_{t=1}^{T} \tilde{X}_{it}^2}$$

The standard error is:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{nT}\frac{\hat{\sigma}_\eta^2}{\hat{Q}_{\tilde{X}}^4}}$$

where $\hat{Q}_{\tilde{X}}^2 = \frac{1}{nT}\sum\limits_{i=1}^{n}\sum\limits_{t=1}^{T} \tilde{X}_{it}^2$, and $\sigma_\eta^2 = \text{var}\left(\sqrt{\frac{1}{T}}\sum\limits_{t=1}^{T} \tilde{v}_{it}\right)$, where $\tilde{v}_{it} = \tilde{X}_{it}u_{it}$

- If $u_{it}, u_{is}$ are uncorrelated, $\sigma_\eta^2 = \text{var}(\tilde{v}_{it})$.
- if $u_{it}, u_{is}$ are correlated, we can only estimate $\sigma_\eta^2$, which is called clustered standard error.
  $\hat{\sigma}_{\eta,\text{clustered}}^2 = \frac{1}{n}\sum\limits_{i=1}^{n}\left(\sum\limits_{t=1}^{T} \hat{\tilde{v}}_i t\right)^2$, where $\hat{\tilde{v}}_{it} = \tilde{X}_{it}\hat{u}_{it}$

## 9.8 Summary of Panel Data Regression

Advantages:

- You can control for unobserved variables that
  - vary across states but not over time
  - vary over time but not across states
- After applying the "X-1 binary regressors" or "demeaned" trick, the estimation is similar to multiple regression

Limitations:

- Need variation in $X$ over time within states
- Time lag effects can be important
- You should use hetroskedasticity-autocorrelation-considered (clustered) standard errors

# 10 Time Series Regression and Forecasting

## 10.1 Time Series Data and Serial Correlation

### 10.1.1 Notations

- $Y_t$: value of $Y$ in period $t$
- Dataset: $Y_1, \ldots, Y_t, T$ observations on the time series of random variable $Y$
- We consider only consecutive, evenly-spaced observations
- The $j^{\text{th}}$lag is $Y_{t-j}$, the first lag is $Y_{t-1}$
- The first difference is $\Delta Y_t = Y_t - Y_{t-1}$
- The first difference of the logarithm of $Y_t$ is $\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$
- The percentage change of $Y_t$ between periods t-1 and t is approximately $100\Delta \ln(Y_t)$

### 10.1.2 Autocorrlation

Autocorrelation is the correlation of a series with its own lagged values.

- The first autocorrelation of $Y_t$ is $\text{corr}(Y_t, Y_{t-1})$
- The first autocovariance of $Y_t$ is $\text{cov}(Y_t, Y_{t-1})$

$$\rho_1 = \text{corr}(Y_t, Y_{t-1}) = \frac{\text{cov}(Y_t, Y_{t-1})}{\sqrt{\text{var}(Y_t)\text{var}(Y_{t-1})}}$$

The $j^{\text{th}}$autocorrelation is

$$\rho_j = \text{corr}(Y_t, Y_{t-j}) = \frac{\text{cov}(Y_t, Y_{t-j})}{\sqrt{\text{var}(Y_t)\text{var}(Y_{t-j})}}$$

The $j^{\text{th}}$sample autocorrelation is an estimate of the $j^{\text{th}}$population autocorrelation.

$$\hat{\rho}_j = \frac{\text{cov}(Y_t, Y_{t-j})}{\text{var}(Y_t)}$$

where

$$\text{cov}(Y_t, Y_{t-j}) = \frac{1}{T}\sum_{t=j+1}^{T}(Y_t - \bar{Y}_{j+1,T})(Y_{t-j} - \bar{Y}_{1,T-j})$$

where $\bar{Y}_{j+1,T}$ is the sample average of $Y_t$ computed over observations $t = j+1, \ldots, T$.

Note:

- The summation is over $t = j+1$ to $T$
- The divisor is $T$, not $T - j$

### 10.1.3 Stationarity

A time series $Y_t$ is stationary if the join probability distribution $Y_{s+1}, \ldots, Y_{s+T}$ does not depend on $s$. It means that the distribution does not change over time, so that the history is relevant, and the future resembles the past.

## 10.2 Autoregressions

An autoregression is a regression model in which $Y_t$ is regressed against its own lagged values.

The number of lags used as regressors is called the order of the auto regression. In a $j^{\text{th}}$order autoregression, $Y_t$ is regressed against $Y_{t-1}, \ldots, Y_{t-j}$.

### 10.2.1 The First Order Autoregressive (AR(1)) Model

The AR(1) model is

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

Here,

- $\beta$'s do NOT have casual interpretations.
- $\beta$'s can be estimated by OLS
- If $\beta_1 = 0$, $Y_{t-1}$ is not useful for forecasting $Y_t$
- Testing $\beta_1 = 0$ against $\beta_1 \neq 0$ provides a test of the hypotheses that $Y_{t-1}$ is not useful for forecasting $Y_t$

**Forecast: Terminology and Notation**

- Prediction: "in sample" values, the usual definition
- Forecast: "out-of-sample" values, in the future

Notation:

- $Y_{T+1|T}$ = forecast of $Y_{T+1}$ based on $Y_T, Y_{T-1}, \ldots$, using the population (true known) coefficients
- $\hat{Y}_{T+1|T}$ = forecast of $Y_{T+1}$ based on $Y_T, Y_{T-1}, \ldots$, using the estimated coefficients, which are estimated using data through period T

For AR(1):

- $Y_{T+1|T} = \beta_0 + \beta_1 Y_T$
- $\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T$, where the $\hat{\beta}$'s are estimated

**Forecast Errors** The one-period ahead forecast error is $Y_{T+1} - \hat{Y}_{T+1|T}$

Difference between forecast error and residual is "in-sample" or "out-of-sample". The value of $Y_{T+1}$ is not used in the estimation of the regression coefficients.

### 10.2.2 AR(p) Model

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + u_t$$

The coefficients do NOT have a casual interpretation neither. To test that $Y_{t-2}, \ldots, Y_{t-p}$ do not further help forecast $Y_t$, beyond $Y_{t-1}$, use an F-test. This can be used to determine the proper lag order $p$.

### 10.2.3 Additional Predictors and the Autogressive Distributed Lag (ADL) Model

Besides past values of Y, other variables (X) can be added as well.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \delta_1 X_{t-1} + \ldots + \delta_r X_{t-r} + u_t$$

This is an autoregressive distributed lag model with p lags of Y and r lags of X, ADL(p, r).

**Granger Causality Test** The Granger causality statistic is the F-statistic that tests the hypothesis that the coefficients of all X's are zero, beyond lagged values of Y in the ADL model.

The degree of freedom is the number of X's, i.e. the computed F-statistic should be compared to the critical value of $F_{r,\infty}$.

Reject the null hypotheses (some coefficient is not zero) if the computed F-statistic is larger than the critical value. Accept the null hypotheses (all coefficients are zero) if the computed F-statistic is smaller than the critical value.

### 10.2.4 Forecast Uncertainty and Forecast Intervals

For the forecast

$$\hat{Y}_{T+1}|T = \hat{\beta}_0 + \hat{\beta}_1 Y_T + \hat{\beta}_2 X_T$$

The forecast error is

$$Y_{T+1} - \hat{Y}_{T+1|T} = u_{T+1} - \left[(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)Y_T + (\hat{\beta}_2 - \beta_2)X_T\right]$$

The mean squared forecast error (MSFE) is

$$E(Y_{T+1} - \hat{Y}_{T+1|T})^2 = E(u_{T+1})^2 + E\left[(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)Y_T + (\hat{\beta}_2 - \beta_2)X_T\right]^2$$

The root mean squared forecast error (RMSFE) is the square root of the MSFE.

$$\text{RMSFE} = \sqrt{E\left[(Y_{T+1} - \hat{Y}_{T+1|T})^2\right]}$$

The RMSFE is a measure of the spread of the forecast error distribution. Three ways to estimate the RMSFE:

- Use the approximation RMSFE $\approx$ SER, when the sample size is large[1]
- Use the forecast history for $t = t_1, \ldots, T$, then estimate by

$$\text{MSFE} = \frac{1}{T - t_1 + 1} \sum_{t=t_1-1}^{T-1} \left(Y_{t+1} - \hat{Y}_{t+1|t}\right)^2$$

- Use a simulated forecast history, then use method 2. It's called pseudo out-of-sample forecasts.

The RMSFE can be used to construct the forecast intervals.

$$\hat{Y}_{T|T-1} \pm 1.96 \times \text{RMSFE}$$

Note that the forecast interval is not a confidence interval.

[1] Because when the sample size is large, MSFE $\approx$ var$(u_{T+1})$.

## 10.3 Lag Length Selection Using Information Criteria

How to choose the number of lags $p$ in an AR($p$)? Use the Information Criteria (IC)

### 10.3.1 Bayes Information Criterion (BIC)

$$\mathrm{BIC}(p) = \ln\left(\frac{\mathrm{SSR}(p)}{T}\right) + (p+1)\frac{\ln T}{T}$$

- First term: always decrease for larger $p$
- Second term: always increase for larger $p$. It is a penalty for using more parameters which will increase the forecast variance.

Minimizing BIC trades off bias and variance. $p$ chosen by BIC is consistent.

### 10.3.2 Akaike Information Criterion (AIC)

$$\mathrm{AIC}(p) = \ln\left(\frac{\mathrm{SSR}(p)}{T}\right) + (p+1)\frac{2}{T}$$

Minimizing AIC gives a choice of $p$ as well, the penalty term is smaller so it estimates more lags.

$p$ estimated by AIC is not consistent, however it might be desirable if longer lags might be important.

### 10.3.3 BIC for Multivariate (ADL) Models

Let $K$ = the total numbers of coefficients in the model.

$$\mathrm{BIC}(K) = \ln\left(\frac{\mathrm{SSR}(K)}{T}\right) + K\frac{\ln T}{T}$$

# 11 Nonstationarity: Trends

## 11.1 Concept Introduction

### 11.1.1 Definition of Trend

A trend is a long-term movement or tendency in the data.

A deterministic trend is a nonrandom function of time, e.g. $y_t = t^2$. A stochastic trend is random over time, e.g. a random walk.

### 11.1.2 The Random Walk

$$Y_t = Y_{t-1} + u_t$$

where $u_t$ is serially uncorrelated.

- $Y_{T+h|T} = Y_T$, the best prediction [2] of Y in the future is the value of Y today.
- var$(Y_{T+h} - Y_{T+h|T}) = h\sigma_u^2$, the more distant your forecast is, the greater the forecast uncertainty is.

[2] The “hat” is not used as the notation of prediction here, see 10.2.1

A variation, the random walk with drift model is:

$$Y_t = \beta_0 + Y_{t-1} + u_t$$

where $u_t$ is serially uncorrelated.

If $\beta_0 \neq 0$, $Y_t$ follows a random walk around a linear trend.

- $Y_{T_h|T} = \beta_0 h + Y_T$

**If $Y_t$ has a random walk trend, then $\Delta Y_t$ is stationary and regression should be taken using $\Delta Y_t$ instead of $Y_t$.**

### 11.1.3 Unit Autoregressive Roots

**AR(1)**

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

- The special case of $\beta_1 = 1$ is called a unit root
- In the case of unit root, the AR(1) model is equivalently $\Delta Y_t = \beta_0 + u_t$

**AR(2)**

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + u_t$$

It can be rearranged as

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + u_t$$

where $\delta = \beta_1 + \beta_2 - 1$ and $\gamma_1 = -\beta_2$

- The special case of $\beta_1 + \beta_2 = 1$ is called a unit root
- In the case of unit root, the AR(2) model is equivalently $\Delta T_t = \beta_0 + \gamma_1 \Delta Y_{t-1} + u_t$

**AR(p)**

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + u_t$$

This regression can be rearranged as

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \ldots + \gamma_{p-1} \Delta Y_{t-p+1} + u_t$$

where

$$\begin{aligned} \delta &= \beta_1 + \beta_2 + \ldots + \beta_p - 1 \\ \gamma_1 &= -(\beta_2 + \ldots + \beta_p) \\ \gamma_2 &= -(\beta_3 + \ldots + \beta_p) \\ &\ldots \\ \gamma_{p-1} &= -\beta_p \end{aligned}$$

- The special case of $\beta_1 + \beta_2 + \ldots + \beta_p = 1$ is called a unit root
- In the case of unit root, the AR(p) model is equivalently $\Delta T_t = \beta_0 + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \ldots + \gamma_{p-1} \Delta Y_{t-p+1} + u_t$

## 11.2 Problems Caused by Trends

When there are stochastic trends:

- AR coefficients can be badly biased towards zero.
- Some $t$-statistics do not have a standard normal distribution.
- If $Y$ and $X$ both have random walk trends, they can look related and get a significant regression coefficient even if they are not.

## 11.3 Trend Detection

You can plot the data, or use the Dickey-Fuller test for a unit root.

### 11.3.1 Dickey-Fuller Test

In the test introduced below, $H_0$ means the unit root. Rejecting $H_0$ means that it is stationary.

**AR(1)** A hypotheses test $H_0 : \delta = 0$ v.s. $H_1 : \delta < 0$ in[3]

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + u_t$$

or[4]

$$\Delta Y_t = \beta_0 + \mu_t + \delta Y_{t-1} + u_t$$

Compute the $t$-statistic testing $\delta = 0$ and compare with the Dickey-Fuller critical value table of the specification you choose. Reject $H_0$ if the calculated DF $t$-statistic is less than the specified value in the table.

**AR(p)** A hypotheses test $H_0 : \delta = 0$ v.s. $H_1 : \delta < 0$ in the intercept only specification.

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \ldots + \gamma_{p-1} \Delta Y_{t-p+1} + u_t$$

Estimate $\delta$ and compare the computed $t$-statistic testing $\delta = 0$ with the DF critical value in the table.

The intercept & time trend specification is:

$$\Delta Y_t = \beta_0 + \mu_t + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \ldots + \gamma_{p-1} \Delta Y_{t-p+1} + u_t$$

## 11.4 Solution for the Trend

If $Y_t$ has the random walk stochastic trend (has a unit root), the solution is to use $\Delta Y_t$ instead of $Y_t$ in the AR specification.

[3] The specification is called intercept only, or Y is stationary around a constant.

[4] The specification is called intercept & time trend, or Y is stationary around a deterministic linear time trend