Elements of Econometrics.
Lecture 27.
Revision 1

FCS, 2022-2023

Types of Data and Regression Models

Data: cross-sections, time series, panel data.

Model A: cross-sectional data with **nonstochastic regressors**. Their values in the observations are fixed and do not have random components.

Model B: cross-sectional data with **stochastic regressors. T**he regressors' values are drawn randomly and independently from defined populations.

Model C: time series data. The regressors' values may exhibit persistency over time

Panel Data Models. The time series relationships for different units are combined in the same model.

Regression Model A

Cross-sectional data with nonstochastic regressors (strong restriction! Needed for analytical simplicity), simple or multiple regression.

A.1 The model is linear in parameters and correctly specified.

$$Y = \beta_1 + \beta_2 X + u$$

A.2 There is some variation in the regressor in the sample and no exact linear relationship between regressors in the sample.

A.3 *(G-M 1)* The disturbance term has zero expected value in each observation: $E(u_i) = 0$ **for all i (automatically satisfied if intercept is included in regression)**

A.4 *(G-M 2)* The disturbance term is homoscedastic

for all i $\sigma_{u_i}^2 = \sigma_u^2$

A.5 *(G-M 3)* The values of the disturbance term have independent distributions (u_i and u_j are independent for all $j \neq i$)

$$\sigma_{u_i u_j} = E[(u_i - \mu_u)(u_j - \mu_u)] = E(u_i u_j)$$

$$= E(u_i)E(u_j) = 0$$

A.6 The disturbance term has a normal distribution

(G-M 4) Disturbance term and regressors are independent (satisfied automatically in Model A)

UNDER MODEL A ASSUMPTIONS OLS GIVES BLUE ESTIMATES

DERIVING SIMPLE LINEAR REGRESSION COEFFICIENTS

$$RSS = \hat{u}_1^2 + \ldots + \hat{u}_n^2 = (Y_1 - b_1 - b_2 X_1)^2 + \ldots + (Y_n - b_1 - b_2 X_n)^2$$

$$= \sum Y_i^2 + nb_1^2 + b_2^2 \sum X_i^2 - 2b_1 \sum Y_i - 2b_2 \sum X_i Y_i + 2b_1 b_2 \sum X_i$$

$$\frac{\partial RSS}{\partial b_1} = 0 \quad \Rightarrow \quad 2nb_1 - 2\sum Y_i + 2b_2 \sum X_i = 0$$

$$b_1 = \overline{Y} - b_2 \overline{X}$$

$$b_2 = \hat{\beta}_2 = \frac{\sum \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right)}{\sum \left(X_i - \overline{X} \right)^2} = \frac{Cov(X,Y)}{Var(X)} = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \sum a_i y_i$$

PRECISION OF THE REGRESSION COEFFICIENTS

$$\sigma^2_{\hat{\beta}_2} = E\left\{\left(\hat{\beta}_2 - E(\hat{\beta}_2)\right)^2\right\} = E\left\{\left(\hat{\beta}_2 - \beta_2\right)^2\right\} = E\left\{\left(\sum_{i=1}^{n} a_i u_i\right)^2\right\} =$$

$$= E\left\{\sum_{i=1}^{n} a_i^2 u_i^2 + \sum_{i=1}^{n}\sum_{j \neq i} a_i a_j u_i u_j\right\} = \sum_{i=1}^{n} a_i^2 E\left(u_i^2\right) + \sum_{i=1}^{n}\sum_{j \neq i} a_i a_j E\left(u_i u_j\right) =$$

$$= \sum_{i=1}^{n} a_i^2 \sigma_u^2 = \sigma_u^2 \sum_{i=1}^{n} a_i^2 = \frac{\sigma_u^2}{\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2}$$

$$\sigma^2_{\hat{\beta}_1} = \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right)$$

PRECISION OF THE MULTIPLE REGRESSION COEFFICIENTS

True model	Fitted model
$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$	$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$

$$\sigma_{b_j}^2 = \frac{\sigma_u^2}{\sum \left(X_{ji} - \overline{X}_j \right)^2} \times \frac{1}{1 - R_j^2} = \frac{\sigma_u^2}{\sum x_{ji}^{\;2}} \times \frac{1}{1 - R_j^2}$$

$$E\left(\frac{1}{n} \sum \hat{u}_i^2 \right) = \frac{n-k}{n} \sigma_u^2 \qquad s_u^2 = \frac{1}{n-k} \sum \hat{u}_i^2$$

$$\text{s.e.}\left(\hat{\beta}_j \right) = \sqrt{\frac{s_u^2}{\sum \left(X_{ji} - \overline{X}_j \right)^2} \times \frac{1}{1 - R_j^2}}$$

Where ${R^2}_j$ *is* determination coefficient of the regression of X_j on all X_m $(m \neq j)$

CONFIDENCE INTERVALS FOR REGRESSION COEFFICIENTS

Model $$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

Null hypothesis: $$H_0 : \beta_2 = \beta_2^0$$

Alternative hypothesis: $$H_1 : \beta_2 \neq \beta_2^0$$

d.f. = n-k

Reject *H*$_0$ if $$\frac{\hat{\beta}_2 - \beta_2^0}{\text{s.e.}\left(\hat{\beta}_2\right)} > t_{\text{crit}}$$ **or** $$\frac{\hat{\beta}_2 - \beta_2^0}{\text{s.e.}\left(\hat{\beta}_2\right)} < -t_{\text{crit}}$$

Reject *H*$_0$ if $$\hat{\beta}_2 - \beta_2^0 > \text{s.e.}\left(\hat{\beta}_2\right) \times t_{\text{crit}}$$ **or** $$\hat{\beta}_2 - \beta_2^0 < -\text{s.e.}\left(\hat{\beta}_2\right) \times t_{\text{crit}}$$

Reject *H*$_0$ if $$\hat{\beta}_2 - \text{s.e.}\left(\hat{\beta}_2\right) \times t_{\text{crit}} > \beta_2^0$$ **or** $$\hat{\beta}_2 + \text{s.e.}\left(\hat{\beta}_2\right) \times t_{\text{crit}} < \beta_2^0$$

Do not reject *H*$_0$ if $$\hat{\beta}_2 - \text{s.e.}\left(\hat{\beta}_2\right) \times t_{\text{crit}} \leq \beta_2 \leq \hat{\beta}_2 + \text{s.e.}\left(\hat{\beta}_2\right) \times t_{\text{crit}}$$

$$(\hat{\beta}_2 - \text{s.e.}\left(\hat{\beta}_2\right) \times t_{\text{crit}} ; \hat{\beta}_2 + \text{s.e.}\left(\hat{\beta}_2\right) \times t_{\text{crit}})$$

- Confidence interval; same for i≠2

Problems that may arise with OLS method within Model A assumptions.

A.1 - violation: The model is linear in parameters and correctly specified

- **Non-linear models**
- **Variables Mispecification**

Omitted variables **Including irrelevant variable**

A.2- violation: There is some variation in the regressor in the sample and no exact linear relationship between regressors in the sample.

- **X is a constant: cannot calculate the relationship**
- **Perfect multicollinearity** $X_3 = \lambda + \mu X_2$

A.3 - (G-M 1)The disturbance term has zero expected value in each observation: $E(u_i) = 0$

(automatically satisfied if intercept is included in regression)

A.4- violation: (G-M 2)The disturbance term is homoscedastic

- **Heteroscedasticity** $\sigma_{u_i}^2 \neq \sigma_u^2$

A.5 – violation: (G-M 3)The values of the disturbance term have independent distributions

- **Autocorrelation**

A.6 The disturbance term has a normal distribution

- **Need to perform statistical tests**

A.1 violation

A.1 - violation: The model is linear in <u>parameters</u> (1)

- **non-linear models (give economic interpretations!)**

Types of nonlinearity

1. **Non-linear in variables** $y = \alpha + \beta \cdot f(x) + u$
   - **Solution: introduce new variable** $z = f(x)$, $y = \alpha + \beta z + u.$
2. **Non linear in parameters**

2.1 of type $y = \alpha x^{\beta} v$ **(elasticities!)**

- **Solution: take log** $\log y = \log \alpha + \beta \log x + \log v$

2.2 Of type $y = \alpha e^{\beta x}$

- **Solution: take log** $\log y = \log \alpha + \beta x$

2.3 Other types

3. Cannot be linearized

- **Non-linear** $y = A(u \cdot K^{-\rho} + (1-u) \cdot L^{-\rho})^{-\frac{n}{\rho}}$

A.2- violation: no exact linear relation between *Xs*

A.2- violation: no exact linear relationship between regressors

This assumption excludes the case of <u>perfect multicollinearity</u>

$$b_2 = \frac{\mathrm{Cov}(X_2,Y)\mathrm{Var}(X_3) - \mathrm{Cov}(X_3,Y)\mathrm{Cov}(X_2,X_3)}{\mathrm{Var}(X_2)\mathrm{Var}(X_3) - \left[\mathrm{Cov}(X_2,X_3)\right]^2} =$$

$$= \frac{\mathrm{Cov}(X_2,Y)\mathrm{Var}(\lambda+\mu X_2) - \mathrm{Cov}(\lambda+\mu X_2,Y)\mathrm{Cov}(X_2,\lambda+\mu X_2)}{\mathrm{Var}(X_2)\mathrm{Var}(\lambda+\mu X_2) - \left[\mathrm{Cov}(X_2,\lambda+\mu X_2)\right]^2} = \frac{0}{0}$$

Non-perfect multicollinearity can be present

Multicollinearity (non-perfect)

- Arises in multiple regression
- Regressors are highly correlated
- s.e. become large
- t-stat becomes low
- Wrong conclusion due to type II error!

Note: estimated still unbiased and s.e. are valid

Multicollinearity (non-perfect)

- It is not a problem in fact, it is a typical situation when t and F tests have low power
- Reasons: nature of the data
- How to check: correlation matrix, R squares from auxiliary regressions (1/(1-R_sq_aux))
- Consequences:
  - s.e. of estimators goes up (**No inefficiency!**)
  - Estimates become very sensitive to changes in data and specification of the model
- How to overcome
  - Increase N (observations)
  - Reduce correlation between regressors
  - Combine correlated variables
  - Drop one of correlated X (But omitted variable bias may arise)
  - **Impose restrictions on parameters**

Multiple Linear Regression Model:

_F_ TEST OF GOODNESS OF FIT FOR THE WHOLE EQUATION

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

$$H_0 : \beta_2 = \ldots = \beta_k = 0$$

H_1 : at least one $\beta \neq 0$

$$F(k-1, n-k) = \frac{(RSS_r - RSS_{ur})/(k-1)}{RSS_{ur}/(n-k)} = \frac{(TSS - RSS)/(k-1)}{RSS/(n-k)} =$$

$$= \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{\frac{ESS}{TSS}\Big/(k-1)}{\frac{RSS}{TSS}\Big/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

F TESTS RELATING TO GROUPS OF EXPLANATORY VARIABLES

$$Y = \beta_1 + \beta_2 X_2 + u \qquad SSR_1$$

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u \qquad SSR_2$$

$$H_0 : \beta_3 = \beta_4 = 0$$

$H_1 : \beta_3 \neq 0$ or $\beta_4 \neq 0$ or both β_3 and $\beta_4 \neq 0$

$$F(\text{cost in d.f., d.f. unrestricted}) = \frac{\text{reduction in } SSR \;/\; \text{cost in d.f.}}{SSR \text{ unrestricted} \;/\; \text{degrees of freedom unrestricted}}$$

$$F(2, n-4) = \frac{(SSR_1 - SSR_2)/2}{SSR_2/(n-4)} = \frac{(R_2^2 - R_1^2)/2}{(1 - R_2^2)/(n-4)}$$

Restricted and unrestricted models

- Why we need restrictions?
  - Because our model in these cases becomes less complicated
  - Testing of economic theories (like constant returns to scale or unit elasticity)
- We have some trade-off:
  - If restrictions are valid we can estimate less parameters, then estimators are nonetheless unbiased and more efficient
  - If restrictions are invalid our estimators become (asymptotically) biased

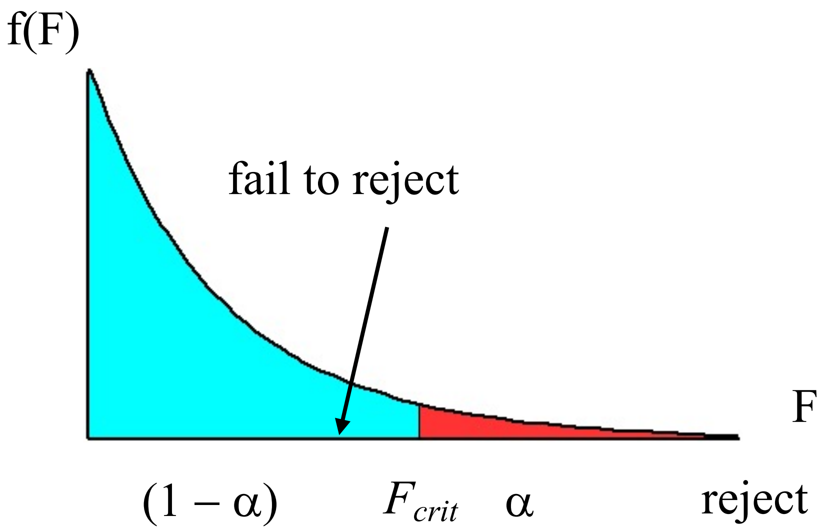
F TESTS FOR LINEAR RESTRICTIONS IN GENERAL, AND FOR GROUPS OF EXPLANATORY VARIABLES

$$F\text{ (cost in d.f., d.f. unrestricted)} = \frac{\text{reduction in } RSS \;/\; \text{cost in d.f.}}{RSS \text{ unrestricted} \;/\; \text{degrees of freedom unrestricted}}$$

$$F(q, n-k) = \frac{\left(RSS_r - RSS_{ur}\right)/q}{RSS_{ur}/(n-k)} = \frac{\left(R^2_{ur} - R^2_r\right)/q}{\left(1 - R^2_{ur}\right)/(n-k)},$$

(*r - restricted, ur - unrestricted.*

q - the number of linear restrictions on coefficients).



LINEAR RESTRICTION: EDUCATIONAL ATTAINMENT FUNCTION EXAMPLE

S – years of schooling;
SM – years of schooling of mother;
SF – years of schooling of father.

If SM and SF are strongly correlated, the coefficients may be insignificant due to multicollinearity.

Options for a restriction: 1) $\beta_3 = 0;$ **2)** $\beta_4 = 0;$ **3)** $\beta_3 = \beta_4.$

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$\beta_3 = \beta_4$$

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 (SM + SF) + u = \beta_1 + \beta_2 ASVABC + \beta_3 SP + u$$

Here we define *SP* as the sum of *SM* and *SF* (total parental schooling as the indicator of family background*).* The problem caused by multicollinearity has been eliminated.