A Brief Introduction to Linear Regression Model with Panel Data

- A brief overview of linear regression model assumptions.
- A brief overview of cross-section and panel data.
- Test of hypothesis for model specification test.
- Brief overview of heterogeneity and endogeneity issues with panel data.
- Between groups and between times model.
- One way and Two way fixed and random effect models overview.

Pinto Bhusan Datta, M.Sc. (Statistics), M.Sc. (Operation Research and Business Analytics (Pursuing))
Working Student, HVB.

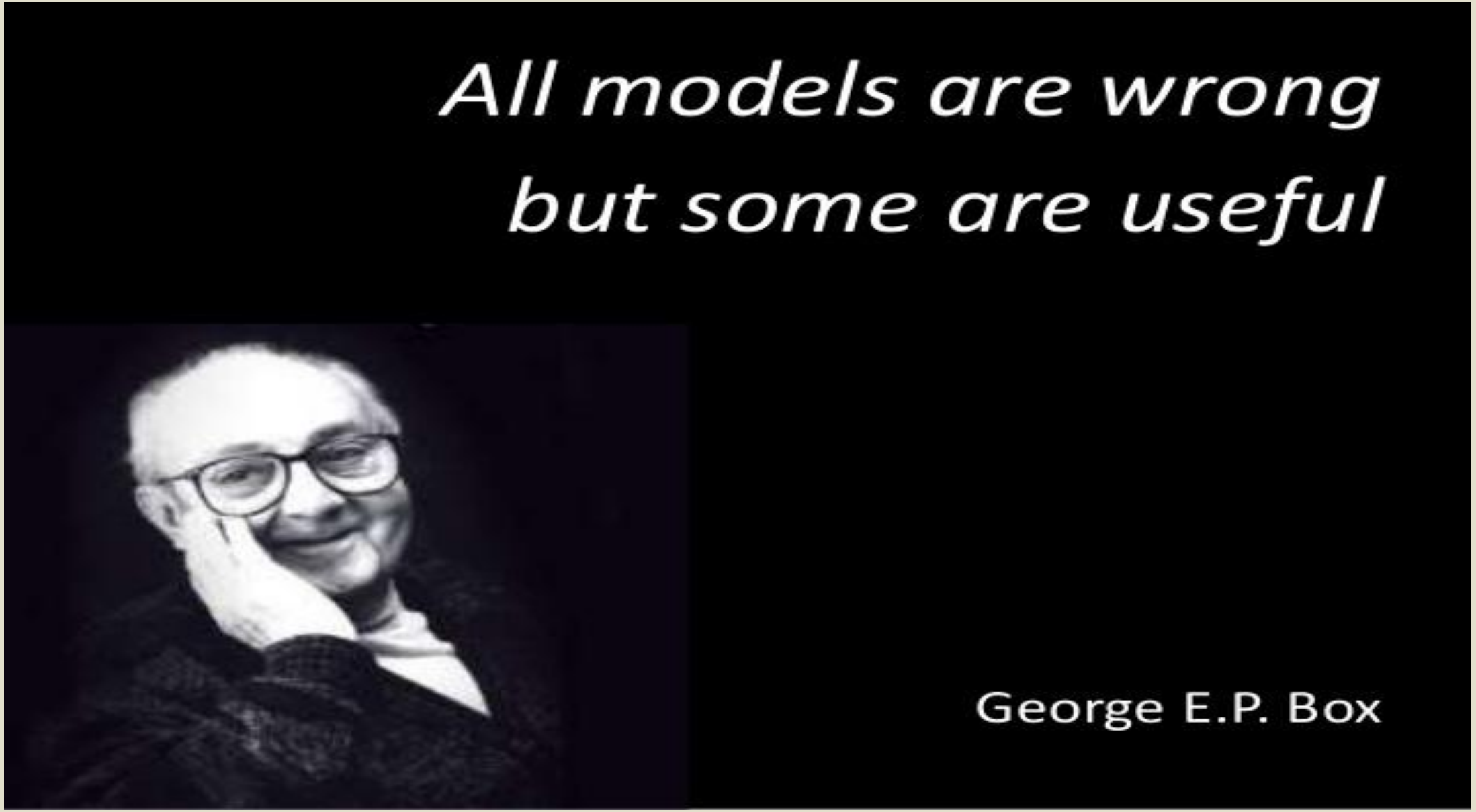


Image source: "All models are wrong, but some are useful", George E. P. Box – AdMoRe ITN (upc.edu)
Details: All models are wrong - Wikipedia

"No Free Lunch" :(

D. H. Wolpert. The supervised learning no-free-lunch theorems. In Soft Computing and Industry, pages 25–42. Springer, 2002.

Our model is a simplification of reality

Simplification is based on assumptions (model bias)

Assumptions fail in certain situations

Roughly speaking:

"No one model works best for all possible situations."

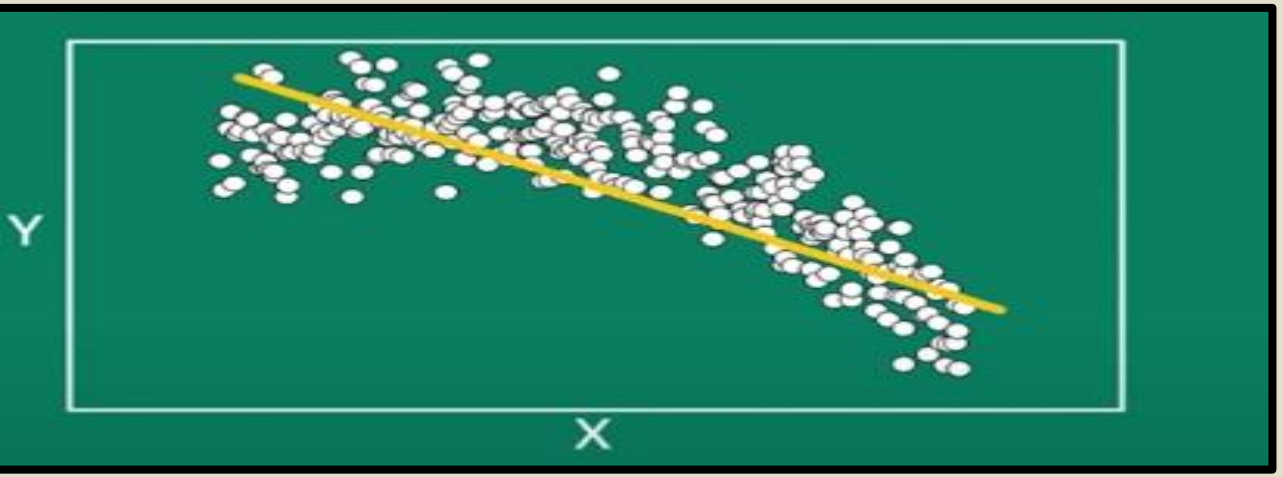
Linear Regression Assumptions

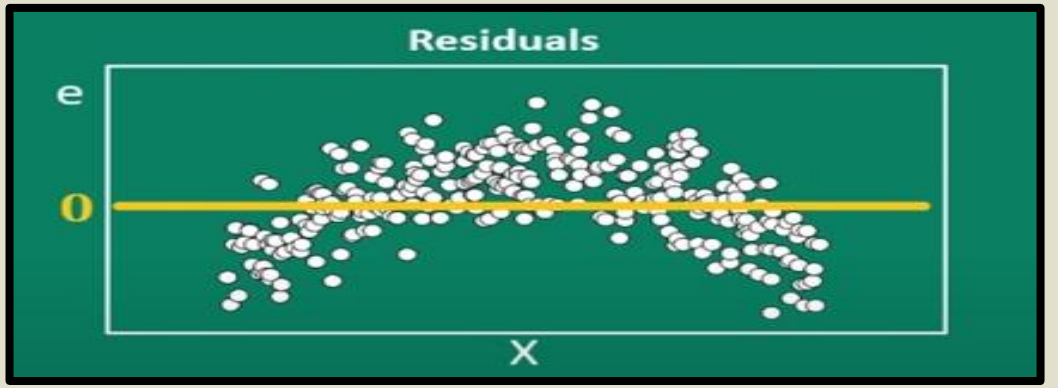
1. Linearity

(Correct functional form)

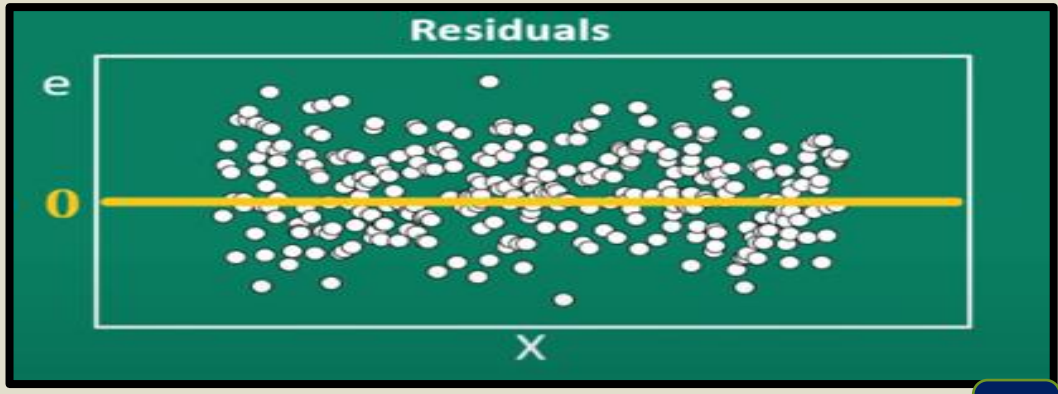
Consider the following model:

$$Lung\ Function_i = \beta_0 + \beta_1(age)_i + \varepsilon_i$$





$$Lung\ Function_i = \beta_0 + \beta_1(age)_i + \beta_2(age^2)_i + \varepsilon_i$$



Source: Regression assumptions explained! - YouTube

Other forms

Linear in parameter

$$y = Ax^{\beta}e^{\varepsilon}$$

Elasticity from Linear Regression Model

Example 2: lnY = a + blnX

$\frac{1}{Y}dY = \frac{b}{X}dX$ (divide through by dX)

$\frac{dY}{dX}\frac{1}{Y} = \frac{b}{X}$ (multiply through by X)

$$\frac{dY}{dX}\frac{X}{Y} = b$$

Non-linear in parameter

$$y = Ax^{\beta} + \varepsilon$$

What's the issue

Due to wrong functional form there is a great chance that both the coefficients and standard error are biased

Detection Methods

- **Residual plots**
- **Likelihood Ratio (LR) Test**

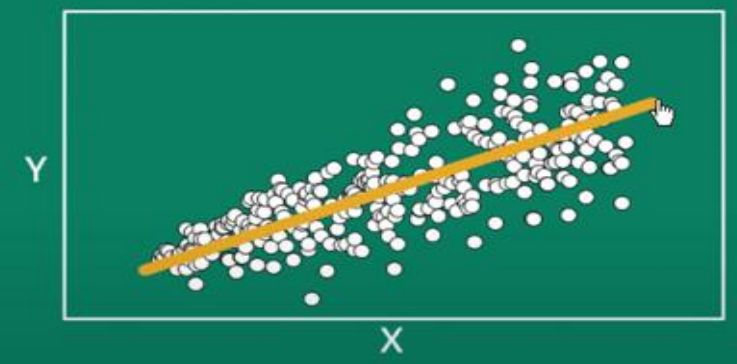
Treatments

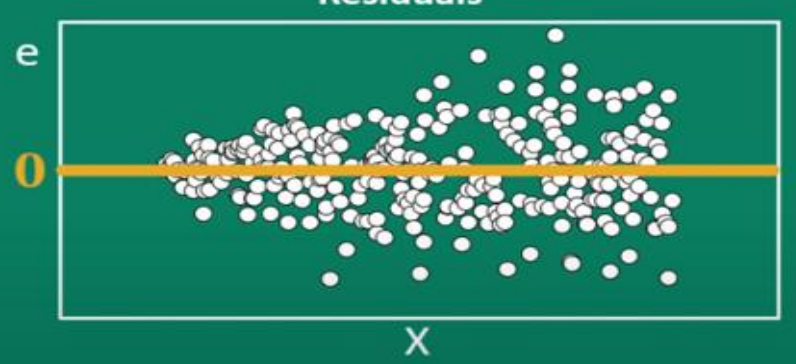
- **Get the correct specification**
- **Correct (Trial and error)**

Machine Learning **methods are more prudent here!!!** 

2. Constant Error Variance

$$Expenditure_i = \beta_0 + \beta_1(Income)_i + \varepsilon_i$$





$$\text{Var}[\varepsilon_i \mid \mathbf{X}] = \sigma^2, \qquad \text{for all } i = 1, \ldots, n,$$

What's the issue

Due to heteroscedasticity, standard error will be biased

Detection Methods

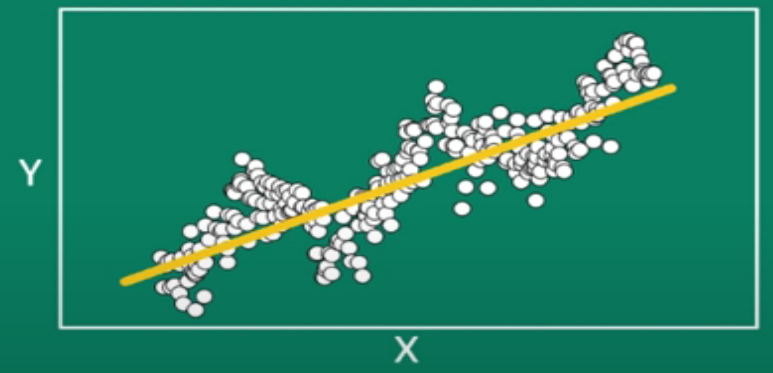
- **Residual plots**
- **Goldfeldt-Quant test**
- **Breusch-Pagan test**

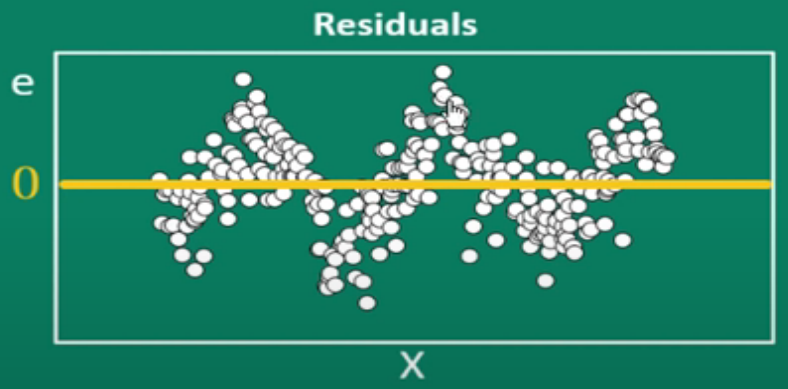
Treatments

- Whits standard errors
- Weighted least squares
- Log transformation and many more

3. Independent Error Terms

$$Stock\ Index_i = \beta_0 + \beta_1(Time)_i + \varepsilon_i$$





$$\mathrm{Cov}[\varepsilon_i, \varepsilon_j \mid \mathbf{X}] = 0, \qquad \text{for all } i \neq j.$$

Homoscedasticity and No-Autocorrelation

$$E[\varepsilon\varepsilon' \mid \mathbf{X}] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_1\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_1\varepsilon_n \mid \mathbf{X}] \\ E[\varepsilon_2\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_2\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_2\varepsilon_n \mid \mathbf{X}] \\ \vdots & \vdots & \vdots & \vdots \\ E[\varepsilon_n\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_n\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_n\varepsilon_n \mid \mathbf{X}] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix},$$

What's the issue

Due to autocorrelation, standard error will be biased as well as high prediction bias.

Detection Methods

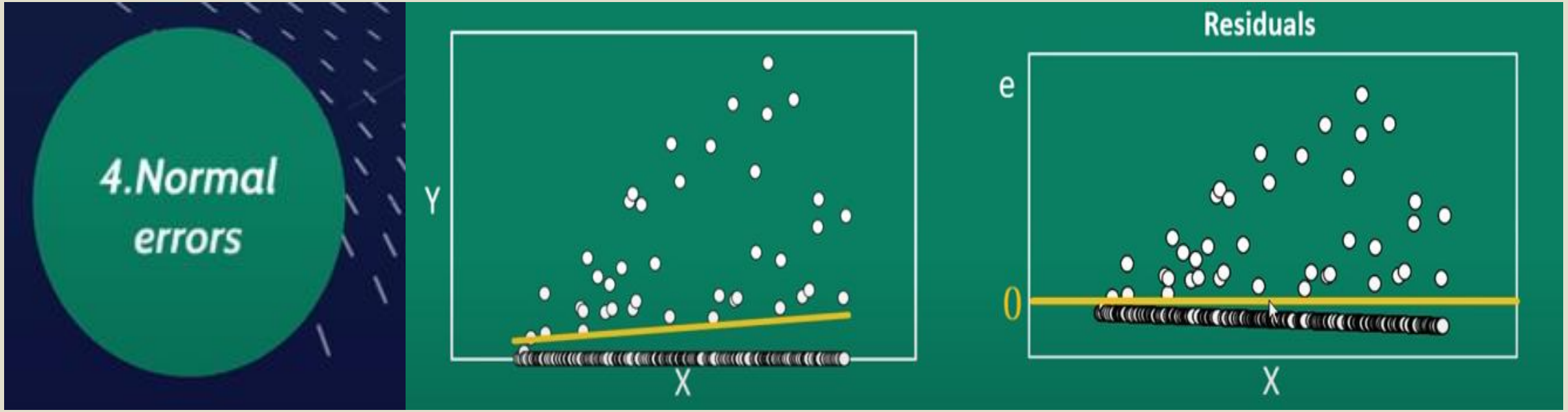
- **Residual plots**
- **Durbin-Watson test**
- **Breusch-Godfrey test**

Treatments

- **Searching omitted variables**
- **Generalized difference equations**
- **Advance modelling techniques**

  Univariate: ARIMA, ARIMAx, SARIMA, SARIMAx, ARCH , GARCH etc.

  Multivariate: like VAR, machine learning (RNN, LSTM, GRU etc.)



$$\varepsilon \mid \mathbf{X} \sim N[\mathbf{0}, \sigma^2 \mathbf{I}].$$

What's the issue

Due to normality violation, standard error will be biased

Detection Methods

- Histogram or Q-Q plot
- Shapirio-wilk test
- Kolmogrov-Smirnob test
- Anderson-Darling test

Treatments

- Take log transformation or other variables transformations
- Advance modelling techniques non-parametric methods: Stochastic Frontier.

$$Motor\ Accidents_i = \beta_0 + \beta_1(Num\ cars)_i + \beta_2(Num\ residents)_i + \varepsilon$$

X2 = A + B * X1

What's the issue

Due to multicolinearity, coefficients and standard error will be biased of the effected variables

Detection Methods

- Scatter plot Matrix or heatmap
- Variance inflation factor (VIF)

Treatments

- Delete strong correlated variables except one

$$Salary_i = \beta_0 + \beta_1(Years\ of\ education)_i + \varepsilon_i$$

Socio-economic status affects **both** X and Y variables, thus would cause **omitted variable bias.**

TECHNICALLY - Socio-economic status would affect ε_i in the model, thus, Education is no longer wholly exogenous as it can be explained in part by the error term.

$$E[\varepsilon_i \mid \mathbf{X}] = 0.$$

$$E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \begin{bmatrix} E[\varepsilon_1 \mid \mathbf{X}] \\ E[\varepsilon_2 \mid \mathbf{X}] \\ \vdots \\ E[\varepsilon_n \mid \mathbf{X}] \end{bmatrix} = \mathbf{0}.$$

What's the issue

Due to endogeneity problem model can use for prediction but not causation

Detection Methods

- **Model speciation test**
- **Test of hypothesis**

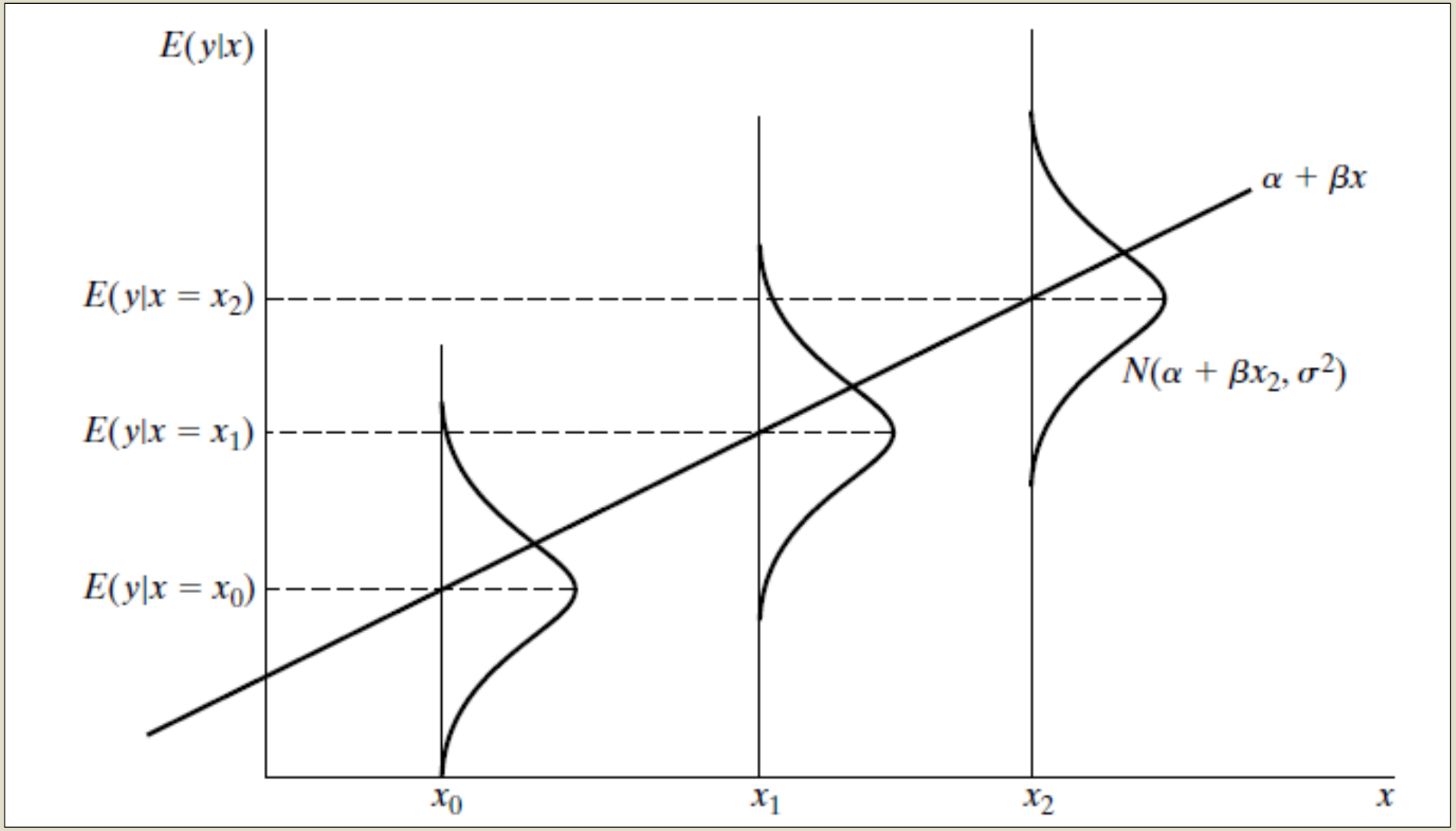
Treatments

- Using Instrumental variables

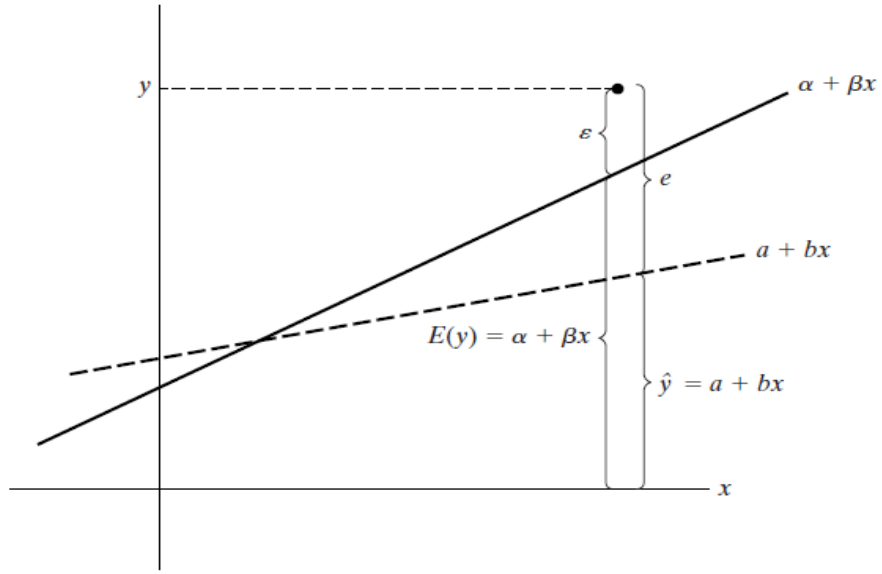
Other assumptions:

- **Full-Rank:**
  The number of parameters must be less than number of observations.
- **Data Generation:**
  X values fixed in repeated samples i.e. non-stochastic.
- **The explanatory variable in the given sample must not be the same i.e Var(X) must be finite positive value.**
- **Zero covariance between error and explanatory variables.**

The Normal Linear Regression Model



Least Square Regression



THE LEAST SQUARES COEFFICIENT VECTOR

The necessary condition for a minimum is

$$\frac{\partial S(\mathbf{b}_0)}{\partial \mathbf{b}_0} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}_0 = \mathbf{0}.$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

For this solution to minimize the sum of squares,

$$\frac{\partial^2 S(\mathbf{b})}{\partial \mathbf{b}\, \partial \mathbf{b}'} = 2\mathbf{X}'\mathbf{X}$$

THEOREM 3.2 Orthogonal Partitioned Regression
In the multiple linear least squares regression of $\mathbf{y}$ *on two sets of variables* $\mathbf{X}_1$ *and* $\mathbf{X}_2$*, if the two sets of variables are orthogonal, then the separate coefficient vectors can be obtained by separate regressions of* $\mathbf{y}$ *on* $\mathbf{X}_1$ *alone and* $\mathbf{y}$ *on* $\mathbf{X}_2$ *alone.*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

coefficient of determination: $\dfrac{\text{SSR}}{\text{SST}} = \dfrac{\mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = 1 - \dfrac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}}.$

$$\bar{R}_j^2 = 1 - \frac{n + K_j}{n - K_j}(1 - R_j^2).$$

$$\text{AIC}(K) = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{2K}{n}$$

$$\text{BIC}(K) = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{K \log n}{n}.$$

Law of large numbers and Central limit Theorem

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{X_i}{n} = \overline{X}$$

$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$

converges to the expected value:

$\overline{X}_n \to \mu \quad \text{as } n \to \infty.$

Weak law [edit]

The **weak law of large numbers** (also called Khinchin's law) states that the sample average converges in probability towards the expected value[17]

$$\overline{X}_n \xrightarrow{p} \mu \qquad \text{when } n \to \infty.$$

That is, for any positive number ε,

$$\lim_{n\to\infty} \Pr\left(|\overline{X}_n - \mu| < \varepsilon\right) = 1.$$

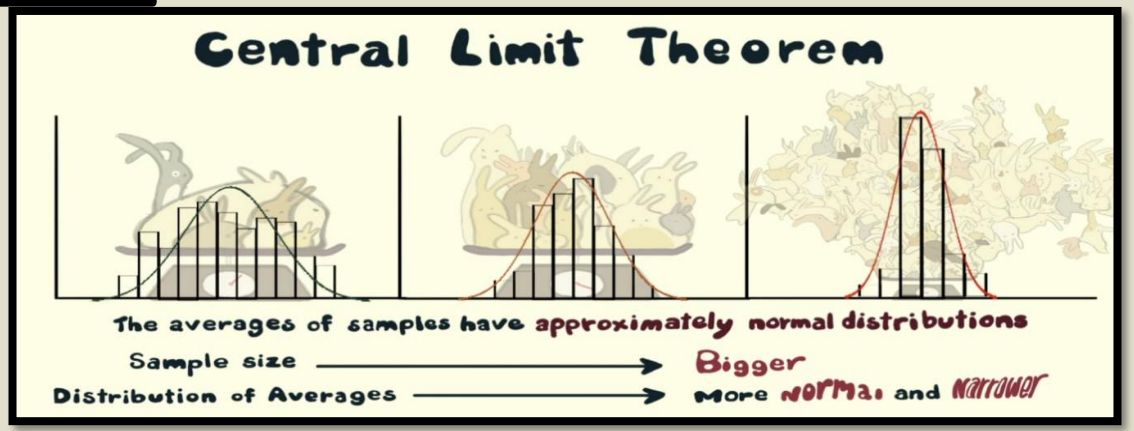
Strong law [edit]

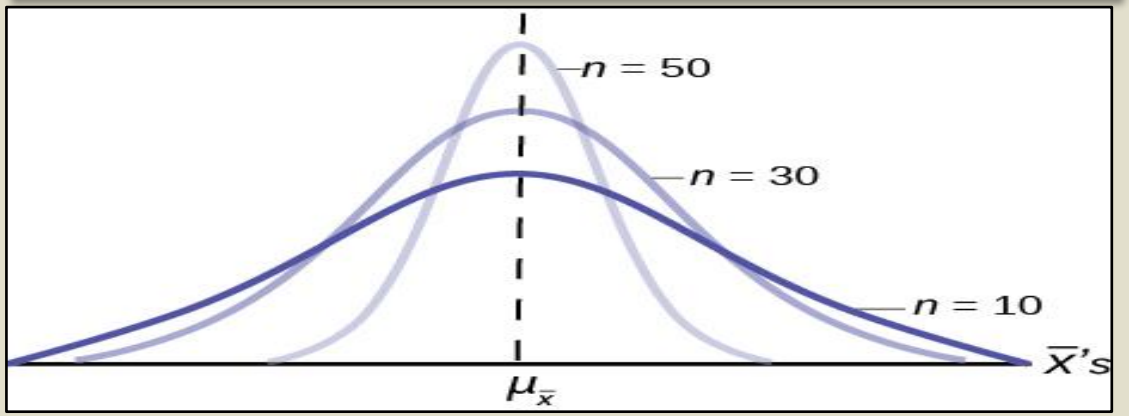
The **strong law of large numbers** (also called Kolmogorov's law) states that the sample average converges almost surely to the expected value[18]

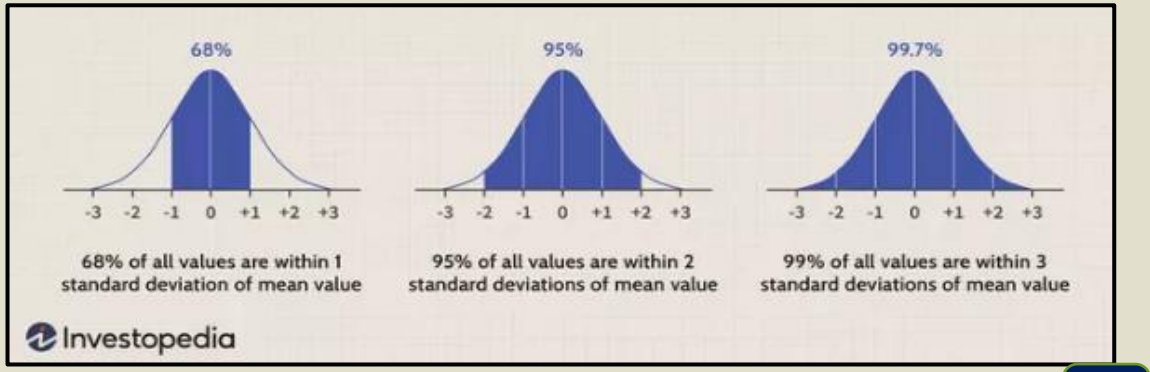
$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu \qquad \text{when } n \to \infty.$$

That is,

$$\Pr\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$







Finite sample properties of Least square estimator

Terms of Art

- Estimates and estimators
- Properties of an estimator - the sampling distribution
- "Finite sample" properties as opposed to "asymptotic" or "large sample" properties

Ordinary Least squares estimator and its variance

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$$

$$E[\mathbf{b}|\mathbf{X}] = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}[\varepsilon \mid \mathbf{X}] = \beta \text{ as } \mathbf{E}[\varepsilon \mid \mathbf{X}] = \mathbf{0}$$

$$\begin{aligned}\text{Var}[\mathbf{b} \mid \mathbf{X}] &= E[(\mathbf{b}-\beta)(\mathbf{b}-\beta)' \mid \mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}[\varepsilon\varepsilon' \mid \mathbf{X}]\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Testing a Hypothesis about coefficient and Regression

$$t_k = \frac{(b_k - \beta_k)/\sqrt{\sigma^2 S^{kk}}}{\sqrt{[(n-K)s^2/\sigma^2]/(n-K)}} = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}} \qquad (4\text{-}13)$$

$$t = \frac{b_k}{s_{b_k}} \qquad (4\text{-}14)$$

$$F[K-1, n-K] = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}. \qquad (4\text{-}15)$$

Large sample properties and instrumental variables

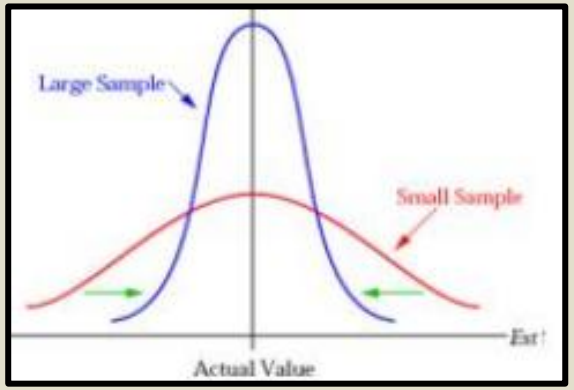
Taking Stock: The Ordinary Least Squares (OLS) Estimation Procedure

- OLS Bias Question
- OLS Reliability Question

Estimation Procedures: Unbiased, Biased, Consistent, and Inconsistent

- Unbiased and Consistent Estimation Procedure
- Unbiased but Inconsistent Estimation Procedure
- Biased but Consistent Estimation Procedure

Instrumental Variables (IV): A Two-Step Estimation Procedure



OLS Bias Question: Are the model's explanatory variable and error term independent or correlated?

- **Independent**
  - OLS estimation procedure for the value of the Coefficient is Unbiased
  - **OLS Reliability Question:** Are the OLS standard error term premises satisfied or violated?
    - If "Violated" then use alternative approach like GLS.
    - Satisfied
      - The OLS estimation procedure is BLUE and the coefficient standard error may be trusted.
- **Correlated**
  - Use alternative approach which may be biased but consistent.

Others measures for assessing prediction accuracy

$$\text{RMSE} = \sqrt{\frac{1}{n^0}\sum_i (y_i - \hat{y}_i)^2}$$

$$\text{MAE} = \frac{1}{n^0}\sum_i |y_i - \hat{y}_i|,$$

$$U = \sqrt{\frac{(1/n^0)\sum_i (y_i - \hat{y}_i)^2}{(1/n^0)\sum_i y_i^2}}.$$

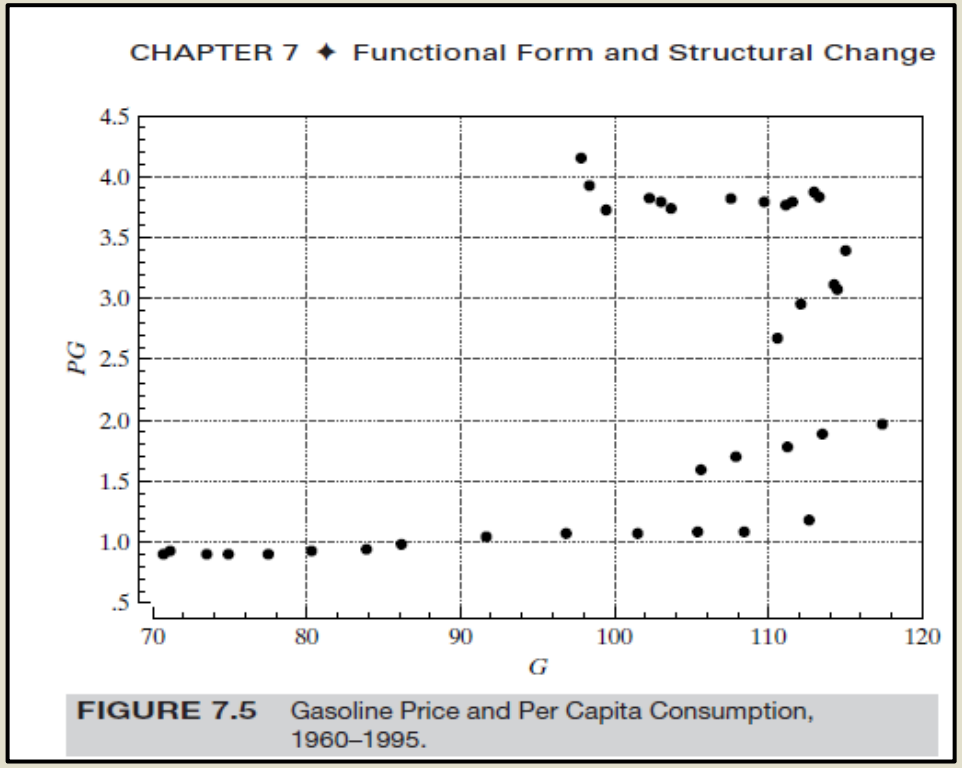


FIGURE 7.5 Gasoline Price and Per Capita Consumption, 1960–1995.

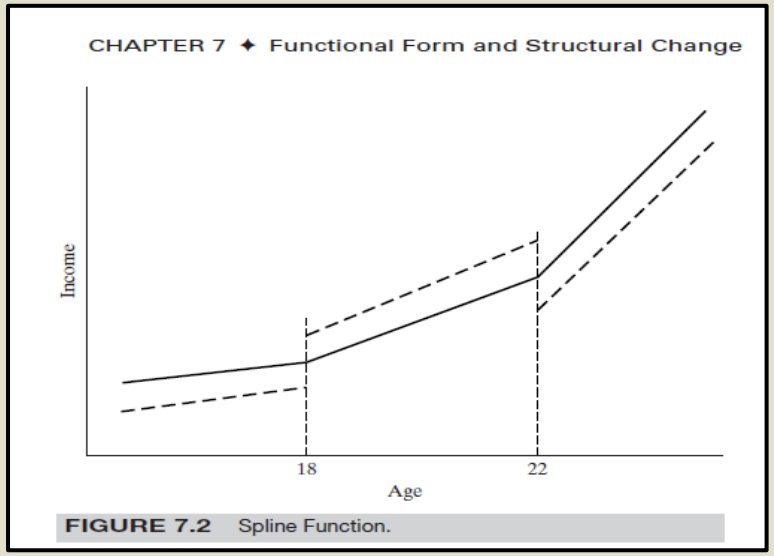


FIGURE 7.2 Spline Function.

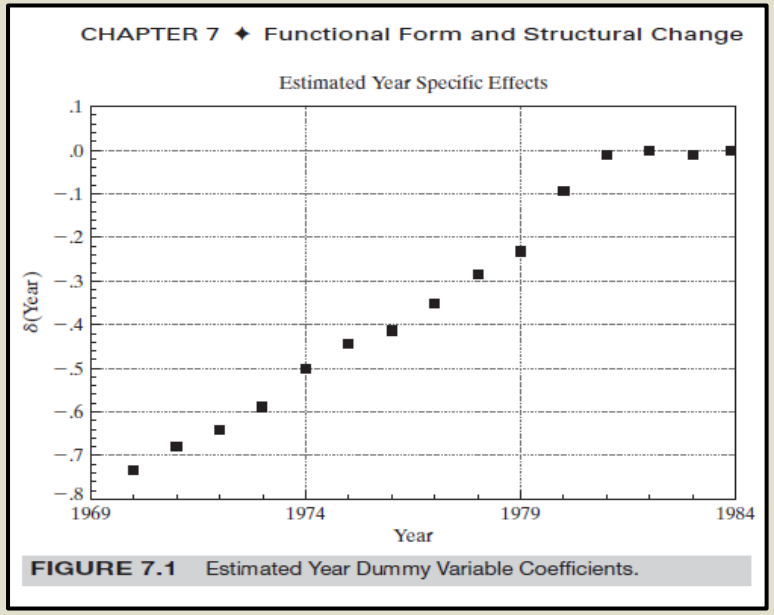


FIGURE 7.1 Estimated Year Dummy Variable Coefficients.

Model Specification Tests

Functional form specification

- The functional form for the regression model needs to be correctly specified. The functional form may include square terms, interactions terms, logs of variables, etc.
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$
- $log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 + u$
- $log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 + u$

RESET – regression specification error test

- RESET (regression specification error test) includes squares, cubes, and possibly higher order of the fitted values for the dependent variable in the regression model and tests for their joint coefficient significance.
- Regression model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$
- Obtain fitted values $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$
- Calculate the squares $\hat{y}^2$ and cubes $\hat{y}^3$.
- Estimate the regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + e$
- H_0 $\delta_1 = 0$ and $\delta_2 = 0$ (correctly specified model)
- H_a $\delta_1 \neq 0$ or $\delta_2 \neq 0$ (misspecified model)

Omitted variable bias (review)

- The "true" population regression model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
- We need to estimate: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
- But instead we estimate a misspecified model: $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$, where x_2 is the omitted variable from this model. The coefficient $\tilde{\beta}_1$ will be biased.

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

Substitute in above equation to get:

$$y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u = (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) x_1 + (\beta_2 v + u)$$

The coefficient that will be estimated for x_1 when x_2 is omitted will be biased.

Proxy variables

- The true population regression model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
- If we do not have data on x_2 and it is omitted from the regression, there will be omitted variable bias.
- Find a "proxy" variable x_2^P.
- If x_2^P and x_2 are correlated, there will be a relationship between them: $x_2 = \delta_0 + \delta_2 x_2^P + v$

Others Specification Tests

Lagrange Multiplier (LM) Test

Wald Test Tests

Classical Model Selection

- ❑ Parametric Vision.
- ❑ Assuming a true data-generating process.
- ❑ Evaluation based on fit.
- ❑ Ignoring Model uncertainty.

Source: Regression Specification - YouTube

Cross-Sectional Data

Time	Y_1	Y_2	Y_N
1	Y_{11}	Y_{21}	Y_{N1}

Panel Data

Time	Y_1	Y_2	Y_N
1	Y_{11}	Y_{21}	Y_{N1}
2	Y_{12}	Y_{22}	Y_{N2}
3	Y_{13}	Y_{23}	Y_{N3}
4	Y_{14}	Y_{24}	Y_{N4}
⋮	⋮	⋮	⋮
T-1	Y_{1T-1}	Y_{2T-1}	Y_{NT-1}
T	Y_{1T}	Y_{2T}	Y_{NT}

Source: [Lesson 1] The Charm of DSLR Cameras (canon-asia.com)
Datatabel: Introduction to the Fundamentals of Panel Data - Aptech

Balanced Panel Data

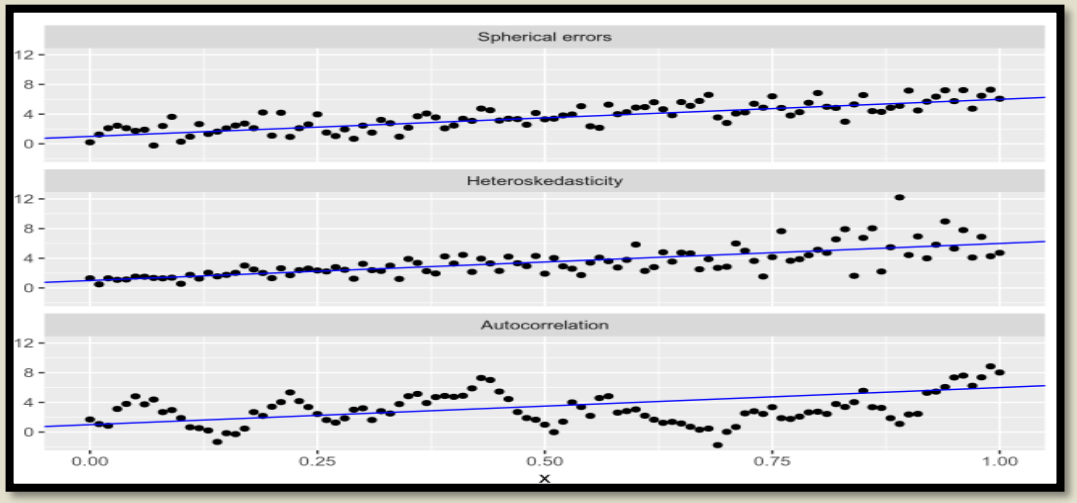
Person	Year	Income	Age	Sex
1	2013	20,000	23	F
1	2014	25,000	24	F
1	2015	27,500	25	F
2	2013	35,000	27	M
2	2014	42,500	28	M
2	2015	50,000	29	M

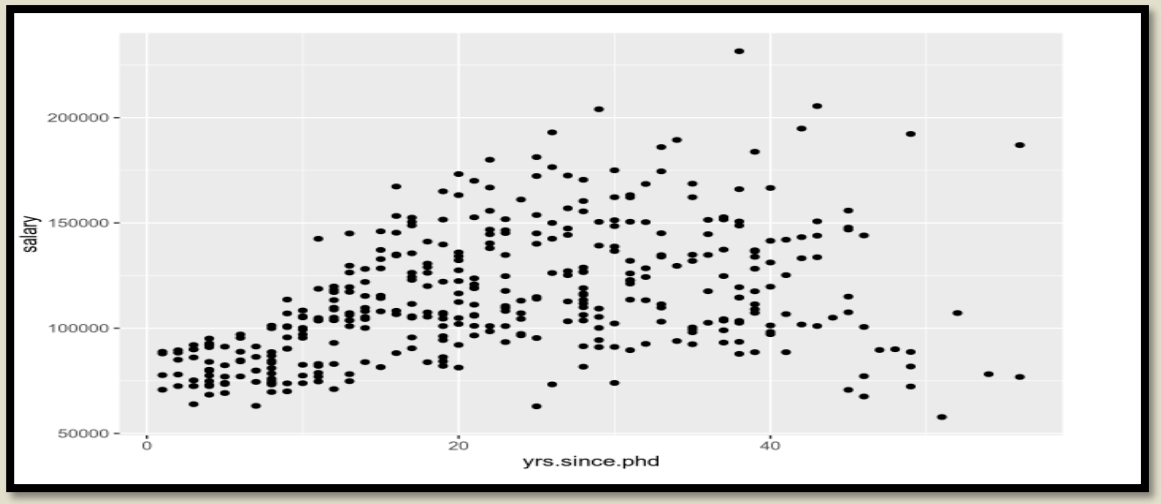
Unbalanced Panel Data

Person	Year	Income	Age	Sex
1	2013	20,000	23	F
1	2014	25,000	24	F
2	2013	35,000	27	M
2	2014	42,500	28	M
2	2015	50,000	29	M
3	2014	46,000	25	F


COFFEE
BREAK

NONSPHERICAL DISTURBANCES — THE GENERALIZED REGRESSION MODEL (GLS)





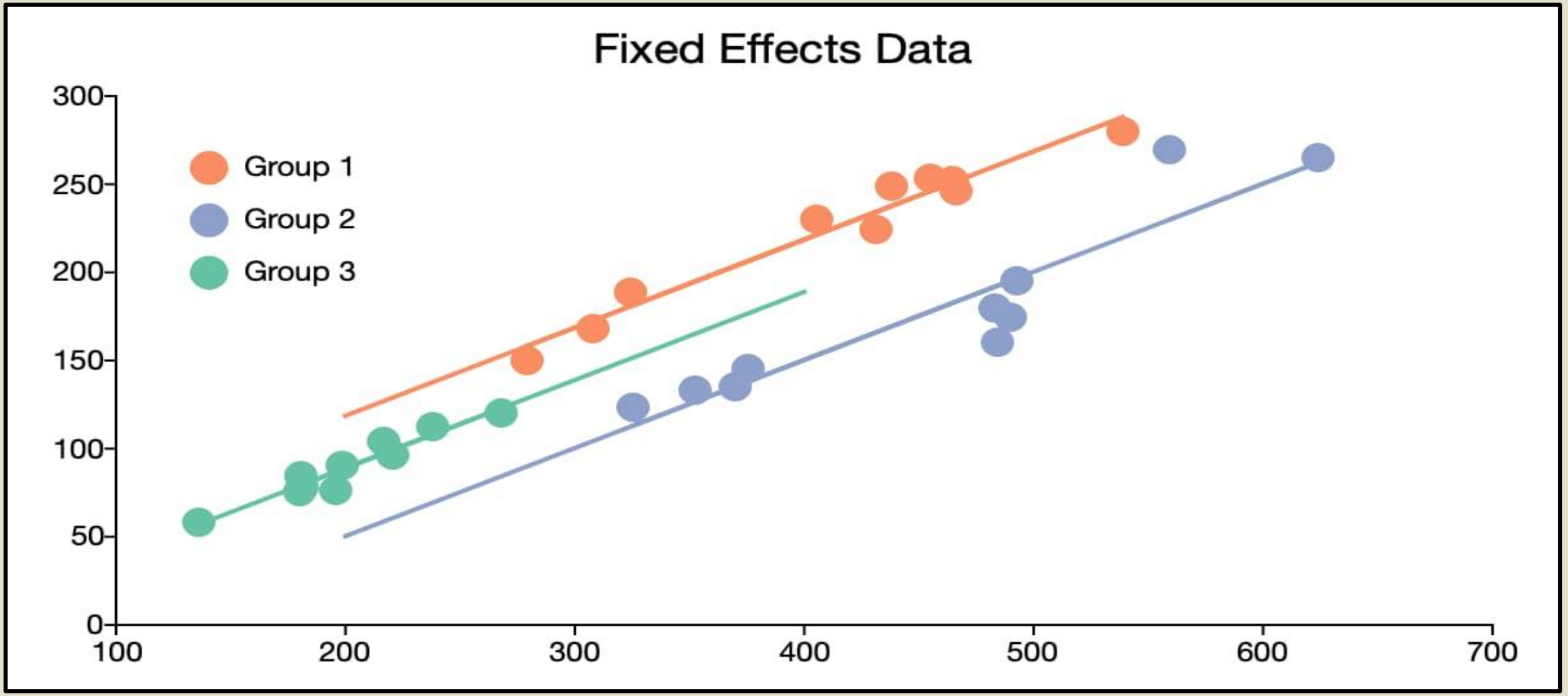
$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} \omega_{11} & 0 & \cdots & 0 \\ 0 & \omega_{22} & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \omega_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

$$\sigma^2\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ & & \vdots & \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}.$$

- True GLS uses $[\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$ which converges in probability to $\boldsymbol{\beta}$.
- We seek a vector which converges to the same thing that this does. Call it "feasible" GLS, FGLS, based on $\left[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}$

One Way Fixed Effect Model

$$y_{it} = \mathbf{x}_{it}'\beta + \alpha_i + \varepsilon_{it}, \quad \text{where } \alpha_i = \mathbf{z}_i'\alpha$$



$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{i} & \cdots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{i} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_n \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

Fixed Effect Model (Within Transformation)

The within transformation demeans the variables:

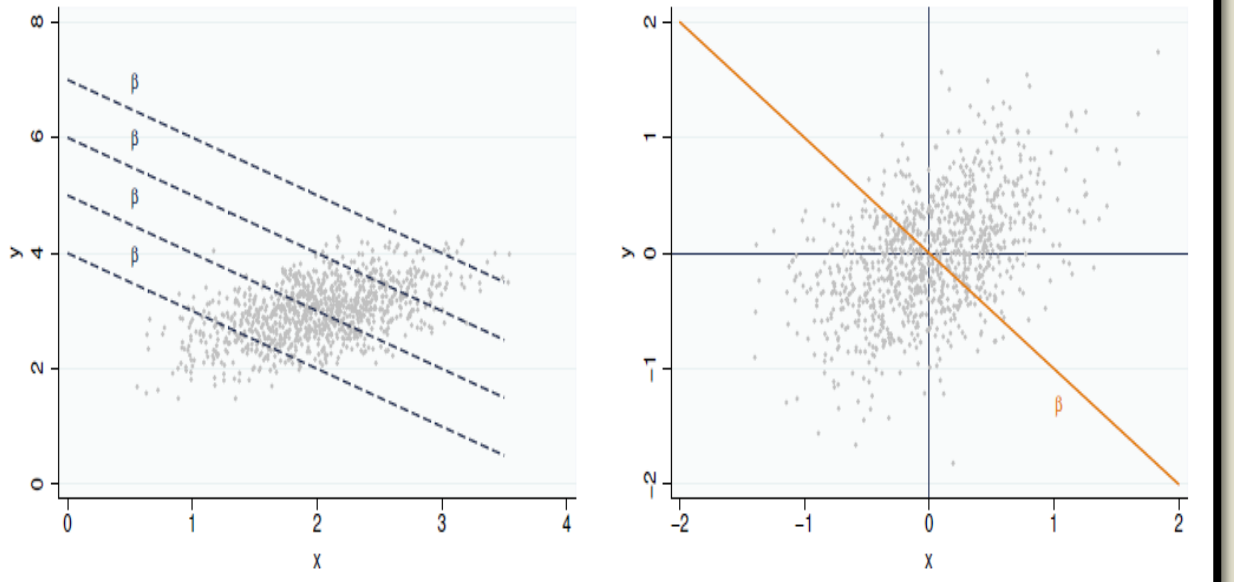
$$\tilde{y}_{it} = \beta_0 + \beta_1 \tilde{x}_{1it} + \ldots + \beta_K \tilde{x}_{Kit} + \tilde{\epsilon}_{it}$$

$$\begin{pmatrix} y_{11} - \bar{y}_1 \\ y_{12} - \bar{y}_1 \\ y_{21} - \bar{y}_2 \\ y_{22} - \bar{y}_2 \\ \vdots \\ y_{N1} - \bar{y}_N \\ y_{N2} - \bar{y}_N \end{pmatrix} = \begin{pmatrix} 1 & x_{1,11} - \bar{x}_{1,1} & \cdots & x_{K,11} - \bar{x}_{K,1} \\ 1 & x_{1,12} - \bar{x}_{1,1} & \cdots & x_{K,12} - \bar{x}_{K,1} \\ 1 & x_{1,21} - \bar{x}_{1,2} & \cdots & x_{K,21} - \bar{x}_{K,2} \\ 1 & x_{1,22} - \bar{x}_{1,2} & \cdots & x_{K,22} - \bar{x}_{K,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n1} - \bar{x}_{1,N} & \cdots & x_{K,N1} - \bar{x}_{K,N} \\ 1 & x_{1,n2} - \bar{x}_{1,N} & \cdots & x_{K,N2} - \bar{x}_{K,N} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_{11} - \bar{\epsilon}_1 \\ \epsilon_{12} - \bar{\epsilon}_1 \\ \epsilon_{21} - \bar{\epsilon}_2 \\ \epsilon_{22} - \bar{\epsilon}_2 \\ \vdots \\ \epsilon_{N1} - \bar{\epsilon}_N \\ \epsilon_{N2} - \bar{\epsilon}_N \end{pmatrix}$$

Two Way Fixed Effect Model (Within Transformation + Dummy Technique)

$$\tilde{y}_{it} = \beta_0 + \beta_1 \tilde{x}_{1it} + \ldots + \beta_K \tilde{x}_{Kit} + \sum_{t=1}^{T-1} \tau_t TD_t + \epsilon_{it}$$

What's the difference between dummies and within transformation



First deference estimator

$$\Delta y_{it} = \beta \Delta x_{it} + \Delta \epsilon_{it}$$

with

- $\Delta y_{it} = (y_{it} - y_{i,t-1})$
- $\Delta x_{it} = (x_{it} - x_{i,t-1})$
- $\Delta \epsilon_{it} = (\epsilon_{it} - \epsilon_{i,t-1})$

A dynamic model

$$y_{it} = \gamma_1 y_{i,t-1} + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \alpha_i + \epsilon_{it}$$

$$\Delta y_{it} = \gamma_1 \Delta y_{i,t-1} + \Delta \boldsymbol{x}'_{it}\boldsymbol{\beta} + \Delta \epsilon_{it}$$

One Way Random Effect Model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + E[\mathbf{z}_i'\boldsymbol{\alpha}] + \{\mathbf{z}_i'\boldsymbol{\alpha} - E[\mathbf{z}_i'\boldsymbol{\alpha}]\} + \varepsilon_{it}$$
$$= \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha + u_i + \varepsilon_{it},$$

Two Way Random Effect Model

$$u_{it} = v_i + e_t + \epsilon_{it}$$

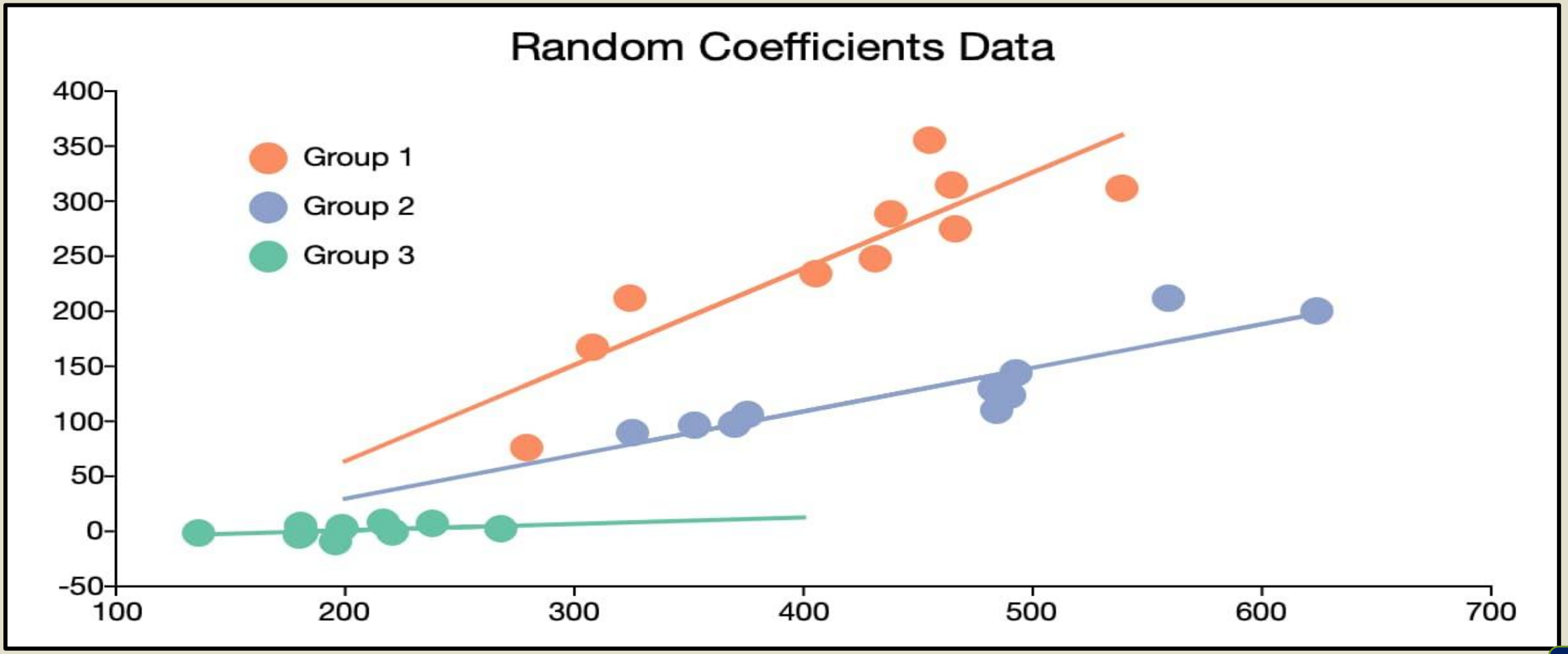


Image source: Introduction to the Fundamentals of Panel Data - Aptech

Estimation methods for One Way and Two Way random effect model.

- **Fuller and Battese Method**
- **Wansbeek and Kapteyn Method**
- **Wallace and Hussain Method**
- **Nerlove Method**

Specification Tests

***F* Test**

Breusch-Pagan (BP) test for one-way random effects

Breusch-Pagan (BP) test for two-way random effects

Advance Methods (Hybrid Method)

- **Hausman-Taylor Estimation**
- **Amemiya-MaCurdy Estimation**
- **Parks Method (Autoregressive Model)**
- **Da Silva Method (Variance-Component Moving Average Model)**

Panel Data Poolability Test

Null hypothesis: Poolability assumes homogeneous slope coefficients i.e. There is no fixed effect.

***F* Test**

$$F = \frac{(SSE_r - SSE_u)/q}{SSE_u/df_u} \sim F(q, df_u)$$

***LR* Test**

$$LR = -2log\left((1 + qF/df_u)^{-NT/2}\right)$$

Panel Data Cross-Sectional Dependence Test

Null hypothesis: zero cross-sectional error correlations.

$$\mathrm{BP}s = \sqrt{\frac{1}{N\,(N-1)}} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(T_{ij}\hat{\rho}_{ij}^{2} - 1\right)$$

Unit Root Test

Test all time series analysis related tests

Lagrange Multiplier (LM) Tests for Cross-Sectional and Time Effects.

Tests for Serial Correlation and Cross-Sectional Effects

Wooldridge Test for the Presence of Unobserved Effects

Bera, Sosa Escudero, and Yoon Modified Rao's Score Test in the Presence of Local Misspecification.

LM Test for First-Order Correlation under Fixed Effects

Baltagi and Li Joint LM Test for Serial Correlation and Random Cross-Sectional Effects

Getting Started: PANEL Procedure

```
data greene;
   input firm year production cost @@;
datalines;
1 1955   5.36598   1.14867  1 1960   6.03787   1.45185
1 1965   6.37673   1.52257  1 1970   6.93245   1.76627
2 1955   6.54535   1.35041  2 1960   6.69827   1.71109
2 1965   7.40245   2.09519  2 1970   7.82644   2.39480
3 1955   8.07153   2.94628  3 1960   8.47679   3.25967
3 1965   8.66923   3.47952  3 1970   9.13508   3.71795
4 1955   8.64259   3.56187  4 1960   8.93748   3.93400

... more lines ...
```

$$C_{it} = \text{Intercept} + \boldsymbol{\beta} P_{it} + v_i + e_t + \epsilon_{it} \quad i = 1, \ldots, N;\ t = 1, \ldots, T$$

```
proc sort data=greene;
   by firm year;
run;
```

```
proc panel data=greene;
      model cost = production / rantwo vcomp = fb;
      id firm year;
run;
```

Figure 26.1 The Variance Components Estimates

The PANEL Procedure
Fuller and Battese Variance Components (RanTwo)

Dependent Variable: cost

Model Description	
Estimation Method	RanTwo
Number of Cross Sections	6
Time Series Length	4

Example 26.2: The Airline Cost Data: Fixtwo Model

$$TC_{it} = exp\left(\alpha_i + \gamma_t + \beta_3 LF_{it} + \epsilon_{it}\right) Q_{it}^{\beta_1} PF_{it}^{\beta_2}$$

$$\begin{aligned} ln\left(TC_{it}\right) = \alpha_N + \gamma_T \quad &+ \quad \left(\alpha_i - \alpha_N\right) + \left(\gamma_t - \gamma_T\right) \\ &+ \quad \beta_1 ln\left(Q_{it}\right) + \beta_2 ln\left(PF_{it}\right) + \beta_3 LF_{it} + \epsilon_{it} \end{aligned}$$

```
data airline;
   set airline;
   lC = log(C);
   lQ = log(Q);
   lPF = log(PF);
   label lC = "Log Transformation of Costs";
   label lQ = "Log Transformation of Quantity";
   label lPF= "Log Transformation of Price of Fuel";
run;
```

```
proc panel data=airline printfixed;
   id i t;
   model lC = lQ lPF LF / fixtwo;
run;
```

Output 26.2.1 The Airline Cost Data—Model Description

The PANEL Procedure
Fixed Two-Way Estimates

Dependent Variable: lC (Log Transformation of Costs)

Model Description	
Estimation Method	FixTwo
Number of Cross Sections	6
Time Series Length	15

Output 26.2.2 The Airline Cost Data—Fit Statistics

Fit Statistics			
SSE	0.1768	DFE	67
MSE	0.0026	Root MSE	0.0514
R-Square	0.9984		

Output 26.2.3 The Airline Cost Data—Test for Fixed Effects

F Test for No Fixed Effects			
Num DF	Den DF	F Value	Pr > F
19	67	23.10	<.0001

Output 26.2.4 The Airline Cost Data—Parameter Estimates

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
CS1	1	0.174237	0.0861	2.02	0.0470	Cross Sectional Effect 1
CS2	1	0.111412	0.0780	1.43	0.1576	Cross Sectional Effect 2
CS3	1	-0.14354	0.0519	-2.77	0.0073	Cross Sectional Effect 3
CS4	1	0.18019	0.0321	5.61	<.0001	Cross Sectional Effect 4
CS5	1	-0.04671	0.0225	-2.08	0.0415	Cross Sectional Effect 5
TS1	1	-0.69286	0.3378	-2.05	0.0442	Time Series Effect 1
TS2	1	-0.63816	0.3321	-1.92	0.0589	Time Series Effect 2
TS3	1	-0.59554	0.3294	-1.81	0.0751	Time Series Effect 3
TS4	1	-0.54192	0.3189	-1.70	0.0939	Time Series Effect 4
TS5	1	-0.47288	0.2319	-2.04	0.0454	Time Series Effect 5
TS6	1	-0.42705	0.1884	-2.27	0.0267	Time Series Effect 6
TS7	1	-0.39586	0.1733	-2.28	0.0255	Time Series Effect 7
TS8	1	-0.33972	0.1501	-2.26	0.0269	Time Series Effect 8
TS9	1	-0.2718	0.1348	-2.02	0.0478	Time Series Effect 9
TS10	1	-0.22734	0.0763	-2.98	0.0040	Time Series Effect 10
TS11	1	-0.1118	0.0319	-3.50	0.0008	Time Series Effect 11
TS12	1	-0.03366	0.0429	-0.78	0.4354	Time Series Effect 12
TS13	1	-0.01775	0.0363	-0.49	0.6261	Time Series Effect 13
TS14	1	-0.01865	0.0305	-0.61	0.5430	Time Series Effect 14
Intercept	1	12.93834	2.2181	5.83	<.0001	Intercept
lQ	1	0.817264	0.0318	25.66	<.0001	Log Transformation of Quantity
lPF	1	0.168732	0.1635	1.03	0.3057	Log Transformation of Price of Fuel
LF	1	-0.88267	0.2617	-3.37	0.0012	Load Factor (utilization index)

```
proc panel data=airline;
   id i t;
   model lC = lQ lPF LF / fixone;
run;
```

Output 26.3.2 The Airline Cost Data—Test for Fixed Effects

F Test for No Fixed Effects

Num DF	Den DF	F Value	Pr > F
5	81	57.74	<.0001

Output 26.3.3 The Airline Cost Data—Parameter Estimates

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
Intercept	1	9.79304	0.2636	37.15	<.0001	Intercept
IQ	1	0.919293	0.0299	30.76	<.0001	Log Transformation of Quantity
IPF	1	0.417492	0.0152	27.47	<.0001	Log Transformation of Price of Fuel
LF	1	-1.07044	0.2017	-5.31	<.0001	Load Factor (utilization index)

```
proc panel data=airline;
   id I T;
   model "One-Way, FB"     lC = lQ lPF lF / ranone vcomp=fb;
   model "One-Way, WK"     lC = lQ lPF lF / ranone vcomp=wk;
   model "One-Way, WH"     lC = lQ lPF lF / ranone vcomp=wh;
   model "One-Way, NL"     lC = lQ lPF lF / ranone vcomp=nl;
   model "Two-Way, FB"     lC = lQ lPF lF / rantwo vcomp=fb;
   model "Two-Way, WK"     lC = lQ lPF lF / rantwo vcomp=wk;
   model "Two-Way, WH"     lC = lQ lPF lF / rantwo vcomp=wh;
   model "Two-Way, NL"     lC = lQ lPF lF / rantwo vcomp=nl;
   model "Pooled"          lC = lQ lPF lF / pooled;
   model "Between Groups"  lC = lQ lPF lF / btwng;
   model "Between Times"   lC = lQ lPF lF / btwnt;
   compare / pstat(estimate) mstat(varcs varts vararr);
run;
```

Output 26.4.1 Parameter Estimates

The PANEL Procedure
Model Comparison

Dependent Variable: lC (Log Transformation of Costs)

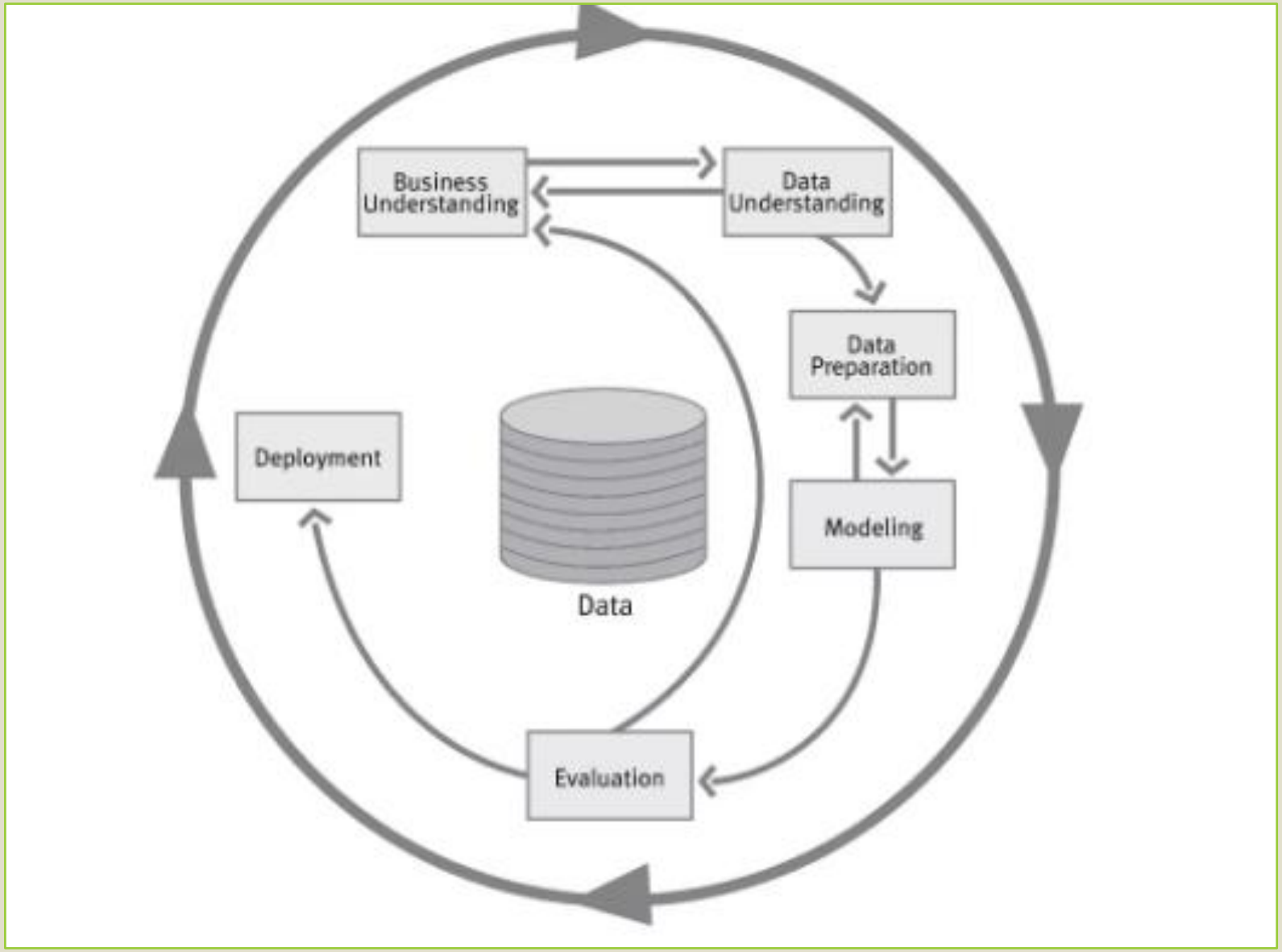
Comparison of Model Parameter Estimates

Variable		One-Way, FB RanOne	One-Way, WK RanOne	One-Way, WH RanOne	One-Way, NL RanOne	Two-Way, FB RanTwo	Two-Way, WK RanTwo
Intercept	Estimate	9.637027	9.629542	9.643869	9.640560	9.362705	9.643579
lQ	Estimate	0.908032	0.906926	0.909042	0.908554	0.866458	0.843341
lPF	Estimate	0.422199	0.422676	0.421766	0.421975	0.436160	0.409662
LF	Estimate	-1.064733	-1.064564	-1.064966	-1.064844	-0.980482	-0.926308

Comparison of Model Parameter Estimates

Variable		Two-Way, WH RanTwo	Two-Way, NL RanTwo	Pooled Pooled	Between Groups BtwGrps	Between Times BtwTime
Intercept	Estimate	9.379328	9.972603	9.516907	85.809402	11.184905
lQ	Estimate	0.869214	0.838724	0.882740	0.782455	1.133318
lPF	Estimate	0.435317	0.382904	0.453978	-5.524011	0.334268
LF	Estimate	-0.985181	-0.913357	-1.627511	-1.750949	-1.350947

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Limitations

1. There is no details about mentioned algorithms.
2. Little amount of data visualizations.
3. Little amount of mathematical derivations.

References

- Greene, William H. *Econometric analysis*. Pearson Education India, 2003.
- Wooldridge, Jeffrey M. *Econometric analysis of cross section and panel data*. MIT press, 2010.
- Others mentioned websites and lecture notes from my current study.


Thank you for
your attention