Paper 3:
ECONOMETRICS

SUMMARY NOTES

Luca F. Righetti (lfr29)

University Of Cambridge | Economics: Part IIA (2018/19)

Table of Contents

Multiple Regression Analysis ***4***

Basics **4**

Large Sample Properties 4

Interpretation 4

$\boldsymbol{R2}$- Goodness of Fit 4

Simple ($\boldsymbol{\beta i}$) vs Multiple ($\boldsymbol{\beta i}$) OLS 4

Assumptions **5**

MLR 1. Linear in Parameters 5

MLR 2. IID Random Sampling: (Independent $\boldsymbol{yi}$) 5

MLR 3. No Perfect Collinearity: $\boldsymbol{EuiujXiXj = 0}$ 5

MLR 4. Zero Conditional Mean: $\boldsymbol{Euix = 0}$ 5

MLR 5. Homoskedasticity: $\boldsymbol{Varyix}$= $\boldsymbol{Varuix = Eu2|x = \sigma 2 < \infty}$ 5

MLR 6. Normality: $\boldsymbol{u \sim N(0, \sigma 2)}$ 6

Results **6**

Implication of MLR 6

Stochastic Adjustments 6

Asymptotic Efficiency 6

Statistical Inference **7**

Multiple $\boldsymbol{\beta i}$ in single Linear Restriction 7

F-Test: Multiple Linear Restrictions 7

Relationship of test statistics $\boldsymbol{z, \chi 2, F}$ 7

Odd Bits 8

Extensions **8**

Dummy Variables 8

Interaction Terms 8

Log Form 8

Problems ***9***

Misspecification **9**

Detection 9

Heteroscedasticity **9**

Detection 9

Solutions 9

Endogeneity: Issues **10**

Omitted Variable Bias 10

Measurement Errors 10

Simultaneous Equations 10

Endogeneity: Solutions **11**

Proxy Variable 11

Instrumental Variables (IV) **11**

Set-Up 11

Properties 11

2 Stage Least Squares (2SLS) **11**

Set-Up 11

Properties 12

Detection 12

Limited Dependent Variables 12

Basic: Linear Probability Model 12

Complicated: Limited Dependent Variables 12

Maximum Likelihood Estimation 13

Interpretation 13

Time-Series Analysis 14

Basics 14

White Noise 14

Assumptions 14

Perfect Approach: Unbiasedness 14

Auto/Serial Correlation 15

Second-Best Approach: Consistency 15

Stationarity & Weakly Dependent 16

Wold's Theorem 16

The Theorem 16

The Alternative 16

Structure of Time Series 17

Static & Finite Distributed Lag Models 17

Moving Average Model 17

Autoregressive Model 17

Autoregressive and Moving Average Model 18

A Note on Unit Roots 18

Model Selection Criteria 19

Mean Squared Error & R^2 19

Akaike and Bayesian Information Criteria 19

Forecasting Cycles 19

$\boldsymbol{MA2}$: $\boldsymbol{yt = \epsilon t + \theta 1 \epsilon t - 1 + \theta 2 \epsilon t - 2}$ 19

$\boldsymbol{MAq}$: $\boldsymbol{yt = \epsilon t + \theta 1 \epsilon t - 1 + \cdots + \theta q \epsilon t - q}$ 20

$\boldsymbol{MA\infty}$: $\boldsymbol{yt = \epsilon t + b1 \epsilon t - 1 + \cdots + bh \epsilon T + bh + 1 \epsilon T - 1 + \cdots}$ 20

Non-MA 20

Putting it all Together 20

Making the Forecasts Operational 20

Recursive Estimate 20

Cross Validation 21

Random Walks 21

Motivation 21

Form 21

Test 22

Types of Forecasting 22

Conditional Forecasting 22

Unconditional Forecasting 23

Causality 23

Weak & Strong Exogeneity 23

Granger Causality 23

Dynamic Causal Effect 23

Panel Data ... ***24***

Basics ... **24**

FD & FE ... **24**

(1) First Differences ... 24

(2) Fixed Effects ... 24

FD vs. FE ... 25

Difference-In-Difference ... 25

Arellano Bond Instrument (?) ... 25

Multiple Regression Analysis

Basics

Large Sample Properties

Central Limit Theorem

- Sample mean converges in distribution to the normal distribution
- If X_i are IID with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$, then $X \sim^a N\left(\mu, \frac{\sigma^2}{n}\right)$ as $n \to \infty$
- No matter what distribution u, $\hat{\beta}_1 \sim^a N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$ when n is large. Key to statistical inference and allows us to relax MLR 6.

Law of Large Numbers

- Sample mean converges in probability to the population mean.
- If $X_1 \dots X_n \sim E(X_i) = \mu, Var(X_i) = \sigma^2$, then $\bar{X} = \frac{1}{n}\sum X_i$ where $E(\bar{X}_i) = \mu$ and $Var(\bar{X}_i) = \frac{\sigma^2}{n}$
- $X_i, i \in [1, n]$: as $n \to \infty, \bar{X} \to \mu$ (i.e. $Var(\bar{X}) \to 0$). Key to consistency!

Interpretation

- A linear model ($y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$) is estimated via regression ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$)
  - We do not estimate an OLS model. We do estimate a linear model by OLS.
  - Enables us to draw "ceteris paribus" conclusions even if data was not collected in such a fashion
  - For linear variables $\frac{\Delta y}{\Delta x_1} = \beta_1$. For quadratic variables ($\beta_1 x_1 + \beta_2 x_1^2$): $\frac{\Delta \hat{y}}{\Delta x_1} = \hat{\beta}_1 + 2\hat{\beta}_2 x_1$ etc.
- u (error) is unobservable. $\hat{u}_i$ (residual) $= y_i - \hat{y}_i$ where $\sum \hat{u}_i = 0$ and $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2$

- We choose $\hat{\beta}$ to $\min \sum_{i=1}^n \hat{u}_i{}^2 = \sum_{i=1}^n (y - \hat{y}_i)^2 = \sum_{i=1}^n \left(y - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_{ki} x_{ki}\right)^2$
  - Gives FOCs, which need a unique solution to solve:
    - $\frac{\partial}{\partial \hat{\beta}_0} = \sum_{i=1}^n \left(y - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_{ki} x_{ki}\right)^2 = 0$
    - $\frac{\partial y}{\partial \hat{\beta}_1} = \sum_{i=1}^n x_{1i}\left(y - \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_{ki} x_{ki}\right)^2 = 0$ etc.
    - $\frac{\partial y}{\partial \hat{\beta}_k} = \sum_{i=1}^n x_{ki}\left(y - \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_{ki} x_{ki}\right)^2 = 0$
- Frisch-Waugh Theorem: $\hat{\beta}_1 = \frac{Cov(x_1, y)}{Var(x_1)} = \frac{\sum_{i=1}^n \hat{r}_{i1}^2 y_i}{\sum_{i=1}^n \hat{r}_{i1}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$

R^2 - Goodness of Fit

- $R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$ where SST = SSE + SSR [i.e. $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y - \hat{y}_i)^2$]
- There are several limitations:
  - R^2 never decreases when independent variables are added. No punishment for over-specifying model even though it increases variance. Hence adjusted $\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - \frac{\hat{\sigma}^2}{SST/(n-1)}$
  - A low R^2 indicates model is hard to predict y accurately but each β_i included may still be accurate

Simple ($\tilde{\beta}_i$) vs Multiple ($\hat{\beta}_i$) OLS

- Note *Simple* $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$ vs. *Multiple* $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
  - Can be related by $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$ (where $\tilde{\delta}_1$= slope coefficient from x_{2i} regressed on x_{1i})

  - Only equivalent if $\hat{\beta}_2 = 0$ (x_2 does not matter) or $\tilde{\delta}_1 = 0$ ($\text{corr}(x_1, x_2) = 0$)
- Bias-Variance Trade-off: $\tilde{\beta}_i$ is susceptible to OVB but $\hat{\beta}_i$ has higher variance $\frac{\sigma^2}{SST_j(1-R_j^2)} > \frac{\sigma^2}{SST_j}$
  - But… var shrink with n but bias does not! Also in $\tilde{\beta}_i$, σ^2 is likely to be larger. $\hat{\beta}_i$ generally better

Assumptions

MLR 1. Linear in Parameters

<u>Is it linear?</u>

- $y = \beta_1 x_1 + \beta_2 x_1^2 + +\beta_3^2 x_3$ – Yes, since can rewrite $\beta_3^2 = \alpha_3$
- $y = \beta_1 \beta_2 x_1$– No
- $y = \beta_1 x_1 + \beta_2 (x_1 + x_2)$ – No, since can rewrite $y = (\beta_1 + \beta_2) x_1 + \beta_2 x_2$
- $y = \beta_1 \log(x_1) + \beta_2 \log(x_1)^2$ – Yes
- $y = \beta_1 \log(x_1) + \beta_2 \log(x_1^2)$ – No, since can rewrite $y = (\beta_1 + 2\beta_2) \log(x_1)$

<u>Quadratic Properties:</u>

- Let $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 +$. If $\beta_1 > 0$ y is increasing and if $\beta_2 > 0$ y is convex

MLR 2. IID Random Sampling: (Independent y_i)

MLR 3. No Perfect Collinearity: $E(u_i u_j | X_i X_j) = 0$

- i.e. $Cov(u_i u_j | X_i X_j) = E(u_i u_j | X_i X_j) - E(u_i | X_i) E(u_j | X_j) = E(u_i u_j | X_i X_j) - 0$
- Otherwise cannot estimate $\hat{\beta}_i$ via FOCs. Nothing to do with u and testable
- Can also randomly fail due to sampling (e.g. happens that $exp = 2edu$). Hence prefer large n.
- Also incalculable if there are fewer observations than variables ($n < k + 1$)

MLR 4. Zero Conditional Mean: $E(u_i | x) = 0$

- Anything not in the reg. affecting y must be constant across $\boldsymbol{x}$. Set β_0 so this then equals zero.
- Otherwise $E(\hat{\beta}_i) \neq \beta_i$. Everything to do with u and unknowable
- Many causes (see Sources of Endogeneity). Note that a procedure can be bias, an estimate cannot.

MLR 5. Homoskedasticity: $Var(y_i | x) = Var(u_i | x) = E(u^2 | x) = \sigma^2 < \infty$

- Allows us to calculate the variance of estimators. Thus get valid standard errors and test statistics
- In theory: $Var(\hat{\beta}_j) = sd^2(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}$ (Note that MLR 3. precludes $SST_j = 0$ and $R_j^2 = 1$)
  - σ^2 is feature of pop. so not dependent on n! Variance of unobservable.
  - SST_j increases via larger n. High var. in x_j means we can in turn explain a lot of var. in y.
  - R_j^2 (R^2 when x_j is regressed on $\boldsymbol{x}_{j'}$) is a measure of multi-collinearity. Note R_j^2 differs for different j and we may not care about high correlations amongst controls
- In practice: u is unobserved so…
  - Estimate σ^2 via $\hat{\sigma}^2 = \frac{SSR}{df} = \frac{\sum \hat{u}^2}{(n-k-1)}$ [trade-off: More variables decreases SSR but increases df]
  - Estimate $sd^2(\hat{\beta}_j)$ via $se^2(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{SST_j \cdot (1-R_j^2)} = \frac{\hat{\sigma}^2}{n sd^2(x_j) \cdot (1-R_j^2)}$ [larger n causes smaller $se(\quad)$]

MLR 6. Normality: $\boldsymbol{u \sim N(0, \sigma^2)}$

- Required for statistical inference. If it holds then $\hat{\beta}_j \sim N(\beta_j, \text{Var}(\beta_j))$
- Ergo $\frac{\hat{\beta}_j - \beta_j}{sd(\beta_j)} \sim \text{N}(0,1)$ and $\frac{\hat{\beta}_j - \beta_j}{se(\beta_j)} \sim \text{t}_{n-k-1}$
- Very strong assumption. Somewhat reasonable as u is the sum of unobserved factors, so we can apply CLT. Taking log can help. Not applicable in some obvious cases: cut offs (e.g. min wage) and few outputs (e.g. crimes/person). Also u-factors may not be additive

Results

Implication of MLR

- If MLR 1-4. fulfilled then OLS estimators are unbiased
- If MLR 1-5. fulfilled (Gauss-Markov) then OLS estimators are BLUE.
- If MLR 1-6. fulfilled (Classical Linear Model) then OLS estimators are BUE

Stochastic Adjustments

	Stochastic: conditional $E(\hat{\beta}_i \vert X)$	Fixed: unconditional $E(\hat{\beta}_i)$
MLR 3.	$E(u_i u_j \vert X_i X_j) = 0$	$E(u_i u_j) = 0$
MLR 4.	$E(u_i \vert X) = 0$	$E(u_i) = 0$
MLR 5.	$E(u_i^2 \vert X) = \sigma^2$	$E(u_i^2) = \sigma^2$

- What if stochastic? $\hat{\beta}_i$ can still be BLUE due to Law of Iterated Expectations $E(X) = E[E(X|Y)]$
  - $E(\hat{\beta}_i) = E[E(\hat{\beta}_i|X)] = E\left[E\left(\beta_j + \frac{\sum(X_i - \bar{X})u_i}{\sum(X_i - \bar{X})^2}|X\right)\right] = \beta_i + E\left[E\left(\frac{\sum(X-\bar{X})u_i}{\sum(X-\bar{X})^2}|X\right)\right] = \beta_i + E\left[\frac{\sum(X-\bar{X})}{\sum(X-\bar{X})^2}E(u_i|X)\right] = \beta_i$ due to MLR 3.

Asymptotic Efficiency

Consistency

- Definition: As $n \to \infty$, distribution of $\hat{\beta}_j$ collapses to 0. Although not all estimates are unbiased, consistency is seen as a minimal requirement.
- Difference between inconsistency and bias is that inconsistency is expressed in terms of population variance and population covariance, whilst bias is based on their sample counterparts

General Case for $\boldsymbol{y = \beta_0 + \beta_1 x + u}$

- $\tilde{\beta}_1 = \frac{\sum([g(x) - \overline{g(x)}]y_i)}{\sum([g(x) - \overline{g(x)}]x_i)}$
- $\tilde{\beta}_1 = \beta_1 + \frac{n^{-1}\sum([g(x) - \overline{g(x)}]u_i)}{n^{-1}\sum([g(x) - \overline{g(x)}]x_i)}$
- $\text{plim}\,\tilde{\beta}_1 = \beta_1 + \frac{Cov(g(x),u)}{Cov(g(x),x)}$ [applying LLN] $= \beta_1$
- Consistent provided $Corr(x, g(x)) = 0$ [MLR4] and $Corr(x, g(x)) \neq 0$
- For OLS $\text{plim}\,\hat{\beta}_i - \beta_i = \frac{Cov(x_i,u)}{Var(x_i)} = 0$
  - In fact we can use weaker MLR 4' $Cov(x_i, u) = 0$ [each x, as opposed to any x *function*, is uncorrelated with u]
  - Need a stronger assumption for finite sample, with MLR 4. ensuring we have properly modeled the Population Regression Function
  - Note that any correlation between u and any x causes <u>all</u> OLS estimators to be inconsistent

<u>Normality</u>

General Case for $\boldsymbol{y = \beta_0 + \beta_1 x + u}$

- $\sqrt{n}(\tilde{\beta}_j - \beta_j) \sim^a N(0, \frac{\sigma^2 Var(g(x))}{[Cov(g(x),x)]^2})$
- For OLS $\sqrt{n}(\tilde{\beta}_j - \beta_j) \sim^a N\left(0, \frac{\sigma^2 Var(x)}{[Cov(x,x)]^2}\right) = \left(0, \frac{\sigma^2}{Var(x)}\right)$
- Cauchy-Schwartz Inequality: $[Cov(g(x),x)]^2 \leq Var(g(x))Var(x)$ thus $\frac{\sigma^2}{Var(x)} \leq \frac{\sigma^2 Var(g(x))}{[Cov(g(x),x)]^2}$ (i.e. under GM assumptions, OLS has at least smallest asymptotic variance)

- $\hat{\sigma}^2$ is a consistent estimator of σ^2
- $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim^a N(0,1)$ and ergo $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim^a t_{n-k-1}$

Statistical Inference

Multiple $\boldsymbol{\beta_i}$ in single Linear Restriction

<u>Option A</u>

- If $H_0: \beta_1 - \beta_2 = 0$ then $t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$
- Note that $se(\hat{\beta}_1 - \hat{\beta}_2) \neq se(\hat{\beta}_1) - se(\hat{\beta}_2)$ but is instead difficult to estimate…

  $Var(\hat{\beta}_1 - \hat{\beta}_2) = Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov$ so $se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{se(\hat{\beta}_1)^2 + se(\hat{\beta}_2)^2 - 2Cov}$

<u>Option B</u>

- Rewrite as $y = \beta_0 + \beta_1 x_1 - \beta_2 x_1 + \beta_2 x_1 + \beta_2 x_2 = \beta_0 + (\beta_1 - \beta_2)x_1 + +\beta_2 x_1 + \beta_2 x_2$
- Let $\theta = \beta_1 - \beta_2$ and so test $H_0: \theta = 0$ with $t = \frac{\theta}{se(\theta)}$

F-Test: Multiple Linear Restrictions

- Test if jointly sig. ($H_0: \beta_1 = 0, \beta_2 = 0$) via comparing restricted (exc. $\beta_{1,2}$) and unrestricted model. Intuitively, asks if adding independent variable adds enough explanatory power to justify losing df?
- Useful if $\beta_{1,2}$ are highly correlated (e.g. different measures of intelligence). Possible none is individually sig. but jointly so. Possible one is individually sig. but not jointly so.
- $F \equiv \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} = \frac{(R_u^2 - R_r^2)/q}{(1-R_u^2)/(n-k-1)} \sim F_{q,n-k-1}$ (where $t^2 = F$ if two-sided test)
- Suppose $H_0: \beta_1 = 1, \beta_2 = 0$ then restricted model becomes $y = \beta_0 + x_1 + u$. Obtain SSR (cannot use R^2 equation because different dependent variable i.e. TSS)

Relationship of test statistics $\boldsymbol{z, \chi^2, F}$

- Notation: $t_{\text{prob,df}}$ $F_{\text{prob,numerator df,denominator df}}$
- If $z_i \sim N(0,1)$, then $v = \left[\sum_{i=1}^{k} z_i\right] \sim \chi_k^2$ (where k =d.o.f.) and $T = \frac{z}{\sqrt{v/k}} \sim t_k$
- If $v_1 \sim \chi_m^2$ and $v_2 \sim \chi_n^2$, then $\frac{v_1/m}{v_2/n} \sim F_{m,n}$:
- χ^2 is like a N, with left side flipped onto right and then squared (i.e. a-symmetrical)

Odd Bits

- p-value: smallest sig-level at which H_0 would be rejected
- Rise in n → Fall in $se()$ → Rise in t. Hence use smaller sig-level for larger samples.
- Statistical significance: t is large. Economic significance: $|\beta|$ is large.
- Heteroskedasticity causes wrong $se(\hat{\beta}_i)$. Too small means we reject too often., too large too seldom.
- Confidence Intervals [$\overline{\underline{\beta}} = \hat{\beta} \pm c \cdot se(\beta)$] are only as good as the underlying regression

Extensions

Dummy Variables

- Dummy Variable: β_i can only take on $0{,}1$. Interpreted as an intercept shift
- If we use two dummy variables we get perfect collinearity
  - $y = \beta_0 + \beta_1 male + \beta_2 female + u$ – No
  - $y = \beta_1 male + \beta_2 female + u$ – Yes

Using Multiple Dummies: g groups require $g - 1$ dummy variables

- E.g. $y = \beta_0 + \beta_1 married\_male + \beta_2 married\_fem + \beta_1 singe\_male + \beta_2 single\_fem + u$
- Else assumes marriage effect is same for genders! E.g. $y = \beta_0 + \beta_1 female + \beta_2 married + u.$

Ordinal Information: AAA (3) BBB (2) Junk (1). Assumes a one-unit increase has constant effect. If not multiple dummies

Interaction Terms

- Allows for different slopes. E.g. $y = \beta_0 + \delta_0 fem + \beta_1 educ + \delta_1 fem \cdot educ + u$
- If $\delta_1 > 0$ complements. If $\delta_1 < 0$ subsitutes.

Chow Test

- With few variables, F-test for interaction terms is feasible. If not use Chow Test…
- $y = \beta_{g,0} + \beta_{g,1} x_1 + \beta_{g,2} x_2 + \cdots + u$ for $g = 1{,}2$ and H_0: $\beta_{1,i} = \beta_{2,i}$ for all i
- Chow Statistic: $F = \frac{SSR_P - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \times \frac{n - 2(k+1)}{k+1}$
  - SSR_1 for $g = 1$ (n_1 observations)
  - SSR_2 for $g = 2$ (n_2 observations)
  - SSR_P for $g = 1{,}2$ (n observations)
  - Unrestricted regression has $n - 2(k+1)$ degrees of freedom
- Notes: Requires homoscedasticity but not normality; R^2 form is only applicable if interactions have been included to create unrestricted model
- What if we just want test for slope differences? Test joint sig. of interaction terms only (exc. β_0). Note that SSR_P obtained contains only intercept shift so k (not $k+1$) restrictions

Log Form

Model	Regression	Interpretation of β_1
Level-level	y on x	x up 1, then y up β_1
Level-log	y on log(x)	x up 1%, then y up (β_1/100)
Log-level	log(y) on x	x up 1, then y up ≈(100β_1)% (y actually up (100(e^{β_1}-1))%
Log-log	log(y) on log(x)	Elasticity equal to β_1 x up 1%, then y up (β_1)%

- Log model is less sensitive to outliers and less heteroskedastic, but prediction becomes harder…
  - $\ln y = \beta_0 + \beta_1 x_1 + u$ can be rewritten as $y = e^{\beta_0+\beta_1 x_1+u} = e^{\beta_0+\beta_1 x_1} \cdot e^{u}$
  - If $E(u) = 0$ then $E(e^u) \neq 0$ but instead $E(e^u) = e^{\frac{\sigma^2}{2}}$.
  - So need to adjust model… as $y = e^{\hat{\beta}_0+\hat{\beta}_1 x_1} \cdot e^{\frac{SSR}{(n-2)2}}$

Problems

Misspecification

Detection

<u>Ramsay's Reset</u>

- Should I add quadratics (x_1^2, x_1x_2, etc.)? Testing each is a lot of work so here is a short cut… (but note it cannot detect OVB or distinguish diminishing quadratic from logarithmic)

1. Let $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + u$. Regress to calculate $\hat{\alpha}_i$
2. Define $\hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \hat{\alpha}_3 x_3 + u$; $\hat{y}_i^2 = (\quad)^2$; $\hat{y}_i^3 = (\quad)^3$ etc.
3. Regress $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \delta_1 \hat{y}_i^2 + \delta_2 \hat{y}_i^3$ etc.
4. F-Test H_0: $\delta_1 = 0$, $\delta_2 = 0$ etc. If do not reject then do not need to include quadratics.

<u>Mizon-Richards Test</u>

- Non-nested (e.g. linear $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ vs. log $y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + u$)

1. Estimate $y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log x_1 + \gamma_4 \log x_2 + u$
2. F-Test H_0: γ_3, γ_4. If do not reject then the level-log model must be incorrect.

Heteroscedasticity

Detection

Breusch-Pagan-Godfrey Test: Looks for linear relationship

1. Run OLS regression and obtain $\hat{u}_i = y - \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
2. Run new regression $\hat{u}_i^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + v_i$
3. F-Test $\alpha_1 = 0, \alpha_2 = 0$. If rejected, then heteroskedastic

Augmented White Test: Looks for higher order polynomial relationships

1. Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ and so define $\hat{y}_i^2$
2. Run OLS regression $\hat{u}_i^2 = \alpha_0 + \alpha_1 \hat{y}_i + \alpha_2 \hat{y}_i^2$
3. F-Test $\alpha_1 = 0, \alpha_2 = 0$. If rejected, then heteroskedastic.

Solutions

<u>Remove from unknown form: Robust Standard Errors</u>

- Asymptotically $\hat{\beta}_j = \beta_j + \frac{\sum(x_i-\bar{x})u_i}{\sum(x_i-\bar{x})^2}$ so $sd^2(\hat{\beta}_j) = \frac{\sum(x_i-\bar{x})^2\sigma_i^2}{SST_j}$ and $se^2(\hat{\beta}_j) = \frac{\sum(x_i-\bar{x})^2\hat{u}_i{}^2}{SST_j}$

<u>Remove from known form: Weighted Least Squares</u>

- E.g. if know $Var(u|x_i) = \sigma^2 x_i$ then $Var\left(\frac{u}{\sqrt{x_i}}|x_i\right) = \sigma^2$. Hence divide all series by roo of series

Endogeneity: Issues

Omitted Variable Bias

- Let True Model be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ and False Model be $y = \alpha_0 + \alpha_1 x_1 + v$
- $E(\alpha_1) = \frac{Cov(y,x_1)}{Var(x_1)} = \frac{Cov(\beta_0, x_1)+\beta_1 Cov(x_1, x_1)+\beta_2 Cov(x_2, x_1)+Cov(u, x_1)}{Var(x_1)} = \frac{\beta_1 Cov(x_1, x_1)+\beta_2 Cov(x_2, x_1)}{Var(x_1)} = \beta_1 + \beta_2 \frac{Cov(x_2, x_1)}{Var(x_1)} = \beta_1 + \beta_2 \frac{\sum(x_1 - \bar{x})x_2}{\sum[(x_1-\bar{x})^2]} = \beta_1 + \beta_2 \tilde{\delta}_1$
- Do not know size of $\tilde{\delta}_1$ but can intuitively guess sign. Hence know if bias is upwards or downwards.
    - If correlation between x_1, x_2 and x_1, y is the same direction, bias is +. If opposite, -
    - Can only sign the bias for the more general case if all of the included xs are uncorrelated

Measurement Errors

- Can account for systematic (e.g. pretend to be normal so reversion to the mean) but not random
- Likewise for missing data (and then deciding on choosing a restricted sample).

Measurement Error in y

- Let $y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ and $y = y^* + e$
- Hence $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (e + u)$
    - Unbiased if $Cov(y^*, e) = 0$ unless $E(e) \neq 0$
    - Even if unbiased, larger standard errors as $Var(u + e) > Var(u)$

Measurement Error in x

- Let $y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u$ and $x_1 = x_1^* + e$ where
- Hence $y = \beta_0 + \beta_1(x_1 - e) + \beta_2 x_2 + u = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e)$
    - Unbiased if $Cov(x_1, e) = 0$ but variance remains larger
    - Attenuation bias (i.e. towards 0) and inconsistent if $Cov(x_1^*, e) = 0$ [since $Cov(x_1, e) = \sigma_e^2$]
        - $E(\hat{\beta}_1) = \beta_1 + \frac{Cov(x_1, u-\beta_1 e)}{Var(x_1)} = \beta_1 - \beta_1 \frac{Var(e)}{Var(x_1^*)+Var(e)} = \beta_1 \left[\frac{Var(x_1^*)}{Var(x_1^*)+Var(e)}\right]$
        - An insignificant result could always be because measurement error!
- Note $Cov(x_1, u - \beta_1 e) = Cov(x_1^* + e, u - \beta_1 e) = Cov(x_1^*, u) - \beta_1 Cov(e, x_1^*) + Cov(e, u) - \beta_1 Cov(e, e) = \beta_1 Cov(e, e) = \beta_1 Var(e) \neq 0$ even if $Cov(x_1^*, e) = 0$ and e is random

Simultaneous Equations

- E.g. $\hat{\alpha}_1$ for $cr = \alpha_0 + \alpha_1 pol + \alpha_2 tax + u$ and $pol = \beta_0 + \beta_1 cr + \beta_2 tax + v$
- Each equation should have a causal interpretation in isolation of the other. Just because two variables are determined simultaneously does not mean SEM is suited!
- The key identification condition is that each explanatory variable is uncorrelated with the error term. This does not fold for SEM:
    - Sub: $pol = \beta_0 + \beta_1 \alpha_0 + \alpha_1 \beta_1 pol + (\beta_2 + \beta_1 \alpha_2) tax + \beta_1 u + v = \frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\beta_2 + \beta_1 \alpha_2}{1 - \alpha_1 \beta_1} tax + \frac{\beta_1 u + v}{1 - \alpha_1 \beta_1}$
    - Note pol is function of u and v, thus $cov(pol, u) \neq 0$ thus endogeneity problem thus OLS estimator $\hat{\alpha}_1$ will be inconsistent
- Rank Condition: The first equation in a two-equation SEM is identified iff the second equation contains at least on exogenous variable that is excluded from the first equation.
    - Thus need $\beta_2 \neq 0$ (relevance) and $\alpha_2 = 0$ (exclusion) to use tax as IV for pol in equation (1)
    - tax shifts supply without affecting demand. Given tax variation we can trace out demand curve (i.e. use tax as IV)

Endogeneity: Solutions

Proxy Variable

- E.g. $wage = \beta_0 + \beta_1 edu + \beta_2 IQ + u$. But most people don't know IQ.
- $IQ = \alpha_0 + \alpha_1 Test + v$ (assuming $\alpha_1 \neq 0$) so…
- $wage = \beta_0 + \beta_1 edu + \beta_2(\alpha_0 + \alpha_1 Test + v) + u = (\beta_0 + \beta_2\alpha_0) + \beta_1 edu + \beta_2\alpha_1 Test + (\beta_2 v + u)$
- If $Cov(edu, u) = Cov(Test, u) = Cov(Test, v) = 0$ [usual OLS] and $Cov(edu, v) = 0$ [IQ no correlation with educ once Test is partialled out; testable] then $\hat{\beta}_1$ is unbiased and consistent.
    - Even if not, including a proxy may still be better than Omitted Variable Bias.
- Note we only care about β_1. Holds even if Test is an imperfect measure of IQ!

Instrumental Variables (IV)

Set-Up

- $y = \beta_0 + \beta_1 x_1 + u$ assumes $Cov(x, u) = 0$. IV offers a way how to deal when this is not the case. Can also use IV for measurement error if uncorrelated with e
- Require a consistent Instrument z that satisfies
    - Exogeneity: $Cov(z, u) = 0$ [untestable]
        - $Cov(z, y) = \beta_1 Cov(z, x) + Cov(z, u) = \beta_1 Cov(z, x) + 0$
        - Hence $\beta_1 = \frac{Cov(z,y)}{Cov(z,x)}$ and $\hat{\beta}_1 = \frac{\sum(z-\bar{z})(y-\bar{y})}{\sum(z-\bar{z})(x-\bar{x})}$. Note IV is biased. Thus we need a large sample.
        - Intuitively, z should only influence y through x. If added to original reg. it is insig Thus, a variable that is a good proxy is a bad IV, and vice versa!
    - Relevance: $Cov(z, x) \neq 0$ [testable as $x = \pi_0 + \pi_1 z + v$ H0: $\pi_1 = 0$]
        - Stock-Staiger: Correlation shrinks to zero at $1/\sqrt{n}$, so keep sample size in mind!

Properties

- Requires stronger homoskedasticity assumption: $Var(u|z) = Var(u) = \sigma^2$
- [Asymptotic] $Var(\hat{\beta}_{i,IV}) = \frac{\sigma^2}{n\sigma_x^2 p_{x,z}^2} = \frac{Var(\beta_{OLS})}{R_{x,z}^2}$ decreasing to 0 at $\frac{1}{n}$ and $se(\hat{\beta}_{i,IV})^2 : \frac{\hat{\sigma}^2}{SST_x R_{x,z}^2}$
    - Note that it is always larger than OLS!
- $\text{plim}\, \hat{\beta}_{i,IV} = \beta_i + \frac{Corr(z,u)}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$ so even small $Corr(z, x)$ creates big inconsistency if $Corr(z, u) \neq 0$
    - It could be even worse than OLS, $\text{plim}\, \hat{\beta}_{i,OLS} = \beta_i + Corr(x, u) \cdot \frac{\sigma_u}{\sigma_x}$, if $\frac{Corr(z,u)}{Corr(z,x)} > Corr(x, u$

2 Stage Least Squares (2SLS)

Set-Up

- IV with multiple z is called Two-Stage-Least Squares (2SLS)
- Order Condition: Need at least as many exogenous as endogenous variables in structural equation
- Let $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u$ where z_1 is exog. and y_2 is endog. determined by z_2, z_3
- Each z_i is uncorrelated with u, so any linear combination is also uncorrelated. Choose best(*) fit IV:
    - $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v$
    - $y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$ where π_2 or $\pi_3 \neq 0$ (so z_1 is not perfectly correlated to y_2^*)
- If First Stage has low R^2, either exog. do no explain endog. much OR we have a weak instrument
- We can think of 2SLS as purging y_2 of $Corr(y_2, u)$:
    - $y_2 = y_2^* + v$ hence $y_1 = \beta_0 + \beta_1 y_2^* + \beta_2 z_1 + [u + \beta_1 v_2]$

Properties

- To be consistent $Corr(z_j, u) = 0$ and at least one z is corr with y_2
- To have asymptotically valid t, F stats we require $Var(z_j, u) = k$
  - Because of the nature of the regression, R^2 is negative so no guarantee of positive F-statistic.
- [Asymptotic] $se(\beta_1)^2 \approx \frac{\sigma^2}{\widehat{SST}_2\left(1-\hat{R}_2{}^2\right)}$. High multicollinearity causes large se. Large samples offset this
- Stock and Yogo: Don't use 5% level to test if z is relevant, use $\sqrt{10}$ for t and 10 for F.

Detection

<u>Wu-Hausman Test</u>

- Estimate reduced form y_2 by regressing it on all exogenous variables: $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v$
- Obtain residuals $\hat{v}_2$ and add to structural equation: $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v} + error$
- y_2 is only endogenous if the residuals are significant. Thus test H_0: $\delta_1 = 0$
- Note: estimate of β_j does not change (use to check answer); adding $\hat{v}_2$ purges endog. of y_2

<u>Hausman Test</u>

- If y_2 is endog and zs is a valid IV then $\beta_{2SLS} \rightarrow \beta; \beta_{OLS} \nrightarrow \beta$

  $F = \frac{\left[\hat{\beta}_{2SLS}-\hat{\beta}_{OLS}\right]^2}{Var(\hat{\beta}_{2SLS}-\hat{\beta}_{OLS})}$. If H_0 cannot be rejected then β_{2SLS} and β_{OLS} converge to the same limit.

  Either all zs are valid or both are invalid instruments

<u>Testing Overidentifying Restrictions</u>

- If we have multiple instruments, it is possible to test overidentifying restrictions (i.e. to see if some of the instruments are correlated with the error) (i.e. test if all IVs give the same coefficients)
  - Estimate the structural model using 2SLS and obtain the residuals $\hat{u}_i$
  - Regress the residuals on all the exogenous variables: $\hat{u}_i = \beta_0 + \beta_1 z_1 + \delta_1 z_2 + \delta_2 z_3 + error$
  - Obtain R^2 (i.e. F-test H_0: $\delta_i = 0\ \forall i$) and form $J = mR^2$ (where n is m instruments)
  - Test H_0: $J \sim \chi^2_{m-q}$ (where $m - q$ is the #instruments - #endog. Regressors).
    - If reject then at least one instrument is not exogenous

Limited Dependent Variables

Basic: Linear Probability Model

- $P(y = 1|x) = E(y|x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$
- Since $y < 0, y > 1$ define $\tilde{y} = \begin{cases} 0\ if\ \hat{y} < 0.5 \\ 1\ if\ \hat{y} > 0.5 \end{cases}$. Compare $\tilde{y}$ with y to get % correctly predicted.
- Note: IRL change may not be linear (e.g. first vs. second kid effect)
- Note LPM violates homoskedasticity since $Var(y|x) = p(x)[1 - p(x)]$. Hence no t, F tests!

Complicated: Limited Dependent Variables

- $P(y = 1 \mid \boldsymbol{x}) = G(\beta_0 + \boldsymbol{x\beta})$ where $0 < G(\zeta) < 1$
  - <u>Logit</u>: $\Lambda(\zeta) = \frac{e^{\zeta}}{1+e^{\zeta}}$ (i.e. cumulative density function)
  - <u>Probit</u>: $\Phi(\zeta) = \int_{-\infty}^{\zeta} (2\pi)^{-\frac{1}{2}} e^{-\frac{\zeta^2}{2}}\, dv$ (i.e. standard normal density)
- $P(y = 1 \mid \boldsymbol{x}) = P(y^* > 0|\boldsymbol{x}) = P(u > -[\beta_0 + \boldsymbol{x\beta}]|\boldsymbol{x}) = 1 - G(-[\beta_0 + \boldsymbol{x\beta}]) = G(\beta_0 + \boldsymbol{x\beta})$
  - Derived from an underlying latent variable model $y^* = \beta_0 + \boldsymbol{x\beta} + u$ where $y = 1[y^* > 0]$

(i.e. $y = 1$ if $y^* > 0$ and $y = 0$ if $y^* \leq 0$)

  - Assume u is independent and symmetrically distributed around zero. Thus $1 - G(-\zeta) = G(\zeta)$
- Note: Heteroscedasticity seen in LPM not applicable (MLE is based on the distribution of *y* given the *xi*) and assume **errors are distributed according to the pdf of the probability transformation**

Maximum Likelihood Estimation

- Cannot use OLS because of nonlinear nature of $E(y|\boldsymbol{x})$. Instead…
- MLE used in Bayesian Inference, and estimates the parameter β_1 that is most likely to be the parameter that gave the sample data. Maximizes the log-likelihood function:
  - $f(y|\boldsymbol{x_i}; \boldsymbol{\beta}) = [G(\boldsymbol{x_i\beta})]^y[1 - G(\boldsymbol{x_i\beta})]^{1-y}$ (where $y = 0{,}1$)
  - $\ell(\boldsymbol{\beta}) = y_i \log[G(\boldsymbol{x_i\beta})] + (1 - y_i) \log[1 - G(\boldsymbol{x_i\beta})]$
  - $\widehat{\boldsymbol{\beta}}$ s.t. $\max \mathcal{L}(\boldsymbol{\beta}) = \sum \ell_i(\boldsymbol{\beta})$

Likelihood Ratio Test

- $\text{LR} = 2(\mathcal{L}_{\text{ur}} - \mathcal{L}_{\text{r}}) \sim a\chi_q^2$ (multiply by 2 so LR has approx.. chi-square distribution under H0)
- Same intuition as F-test: Is the fall in the log-likelihood large enough to conclude that the dropped variables are important?

% correctly predicted

- Let $\tilde{y}_i = \begin{cases} 0 \; if \; G(\hat{\beta}_0 + \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}) < 0.5 \\ 1 \; if \; (\hat{\beta}_0 + \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}) \geq 0.5 \end{cases}$. Hence % correctly predicted is the % $y = \tilde{y}_i$
- But can be misleading if the poorly predicted outcome is rare in sample

Pseudo R-Sqaured: McFadden: $1 - \frac{\mathcal{L}_{\text{ur}}}{\mathcal{L}_0}$

Interpretation

- Direction of effect: $E(y|\boldsymbol{x}) = G(\beta_0 + \boldsymbol{x\beta})$ and $E(y^*|\boldsymbol{x}) = \beta_0 + \boldsymbol{x\beta}$ are same
- Magnitude of effect: y^* rarely has a well-defined unit of measurement so magnitude of β not useful
- Want to estimate effect of x on $P(\quad)$ but this is complicated by nonlinear nature of $G(\quad)$
  - e.g. $P(y = 1|\boldsymbol{z}) = G(\beta_0 + \beta_1 z_1 + \beta_2 z_1^2)$ has $\frac{dP(y = 1|\boldsymbol{z})}{dz_1} = g(\beta_0 + \boldsymbol{x\beta})(\beta_1 + 2\beta_2 z_1)$
  - $\Delta\widehat{\text{P}}(y = 1|\boldsymbol{x}) \approx [g(\hat{\beta}_0 + \boldsymbol{x}\widehat{\boldsymbol{\beta}})\hat{\beta}_j]\Delta x_j$
- Differences: (useful for discrete variables) e.g. $G(\beta_0 + \beta_1 + \beta_2 x_2) - G(\beta_0 + \beta_2 x_2)$
- Partial Effects: $P(y = 1|x) = \frac{\partial G(\beta_0 + X\beta)}{\partial x_j} = \beta_j g(\beta_0 + X\beta)$ [continuous] OR $G[\beta_0 + \beta_1 x_{1_.} + \cdots \beta_j(\alpha + 1) + \cdots + \beta_k x_k] - G[\beta_0 + \beta_1 x_{1_.} + \cdots \beta_j \alpha + \cdots + \beta_k x_k]$ [discrete]
  - As $g > 0$ β_j determines the sign of the partial effect
  - Note this is not marginal since change in x_k from c_k to $c_k + 1$ may not be small
  - $g(\hat{\beta}_0 + \boldsymbol{x}\widehat{\boldsymbol{\beta}})\hat{\beta}_j$ depends on $\boldsymbol{x}$: the partial effect of one variable depends on all of the variables
  - (i) replace variable with sample average: Partial Effect at the Average: $g(\hat{\beta}_0 + \bar{\boldsymbol{x}}\widehat{\boldsymbol{\beta}})\hat{\beta}_j$
  - (ii) average the partial effects across the sample: Average Partial Effect: $[n^{-1}\sum g(\hat{\beta}_0 + \boldsymbol{x_i}\widehat{\boldsymbol{\beta}})]\hat{\beta}_j$

Time-Series Analysis

Basics

- @ time t: $y_t = \beta_0 + \beta_1 TIME_t + \epsilon_t$
- @ h periods after T: $y_{T+h} = \beta_0 + \beta_1 TIME_{T+h} + \epsilon_{T+h}$
- Feasible Point Forecast: $\hat{y}_{T+h,T} = \hat{\beta}_0 + \hat{\beta}_1 TIME_{T+h}$
- 95% Interval Forecast: $[\hat{y}_{T+h,T} - 1.96\hat{\sigma}, \hat{y}_{T+h,T} + 1.96\hat{\sigma}]$ where $\hat{\sigma}$ is Standard Error

- OLS (reg): $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \arg\min \sum [y_t - \beta_0 - \beta_1 TIME_t - \beta_2 TIME_t^2]^2$
- NLS (nl): $(\hat{\beta}_0, \hat{\beta}_1) = \arg\min \sum [y_t - \beta_0 \exp\{\beta_1 TIME_t\}]^2$ or $[\ln y_t - \ln \beta_0 - \beta_1 TIME_t]^2$

- Seasonal Dummy: $y_t = \beta_1 TIME_t + \sum_{i=1}^{s} \gamma_i D_{it} + \sum_{i=1}^{s} \delta_i HDV_{it} + +\epsilon_t$
  - Never include seasonal dummies and intercept (either s-1 and intercept or s and none)
- Cycles: Any sort of dynamics not captured by trends or seasonal that is not perfectly predictable
  - In theory follow an infinite time series $\ldots y_{-2}, y_{-1}, y_0, y_1, y_2, \ldots$ In practice observe a sample path (finite subset of a realization) $y_1, \ldots, y_T$.
- Lag Operator: L transforms $\{\ldots y_{-1}, y_0, y_1, \ldots\}$ into $\{\ldots\ y_{-2}, y_1, y_0, \ldots\}$ or generally $L^k y_t = y_{t-k}$
  - Can be combined into lag polynomials: $B(L) = b_0 + b_1 L + \cdots + b_m L^m \ldots$
  - … which gives Distributed Lag: $B(L)y_t = b_0 y_t + b_1 L y_{t-1} + \cdots + b_m L^m y_{t-m}$
- Stochastic Process: A sequence of random variables

White Noise

- Zero mean, constant variance, no serial correlation: $y_t = \epsilon_t \sim (0, \sigma^2)$ where $\sigma^2 < \infty$
  - If y is serially independent then $y_t \sim \text{iid}(0, \sigma^2)$ (i.e. independent WN)
  - If y is serially independent and normally distributed $y_t \sim \text{iid } N(0, \sigma^2)$ (i.e. Gaussian WN)
- $\gamma(\tau) = \begin{cases} \sigma^2 \text{ if } \tau = 0 \\ 0 \text{ if } \tau \geq 1 \end{cases}$ | $\rho(\tau) = \begin{cases} 1 \text{ if } \tau = 0 \\ 0 \text{ if } \tau \geq 1 \end{cases}$ | $p(\tau) = \begin{cases} 1 \text{ if } \tau = 0 \\ 0 \text{ if } \tau \geq 1 \end{cases}$
- For Independent White Noise:
  - $E[y_{T+h} | \Omega_T] = 0$
  - $Var(y_{T+h} | \Omega_T) = E\left[\left(y_{T+h} - E(y_{T+h} | \Omega_T)\right)^2 | \Omega_T\right] = \sigma^2$
  - Identical to unconditional since there are no dynamics
- 1-s-a forecast errors should be WN since else serially correlated, thus forecastable, thus improvable

Assumptions

Perfect Approach: Unbiasedness

- We still require GM assumptions. Random sampling is dropped. We add **no serial correlation** and zero conditional mean is strengthened **from contemporaneously to strict exogeneity**.

$$\left.\begin{array}{l} \circ\ \text{Model is linear in parameters} \\ \circ\ E(u_t|\boldsymbol{X_t}) = 0 \\ \circ\ \text{No perfect multicollinearity} \end{array}\right\} \text{Estimates are Unbiased}$$

$$\left.\begin{array}{l} \circ\ Var(u|x) = \sigma^2 \\ \circ\ \boldsymbol{Corr(u_t, u_s|x) = 0} \end{array}\right\} \text{Estimates are BLUE}$$

$$\left.\begin{array}{l} \circ\ \text{Estimators normally distributed} \end{array}\right\} \text{Standard inference is valid}$$

Auto/Serial Correlation

- Auto-Correlation Function: $\rho(\tau) = \frac{cov(y_t, y_{t-\tau})}{\sqrt{var(y_t)}\sqrt{var(y_{t-\tau})}} = \frac{E[(y_t-\mu)(y_{t-\tau}-\mu)]}{E[(y_t-\mu)^2]} = \frac{\gamma(\tau)}{\gamma(0)}$
    - Simple correlation between y_t and $y_{t-\tau}$. (Note $\rho(0) \equiv 1$)
    - In sample: $\hat{\rho}(\tau) = \frac{\frac{1}{T}\sum_{t=\tau+1}^{T}[(y_t-\bar{y})(y_{t-\tau}-\bar{y})]}{\frac{1}{T}\sum_{t=\tau+1}^{T}[(y_t-\bar{y})^2]}$
    - If series is white noise then asymptotically as $T \to \infty$ for any τ $\hat{\rho}(\tau) \sim N\left(0, \frac{1}{T}\right)$
- Partial Auto-Correlation: $p(\tau)$ =coefficient on $y_{t-\tau}$ in pop. linear regression of y_t on $y_{t-1} \dots y_{t-\tau}$
    - Correlation between y_t and $y_{t-\tau}$ after controlling for the effects of $y_{t-1}, \dots, y_{t-\tau+1}$
    - In sample: $\hat{p}(\tau) = \hat{\beta}_\tau$on $y_{t-\tau}$ in reg. of y_t on constant $y_{t-1}, \dots, y_{t-\tau}$

- Consider model $y_t = \beta x_t + \epsilon_t$ when $\epsilon_t = u_t + \theta u_{t-1}$
- No longer $Var(\hat{\beta}_{OLS}) = \frac{\sigma^2}{SST_x}$ but instead $Var(\hat{\beta}_{OLS}) = \frac{\sigma^2}{SST_x} + \frac{2\sigma^2}{(SST_x)^2}\sum_{t=1}^{n-1}\sum_{j=1}^{n-t}\theta^j x_t x_{t+j}$
    - Most econ variables have ve+ serial correlation, so OLS variance estimators underestimates
- Do not cause bias here but can do in some cases. E.g. **lagged dependent variable** model ($y_t = \beta_0 + \beta_1 y_{t-1} + u_t$) where $Corr(u_t|y_{t-2}) \neq 0$ and $u_t = \phi u_{t-1} + \epsilon_t$ results in $E(u_t|y_{t-1}) \neq 0$
    - To solve we need to re-specify the model: $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + u_t$

Tests

- Durbin-Watson Test: Note that this can only detect for first order serial correlation
    - Suppose $y_t = \beta_0 + \sum_{i=1}^{k}\beta_i x_{it} + \epsilon_t$ where $\epsilon_t = \phi\epsilon_{t-1} + v_t$ and v_t~iid $N(0, \sigma^2)$. ϵ_t is non-serially correlated iff $\phi = 0$.
    - Test via $DW = \frac{\sum_{t=2}^{T}(\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{T}\hat{\epsilon}_t^2} \approx 2(1-\hat{\rho}(1))$ and H$_0$: $\phi = 0$. Reject in favor of positive serial if $DW < 2$. Reject in favor of negative serial if $DW > 2$. Do not reject if $DW \approx 2$.
- Breusch-Godfrey Test:
    - Run regression and obtain estimates for residuals, $\hat{u}_t$.
    - Run new regression $\hat{u}_t = \theta_0 + \theta_1\hat{u}_{t-1} + \theta_2\hat{u}_{t-2} + \cdots + e_t$
    - t/F-test H$_0$: $\theta_i = 0$. If do not reject there is no serial correlation
    - Note, if we believe that one of the regressors is endogenous, just include in equation
- Box-Pierce Q-Statistic:
    - Tests whether all of the autocorrelations are jointly zero i.e. H$_0$: y is white noise
    - $\hat{\rho}(\tau) \sim N\left(0, \frac{1}{T}\right)$ hence $T\hat{\rho}(\tau)^2 \sim \chi^2(1)$ and is asymptotically indifferent for different τ
    - $Q_{BP} = T\sum_{\tau=1}^{m}\hat{\rho}(\tau)^2$ is asymptotically $\chi^2(m)$ under H$_0$
- Ljung-Box Q Statistic:
    - Like Q_{BP} but distributed closer to $\chi^2(m)$ in finite samples (Value of $m \approx \sqrt{T}$ is reasonable)
    - $Q_{LB} = T(T+2)\sum_{\tau=1}^{m}\left(\frac{1}{T-\tau}\right)\hat{\rho}(\tau)^2$

Second-Best Approach: Consistency

- IRL strict exogeneity $E(u_t|X_t) = 0$ fails. Instead, use weaker assumptions to get consistency. Revert back to **contemporaneously exogeneity**, assume **stationarity** and **weakly dependent process**.
    - Model is linear in parameters
    - Series are stationary and weakly dependent
    - $E(u_t|x_t) = 0$
    - No perfect multicollinearity

  Estimates are Consistent

  - $Var(u|x_t) = \sigma^2$
  - $Corr(u_t, u_s|x_t, x_s) = 0$

  Estimates are Asymptotically Normally Distributed (for standard inference)

- Intuitively, fixing distribution across t (stationarity) allows inferences with weaker assumptions. To replace random sampling we also need distribution to have certain characteristics (weakly dependent)

Stationarity & Weakly Dependent

- <u>(Strict) Stationary:</u> $f(y_{t_1}, y_{t_2}, \dots, y_{t_n})$ is identical to $f(y_{t_{h+1}}, y_{t_{h+2}}, \dots, y_{t_{h+n}})$ for all h and n. Sometimes this assumption is stronger than necessary, hence…
  - Note: If joint distribution is normal then it is fully described by $\mu_t = E(y_t)$ and $\gamma(t,\tau) = Cov(y_t, y_{t-\tau}) = E[(y_t - \mu_t)(y_{t-\tau} - \mu_{t-\tau})]$
- <u>Weak/Covariance Stationarity:</u> Only concerns itself with being stationary up to the second moment
  - Mean is stable over time $E(y_t) = \mu$.
  - Autocovariance depends only on displacement τ: $Cov(y_t, y_{t-\tau}) = \gamma(\tau)$
    - Note that $\gamma(\tau) = \gamma(-\tau)$ and $\gamma(0) = Var(y_t)$
    - If only second condition is fulfilled, the process is weakly stationary
- Note: Weak stationarity (μ and Cov independent of time) + normality (entirely characterized my μ and Var) = Strict stationarity (entire unconditional joint distribution is stationary over time)
- <u>Weakly Dependent:</u> y_{t+h} and y_{t+h} are asymptotically uncorrelated: $\lim_{h\to\infty} Corr(y_t, y_{t+h}) = 0$

Wold's Theorem

The Theorem

- Let y_t be any zero-mean covariance stationary process that does not contain deterministic components. Then we can write it as $y_t = B(L)\epsilon_t = \sum_{i=0}^{\infty} b_i \epsilon_{t-i}$ where $b_0 = 1$ and $\sum_{i=0}^{\infty} {b_i}^2 <$
- I.e. that the General Linear Process ($y_t = \sum_{i=0}^{\infty} b_i \epsilon_{t-i}$) is sufficiently flexible to represent any zero mean covariance stationarity process.
- Note that since ϵ_t are serially uncorrelated, $var(y_t) = \sum_{i=0}^{\infty} b_i^2 var(\epsilon_{t-i}) = \sigma^2 \sum_{i=0}^{\infty} b_i^2$

<u>Mean and Variance</u>

- $E(y_t) = E(\sum b_i \epsilon_{t-i}) = \sum b_i E(\epsilon_{t-i}) = 0$
- $E(y_t|\Omega_{t-1}) = E(\sum_{i=0}^{\infty} b_i \epsilon_{t-i} \,|\Omega_{t-1}) = 0 + b_1\epsilon_{t-1} + b_2\epsilon_{t-2} + \cdots = +\sum_{i=0}^{\infty} b_i \epsilon_{t-i}$
  - Note conditional mean moves over time in response to the evolving information set
- $Var(y_t) = Var(\sum b_i \epsilon_{t-i}) = \sum b_i^2 Var(\epsilon_{t-i}) = \sigma^2 \sum b_i^2$
- $Var(y_t|\Omega_{t-1}) = E((y_t - E(y_t|\Omega_{t-1}))^2|\Omega_{t-1}) = E(\epsilon_t^2|\Omega_{t-1}) = \sigma^2$
- Where $\Omega_{t-1} = \{\epsilon_{t-1}, \epsilon_{t-2} \dots\}$

The Alternative

- GLP impractical because depends on infinite number of parameters. But… infinite polynomials in L need not contain infinitely many free parameters! Can just be a ratio of finite-order polynomials…
- Rational Distributed Lag: $B(L) = \Theta(L)/\Phi(L)$ where $\Theta(L) = \sum_{i=0}^{q} \theta_i L^i$ and $\Phi(L) = \sum_{i=0}^{p} \phi_i L^i$
  - Taylor Expansion still has infinite terms but all Taylor coefficients are function of $\{\theta_i\}_{i=0}^{q}$ and $\{\phi_i\}_{i=0}^{p}$. That is $\Theta(L)/\Phi(L)$ is a rational function
  - Many $\sum_{i=0}^{\infty} b_i L^i$ can be approximated by $\Theta(L)/\Phi(L)$ with relatively small q and p

Structure of Time Series

Static & Finite Distributed Lag Models

- Static model: $y_t = \beta_0 + \beta_1 z_t + u_t$
- Finite distributed lag model: $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t$
    - Note: Necessary to satisfy stronger strict exogeneity assumption here since, if lags are omitted, error in one period is be correlated with the x_i in another.

Moving Average Model

- $MA(q)$: $y_t = \sum_{j=0}^{q} \theta_j \epsilon_{t-j} = \Theta(L)\epsilon_t$. May be viewed as approx. Wold representation of a general weakly stationary process
    - E.g. $MA(1)$: $y_t = \epsilon_t + \theta\epsilon_{t-1}$ or $y_t = (1 + \theta L)\epsilon_t$
        - $E(y_t) = E(\epsilon_t) + \theta E(\epsilon_{t-1}) = 0$
        - $E(y_t|\Omega_{t-1}) = E[(\epsilon_t + \theta\epsilon_{t-1})|\Omega_{t-1}] = E[\epsilon_t|\Omega_{t-1}] + \theta E[\epsilon_{t-1}|\Omega_{t-1}] = \theta\epsilon_{t-1}$
        - $\text{Var}(y_t) = \text{Var}(\epsilon_t) + \theta^2\text{Var}(\epsilon_{t-1}) = \sigma^2(1 + \theta^2)$
        - $\text{Var}(y_t|\Omega_{t-1}) = E\left[\left(y_t - E(y_t|\Omega_{t-1})\right)^2\right] = E(\epsilon_t^2|\Omega_{t-1}) = \sigma^2$

Properties

- *Invertibility*: value of unobservable ϵ_t can be "recovered" for the observable values y_t, y_{t-1} ... In practice turns a finite MA into an infinite AR
    - For $MA(1)$: True if $|\theta| < 1$. We note that $\epsilon_t = y_t - \theta y_{t-1} + \theta^2 y_{t-2}$ ... and $|\theta| < 1$ needed for convergence
    - For $MA(q)$ (i.e. $y_t = \Theta(L)\epsilon_t$): True if all inverse roots of polynomial $\Theta(L)$ (i.e. values of L that yield $\Theta(L) = 0$) are absolute less than 1 (i.e. outside unit circle on complex plane).
- *Autocovariance*: of $MA(q)$: $\text{Cov}(y_t, y_{t-\tau}) = \text{Cov}\left(\sum_{j=0}^{q} \theta_j \epsilon_{t-j}, \sum_{j=0}^{q} \theta_j \epsilon_{t-\tau-j}\right)$
    - If $\tau > q$ the two linear combinations do not contain the same ϵ's so $\text{Cov}(y_t, y_{t-\tau}) = 0$
    - For $0 \leq \tau \leq q$ $\text{Cov}(y_t, y_{t-\tau}) = \left(\theta_\tau\theta_0 + \theta_{\tau+1}\theta_1 + \cdots + \theta_q\theta_{q-\tau}\right)\sigma^2$
    - Neither mean (which is zero) or autocovariance depends on t. Thus weakly stationary!

Autoregressive Model

- $MA(q)$: $y_t = \sum_{j=1}^{p} \phi_j y_{t-j} + \epsilon_t$. May be viewed as approx. of Wold representation of a general weakly stationary process. Can be consistently estimated by OLS!
    - E.g. $AR(1)$: $y_t = \phi y_{t-1} + \epsilon_t$ or $(1 - \phi L)y_t = \epsilon_t$
        - $E(y_t) = E(\epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \cdots) = E(\epsilon_t) + \phi E(\epsilon_{t-1}) + \phi^2 E(\epsilon_{t-2}) \ldots = 0$
        - $E(y_t|y_{t-1}) = E(\phi y_{t-1} + \epsilon_t|y_{t-1}) = \phi E(y_{t-1}|y_{t-1}) + E(\epsilon_t|y_{t-1}) = \phi y_{t-1}$
        - $\text{Var}(y_t) = \text{Var}(\epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \cdots) = \sigma^2 + \phi^2\sigma^2 + \phi^4\sigma^2 \ldots = \sigma^2\sum\phi^{2i} = \frac{\sigma^2}{1-\phi^2}$
        - $\text{Var}(y_t|y_{t-1}) = \text{Var}(\phi y_{t-1} + \epsilon_t|y_{t-1}) = \phi^2\text{Var}(y_{t+1}|y_{t+1}) + \text{Var}(\epsilon_t|y_{t+1}) = 0 + \sigma^2$

Properties

- *Stability/Convergence*:
    - For $AR(1)$: True if $|\phi| < 1$. We note that $y_t = \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \cdots$ and $|\phi| < 1$ is needed for convergence
- *Invertibility:*
    - For $AR(p)$: $\Phi(L)y_t = \epsilon_t$ is stated in terms of the roots of polynomial $\Phi(L)$ (that is the values of L that yield $\Phi(L) = 0$). True if all inverse roots of $\Phi(L)$ are absolute less than one.

- *Autocovariance:* We use the Yule Walker equation.
  - Take variances of both parts
  - $y_t = \phi y_{t-1} + \epsilon_t$
  - $y_t y_{t-\tau} = \phi y_{t-1} y_{t-\tau} + \epsilon_t y_{t-\tau}$
  - $E(y_t y_{t-\tau}) = E(\phi y_{t-1} y_{t-\tau} + \epsilon_t y_{t-\tau})$
  - $\gamma(\tau) = \phi\gamma(\tau - 1)$. Find recursively where $\gamma(0) = \mathrm{Var}(y_t) = \frac{\sigma^2}{1-\phi^2}$ so $\gamma(\tau) = \phi^\tau \frac{\sigma^2}{1-\phi^2}$
  - Hence $\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} = \phi^\tau$ but $p(\tau) = \begin{cases} \phi \text{ if } \tau = 1 \\ 0 \text{ if } \tau \geq 1 \end{cases}$

Autoregressive and Moving Average Model

- $ARMA(p,q)$: $y_t = \sum_{j=1}^{p} \phi_j y_{t-j} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$ or $\Phi(L)y_t = \Theta(L)\epsilon_t$ or $y_t = \frac{\Theta(L)}{\Phi(L)}\epsilon_t$
  - AR and MA are combined to obtain better/more-parsimonious approx. to Wold
  - Many origins: random shock driving AR is MA; sum of AR (see below); AR with WN etc.
- To estimate for ARMA with nontrivial MA part, use maximum likelihood assuming $\epsilon_t \sim N(0, \sigma^2)$
  - e.g. For MA(1) $y_t = \epsilon_t + \theta\epsilon_{t-1}$ write joint density of data as $f(\theta, \sigma^2)$ and maximize.

- E.g. $ARMA(1,1)$: $y_t = \phi y_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$
  - Where $|\phi| < 1$ for stationarity and $|\theta| < 1$ for invertibility
  - If covariance stationarity is satisfied then $y_t = \frac{1+\theta(L)}{1-\phi(L)}\epsilon_t$
  - If invertibility condition is satisfied then $\frac{1-\phi(L)}{1+\theta(L)} y_t = \epsilon_t$
- More generally for $ARMA(p,q)$:
  - If covariance stationarity is satisfied then $y_t = \frac{\Theta(L)}{\Phi(L)}\epsilon_t$
  - If invertibility condition is satisfied then $\frac{\Phi(L)}{\Theta(L)} y_t = \epsilon_t$

<u>Sum of AR</u>

- Consider two AR: $y_{1,t} = \phi y_{1,t-1} + \epsilon_{1,t} \implies (1 - \phi L)y_{1,t} = \epsilon_{1,t}$; $y_{2,t} = \gamma y_{2,t-1} + \epsilon_{2,t} \implies (1 - \gamma L)y_{2,t} = \epsilon_{2,t}$
- Thus sum of AR $W(t) = y_{1,t} + y_{2,t} = \frac{\epsilon_{1,t}}{(1-\phi L)} + \frac{\epsilon_{2,t}}{(1-\gamma L)}$
- Rearrange to $(1 - \phi L)(1 - \gamma L)W(t) = (1 - \gamma L)\epsilon_{1,t} + (1 - \phi L)\epsilon_{2,t}$
- i.e. $AR(2) = MA(1) + MA(1) = MA(1)$ so $ARMA(2,1)$

A Note on Unit Roots

- Consider ARMA(p,q): $\Phi(L)y_t = \Theta(L)\epsilon_t$ with AR polynomial having unit root and all other absolute larger than 1 (i.e. root of corresponding AR equation $1 - \phi L = 0$ is unit root $\phi^{-1} = 1$)
- Note that $\Phi(L) = \widetilde{\Phi}(L)(1 - L)$ where $\widetilde{\Phi}(L)$ is a polynomial of degree $p - 1$ with all roots being outside of the unit circle in the complex plane.
- Hence rewrite ARMA as $\widetilde{\Phi}(L)(1 - L)y_t = \Theta(L)\epsilon_t$ and thus as $\widetilde{\Phi}(L)\Delta y_t = \Theta(L)\epsilon_t$. Whilst y_t is not covariance stationary, Δy_t is covariance stationarity.

Model Selection Criteria

Mean Squared Error & R²

- $\text{MSE} = \frac{\sum_{t=1}^{T} \hat{e}_t^2}{T} \to \min$ or $\text{R}^2 = \frac{SSR}{SST} = \frac{\sum_{t=1}^{T} \hat{y}_t^2}{\sum_{t=1}^{T} y_t^2} \to \min$
- Leads to in-sample overfitting since MSE/R^2 cannot rise/fall when more variables are added
- Including more variables in forecasting model won't necessarily improve out-of-sample forecasting performance but will improve "fit" on historical data
  - Parsimonious: have a small number of parameters

- Base selection on out-of-sample 1-step-ahead expected squared prediction error $E\left(\hat{y}_{T+1,T} - y_{T+1}\right)^2$
- E.g. $y = \beta_0 + \epsilon_t$ becomes $E\left(\hat{\beta}_0 - \beta_0 - \epsilon_{T+1}\right)^2 = \text{Var}\left(\hat{\beta}_0\right) + \sigma^2 = \frac{\sigma^2}{T} + \sigma^2 = \frac{T+1}{T}\sigma^2$
- We estimate using $s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T-k} = \frac{T}{T-k}\text{MSE}$ (with k dof): $\approx \frac{T+1}{T}s^2 \approx \frac{T+1}{T-1}\text{MSE}$

Akaike and Bayesian Information Criteria

- "To get accurate est. of prediction error variance, penalize in-sample residual variance to reflect degrees of freedom" (i.e. penalize model complexity by multiplying MSE by some factor)
- $\text{AIC} = \exp\{2k/T\}\,MSE$
  - Asymptotically efficient so as T gets large its 1-s-a forecast error variance approaches "true" model's at a rate at least as fast as any other criterion. Useful when "true" is too complicated.
- $\text{BIC} = \exp\{k \ln T\,/\,T\}\,MSE$
  - Stricter hence consistent when true model is in the fixed set of models considered
- Cannot compare MSE or AIC/BIC between models with different dependent variables

Forecasting Cycles

- Time T information set $\Omega_T = \{y_T, y_{T-1} \dots\}$. For stable invertible processes we can also write as $\Omega_T = \{\epsilon_T, \epsilon_{T-1} \dots\}$ because we can recover ϵ from y and vice-versa

- The optimal forecast $y_{T+h,T}$ is the one that minimizes expected loss:
  - For the quadratic loss, $\left(y_{T+h} - y_{T+h,T}\right)^2$ this is the conditional mean $E(y_{T+h}|\Omega_t)$
  - The best linear approximation of $E(y_{T+h}|\Omega_t)$ is linear projection $P(y_{T+h}|\Omega_t)$
  - For Gaussian processes $E(y_{T+h}|\Omega_t) = P(y_{T+h}|\Omega_t)$
- Find optimal forecast using certain expression estimate error $E[y_{T+1} - cy_T]^2$. Solve in terms of c and then minimize expression

$\boldsymbol{MA(2)}$: $\boldsymbol{y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2}}$

- For $t = T + 1$ we have $y_{T+1} = \epsilon_{T+1} + \theta_1\epsilon_T + \theta_2\epsilon_{T-1}$ and…

Time	Optimal Forecast	Error	Variance
$t = T + 1$	$y_{T+1,T}$ $= \theta_1\epsilon_T + \theta_2\epsilon_{T-1}$	$e_{T+1,T} = y_{T+1} - y_{T+1,T} = \epsilon_{T+1}$ [WN]	$\sigma_1^2 = \sigma^2$
$t = T + 2$	$y_{T+2,T} = \theta_2\epsilon_{T-2}$	$e_{T+2,T} = y_{T+2} - y_{T+2,T} = \epsilon_{T+2} + \theta_1\epsilon_{T+2}$ $[MA(1)]$	$\sigma_2^2 = \sigma^2(1 + \theta_1^2)$
$t = T + h$ $h > 2$	$y_{T+h,T} = 0$	$e_{T+h,T} = y_{T+h} - y_{T+h,T} = \epsilon_{T+h} + \theta_1\epsilon_{T+h-1} + \theta_2\epsilon_{T+h-2}$ $[MA(2)]$	σ_h^2 $= \sigma^2(1 + \theta_1^2 + \theta_2^2)$

$\boldsymbol{MA(q): y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q}}$

- If $h < q$: $y_{T+h,T} = 0 + \text{"adjustment"}$ where $\sigma^2 \leq \text{Var}(y_t)$ and $e = MA(h-1)$
- If $h > q$: $y_{T+h,T} = 0$ where $\sigma^2 = \text{Var}(y_t)$ and $e = MA(q)$

$\boldsymbol{MA(\infty): y_t = \epsilon_t + b_1\epsilon_{t-1} + \cdots + b_h\epsilon_T + b_{h+1}\epsilon_{T-1} + \cdots}$

- For $t = T + h$ we have $y_{T+h} = \epsilon_{T+h} + b_1\epsilon_{T+h-1} + \cdots + b_h\epsilon_T + b_{h+1}\epsilon_{T-1} + \cdots$
- Projecting on Ω_T we get $y_{T+h,T} = b_h\epsilon_T + b_{h+1}\epsilon_{T-1} + \cdots$
- Hence $e_{T+h} = y_{T+h,T} - y_{T+h} = \sum^{h-1} b_i\epsilon_{T+h-i}$. Serially correlated by cannot be forecasted!
- Where $\mu = 0$ and $\sigma_h^2 = \sigma^2 \sum^{h-1} b_i^2$

Non-MA

- Autoregressions: Transform stable AR(q) or AR(p, q)into MA(∞) and then use MA methods above
- But… very simple if autoregression case:

Chain Rule of Forecasting

- E.g. $AR(1)$ $y_t = \phi y_{t-1} + \epsilon_t$. Considering $t = T + h$ with $h = 1,2 \ldots$ and projecting on Ω_T:
  - $y_{T+1} = \phi y_T + \epsilon_{T+1}$ so that $y_{T+1,T} = \phi y_T$
  - $y_{T+2} = \phi y_{T+1} + \epsilon_{T+2}$ so that $y_{T+2,T} = \phi y_{T+1,T}$ etc.
- E.g. ARMA(1,2) $y_t = \phi y_{t-1} + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2}$. Considering $t = T + h$ with $h = 1,2 \ldots$ and projecting on Ω_T:
  - $y_{T+1,T} = \phi y_T + \theta_1\epsilon_T + \theta_2\epsilon_{T-1}$
  - $y_{T+2,T} = \phi y_{T+1} + \theta_2\epsilon_T$
  - $y_{T+3,T} = \phi y_{T+2,T}$
  - $y_{T+4,T} = \phi y_{T+3,T}$ etc.

Putting it all Together

Making the Forecasts Operational

- Full Model: $y_t = T_t(\theta) + \sum_{i=1}^{s} \gamma_i D_{it} + \sum_{i=1}^{s} \delta_i HDV_{it} + \epsilon_t$
  - where $\Phi(L)\epsilon_t = \Theta(L)v_t$ and $T_t(\theta)$ =e.g. $\beta_1 Time$
- Replace the unknown parameter values with estimates, and unobservable innovations with residuals
  - E.g. MA(2) $y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2}$
  - Optimal 2-s-a forecast: $y_{T+2,T} = \theta_2\epsilon_t$ with forecast-error variance $\sigma_2^2 = \sigma^2(1 + \theta_1^2)$
  - Operational 2-s-a forecast: $\hat{y}_{T+2,T} = \hat{\theta}_2\hat{\epsilon}_T$ with the forecast error variance est. $\hat{\sigma}_2^2 = \hat{\sigma}^2\left(1 + \hat{\theta}_1^2\right)$ and $\hat{e}_{T+2,T} = y_{T+2} - \hat{y}_{T+2,T} = \epsilon_{T+2} + \theta_1\epsilon_{T+2} + \left(\theta_2\epsilon_{t-2} - \hat{\theta}_2\hat{\epsilon}_T\right)$

Recursive Estimate

- Forecast model selection strategies examined effectively estimate out-of-sample 1-s-a mean squared forecast by adjusting the in-sample MSE with Degrees-Of-Freedom penalty.
- Recursive estimate out of sample forecast using out of sample forecast errors. More direct and flexible.

- Instead of using all data to estimate parameters, begin with a small subset. If model contains k parameters, use first k observations, then $k + 1, k + 2$ etc. Gives set of recursive parameter estimates $\hat{\beta}_{i,t}$ to test for parameter stability.

- Compute a 1-s-a forecast $\hat{y}_{t+1,t} = \sum \hat{\beta}_{i,t} x_{i,t+1}$ and forecast errors $\hat{e}_{t+1,t} = y_{t+1} - \hat{y}_{t+1,t} \sim N(0, \sigma^2 r_t)$.
  - Make a recursive 1-s-a 95% interval forecasts to check where realizations fall
  - Standardized recursive residuals: $w_{t+1,t} = \frac{\hat{e}_{t+1,t}}{\sigma\sqrt{r_t}} \sim \text{iid } N(0,1)$ gives $\text{CUSUM} = \sum w_{t+1,t}$

Cross Validation

- <u>Cross Validation:</u> "Leave one out" only makes sense in absence of dynamics. Thus used in cross section not time series
  - Use all observations except the first one to forecast the first observations and compute MSE
  - Use all observations except the second one to forecast the first observations and compute MSE
  - … repeat with sequential deletion and average squared errors
  - Select the model with the smallest average MSE
- <u>Recursive Cross Validation:</u>
  - Let initial estimation sample run from $t = 1, \dots, T^*$ and "hold out sample" $t = T^* + 1, \dots, T$.
  - Use to forecast $T^* + 1$ and compue associated squared error
  - Update the sample by one observation $(T^* + 1)$ and forecast $T^* + 2$
  - Continue recursive re-estimation until sample is exhausted. Average squared errors $T^* + 1 \dots T$.

<u>Chow Forecast Error Test</u>

- Test how good the model would have been at predicting said future realizations of the variables of interest. E.g. Suppose $y_t = \beta_0 + \beta_1 x_1 + u_t$ for t = 1 to 10 but also have t = 11,12
- By including dummies move exactly to fit new data points $y_t = \beta_0 + \beta_1 x_1 + \mu_1 t_{11} + \mu_2 t_{12} + u_t$. F-test $\mu_i = 0 \ \forall i$ (if model is a good predictor of t = 11 and 12, would expect μ to not be significant)

Random Walks

Motivation

- Opposite of a weakly dependent time series is a highly persistent one ($\lim_{h\to\infty} Corr(y_t, y_{t+h}) \neq 0$)
- A series is "integrated" if its non-stationarity can be undone by the differencing. If the non-stationarity is undone only after d differencing, we say that the series is integrated of order d, or $I(d)$
  - White Noise is the simplest form of $I(0)$ and Random Walks is the simplest form of $I(1)$
  - General ARIMA(p,d,q): $\Phi(L)(1-L)^d y_t = c + \Theta(L)\epsilon_t$ has all roots are outside the unit circle
    - ARIMA(p,d,q) are stationary and invertible ARMA(p,q) after differencing d times
    - When $d = 0$, y is $I(0)$ or covariance stationary with mean $c\Phi^{-1}(1)$.
    - When $d = 1$, y is $I(1)$ with drift or stochastic linear trend, of $c\Phi^{-1}(1)$ per period.

Form

Random walks are the simplest example of non-stationarity $I(1)$

<u>Basic</u>

- $y_t = y_{t-1} + \epsilon_t$ is the special case of $AR(1)$ where $\phi = 1$. Iterating we get $y_t = y_0 + \sum_{j=1}^{t} \epsilon_j$.
  - $E(y_t) = E(e_t) + E(e_{t-1}) + \dots + E(e_1) + E(y_0) = E(y_0)$. Constant over time.
  - $Var(y_t) = Var(e_t) + Var(e_{t-1}) + \dots + Var(e_1) = \sigma_e^2 t$. Increases with time.
- Using $AR(1)$ rules:
  - Optimal forecast: $y_{T+h,T} = \phi^h y_T$ so y_T
  - Error: $e_{T+h,T} = \epsilon_{T+h} + \phi\epsilon_{T+h-1} + \dots + \phi^{h-1}\epsilon_{T+1}$ so $\epsilon_{T+h} + \epsilon_{T+h-1} + \dots + \epsilon_{T+1}$
  - Error Variance: $\sigma^2 \sum_{i=0}^{h-1} \phi^{2i}$ so $\sigma^2 h$. Increases without bound as forecast horizon goes to infinity

With Drift

- $y_t = \delta + y_{t-1} + \epsilon_t$. Iterating we get $y_t = \delta t + y_0 + \sum_{j=1}^{t} \epsilon_j$.
  - $E(y_t) = E(e_t) + E(e_{t-1}) + \cdots + E(e_1) + E(y_0) = E(y_0)$. Constant over time.
  - $Var(y_t) = Var(e_t) + Var(e_{t-1}) + \cdots + Var(e_1) = \sigma_e^2 t$. Increases with time.
- $y_t = \delta t + y_0 + \sum_{j=1}^{t} \epsilon_j$ vs. $z_t = \alpha + \delta t + \epsilon_t$
  - Intercept is y_0 which is not fixed parameter
  - Accumulated change $\sum_{j=1}^{t} \epsilon_j$ is not stationary and variance increases over time

Test

- Random walks cannot be stationary. Since they violate OLS assumptions, they produce a spurious regression if estimated. Thus wish to test for, and then if necessary remove, unit root processes

Dickey-Fuller Test

- Rewrite equation into appropriate form [where $\beta = (\phi - 1)$]
  - *Basic:* $y_t = \phi y_{t-1} + \epsilon_t$ to $\Delta y_t = \beta y_{t-1} + \epsilon_t$
  - *Intercept:* $(y_t - \mu) = \phi(y_{t-1} - \mu) + \epsilon_t$ to $\Delta y_t = a + \beta y_{t-1} + \epsilon_t$
  - *Intercept & Trend:* $(y_t - a - bt) = \phi(y_{t-1} - a - bt) + \epsilon_t$ to $\Delta y_t = a + bt + \beta y_{t-1} + \epsilon_t$
- Test $H_0 : \beta = 0$ [random walk and hence non-stationarity] using correct DF distribution
  - Note: if $\beta \leq\approx -2$ process is non-stationary in a peculiar oscillating way. Thus, rejecting H_0 does not automatically imply stationarity but oscillations IRL are rare.
  - Note: t-distribution has larger critical values than DF-test when adding trend, reducing its power. Thus only include a deterministic trend if we anticipate it.

Augmented DF

- Basic DF test assumes series γ_t is generated by an AR(1) process. If ϵ_t is $AR(p \geq 2)$ process, then $SE(\hat{\beta})$ is incorrect so DF is distorted
- Solve this by running an augmented regression: $\Delta y_t = a + bt + \beta \gamma_{t-1} + \sum_{j=1}^{p} \theta_j \Delta y_{t-j} + \epsilon_t$ where p is chosen by AIC or BIC. Same testing of hypothesis as before

Types of Forecasting

Conditional Forecasting

- Scenario analysis: Assuming h -step ahead value of x is x_{T+h}^*: $y_{T+h,T} \mid x_{T+h}^* = \beta_0 + \beta_1 x_{T+h}^*$
- Thus $\hat{e}_{T+h,T} = y_{T+h} - \hat{y}_{T+h,T} = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_{T+h}^* + \epsilon_{T+h}$. We note that there are two sources of error: [i] innovation ϵ_{T+h} and [ii] parameter uncertainty in β_0, β_1
- $Var(\hat{e}_{T+h,T}) = Var\left((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_{T+h}^*\right) + Var(\epsilon_{T+h}) = Var(\hat{\beta}_0) + Cov(\hat{\beta}_0, \hat{\beta}_1)x_{T+h}^* + Var(\hat{\beta}_1){x_{T+h}^*}^2 + \sigma^2$
  - If x_t has zero mean, $Cov(\hat{\beta}_0, \hat{\beta}_1) \approx 0$, $Var(\hat{\beta}_0) \approx \frac{\sigma^2}{T}$ and $Var(\hat{\beta}_1) \approx \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2}$
  - Thus: $Var(\hat{e}_{T+h,T}) = \frac{\sigma^2}{T} + \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2}{x_{T+h}^*}^2 + \sigma^2$
  - Generally, the closer x_{T+h}^* is to its mean, the smaller is the prediction error variance.

Unconditional Forecasting

- Unconditional forecasting: $y_{T+h,T} = \beta_0 + \beta_1 x_{T+h,T}$
- Instead of forecasting x and using this to forecast y, we model $x_{T+h,T}$ as a linear combination of $x_T, x_{T-1}, \ldots$ and estimate this via [autoregressive] distributed lag: $y_{T+h,T} = \beta_0 + \left[\sum_{i=h}^{p} \alpha_i y_{t-i}\right] + \sum_{i=h}^{q} \delta_i x_{t-i} + \epsilon_t$

Causality

Weak & Strong Exogeneity

- Weakly Exogenous: All of the information needed to estimate y_t from x_t is embodied in the series x_t and **any past data alone**.
  - Note simultaneous equations do not fulfill this, for example, since x_t itself requires contemporaneous data about y_t in order to be estimated in the first place
- Weak exogeneity is sufficient for unbiasedness, it is not sufficient for determining the long-run relationship between two variables….
  - E.g. $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + u_t$ and $x_t = \gamma_0 + \gamma_1 x_{t-1} + \gamma_2 y_{t-1} + e_t$
  - x_t is weakly exogenous but using only the first equation mistakes the relationship
- Strongly Exogenous: all of the information to estimate y_t from x_t is embodied **only in x_t and past values of x**. Formally, $E(x_t | x_{t-1}, x_{t-2} \ldots) = E(x_t | x_{t-1}, y_{t-1}, x_{t-2} \ldots)$

Granger Causality

- If strong exogeneity is violated [i.e. $E(x_t | x_{t-1}, x_{t-2} \ldots) \neq E(x_t | x_{t-1}, y_{t-1}, x_{t-2} \ldots)$], y Granger causes x: y Granger causes x iff after controlling for past x, past y still help to forecast x.
- But serious limitations: says nothing about direction of causality, does not chain-well, two variables could be seen to Granger-cause each other, does not imply a change in y will cause a change in x!
- Conditional Granger causality: $E(x_t | x_{t-1}, z_{t-1} \ldots) \neq E(x_t | x_{t-1}, y_{t-1}, z_{t-1} \ldots)$ so y Granger causes x conditional on z. This allows us to test for the 'true' connectivities between variables.

Test

- Hypothesis that x does not Granger cause y tested using the F-test in an ADL Model: $y_{T+h,T} = \beta_0 + \sum_{i=h}^{p} \alpha_i y_{t-i} + \sum_{i=h}^{q} \delta_i x_{t-i} + \epsilon_t$ where $\delta_1 = \cdots = \delta_q = 0$
  - i.e. regress x on past values of x and y [and z] to see if past values of y are jointly sig.
  - Need to assume homoskedasticity, ϵ_t is uncorrelated to past y, and no serial correlation

Dynamic Causal Effect

- For causal DL model $y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_r x_{t-r} + \epsilon_t$…
  - β_0 is impact multiplier
  - β_h is h-period dynamic multiplier
  - $\beta_0 + \cdots + \beta_r$ is long-run multiplier
- For causal ADL model $y_t = \phi y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \epsilon_t$
  - β_0 is impact multiplier
  - $\beta_1 + \phi\beta_0$ is first-period dynamic multiplier
  - $\phi^{h-1}(\beta_1 + \phi\beta_0)$ is h-period dynamic multiplier
  - $\beta_0 + \sum_{h=1}^{\infty} \phi^{h-1}(\beta_1 + \phi\beta_0) = \frac{\beta_0 + \beta_1}{1-\phi}$ is long-run multiplier

Panel Data

Basics

- Panel Data: Cross sectional sample followed over time:
  - E.g. $Crime_{it} = \beta_0 + \beta_1 Police_{it} + \beta_2 Unemployed_{it} + (\alpha_i + \epsilon_{it})$
  - where α_i is time invariant unobservable and ϵ_{it} is time varying unobservable (idiosyncratic error)
  - There are many ways to eliminate α_i (and thus OVB resulting from $Cov(\beta_i, \alpha_i)$…
- Unbalanced Panel: missing years or cross-sectional units in sample. Not a problem unless the reason why data is missing is endogenous.

FD & FE

(1) *First Differences*

- Two-period Model (i.e. $T = 2$)
  - Note $score_{it} = \delta_0 d2_t + \beta_1 hrs_{it} + (\alpha_i + \epsilon_{it})$ for $t = 1{,}2$
  - Hence $score_{i2} - score_{i1} = \delta_0 + \beta_1(hrs_{i2} - hrs_{i1}) + (\epsilon_{i2} - \epsilon_{i1})$
  - Rewritten as $\Delta score_i = \delta_0 + \beta_1 \Delta hrs_i + \Delta\epsilon_i$
  - We need strictly exogenous regressors, which may not be the case:
    - $Cov(\Delta hrs_i, \Delta\epsilon_i) = Cov(hrs_{i2} - hrs_{i1}, \epsilon_{i2} - \epsilon_{i1})$
    - $= Cov(hrs_{i2}, \epsilon_{i2}) - Cov(hrs_{i2}, \epsilon_{i1}) - Cov(hrs_{i1}, \epsilon_{i2}) + Cov(hrs_{i1}, \epsilon_{i1}) =$
    - $= -Cov(hrs_{i2}, \epsilon_{i1}) \neq 0$ since bad results last year induce more effort!
- General Model (e.g. $T = 3$)
  - $y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 hrs_{it} + (\alpha_i + \epsilon_{it})$ with key assumption $Cov(x_{itj}, u_{is}) = 0$
  - $\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta hrs_{it} + \Delta\epsilon_{it}$
  - Must now assume Δu_{it} are serially uncorrelated (i.e. u_{it} must follow a random walk) for Standard Errors and test statistics to be valid
- Drawbacks: Hard to collect data; greatly reduces variation in explanatory variables (i.e. Δx_i less variation than x_{it}) and thus large standard errors

(2) *Fixed Effects*

- General Model
  - Note $y_{it} = \beta_1 x_{it} + a_i + u_{it}$ and $\bar{y}_{it} = \beta_1 \bar{x}_{it} + a_i + \bar{u}_{it}$
  - Hence $y_{it} - \bar{y}_{it} = \beta_1(x_{it} - \bar{x}_{it}) + (u_{it} - \bar{u}_{it})$
  - Rewritten as $\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it}$ (time-demeaning)
  - Need strict exogeneity for FE estimator to be unbiased
- FE Assumptions:

  [i.] takes form $y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}$;
  [ii.] random sample from cross section;
  [iii.] each explanatory variable changes over time and no perfect linear relationships;
  [iv.] $E[u_{it}|\boldsymbol{X_i}, a_i] = 0$ [strict exog.; for consistency only need $Corr(\Delta x_{itj}, \Delta u_{it})$]
  -Needed for unbiased estimators and consistency
  [v.] $Var[u_{it}|\boldsymbol{X_i}, a_i] = Var(u_{it}) = \sigma_u^2$
  [vi.] $Cov(u_{it}, u_{is}|\boldsymbol{X_i}, a_i) = 0$
  -Needed for BLUE (vi. Makes it better than FD!)

- For each cross-sectional observation i we lose one df because of the time-demeaning. Thus $dof = NT - N - k = N(T-1) - k$. To apply OLS, u_{it} must be serially uncorrelated and homoscedastic
- Alternative way to estimate is via Dummy Variable Regression
  - Put dummy variable for each cross-sectional observation i. Note that this is impractical for large i
  - Gives exact same estimates as time-demeaning and can estimate $\hat{\alpha}_i = \bar{y}_i - \hat{\beta}_1 \bar{x}_{i1} - \cdots - \hat{\beta}_k \bar{x}_{ik}$
  - Unbiased but inconsistent when T is fixed and $N \to \infty$ (For each added observation, added a_i)
- Note the following limitations of FE:
  - Time-constant variables cannot be included by themselves but they can be interacted with variables that change over time (including time itself). E.g. gender
  - Cannot estimate the effect of any variable whose change across time is constant (e.g. years of experience) since it cannot be distinguished from aggregate time effects

FD vs. FE

- If $T = 2$ then $FE = FD$ and both are consistent as long as we estimate the same model. (I.e. if intercept in FD then FE must include dummy variable for the second time period)
- If $T \geq 3$ then $FE \approx FD$ both are unbiased and consistent under strict exogeneity
  - FE model is more efficient if the idiosyncratic error is not serially correlated
    - Note: can only estimate said errors in the FD model anyway
  - FE model has a tendency to misbehave if many time periods and small cross section
  - FE model is less sensitive to measurement error for large T.

Difference-In-Difference

- $Y_i = \beta_0 + \beta_1 After_i + \beta_2 Treatment_i + \beta_3 Treatment_i * After_i + u_i$
- Test $\beta_3 = 0$ via regression

	Treatment	Control	Diff
Before	β_0	$\beta_0 + \beta_2$	β_2
After	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
Diff	β_1	$\beta_1 + \beta_3$	β_3

Arellano Bond Instrument (?)

- $score_{i2} - score_{i1} = \beta_1(hrs_{i2} - hrs_{i1}) + (\epsilon_{i2} - \epsilon_{i1})$ | $score_{i3} - score_{i2} = \beta_1(hrs_{i3} - hrs_{i2}) + (\epsilon_{i3} - \epsilon_{i2})$
- 2n observations. ϵ_{i2} present in both so not IID across individuals. Need correct for auto correlation