Empirical Methods in Finance MFE230E
Week 3: GLS, IV, VARs, MLE

Martin Lettau

Spring 2019

Haas School of Business

Outline

1. **Dealing with heteroskedasticity: GLS**
2. When OLS is inconsistent: Regressor endogeneity
   - Omitted variables
   - Measurement error
   - Simultaneous equations
   - Solution: Instrumental variables (IV)
3. VARMA(p, q)
4. Maximum likelihood
   - MLE estimation of MA(1) process
   - MLE estimation of regime-switching models

Dealing with heteroskedasticity

GLS

Heterogeneity in OLS

- Let's go back to the classical finite sample OLS (asymptotic results are analogous)
- We assumed homoskedastic i.i.d. errors: $\mathrm{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}\right] = \sigma^2\mathbf{I}$
- Let's relax this assumption:

$$\mathrm{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}\right] = \sigma^2\mathbf{D}(\mathbf{X})$$

  where $\mathbf{D}(\mathbf{X})$ is nonsingular and known

- Consequences:
  1. OLS estimator is still unbiased but no longer BLUE
  2. Standard errors need to be adjusted (White correction)

Let's assume for the moment that we know $\mathbf{D} = \mathbf{D}(\mathbf{X})$.

Since $\mathbf{D}$ is symmetric and nonsingular, there is a (not necessarily unique) matrix $\mathbf{C}$ such that

$$\mathbf{D}^{-1} = \mathbf{C}'\mathbf{C}$$

We can transform the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ by multiplying by $\mathbf{C}$:

$$\tilde{\mathbf{Y}} = \mathbf{C}\mathbf{Y},\ \tilde{\mathbf{X}} = \mathbf{C}\mathbf{X},\ \tilde{\boldsymbol{\epsilon}} = \mathbf{C}\boldsymbol{\epsilon}$$

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$$

Transformed errors are i.i.d.:

$$
\begin{aligned}
\mathrm{E}\left[\tilde{\mathbf{e}}\tilde{\mathbf{e}}'|\tilde{\mathbf{X}}\right] = \mathrm{E}\left[\tilde{\mathbf{e}}\tilde{\mathbf{e}}'|\mathbf{X}\right] &= \\
&= \mathbf{C}\,\mathrm{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\tilde{\mathbf{X}}\right]\mathbf{C}' \\
&= \sigma^2\mathbf{CDC}' = \sigma^2\mathbf{I}
\end{aligned}
$$

by definition of C

If $\mathrm{E}(\tilde{\boldsymbol{\epsilon}}|\tilde{\mathbf{X}}) = 0$, the OLS estimator of transformed model is BLUE and

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_{GLS} &= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}} \\
&= \left[(\mathbf{CX})'(\mathbf{CX})\right]^{-1}(\mathbf{CX})'\mathbf{CY} \\
&= \left(\mathbf{X}'\mathbf{C}'\mathbf{CX}\right)^{-1}\mathbf{X}'\mathbf{C}'\mathbf{CY} \\
&= \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{Y}
\end{aligned}
$$

$$
\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{GLS}|\mathbf{X}) = \sigma^2\left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}
$$

Weighting matrices

- $\mathbf{C}$ is an example of a **weighting matrix**
- General idea: Let $\mathbf{W}$ be a non-singular $n \times n$ matrix
- Define

$$\tilde{\mathbf{Y}} = \mathbf{W}\,\mathbf{Y},\ \tilde{\mathbf{X}} = \mathbf{W}\,\mathbf{X},\ \tilde{\boldsymbol{\epsilon}} = \mathbf{W}\,\boldsymbol{\epsilon}$$

$$\widehat{\boldsymbol{\beta}}_W = \left(\mathbf{X}'\,\mathbf{W}'\mathbf{W}\,\mathbf{X}\right)^{-1}\mathbf{X}'\,\mathbf{W}'\mathbf{W}\,\mathbf{Y}$$

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_W|\mathbf{X}) = \sigma^2\left(\mathbf{X}'\,\mathbf{W}'\mathbf{W}\,\mathbf{X}\right)^{-1}$$

- OLS: $\mathbf{W} = \mathbf{I} \longrightarrow$ all observations are "weighted" equally
- Question: Which $\mathbf{W}$ is **optimal** (smallest $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_W|\mathbf{X})$)?

  $\longrightarrow$ **GLS**: $\mathbf{W}$ such that $\mathbf{W}'\mathbf{W} = \mathbf{D}^{-1}$
- We will encounter other examples of weighting matrices later in the class

Example: Weighted Least Squares (WLS) I

- Suppose $E(e_i e_j) = 0, \forall i \neq j$ but $\sigma_i^2 \neq \sigma_j^2$:

variances are different
but errors are not correlated

$$\mathbf{D} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

- Then

idea: put more weight
on more precise

$$\mathbf{C} = \begin{bmatrix} 1/\sigma_1 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_n \end{bmatrix}$$

- Let's assume for the moment that we observe σ_i

EXAMPLE: WEIGHTED LEAST SQUARES (WLS) II

- The transformation $\tilde{\mathbf{Y}} = \mathbf{CY}$, $\tilde{\mathbf{X}} = \mathbf{CX}$, $\tilde{\mathbf{e}} = \mathbf{C}\,\mathbf{e}$ is

$$\tilde{y}_i = \frac{y_i}{\sigma_i},\ \tilde{x}_i = \frac{x_i}{\sigma_i},\ \tilde{u}_i = \frac{u_i}{\sigma_i}$$

$D^{-1} = C^T C$

- The WLS estimator is

$$\hat{\beta}_{WLS} = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_i x_i\right)^{-1} \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_i y_i\right)$$

- The observations are weighted according to their volatility.
- In general, GLS puts the "optimal" weight on each observation.
- Note that the optimal weights generally depend on the variance/covariance structure of the residuals.

USING GLS

- The GLS estimator can be computed easily if **D** is known
- If **D** is not known, it has to be estimated
- If there are n observations, **D** is a symmetric $n \times n$ matrix with $n(n+1)/2$ unknown elements

upper diagonal.

- $n(n+1)/2 > n$, so **D** cannot be estimated without further assumptions
- Even in the WLS case **D** has n parameters that cannot be estimated with n observations

$\Rightarrow$ GLS and WLS are **infeasible** without further assumptions

Feasible GLS (FGLS)

- To implement GLS, we need to make further assumptions about **D**
- Feasible WLS: σ_i^2 depends on an observable variable z_i
  - $\sigma_i^2 = \delta z_i$
  - $\sigma_i^2 = \exp(\delta z_i)$
- Estimate one parameter, $\hat{\delta}$, to form $\hat{\sigma}_i^2$
- Alternative: ARCH/GARCH models to estimate variance $\hat{\sigma}_i^2$ (later)

DISCUSSION: OLS VS. GLS

IMPORTANT

GLS is more efficient than OLS but ...

- **D**(**X**) might not be known and has to be estimated
- **D**(**X**) might be close to singular and inverting an $n \times n$ matrix that is close to singular is numerically difficult
- In the classical finite sample setting with exogenous errors, OLS and GLS are unbiased and GLS is BLUE
- OLS is still consistent in large samples even if errors are predetermined but not exogenous
- GLS requires exogeneity of errors to be consistent, i.e. GLS can be inconsistent if regressors are predetermined but not exogenous!

Conclusion: **GLS might be more efficient but OLS is more robust**

Summary: Two ways to deal with heteroskedasticity

IMPIRTANT

1. (F)GLS:

$$\widehat{\boldsymbol{\beta}}_{GLS} = \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{Y}$$

2. OLS with White correction of standard errors of OLS estimator:

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\widehat{\mathbf{V}}^{W}_{\widehat{\boldsymbol{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\hat{e}_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}$$

Note: The White correction does **not** require the estimation of **D**!

Outline

1. Dealing with heteroskedasticity: GLS
2. When OLS is inconsistent: Regressor endogeneity
    - Omitted variables
    - Measurement error
    - Simultaneous equations
    - Solution: Instrumental variables (IV)
3. VARMA(p, q)
4. Maximum likelihood
    - MLE estimation of MA(1) process
    - MLE estimation of regime-switching models

Regressor Endogeneity

A (potentially) serious issue for OLS!

- OLS assumption (classical and asymptotic) : Predetermined regressors
  → **Errors are orthogonal to regressors**:

$$\mathrm{E}[e_i \mathbf{x}_i] = 0$$

ALWAYS

- Example: If the "true" model is an AR(p) and we estimate an AR(p) by OLS, then this condition is satisfied.

Definition 1 (Regressor endogeneity).

If the OLS condition of predetermined regressors is violated, i.e.

$$\mathrm{E}[e_i \mathbf{x}_i] \neq 0,$$

then the regressors are called **endogenous**.

EXAMPLE: SERIAL CORRELATION IN ERRORS

- Suppose z_t is an AR(1) :

$$z_t = \phi z_{t-1} + e_t$$

and we estimate an AR(1) model: OLS is consistent.

- However, if z_t is an **ARMA(1,1)** process

$$z_t = \phi z_{t-1} + v_t + \theta v_{t-1}$$

and we (falsely) estimate an AR(1) process

$$z_t = \phi z_{t-1} + e_t,$$

then

X variable

$$E[z_{t-1} e_t] = E[z_{t-1}(v_t + \theta v_{t-1})] = \theta \sigma_e \neq 0!!$$

Regressor endogeneity

- Examples of regressor endogeneity:
  1. Serial correlation in errors
  2. Omitted variables
  3. Simultaneous equations
  4. Measurement error
- Question: What happens to the OLS estimator if regressors are endogenous?

Regressor endogeneity I: Serial correlation in errors

- If $\mathrm{E}[e_i x_i] \neq 0$: **OLS is no longer consistent, the cardinal sin in econometrics!**
- Recall:

$$\boxed{\begin{aligned} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left(\frac{1}{n}\sum \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \left(\frac{1}{n}\sum \mathbf{x}_i e_i\right) \\ \Rightarrow \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left(\frac{1}{n}\sum \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum \mathbf{x}_i e_i\right) \end{aligned}}$$

- For $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ to converge to 0, $\frac{1}{\sqrt{n}}\sum \mathbf{x}_i e_i$ has to converge to 0.
- If $\mathbf{x}_i$ and e_i are correlated, then OLS will not be consistent!
- Ignoring the MA(1) serial correlation in the error in the OLS regression

$$z_t = \phi z_{t-1} + e_t.$$

will produce an **inconsistent** estimate of ϕ!

- $E[e|X] = 0$ exogenous errors
- predetermined
- non-predetermined

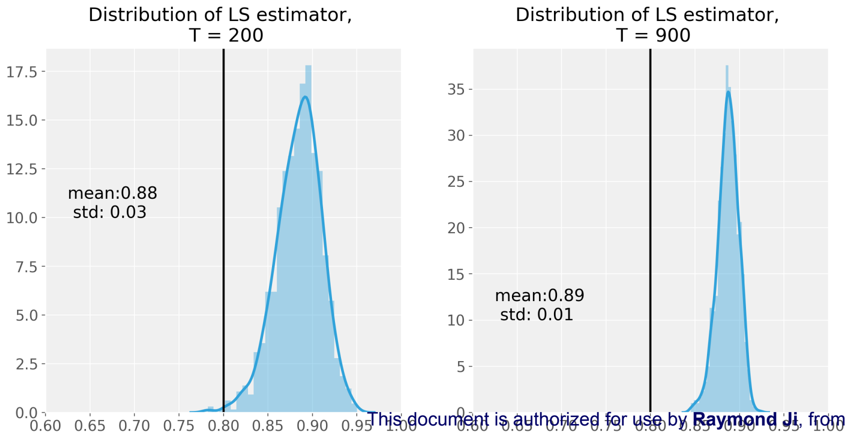
Regressor endogeneity I: Serial correlation in errors

True model: $z_t = 0.8\,z_{t-1} + v_t + 0.5\,v_{t-1},\ v_t \sim N(0,2)$

Estimated model: $z_t = \phi z_{t-1} + e_t$

bias upwards

OLS is simply wrong



test: ACF of errors

$$\text{Regression:} \quad y_t = \mathbf{x}_t'\boldsymbol{\beta} + e_t$$

- If $\mathbf{x}_i$ includes lagged dependent variable(s) y_{t-j} $\longrightarrow$ $\mathrm{E}[e_i \mathbf{x}_i] \neq 0$ and OLS is inconsistent
- No lagged dependent variables: $\mathbf{x}_t$ does not include and y_{t-j} $\longrightarrow$ $\mathrm{E}[e_i \mathbf{x}_i] = 0$ and OLS is consistent even with serially correlated errors

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + v_t + \theta v_{t-1}$$

$$\text{OLS: } y_t = \mathbf{x}_t'\boldsymbol{\beta} + e_t$$

$\longrightarrow$ $\hat{\boldsymbol{\beta}}$ is consistent as long as $\mathbf{x}_t$ does not include any y_{t-j}.

Pb: AR model always included
So OLS appropriate only if error are not serially correlated

- Serial correlation in errors is often a symptom of model misspecification
- Models **without** lagged dependent variables (i.e. no AR lags)
  $\Rightarrow$ Error serial correlation causes **loss of efficiency**
- Models **with** lagged dependent variables (i.e. with AR lags)
  $\Rightarrow$ **OLS is inconsistent!**
- Lesson: When you estimate AR(p) models, **ALWAYS check whether the residuals are (close to) i.i.d.!!**

$$\text{True model:}\quad y = \beta x + \gamma z + e$$

$$\text{Regression:}\quad y = \beta_x x + u$$

$$\begin{aligned}
\hat{\beta}_x &= \mathrm{E}(xx)^{-1}\mathrm{E}(xy) \\
&= \mathrm{E}(xx)^{-1}\mathrm{E}(x(\beta x + \gamma z + e)) \\
&= \beta + \mathrm{E}(xx)^{-1}\mathrm{E}(xz)\gamma \\
&\neq \beta \quad \text{if } \mathrm{E}(xz) \neq 0
\end{aligned}$$

$$\text{OLS is inconsistent:}\quad \hat{\beta}_x \not\rightarrow \beta_x$$

Regressor endogeneity III: Measurement error

Suppose that the true model is

$$y_t = x_t^* \beta + e_t^* .$$

Assume that the models satisfies all assumptions of the classical regression model.

Suppose that data on x_t^* is observed with error. Instead of observing x_t^*, we observe x_t where

$$x_t = x_t^* + u_t.$$

u_t represents **measurement error**.

Standard assumption: u_t is orthogonal to y_t, x_t^* and $e_t^\star$.

What is the consequence of measurement error for OLS of y_t on x_t?

$$\begin{aligned} y_t &= x_t^* \beta + e_t^* \\ &= (x_t - u_t)\beta + e_t^* \\ &= x_t\beta + e_t \quad \text{where} \quad e_t = -u_t\beta + e_t^* \end{aligned}$$

The problem is that x_t and e_t are correlated:

$$E(x_t e_t) = E((x_t^* + u_t)(-u_t\beta + e_t^*)) = -\beta\sigma_u^2$$

Therefore, OLS is **biased and inconsistent!**

Note: Measurement error in the dependent variable y_t does NOT imply the OLS is inconsistent!

Regressor endogeneity III: Simultaneous equations

Example: Link between firm value (V_t) and the firm's debt (D_t)

Possible regression:

$$V_t = \alpha + \beta D_t + e_t$$

Question: Is debt an exogenous variable and uncorrelated with the error?

Probably not: Young firms and firms in distress might not have access to credit markets. So we might consider the **opposite** regression:

$$D_t = \delta + \gamma V_t + \upsilon_t$$

$$\Rightarrow E\left(D_t e_t\right) = E\left(\left(\gamma V_t + \upsilon_t\right) e_t\right) \neq 0$$

since the firm value V_t is clearly correlated with e_t

Regressor endogeneity VI: Simultaneous equations

- Goal: Estimate the effect of price on quantity of a good.

$$\text{Supply:} \quad q_t^s = \gamma p_t + u_t^s, \ \gamma > 0$$
$$\text{Demand:} \quad q_t^d = \theta p_t + u_t^d, \ \theta < 0.$$

- q_t^s, q_t^d are demeaned
- Note that supply and demand both depend on the price p_t.
- In equilibrium, $q_t^s = q_t^d = q_t$.
- Solving for p_t and q_t yields

$$p_t = \frac{1}{\gamma - \theta} u_t^d - \frac{1}{\gamma - \theta} u_t^s$$
$$q_t = \frac{\gamma}{\gamma - \theta} u_t^d - \frac{\theta}{\gamma - \theta} u_t^s$$

- Demand and supply shocks u_t^d and u_t^s affect both price and quantity!

Now, suppose we estimate a regression of quantities on prices via OLS (assume that supply and demand shocks are uncorrelated):

$$
\begin{aligned}
q_t &= \beta p_t + e_t \\
\hat{\beta} &= \frac{\mathrm{Cov}\left(q_t, p_t\right)}{\mathrm{Var}\left(p_t\right)} \\
&= \frac{\mathrm{Cov}\left(\frac{1}{\gamma-\theta} u_t^d - \frac{1}{\gamma-\theta} u_t^s, \frac{\gamma}{\gamma-\theta} u_t^d - \frac{\theta}{\gamma-\theta} u_t^s\right)}{\mathrm{Var}\left(\frac{1}{\gamma-\theta} u_t^d - \frac{1}{\gamma-\theta} u_t^s\right)} \\
&= \frac{\gamma\sigma_d^2 + \theta\sigma_s^2}{\sigma_d^2 + \sigma_s^2}
\end{aligned}
$$

The OLS estimator

$$\hat{\beta} = \frac{\gamma\sigma_d^2 + \theta\sigma_s^2}{\sigma_d^2 + \sigma_s^2}$$

is an average of the demand and supply coefficients γ and θ.

Problem:

The demand shocks u_t^d and supply shocks u_t^s affect the price p_t and the error term e_t, thus the error term is correlated with the explanatory variable!

Survey of regressor endogeneity in cross-sectional **corporate finance**:
Michael Roberts and Toni Whited "Endogeneity in Empirical Corporate Finance", in George Constantinides, Milton Harris, and Rene Stulz, eds. Handbook of Economics of Finance, Volume 2, 2012, Elsevier.

A working paper version is available at
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1748604

- Consider the linear model with endogenous errors:

$$y_t = \beta x_t + e_t$$
$$\mathrm{E}[x_t e_t] \neq 0$$

- Suppose we have a variable, z_t which is:
  - correlated with x_t
  - uncorrelated with $e_t : \mathrm{E}[z_t e_t] = 0$.
- z_t is called and **instrument**.

- OLS moment condition $\mathrm{E}(x_t e_t) = 0$ is not satisfied

  ... but the orthogonolity moment condition with instrument z_t is:

$$\mathrm{E}(z_t e_t) = 0$$

orthogonal to error term

$$\mathrm{E}(z_t e_t) = \mathrm{E}\left(z_t(y_t - \beta x_t\right) = 0$$

$$\mathrm{E}(z_t y_t) = \beta \mathrm{E}(z_t x_t)$$

$$\beta = \mathrm{E}(z_t x_t)^{-1} \mathrm{E}(z_t y_t)$$

$$\boldsymbol{\beta} = (\mathbf{Z'X})^{-1}\mathbf{Z'Y}$$

instrumental variable

Definition 2.

Let $\mathbf{z}_t$ be a vector of instruments that are uncorrelated with the regression error, i.e. $\mathrm{E}(\mathbf{z}_t e_t) = 0$. The estimator

$$\hat{\boldsymbol{\beta}}_{IV} = \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{z}_t' \mathbf{x}_t \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{z}_t' y_t \right) = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{Y}$$

is called the **Instrumental Variable (IV) estimator.**

Theorem 1.

*The IV estimator is **consistent***

Proof:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{IV} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y} \\
&= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) \\
&= \boldsymbol{\beta} + (\mathbf{Z}'\mathbf{X})^{-1}\underbrace{\mathbf{Z}'\mathbf{e}}_{\frac{1}{n}\sum z_t e_t \to 0} \to \boldsymbol{\beta}
\end{aligned}
$$

Recall the AR(1) model with MA errors: $\nearrow$ ARMA(1,1)

$$\boxed{y_t = \phi y_{t-1} + v_t + \theta v_{t-1}}$$

$y_{t-1} = \phi y_{t-2} + v_{t-1} + \theta v_{t-2}$

$$y_t = \phi y_{t-1} + v_t + \theta v_{t-1} = \phi(\phi y_{t-2} + v_{t-1} + \theta v_{t-2}) + v_t + \theta v_{t-1}$$

$$= \phi^2 y_{t-2} + \underline{v_t} + (\theta + \phi) v_{t-1} + \theta\phi v_{t-2}$$

$y_{t-2} = v_{t-2} + (\ldots) \leftarrow t-3, t-4 \ldots$

$\perp v_{t-2}$

$\rightarrow$ y_{t-2} is correlated with y_{t-1} and orthogonal to v_t, v_{t-1}.

Hence it is a valid instrument: $\boxed{z_t = y_{t-2}}$

$$E(y_t y_{t-2}) = E(y_{t-2}(\phi^2 y_{t-2} + v_t + (\theta + \phi) v_{t-1} + \theta\phi v_{t-2}))$$

$$= \phi \sigma_y^2 + \theta\phi \sigma_v^2$$

$$\hat{\beta}_{IV} = \frac{E[y_t y_{t-2}]}{E[y_{t-1} y_{t-2}]} = \frac{\phi^2 \sigma_y^2 + \phi\theta\sigma_v^2}{\phi\sigma_y^2 + \theta\sigma_v^2} = \phi$$

Suppose we are analyzing the supply and demand of orange juice. Let w_t be the number of days below freezing in Florida.

Since weather affects orange supply:

$$q_t^s = \gamma p_t + h w_t + u_t^s, \ \gamma > 0.$$

The demand equation is unchanged:

$$q_t^d = \theta p_t + u_t^d, \ \theta < 0.$$

Assumption: w_t is uncorrelated with demand shocks u_t^d and other supply shocks u_t^s.

Instrument: Supply and demand II

Solving for p_t and q_t yields

$$p_t = \frac{h}{\gamma - \theta} w_t + \frac{1}{\gamma - \theta} u_t^d - \frac{1}{\gamma - \theta} u_t^s$$

$$q_t = \frac{h\beta}{\gamma - \theta} w_t + \frac{\gamma}{\gamma - \theta} u_t^d - \frac{\theta}{\gamma - \theta} u_t^s$$

Note: w_t affects supply but not demand!

The (inconsistent) OLS regression is

$$q_t = \beta p_t + e_t.$$

- Consider the IV estimator with w_t as an instrument
- w_t is uncorrelated with other demand and supply shocks.

  $\longrightarrow$ w_t is a valid instrument

$$
\begin{aligned}
\beta_{IV} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y} \\
&= \frac{\mathrm{Cov}(w_t, q_t)}{\mathrm{Cov}(w_t, p_t)} \quad \text{since regression is univariate} \\
&= \frac{\mathrm{Cov}(w_t, \frac{\beta h}{\gamma-\beta} w_t + \frac{1}{\gamma-\beta} u_t^d - \frac{1}{\gamma-\beta} u_t^s)}{\mathrm{Cov}(w_t, \frac{h}{\gamma-\beta} w_t + \frac{1}{\gamma-\beta} u_t^d - \frac{1}{\gamma-\beta} u_t^s)} \\
&= \beta
\end{aligned}
$$

Thus the IV regressions correctly estimates the effect of price on quantity!

KEY: FINDING AN APPROPRIATE INSTRUMENT

- Sometimes, finding the appropriate instrument is obvious, example: lags for MA(1) error in residual
- Often, it is difficult to find a good instrument, example: shock on demand (# frost days)
- Problem: The key assumption $\mathrm{E}(z_t e_t) = 0$ is untestable since e_t is not observable
- The Roberts-Whited paper mentioned above is a great reference on how to pick an appropriate instrument

Summary: Violations of OLS assumption/Endogenous regressors

- $E(xe) \neq 0 \longrightarrow$ OLS is biased and inconsistent
- Examples:
  1. Serial correlation in errors
  2. Omitted variables
  3. Simultaneous equations
  4. Measurement error
- (Partial) solution: IV estimator
- Key: “Good” instruments

Outline

1. Dealing with heteroskedasticity: GLS
2. When OLS is inconsistent: Regressor endogeneity
    - Omitted variables
    - Measurement error
    - Simultaneous equations
    - Solution: Instrumental variables (IV)
3. **VARMA(p, q)**
4. Maximum likelihood
    - MLE estimation of MA(1) process
    - MLE estimation of regime-switching models

Multivariate AR models: VARs

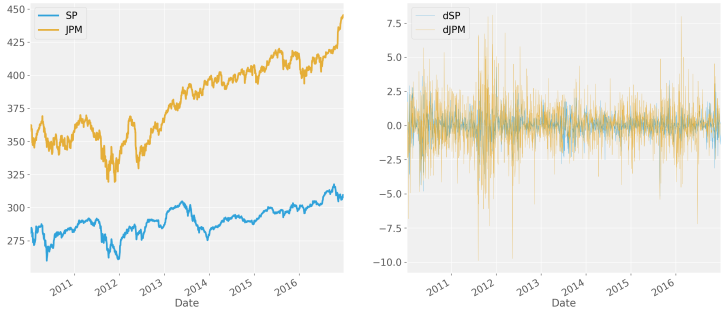
Don’t confuse VAR with VaR!

Vector AutoRegressions (VAR)

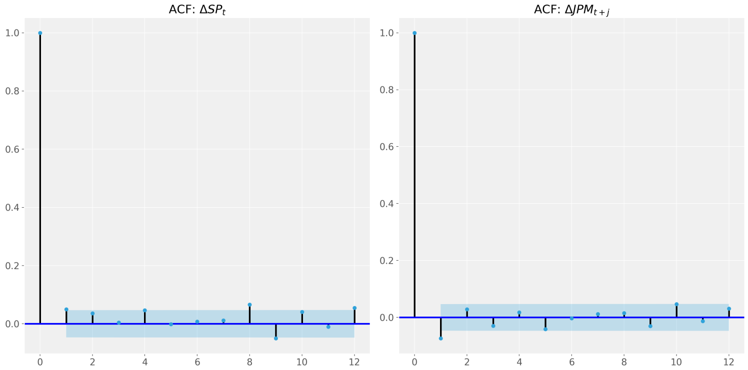
- The VAR is a generalization of the autoregressive process for multivariate series.
- Suppose we are interested in the joint dynamic interaction between two time series: S&P 500 index and JP Morgan.
- Let SP_t and JPM_t be log stock prices.
- We will work with log first difference of stock prices, i.e. returns: ΔSP_t, ΔJPM_t (Why?)
- Let

$$\mathbf{Y}_t = \begin{bmatrix} \Delta SP_t \\ \Delta JPM_t \end{bmatrix}$$

VECTOR AUTOREGRESSIONS (VAR)



VECTOR AUTOREGRESSIONS (VAR): INDIVIDUAL ACFS



VARS: The cross-correlation function

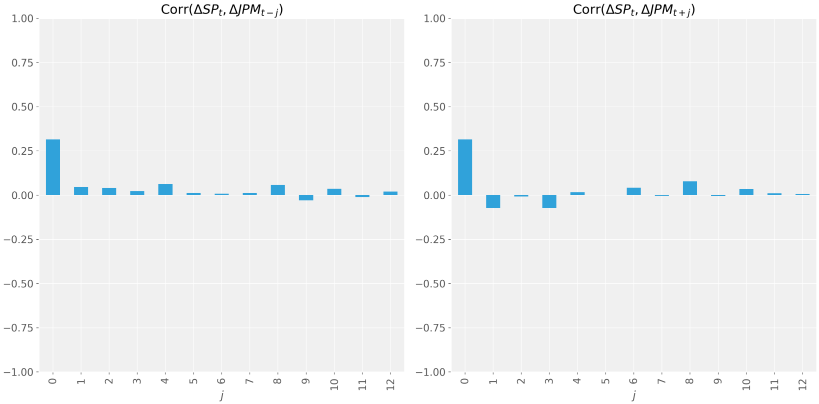
The **cross-correlation function (CCF)** is defined as

$$\rho_{i,j}(k) = \text{Corr}(Y_{it} Y_{jt-k}).$$

Note: Unlike the ACF, the CCF is not necessarily symmetric:

$$\rho_{i,j}(k) \neq \rho_{i,j}(-k)$$

The cross-correlation function



VARs

We can write the joint bivariate VAR(1) as

$$\begin{aligned}\Delta SP_t &= \phi_{10} + \phi_{11}\Delta SP_{t-1} + \phi_{12}\Delta JPM_{t-1} + e_{1,t} \\ \Delta JPM_t &= \phi_{20} + \phi_{21}\Delta SP_{t-1} + \phi_{22}\Delta JPM_{t-1} + e_{2,t}\end{aligned}$$

- Both series depend on their own lagged values (as in an AR process)
- Both series depend on the lagged realization of the other variable(s)
- The residuals $\mathbf{e}_t = [e_{1t}, e_{2t}]$ have a covariance matrix $\boldsymbol{\Sigma}$.
- This system can be written as:

$$\mathbf{Y}_t = \boldsymbol{\Phi}\mathbf{Y}_{t-1} + \mathbf{e}_t$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}$$

Many properties of AR processes extend to VARs:

- Forecasts:

$$E_t(\mathbf{Y}_{t+k}|\mathbf{Y}_t) = \mathbf{\Phi}^k \mathbf{Y}_t$$

two different regressions

- VARs can be estimated OLS equation by equation:

Regression 1: $\Delta SP_t = \phi_1 + \phi_{11}\Delta SP_{t-1} + \phi_{12}\Delta JPM_{t-1} + e_{1,t}$

Regression 2: $\Delta JPM_t = \phi_{20} + \phi_{21}\Delta SP_{t-1} + \phi_{22}\Delta JPM_{t-1} + e_{2,t}$

VAR estimation in `statsmodels`:

```
_fitdX = sm.tsa.VAR(dX).fit(1, trend = "c")
        var_fitdX.summary()

```

VAR(1) ESTIMATION

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:             Fri, 07, Apr, 2017
Time:                     13:13:07
--------------------------------------------------------------------
No. of Equations:         2.00000    BIC:                   0.711410
Nobs:                     1760.00    HQIC:                  0.699647
Log likelihood:          -5598.28    FPE:                    1.99921
AIC:                     0.692751    Det(Omega_mle):         1.99241
--------------------------------------------------------------------
Results for equation dSP
==========================================================================
             coefficient       std. error           t-stat            prob
--------------------------------------------------------------------------
const           0.013925         0.020822            0.669           0.504
L1.dSP          0.039470         0.025093            1.573           0.116
L1.dJPM         0.016906         0.012797            1.321           0.187
==========================================================================

Results for equation dJPM
==========================================================================
             coefficient       std. error           t-stat            prob
--------------------------------------------------------------------------
const           0.052430         0.040722            1.288           0.198
L1.dSP         -0.105707         0.049074           -2.154           0.031
L1.dJPM        -0.056634         0.025028           -2.263           0.024
==========================================================================

Correlation matrix of residuals
             dSP      dJPM
dSP     1.000000  0.322104
dJPM    0.322104  1.000000
```

Causality in a bivariate VAR:

$$\begin{aligned} \Delta SP_t &= \phi_{11}\Delta SP_{t-1} + \phi_{12}\Delta JPM_{t-1} + e_{1,t} \\ \Delta JPM_t &= \phi_{21}\Delta SP_{t-1} + \phi_{22}\Delta JPM_{t-1} + e_{2,t} \end{aligned}$$

- Suppose that $\phi_{12} = 0 : \Delta SP_t = \phi_{11}\Delta SP_{t-1} + e_{1,t}$

  $\longrightarrow$ ΔSP_t is not affected by ΔJPM_{t-1}. In such a case, we say that ΔJPM_{t-1} **does not Granger-cause (or just cause)** ΔSP_t.
- It is important to understand that Granger-causality gives us the timing but not (necessarily) the economic causality.
- This is a very common mistake in academia, in practice, and in everyday life.

Test for Granger-causality

- We run the two regressions
  - For the hypothesis: "ΔJPM_{t-1} Granger-causes ΔSP_t", we test $\phi_{12} = 0$
  - For the hypothesis: "ΔSP_{t-1} Granger-causes ΔJPM_t", we test $\phi_{21} = 0$
- Note: In VAR(p), we test whether all coefficients are jointly zero
- It is always a good idea to start an empirical work with some background Granger-causality tests, be it only to get a feel for the data.
- But it is inappropriate economic causality based on that evidence alone!

GRANGER CAUSALITY IN $\mathbf{Y} = (\Delta SP, \Delta JPM)$

?

Lag	Null Hypothesis	F-statistics	Prob.
1	JPM does not Granger Cause SP	1.75	0.19
1	SP does not Granger Cause JPM	4.64	0.03
2	JPM does not Granger Cause SP	2.01	0.13
2	SP does not Granger Cause JPM	2.79	0.06
10	JPM does not Granger Cause SP	1.46	0.15
10	SP does not Granger Cause JPM	2.60	0.00

STATISTICAL (GRANGER) CAUSALITY VS. STRUCTURAL (ECONOMIC) CAUSALITY

- Example: Martian MFE students are asked to investigate what causes rain on earth. They observe that when people bring umbrellas in the morning, it tends to start raining in the afternoon.
- Can they conclude that people bringing umbrellas causes rain?
- No! It is the other way around! The anticipation of rain makes people bring their umbrellas.
- In this example, the statistical causality runs from "umbrellas" to "rain"
- The structural causality runs from "rain" to "umbrellas".
- Problem: In empirical work, if we don't have a model (or more information about the system), we cannot distinguish between the two alternatives.
  - No amount of data can settle the issue
  - We need a model
  - We need more data, i.e. an instrument

ANOTHER EXAMPLE

- Fact: The effect of the Fed actions on the stock market is negative. In other words, a contractionary monetary policy (aimed at curbing inflation) will result in lower returns for some time in the future.
- Question: Why is this so? There are two possibilities
  - Fed has a better forecast of the state of the economy, but its policy has no real effect on stock fundamentals (the umbrella). This is only Granger-causation without structural effect.
  - Fed has an impact on the economy and by contracting the economy, cash flows go down, returns decrease (the cloud). This is Granger-causation and a structural effect.

- VARMA(p, q)

$$
\begin{aligned}
\mathbf{\Phi}(L)\mathbf{Y}_t &= \mathbf{\Theta}(L)e_t \\
\mathbf{\Phi}(L) &= \mathbf{I} - \Phi_1 L - \Phi_2 L^2 - \dots \\
\mathbf{\Theta}(L) &= \mathbf{I} + \Theta_1 L + \Theta_2 L^2 + \dots,
\end{aligned}
$$

$$
\Phi_j = \begin{bmatrix} \phi_{j,yy} & \phi_{j,yz} \\ \phi_{j,zy} & \phi_{j,zz} \end{bmatrix},
$$

- VARMA(p, q) models can be inverted just like ARMA(p, q) :

$$(\mathbf{I}-\mathbf{\Phi}L)\,Y_t=\mathbf{e}_t \Leftrightarrow x_t=(\mathbf{I}-\mathbf{\Phi}L)^{-1}\,e_t=\sum_{j=0}^{\infty}\mathbf{\Phi}^j\mathbf{e}_{t-j}.$$

- VAR(1) forecasts: $E_t[\mathbf{Y}_{t+j}]=\mathbf{\Phi}^j x_t$

Theorem 2.

A VARMA(p, q) process

$$\begin{aligned}
\mathbf{\Phi}(L)\,\mathbf{Y}_t &= \mathbf{\Theta}(L)\,\mathbf{e}_t \\
\mathbf{\Phi}(L) &= \mathbf{I} - \Phi_1 L - \Phi_2 L^2 - \ldots \\
\mathbf{\Theta}(L) &= \mathbf{I} + \Theta_1 L + \Theta_2 L^2 + \ldots,
\end{aligned}$$

is stationary if

$$det(\mathbf{I} - \Phi_1 z - \Phi_2 z^2 - \ldots) \neq 0 \quad for\ |z| < 1.$$

roots of determinant are outside the unit circle

Example 1.

$$\mathbf{Y}_t = \begin{pmatrix} .5 & 0 & 0 \\ .1 & .1 & .3 \\ 0 & .2 & .3 \end{pmatrix} \mathbf{Y}_{t-1} + \mathbf{e}_t$$

$$\det\left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} .5 & 0 & 0 \\ .1 & .1 & .3 \\ 0 & .2 & .3 \end{pmatrix} z \right) = (1 - .5z)(1 - .4z - .03z^2)$$

with roots $z_1 = 2, z_2 = 2.12, z_3 = -15.48.$ Hence $\mathbf{Y}_t$ is stationary.

Summary: VARs

- Univariate AR models generalize to multivariate VAR models
- VARs can be estimated by OLS equation by equation
- Granger causality vs. structural causality

Nonlinear Models

Models that cannot be estimated by OLS

NONLINEAR MODELS

- Regressions are the go-to tool for estimating linear models: simple, robust, easy to add bells and whistles
- Limited to linear models
- How can nonlinear models be estimated?
- Examples of nonlinear models:
  GARCH, stochastic volatility, term structure models
- We will discuss maximum likelihood (MLE) and generalized method of moments (GMM) techniques
- Recall: We covered MLE and MM (but not GMM) in the pre-program Stats class.

MA(q):

$$x_t = e_t + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q}$$

We do not observe e_{t-j}, hence we cannot specify a regression.

EXAMPLE: ARCH

Autoregressive Conditional Heteroskedasticity (ARCH) model:

$$y_t = c + \phi y_{t-1} + u_t$$

$$u_t = \sqrt{h_t} v_t \quad \Rightarrow \text{problem}$$

$$h_t = \zeta + \alpha\, u_{t-1}^2$$

$$v_t \sim N(0, 1)$$

past shocks affect volatility!

If $\alpha = 0$ the variance of u_t is constant and we can apply OLS.

If $\alpha \neq 0$, does the model satisfy the OLS assumptions? Is the OLS $\hat{\phi}$ unbiased and/or consistent?

How do we compute the standard error of $\hat{\phi}$?

How can we estimate α?

Outline

1. Dealing with heteroskedasticity: GLS
2. When OLS is inconsistent: Regressor endogeneity
   - Omitted variables
   - Measurement error
   - Simultaneous equations
   - Solution: Instrumental variables (IV)
3. VARMA(p, q)
4. **Maximum likelihood**
   - **MLE estimation of MA(1) process**
   - **MLE estimation of regime-switching models**

Maximum likelihood estimation (MLE)

- The most common method for estimating parameters in a parametric model is the **maximum likelihood method**.
- ML estimators have many appealing properties
- In particular: ML estimators are **optimal** in the sense that they have the smallest possible variance in the set of "well behaved" estimators
- Drawback: They rely on strong distributional assumptions
  $\longrightarrow$ MLE is efficient but not necessarily robust!

Maximum likelihood estimation (MLE)

Let $X_1, \ldots, X_n$ be an iid sample with pdf $f(x_i; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a $(k \times 1)$ vector of parameters that characterize $f(x_i; \boldsymbol{\theta})$.

The *joint density* is, by independence, equal to the product of the marginal densities

$$f(x_1, \ldots, x_n; \boldsymbol{\theta}) = f(x_1; \boldsymbol{\theta}) \times \ldots \times f(x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}).$$

The likelihood function

The **likelihood function** is defined as the joint density treated as a function of the parameters $\boldsymbol{\theta}$ given the data:

$$L(\boldsymbol{\theta}|x_1,\ldots,x_n) = f(x_1,\ldots,x_n;\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}).$$

Notice that $L(\boldsymbol{\theta}|x_1,\ldots,x_n)$ is a function of $\boldsymbol{\theta}$ given the data $x_1,\ldots,x_n$.

If x_i is iid, the log-likelihood has the particularly simple form

$$\log L(\boldsymbol{\theta}|\mathbf{x}) = \log\left(\prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})\right) = \sum_{i=1}^{n} \log f(x_i;\boldsymbol{\theta})$$

The MLE estimator maximizes the (log) likelihood function:

$$\hat{\boldsymbol{\theta}}_{mle} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log L(\boldsymbol{\theta}|\mathbf{x})$$

$$\frac{\partial \log L(\hat{\boldsymbol{\theta}}_{mle}|\mathbf{x})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

The vector of derivatives is called the **score vector**:

$$S(\boldsymbol{\theta}|\mathbf{x}) = \frac{\partial \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}} = \frac{\partial \log L(\boldsymbol{\theta}|x_i)}{\partial \boldsymbol{\theta}}.$$

is zero in $\hat{\theta}_{MLE}$

Since x_i are iid:

$$S(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \frac{\partial \log f(x_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} S(\boldsymbol{\theta}|\mathbf{x}_i)$$

$$= \frac{\partial}{\partial \theta}\left[\log \prod f(x_i,\theta)\right] = \frac{\partial}{\partial \theta} \sum \log f(x_i,\theta) = \sum S(\theta, x_i)$$

sum of score wrt each x_i because iid!

THE HESSION SECOND DERIVATIVE MATRIX

The curvature of the log-likelihood is measured by its second derivative matrix (*Hessian*) (named after Ludwig Hesse, 19th cent. mathematician)

$$H(\boldsymbol{\theta}|\mathbf{x}) = \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1} & \cdots & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_1} & \cdots & \frac{\partial^2 \log L(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k} \end{bmatrix}$$

Since x_i are iid:

$$H(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \frac{\partial^2 \log f(x_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \sum_{i=1}^{n} H(\boldsymbol{\theta}|\mathbf{x}_i)$$

sum of Hessian wrt each x_i!

$$I(\boldsymbol{\theta}|\mathbf{x}) = -\sum_{i=1}^{n} E[H(\boldsymbol{\theta}|x_i)] \overset{\text{iid}}{=} -nE[H(\boldsymbol{\theta}|x_i)] = nI(\boldsymbol{\theta}|x_i)$$

- expected value of Hessian

Now we are ready to derive the asymptotic properties of the MLE.

$$I(\theta|x) = nI(\theta|x_i)$$

Theorem 3.

Let $X_1, \dots, X_n$ an iid sample with pdf $f(x_i; \boldsymbol{\theta})$ and let $\boldsymbol{\theta}_\star$ denotes the true value of the parameter $\boldsymbol{\theta}$. Then under certain regularity conditions, the ML estimator of $\boldsymbol{\theta}$ has the following asymptotic properties:

1. $\widehat{\boldsymbol{\theta}}_{mle}$ is ***consistent***: $\widehat{\boldsymbol{\theta}}_{mle} \xrightarrow{p} \boldsymbol{\theta}_\star$
2. $\widehat{\boldsymbol{\theta}}_{mle}$ is ***asymptotically normal***:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{mle} - \boldsymbol{\theta}_\star) \xrightarrow{d} N(0, I(\boldsymbol{\theta}|x_i)^{-1})$$

Suppose that $X_1, \dots, X_n \sim N(\theta, \sigma^2)$. The MLE is $\widehat{\theta}_n = \overline{X}_n$.

Another reasonable estimator is the sample median $\widetilde{\theta}_n$.

The MLE satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

It can be proved that the median satisfies

$$\sqrt{n}(\widetilde{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2 \frac{\pi}{2}).$$

This means that the median converges to the right value but has a larger variance than the MLE .

More generally, consider two estimators T_n and U_n and suppose that

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, t^2).$$

and that

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, u^2).$$

We define the **asymptotic relative efficiency** of U to T by $ARE(U, T) = t^2/u^2$.

In the Normal example, $ARE(\widetilde{\theta}_n, \widehat{\theta}_n) = 2/\pi = 0.63$.

Interpretation: If you use the median, you are only using about 63 percent of the available information.

Optimality of MLE

Theorem 4.

If $\widehat{\theta}_n$ is the MLE and $\widetilde{\theta}_n$ is any other estimator then

$$ARE(\widetilde{\theta}_n, \widehat{\theta}_n) \leq 1.$$

Thus, the MLE has the smallest (asymptotic) variance and we say that MLE is ***efficient or asymptotically optimal****.*

Note: This result holds if the assumed model being correct. If the model is wrong, the MLE may no longer be optimal (and/or asymptotically normal).

LIMITATIONS OF MLE

- ML depends on strong assumptions on the distribution of the underlying data (recall that we made no distributional assumption in MM)
- The MLE relies on iid data
- The MLE estimator might not exist (DGS Example 7.5.8)
- The MLE estimator might not be unique (DGS example 7.5.9)
- The likelihood function can be very complicated. Often it is has to be evaluated numerically.
- Finding the global maximum numerically can be difficult

MLE Examples

1. The classical regression model with normal errors
2. Normal regression with time-varying variance (ARCH)
3. Regression with varying coefficients: Regime-switching models

- Classical regression model with normal errors and $(y_i, \mathbf{x}_i')$ are i.i.d.:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i$$

$$y_i \mid \mathbf{x}_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$$

- The log-likelihood function for the normal regression model is

$$\log L(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n} \log\left(\frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}\right)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\mathrm{SSE}_n(\boldsymbol{\beta})$$

- Thus maximizing log likelihood is equivalent to minimizing $\mathrm{SSE}_n(\beta)$, i.e.

$$\widehat{\boldsymbol{\beta}}_{\mathrm{mle}} = \widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$$

The classical regression model with normal errors

- We can also find the MLE for σ^2. Substituting $\widehat{\boldsymbol{\beta}}$ into the log likelihood:

$$\log L(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\text{SSE}_n(\widehat{\boldsymbol{\beta}}_{\text{mle}})$$

- The first order condition with respect to σ^2 is given by

$$\frac{\partial}{\partial\sigma^2}\log L(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \sigma^2) = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\text{SSE}_n(\widehat{\boldsymbol{\beta}}_{\text{mle}}) = 0$$

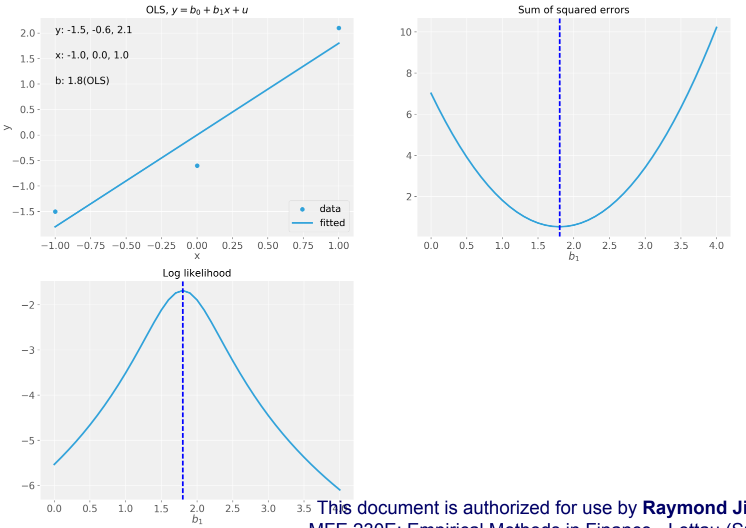
- Thus the MLE for σ^2 is

$$\hat{\sigma}^2_{\text{mle}} = \frac{\text{SSE}_n(\widehat{\boldsymbol{\beta}}_{\text{mle}})}{n} = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2$$

and the maximized log likelihood is

$$\log L(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\sigma}^2_{\text{mle}}) = -\frac{n}{2}(\log(2\pi)+1) - \frac{n}{2}\log(\hat{\sigma}^2_{\text{mle}})$$

MLE=OLS



OLS, $y = b_0 + b_1 x + u$
y: -1.5, -0.6, 2.1
x: -1.0, 0.0, 1.0
b: 1.8(OLS)
data
fitted
x
y
Sum of squared errors
b_1
Log likelihood
b_1

So far we assumed that the x_i are independent, so that

$$f(x_1, \dots, x_n; \boldsymbol{\theta}) = f(x_1; \boldsymbol{\theta}) \cdots f(x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}).$$

In finance, this assumption will usually be violated.

Trick: **Conditional densities**

$$\begin{aligned}
f(x_1) &: \text{density of initial observation} \\
f(x_1, x_2) &= f(x_2|x_1)f(x_1) \\
f(x_1, x_2, x_3) &= f(x_3|x_2, x_1)f(x_2|x_1)f(x_1) \\
&\vdots \\
f(x_1, \dots, x_n) &= f(x_n|x_{n-1}, \dots, x_1)f(x_{n-1}|x_{n-2}, \dots, x_1)\dots f(x_2|x_1)f(x_1)
\end{aligned}$$

ARCH(1) model:

$$y_t = c + \phi y_{t-1} + u_t$$

$$u_t = \sqrt{h_t} v_t \quad = v_t \sqrt{\zeta + \alpha u_{t-1}^2}$$

$$h_t = \zeta + \alpha u_{t-1}^2$$

$$v_t \sim N(0, 1)$$

If $\alpha = 0$ the variance of u_t is constant and we can apply OLS.

If $a \neq 0$, the model is nonlinear.

Start with y_1 and y_2: Since (y_1, y_2) are jointly normal, the marginals are also normal

$$f_{y_2|y_1} \sim N\left((c + \phi y_1), h_2\right)$$

$$f_{y_2|y_1} = \frac{1}{\sqrt{2\pi\left(\zeta + \alpha u_1^2\right)}} \exp\left[\frac{-(y_2 - c - \phi y_1)^2}{2\left(\zeta + \alpha u_1^2\right)}\right]$$

$y_2 = c + \phi y_1 + u_2$

$u_2 = \sqrt{h_2}\, v_2$

$= \sqrt{\zeta + \alpha u_1^2}$

Solve forward:

$$f_{y_3|y_2} = \frac{1}{\sqrt{2\pi\left(\zeta + \alpha u_2^2\right)}} \exp\left[\frac{-(y_3 - c - \phi y_2)^2}{2\left(\zeta + \alpha u_2^2\right)}\right]$$

ARCH, CONT.

The unknown parameters are collected in $\boldsymbol{\theta} = (c, \phi, \zeta, \alpha)$

Then, the conditional log likelihood can be written as

$$
\begin{aligned}
L(\boldsymbol{\theta}|y_1) &= \sum_{t=2}^{T} \log f_{y_t|y_{t-1}} \\
&= -\frac{(T-1)}{2}\log(2\pi) - \frac{1}{2}\sum_{t=2}^{T}\log\left(\zeta + \alpha u_{t-1}^2\right) \\
&\quad -\sum_{t=2}^{T}\frac{(y_t - c - \phi y_{t-1})^2}{2\left(\zeta + \alpha u_{t-1}^2\right)}
\end{aligned}
$$

We can maximize $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta} = (c, \phi, \zeta, \alpha)$ and find the estimates that maximize the probability of having observed such a sample.

Details later when we talk about ARCH/GARCH volatility models.

- Consider the MA(1) process with normal errors:

$$y_t = e_t + \psi e_{t-1},\ e_t \sim N(0, \sigma^2)$$
$$\Rightarrow\ y_t | e_{t-1} \sim N(\psi e_{t-1}, \sigma^2)$$

- Assume $e_0 = 0$
- Hence

$$e_1(\psi) = y_1$$
$$e_2(\psi) = y_2 - \psi y_1$$
$$e_3(\psi) = y_3 - \psi y_2 + \psi^2 y_1$$
$$\vdots$$
$$e_t(\psi) = y_t - \psi y_{t-1} + \ldots + (-\psi)^{t-1} y_1$$

MLE ESTIMATION OF MA(1) PROCESS

- Since $y_t|e_{t-1} \sim N(\psi e_{t-1}, \sigma^2)$:

$$f_{y_t|e_{t-1}} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(y_t - \psi e_{t-1}(\psi)\right)^2}{2\sigma^2}\right]$$

- The log-likelihood function is

$$\begin{aligned} L\left(\boldsymbol{\theta}|y_0\right) &= -\frac{T}{2}\log\left(2\pi\right) - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_t - \psi e_{t-1}(\psi)\right)^2 \\ &= -\frac{T}{2}\log\left(2\pi\right) - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T} e_t(\psi)^2 \end{aligned}$$

- `statsmodels.tsa.arima_model.ARMA` uses MLE.
- Problem set: Numerically maximize $L\left(\boldsymbol{\theta}|y_0\right)$.

MLE ESTIMATION OF REGIME-SWITCHING MODELS

So far: Parameter are constant over the entire sample

Alternative: **Regime-switching** models (RGS)

- Parameters might switch back and forth between certain values
- Example: Political party in power
- Popular model: Regime switching of Hamilton (Econometrica, 1989)
- Parameters depend on a (potentially unobservable) state variable s_t
- The state s_t follows a pre-specified process that can be estimated

- Example: Hamilton RGS model

$$y_t = \alpha_{s_t} + \phi_{s_t} y_{t-1} + \sigma_{s_t} e_t,$$

where s_t follows a **Markov chain**: $s_t = 0, 1$.

- See Stats notes for an introduction to Markov chains.
- All parameters might depend on s_t or only a subset.
- Goal: Estimate $\alpha_0, \alpha_1, \phi_0, \phi_1, \sigma_0, \sigma_2$ and the process for s_t.
- Hamilton ch. 22 covers RS models in detail.

Example: MLE estimation of regime-switching models

- A **Markov chain** is a sequence of discrete rvs $s_1, s_2, \ldots$ s.t.

$$\mathbf{Pr}\left(s_{n+1}=j|s_n=i,\ldots,s_1=k\right)=\mathbf{Pr}\left(s_{n+1}=j|s_n=i\right)=p_{ij}$$

- The matrix $P=[p_{ij}]$ is called the **transition matrix**.
- We derived properties of Markov chains in the Stats class:
  - Forecasting
  - Ergodicity
  - Unconditional and conditional distributions
  - MLE estimation
- Python package: `from quantecon.markov import MarkovChain`

MLE ESTIMATION OF REGIME-SWITCHING MODELS

- Model with 2-states for the mean (ϕ and σ are constant):

$$y_t = \alpha_{s_t} + \phi y_{t-1} + \sigma e_t$$

- We as econometrician observe only $\{y_1, y_2, \ldots, y_T\}$
- Objective: Estimate
  - parameters of the model $\theta = (\sigma, \phi, \alpha_1, \alpha_2, p_{11}, p_{22})$
  - the probability of being in state 1 or 2 at each t:

$$\xi_{j,t} = \mathbf{Pr}(s_t = j | \Omega_t, \theta)$$

  where $\Omega_t = \{y_1, y_2, \ldots, y_t\}$

- Hamilton proposes a recursive 2-step procedure:
- Idea: Bayesian updating (recall Stats class)

Step 1: Suppose we know parameters θ with certainty, let's calculate $\xi_{j,t}$ recursively via Bayesian updating (recall Stats class):

1. Fix starting values $\xi_{j,0}$
2. Take first observation y_1 and compute $\xi_{j,1}$
3. Given $\xi_{j,1}$, take y_2 and compute $\xi_{j,2}$, and so on.

Step 2: $\xi_{j,t}$ are functions of θ. So, given the $\xi_{j,0}, \ldots, \xi_{j,T}$, estimate θ by MLE.

Step 1: Given ξ_{jt-1} and y_t, compute ξ_{jt}:

1. The densities of y_t under the two regimes are:

$$\eta_{jt} = f(y_t | s_t = j, \Omega_{t-1}, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_t - \alpha_j - \phi y_{t-1})^2}{2\sigma^2}\right]$$

2. We need the density of y_t.
   - If the state at time $t-1$ is i, then $p_{ij}\eta_{jt}$ is the expected density at t.
   - Recall that $\xi_{i,t-1}$ is the probability of being in state i at $t-1$.
   - Thus, the conditional density of the t-th observation is therefore

$$f(y_t | \Omega_{t-1}, \theta) = \sum_{i=1}^{2}\sum_{j=1}^{2} \xi_{i,t-1} p_{ij} \eta_{jt}$$

- Finally, the marginal probability of being in state j at time t is

$$\xi_{j,t} = \frac{\sum_{i=1}^{2} p_{ij}\xi_{i,t-1}\eta_{jt}}{f(y_t|\Omega_{t-1},\theta)}$$

Step 2:

- The recursive marginal densities ξ_{jt} were computed assuming that we know all model parameters $\theta = (\sigma, \phi, \alpha_1, \alpha_2, p_{11}, p_{22})$
- Given $\xi_{j1}, \ldots, \xi_{jT}$ for $j = 1, 2$, the parameters can be estimated by numerically maximizing the conditional likelihood function of the observed data:

$$\log f(y_1, \ldots, y_T | y_0, \theta) = \sum_{t=1}^{T} \log f(y_t | \Omega_{t-1}, \theta)$$

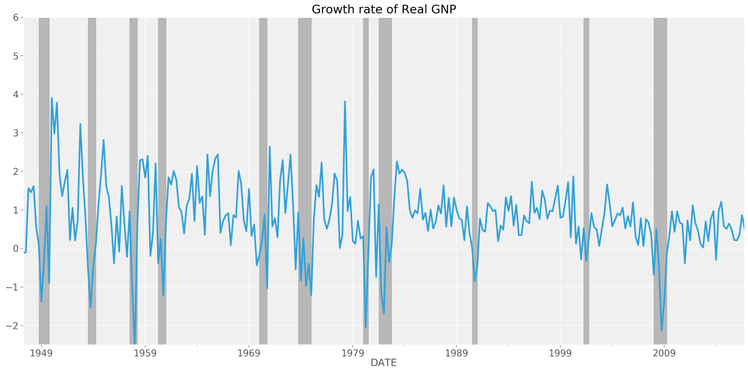
The likelihood function depends on the initial value ξ_{j0}. ξ_{j0} can be set to

- the unconditional probability of state j
- set $\xi_{j0} = 1/2$
- estimate ξ_{j0} by MLE as an additional parameter

Note:

- ξ_{jt} is the estimated probability of being in state j at time t given data $y_1, \ldots, y_t$. The ξ_t's' is called the **filtered**, or "real-time", probabilities.
- We can compute the probability of being in state j at time t given the entire dataset $y_1, \ldots, y_T$. These **smoothed**, or "ex-post", probabilities are denoted $\xi_{jt|T}$.

Regime switching models: GDP growth



Shaded periods: "Official" recessions dated by the NBER ex-post.

$y_t = \Delta \log \mathrm{GDP}_t$: Real quarterly postwar data

$$y_t = \alpha_{s_t} + \sum_{p=1}^{4} \phi_p y_{t-p} + e_t$$

```python
df = data.DataReader('GDPC1', 'fred',
        start=datetime.datetime(1947, 1, 1),
        end=datetime.datetime(2017, 3, 1))
df['dlgnp'] = np.log(df['GDPC1']).diff()*100

mod = sm.tsa.MarkovAutoregression(df['dlgnp'], k_regimes=2, order=4,
        switching_ar=False)
res = mod.fit()
```

Model results

Regime 0 parameters

	coef	std err	z	P>|z|	[0.025	0.975]
const	−1.2561	0.268	−4.683	0.000	−1.782	−0.730

Regime 1 parameters

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.8840	0.075	11.797	0.000	0.737	1.031

Non−switching parameters

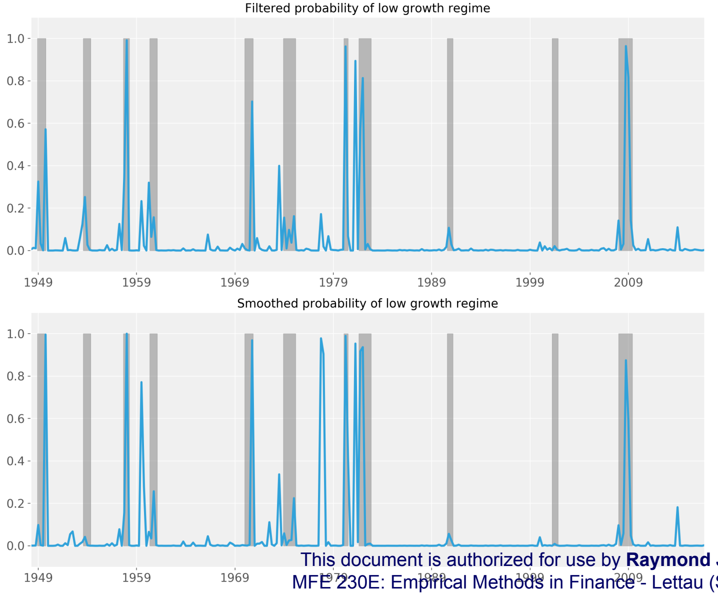
	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	0.5264	0.051	10.259	0.000	0.426	0.627
ar.L1	0.4233	0.068	6.226	0.000	0.290	0.557
ar.L2	0.2448	0.072	3.405	0.001	0.104	0.386
ar.L3	−0.1846	0.076	−2.431	0.015	−0.333	−0.036
ar.L4	−0.0953	0.069	−1.383	0.167	−0.230	0.040

Regime transition parameters

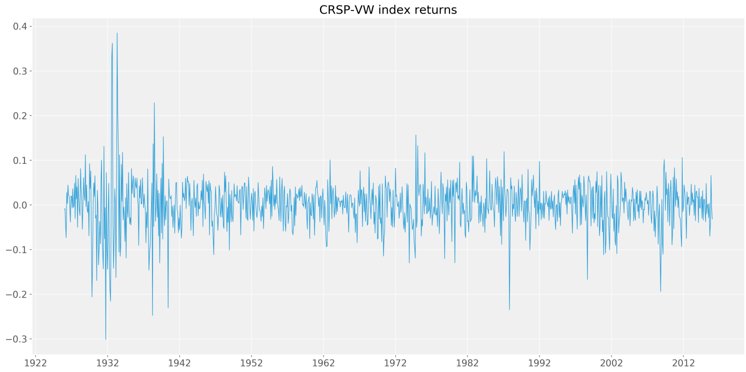
	coef	std err	z	P>|z|	[0.025	0.975]
p[0−>0]	0.2637	0.147	1.795	0.073	−0.024	0.552
p[1−>0]	0.0401	0.015	2.639	0.008	0.010	0.070

Implied duration of regimes: R0 1.36, R1 24.93 (quarters)

Regime switching models: GDP growth



Regime switching models: Stock return volatility



Model with 3 volatility regimes:

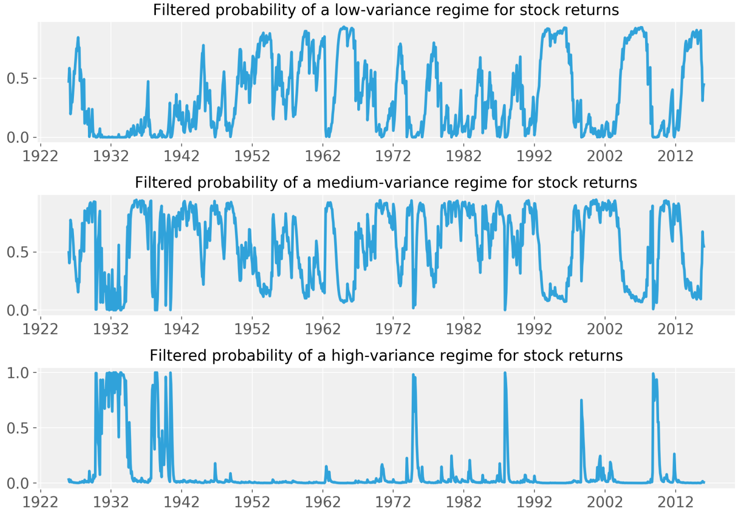
$$r_t = \mu + e_t$$

$$e_t = N(0, \sigma^2_{S_t})$$

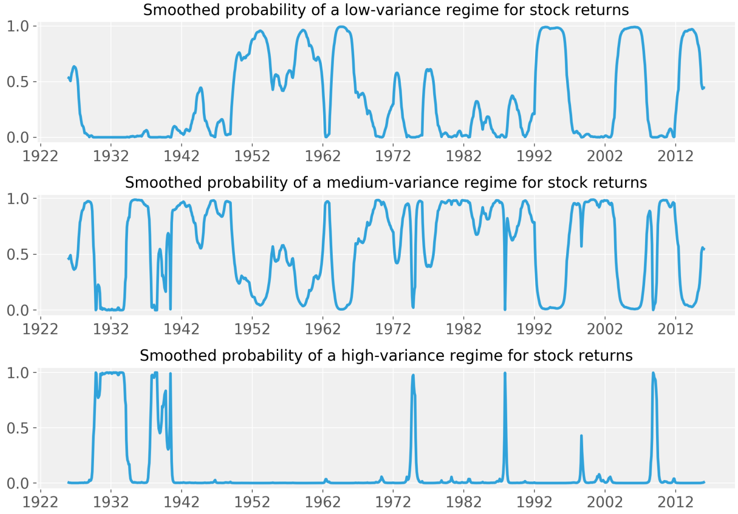
```
mod = sm.tsa.MarkovRegression(df, k_regimes=3, trend='nc',
        switching_variance=True)
        res = mod.fit()

```

Regime switching models: Volatility – filtered probabilities



Regime switching models: Volatility – smoothed probabilities



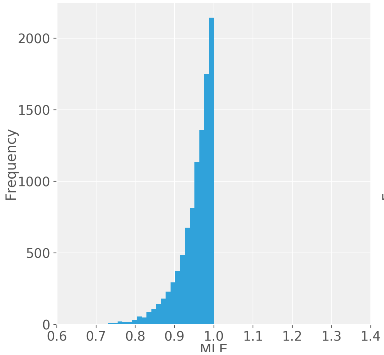
MFE Stats class:

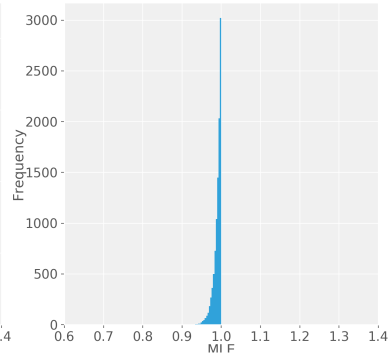
Let $X_1, ...X_n \sim Unif[0, \theta]$. Recall that

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & otherwise. \end{cases}$$

$$\widehat{\theta}_{mle} = \max(X_1, \ldots, X_n) < \theta$$

The histograms of 10,000 MLE estimators from 20 and 100 draws





The MLE estimator is **NOT** asymptotically normal!

MLE: Summary

- Maximum likelihood principle:
  - Make a distributional assumption about the data.
  - Use conditioning to write the joint likelihood function recursively.
  - Maximize the log likelihood function with respect to the parameters
- Potential issues:
  - Strong distributional assumptions are required
  - We had to make an assumption about the dependence in the series
  - We need to specify the unconditional distribution of the first observation
  - If the model is not correctly specified, MLE looses its optimality property
- MLE is efficient but often not robust!
- But sometimes, MLE is the only way to go.
- MLE is particularly appealing if we know the distribution of the series. Most other deficiencies can be circumvented.

