ARMA Handout

Jialin Yu

1 Optimal Prediction

Consider two random variables X and Y. Suppose we want to predict Y using X where Y is a scalar and X is $k \times 1$ vector, i.e. we want to use $\widehat{Y} = h(X)$ as a predictor of Y. How should $h(\cdot)$ be chosen? The answer depends on the forecaster's loss function. Suppose we solve the problem for squared loss, i.e. we choosen $h(X)$ to minimize the mean square prediction error

$$E\left[\left(Y - \widehat{Y}\right)^2\right]$$

In this case the optimal $h(X)$ is

$$h(X) = E(Y|X)$$

1.1 Linear Predictors

Now, suppose we restrict the predictor to be linear. That is, we restrict the predictor to be of the form

$$\widehat{Y} = \alpha + X^T B$$

where α is scalar and β is a $k \times 1$ vector. What values of α and β minimize the prediction squared error?

Some useful notation. Let

$$\begin{aligned}
\mu_Y &= E(Y) \\
\mu_X &= E(X) \\
\Sigma_{XX} &= Var(X) \\
\Sigma_{YY} &= Var(Y) \\
\Sigma_{XY} &= Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]
\end{aligned}$$

Then we write the optimal linear prediction problem as

$$\min_{a,b} E\left[\left(Y - a - X^T b\right)^2\right]$$

The first ordere necessary conditions are

$$\begin{aligned} E\left[Y - a - X^T b\right] &= 0 \\ E\left[X\left(Y - a - X^T b\right)\right] &= 0 \end{aligned}$$

which yields the solutions

$$\begin{aligned} \beta &= \Sigma_{XX}^{-1}\Sigma_{XY} \\ \alpha &= \mu_Y - \mu_X^T \beta \end{aligned}$$

Some useful facts about the linear predictor:

- The prediction error $Y - \widehat{Y}$ has a mean of 0. (This follows from the first, first order condition.)
- The prediction error is uncorrelated with X. (This follows from the second, second order condition.)
- $Var\left(Y - \widehat{Y}\right) = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$. (You should verify this.) This formula should look familiar – it is the variance of $Y|X$ when Y and X are jointly normally distributed. (Why must this be true?)

What's the best predictor if the goal is to minimize the following?

$$E\left[\left|Y - \widehat{Y}\right|\right]$$

optimal choice: $\widehat{Y} = median\,(Y)$. Why? Consider the example where Y can possibly take 1,2,3 with 1/3 probability each. Compare this to the case where Y can take 1,2,5 with 1/3 probability each. How about Y can take 0,1,2,5,1000 with 1/5 probability each?

1.2 Optimal Predictors for ARMA processes

Consider the invertible MA process

$$Y_t = c(L)\,\varepsilon_t = \varepsilon_t + c_1\varepsilon_{t-1} + c_2\varepsilon_{t-2} + ...$$

where $\varepsilon_t \sim iid\left(0, \sigma^2\right)$. $c(L) = 1 + c_1 L + c_2 L^2 + ...$ Clearly, the best forecast of Y_t given $\{Y_\tau\}_{\tau=-\infty}^{t-1}$ is

$$\begin{aligned} Y_{t|t-1} &= c_1\varepsilon_{t-1} + c_2\varepsilon_{t-2} + ... \\ &= \left(c_1 + c_2 L + c_3 L^2 + ...\right)\varepsilon_{t-1} \end{aligned}$$

where

$$\varepsilon_{t-i} = c(L)^{-1} Y_{t-i}$$

It is convenient to write $c_1 + c_2 L + c_3 L^2 + ...$ as

$$c_1 + c_2 L + c_3 L^2 + ... = \left[L^{-1} c(L)\right]_+$$

where $(\cdot)_+$ means "ignore negative powers of L in the bracketed polynomial".

Using this notation,

$$Y_{t|t-1} = \left[L^{-1} c(L)\right]_+ \varepsilon_{t-1}$$

so that

$$Y_{t+1|t} = \left[L^{-1} c(L)\right]_+ \varepsilon_t$$

Similarly

$$Y_{t+k|t} = \left[L^{-k} c(L)\right]_+ \varepsilon_t$$

For the ARMA model

$$c(L) = \frac{\theta(L)}{\phi(L)}$$

and thus

$$\begin{aligned} Y_{t+1|t} &= \left[L^{-1}\frac{\theta(L)}{\phi(L)}\right]_+ \varepsilon_t \\ &= \left[L^{-1}\frac{\theta(L)}{\phi(L)}\right]_+ \frac{\phi(L)}{\theta(L)} Y_t \end{aligned}$$

where $\varepsilon_t = \frac{\phi(L)}{\theta(L)} Y_t$.

2 Estimation and Inference in ARMA Models

2.1 Some Background

When observations are auto-correlated, the usual central limit theorem may not apply. Solution? Ergodic central limit theorem. We will first give a technical definition of ergodicity, and then give a condition that can easily verify if ergodicity holds in practice.

- If $x = \{x_0, x_1, x_2, ...\}$ is a real sequence. Let Tx denote the shift operator satisfying

  $$Tx = \{x_1, x_2, x_3, ...\}$$

  A set A of real sequences is called shift invariant when $Tx \in A$ if and only if $x \in A$. Example

  $$A_1 = \{x : \text{for some } k = 1, 2, ..., x_k = x_{k+1} = ... = 0\}$$

  $$A_2 = \left\{x : \lim n^{-1}(x_1 + x_2 + ... + x_n) = b\right\}$$

- A stationary process is said to be ergodic if

  $$P\{(X_0, X_1, ...) \in A\}$$

  is either zero or one whenever A is shift invariant. Though not covered in this course, if you are interested, you can find additional details on ergodicity in [1].

- Why non-ergodicity can be a problem? Consider the following example: with probability 1/2, $(X_0, X_1, ...) \in A_1$ and with probability 1/2, $(X_0, X_1, ...) \in A_2$. Where

$$A_1 = \{X_0 = X_1 = ... = 0\}$$
$$A_1 = \{X_0 = X_1 = ... = 1\}$$

  In this case, does the $\frac{1}{n}\sum_{i=1}^{n} X_i$ converge to the mean of X? No, 50% chance it converges to 0, 50% chance to 1.

- Let Y_t be a covariance stationary process with

$$E(Y_t) = \mu$$

$$E(Y_t - \mu)(Y_{t-j} - \mu) = \gamma_j$$

  If

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty$$

  then $\{Y\}$ is ergodic.

- Strong Ergodic Theorem: Suppose $\{z_t\}$ is strict stationary and ergodic with $E(z_t) = \mu$. Then $T^{-1}\sum_{t=1}^{T} z_t \overset{a.s.}{\to} \mu$. Which is a generalization to SLLN.

- Mean Square Ergodic Theorem: Suppose $\{z_t\}$ is covariance stationary and ergodic with $E(z_t) = \mu$. Then $T^{-1}\sum_{t=1}^{T} z_t \overset{m.s.}{\to} \mu$.

- For many applications, stationarity and ergodicity turn out to have the same requirement. However, stationarity does not imply ergodicity. Example: $Y_t = \varepsilon$ where $\varepsilon \sim N(0,1)$. Clearly Y_t is stationary because Y_t and Y_{t+j} for $j \neq 0$ are literally the same. However, $T^{-1}\sum_{t=1}^{T} Y_t = \varepsilon \nrightarrow 0 = E(Y_t)$ as would have implied by ergodicity.

- If z_t is stationary and ergodic, then so is $x_t = f(z_t)$ for *arbitrary* function f.

- We will need a CLT. Recall a *martingale* and a *martingale difference sequence*.
  - $\{z_t\}$ is a martingale if

$$E\left[z_t | \{z_i\}_{i=1}^{t-1}\right] = z_{t-1} \text{ for } t = 2, 3, ...$$

  - $\{w_t\}$ is a martingale difference sequence (abbreviated *mds*) if

$$E\left[w_t | \{w_i\}_{i=1}^{t-1}\right] = 0 \text{ for } t = 2, 3, ...$$

    (Think of $w_t = z_t - z_{t-1}$ as an explanation of the use of the term "martingale difference")
  - Note that a martingale difference sequence is serially uncorrelated

$$E(w_t w_{t-j}) = 0 \text{ for } j \neq 0 \text{ and all } t$$

- CLT for mds. Let $\{w_t\}$ be a (possibly vector-valued) mds that is stationary and ergodic with $E\left(w_t w_t^T\right) = \Sigma$. Then

$$\sqrt{T}\overline{w} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} w_t \xrightarrow{d} N(0, \Sigma)$$

- Notes:
  1. When $\{w_t\}$ is serially uncorrelated, it may be serially dependent (through higher order moments).
  2. $E\left(w_t w_t^T\right) = \Sigma$ concerns the "unconditional" variance. The conditional variance may be non-constant.

2.2 Estimating Autoregressive Models

Some Preliminaries:

- Suppose y_t follows the MA process

$$y_t = \theta(L)\,\varepsilon_t = \sum_{i=0}^{\infty} \theta_i \varepsilon_{t-i}$$

where ε_t is $\text{iid}\left(0, \sigma^2\right)$. Let

$$\gamma_i = E\left(y_t y_{t-i}\right)$$

denote the i-th autocovariance of $\{y_t\}$. if

$$\sum_{i=0}^{\infty} |\theta_i| < \infty$$

then

$$\sum_{i=0}^{\infty} |\gamma_i| < \infty$$

and the process is stationary and ergodic. Proof: see Hamilton page 69-70.

- Suppose $\phi(L)\,y_t = \varepsilon_t$ where ε_t is $\text{iid}\left(0, \sigma^2\right)$, $\phi(L) = 1 - \phi_1 L - ... - \phi_p L^p$ with roots outside of the unit circle. Then

$$y_t = \theta(L)\,\varepsilon_t$$

with $\sum_{i=0}^{\infty} |\theta_i| < \infty$. (straightforward calculation, to see the intuition, use AR(1) model)

Now, consider the AR model

$$\phi(L)\,y_t = \varepsilon_t$$

Suppose that ε_t is $\text{iid}\left(0, \sigma^2\right)$ and that the roots of $\phi(z)$ are outside the unit circle (this means for the case of AR(1) $y_t = \phi y_{t-1} + \varepsilon_t$ that $|\phi| < 1$). Write the model as

$$y_t = x_t^T \beta + \varepsilon_t$$

where $x_t = (y_{t-1}, y_{t-2}, ..., y_{t-p})$ and $\beta = (\phi_1, \phi_2, ..., \phi_p)$. Then

$$\widehat{\beta} \xrightarrow{p} \beta$$

and

$$\sqrt{T}\left(\widehat{\beta} - \beta\right) \rightarrow N\left(0, V_{\widehat{\beta}}\right)$$

where

$$V_{\widehat{\beta}} = \sigma^2 \Sigma_{xx}^{-1}$$

with $\Sigma_{xx} = E\left(x_t x_t^T\right)$.

Proof: Key Points

- $\{y_t, x_t\}$ is stationary and ergodic
- $\left[E\left(x_t x_t^T\right)\right]_{ij} = E\left(y_{t-i} y_{t-j}\right) = \gamma_{|i-j|}$
- $w_t = \varepsilon_t x_t$, ε_t is independent of x_{t-i} for $i \geq 0$, independent of ε_{t-i} for $i > 0$ and $E\varepsilon_t = 0$. Thus $E\left(\varepsilon_t | \left\{\varepsilon_i, x_i\right\}_{i=1}^{t-1}, x_t\right) = 0$ and so w_t is a mds.
- $E\left(w_t w_t^T\right) = E\left(\varepsilon_t^2 x_t x_t^T\right) = E\left[E\left(\varepsilon_t^2 x_t x_t^T | x_t\right)\right] = \sigma^2 \Sigma_{xx}$

And the results follow from the general results given above.

- AR(1) Example

  $$y_t = \beta x_t + \varepsilon_t$$

  with $\beta = \phi$ and $x_t = y_{t-1}$. Then

  $$\Sigma_{xx} = Var\left(y_{t-1}\right) = \frac{\sigma^2}{1 - \phi^2}$$

  and

  $$\sqrt{T}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, V_{\widehat{\beta}}\right)$$

  with

  $$V_{\widehat{\beta}} = \sigma^2 \Sigma_{xx}^{-1} = 1 - \phi^2$$

so that

$$\widehat{\phi} \overset{a}{\sim} N\left(\phi, \frac{1}{T}\left(1-\phi^2\right)\right)$$

and an approximate 95% confidence interval for ϕ is given by

$$\widehat{\phi} \pm 1.96\left[\frac{1}{T}\left(1-\widehat{\phi}^2\right)\right]^{1/2}$$

(Discuss what happens when ϕ is close to unity – small sample implications).

- Relationship to the Gaussian MLE for AR(p)

Suppose ε_t is $Niid\left(0, \sigma^2\right)$. Let $\boldsymbol{Y}_T = (y_1, ..., y_T)$ and denote the joint density of $\boldsymbol{Y}_T$ by $f\left(\boldsymbol{Y}_T\right)$. Recall that $f(a,b) = f(a|b) f(b)$ for arbitrary random variable a and b. Thus

$$\begin{aligned} f\left(\boldsymbol{Y}_T\right) &= f\left(y_T|\boldsymbol{Y}_{T-1}\right) f\left(\boldsymbol{Y}_{T-1}\right) \\ &= f\left(y_T|\boldsymbol{Y}_{T-1}\right) f\left(y_{T-1}|\boldsymbol{Y}_{T-2}\right) f\left(\boldsymbol{Y}_{T-2}\right) \\ &= ... \\ &= f\left(\boldsymbol{Y}_p\right) \prod_{t=p+1}^{T} f\left(y_t|\boldsymbol{Y}_{t-1}\right) \end{aligned}$$

Since ε_t is $Niid\left(0, \sigma^2\right)$

$$y_t|\boldsymbol{Y}_{t-1} \sim N\left(\mu_{t-1}, \sigma^2\right)$$

where

$$\mu_{t-1} = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p}$$

and so

$$f\left(\boldsymbol{Y}_T\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{T-p}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=p+1}^{T}\left(y_t - \phi_1 y_{t-1} - ... - \phi_p y_{t-p}\right)^2\right] \times f\left(\boldsymbol{Y}_p\right)$$

The least squares estimator is then seen as the MLE after ignoring the term $f(\boldsymbol{Y}_p)$. For stationary models this term is asymptotically negligible.

- – Since $\mu_{t-1} = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} = E(y_t | \{y_{t-i}\}_{i=1}^{\infty})$, $f(y_t|\boldsymbol{Y}_{t-1})$ can be interpreted as the density of the one-step-ahead forecast error in y_t. Hence the factorization

$$f(\boldsymbol{Y}_T) = f(\boldsymbol{Y}_p) \prod_{t=p+1}^{T} f(y_t|\boldsymbol{Y}_{t-1})$$

  is often called the *prediction error decomposition of the likelihood.*

- Estimation of Impulse Response (this is studied before in the asymptotics handout, but note the issue of joint confidence interval)

Thus, let β denote the VAR parameters, and let δ denote the set of impulse response parameters. We know that $\delta = f(\beta)$. Then, using the delta-method, we know that if

$$\sqrt{T}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N(0, V)$$

then

$$\sqrt{T}\left(\widehat{\delta} - \delta\right) \xrightarrow{d} N\left(0, DVD^T\right)$$

where

$$D = \frac{\partial f(\beta)}{\partial \beta^T}$$

As an example, consider the univariate AR(1) model

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

From the usual calculation we know that

$$\sqrt{T}\left(\widehat{\phi} - \phi\right) \xrightarrow{d} N\left(0, 1 - \phi^2\right)$$

Let δ denote the 5'th impulse response, that is

$$\delta = \frac{\partial y_{t+5}}{\partial \varepsilon_t} = \phi^5$$

Then $D = 5\phi^4$,

$$\sqrt{T}\left(\widehat{\delta} - \delta\right) \xrightarrow{d} N\left(0, 25\phi^8\left(1-\phi^2\right)\right)$$

or

$$\widehat{\delta} \overset{a}{\sim} N\left(\delta, \frac{25}{T}\phi^8\left(1-\phi^2\right)\right)$$

Thus, an approximate 95% confidence interval for δ can be formed as

$$\widehat{\phi}^5 \pm 1.96\sqrt{\frac{25\widehat{\phi}^8\left(1-\widehat{\phi}^2\right)}{T}}$$

where $\widehat{\phi}$ is the OLS estimator of ϕ. Note that this gives the approximate marginal distribution of the impulse responses. What is the probability that the confidence band for different impulse reponse lag covers the true impulse response function? How, then, would you calculate the joint distribution for impulse reponse estimates at two different time lags?

2.3 Estimation of MA Models

To concretely focus on some issues, we consider the MA(1) Model

$$y_t = \varepsilon_t - \theta\varepsilon_{t-1}$$

1. GMM

   For simplicity, assume that the variance of ε is known and we are estimating θ. Let $w_t = \varepsilon_t z_t$ with $z_t = y_{t-1}$. (Discuss the use of z_t as instrument. Is there a better instrument? $z_t = (y_{t-1}, y_{t-2})$? Even better, Gaussian Score?) Use GMM

   $$\min V_T\left(\widehat{\theta}\right) = \min \overline{w}\left(\widehat{\theta}\right)^T \widehat{S}^{-1}\overline{w}\left(\widehat{\theta}\right)$$

(note optimal weighting matrix $\widehat{S}$ is used). Write

$$\begin{aligned}\varepsilon_t\left(\widehat{\theta}\right) &= \left(1-\widehat{\theta}L\right)^{-1} y_t \\ &= \left(1-\widehat{\theta}L\right)^{-1}(1-\theta L)\,\varepsilon_t \\ &= \varepsilon_t + \left(\widehat{\theta}-\theta\right)\varepsilon_{t-1} + \left(\widehat{\theta}-\theta\right)\widehat{\theta}\varepsilon_{t-2} + ...\end{aligned}$$

Since $z_t = y_{t-1} = \varepsilon_{t-1} - \theta\varepsilon_{t-2}$, then

$$E\left(w_t\left(\widehat{\theta}\right)\right) = E\left(\varepsilon_t\left(\widehat{\theta}\right)z_t\right) = \sigma^2\left(\widehat{\theta}-\theta\right)\left(1-\widehat{\theta}\theta\right)$$

and

$$\begin{aligned}V_T\left(\widehat{\theta}\right) &= \overline{w}\left(\widehat{\theta}\right)^T \widehat{S}^{-1}\overline{w}\left(\widehat{\theta}\right) \\ &\to \sigma^4\left(\widehat{\theta}-\theta\right)^2\left(1-\widehat{\theta}\theta\right)^2 S^{-1}\end{aligned}$$

which is uniquely minimized at $\widehat{\theta} = \left(\theta, \theta^{-1}\right)$. The minimizer θ^{-1} can be ruled out by imposing the invertibility restriction. This proves consistency. (Recall how to prove consistency in the asymptotics handout).

Now for asymptotic normality (except that the central limit theorems are based on ergodic central limit theorems, the following algebras are similar to what we studied in the asymptotics handout and are listed for formality).

Step 1:

Write

$$\sqrt{T}\frac{\partial V_T(\theta)}{\partial\theta} = 2\left[\frac{\partial\overline{w}(\theta)}{\partial\theta}\right]^T \widehat{S}^{-1}\left[\sqrt{T}\overline{w}(\theta)\right]$$

and note

$$\sqrt{T}\overline{w}(\theta) = \frac{1}{\sqrt{T}}\sum\varepsilon_t y_{t-1} \xrightarrow{d} N\left(0, \sigma^4\left(1+\theta^2\right)\right)$$

(so that the optimal weighting matrix $S = \sigma^4\left(1+\theta^2\right)$.)

Also

$$\begin{aligned}\overline{w}(\theta) &= \frac{1}{T}\sum \varepsilon_t y_{t-1} \\ &= \frac{1}{T}\sum \left[(1-\theta L)^{-1} y_t\right] y_{t-1}\end{aligned}$$

so that

$$\begin{aligned}\frac{\partial \overline{w}(\theta)}{\partial \theta} &= -\frac{1}{T}\sum \left[(1-\theta L)^{-2} L y_t\right] y_{t-1} \\ &= -\frac{1}{T}\sum \left[y_{t-1} + 2\theta y_{t-2} + 3\theta^2 y_{t-3} + ...\right] y_{t-1} \\ \xrightarrow{p} -(\gamma_0 + 2\theta\gamma_1) &= -\sigma^2\left(1+\theta^2-2\theta^2\right) = -\sigma^2\left(1-\theta^2\right)\end{aligned}$$

and thus

$$\sqrt{T}\frac{\partial V_T(\theta)}{\partial \theta} \xrightarrow{d} N(0, A)$$

where

$$A = \left[-2\frac{\sigma^2\left(1-\theta^2\right)}{\sigma^4\left(1+\theta^2\right)}\right]^2 \sigma^4\left(1+\theta^2\right) = 4\frac{\left(1-\theta^2\right)^2}{\left(1+\theta^2\right)}$$

Step 2 is a Taylor expansion

$$0 = \frac{\partial V_T\left(\widehat{\theta}\right)}{\partial \theta} = \frac{\partial V_T(\theta_0)}{\partial \theta} + \frac{\partial^2 V_T(\theta^*)}{\partial^2 \theta}\left(\widehat{\theta} - \theta_0\right)$$

Step 3:

We need to find H which is the limiting value of $\left[\frac{\partial^2 V_T(\theta)}{\partial\theta\partial\theta^T}\right]^{-1}$. Write

$$\frac{\partial^2 V_T(\theta)}{\partial^2 \theta} = 2(a+b)$$

where

$$\begin{aligned}a &= \left[\frac{\partial \overline{w}(\theta)}{\partial \theta}\right]^T S^{-1}\frac{\partial \overline{w}(\theta)}{\partial \theta} \\ b &= \frac{\partial^2 \overline{w}(\theta)}{\partial\theta\partial\theta^T} S^{-1}\overline{w}(\theta)\end{aligned}$$

Now

$$\left[\frac{\partial \overline{w}(\theta)}{\partial \theta}\right]^T S^{-1} \frac{\partial \overline{w}(\theta)}{\partial \theta} \xrightarrow{p} \frac{\left(1-\theta^2\right)^2}{1+\theta^2}$$

from the analysis above. Also

$$\frac{\partial^2 \overline{w}(\theta)}{\partial \theta \partial \theta^T} = \frac{1}{T}\sum \left[2y_{t-2} + 6\theta y_{t-3} + 12\theta^2 y_{t-4} + ...\right] y_{t-1} \xrightarrow{p} 0$$

and

$$\overline{w}(\theta) \xrightarrow{p} 0$$

so that

$$b \xrightarrow{p} 0$$

so that

$$H = \left[2\frac{\left(1-\theta^2\right)^2}{1+\theta^2}\right]^{-1}$$

Putting steps 1-3 together

$$\sqrt{T}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} N\left(0, HAH^T\right)$$

where

$$HAH^T = \frac{1+\theta^2}{\left(1-\theta^2\right)^2}$$

3 Robust Standard Error Estimates

- OLS estimator of (for simplicity, assume x is non-random below)

$$y = x\beta + \varepsilon$$

$$\begin{aligned}\widehat{\beta} &= \left(x'x\right)^{-1} x'y \\ &= \beta + \left(x'x\right)^{-1} x'\varepsilon\end{aligned}$$

$$Var\widehat{\beta} = \left(x'x\right)^{-1} x'Vx \left(x'x\right)^{-1}$$

- OLS standard error
  - Homoskedasticity $Var(\varepsilon_i) = \sigma^2$
  - Zero correlation $Corr(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$

$$\widehat{\beta} = \beta + \left(\frac{1}{n}\sum x_i^2\right)^{-1}\left(\frac{1}{n}\sum x_i\varepsilon_i\right)$$

$$\begin{aligned} Var\left(\widehat{\beta}\right) &= \left(\frac{1}{n}\sum x_i^2\right)^{-2}\left(\frac{1}{n^2}\sum x_i^2\sigma^2\right) \\ &= \sigma^2\left(\frac{1}{n}\sum x_i^2\right)^{-1} \end{aligned}$$

- Relax OLS: OLS + Heteroskedasticity

$$\widehat{\beta} = \beta + \left(\frac{1}{n}\sum x_i^2\right)^{-1}\left(\frac{1}{n}\sum x_i\varepsilon_i\right)$$

$$Var\left(\widehat{\beta}\right) = \left(\frac{1}{n}\sum x_i^2\right)^{-2}\left(\frac{1}{n^2}\sum x_i^2 Var(\varepsilon_i)\right)$$

and estimate $Var(\varepsilon_i)$ via $\widehat{\varepsilon_i}^2$ i.e.

$$\widehat{Var\left(\widehat{\beta}\right)} = \left(\frac{1}{n}\sum x_i^2\right)^{-2}\left(\frac{1}{n^2}\sum x_i^2\widehat{\varepsilon_i}^2\right)$$

- Relax OLS: OLS + heteroskedasticity + clustered correlation. Assume observations $x_{i,t}$ are indexed by two variables i, t. Assume that for given i, $\varepsilon_{i,t}$ are correlated across t. Assume ε_i are uncorrelated across i.

$$\begin{aligned} \widehat{\beta} &= \beta + \left(\frac{1}{n}\sum_{i,t} x_{i,t}^2\right)^{-1}\left(\frac{1}{n}\sum_{i,t} x_{i,t}\varepsilon_{i,t}\right) \\ &= \beta + \left(\frac{1}{n}\sum_{i,t} x_{i,t}^2\right)^{-1}\left(\frac{1}{n}\sum_{i}\sum_{t} x_{i,t}\varepsilon_{i,t}\right) \end{aligned}$$

$$Var\left(\widehat{\beta}\right) = \left(\frac{1}{n}\sum_{i,t} x_i^2\right)^{-2} \left(\frac{1}{n^2}\sum_i Var\left(\sum_t x_{i,t}\varepsilon_{i,t}\right)\right)$$

and estimated by

$$\widehat{Var\left(\widehat{\beta}\right)} = \left(\frac{1}{n}\sum_{i,t} x_i^2\right)^{-2} \left(\frac{1}{n^2}\sum_i \left(\sum_t x_{i,t}\widehat{\varepsilon}_{i,t}\right)^2\right)$$

This standard error estimator allows for heteroskedasticity and correlation among observations in clusters indexed by variable i. One can also switch the role of i and t and obtain s.e. estimators allowing correlation within clusters indexed by variable t.

References

[1] Davidson, James, Stochastic Limit Theory, Oxford University Press.