OLS Regression

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara

Overview

Reference: B. Hansen Econometrics Chapter 4.8-4.17

- In-Sample Prediction errors
  - conditional mean and variance
- Out-of-sample Prediction Errors (Forecast Errors)
  - unconditional MSE
- Heteroskedasticity-Robust Covariance Matrix Estimators
- Standard Errors
- Measures of Fit
- Normal Regression Model

Prediction Errors

- prediction errors differ from residuals
  - residual for observation i

$$\widehat{u}_i = y_i - x_i^{\mathrm{T}} \widehat{\beta}$$

  - prediction error for observation i

$$\widetilde{u}_i = y_i - x_i^{\mathrm{T}} \widehat{\beta}_{(-i)}$$

    - ⋆ observation i is not used to estimate β

- simple construction of prediction errors

$$\widetilde{u}_i = (1 - h_{ii})^{-1} \widehat{u}_i$$

  - $h_{ii} = x_i^{\mathrm{T}} \left(X^{\mathrm{T}} X\right)^{-1} x_i$

Conditional Mean and Variance of Prediction Errors

- vector form

$$\begin{aligned} \widetilde{u} &= M^{*}\widehat{u} = M^{*}M \cdot u \\ M^{*} &= diag\left((1-h_{11})^{-1}, \ldots, (1-h_{nn})^{-1}\right) \end{aligned}$$

- conditional mean

$$\mathbb{E}\left(\widetilde{u}|X\right) = M^{*}M\mathbb{E}\left(u|X\right) = 0$$

- conditional variance

$$Var\left(\widetilde{u}|X\right) = M^{*}MDMM^{*}$$

- under conditional homoskedasticity

$$Var\left(\widetilde{u}|X\right) = M^{*}MM^{*}\sigma^{2}$$

  - variance of *i'th* prediction error

$$\begin{aligned} Var\left(\widetilde{u}_i|X\right) &= \mathbb{E}\left(\widetilde{u}_i^2|X\right) = (1-h_{ii})^{-1}(1-h_{ii})(1-h_{ii})^{-1}\sigma^2 \\ &= (1-h_{ii})^{-1}\sigma^2 \end{aligned}$$

Out of Sample Prediction

- goal: predict y_{n+1}
  - observe x_{n+1} so predict $\mathbb{E}\left(y_{n+1}|x_{n+1}\right)$
  - have $\hat{\beta}$ from sample of n observations
- prediction (also called a forecast)

$$\widetilde{y}_{n+1} = x_{n+1}^{\mathrm{T}}\widehat{\beta}$$

- key measure of accuracy: mean-square forecast error

$$MSFE_n = \mathbb{E}\left(\widetilde{u}_{n+1}^2\right)$$

  - forecast error $\widetilde{u}_{n+1} = y_{n+1} - \widetilde{y}_{n+1}$
  - forecast based on a sample of size n
    - ★ $MSFE_{n-1}$ - forecast based on sample of size $n-1$

Mean-Square Forecast Error

Theorem (MSFE). (A1 is Assumption 1 from lecture 8)
In the heteroskedastic linear regression model (A1)

$$MSFE_n = \sigma^2 + \mathbb{E}\left(x_{n+1}^{\mathrm{T}} V_{\widehat{\beta}} x_{n+1}\right),$$

where $V_{\widehat{\beta}} = Var\left(\widehat{\beta}|X\right)$.
If the errors are homoskedastic (A2)

$$MSFE_n = \sigma^2\left(1 + \mathbb{E}\left(x_{n+1}^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1} x_{n+1}\right)\right).$$

Further, $\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widetilde{u}_i^2$ with $\widetilde{u}_i = y_i - x_i^{\mathrm{T}}\widehat{\beta}_{(-i)}$, is an unbiased estimator of $MSFE_{n-1}$:

$$\mathbb{E}\left(\widetilde{\sigma}^2\right) = MSFE_{n-1}.$$

Mean-Square Forecast Error Theorem Interpretation

1. two components σ^2 and $\mathbb{E}\left(x_{n+1}^{\mathrm{T}} V_{\widehat{\beta}} x_{n+1}\right)$
   1. σ^2 component due to unknown u_{n+1}
   2. $V_{\widehat{\beta}}$ component due to estimation of β
      1. averaged over all realizations of X and x_{n+1}
   3. second component includes $V_{\widehat{\beta}}$, part due to estimation of β
2. $\widetilde{\sigma}^2$ is constucted from $\widehat{\beta}_{(-i)}$, which is calculated from a sample of size $n-1$
   1. unless n is very small, we expect $MSFE_n$ to be close to $MSFE_{n-1}$
      1. hence $\widetilde{\sigma}^2$ should be a reasonable estimator of $MSFE_n$

MSFE Theorem

Covariance Matrix Estimation: Homoskedasticity

- to estimate $V_{\widehat{\beta}} = \left(X^{\mathrm{T}}X\right)^{-1}\sigma^2$

$$\widehat{V}^0_{\widehat{\beta}} = \left(X^{\mathrm{T}}X\right)^{-1} s^2$$

- unbiased

$$\mathbb{E}\left(\widehat{V}^0_{\widehat{\beta}}|X\right) = \left(X^{\mathrm{T}}X\right)^{-1}\mathbb{E}\left(s^2|X\right) = V_{\widehat{\beta}}$$

- substantial bias if the error is heteroskedastic
- suppose $\sigma_i^2 = x_i^2$ and $k = 1$

$$\frac{V_{\widehat{\beta}}}{\mathbb{E}\left(\widehat{V}^0_{\widehat{\beta}}|X\right)} = \frac{\sum_{i=1}^n x_i^4}{\sigma^2\sum_{i=1}^n x_i^2} \approx \frac{\mathbb{E}\left(x_i^4\right)}{\left(\mathbb{E}\left(x_i^2\right)\right)^2} = \kappa$$

- κ is the kurtosis (standardized fourth moment) of x_i
  - if x_i is $\mathcal{N}(0,1)$, $\kappa = 3$
    - true variance is 3 times larger than the expected $\widehat{V}^0_{\widehat{\beta}}$

Covariance Matrix Estimation: Heteroskedasticity

- to estimate $V_{\widehat{\beta}} = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}DX \left(X^{\mathrm{T}}X\right)^{-1}$
  - $D = diag(\sigma_1^2, \ldots, \sigma_n^2)$
  - $\widehat{D}^{ideal} = diag(u_1^2, \ldots, u_n^2)$
- $\widehat{V}_{\widehat{\beta}}^{ideal} = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\widehat{D}^{ideal}X \left(X^{\mathrm{T}}X\right)^{-1}$ is unbiased

$$\begin{aligned} \mathbb{E}\left(\widehat{V}_{\widehat{\beta}}^{ideal}|X\right) &= \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\mathbb{E}\left(\widehat{D}^{ideal}|X\right) X \left(X^{\mathrm{T}}X\right)^{-1} \\ \mathbb{E}\left(u_i^2|X\right) &= \sigma_i^2 \quad \Rightarrow \quad \mathbb{E}\left(\widehat{D}^{ideal}|X\right) = D \end{aligned}$$

- feasible estimators replace u_i^2 with $\widehat{u}_i^2$ (Eicker 1963, White 1980)

Feasible Covariance Matrix Estimators

Heteroskedasticity-Robust Estimators

- no bias correction

$$\widehat{V}_{\widehat{\beta}}^{W} = \left(X^{\mathrm{T}}X\right)^{-1}\left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \widehat{u}_i^2\right)\left(X^{\mathrm{T}}X\right)^{-1}$$

  - yet $\widehat{u}_i^2$ is biased toward zero

- bias-correction (termed Eicker-White)

$$\widehat{V}_{\widehat{\beta}} = \left(\frac{n}{n-k}\right)\left(X^{\mathrm{T}}X\right)^{-1}\left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \widehat{u}_i^2\right)\left(X^{\mathrm{T}}X\right)^{-1}$$

  - correction is ad hoc but preferable to $\widehat{V}_{\widehat{\beta}}^{W}$ (default method in Stata)
  - $X^{\mathrm{T}}DX = \sum_{i=1}^{n} x_i^2 \sigma_i^2$
    - ★ weighted version of $X^{\mathrm{T}}X$

Alternative HR Estimators

- Horn,Horn, Duncan 1975, Stata vce(hc2)

$$\overline{V}_{\widehat{\beta}} = \left(X^{\mathrm{T}}X\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-1}x_i x_i^{\mathrm{T}}\widehat{u}_i^2\right)\left(X^{\mathrm{T}}X\right)^{-1}$$

- Andrews 1991, based on cross-validation, Stata vce(hc3)

$$\widetilde{V}_{\widehat{\beta}} = \left(X^{\mathrm{T}}X\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-2}x_i x_i^{\mathrm{T}}\widehat{u}_i^2\right)\left(X^{\mathrm{T}}X\right)^{-1}$$

- relation among bias corrected estimators
  - because $\left(1-h_{ii}\right)^{-2} > \left(1-h_{ii}\right)^{-1} > 1$

$$\widehat{V}_{\widehat{\beta}}^{W} < \overline{V}_{\widehat{\beta}} < \widetilde{V}_{\widehat{\beta}}$$

  - for matrices $A < B$ means the matrix $B - A$ is positive definite

Bias of HR Estimators (Student Annotation)

Standard Errors

- $\widehat{V}_{\widehat{\beta}}$ is an estimator of the variance of the distribution of $\widehat{\beta}$
- A standard error $s\left(\widehat{\beta}\right)$ for a real-valued estimator $\widehat{\beta}$ is an estimate of the standard deviation of the distribution of $\widehat{\beta}$
- if β is a vector with estimate $\widehat{\beta}$ and covariance matrix estimate $\widehat{V}_{\widehat{\beta}}$
  - standard error for $\widehat{\beta}_j$ is square-root of diagonal element $[j, j]$

$$s\left(\widehat{\beta}_j\right) = \sqrt{\widehat{V}_{\widehat{\beta}_j}} = \sqrt{\left[\widehat{V}_{\widehat{\beta}}\right]_{jj}}$$

Measures of Fit

- classic

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2} = \frac{\widehat{\sigma}^2}{\widehat{\sigma}_y^2}$$

  - estimates $\rho^2 = Var\left(x_i^{\mathrm{T}}\beta\right) / Var\left(y_i\right) = 1 - \sigma^2 / \sigma_y^2$

- $\widehat{\sigma}^2$ and $\widehat{\sigma}_y^2$ are biased estimators, Theil (1961) used unbiased estimators s^2 and $\widetilde{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$

$$\overline{R}^2 = 1 - \frac{s^2}{\widetilde{\sigma}_y^2} = 1 - \frac{(n-1)\sum_{i=1}^{n} \widehat{u}_i^2}{(n-k)\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

- improved measure of fit is based on prediction errors

$$\widetilde{R}^2 = 1 - \frac{\sum_{i=1}^{n} \widetilde{u}_i^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2} = 1 - \frac{\widetilde{\sigma}^2}{\widehat{\sigma}_y^2}$$

    - fully corrects problem that R^2 necessarily increases when regressors are added
        - $\overline{R}^2$ only partially corrects this
        - $\widetilde{R}^2$ can be negative, if an intercept only model is a better predictor

- $\widetilde{\sigma}^2$ is the MSPE from leave-one-out cross validation - modern version of model selection
    - report $\widetilde{R}^2$

Multicollinearity

- *strict* multicollinearity: $X^{\mathrm{T}}X$ is singular
  - columns of X are linearly dependent
    - there exists some $\alpha \neq 0$ such that $Xa = 0$
  - $\left(X^{\mathrm{T}}X\right)^{-1}$ and $\widehat{\beta}$ are not defined
  - arises only through mistakes, include hourly and weekly wages, everyone works 40 hours each week
- more relevant, *near* multicollinearity
  - columns of X are nearly linearly dependent
  - not clear what it means to be near
- affects precision of estimation
- if $\frac{1}{n}X^{\mathrm{T}}X = \frac{1}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

$$Var\left(\widehat{\beta}|X\right) = \frac{\sigma^2}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n\left(1-\rho^2\right)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

  - as $\rho \rightarrow 1$ variance grows

Normal Regression Model

- Assume $u_i | x_i \sim \mathcal{N}\left(0, \sigma^2\right)$ implies
$$u | X \sim \mathcal{N}\left(0, I_n \sigma^2\right)$$
  - u is independent of X and normally distributed
- because linear functions of normal random variables are normal
$$\begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{u} \end{pmatrix} = \begin{pmatrix} \left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}} \\ M \end{pmatrix} u \sim \mathcal{N}\left(0, \begin{pmatrix} \left(X^{\mathrm{T}} X\right)^{-1} \sigma^2 & 0 \\ 0 & M\sigma^2 \end{pmatrix}\right)$$
  - because uncorrelated jointly normals are independent, $\widehat{\beta}$ is independent of any function of $\widehat{u}$
    - ★ in particular, $\widehat{\beta}$ is independent of s^2, $\widehat{\sigma}^2$, prediction errors $\widetilde{u}$

- spectral decomposition of M yields $M = H \begin{pmatrix} I_{n-k} & 0 \\ 0 & 0 \end{pmatrix} H^{\mathrm{T}}$ where $H^{\mathrm{T}} H = I_n$
- let $v = \sigma^{-1} H^{\mathrm{T}} u \sim \mathcal{N}\left(0, H^{\mathrm{T}} H\right) \sim \mathcal{N}\left(0, I_n\right)$
- $\frac{n\widehat{\sigma}^2}{\sigma^2} = \frac{(n-k)s^2}{\sigma^2}$
- $= \frac{1}{\sigma^2} \widehat{u}^{\mathrm{T}} \widehat{u}$
- $= \frac{1}{\sigma^2} u^{\mathrm{T}} M u$
- $= \frac{1}{\sigma^2} u^{\mathrm{T}} H \begin{pmatrix} I_{n-k} & 0 \\ 0 & 0 \end{pmatrix} H^{\mathrm{T}} u$
- $= v^{\mathrm{T}} \begin{pmatrix} I_{n-k} & 0 \\ 0 & 0 \end{pmatrix} v$
- $\sim \chi^2_{(n-k)}$

Test Statistic

if standard errors are calculated using homoskedastic formula

$$
\begin{aligned}
\frac{\widehat{\beta}_j - \beta}{s\left(\widehat{\beta}_j\right)} &= \frac{\widehat{\beta}_j - \beta}{s\sqrt{\left[\left(X^{\mathrm{T}}X\right)^{-1}\right]_{jj}}} \sim \frac{\mathcal{N}\left(0, \sigma^2\left[\left(X^{\mathrm{T}}X\right)^{-1}\right]_{jj}\right)}{\sqrt{\frac{\sigma^2}{(n-k)}\chi^2_{(n-k)}}\sqrt{\left[\left(X^{\mathrm{T}}X\right)^{-1}\right]_{jj}}} \\
&= \frac{\mathcal{N}(0,1)}{\sqrt{\frac{1}{(n-k)}\chi^2_{(n-k)}}} \sim t_{n-k}
\end{aligned}
$$

Finite Sample Distribution

Theorem (Finite Sample Distribution).
In the linear regression model of Assumption 1, if u_i is independent of x_i and distributed $\mathcal{N}\left(0, \sigma^2\right)$ then

- $\widehat{\beta} - \beta \sim \mathcal{N}\left(0, \sigma^2 \left(X^{\mathrm{T}}X\right)^{-1}\right)$
- $\frac{n\widehat{\sigma}^2}{\sigma^2} = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{(n-k)}$
- $\frac{\widehat{\beta}_j - \beta}{s\left(\widehat{\beta}_j\right)} \sim t_{n-k}$

Derivation of Mean-Square Forecast Error

- $\widetilde{u}_{n+1} = u_{n+1} - x_{n+1}^{\mathrm{T}} \left(\widehat{\beta} - \beta \right)$

$$MSFE_n = \mathbb{E}\left(u_{n+1}^2\right) + \mathbb{E}\left(x_{n+1}^{\mathrm{T}} \left(\widehat{\beta} - \beta \right) \left(\widehat{\beta} - \beta \right)^{\mathrm{T}} x_{n+1} \right)$$

  - $2\mathbb{E}\left(u_{n+1} x_{n+1}^{\mathrm{T}} \left(\widehat{\beta} - \beta \right) \right) = 0$
    - $u_{n+1} x_{n+1}^{\mathrm{T}}$ independent of $\left(\widehat{\beta} - \beta \right)$ and both are mean zero

- third term equals $\mathbb{E}\left(tr \left(x_{n+1}^{\mathrm{T}} \left(\widehat{\beta} - \beta \right) \left(\widehat{\beta} - \beta \right)^{\mathrm{T}} x_{n+1} \right) \right)$
- $= \mathbb{E}\left(tr \left(x_{n+1} x_{n+1}^{\mathrm{T}} \left(\widehat{\beta} - \beta \right) \left(\widehat{\beta} - \beta \right)^{\mathrm{T}} \right) \right)$
- $= tr \left(\mathbb{E}\left(x_{n+1} x_{n+1}^{\mathrm{T}} \right) \mathbb{E}\left(V_{\widehat{\beta}} \right) \right)$ because x_{n+1} is independent of $\widehat{\beta}$
- $= \mathbb{E}\left(tr \left(x_{n+1} x_{n+1}^{\mathrm{T}} V_{\widehat{\beta}} \right) \right) = \mathbb{E}\left(x_{n+1}^{\mathrm{T}} V_{\widehat{\beta}} x_{n+1} \right)$

Unbiased Estimator of MSFE

- $\mathbb{E}\left(\widetilde{u}_i^2\right) = \mathbb{E}\left(u_i - x_i^{\mathrm{T}}\left(\widehat{\beta}_{(-i)} - \beta\right)\right)^2$ averaging over i as well as u and x
- $= \sigma^2 + \mathbb{E}\left(x_i^{\mathrm{T}}\left(\widehat{\beta}_{(-i)} - \beta\right)\left(\widehat{\beta}_{(-i)} - \beta\right)^{\mathrm{T}} x_i\right)$
- $= \sigma^2 + \mathbb{E}\left(x_i^{\mathrm{T}} V_{\widehat{\beta}_{(-i)}} x_i\right)$
- $\mathbb{E}\left(\widetilde{\sigma}^2\right) = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(x_i^{\mathrm{T}} V_{\widehat{\beta}_{(-i)}} x_i\right)$
- $= MSFE_{n-1}$

Return to MSFE Theorem

Spectral Decomposition

- let A be an $n \times n$ square matrix
- let Λ be a diagonal matrix with eigenvalues of A
- let $H = [h_1 \cdots h_k]$ contain the eigenvectors of A
- if A is symmetric, then $A = H\Lambda H^{\mathrm{T}}$ - called the spectral decomposition of A

Return to Properties