Diagnostics and Model Selection

Junhui Qian

November 16, 2018

Outline

- Diagnostics
  - Introduction
  - Nonlinearity
  - Correlation
  - Heteroscedasticity
- Model Selection
- Structural Change

The Objectives

- After estimating a model, we should always perform diagnostics on the model. In particular, we should check whether the assumptions we made are valid.
- For OLS estimation, we should usually check:
  - Is the relationship between x and y linear?
  - Are the residuals serially uncorrelated?
  - Are the residuals uncorrelated with explanatory variables? (endogeneity)
  - Does homoscedasticity hold?

Estimation of Residuals

- Residuals are unobservable. But they can be estimated:

$$\hat{u}_i = y_i - x_i' \hat{\beta}.$$

- Using matrix language,

$$\hat{u} = (I - P_X) Y.$$

- If $\hat{\beta}$ is close to β, then $\hat{u}_i$ is close to u_i.

Outline

- Diagnostics
  - Introduction
  - **Nonlinearity**
  - Correlation
  - Heteroscedasticity
- Model Selection
- Structural Change

Residual Plots

We can plot

- Residuals
- Residuals versus Fitted Value
- Residuals versus Explanatory Variables

Any pattern in residual plots suggests nonlinearity or endogeneity.

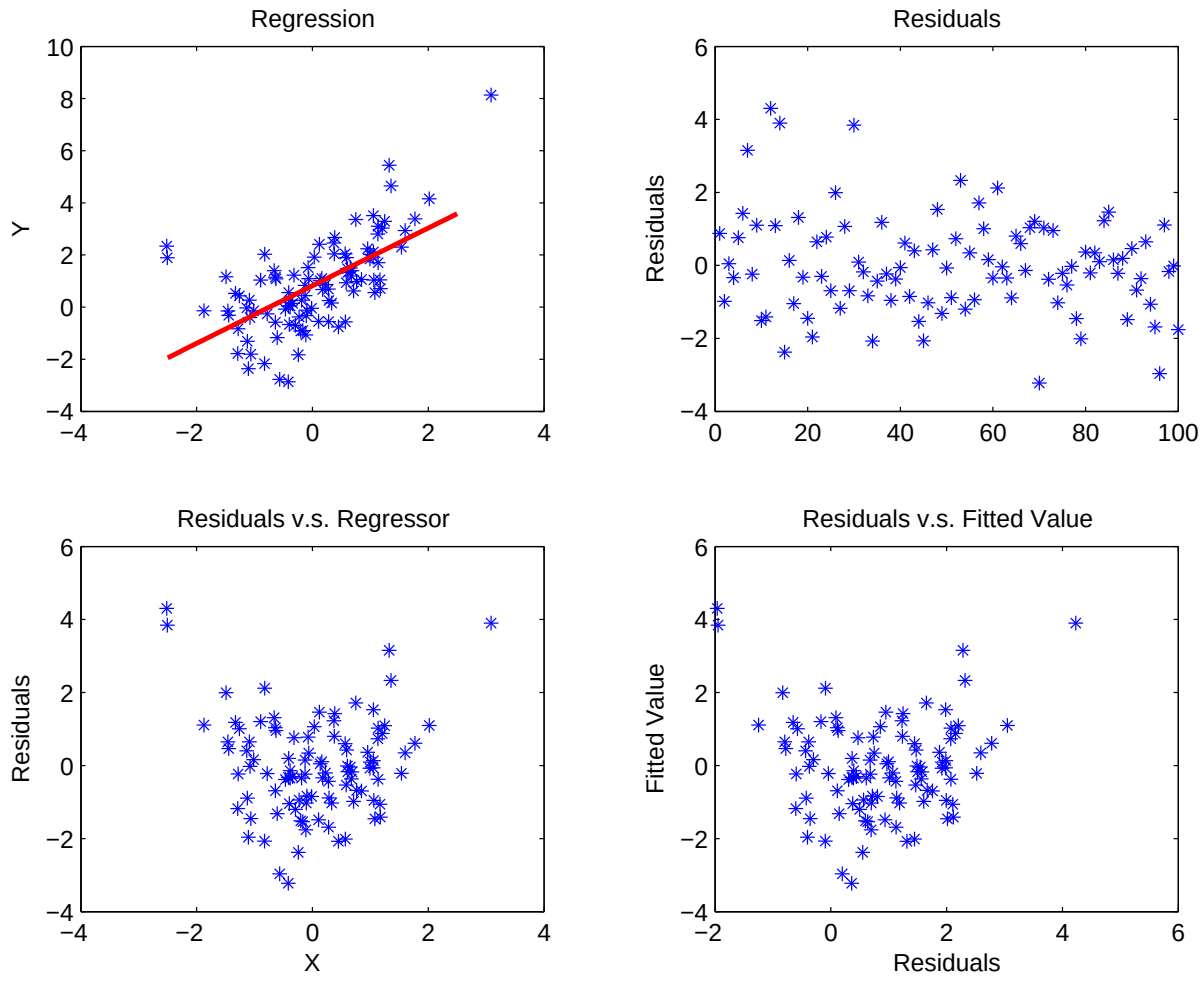


Figure: Residual Plots. DGP: $y = 0.2 + x + 0.5x^2 + u$

Partial Residual Plots

- To see whether there exists nonlinearity in a regressor, say the j-th explanatory variable x_j, We can plot

$$\hat{u} + \hat{\beta}_j x_j \quad \text{versus} \quad x_j,$$

  where $\hat{u}$ is residual from the full model.
- Partial residual plots may help us find the true (nonlinear) functional form of x_j.

Partial Residual Plots: Example

Suppose the true model is

$$y = \beta_0 + \beta_1 x + \beta_2 z + g(z) + u,$$

where $g(z)$ is a nonlinear function. We mistakenly estimate:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z + \hat{u}.$$

If we plot $\hat{\beta}_2 z + \hat{u}$ versus z, we may probably be able to detect nonlinearity in $g(z)$.

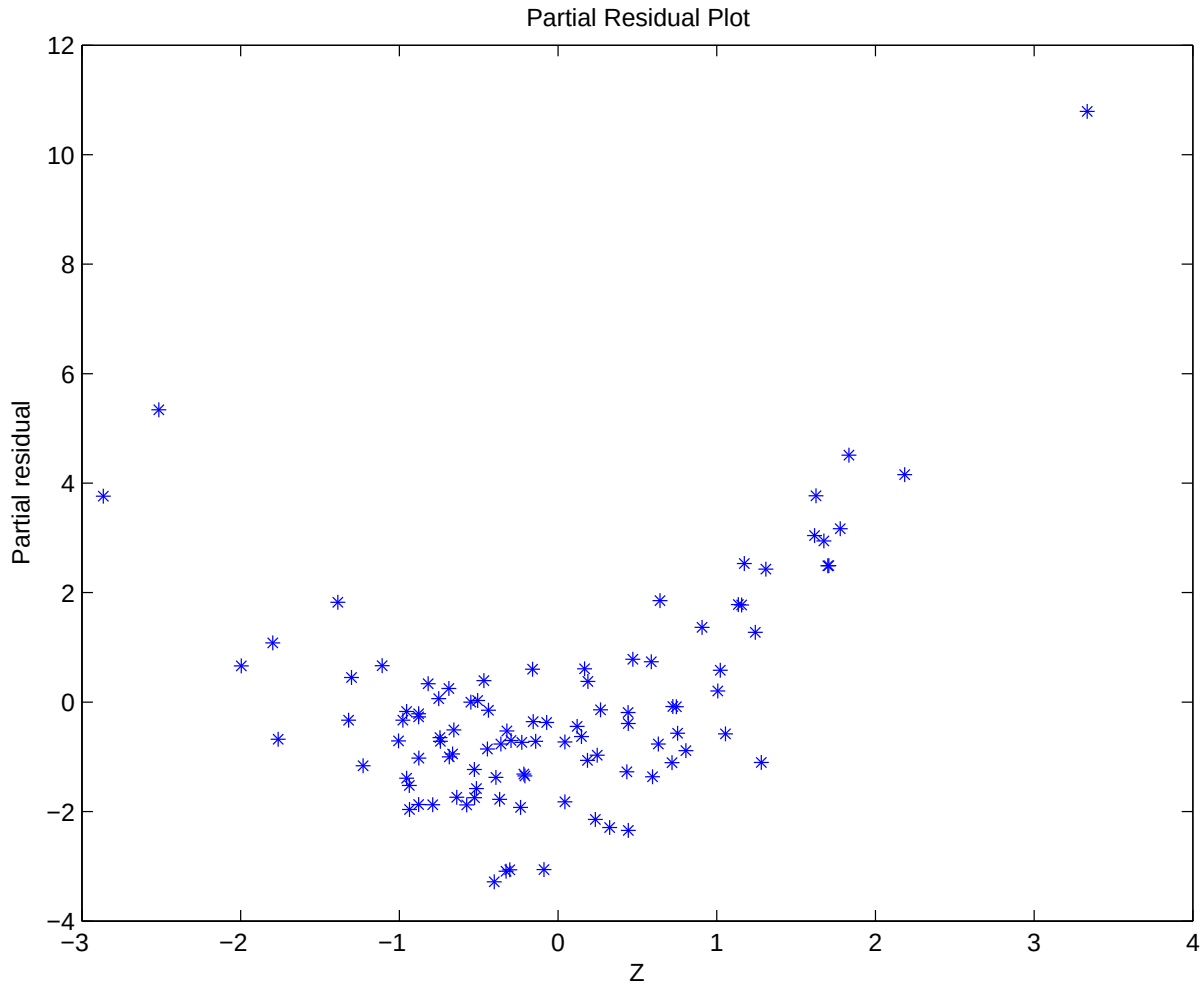


Figure: Residual Plots. DGP: $y = 0.2 + x + 0.5z + z^2 + u$

Outline

- Diagnostics
  - Introduction
  - Nonlinearity
  - **Correlation**
  - Heteroscedasticity
- Model Selection
- Structural Change

The iid Assumption

- The CLR assumption dictates that residuals should be iid.
- It is generally difficult to determine whether a given number of observations are from the same distribution.
- If there is a natural order of the observations (e.g., time), then we may check whether the residuals are correlated.
- If there is correlation, then the iid assumption is violated.

Residuals with Time

- When we deal with time series regression, for example,

$$\pi_t = \beta_0 + \beta_0 m_t + u_t,$$

  where π_t is the inflation rate and m_t is the growth rate of money supply, both indexed by time t.
- Now the "natural order" is time, and a time series plot of the estimated residual contains information.

Residual Plots

We can plot:

- Residuals over time
- Residuals v.s. previous residual
- Correlogram

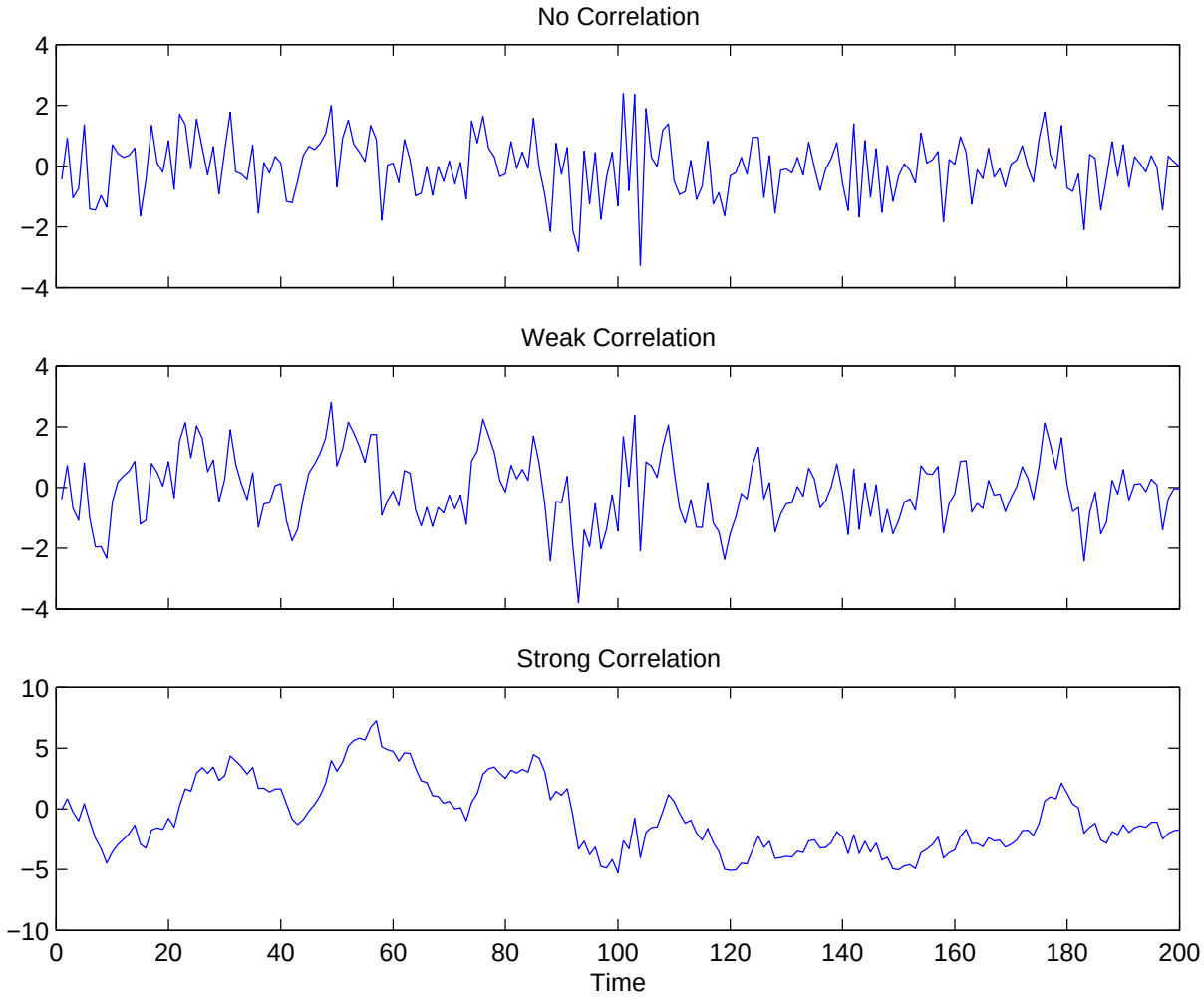


Figure: Residuals over time: $u_t = \alpha u_{t-1} + \varepsilon_t$, $\alpha = 0, 0.5, 0.95$, from top to bottom.

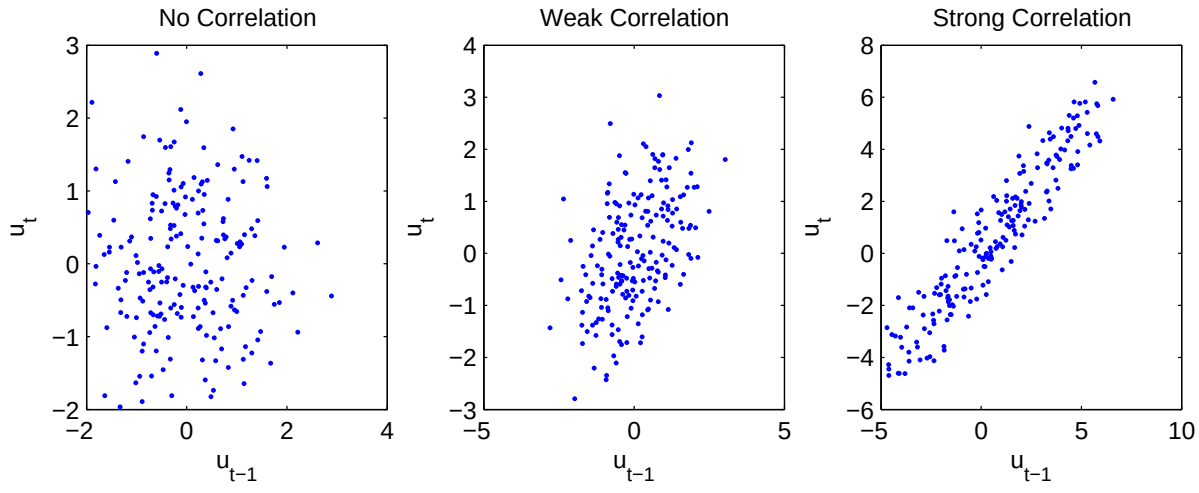


Figure: Residuals v.s. previous residual: $u_t = \alpha u_{t-1} + \varepsilon_t$, $\alpha = 0, 0.5, 0.95$, from left to right.

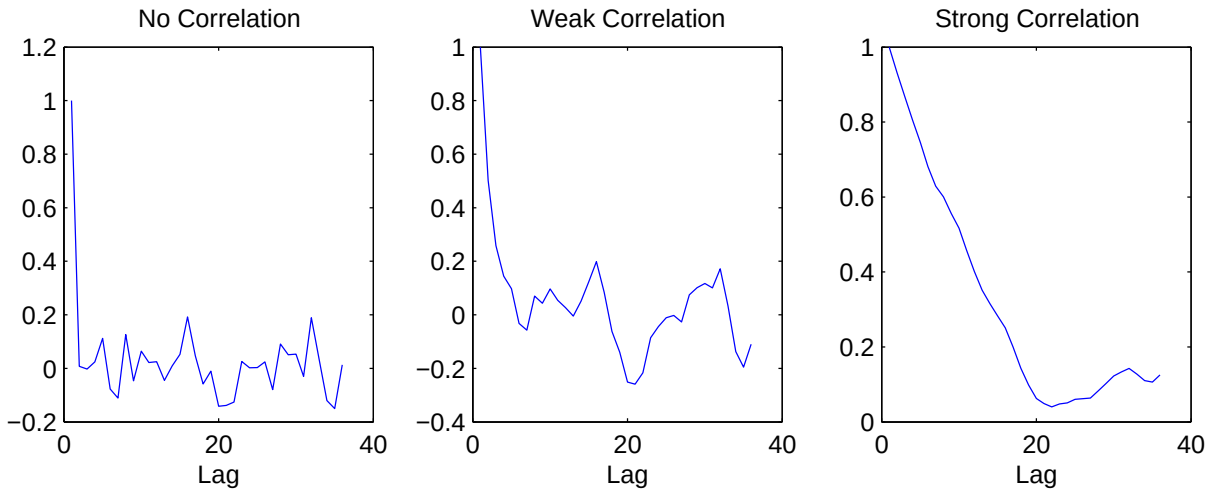


Figure: Correlograms: $u_t = \alpha u_{t-1} + \varepsilon_t$, $\alpha = 0, 0.5, 0.95$, from left to right.

Durbin-Watson Test

- Durbin-Watson is the formal test for independence, or more precisely, non-correlation.
- It assumes a AR(1) model for u_t, $u_t = \alpha u_{t-1} + \varepsilon_t$.
- The null hypothesis is: $H_0 : \rho = \alpha = 0$.
- The test statistic is

$$DW = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_{t-1}^2}.$$

Durbin-Watson Test

- $DW \in [0, 4]$.
- $DW = 2$ indicates no autocorrelation.
- If DW is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if DW is less than 1.0, there may be cause for alarm.
- Small values of DW indicate successive error terms are, on average, close in value to one another, or positively correlated.
- Large values of DW indicate successive error terms are, on average, much different in value to one another, or negatively correlated.

Fixing Correlation

- It's most likely that the model is misspecified.
- The usual practices are:
  - Add more explanatory variables
  - Add more lags of the existing explanatory variables

Outline

- Diagnostics
  - Introduction
  - Nonlinearity
  - Correlation
  - **Heteroscedasticity**
- Model Selection
- Structural Change

Checking Heteroscedasticity

- If $\mathrm{var}(u_i|x) = \sigma^2$, we call the model "homoscedastic". If not, we call it "heteroscedastic".
- If homoscedasticity does not hold, but CLR Assumptions 1-4 still hold, the OLS estimator is still unbiased and consistent. However, OLS is no longer BLUE.
- We can detect heteroscedasticity by looking at the residuals v.s. regressors.
- For simple regressions, we can look at regression lines.
- And we can formally test for homoscedasticity.
  - White test
  - Breusch-Pagan test

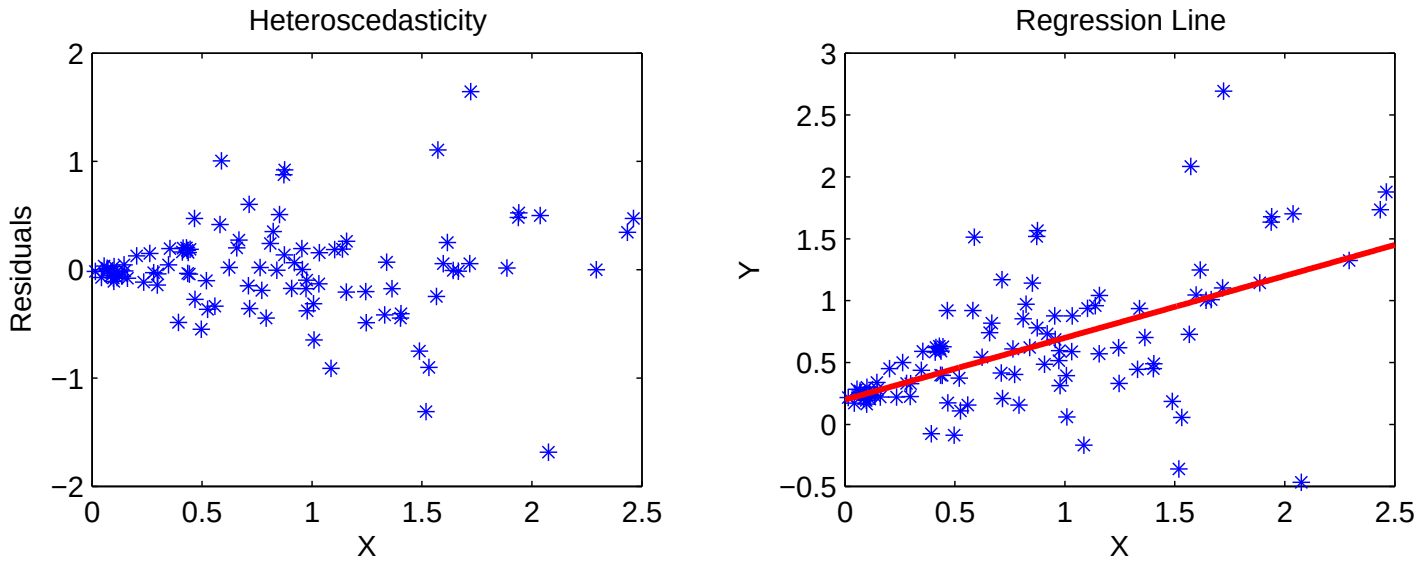


Figure: Heteroscedasticity. DGP: $y_i = \beta_0 + 0.5x_i + x_i\varepsilon_i$.

Fixing Heteroscedasticity

- Use a different specification for the model (different variables, or perhaps non-linear transformations of the variables).
- Use GLS (Generalized Least Square).

Outline

- Diagnostics
- **Model Selection**
  - Introduction
  - Hypothesis testing
  - Information criterion
  - Cross-validation
  - Lasso
- Structural Change

Model Selection

- Models simplify. As a result, there is no such thing as a "true model" except in simulation studies.
- The problem of selecting the best model among a set of models is called "model selection". But how do we define "best"?
- In theoretical studies, a good model is a set of assumptions (open unrealistic) that isolate crucial features for a particular problem and yield predictions that are testable.
- In empirical studies, a good model is one that yields good performance in out-of-sample predictions.
  - Predictions on individuals who are not in the sample.
  - Forecast the future value of a time series.

Nested and Non-nested Models

- Two models are "nested" when one can be obtained by imposing restrictions on the other. For example, the following model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

nests the following simple regression,

$$y = \beta_0 + \beta_1 x_1 + u.$$

- Otherwise, we call these two models "non-nested".
- Model selection requires different treatment in the context of nested or non-nested models.

Tradeoffs in Model Selection

- Model selection is challenging since we generally do not know whether improvement of fitness reflects genuine improvement of modeling, that is, producing better prediction/forecast.
  - Estimation error in complex models v.s. mis-specification risk in parsimonious models.
  - Goodness of fit within sample v.s. accuracy of out-of-sample prediction/forecast.

Outline

- Diagnostics
- Model Selection
  - Introduction
  - **Hypothesis testing**
  - Information criterion
  - Cross-validation
  - Lasso
- Structural Change

Hypothesis Testing

- If we deal with nested models, we may use hypothesis testing to select model, especially the likelihood ratio (LR) test, which can be applied to models more general than the linear regression.
- If the dimension of the full model is large, we have to rely on some ad hoc procedure.
  - Backward elimination.
  - Forward selection.
- For non-nested models, the classical LR test does not work. However, we can use modified one as in Vuong (1989).

Likelihood Ratio Test for Model Selection

Suppose we have a full model $f_1(\theta_1)$ and a restricted model $f_2(\theta_2)$, where f_1 and f_2 are likelihood functions, and $\theta_2 \subset \theta_1$. Then we construct LR statistic as follows,

$$LR = 2\left(\log f_1\left(\hat{\theta}_1\right) - \log f_2\left(\hat{\theta}_2\right)\right) \rightarrow_d \chi^2_j,$$

where j is the difference in the dimensions of θ_1 and θ_2.

- If LR is bigger than a critical value (or the p-value of LR is small), then it says that the full model is superior.
- If LR is smaller than a critical value (or the p-value of LR is big), then it says that there is insufficient evidence for the full model to be superior, hence the simpler model should be selected.

Likelihood Ratio Test for Model Selection

- In linear regression, the LR test is reduced to F test (or t test if $j = 1$).
- Note that a higher likelihood does not imply a better model, otherwise the full model would always be the best one.
- A more complex model must have sufficiently higher likelihood to be justified.
- And this requirement of "higher likelihood" is proportional to j, since if $X \sim \chi^2_j$, then $\mathbb{E}X = j$. This is how LRT achieves the penalty against complexity.

Backward and Forward Elimination

- Backward elimination starts from the full model.
  - At each step, eliminate the variable with the least effect on the model, using Student t or F statistic.
  - The process stops when some threshold is reached (e.g., a threshold on t or F statistic, minimum number of regressors).
- Forward elimination starts from the null model ($y = \beta_0 + u$).
  - At each step, add the variable with the largest effect on the model.
  - The process stops when some threshold is reached (e.g., a threshold on t or F statistic, minimum number of regressors).
- The backward and forward elimination procedures may not converge to a unique model. Neither guarantees convergence to the "true model".

Outline

- Diagnostics
- Model Selection
  - Introduction
  - Hypothesis testing
  - **Information criterion**
  - Cross-validation
  - Lasso
- Structural Change

Akaike Information Criterion

- The Akaike information criterion (AIC) is named after the statistician Hirotugu Akaike, its inventor.
- AIC is founded on information theory. When a model is constructed to characterize the data generating process (DGP), the model must always simplify. Hence some information will be lost by using the model to represent the reality. AIC is an estimator of the relative information lost by a given model.
- Let $f(\hat{\theta})$ be the maximum likelihood of a given model and let k be the number of parameters and n be the sample size. Then AIC is defined by

$$AIC = \frac{2}{n}k - \frac{2}{n}\log f(\hat{\theta}).$$

- One would select the model with smaller AIC (thus smaller information loss). Obviously, AIC achieves a certain balance between model fit and complexity.

AIC for Linear Regression

Consider a predictive linear regression,

$$y_{t+1} = x_t'\beta + u_{t+1},$$

where $\beta \in \mathbb{R}^k$ and $u_t \sim i.i.d.\ N(0, \sigma^2)$. Then

$$\begin{aligned} \log f(\hat{\theta}) &= -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{t=0}^{n-1}\hat{u}_{t+1}^2 \\ &= -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}. \end{aligned}$$

Hence

$$AIC = \log(\hat{\sigma}^2) + \frac{2}{n}k + C,$$

where C is a constant that does not change with k and hence can be omitted.

Bayesian Information Criterion

- The Bayesian information criterion (BIC) is also called Schwarz criterion (SIC or SBIC), developed by Gideon E. Schwarz using a Bayesian argument.
- BIC selects the model with the highest posterior probability given the data. It is, however, independent of the prior.
- Let $f(\hat{\theta})$ be the maximum likelihood of a given model and let k be the number of parameters and n be the sample size. Then BIC is defined by

$$BIC = \frac{\log n}{n} k - \frac{2}{n} \log f(\hat{\theta}).$$

- One would select the model with smaller BIC. Obviously, BIC penalizes complexity more severely than AIC.
- For a predictive linear regression, we have

$$BIC = \log(\hat{\sigma}^2) + \frac{\log n}{n} k.$$

The Relationship between Information Criteria and the LR Test

- For nested models, both IC and LRT may work. But they do not necessarily lead to the same choice.
- However, model selections based on IC and LRT have different meanings.
  - We conclude from IC that one model is better than the other.
  - While in LRT, we conclude either that the complex model is better (rejection), or that there is no sufficient evidence to differentiate between the complex and the simple.
- IC works for non-nested models, too. But LRT requires a modified version to work for non-nested models.

Outline

- Diagnostics
- Model Selection
  - Introduction
  - Hypothesis testing
  - Information criterion
  - **Cross-validation**
  - Lasso
- Structural Change

Cross-validation for Independent Sample

- Leave-p-out cross-validation.
  - Given a model specification, estimate the model using $n - p$ observations;
  - Predict the remaining p observations using the estimated model;
  - Calculate the prediction error;
  - Repeat the process for C_p^n times, calculate the MSE (mean squared error);
  - Select the model specification with the smallest MSE.
- Special cases:
  - Leave-one-out cv
  - k-fold cv

Cross-validation for Time Series

- Leave-p-out cross-validation, where $p = 2v + 1$.
  - Given a model specification, estimate the model using all observations apart from those in $[t - v, t + v]$;
  - Predict the remaining p observations using the estimated model, and calculate the prediction error;
  - Repeat the process for $n - p$ times, calculate the MSE (mean squared error).
  - Select the model specification with the smallest MSE. That is, select the model which minimizes the following

$$CV = \frac{1}{n - 2v - 1} \sum_{t=v+1}^{n-v} \left[\frac{1}{2v + 1} \sum_{s=t-v}^{t+v} \left(y_{s+1} - x_s' \hat{\beta}_{\{t-v:t+v\}} \right)^2 \right],$$

where $\hat{\beta}_{\{t-v:t+v\}}$ represents the estimated model using sample apart from those in $[t - v, t + v]$.

Cross-validation

- Cross-validation naturally avoids the problem of overfit in model selection.
- Asymptotically, minimizing the CV value is equivalent to minimizing the AIC. This is true for any model (Stone 1977), not just linear models.
- CV can be used to select variables in linear models, or tuning parameters in nonparametric models.
- CV does not always work.
  - It requires sampling from the same joint distribution. Or in time series setting, it requires joint stationarity.
  - If there are identical observations, then leave-one-out cv does not work. Leave-p-out cv can solve this problem, at heavier computation costs.
  - The outcome of cv is often sensitive to small variations in data. Again, leave-p-out can be more robust.
- Be conservative with statistical tests after model selection using cross-validation.

Outline

- Diagnostics
- Model Selection
  - Introduction
  - Hypothesis testing
  - Information criterion
  - Cross-validation
  - **Lasso**
- Structural Change

Lasso

For the purpose of selecting a small number of relevant regressors among a possibly large number of variables, Lasso (Least absolute shrinkage and selection operator) solves the following problem,

$$\min_{\{\beta_j\}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where λ is the penalty on the absolute values of β_j.

- If $\lambda = 0$, Lasso reduces to the usual OLS. If $\lambda = \infty$, Lasso forces all parameters to *exactly* zero.
- With an appropriate value of λ, Lasso can force the parameters on the irrelevant regressors to exactly zero, while keeping the relevant ones.
- Lasso has become very popular for the estimation of high-dimensional models with some built-in sparseness.

Lasso

To see why Lasso can force parameters to exactly zero, we consider the case where $p = 2$,

$$\min_{\{\beta_1,\beta_2\}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 + \lambda \left(|\beta_1| + |\beta_2| \right).$$

This is equivalent to

$$\min_{\{\beta_1,\beta_2\}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \quad \text{subject to} \quad |\beta_1| + |\beta_2| \leq s.$$

Graphically, $\sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 = z$ are ellipses in the (β_1, β_2) plane, while $(|\beta_1| + |\beta_2|) \leq s$ is a diamond around 0. Hence corner solutions can be easily produced.

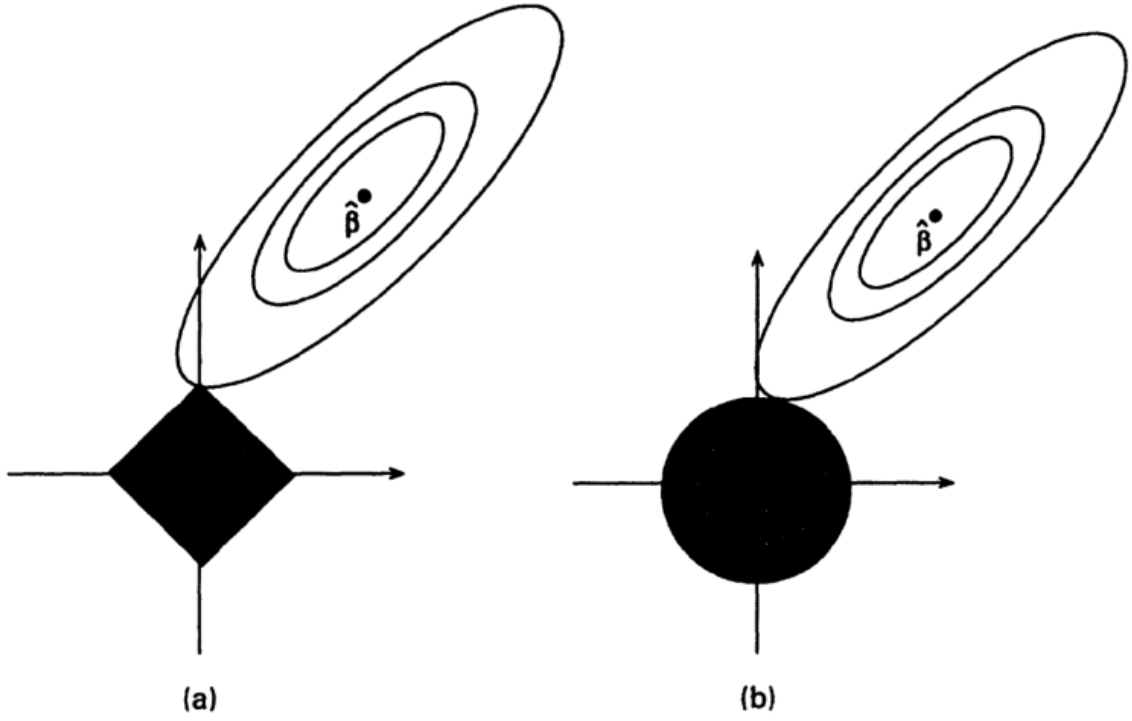


Figure: (a) Lasso; (b) Ridge regression. Graph is from Tibshirani (1996)

Ridge Regression

- While Lasso penalizes the absolute values of the parameters, the ridge regression penalizes the squared values of the parameters.
- When $p = 2$, the ridge regression solves

$$\min_{\{\beta_1,\beta_2\}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 + \lambda \left(\beta_1^2 + \beta_2^2\right).$$

This is equivalent to

$$\min_{\{\beta_1,\beta_2\}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \quad \text{subject to} \quad \beta_1^2 + \beta_2^2 \leq s.$$

Graphically, $\frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 = z$ are ellipses in the (β_1, β_2) plane, while $\beta_1^2 + \beta_2^2 \leq s$ is a circle around 0.

Elastic net

The elastic net combines Lasso and the ridge regression. Let $\|\beta\|_1 = |\beta_1| + \cdots + |\beta_p|$ and $\|\beta\|_2^2 = \left(\beta_1^2 + \cdots + \beta_p^2\right)^2$. The elastic net solves

$$\min_{\{\beta\}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \lambda P_\alpha(\beta),$$

where

$$P_\alpha(\beta) = \frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1.$$

- If $\alpha \to 1$, then the elastic net reduces to Lasso.
- If $\alpha \to 0$, then it approaches the ridge regression.

The Application of Lasso and Elastic Net

- λ can be selected using cross-validation.
- After model selection using Lasso or elastic net, one can estimate the selected model using OLS. This step is called post-Lasso estimation.
- When highly correlated variables are present, Lasso can be unstable and elastic net may yield better out-of-sample forecasts.
- The Matlab command for Lasso and elastic net is "lasso".

Outline

- Diagnostics
- Model Selection
- **Structural Change**
  - Introduction
  - Test for a known break date
  - Test for an unknown break date
  - Test for multiple breaks
  - Test for an unknown number of breaks
  - The shrinkage approach

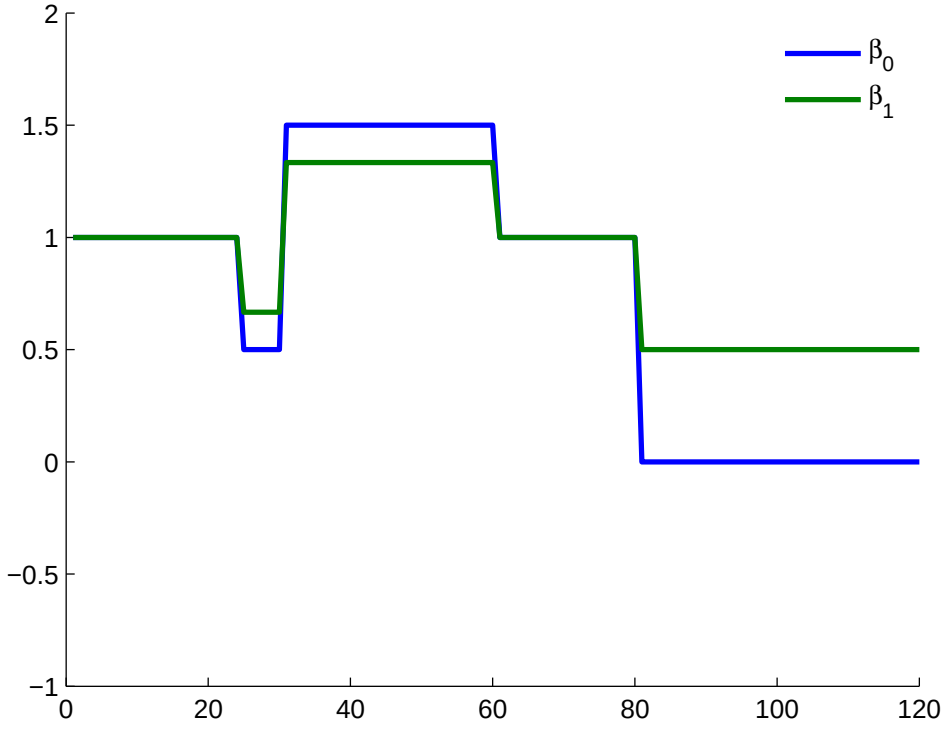
The time instability of regression analysis

- Kenyes: The (regression) coefficients arrived at are apparently assumed to be constant for 10 years or for a larger period. Yet, surely we know that they are not constant. There is no reason at all why they should not be different every year.
- Lucas: Given that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models.

Econometric models that allow time-varying coefficients

- Regression with piecewise constant coefficients (structural breaks)
- State-space models
- Random-coefficient models
- Functional-coefficient models

A limited form of time-varying-coefficient regression model



Dealing with piecewise-constant coefficients

- Regression with a known break date
- Regression with an unknown break date
- Regression with multiple breaks
- Testing for an unknown number of structural breaks (e.g., Bai and Perron (1998, 2003))
- Shrinkage-based estimation (e.g., Qian and Su (2015, 2014, 2016), Li, Qian, and Su (2016))

Regression with a Known Break Date

Consider the following linear regression with a possible structural break at k_0,

$$\begin{aligned} y_t &= x_t'\beta_1 + u_t, \quad t = 1, \ldots, k_0 - 1 \\ y_t &= x_t'\beta_2 + u_t, \quad t = k_0, \ldots, n, \end{aligned}$$

where x_t is $p \times 1$.

- If $\beta_1 \neq \beta_2$, then there is a common break in all coefficients.

Testing for a Break at a Known Date

Let $X_1 = (x_1, \ldots, x_{k_0-1})'$, $X_2 = (x_{k_0}, \ldots, x_n)'$. The model can be re-written as

$$Y = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + u.$$

Then the test for a known break at k_0 is:

$$\mathrm{H}_0 : \beta_1 = \beta_2, \quad \mathrm{H}_1 : \mathrm{Otherwise}.$$

- We may conduct the usual F test (the Chow test).
- Or equivalently, we may use the Wald statistic.

The Wald Test

The Wald test compares the estimate $\hat{\theta}$ of the unrestricted model to the hypothetical value θ_0, assuming that $\sqrt{n}\left(\hat{\theta}-\theta_0\right) \to_d N(0, \Sigma_{\theta_0})$.

- Univariate case: $W = \frac{\left(\hat{\theta}-\theta_0\right)^2}{\mathrm{var}\left(\hat{\theta}\right)} \to_d \chi^2_1$.
- Multivariate case: let $\theta \in \mathbb{R}^p$, consider the following test,

$$\mathrm{H}_0 : R\theta = r, \quad \mathrm{H}_1 : R\theta \neq r,$$

where R is $d \times p$ and r is $d \times 1$. Then the Wald test is given by

$$W = (R\hat{\theta} - r)' \left(R\left(\hat{\Sigma}_{\theta_0}/n\right)R'\right)^{-1} (R\hat{\theta} - r) \to_d \chi^2_d.$$

- In finite sample, under normality, $W/d \sim F_{d,n-p}$.

The Wald Test of a Break at $t = k_0$

Let

$$X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Then

$$\Sigma_{\hat{\beta}} = \sigma^2 \left(X'X\right)^{-1} = \sigma^2 \begin{bmatrix} \left(X_1'X_1\right)^{-1} & 0 \\ 0 & \left(X_2'X_2\right)^{-1} \end{bmatrix}.$$

Since $\hat{\beta}_1 - \hat{\beta}_2 = [I - I] \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$,

$$\Sigma_{\hat{\beta}_1 - \hat{\beta}_2} = [I - I]\Sigma_{\hat{\beta}}[I - I]' = \sigma^2 \left[\left(X_1'X_1\right)^{-1} + \left(X_2'X_2\right)^{-1}\right].$$

Then

$$W = \left(\hat{\beta}_1 - \hat{\beta}_2\right)' \left\{s^2 \left[\left(X_1'X_1\right)^{-1} + \left(X_2'X_2\right)^{-1}\right]\right\}^{-1} \left(\hat{\beta}_1 - \hat{\beta}_2\right) \rightarrow_d \chi^2_p,$$

where $s^2 = \hat{u}'\hat{u}/(n - 2p)$ is the estimate of σ^2.

Testing for a Break with an Unknown Date

Let $X_1 = (x_1, \ldots, x_{k-1})'$, $X_2 = (x_k, \ldots, x_n)'$, where k is unknown. Note that the Wald statistic is now a function of k,

$$W(k) = \left(\hat{\beta}_1 - \hat{\beta}_2\right)' \left\{s^2 \left[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}\right]\right\}^{-1} \left(\hat{\beta}_1 - \hat{\beta}_2\right),$$

where X_1, X_2, $\hat{\beta}_1$, $\hat{\beta}_2$, and s^2 are all dependent on k. Then a natural test statistic for testing $\beta_1 = \beta_2$ is

$$SupW = \sup_{\epsilon n \leq k \leq (1-\epsilon)n} W(k)/p,$$

where ϵ is a small number called "trimming parameter".

- We reject the null hypothesis (no break) when $SupW$ is sufficiently large.

Testing for a Break with an Unknown Date

- The distribution of *SupW*, however, is not standard. The break date can be estimated by

$$\hat{k} = \text{argmax}_{p \leq k \leq n-p} W(k)/p.$$

- The power of *SupW* (the probability of rejecting H_0 when H_1 is true) is not optimal.
- Andrews and Ploberger (1994) proposed a class of optimal tests. For example

$$\begin{aligned} \mathit{MeanW} &= \frac{1}{n} \sum_{k=\epsilon n}^{(1-\epsilon)n} W(k)/p \\ \mathit{ExpW} &= \log \left(\frac{1}{n} \sum_{k=\epsilon n}^{(1-\epsilon)n} \exp \left(\frac{1}{2} W(k)/p \right) \right). \end{aligned}$$

The Linear Regression with Multiple Breaks

Consider the linear regression with m structural breaks,

$$y_t = x_t'\beta_j + u_t, \quad t \in \{T_{j-1}, \ldots, T_j - 1\}, \quad j = 1, \ldots, m+1,$$

where we define $T_0 \equiv 0$, $T_{m+1} \equiv n+1$, and $\min_j\{T_j - T_{j-1}\} \geq p$.

- We assume that m is known, then there are $m+1$ regimes. In each regime, the coefficient β_j is constant and hence can be estimated by OLS.
- To estimate the breaks, we may solve the following problem,

$$\min_{T_1,\ldots,T_m} SSR(T_1, \ldots, T_m),$$

  where SSR is the sum of squared residuals associated with the partition $\mathcal{T} = \{T_1, \ldots, T_m\}$.
- The computation is intensive, but there is a fast algorithm based on dynamic programming (Bai & Perron, 2003).

Testing for a Known Number of Breaks

Consider the following hypothesis,

$$\mathrm{H}_0 : m = 0, \quad \mathrm{H}_1 : m = m^*.$$

- Let $X_1 = (x_1, \ldots, x_{T_1-1})'$, $X_2 = (x_{T_1}, \ldots, x_{T_2-1})'$,..., $X_{m^*+1} = (x_{T_{m^*}}, \ldots, x_n)'$, where $\mathcal{T} = \{T_1, \ldots, T_{m^*}\}$ is an unknown partition.
- Let $X = \mathrm{diag}(X_1, \ldots, X_{m^*+1})$ and $\beta = (\beta_1', \ldots, \beta_{m^*+1}')'$. Then the unrestricted model is

$$Y = X\beta + u.$$

- Now define $\theta_j = \beta_j - \beta_{j-1}$, $j = 2, \ldots, m^* + 1$, and let $\theta = (\theta_2', \ldots, \theta_{m^*+1}')'$. Then the above hypothesis is equivalent to

$$\mathrm{H}_0 : \theta = 0, \quad \mathrm{H}_1 : \mathrm{Otherwise}.$$

Testing for a Known Number of Breaks

There exists a matrix R such that $R\beta = \theta$. Hence the Wald statistic is

$$W(\mathcal{T}) = \left(R\hat{\beta}\right)' \left\{s^2 \left[R(X'X)^{-1}R'\right]\right\}^{-1} \left(R\hat{\beta}\right),$$

where $s^2 = \hat{u}'\hat{u}/(n - (m^* + 1)p)$, and $\hat{\beta}$, $\hat{u}$, and X are all dependent on the partition $\mathcal{T} = \{T_1, \ldots, T_{m^*}\}$.

- The SupW statistic can be constructed as usual,

$$SupW = \sup_{\mathcal{T}} W(\mathcal{T})/(m^* p) = W(\hat{\mathcal{T}})/(m^* p),$$

  where $\hat{\mathcal{T}}$ is the estimated partition obtained by minimizing the SSR of the unrestricted regression.
- We reject the null hypothesis ($m = 0$) if $SupW$ is sufficiently large.

Testing for an Unknown Number of Breaks

It is obviously better to state the alternative hypothesis as the existence of an unknown number of breaks, something like this:

$$\mathrm{H}_0 : m = 0, \quad \mathrm{H}_1 : m > 0.$$

The double maximum tests proposed in Bai and Perron (1998) is a natural choice. Let $\mathcal{T}_m$ be any partition when the number of breaks is m. One version of the double maximum test statistic is called "equally weighted double maximum test",

$$UDMax = \max_{m \leq M} SupW(m) = \max_{m \leq M} \sup_{\mathcal{T}_m} W(\mathcal{T}_m)/(mp).$$

- It is assumed that the number of breaks is bounded (with the upper bound M).
- If $UDMax$ is sufficiently large, then the null hypothesis of no breaks is rejected.
- When the null hypothesis is rejected, the number of breaks is unknown and needs to be estimated in most applications.

A Test of ℓ versus $\ell + 1$ Breaks

One way to determine the number of breaks is to test whether there is an additional break, given the fact we have established that there are ℓ breaks,

$$\mathrm{H}_0 : m = \ell, \quad \mathrm{H}_1 : m = \ell + 1.$$

- Under the null hypothesis, the restricted model has ℓ breaks, which we can estimate by minimizing the SSR,

$$\hat{\mathcal{T}}_\ell = \operatorname{argmin}_{T_1,\ldots,T_\ell} SSR(T_1, \ldots, T_\ell).$$

- The unrestricted model has $\ell + 1$ breaks. We insert additional break τ into each segment of $\hat{\mathcal{T}}_\ell$ and obtain $\hat{\mathcal{T}}_{\ell+1}$ that achieves the overall minimal value of SSR. We reject the null hypothesis if the SSR associated with $\hat{\mathcal{T}}_{\ell+1}$ is sufficiently lower than associated with $hat\mathcal{T}_\ell$. Specifically, the test statistic is given by

$$\begin{aligned} F(\ell + 1|\ell) &= s^{-2} \Big[SSR(\hat{T}_1, \ldots, \hat{T}_\ell) - \\ &\qquad \min_{1 \le j \le \ell+1} \inf_{\tau \in \Lambda_j} SSR(\hat{T}_1, \ldots, \hat{T}_{j-1}, \tau, \hat{T}_j, \ldots, \hat{T}_\ell) \Big], \end{aligned}$$

  where Λ_j is the j-th segment. The distribution of $F(\ell + 1|\ell)$ under H_0 is non-standard but available (Bai and Perron, 1998).

- We may repeatedly apply the test to determine the number of breaks.

The Shrinkage Approach to the Estimation of Breaks

Suppose that the regressor in our model is a scalar,

$$y_t = \beta_t x_t + u_t.$$

Since $\{(\beta_t - \beta_{t-1}), t = 2, \ldots, T\}$ are mostly zero (assume that there are a small number of breaks), we solve the following problem,

$$\min_{\{\beta_t\}} \sum_{t=1}^{T} (y_t - \beta_t x_t)^2 + \lambda \sum_{t=2}^{T} |\beta_t - \beta_{t-1}|.$$

- If $\lambda = 0$, the problem reduces to the least square estimation of a time-varying-coefficient regression without constraints. If $\lambda = \infty$, the approach forces all consecutive changes in coefficients to *exactly* zero, and the problem reduces to OLS.
- With an appropriate value of λ, the approach can force "false jumps" to exactly zero, while keeping the true ones.
- The above problem is a special case of the fused Lasso (Tibshirani et al. (2005)).

The Penalized Least Square Estimation

- More generally, we estimate $\{\beta_t\}$ by minimizing the following penalized least squares (PLS) objective function

$$\frac{1}{n}\sum_{t=1}^{n}\left(y_t - \beta_t' x_t\right)^2 + \lambda \sum_{t=2}^{n} \|\beta_t - \beta_{t-1}\| \tag{1}$$

  where λ is a positive tuning parameter and $\|\cdot\|$ denotes the matrix norm.
- If $\hat{\beta}_\tau \neq \hat{\beta}_{\tau-1}$, then a structural break (change) occurs at τ.
- The above penalized least square is a convex problem, which can be solved using a general-purpose convex solver (e.g., CVX). For more efficient computation, we use the block-coordinate descent algorithm.
- Choice of λ
  - Trial and error.
  - More systematically, we may choose λ by minimizing some information criterion. For example,

$$IC(\lambda) = \log(\hat{\sigma}_\lambda^2) + n^{-1/2} p\,(\hat{m}_\lambda + 1).$$

References

- Andrews, D. W. K., W. Ploberger, 1994, Optimal tests when a nuisance parameter is present only under the alternative. Econometrica, 62, 1383-1414.
- Bai, J., P. Perron, 1998, Estimating and testing liner models with multiple structural changes. Econometrica, 66, 47-78.
- Bai, J., P. Perron, 2003, Computation and analysis of multiple structural change models. Journal of Applied Econometrics, 18, 1-22.
- Qian, J., L. Su, 2016, Shrinkage estimation of regression models with multiple structural Changes. Econometric Theory, 32 (6), 1376-1433.