8. Structural Breaks

author: ECONOMIC FORECASTING date: SUMMER 2017 autosize: true

Binary Variables

Binary/Indicator/Dummy variables are a way to control for qualitative dfferences between observations

Examples: - Race - Gender - Quarters - Developing vs. Industrialized Countries - Marriage Status

Example

Consider the following model for wages of NBA players

$$wage_i = \alpha + \beta_1 ppg_i + \beta_2 center_i + \epsilon_i$$

where

$$center_i = \begin{cases} 1, \text{ if the player's position is center} \\ 0, \text{ otherwise} \end{cases}$$

So,

$$wage_i = \alpha + \beta_1 ppg_i + \epsilon_i \quad \text{for a non-center} \tag{1}$$

$$wage_i = \alpha + \beta_1 ppg_i + \beta_2 center_i + \epsilon_i \quad \text{for a center} \tag{2}$$

For Equation (2)

$$wage_i = 0.195 + 0.113 ppg_i + 0.518 center_i + \epsilon_i$$

If a center…

$$wage_i = 0.195 + 0.113 ppg_i + 0.518(1)$$
$$wage_i = 0.713 + 0.113 ppg_i$$

If a non-center…

$$wage_i = 0.195 + 0.113 ppg_i + 0.518(0)$$
$$wage_i = 0.195 + 0.113 ppg_i$$

What about this model

$$wage_i = \alpha + \beta_1 ppg_i + \beta_2 center_i + \beta_3 noncenter_i + \epsilon_i$$

Model will not run as specified: - **Model is not full rank** (violation of one of the CLR assumptions) - model must have a reference group

What is the reference group of the original model?

$$wage_i = \alpha + \beta_1 ppg_i + \beta_2 center_i + \epsilon_i$$

Typically, R will kick out one of the dummy variables - Do NOT rely on this - you should choose which variable to include

== So far, we've only considered the "standard" use of dummy variables

$$wage_i = \alpha + \beta_1 ppg_i + \beta_2 center_i + \epsilon_i$$

- The implicit assumption is that both centers and non-centers have the same return to *ppg*
- We can consider an **interaction** term to allow for different returns to *ppg*

$$wage_i = \alpha + \beta_1 ppg_i + \beta_3 (ppg_i \cdot center_i) + \epsilon_i$$

- β_1 is the return to *ppg* for non-centers, while $(\beta_1 + \beta 3)$ is the return to *ppg* for centers

== We can also combine the interaction term with the standard dummy variable to get:

$$wage_i = \alpha + \beta_1 ppg_i + \beta_2 center_i + \beta_3 (ppg_i \cdot center_i) + \epsilon_i$$

- This allows for a shift in both the intercept and slope.
- Intercept
  - non-centers: α
  - centers: $\alpha + \beta_2$
- Slope
  - non-centers: β_1
  - centers: $\beta_1 + \beta_3$

== To test for the significance of the interaction term, we use a Cramer test

- This test has an F-Distribution

$$wage_i = \alpha + \beta_1 ppg_i + \beta_2 center_i + \beta_3 (ppg_i \cdot center_i) + \epsilon_i$$

Hypotheses:

$$H_0 : \beta_1 = \beta_3$$
$$H_1 : \beta_1 \neq \beta_3$$

Or...

$$F_{r,n-k} = \frac{(SSR_r - SSR_{ur})/r}{SSR_{ur}/n - k}$$

Final Notes

Dummy variables can be extremely useful; however, as is typically the case they are not a panacea - Also, care needs taken when using dummy variables, especially with their definition

Example:

$$y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i \qquad \text{(your model)}$$

but

$$X_{2,i} = Z_{1,i} + Z_{2,i}$$

and thus

$$y_i = \beta_1 X_{1,i} + \beta_2 (Z_{1,i} + Z_{2,i}) + \epsilon_i \qquad \text{(your model)}$$

but

$$y_i = \beta_1 X_{1,i} + \beta_2 Z_{1,i} + \beta_3 Z_{2,i} + \epsilon_i \qquad \text{(true model)}$$

Structural Breaks

One of the assumptions of the CLR is that parameters are constant over time (i.e. time invariant)

$$y_t = \alpha_t + \beta_t y_{t-1} + \epsilon_t$$

- $\beta_t = \beta$ and $\alpha_t = \alpha$
- As you might suspect, this seems like a strong assumption fot time series data
- This is important for ARIMA as well as OLS

What Causes a Structural Break?

Imagine you have a model for US GDP growth from 1950 to 2010…

- Historical events could cause changes in the parameters of independent variables in this model. For example:
  - Vietnam
  - Oil price shocks of the 1970s
  - September 11th
  - Financial Crisis of 2008

What Causes a Structural Break?

Historical events are not necessary to cause parameter instability.

For example, what phenomenon might be able to cause parameter instability (and a structural break) is related to: - Crop yields? - Rainfall? - Number of ski days in Aspen?

Global Warming

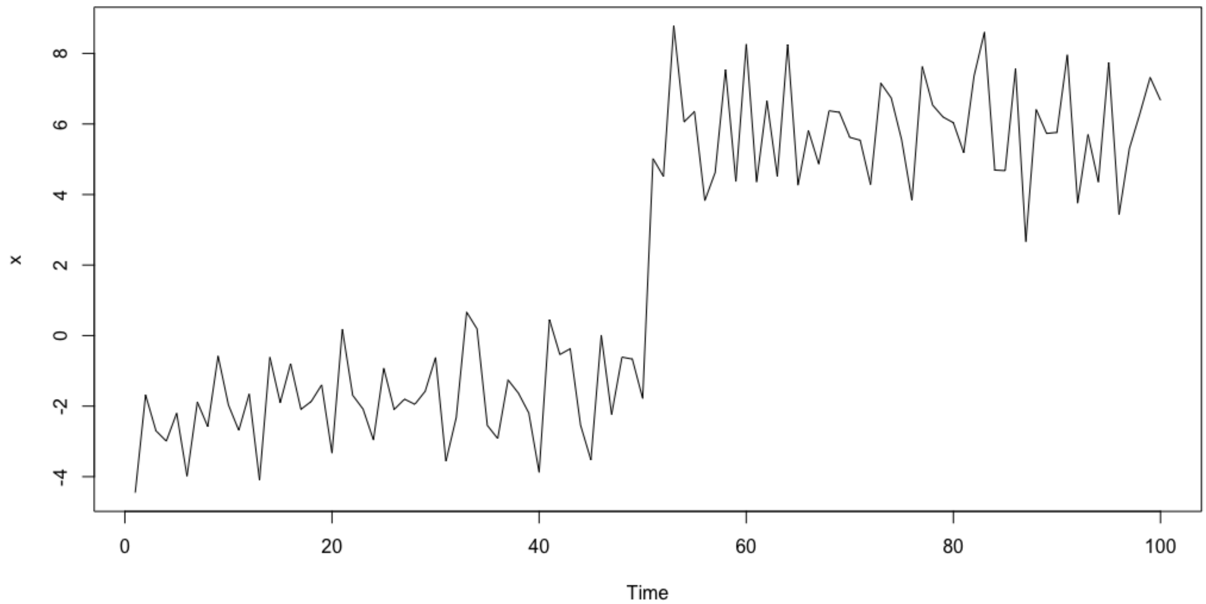

Figure 1:

Series with an obvious structural break

Testing for a Structural Break

Once we know where a structural break occurs, testing is is relatively easy. - Chow Test: - Fit the same model pre- and post-break data

$$t_{post} = T - t_{pre}$$

- If the models are NOT sufficiently different, then there is no structural break in the data generating process.

Chow Test

The Chow Test is an F-Test

$$F_{r,n-k} = \frac{(SSR - SSR_1 - SSR_2)/n}{(SSR_1 + SSR_2)/(T - 2n)}$$

- SSR, SSR_1 and SSR_2 are the residual sum of squares from the full model, the pre- and post-break models, respectively
- n is the number of estimated parameters (including the intercept)
- The larger the F value, the more restrictive is the full model
- NOTE: the number of observations in each break period should be "reasonable"

ALternative Chow Test

Another way to perform a Chow Test is to use a dummy variable:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \gamma D_t + \epsilon_t$$

where

$$D_t = \begin{cases} 1, & \text{if the time period is after the break} \\ 0, & \text{otherwise} \end{cases}$$

Test for a break using a t-test with:

$$\begin{aligned} H_0 &: \gamma = 0 \quad \text{(no break)} \\ H_1 &: \gamma \neq 0 \quad \text{(structural break)} \end{aligned}$$

Adding an Interaction

Perhaps the relationship between y_{t-1} and y_t changes after a break. If we don't account for this, our other parameter estimates will be biased (why?). An attempt to deal with this:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-1} D_t + \gamma D_t + \epsilon_t$$

Can use a Cramer test to test this hypothesis:

$$\begin{aligned} H_0 &: \alpha_1 = \alpha_2 \quad \text{(no difference before or after)} \\ H_1 &: \alpha_1 \neq \alpha_2 \quad \text{(there is a difference)} \end{aligned}$$

Edogenous Breaks

An **endogenous break** is a non-speci ed break point - A break exists, but it is not obvious where it occurs

To see if a break exists anywhere in your data, you could perform a Chow test for every potential break date - Date that results in larges F value (if significant) provides a consistent estimate of the actual break date, if any exists

Issues with this approach

We need an adequate number of observations in each sample - at least 10% of the overall sample

Since we're scanning for the largest F value, we actually cannot rely on the standard F distribution. - Andrews (1993, Econometrica) - Andrews and Ploberger (1994, Econometrica)

Parameter Instability

Sometimes it's not reasonable for a break to manifest itself at a single point in time

- **For example:** microprocessor, internet, etc. . .

Even if we knew when break occurred, the full effect happens only gradually

We can use a recursive technique to identify changes in parameters over time

$$y_t = \alpha + X_t\beta + \epsilon_t$$

Estimate the model at some point in time (say t = 10) and then re-estimate the model for every period (at t = 11, t = 12, etc. . .) and collect the parameter estimates

Continued. . .

Take estimates of the individual parameters and plot (individually) against time

Plots of these variables require a "burn in" period - since we need a "large" n for parameter estimate consistency - make sure to plot con dence intervals as well

IF the magnitude of the parameter estimates suddenly changes after the burn in period, you could have a structural break

CUSUM

Another way to identify structural breaks and parameter instability is using a CUSUM graph - At each step of a process it is possible to calculate a one-step ahead forecast:

$$\begin{aligned} e_t(1) &= y_{t+1} - \mathbf{E}_t y_{t+1} \\ e_t(10) &= y_{11} - \mathbf{E}_{10} y_{11} \\ e_t(100) &= y_{101} - \mathbf{E}_{10} y_{101} \end{aligned}$$

Then we can calculate the CUSUM:

$$CUSUM_N = \sum_{i=n}^{N} \frac{e_i(1)}{\sigma_e}$$

Continued. . .

If the model fits the data well, then forecasts should be unbiased

After calculating the CUSUM, plot it - If the CUSUM value lies outside your confidence interval, then this suggests a structural break

NOTE: CUSUM plots are bad at identifying structural breaks that occur near the end of the sample - due to the cumulative nature of the plot and how standard errors are calculated

Graphically

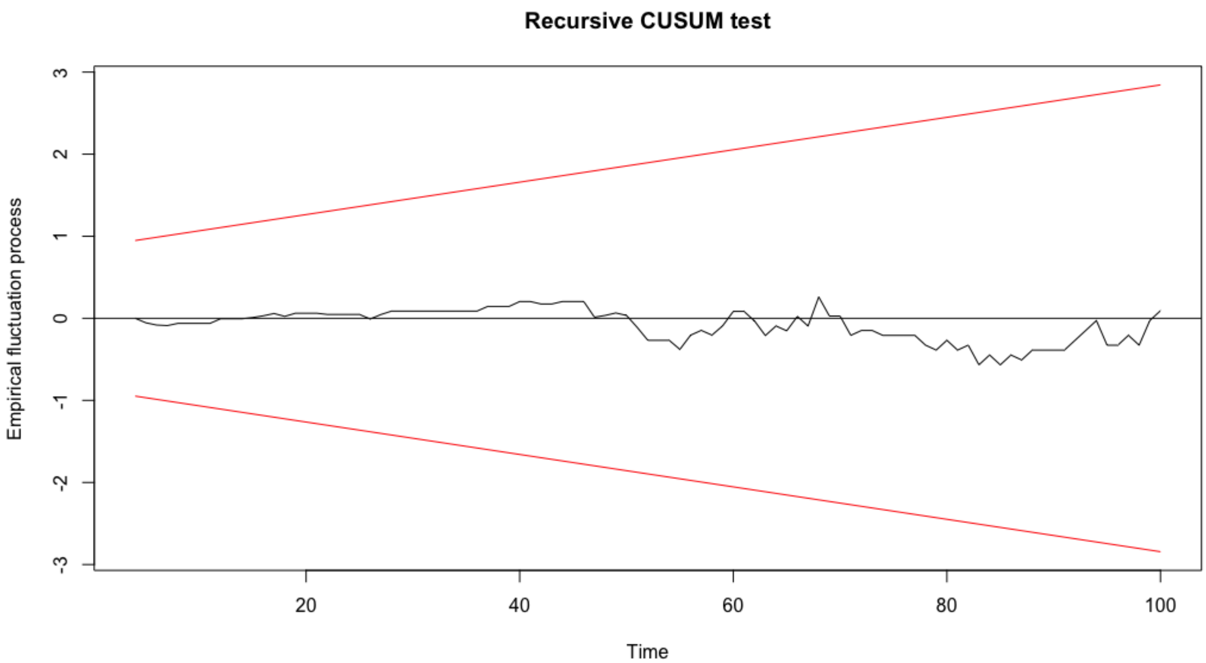

Figure 2: