Topic 4: Model Stability & Specification Analysis

1. Our results to date presume that our regression model holds for the full sample that we are working with.
2. Our results also presume that the model is correctly specified, in the following sense:
   a) The functional form is correct.
   b) All of the relevant regressors have been included.
   c) No redundant regressors have been included.
   d) The only "unexplained" variation in the dependent variable is purely random "noise", as represented by a "well-behaved" error term.

- In this section we'll re-consider item 1, above, and items 2 (b) & (c).
- The other items will be considered later.

Specification Analysis (*Henri Theil, 1957*)

We'll consider various issues associated with the choice of regressors in our linear regression model.

Omission of Relevant Regressors

D.G.P.: $\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$; $E[\boldsymbol{\varepsilon}] = \boldsymbol{0}$

F.M.: $\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{u}$

So,
$$\boldsymbol{b}_1 = (X_1'X_1)^{-1}X_1'\boldsymbol{y}$$
$$= (X_1'X_1)^{-1}X_1'(X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon})$$
$$= \boldsymbol{\beta}_1 + (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_2 + (X_1'X_1)^{-1}X_1'\boldsymbol{\varepsilon}$$

Let's consider the bias of this estimator –

$$E[b_1] = \boldsymbol{\beta}_1 + (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_2$$
$$\neq \boldsymbol{\beta}_1 \quad ; \quad \text{unless } X_1'X_2 = 0 \ ; \ \text{or } X_2\boldsymbol{\beta}_2 = \boldsymbol{0}$$

- So, in general, this estimator will be **Biased**.
- This is just an example of imposing false restrictions on some elements of the $\boldsymbol{\beta}$ vector.
- The estimator, b_1, will also be **inconsistent**.
- However, there will be a reduction in the variance of the estimator, through the imposition of the restrictions, even though they are *false*.

Now consider the converse situation –

Inclusion of Irrelevant Regressors

D.G.P.: $\quad \boldsymbol{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \quad ; \quad E[\boldsymbol{\varepsilon}] = \mathbf{0}$

F.M.: $\quad \boldsymbol{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{u} = X\boldsymbol{\beta} + \boldsymbol{u}$

where,

$$X = [X_1\,, X_2] \quad ; \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$$

So,

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} = (X'X)^{-1}X'\boldsymbol{y}$$

$$= (X'X)^{-1}X'(X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}) \ .$$

Now, we can write: $\quad X_1 = (X_1\,, X_2)\begin{pmatrix} I \\ 0 \end{pmatrix} = XS \quad ,$ say.

So,

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} = (X'X)^{-1}X'XS\boldsymbol{\beta}_1 + (X'X)^{-1}X'\boldsymbol{\varepsilon}$$

$$= S\boldsymbol{\beta}_1 + (X'X)^{-1}X'\boldsymbol{\varepsilon} \ .$$

Then,

$$E[\boldsymbol{b}] = E\begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} = S\boldsymbol{\beta}_1 = \begin{pmatrix} I \\ 0 \end{pmatrix}\boldsymbol{\beta}_1 = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix} \ .$$

That is,

$$E[\boldsymbol{b}_1] = \boldsymbol{\beta}_1 \quad ; \quad \text{and} \quad E[\boldsymbol{b}_2] = \mathbf{0} \ \ (= \boldsymbol{\beta}_2) \ .$$

So, in this case the LS estimator is **Unbiased** (and also **Consistent**).

- In the case where we include irrelevant regressors, we are effectively *ignoring some valid restrictions* on $\boldsymbol{\beta}$.
- Although the LS estimator is Unbiased, it is also **Inefficient**.
- The "costs" of wrongly omitting regressors usually exceed those of wrongly including extraneous ones.
- This suggests that a "***General-to-Specific***" model building strategy may be preferable to a "Specific-to-General" one. (*David Hendry*)
- Over-fit the model, then simplify it on the basis of significance and specification testing.
- Generally this involves a *sequence* of "nested" hypotheses – increasingly restrictive. Stop when restrictions are rejected.
- **Issues:** (a) Degrees of freedom; (b) Loss of precision; (c) Dependence of test statistics, and distortions due to "pre-testing".

Testing for Structural Change

- Suppose that a "shift" in the model occurs due to some exogenous "shock".
- Define a *Dummy Variable*:

  $D_t = 0$; before the shock

  $D_t = 1$; after the shock
- Need not involve "time". Could be 2 regions, for example.
- Could be more than one "shift".
- Do the values of the Dummy variable have to be 0 and 1?
- Then, consider a model of the form:

$$y_t = \beta_1 + \beta_2 x_{2t} + \cdots + \beta_k D_t + \varepsilon_t$$

- We could then think of testing

$$H_0 : \beta_k = 0 \quad vs. \quad H_A : \beta_k \neq 0$$

- Rejection of H_0 implies there is a particular type of ***structural change*** in the model. (*A shift in the level.*)
- Or, more generally, consider a model of the form:

$$y_t = \beta_1 + \beta_2 x_{2t} + \cdots + \beta_{k-1}(D_t \times x_{2t}) + \beta_k D_t + \varepsilon_t$$

- We could then think of testing

$$H_0: \beta_{k-1} = \beta_k = 0 \quad vs. \quad H_A: Not\ H_0$$

- Rejection of H_0 implies there is a different type of ***structural change*** in the model. (*A shift in the level and one of the marginal effects.*)
- Using the dummy variable *fully*, in this way (with intercept and *all* slope coefficients) turns out to be equivalent to the following –

The Chow Test (*Gregory Chow, 1960*)

- Suppose there is a natural break-point in the sample after n_1 observations, and we have:

$$\boldsymbol{y}_1 = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \quad ; \quad \boldsymbol{\varepsilon}_1 \sim N[0\,,\sigma^2 I_{n_1}] \qquad (n_1)$$

$$\boldsymbol{y}_2 = X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2 \quad ; \quad \boldsymbol{\varepsilon}_2 \sim N[0\,,\sigma^2 I_{n_2}] \qquad (n_2)$$

- X_1 and X_2 relate to the same regressors, but different sub-samples. Similarly for $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. Let $n = (n_1 + n_2)$.
- We can write the full model as:

$$\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix} = \begin{bmatrix}X_1 & 0\\ 0 & X_2\end{bmatrix}\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix} + \begin{pmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{pmatrix}$$

$$(n \times 1) \qquad (n \times 2k) \quad (2k \times 1) \quad (n \times 1)$$

or,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad ; \quad \boldsymbol{\varepsilon} \sim N[0\,,\sigma^2 I_n]$$

- If we estimate each part of the model separately, using LS, we get:

$$\boldsymbol{b_1} = (X_1{}'X_1)^{-1}X_1{}'\boldsymbol{y}_1 \quad ; \quad \boldsymbol{e}_1 = \boldsymbol{y}_1 - X_1\boldsymbol{b}_1$$
$$\boldsymbol{b_2} = (X_2{}'X_2)^{-1}X_2{}'\boldsymbol{y}_2 \quad ; \quad \boldsymbol{e}_2 = \boldsymbol{y}_2 - X_2\boldsymbol{b}_2$$

- The total sum of squared residuals for all $n = (n_1 + n_2)$ observations is then:

$$\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{e}_1{}'\boldsymbol{e}_1 + \boldsymbol{e}_2{}'\boldsymbol{e}_2$$

Suppose that we want to test $H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \quad vs. \quad H_A: \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$

- That is we want to test the null hypothesis "*There is no structural break*".
- One way to interpret this problem is as follows:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$H_0: R\boldsymbol{\beta} = \boldsymbol{q} \quad vs. \quad H_A: R\boldsymbol{\beta} \neq \boldsymbol{q}$$

where: $R = [I_k \quad -I_k] \quad ; \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \quad ; \quad \boldsymbol{q} = \boldsymbol{0}$.

If there are k regressors, then $\boldsymbol{q}$ is $(k \times 1)$, and $J = k$.

- Then, we can apply the usual F-test for exact linear restrictions:

$$F = (R\boldsymbol{b} - \boldsymbol{q})'[R(X'X)^{-1}R']^{-1}(R\boldsymbol{b} - \boldsymbol{q})/(ks^2)$$

$F \sim F_{k,n-2k}$ if H_0 is True

- Alternatively, recall that we can write the test statistic as:

$$F = \frac{[(\boldsymbol{e}_*'\boldsymbol{e}_*) - (\boldsymbol{e}'\boldsymbol{e})]/k}{(\boldsymbol{e}'\boldsymbol{e})/(n_1 + n_2 - 2k)}$$

- Here, $\boldsymbol{e}_*$ is the residual vector associated with the RLS estimator , $\boldsymbol{b}_*$, of $\boldsymbol{\beta}$.
- An easy way to obtain $\boldsymbol{b}_*$, and hence $\boldsymbol{e}_*$, is to write:

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix}$$

$$(n \times 1) \quad (n \times 2k)(2k \times 1) \quad (n \times 1)$$

and then restrict $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \overline{\boldsymbol{\beta}}$ (say), yielding the model:

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \overline{\boldsymbol{\beta}} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix}$$

- That is, we just "stack up the $\boldsymbol{y}$ and X data for both sub-samples – that is, estimate the one model for the full sample.
- This will yield $\boldsymbol{b}_*$, and hence $\boldsymbol{e}_*$.
- Notice that we assumed that $\sigma_1^2 = \sigma_2^2$.
- Major complications without this restriction: "**Behrens-Fisher Problem**".

- If we have random regressors, we can still use the **Wald Test**.
- $kF \overset{d}{\rightarrow} \chi^2_{(k)}$; if H_0 is True.

Example

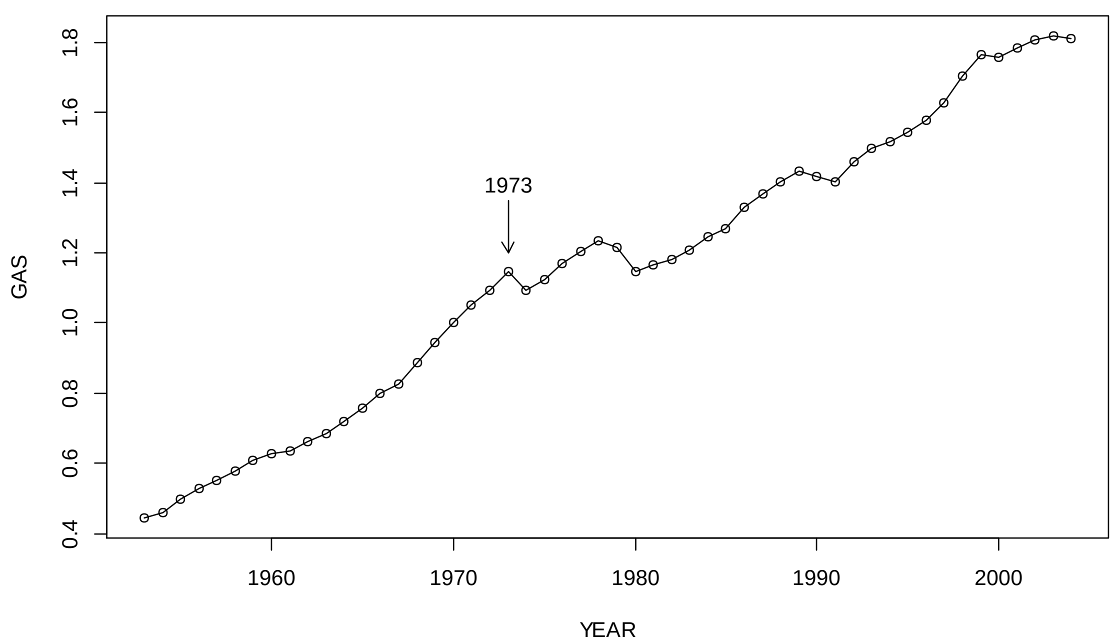
Let's see this illustrated. We'll see *two* equivalent ways of testing for this type of structural change.

Consider the following model for per-capita gasoline consumption[1]:

$$\ln GAS = \beta_1 + \beta_2 YEAR + \beta_3 \ln Income/Pop + \beta_4 \ln GASP + \beta_5 \ln PNC + \beta_6 \ln PUC + \varepsilon$$

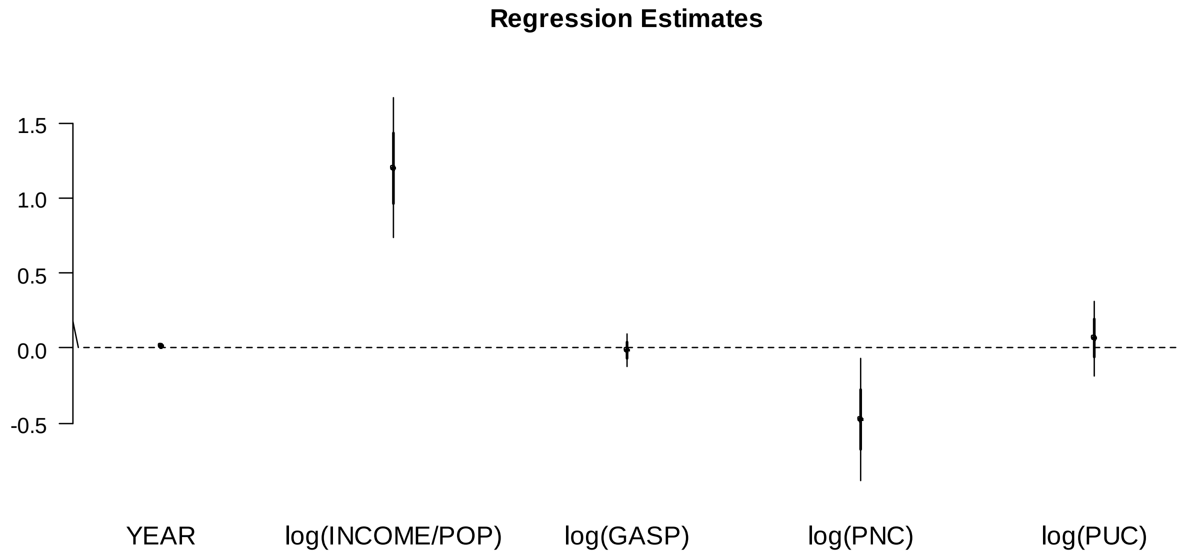
Where *GASP* is the price of gasoline, *PNC* is the price of new cars, and *PUC* the price of used cars. We will consider an exogenous shock for the year 1973.

Per Capita Gasoline Consumption (U.S.A.)



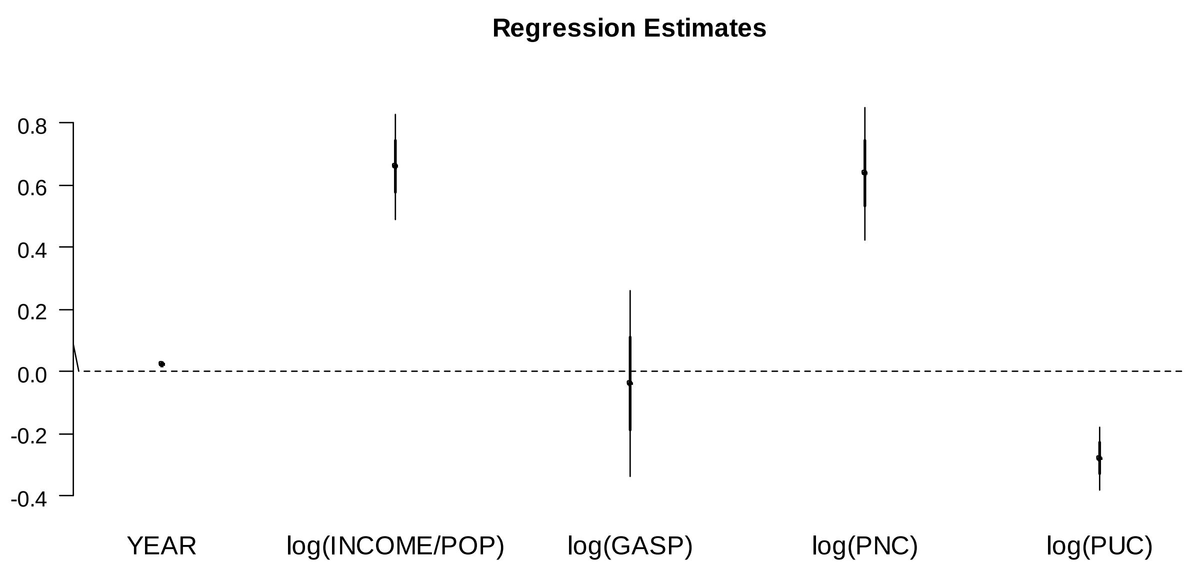
[1] Data from Greene (2012), Table F2.2

Estimate the pooled model (using all observations):



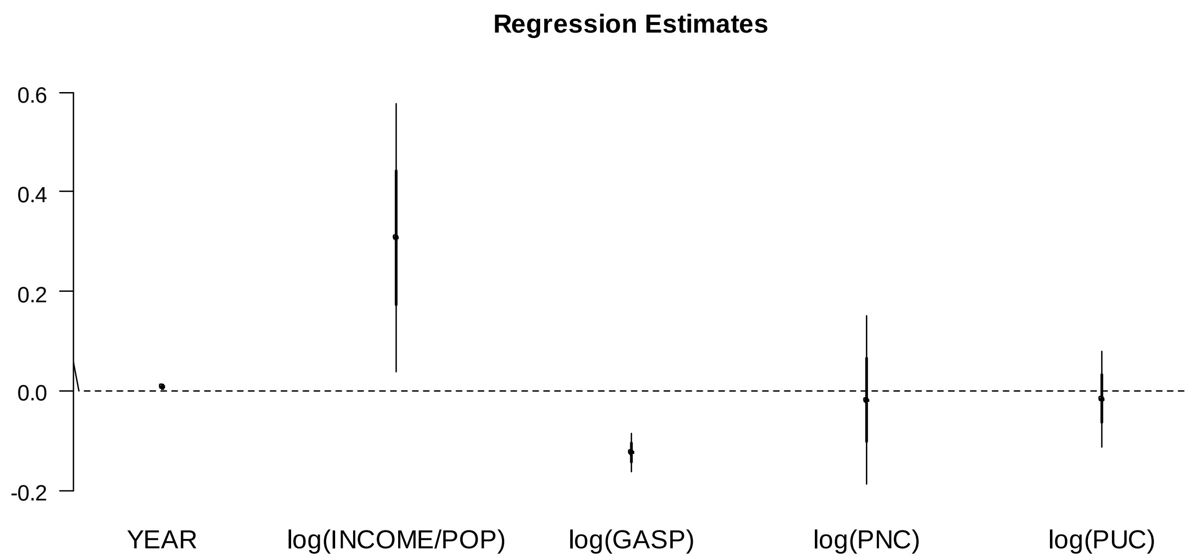
$$e_*' e_* = 0.16302$$

Re-estimate the model using data up to 1973 only (pre-shock data):



$$e_1' e_1 = 0.00184$$

Re-estimate the model using data after 1973 only (post-shock data):



$$e_2'e_2 = 0.00739$$

Chow test:

$$F = \frac{[(\boldsymbol{e}_*{}'\boldsymbol{e}_*) - (\boldsymbol{e}_1'\boldsymbol{e}_1 + \boldsymbol{e}_2'\boldsymbol{e}_2)]/k}{(\boldsymbol{e}_1'\boldsymbol{e}_1 + \boldsymbol{e}_2'\boldsymbol{e}_2)/(n_1 + n_2 - 2k)} = \frac{[0.16302 - 0.00184 - 0.00739]/6}{(0.00184 + 0.00739)/(52 - 12)} = 111.267$$

From an F-distribution with 6 and 40 degrees of freedom, the p-value associated with this test statistic is 0.000.

An alternate way to calculate this test statistic is to estimate a model using dummy variables, and perform an F-test for the joint significance of all coefficients associated with a dummy variable.

DUM = 0 (1953 – 1973) ; = 1 (1974 – 2004)

```
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -60.998170   5.308283 -11.491 3.03e-14 ***
YEAR                  0.024922   0.002960   8.420 2.15e-10 ***
log(INCOME/POP)       0.660168   0.116328   5.675 1.35e-06 ***
log(GASP)            -0.036362   0.205657  -0.177 0.860553
log(PNC)              0.638100   0.146745   4.348 9.18e-05 ***
log(PUC)             -0.279605   0.069318  -4.034 0.000240 ***
DUM                  31.954337   5.859984   5.453 2.77e-06 ***
YEAR:DUM             -0.015663   0.003184  -4.920 1.53e-05 ***
log(INCOME/POP):DUM  -0.352420   0.166666  -2.115 0.040750 *
log(GASP):DUM        -0.087200   0.206332  -0.423 0.674837
log(PNC):DUM         -0.656235   0.164627  -3.986 0.000277 ***
log(PUC):DUM          0.263556   0.081286   3.242 0.002394 **
```

$$e'e = 0.00922$$

Note that $e_1'e_1 + e_2'e_2 = e'e$!

Insufficient Degrees of Freedom

- What if either $n_1 < k$, or $n_2 < k$?
- In this case we can't fit one of the sub-sample regressions, so F can't be computed.
- There is a second version of the Chow test, designed for this situation ("Chow Forecast Test").

Also, note:

- Location of break-point(s) assumed known.
- Situation becomes much more complicated if we have to *estimate* break-point locations(s).

Using the Wald Test

- If *any* of the usual assumptions that underly the F-test for exact linear restrictions are violated, then the usual Chow test is ***not valid***.
- We can, however, still use the Wald test version of the Chow test.
- It will be valid only asymptotically (large *n*).
- It may have poor performance in small samples.
- Examples where we would use the Wald version of the Chow test –
    1. Random regressors (*e.g.*, lagged dependent variable).
    2. Non-Normal errors.

Appendix – R Code

```
#Data is from Greene, Table F2.2
#You will have to install the "arm" package if you wish to use "coefplot".
library(arm)
gasdata = read.csv("http://home.cc.umanitoba.ca/~godwinrt/7010/gas.csv")
attach(gasdata)

#View the break-point:
plot(YEAR,GAS)
lines(YEAR,GAS)
text(1973,1.4,"1973")
arrows(1973,1.35,1973,1.2,length = 0.1)

#Estimate the pooled model:
eq1 = lm(log(GASEXP/GASP/POP) ~ YEAR + log(INCOME/POP) + log(GASP) + log(PNC)
  + log(PUC))
#View the estimated coefficients:
coefplot(eq1,vertical=FALSE,var.las = 1,cex.var=1.2)
#Get the sum of squared residuals from the pooled (restricted) model:
sser = sum(eq1$residuals^2)

#Use only the first 21 observations (up to 1973):
preshock = gasdata[1:21,]
attach(preshock)
eq2 = lm(log(GASEXP/GASP/POP) ~ YEAR + log(INCOME/POP) + log(GASP) + log(PNC)
  + log(PUC))
coefplot(eq2,vertical=FALSE,var.las = 1,cex.var=1.2)
sseu1 = sum(eq2$residuals^2)

#Use only the last 31 observations (after 1973):
postshock = gasdata[22:52,]
```

```r
attach(postshock)
eq3 = lm(log(GASEXP/GASP/POP) ~ YEAR + log(INCOME/POP) + log(GASP) + log(PNC)
  + log(PUC))
coefplot(eq3,vertical=FALSE,var.las = 1,cex.var=1.2)
sseu2 = sum(eq3$residuals^2)

#Calculate Chow test statistic:
chow = ((sser - sseu1 - sseu2)/6)/((sseu1 + sseu2)/(52 - 12))
#p-value:
1 - pf(chow,6,40)

#Estimate the model with dummy variables:
DUM = c(rep(0,21),rep(1,31))
attach(gasdata)
eq4 = lm(log(GASEXP/GASP/POP) ~ YEAR + log(INCOME/POP) + log(GASP) + log(PNC)
  + log(PUC) + DUM + DUM*YEAR + DUM*log(INCOME/POP) + DUM*log(GASP) +
  DUM*log(PNC) + DUM*log(PUC))
summary(eq4)
ssedum = sum(eq4$residuals^2)
```