Econ 423 – Lecture Notes

(These notes are modified versions of lecture notes provided by Stock and Watson, 2007. They are for instructional purposes only and are not to be distributed outside of the classroom.)

Heteroskedasticity and Autocorrelation-Consistent (HAC) Standard Errors

- Consider a generalization of the ***distributed lag model***, where the errors u_t are not necessarily i.i.d., i.e.,

$$Y_t = \beta_0 + \beta_1 X_t + \ldots + \beta_{r+1} \mathrm{X}_{t-r} + u_t \ .$$

- Suppose that u_t is serially correlated; then, OLS will still yield **consistent*** estimators of the coefficients β_0 , β_1 ,….., β_{r+1} (**consistent but possibly biased!*)
- The sampling distribution of $\hat{\beta}_1$, etc., is normal

- ***BUT*** the formula for the variance of this sampling distribution is not the usual one from cross-sectional (i.i.d.) data, because u_t is not i.i.d. in this case since, in particular, u_t is serially correlated!
- This means that the usual OLS standard errors (usual STATA printout) are wrong!
- We need to use, instead, *SE*s that are robust to autocorrelation as well as to heteroskedasticity…
- This is easy to do using STATA and most (but not all) other statistical software.

HAC standard errors, ctd.

The math…for the simplest case with no lags:

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

The OLS estimator: Using the usual regression algebra, we obtain

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})u_t}{\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})^2}$$

$$\cong \frac{\frac{1}{T}\sum_{t=1}^{T} v_t}{\sigma_X^2} \quad \text{(in large samples)}$$

where $v_t = (X_t - \bar{X})u_t$.

HAC standard errors, ctd.

Thus, in large samples,

$$\text{var}(\hat{\beta}_1) = \text{var}\left(\frac{1}{T}\sum_{t=1}^{T} v_t\right)/(\sigma_X^2)^2$$

$$= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\text{cov}(v_t, v_s)/(\sigma_X^2)^2$$

In i.i.d. cross sectional data, $\text{cov}(v_t, v_s) = 0$ for $t \neq s$, so

$$\text{var}(\hat{\beta}_1) = \frac{1}{T^2}\sum_{t=1}^{T}\text{var}(v_t))/(\sigma_X^2)^2 = \frac{\sigma_v^2}{T(\sigma_x^2)^2}$$

This is our usual cross-sectional result.

HAC standard errors, ctd.

But in time series data, $\text{cov}(v_t, v_s) \neq 0$ in general.

Consider $T = 2$:

$$\text{var}\left(\frac{1}{T}\sum_{t=1}^{T} v_t\right) = \text{var}[\tfrac{1}{2}(v_1+v_2)]$$

$$= \tfrac{1}{4}[\text{var}(v_1) + \text{var}(v_2) + 2\text{cov}(v_1,v_2)]$$

$$= \tfrac{1}{2}\sigma_v^2 + \tfrac{1}{2}\rho_1\sigma_v^2 \qquad (\rho_1 = \text{corr}(v_1,v_2))$$

$$= \tfrac{1}{2}\sigma_v^2 \times f_2, \text{ where } f_2 = (1+\rho_1)$$

- In i.i.d. data, $\rho_1 = 0$ so $f_2 = 1$, yielding the usual formula
- In time series data, if $\rho_1 \neq 0$ then $\text{var}(\hat{\beta}_1)$ is ***not*** given by the usual formula.

Expression for var($\hat{\beta}_1$), general T

$$\text{var}\left(\frac{1}{T}\sum_{t=1}^{T} v_t\right) = \frac{\sigma_v^2}{T} \times f_T$$

so

$$\text{var}(\hat{\beta}_1) = \left[\frac{1}{T}\frac{\sigma_v^2}{(\sigma_X^2)^2}\right] \times f_T$$

where

$$f_T = 1 + 2\sum_{j=1}^{T-1}\left(\frac{T-j}{T}\right)\rho_j$$

- Conventional OLS SE's are wrong when u_t is serially correlated (STATA printout is wrong).
- The OLS *SE*s are off by the factor f_T
- We need to use a different *SE* formula!!!

HAC Standard Errors

- Conventional OLS *SEs* (heteroskedasticity-robust or not) are wrong when u_t is autocorrelated
- So, we need a new formula that produces *SE*s that are robust to autocorrelation as well as heteroskedasticity

  *We need **H**eteroskedasticity- and **A**utocorrelation-**C**onsistent (**HAC**) standard errors*
- If we knew the factor f_T, we could just make the adjustment. However, in most practical applications, we must estimate f_T.

HAC SEs, ctd.

$$\text{var}(\hat{\beta}_1) = \left[\frac{1}{T}\frac{\sigma_v^2}{(\sigma_X^2)^2}\right] \times f_T, \text{ where } f_T = 1 + 2\sum_{j=1}^{T-1}\left(\frac{T-j}{T}\right)\rho_j$$

The most commonly used estimator of f_T is:

$$\hat{f}_T = 1 + 2\sum_{j=1}^{m-1}\left(\frac{m-j}{m}\right)\tilde{\rho}_j \quad \text{(Newey-West)}$$

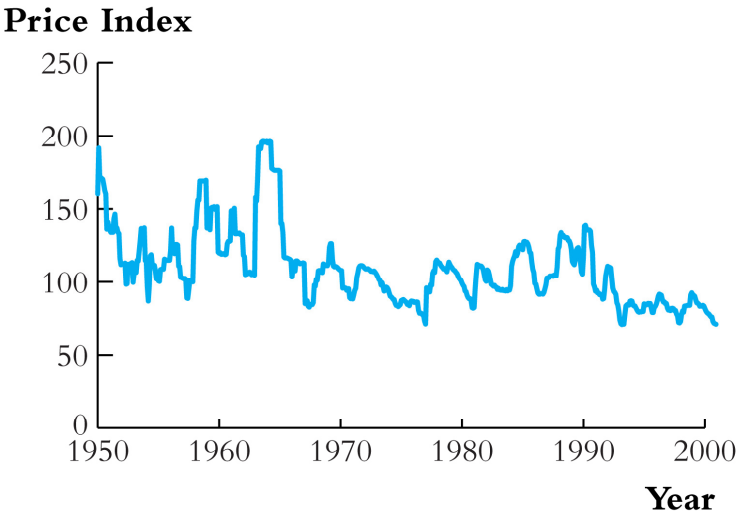
- $\tilde{\rho}_j$ is an estimator of ρ_j
- This is the "Newey-West" HAC *SE* estimator
- m is called the ***truncation parameter***
- Why not just set $m = T$?
- Then how should you choose m?
  - Use the Goldilocks method
  - Or, use the rule of thumb, $m = 0.75T^{1/3}$

Empirical Example
The Orange Juice Data

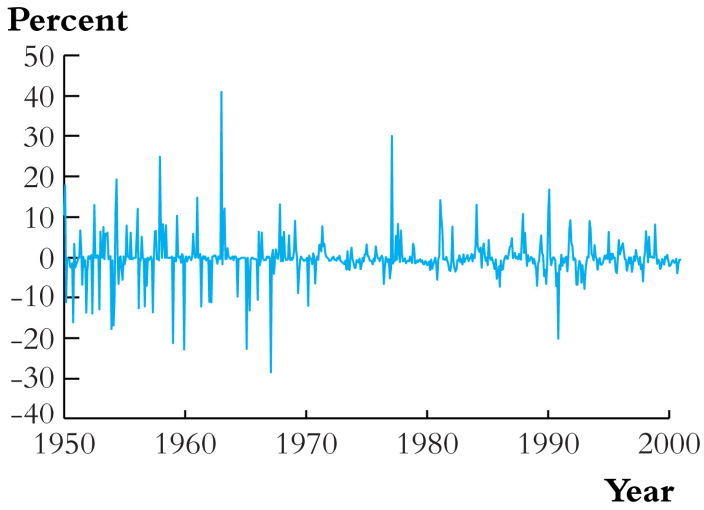
Data

- Monthly, Jan. 1950 – Dec. 2000 ($T = 612$)
- *Price* = price of frozen OJ (a sub-component of the producer price index; US Bureau of Labor Statistics)
- *%ChgP* = percentage change in price at an annual rate, so $\%ChgP_t = 1200\Delta\ln(Price_t)$
- *FDD* = number of freezing degree-days during the month, recorded in Orlando FL
  - Example: If November has 2 days with lows < 32^o, one at 30^o and at 25^o, then $FDD_{Nov} = 2 + 7 = 9$

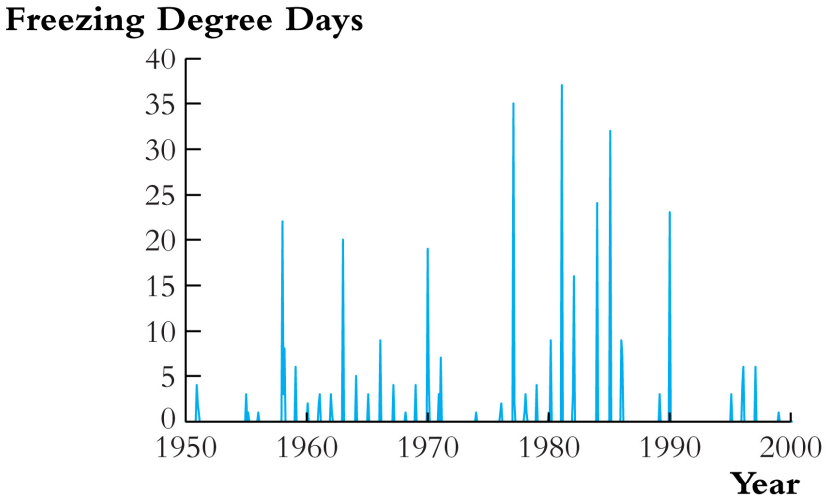
FIGURE 15.1 Orange Juice Prices and Florida Weather, 1950–2000



(a) Price Index for Frozen Concentrated Orange Juice



(b) Percent Change in the Price of Frozen Concentrated Orange Juice



(c) Monthly Freezing Degree Days in Orlando, Florida

Initial OJ regression

$$\widehat{\%ChgP}_t = -.40 + .47FDD_t$$
$$(.22)\quad(.13)$$

- Statistically significant positive relation
- More freezing degree days $\Rightarrow$ price increase
- Standard errors are heteroskedasticity and autocorrelation-consistent (HAC) SE's

Example: OJ and HAC estimators in STATA

```
. gen l0fdd = fdd;                    generate lag #0
. gen l1fdd = L1.fdd;                 generate lag #1
. gen l2fdd = L2.fdd;                 generate lag #2
. gen l3fdd = L3.fdd;                     .
. gen l4fdd = L4.fdd;                     .
. gen l5fdd = L5.fdd;                     .
. gen l6fdd = L6.fdd;

. reg dlpoj fdd if tin(1950m1,2000m12), r;   NOT HAC SEs

Linear regression                                   Number of obs =     612
                                                    F(  1,    610) =   12.12
                                                    Prob > F      =  0.0005
                                                    R-squared     =  0.0937
                                                    Root MSE      =  4.8261

------------------------------------------------------------------------------
             |               Robust
       dlpoj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fdd |   .4662182   .1339293     3.48   0.001     .2031998    .7292367
       _cons |  -.4022562   .1893712    -2.12   0.034    -.7741549   -.0303575
------------------------------------------------------------------------------
```

Example: OJ and HAC estimators in STATA, ctd

Rerun this regression, but with Newey-West *SEs*:

```
. newey dlpoj fdd if tin(1950m1,2000m12), lag(7);

Regression with Newey-West standard errors          Number of obs   =       612
maximum lag: 7                                      F(  1,   610)   =     12.23
                                                    Prob > F        =    0.0005

------------------------------------------------------------------------------
             |             Newey-West
       dlpoj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fdd |   .4662182   .1333142     3.50   0.001     .2044077    .7280288
       _cons |  -.4022562   .2159802    -1.86   0.063    -.8264112    .0218987
------------------------------------------------------------------------------
```

Uses autocorrelations up to m = 7 to compute the SEs

rule-of-thumb: $0.75*(612^{1/3}) = 6.4 \approx 7$, rounded up a little.

OK, in this case the difference in SEs is small, but not always so!

Example: OJ and HAC estimators in STATA, ctd.

```
. global lfdd6 "fdd l1fdd l2fdd l3fdd l4fdd l5fdd l6fdd";

.  newey dlpoj $lfdd6 if tin(1950m1,2000m12), lag(7);

Regression with Newey-West standard errors                Number of obs   =       612
maximum lag : 7                                           F(  7,   604)   =      3.56
                                                          Prob > F        =    0.0009

------------------------------------------------------------------------------
             |             Newey-West
       dlpoj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fdd |    .4693121   .1359686     3.45   0.001     .2022834    .7363407
       l1fdd |    .1430512   .0837047     1.71   0.088    -.0213364    .3074388
       l2fdd |    .0564234   .0561724     1.00   0.316    -.0538936    .1667404
       l3fdd |    .0722595   .0468776     1.54   0.124    -.0198033    .1643223
       l4fdd |    .0343244   .0295141     1.16   0.245    -.0236383    .0922871
       l5fdd |    .0468222   .0308791     1.52   0.130    -.0138212    .1074657
       l6fdd |    .0481115   .0446404     1.08   0.282    -.0395577    .1357807
       _cons |   -.6505183   .2336986    -2.78   0.006    -1.109479   -.1915578
------------------------------------------------------------------------------
```

- global lfdd6 *defines a string which is all the additional lags*
- *What are the estimated dynamic multipliers (dynamic effects)?*

FAQ: Do I need to use HAC *SE*s when I estimate an AR or an ADL model?

A: No, only if one is sure that the true model is an AR or an ADL in the purest sense so that there is no serial correlation or heteroskedasticity in the errors.

- In AR and ADL models with homoskedastic errors, one may argue that the errors will be serially uncorrelated if you include enough lags of Y
  - If you include enough lags of Y, then the error term can't be predicted using past Y, or equivalently by past u – so u is serially uncorrelated
- However, the safer and more robust choice would be to always use HAC SE's.