Lecture 13
Auto/cross-correlation

Generalized Regression Model

• The generalized regression model's assumptions:
  (A1) DGP: $y = X\beta + \varepsilon$ is correctly specified.
  (A2) $E[\varepsilon | X] = 0$
  (A3') $\text{Var}[\varepsilon | X] = \Sigma = \sigma^2 \Omega$.
  (A4) $X$ has full column rank – $\text{rank}(X) = k$, where $T \geq k$.

• We assume that the $\varepsilon$’s in the sample are no longer generated independently of each other. Ignoring heteroscedasticity, we have a new $\Sigma$:
  
  \[
  E[\varepsilon_i \varepsilon_j | X] = \begin{cases} 
    \sigma_{ij} & \text{if } i \neq j \\
    \sigma^2 & \text{if } i = j 
  \end{cases}
  \]
Autocorrelation: Examples

- We keep the linear model: \( y_t = X_t \beta + \varepsilon_t \)
  
  - First-order autoregressive autocorrelation: AR(1)
    \[ \varepsilon_t = \rho \varepsilon_{t-1} + u_t \]
  
  - Fifth-order autoregressive autocorrelation: AR(5)
    \[ \varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \rho_3 \varepsilon_{t-3} + \rho_4 \varepsilon_{t-4} + \rho_5 \varepsilon_{t-5} + u_t \]
  
  - Third-order moving average autocorrelation: MA(3)
    \[ \varepsilon_t = u_t + \lambda_1 u_{t-1} + \lambda_2 u_{t-2} + \lambda_3 u_{t-3} \]

Note: The last example is described as third-order moving average autocorrelation, denoted MA(3), because it depends on the three previous innovations as well as the current one.

Implications for OLS

- Similar to the heteroscedasticity results:
  - OLS is unbiased, consistent (we need additional assumptions), asymptotic normality (we need additional assumptions and definitions), but inefficient.
  - OLS standard errors are incorrect, often biased downwards.

- A very important exception: The lagged dependent variable
  \[ y_t = \beta x_t + \gamma y_{t-1} + \varepsilon_t \]

  Now, \( \text{Cov}[y_{t-1}, \varepsilon_t] \neq 0 \implies \text{IV Estimation} \)

- Useful strategy: OLS estimates with the Newey-West (NW) robust estimation of the covariance matrix. Recall NW’s HAC estimator of \( Q^* \):
  \[ S_T = S_0 + \frac{1}{T} \sum_{j=1}^{T} (\Sigma \varepsilon_t \varepsilon_t') + \frac{1}{T} \sum_{j=1}^{T} (x_t \varepsilon_t) (x_t \varepsilon_t') \]
Newey-West estimator

- The performance of NW estimators depends on the choice of the kernel function –i.e., $k_L$– and truncation lag ($L$). These choices affect the resulting test statistics and render testing results fragile.

- NW SEs perform poorly in Monte Carlo simulations: the finite-sample performance of tests using NW SE is not well approximated by the asymptotic theory (big size problems), especially when show $x_t$ moderate or high persistence:
  - The kernel weighting scheme yields negative bias –i.e., NW SEs are downward biased–, which could be big in finite samples.
  - The tests based on the NW SE usually over-reject $H_0$.
  - A relatively small $L$ is needed to minimize MSE, which leads to considerable bias of the $Q^*$ estimator (&, then, distorts the size the tests). Minimizing size distortions needs a larger $L$.

Newey-West estimator: Implementation

- To implement the HAC estimator, we need to determine: lag order –i.e., truncation lag ($L$) or bandwidth–, and kernel choice ($k_L (L)$).

(1) Truncation lag ($L$)
No optimal formula; though selecting $L$ to minimize MSE is popular.

To determine $L$, we use:
  - Trial and error, informed guess.
  - Rules of thumb. For example: $L = 0.75T^{1/3}$. 1.

The choice of $L$ matters. In general, for ARMA models we have:
  - Shorter lags: Larger Bias, Smaller Variance
  - Longer lags: Smaller Bias, Larger Variance
Newey-West estimator: Implementation

- Usual practical advise regarding $L$: Choose $L$ (lags) a little longer than you might otherwise.

- Sun et al. (2008) give some intuition for a longer $L$ than the optimal MSE $L^*$, by expanding the probability of a test. Simple example:

Let $z \sim N(0, \sigma^2)$, $\hat{s}^2$ is an estimator of $\sigma^2$ (assume independent of $z$).

$$\Pr\left(\frac{z^2}{\hat{s}^2} < c\right) = \Pr(z^2 < s^2 c) = E[I[z^2 < s^2 c]] = E(g(\hat{s}^2))$$

$$\approx E(g(\sigma^2)) + E((s^2 - \sigma^2)g'(\sigma^2)) + \frac{1}{2} E((s^2 - \sigma^2)^2)g''(\sigma^2)$$

$$= F_{\hat{s}^2}(c) + \text{Bias}(\hat{s}^2)g'(\sigma^2) + \frac{1}{2} \text{MSE}(\hat{s}^2)g''(\sigma^2)$$

$\Rightarrow$ Equal weight MSE + Bias. Long $L$ minimizes the Bias; better size!

Newey-West estimator: Implementation

(2) Kernel Choice

- In theory, the kernel choice matters.
- In practice, at least for psd kernels, it does not seem to matter.

![Kernel Choice Graph](image)

- Based on the work of Andrews (1991), where he finds a HAC that minimizes the AMSE of the LRV, the QS kernel tends to be the default kernel in computations of HAC SE.
NW Estimator: Improvements

- Other than finding a good kernel and a (long) $L$, the performance of HAC estimators may be improved by:

1. Pre-whitening the data - Andrews and Monahan (1992). Regress $x_t e_t$ on its lagged values. Some arbitrary choice in the selection of the lag order to do the regression.

2. Computing sample autocovariances based on forecast errors, instead of OLS residuals - Kuan and Hsieh (2006). Replace $e_t$ with one-step-ahead forecast errors: $f e_t = y_t - X_t' b_{t-1}$, where $b_{t-1}$ is the recursive OLS estimators based on the subsample of the first $t - 1$ observations.

NW Estimator: Improvements - Example

**Example:** We compute different NW SE for the 3 factor F-F model for IBM returns, with bandwidth selected as in Andrews (1991):

```r
> library(sandwich)
> reg <- lm(y~x-1)
> reg$coefficients
  x     xx1     xx2     xx3
-0.2331470817  0.0101872239  0.0009802843 -0.0044459013

$\Rightarrow$ OLS $b$

```r
> sqrt(diag(kernHAC(reg, prewhite = 0, bw = bwAndrews, kernel = "Quadratic Spectral", verbose = TRUE)))
  x     xx1     xx2     xx3
 0.020959375 0.002848645 0.003983330 0.005310548

$\Rightarrow$ Bandwidth chosen: 3.035697

```r
> sqrt(diag(kernHAC(reg, prewhite = 0, bw = bwAndrews, kernel = "Bartlett", verbose = TRUE)))
  x     xx1     xx2     xx3
 0.020344074 0.002828663 0.003995942 0.005177482

$\Rightarrow$ Bandwidth chosen: 3.507051

```r
> sqrt(diag(kernHAC(reg, prewhite = 0, bw = bwAndrews, kernel = "Parzen", verbose = TRUE)))
  x     xx1     xx2     xx3
 0.022849506 0.002839034 0.003954436 0.005427730

$\Rightarrow$ Bandwidth chosen: 6.110888
NW Estimator: Improvements - Example

**Example:** Now, we also pre-white the data (prewhite = 1):

```r
> sqrt(diag(kernHAC(reg, prewhite = 1, bw = bwAndrews, kernel = "Quadratic Spectral", verbose = TRUE)))
x  xx1  xx2  xx3
0.043339699 0.002908898 0.004029606 0.005783013  ⟹ & Bandwidth chosen: 0.8118876
```

```r
> sqrt(diag(kernHAC(reg, prewhite = 1, bw = bwAndrews, kernel = "Bartlett", verbose = TRUE)))
x  xx1  xx2  xx3
0.042943572 0.002912273 0.004022336 0.005786720  ⟹ & Bandwidth chosen: 0.516233
```

```r
> sqrt(diag(kernHAC(reg, prewhite = 1, bw = bwAndrews, kernel = "Parzen", verbose = TRUE)))
x  xx1  xx2  xx3
0.040963950 0.002912789 0.004006919 0.005767432  ⟹ & Bandwidth chosen: 1.634337
```

• **Note:** Pre-whitening tends to increase the standard errors (& decrease the bandwidth). Nice result, given that the usual NW SEs tend to be downward biased.

Newey-West estimator: Inconsistency

• Recall that a key assumption in establishing consistency for $S_T$ is that $L \rightarrow \infty$ as $T \rightarrow \infty$, but $L/T \rightarrow 0$.

• In practice, $L/T$ is never equal to 0, but some positive fraction, $b$ ($b \in (0,1]$). Under this situation, the NW estimator is no longer consistent.

• Thus, $t$- and $F$-tests no longer converge in distribution to Normal and $\chi^2$ RVs, but they do converge in distribution to a RV that have non-standard distribution; which do not depend on the unknown value of $\Omega$. Tests are still possible.

• To get asymptotic distributions (& critical values) we use “fixed-$b$” asymptotics. Under fixed-$b$ asymptotics, the truncation parameter, $L$, is treated as proportional to $T$, so $L = bT$, where $b$ is fixed --see, Kiefer, Vogelsang & Bunzell (KVB, 2000), Kiefer & Vogelsang (2002, 2005).
Newey-West estimator: Inconsistency

• Under fixed-\( b \) asymptotics, typically \( S_T \rightarrow Q^{1/2} \Xi Q^{1/2} \), where \( \Xi \) is a RV with \( E[\Xi]=I_p \). \( \Xi \) has a non-standard distribution.

• Kiefer and Vogelsang (2005) derive limiting distribution for \( S_T \), which is complicated, but the 95% critical values (CV) for \( t \)-tests can be constructed using the following polynomial (\( b=L/T \)):

\[
CV \ (L/T) = 1.96 + 2.9694 \ b + 0.416 \ b^2 – 0.05324 \ b^3.
\]

Note: As \( b \rightarrow 0 \), the standard \( t \) critical values apply.

• Since non-standard distributions are not popular, work has been devoted to find simple and intuitive estimators of \( Q^* \) that can be used in tests with traditional distributions (say, \( N(0,1) \) and \( \chi^2 \)).

Newey-West estimator: Inconsistency

• When the frequency domain kernel weights are equal and truncated after the first \( B/2 \) periodogram ordinates (an estimator of the spectrum at frequency \( (2\pi j/T) \)), the limiting fixed-\( b \) distribution of \( S_T \) is a \( \chi^2_{B/B} \).

• This corresponds to the equal-weighted periodogram estimator of \( Q^* \) (the Daniell window):

\[
S_T^{EWP} = \frac{2\pi B^{1/2}}{B} \sum_{j=1}^{B/2} I_{\text{vec}}(2\pi j/T) = \frac{1}{B} \sum_{j=1}^{B/2} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (x_t e^{-i\omega t}) \right) \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (x_t e^{-i\omega t}) \right)^T
\]

Now, the usual \( t \)-test, \( t^{EWP} = \sqrt{T}(\hat{\beta} - \beta_0)/\sqrt{S_T^{EWP}} \), has a \( t_b \) asymptotic distribution under \( H_0 \).

• The EWP estimator has the nice property that fixed-\( b \) asymptotic inference can be conducted using standard \( t \) and \( F \) distributions.
Newey-West estimator: Inconsistency

- In addition, the EWP estimator is psd with probability 1.

- Müller (2007), and Sun (2013) note that other estimators of $Q^*$ can be derived by replacing the Fourier functions in $S_{EWP}^T$ by other basis functions of a general orthonormal set of basis function for $L^2[0,1]$.

- Then, we can see $S_{EWP}^T$ as a especial case of:

$$S_{BF}^T = \frac{1}{B} \sum_{j=1}^B S_j, \quad \text{where} \quad S_j = \hat{\Lambda}_j \hat{\Lambda}_j^T \quad \text{and} \quad \hat{\Lambda}_j = \frac{1}{\sqrt{T}} \sum_{t=1}^T (x_t e_t) \phi_j(t/T)$$

- Different $\phi_j$ basis functions (say, cosine), different estimators.

Note: Since $S_{BF}^T$ is computed using an outer product, it is psd.

Newey-West estimator: KVB

- The (kernel) HAC estimation requires the choices of the kernel function and $L$. Such choices are somewhat arbitrary in practice.

- To avoid these difficulties, Kiefer, Vogelsang, and Bunzel (2000), KVB, proposed an approach that yields an asymptotically pivotal test without consistent estimation of the asymptotic covariance matrix.

- Idea: Use a normalizing matrix to eliminate the nuisance parameters in $Q^{*1/2}$, the matrix square root of $Q^*_T$ & impose no truncation ($b=1$). Let

$$\varphi_t = \frac{1}{\sqrt{T}} \sum_{j=1}^T x_j e_j$$

Normalizing matrix:

$$C_T = \frac{1}{T} \sum_{t=1}^T \varphi_t \varphi_t' = \frac{1}{T^2} \sum_{t=1}^T \left( \sum_{j=1}^T x_j e_j \right) \left( \sum_{j=1}^T c_j x_j' \right)$$
Newey-West estimator: KVB

• Normalizing matrix:

\[ C_T = \frac{1}{T} \sum_{t=1}^{T} \varphi_t \varphi_t' = \frac{1}{T^2} \sum_{t=1}^{T} (\sum_{j=1}^{t} x_j c_j) (\sum_{j=1}^{t} c_j x_j') \]

This normalizing matrix is inconsistent for \( Q^* \) but it is free from the choice of kernel and \( L \). (Note: There is no truncation, \( L=T \) => Good for size of test!)

• We use this \( C_T \) matrix to calculate tests. For example, to test \( r \) restrictions \( H_0: (R \beta - q = 0) \), we have the following statistic

\[ W^+ T = T (R^* b_T - q)^' [R (X'X)^{-1} C_T (X'X)^{-1} R]^{-1} (R^* b_T - q) . \]

Although the asymptotic distribution of \( W^+ T \) is non-standard, it can be simulated - Lobato (2001).

Newey-West estimator: KVB

• KV (2002) showed that \( 2C_T \) is algebraically equivalent to \( Q^{*,B, T} \) (where B stands for Bartlett kernel) without truncation (\( b=1 \)) - i.e., \( L(T)=T \). Then, usual \( W \) based on \( Q^{*,B, T} \) without truncation is the same as \( W^+ T/2 \).

• KVB derive the (non-standard) asymptotic distribution of the conventional \( t \)-test of \( H_0: \beta_j=r; \) but using their robust version, \( t^+ \):

\[ t^+ = \frac{\sqrt{T} (\hat{\delta}_{i,T} - r)}{\sqrt{\hat{\delta}_i}} \frac{W(1)}{\int_0^1 B(r)^2 \, dr}^{1/2} \]

where \( \hat{\delta}_i \) is the i-th diagonal element of \( (X'X)^{-1} C_T (X'X)^{-1} \), \( W \) is a standard Wiener process, and \( B(\hat{r}) \) is a Brownian Bridge - i.e., \( B_\hat{r}(\hat{r}) = W_\hat{r}(\hat{r}) - \hat{r} W_1(1), 0 \leq \hat{r} \leq 1. \) This distribution is symmetric, but more disperse than the \( N(0,1) \).
Newey-West estimator: KVB

• KVB report the quantiles of the asymptotic distribution of the usual t-test, using $C_T$ and using the NW SE, without truncation. (Notation: $Q^* = \Sigma_{\text{kernel}}$)

Table 1: The quantiles of the $t^T$ and $t$ tests based on $\hat{\Sigma}_T$ without truncation.

<table>
<thead>
<tr>
<th>prob.</th>
<th>90%</th>
<th>95%</th>
<th>97.5%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t^T$ with $\hat{C}_T$</td>
<td>3.890</td>
<td>5.374</td>
<td>6.811</td>
<td>8.544</td>
</tr>
<tr>
<td>$t$ with $\hat{\Sigma}_T^B$</td>
<td>2.740</td>
<td>3.764</td>
<td>4.771</td>
<td>6.090</td>
</tr>
<tr>
<td>$t$ with $\hat{\Sigma}_T^P$</td>
<td>2.840</td>
<td>4.228</td>
<td>5.671</td>
<td>8.112</td>
</tr>
<tr>
<td>$t$ with $\hat{\Sigma}_T^Q$</td>
<td>5.188</td>
<td>8.283</td>
<td>12.374</td>
<td>20.380</td>
</tr>
<tr>
<td>$t$ with $\hat{\Sigma}_T^Q$</td>
<td>4.822</td>
<td>7.111</td>
<td>11.573</td>
<td>19.180</td>
</tr>
</tbody>
</table>

Remark: KV (2002) shows that under certain assumptions the t-test with NW's SE without truncation are also asymptotically pivotal.

Newey-West estimator: KVB - Remarks

• An advantage of testing with KVB’s $C_T$ matrix is that its asymptotic distribution usually provides good approximation to its finite-sample counterpart. That is, the empirical size is close to the nominal size ($\alpha$).

• This is not the case for the NW HAC SE: in finite samples, they are downward biased. Tests are usually over-sized –i.e., not conservative.

• KV (2002b) show that, for $Q^*_{k}$ with the truncation lag equal to sample size, $T$, $Q^*_{k} T$ compares favorably with $Q^*_{QS} T$ in terms of power. This is in contrast with the result in HAC estimation, where the latter is usually preferred to other kernels.

Implications for OLS: Relative Efficiency

- We define relative efficiency of GLS against OLS as:

$$RE(\beta) = \left( \frac{(X' \Omega^{-1} X)^{1/2}}{(X' X)^{1/2} X' \Omega X (X' X)^{1/2}} \right)$$

- Let $$y_t = \beta x_t + \gamma y_{t-1} + \varepsilon_t, \quad \varepsilon_t = \rho \varepsilon_{t-1} + u_t$$,

Also, let $$x_t$$ also follow an AR(1) process: $$x_t = \theta x_{t-1} + \xi_t$$.

Then, when $$T$$ is large, it can be shown that

$$RE(\beta) = \frac{\text{Var}(\hat{\beta}_{GLS})}{\text{Var}(\hat{\beta}_{OLS})} \approx \frac{(1 - \rho \theta)(1 - \rho^2)}{(1 + \rho \theta)(1 + \rho^2 - 2 \rho \theta)}$$

- The relative efficiency can be very poor for large $$\rho$$ for any given $$\theta$$.

For example, suppose $$\rho=0.7$$, $$RE \approx 0.3423$$. Then, if the $$SE[\hat{\beta}_{GLS}]=1$$, $$SE[\hat{\beta}_{OLS}]=1.71$$ ($= 0.3423^{-1}$); that is, the OLS standard error is about 71% bigger than its GLS counterpart.

---

Implications for OLS: Relative Efficiency

$$RE(\beta) = \frac{\text{Var}(\hat{\beta}_{GLS})}{\text{Var}(\hat{\beta}_{OLS})} \approx \frac{(1 - \rho \theta)(1 - \rho^2)}{(1 + \rho \theta)(1 + \rho^2 - 2 \rho \theta)}$$

- The OLS estimators can be quite reasonable for a low degree of autocorrelation for any given $$\theta$$, for example, when $$\rho=0.3$$ and $$\theta=0.9$$, then $$RE \approx 0.9510$$.

- The inefficiency of OLS is difficult to generalize. We tend to see increase inefficiency with increasing values of the disturbance variances.

- In practice, it is worst in \textit{low frequency} -i.e., long period (year)- slowly evolving data. Can be extremely bad. GLS vs. OLS, the efficiency ratios can be 3 or more.

- Given the potential efficiency gain, it makes sense to test for autocorrelation.
Testing for Autocorrelation: LM tests

• There are several autocorrelation tests. Under the null hypothesis of no autocorrelation of order \( p \), we have \( H_0: \rho_1 = \ldots = \rho_p = 0 \).

Under \( H_0 \), we can use OLS residuals.

• Breusch–Godfrey (1978) LM test. Similar to the BP test:
  - Step 1. (Auxiliary Regression). Run the regression of \( e_t \) on all the explanatory variables, \( X \). In our example, \( e_t = X_t' \beta + \alpha_1 e_{t-1} + \ldots + \alpha_p e_{t-p} + v_t \)
  - Step 2. Keep the \( R^2 \) from this regression. Let’s call it \( R^2_e \). Then, calculate either
    (a) \( F = \frac{(R^2_e/p)}{[1-(1-R^2_e)/(T-(p+1))]}, \) which follows a \( F_{p,T-(p+1)} \) or
    (b) \( LM = T R^2_e \overset{d}{\rightarrow} \chi^2_p \).

Testing for Autocorrelation: Durbin-Watson

• The Durbin-Watson (1950) (DW) test for AR(1) autocorrelation: \( H_0: \rho_1 = 0 \) against \( H_1: \rho_1 \neq 0 \). Based on simple correlations of \( e \).

\[
    d = \frac{\sum_{i=2}^{T} (e_i - e_{i-1})^2}{\sum_{i=1}^{T} e_i^2}
\]

• It is easy to show that when \( T \rightarrow \infty \), \( d \approx 2(1 - \rho_1) \).

• \( \rho_1 \) is estimated by the sample correlation \( r \).

• Under \( H_0, \rho_1 = 0 \). Then, \( d \) should be distributed randomly around 2.

• Small values of \( d \) lead to rejection of \( H_0 \). The distribution depends on \( X \). Durbin-Watson derived bounds for the test.

• In the presence of lagged dependent variables, Durbin’s (1970) \( b \) test should be used: \( b = r \sqrt{T/(1-T\hat{\rho}^2)} \)
Testing for Autocorrelation: Portmanteau tests

• Box-Pierce (1970) test.
It tests \( H_0: \rho_1 = \ldots = \rho_p = 0 \) using the sample correlation \( r_j \)
\[
r_j = \frac{\sum_{t=1}^{T-j} e_t e_{t-j}}{\sum_{t=1}^{T} e_t^2}
\]
Then, under \( H_0 \)
\[
Q = T \sum_{j=1}^{p} r_j^2 \xrightarrow{d} \chi^2_p
\]

• Ljung-Box (1978) test.
A variation of the Box-Pierce test. It has a small sample correction.
\[
LB = T(T-2) \sum_{j=1}^{p} r_j^2 / (T-j)
\]

• The LB statistic is widely used. But, the Breusch–Godfrey (1978) LM test conditions on \( X \). Thus, it is more powerful.

Testing for Autocorrelation: Examples

• LM-AR Test for the 3 factor F-F model for IBM returns (p=4 lags):

```r
b <- solve(t(x)%*%x)%*%t(x)%*%y # OLS regression
e <- y - x%*%b
p_lag <- 4
e_lag <- matrix(0,T-p_lag,p_lag)
a <- 1
while (a<=p_lag) {
  za <- e[a:(T-p_lag+a-1)]
  e_lag <- e_lag + za
  a <- a+1
}
fit1 <- lm(e[(p_lag+1):T]~e_lag)
r2_e1 <- summary(fit1)$r.squared
lm_t <- (T-p_lag)*r2_e1
df <- ncol(e_lag)
1-pchisq(lm_t,df)
```
Testing for Autocorrelation: Examples

```r
> r2_e1 <- summary(fit1)$r.squared
> r2_e1
[1] 0.8245265
> (T-p_lag)
[1] 316
> lm_t <- (T-p_lag)*r2_e1
> lm_t
[1] 260.5504
> 1-pchisq(lm_t,df)
[1] 0

LM-AR(4) Test: 260.55  \Rightarrow \text{reject } H_0 \text{ at 5\% level (p-value < .00001).}
```

• Durbin-Watson test

```
DW <- sum((e[1:(T-1)]-e[2:T])^2)/RSS  # DW stat
> DW
[1] 0.2577663
⇒ Small DW statistic ⇒ Evidence for autocorrelation.
> 2*(1-cor(e[1:(T-1)],e[2:T]))  # approximate DW stat
[1] 0.2586147
```

Testing for Heteroscedasticity: Examples

• Q and LB tests with p=4 lags for IBM returns:

```
r_sum <- 0
lb_sum <- 0
a <- 1
while (a <= p_lag) {
    za <- as.numeric(t(e[(lag+1):T])%*%e[a:(T-p_lag+a-1)])
    r_sum <- r_sum + (za/RSS)^2  # sum cor(e[(lag+1):T],e[a:(T-p_lag+a-1)])^2
    lb_sum <- lb_sum + (za/RSS)^2/(T-a)
    a <- a + 1
}
Q <- T*r_sum
LB <- T*(T-2)*lb_sum

> Q
[1] 896.4554
> LB
[1] 897.9108
\Rightarrow \text{We strongly reject } H_0 \text{ at 5\% level (p-values < .000001).}
```
GLS: The AR(1) Model

- (A1) holds: \( y = X \beta + \varepsilon \)
  But, \( \varepsilon \) is no longer white noise: \( \varepsilon_t = \rho \varepsilon_{t-1} + u_t, \ |\rho| < 1 \).
  \( u_t \) is white noise error \( \sim D(0, \sigma_u^2) \)

Note: This characterizes the disturbances, not the regressors.

Notation: Let \( L \) be the lag operator, such that \( L_z = z_{t-1} \). Then,
\[
(1 - \rho L) \xi_t = u_t.
\]

- After some algebra, we get
\[
\xi_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3} + ... = \sum_{j=0}^{\infty} \rho^j u_{t-j} = \sum_{j=0}^{\infty} (\rho L)^j u_t \quad \text{(a moving average)}
\]

- \( \text{Var}[\xi_t] = \sum_{j=0}^{\infty} \rho^{2j} \text{Var}[u_{t-j}] = \sum_{j=0}^{\infty} \rho^{2j} \sigma_u^2 = \sigma_u^2 / (1 - \rho^2) \)

GLS: The AR(1) Model

- After some algebra, we get \( \Sigma = \sigma^2 \Omega \)

\[
\begin{pmatrix}
1 & \rho & \rho^2 & L & \rho^{T-1} \\
\rho & 1 & \rho & L & \rho^{T-2} \\
\rho^2 & \rho & 1 & L & \rho^{T-3} \\
M & M & M & O & M \\
\rho^{T-1} & \rho^{T-2} & \rho^{T-3} & L & 1
\end{pmatrix}
\]

(Note, trace \( \Omega = n \) as required.)
GLS: The AR(1) Model

• Then, if we want to estimate this model to gain efficiency, we can use the transformation matrix $P = \Omega^{1/2}$:

$$\Omega^{-1/2} = \begin{bmatrix} \sqrt{1 - \rho^2} & 0 & 0 & \ldots & 0 \\ -\rho & 1 & 0 & \ldots & 0 \\ 0 & -\rho & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & -\rho \end{bmatrix}$$

$$\Omega^{-1/2}y = \begin{bmatrix} (\sqrt{1 - \rho^2})y_1 \\ y_2 - \rho y_2 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{bmatrix}$$

• Problem: To use $\Omega$ we need $\rho$. We need to estimate it $\Rightarrow$ FGLS

FGLS: Eliminating Autocorrelation

• Let’s continue with the first-order autocorrelation:

$$y_t = X_t' \beta + \epsilon_t, \quad \epsilon_t = \rho \epsilon_{t-1} + u_t, \quad |\rho| < 1.$$  

Subtract

$$\rho y_{t-1} = \rho X_{t-1}' \beta + \rho \epsilon_{t-1}$$

Form pseudodifferences:

$$y^*_t = y_t - \rho y_{t-1} = (X_t - \rho X_{t-1})' \beta + \epsilon_t - \rho \epsilon_{t-1}$$

$$= X_t' \hat{\beta} + u_t$$

$\Rightarrow$ Now, $u_t$ is uncorrelated. We can use OLS to estimate $\beta$!

• But, we need to estimate $\rho$ (or know it, rare) to transform the model and do GLS. An iterative two-step algorithm seems natural:
  - First step: Estimate $\rho$
  - Second step: Estimate $\beta_{\text{GLS}}$ i.e., OLS in transformed model.
FGLS Estimation: Cochrane-Orcutt

- \( y_t - \rho y_{t-1} = (X_t - \rho X_{t-1})' \beta + \varepsilon_t - \rho \varepsilon_{t-1} \)
  \( \Rightarrow y_t = \rho y_{t-1} + X_t' \beta - X_{t-1}' \rho \beta + \varepsilon_t \)

- We have a linear model, but it is nonlinear in parameters. This is not a problem: Non-linear estimation is possible.

- Before today’s computer power, Cochrane–Orcutt’s (1949) iterative procedure was an ingenious way to do NLLS. Steps:
  1. Do OLS. Get residuals, \( \varepsilon_t \). Then estimate \( \rho \) with a regression of \( \varepsilon_t \) against \( \varepsilon_{t-1} \). We use \( r \) to denote the estimator of \( \rho \).
  2. FGLS Step. Using \( r \) transform the model to get \( y^* \) and \( X^* \). Do OLS \( \Rightarrow \) get \( \hat{b} \) to estimate \( \hat{\beta} \). Get residuals, \( \varepsilon^* \). Go back to (1).
  3. Iterate until convergence.

FGLS Estimation: Cochrane-Orcutt

Example: Cochrane-Orcutt in R

```r
# C.O. function requires Y, X (with constant), OLS b.
c.o.proc <- function(Y, X, b_0, tol) {
  T <- length(Y)
  e <- Y - X %*% b_0  # OLS residuals
  rss <- sum(e^2); rss1 <- rss; d_rss = rss
  e2 <- e[-1]; e3 <- e[-T]
  ols_e0 <- lm(e2 ~ e3 - 1)
  rho <- ols_e0$coeff[1]  # initial value for rho
  i <- 1
  while (d_rss > tol) {
    rss <- rss1  # RSS at iter (i-1)
    YY <- Y[2:T] - rho * Y[1:(T-1)]  # pseudo-diff Y
    ols_yx <- lm(YY ~ XX - 1)  # adjust if constant
    b <- ols_yx$coeff  # updated b at iter i
    # b[1] < b[1]/(1-rho)  # If constant not pseudo-diff
    e1 <- Y - X %*% b  # update residuals
    e2 <- e1[-1]
    e3 <- e1[-T]
    ols_e1 <- lm(e2 ~ e3 - 1)
    rho <- ols_e1$coeff[1]  # rho i
    rss_1 <- sum(e1^2)  # RSS i
    d_rss <- abs(rss_1 - rss)  # diff in RSS
    i <- i + 1
  }
  result <- list()
  result$Cochrane-Orc.Proc <- summary(ols_yx)
  result$rho.regression <- summary(ols_e1)
  result$Corrected.b_1 <- b[1]
  result$Iterations <- i - 1
  return(result)
}
```
Example: Cochrane-Orcutt for 3 FF factor model for IBM returns:

```r
> reg <- lm(y~x -1)
> reg
Call:
  lm(formula = y ~ x - 1)
Coefficients:
      x    xx1    xx2    xx3
  -0.2331471  0.0101872  0.0009803 -0.0044459
> reg2 <- c.o.proc(y,x,b,.0000001)
> reg2

$Cochrane.Orcutt.Proc
Call:
  lm(formula = YY ~ XX - 1)
Coefficients:
                Estimate  Std. Error   t value   Pr(>|t|)
XX        -0.222459    0.043201   -5.149   4.61e-07 ***
XXx1     0.009998    0.001016    9.841    < 2e-16 ***
XXx2    -0.002767    0.001334   -2.075     0.0388 *
XXx3    -0.003927    0.001565   -2.510     0.0126 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$rho.regression
Call:
  lm(formula = e2 ~ e3 - 1)
Residuals:
   Min       1Q   Median       3Q      Max
-0.30001 -0.05854 -0.00034  0.06156  0.34852
Coefficients:
                Estimate  Std. Error   t value   Pr(>|t|)
e3             0.87501     0.02711   32.27    <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09595 on 318 degrees of freedom
Multiple R-squared:  0.7661,    Adjusted R-squared:  0.7654
F-statistic:  1042 on 1 and 318 DF,  p-value: < 2.2e-16

$Iteractions
[1] 4
```
FGLS Estimation: Cochrane-Orcutt

- $SE[b_{CO}]$ and $SE[r_{CO}]$ are obtained from the regression in the last iteration. If the constant is not pseudo-differentiated, the estimated $b_{CO,0}$ has to be adjusted by $(1 - r_{CO})$. Similar correction for $SE[b_{CO,0}]$.

- If we do not want to lose the first observation, we can use the Prais-Winsten (1945) transformation of the first observation:
  
  \[
  \sqrt{1 - \rho^2} y_1 \quad \& \quad \sqrt{1 - \rho^2} X_1
  \]

- A grid search around $\rho$ can speed up the algorithm considerably. This is the Hildreth-Lu (1960) procedure.

- The iterative two-step estimation procedure can be easily extended to AR(p) models.

FGLS Estimation: Cochrane-Orcutt

**Note:** Cochrane-Orcutt works well if the specified AR(p) structure is correct. Otherwise, we are in the presence of a misspecified model.

**Example:** For the 3 FF factor model for IBM returns we run C-O with an AR(1) process for $\varepsilon_t$: $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$. Then, after the final run, we do an LM-AR(3) test on the residuals, $u_t$. We do this by adding in the C-O procedure (& add to the list the last line:

```r
result$LM.AR3.test_u <- lm_t_u)
## lm_t for AR(3) in u
ols_u <- lm(u[4:T] ~ u[1:(T-3)] + u[2:(T-2)] + u[3:(T-1)])
r2_u <- summary(ols_u)$r.squared
lm_t_u <- (T-1)*r2_u
```

$LM.AR3.test_u

| [1] | 56.29834 |

$\Rightarrow$ Very significant. We need to use a higher AR(p) model.
FGLS & MLE Estimation

- We need to estimate $\Omega \Rightarrow$ We need a model for $\Omega = \Omega(\theta)$.
  $\Rightarrow$ In the AR(1) model, we had $\Omega = \Omega(\rho)$.

- FGLS estimation is done using Cochrane-Orcutt or NLLS.
- MLE can also be done, say assuming a normal distribution for $u_t$, to estimate $\rho$ and $\beta$ simultaneously. For the AR(1) problem, the MLE algorithm works like the Cochrane-Orcutt algorithm.

- For an AR(2) model, Beach-Mackinnon (1978) propose an MLE algorithm that is very fast to converge.

- For an AR(p) models, with $p > 3$, MLE becomes complicated. Two-step estimation is usually done.

MLE Estimation: Example in R

- Log likelihood of ARMA(1,1)-GARCH(1,1) Model:

```r
log_lik_garch11 <- function(theta, data) {
  mu <- theta[1]; delta <- theta[2]; gamma <- theta[3]; alpha0 <- abs(theta[4]);
  alpha1 <- abs(theta[5]); beta1 <- abs(theta[6]); chk0 <- (1 - alpha1 - beta1)
  r <- ts(data)
  n <- length(r)
  u <- vector(length=n);
  u[1] <- r[1]- mu # set initial value for u[t] series
  for (t in 2:n)
    {u[t] = r[t]- mu - delta*r[t-1] - gamma*u[t-1]}
  h <- vector(length=n); h <- ts(h)
  h[1] = alpha0/chk0 # set initial value for h[t] series
  if  (chk0==0) {h[1]=.00001} #check to avoid dividing by 0
  for (t in 2:n)
    {h[t] = abs(alpha0 + alpha1*(u[t-1]^2)+ beta1*h[t-1])
     if  (h[t]==0) {h[t]=.00001} } #check to avoid log(0)
  return(-sum(-0.5*log(2*pi) - 0.5*log(abs(h[2:n])) - 0.5*(u[2:n]^2)/abs(h[2:n])))
}
```
MLE Estimation: Example in R

• To maximize the likelihood we use optim (mln can also be used):

```r
dat_xy <- read.csv("http://www.bauer.ub.edu/rsusmel/phd/datastream-K-DIS.csv",head=TRUE,sep=";")
summary(dat_xy)
names(dat_xy)

z <- dat_xy$SP500 # S&P 500 90-2016 monthly data
theta0 = c(0.01, -0.1, 0.01, -0.001, 0.2, 0.7) # initial values
ml_2 <- optim(theta0, log_lik_garch11, data=z, method="BFGS", hessian=TRUE)
ml_2$par # estimated parameters
I_Var_m2 <- ml_2$hessian
eigen(I_Var_m2) # check if Hessian is pd.
sqrt(diag(solve(I_Var_m2))) # parameters SE
```

Autocorrelation as a Common Factor

• From the first-order autocorrelated model

\[ y_t = \rho y_{t-1} + X_t' \beta - X_{t-1}' \rho \beta + u_t \] (*)

• We can generalize (*) using the lag operator L – i.e., \( L \cdot y_t = y_{t+1} \):

\[ (1- \rho L) y_t = (1- \rho L) X_t' \beta + u_t \]

Then, dividing by \((1- \rho L)\):

\[ y_t = X_t' \beta + u_t / (1- \rho L) = X_t' \beta + \epsilon_t \]

• We can think of a model with autocorrelation as a misspecified model. The common factor \((1- \rho L)\) is omitted. See Mizon (1977).

• We can generalize (*) even more by introducing more common lags:

\[ (1- B(L)) y_t = (1- B(L)) X_t' \beta + u_t \quad B(L): \text{function of } L, L^2, \ldots, L^q; \rho. \]
Common Factor Test

- From the AR(1) model:
  \[ y_t = \rho y_{t-1} + X_t' \beta - X_{t-1}' \rho \beta + u_t \]  

- We can think of (*) as a special case of a more general specification:
  \[ y_t = \lambda_1 y_{t-1} + X_t' \lambda_2 + X_{t-1}' \lambda_3 + u_t \]

Restrictions needed to get (*): \( \lambda_3 = -\lambda_1 \lambda_2 \)

- Hendry and Mizon (1980) propose testing the validity of the restrictions using a LR test, which have an asymptotic \( \chi^2 \) distribution, with degrees of freedom equal to the number of restrictions. 
  \[ LR = T \log \left[ \frac{RSS_R}{RSS_U} \right] \]

- The test is known as the common factor (COMFAC) test.

**Common Factor Test**

- We can use an \( F \)-test or Wald tests. See Mizon (1995) and McGuirk and Spanos (2004).

**Note:** Since the \( H_0 \) and \( H_1 \) models involve lagged \( y_t \)'s, the test statistics do not follow the asymptotic distribution. Bootstraps are a good idea.
Common Factor Test - Example

- Common Factor Test for 3 FF factor model for IBM returns:

(U) - We fit the unrestricted model: \( y_t = \lambda_1 y_{t-1} + X_t' \lambda_2 + X_{t-1}' \lambda_3 + u_t \)

```r
x <- cbind(x0,x1,x2,x3)
x_l <- cbind(x1,x2,x3)
sum(residuals(reg_u)^2)
```

> x <- cbind(x0,x1,x2,x3)
x_l <- cbind(x1,x2,x3)
> sum(residuals(reg_u)^2)

1 2.92264

(R) We fit the restricted model: \( y_t = \rho y_{t-1} + X_t \beta - X_{t-1} \beta + u_t \)

```r
cf_r <- optim(theta0, sum2,x=x, y=y, method="BFGS", hessian=TRUE)
cf_r$par
```

J 0.875011230 -0.027804863 0.009997961 -0.002767329

\[ T \log \left( \frac{2.92264}{3} \right) = 0.1861477 \]

\( \Rightarrow \) cannot reject \( H_0 \) at 5% level.

Note: The restricted model seems OK. But, we need to check that the model is well specified. In this case, does the AR(1) structure is enough to remove the autocorrelation in the errors?
Common Factor Test - Example

- We do an LM-AR(5) test to check the errors in the U Model:

```r
> fit_u <- lm(e_u[(5+1):T]~e_u[1:(T-5)]+e_u[2:(T-4)]+e_u[3:(T-3)]+e_u[4:(T-2)]+e_u[5:(T-1)])
> r2_e_u <- summary(fit_u)$r.squared
> lm_t_u <- (T-4)*r2_e_u
> lm_t_u
[1] 70.75767  \Rightarrow \text{Very significant (p-value: 1.6e-14). An AR(1) structure is not sufficient to remove AR in errors.}
```

In general, if we allow for more dynamics in the U Model we do better. For example, we use 4 lags in \( y_t \) and 2 lags in \( X_t \):

```r
> e_u4 <- residuals(reg_u4)
> fit_u5 <- lm(e_u4[(5+1):T]~e_u4[1:(T-5)]+e_u4[2:(T-4)]+e_u4[3:(T-3)]+e_u4[4:(T-2)]+e_u4[5:(T-1)])
> r2_e_u5 <- summary(fit_u5)$r.squared
> lm_t_u5 <- (T-5)*r2_e_u5
> lm_t_u5
[1] 6.938392  \Rightarrow \text{Not significant (p-value: .139).}
```

Building the Model

- Old (pre-LSE school) view: A feature of the data
  - “Account” for autocorrelation in the data.
  - Different models, different estimators

- Contemporary view: Why is there autocorrelation?
  - What is missing from the model?
  - Build in appropriate dynamic structures
  - Autocorrelation should be “built out” of the model
  - Use robust procedures (OLS with Newey-West SE) instead of elaborated models specifically constructed for the AR errors.