

**A Note on Robust Estimation of  
Repeat Sales Indexes with Serial Correlation  
in Asset Returns**

Kathryn Graddy  
Department of Economics  
and  
International Business School  
Brandeis University  
(kgraddy@brandeis.edu)

Jonathan Hamilton  
Department of Economics  
University of Florida  
(hamilton@ufl.edu)

September 30, 2009

We thank David Ling and Andra Ghent and seminar audiences at Baruch, McMaster and Waterloo for helpful comments. We also thank Chunrong Ai and Mark Watson for advice on the statistical model.

Keywords: repeat sales, heteroskedasticity, serial correlation  
JEL classifications: C13, C29, G12

## **Abstract**

This note studies the second stage of the Case-Shiller repeat sales method under the assumption of serial correlation in the deviations from the mean one-period returns on the underlying individual assets. We propose a flexible GLS methodology using dummy variables for each possible duration length in the second stage.

## 1. Introduction

The repeat sales methodology is an important technique to determine price trends and returns for idiosyncratic assets, including real estate, art, and antique musical instruments. Bailey, Muth, and Nourse [1963] first proposed the method, simply using ordinary least squares. Case and Shiller [1987] developed a three-stage generalized least squares (GLS) method. If deviations from the mean single-period returns for the underlying assets are independently and identically distributed, the variance of returns grows linearly when returns are summed over the holding period of an asset, which leads to heteroskedastic errors. To correct for this, one first estimates OLS regressions using dummy variables for time periods between sales. Then, the squared residuals are regressed against the length of the holding period. Estimates from the second stage provide weights for the third-stage GLS regressions. Our goal in this paper is to study the implications of non-i.i.d. errors for the second-stage regression and to suggest a second stage regression that is robust to a wide range of errors.

Section 2 details the Case-Shiller methodology and explores previous research. Section 3 explores different assumptions regarding the asset return errors. Section 4 applies our results to a repeat sales dataset of violin prices. Section 5 discusses some implications and concludes our analysis.

## 2. The Basic Case-Shiller Model (i.i.d. errors on individual returns)

Each observation consists of the purchase (buy) date,  $b_i$ , the purchase price,  $B_i$ , the sale date,  $s_i$ , and the sale price,  $S_i$ . Define the length of the holding period as

$\tau_i = s_i - b_i$ . Let  $y_i = \log\left(\frac{S_i}{B_i}\right)$  be the log of the compound return on property  $i$ . We can

write this as the sum of the returns to property  $i$  in each period between purchase and sale, or  $y_i = \sum_{t=b_i}^{s_i} r_{i,t}$  where  $r_{i,t} \equiv \log\left(\frac{P_{i,t}}{P_{i,t-1}}\right)$ , and  $P_{i,t}$  is the price of property  $i$  in period  $t$  (only observed for  $t = s_i$  and  $b_i$ ). The standard assumption is that  $r_{i,t} = \mu_t + \varepsilon_{it}$ , where  $\varepsilon_{it}$  is independent and identically normally distributed. Then,  $y_i = \sum_{t=b_i}^{s_i} \mu_t + \sum_{t=b_i}^{s_i} \varepsilon_{it}$ .<sup>1</sup>

Case and Shiller [1987] assumed that  $\log(P_{i,t}) = C_t + H_{i,t} + N_{i,t}$  where  $C_t$  is the value of the index in period  $t$ ,  $H_{i,t}$  is the value of a random walk process for property  $i$  at time  $t$ , and  $N_{i,t}$  is the “sale-specific random error”. This is equivalent to writing the price of property  $i$  in period  $T$  as  $P_{iT} = \exp\left(\sum_1^T \mu_t + \sum_1^T \varepsilon_{it} + v_{iT}\right)$  where  $v_{iT} \neq 0$  only if a

transaction occurs in period  $T$ . Taking logs and differencing prices from two different transactions, we obtain  $\ln(P_{iT}) - \ln(P_{iT-k}) = \sum_{T-k}^T \mu_t + \sum_{T-k}^T \varepsilon_{it} + v_{iT} - v_{i,T-k}$ . Then let

$\kappa_i = \sum_{T-k}^T \varepsilon_{it} + v_{iT} - v_{i,T-k}$  be the residual for property  $i$  in the first-stage regression. Hence,

$E(\kappa_i)^2 = E\left(\sum_1^{\tau} \varepsilon_{it}\right)^2 + 2\sigma_v^2$  where  $\sigma_v^2$  is the expectation of  $(v_{i,t})^2$ , under the assumption

that the  $v_{i,t}$  are i.i.d. Case and Shiller thus suggested first estimating an OLS repeat sales regression. Then, the squared residuals are regressed against the length of the holding period and a constant in the second stage regression. Estimates from the second stage provide weights for the third-stage GLS regressions. Below, we explore the theoretical

---

<sup>1</sup> In the above case, the variance of the error term grows linearly with the length of the holding period. Under this assumption, one can skip the three-stage procedure and simply use  $(s_i - b_i)^{-1}$  as the weights for GLS.

implication of dropping the assumption that the  $\varepsilon_{it}$  are i.i.d., and we propose a second stage regression that is robust to non-i.i.d. errors, using repeat sales data on fine violins as an example.

The Case and Shiller method, with variations, is widely used. Both OFHEO (Office of Federal Housing Enterprise Oversight) and S&P/ Case-Shiller house price indexes use variations of the Case-Shiller method. The OFHEO approach (see Calhoun [1996] for details) fits a quadratic equation—regressing the squared error on time between sales and time squared.<sup>2</sup> Calhoun [1996] states that, in practice, the constant term in the second-stage regression is often negative, which is inconsistent with the Case-Shiller explanation. Calhoun suggests forcing the constant to zero and re-estimating, which is OFHEO's approach. Case and Shiller directly estimate an arithmetic index but still use the standard Case-Shiller correction to correct for heteroskedasticity. Other papers that have focused on modifications of the Case and Shiller method include Quigley [1995], who fits the squared residuals to a quadratic function of elapsed time (without a constant), and Quigley and Hwang [2004], who model autoregression in the errors in price levels rather than returns.

### **3. Individual Asset Errors that Are Serially Correlated Across Periods**

We now drop the assumption that return errors are i.i.d. For Goetzmann's [1992] study of repeat sales regressions using stock market data, the i.i.d. assumption seems appropriate. In contrast, for many asset classes studied in repeat sales regressions, prices

---

<sup>2</sup> See also Abraham and Schauman [1991].

may not adjust quickly.<sup>3</sup> Houses, individual artworks and musical instruments have idiosyncratic features, making simple observations of prices of other assets in the class only signals of the “true price” of an asset. Trading costs are also significant (5-6% commissions plus transactions taxes and other costs for houses in the U.S. and a 10%-20% buyer’s commission plus a seller’s commission for art sold at auction), and short sales are essentially impossible. House price data are also only available with some lag (the interval between contract date and closing date at a minimum).

Note that the statistical issue is whether the error term on the individual asset returns is correlated between periods. In repeat sales data, only the residuals summed over several time periods are observed. This prevents us from uncovering much of the fine structure of the time series processes of the error returns.

In what follows, we shall drop the subscript  $i$  for the individual property since all calculations are with respect to a single property. Let the errors follow the general moving average process,  $\varepsilon_t = \sum_{i=0}^k \mu_i \eta_{t-i}$ , where  $k \in [1, \infty)$ ,  $\eta_t$  is white noise and  $\mu_0 = 1$ .<sup>4</sup> Then  $\kappa_\tau \equiv \sum_{t=1}^\tau \varepsilon_t = \sum_{i=0}^k \mu_i \sum_{t=1}^\tau \eta_{t-i} = \sum_{i=0}^k \mu_i \zeta_{\tau i}$  is the sum of return errors over  $\tau$  periods, and  $E(\kappa_\tau)^2 = \sum_{i=0}^k \mu_i^2 E(\zeta_{\tau i})^2 + 2 \sum_{i=0}^k \sum_{j>i}^k \mu_i \mu_j E(\zeta_{\tau i} \zeta_{\tau j})$ . If the process is stationary, then  $\sum_{i=0}^k \mu_i^2$  is finite. Thus, the first sum equals  $\tau \sigma_\eta^2 \sum_{i=0}^k \mu_i^2$ . Letting,  $s = j - i$ , the second sum equals  $2 \sum_{i=0}^k \sum_{s=1}^{\tau-1} (\tau - s) \mu_i \mu_{i+s} \sigma_\eta^2$ , which is also finite for

---

<sup>3</sup> Shiller [2007] discusses serial dependence in housing price aggregates. Even when repeat sales indexes incorporate a large number of properties, they combine data on diverse subgroups within the asset class (all single-family homes in a large metropolitan area). Since these submarkets may be quite thin and prices across the submarkets may not be closely linked, serial dependence in the errors also seems quite likely.

<sup>4</sup> Since AR and ARMA processes can be represented as infinite-order MA processes, the case where  $k = \infty$  includes them.

stationary processes.  $E(\kappa_\tau)^2$  can be written as a term which is a constant times  $\tau$  and a term which is a nonlinear function of  $\tau$  and  $\mu$ . For particular time series processes on the errors, we can be more specific.

### 3.1 Specific examples

#### *The MA Process*

Suppose that in the above general process,  $\mu_0 = 1$ ,  $\mu_1 = \theta$ , and  $\mu_i = 0$  for  $i > 2$ .

The errors then follow a first-order moving average (MA(1)) process,  $\varepsilon_t = \eta_t + \theta\eta_{t-1}$ ,

where  $-1 < \theta < 1$ . In this case by substitution,  $E\left(\sum_1^\tau \varepsilon_t\right)^2 = \tau(1+\theta^2)\sigma_\eta^2 + 2(\tau-1)\theta\sigma_\eta^2 =$

$\tau(1+\theta)^2\sigma_\eta^2 - 2\theta\sigma_\eta^2$ . Thus, regressing the square of the residual  $\left(\sum_1^\tau e_t\right)^2$  on  $\tau$  and a

constant yields  $\hat{\alpha} = -2\theta\sigma_\eta^2$  and  $\hat{\beta} = (1+\theta)^2\sigma_\eta^2$ . This provides a different explanation for

the constant term than Case and Shiller [1987]. Here,  $\hat{\alpha} < 0$  is not an anomaly, but arises

whenever  $\theta > 0$  (unlike first-order autoregressive processes, there is no presumption that

$\theta > 0$ ). Thus, a negative constant term may be evidence of a non-i.i.d. error process. If

we assume that the  $\varepsilon_t$  follow an MA(1) process without transaction errors, we can

identify point estimates of  $\theta$  and  $\sigma_\eta^2$  from  $\hat{\alpha}$  and  $\hat{\beta}$ .

We can extend this approach to higher-order MA processes. For the MA(2)

process,  $\varepsilon_t = \eta_t + \theta\eta_{t-1} + \gamma\eta_{t-2}$  (or for the general process,  $\mu_0 = 1$ ,  $\mu_1 = \theta$ ,  $\mu_2 = \gamma$  and

$\mu_i = 0$  for  $i > 2$ ), so by substitution,  $E\left(\sum_1^\tau \varepsilon_t\right)^2 =$

$\tau(1+\theta^2+\gamma^2+2\theta+2\theta\gamma+2\gamma)\sigma_\eta^2 - 2\sigma_\eta^2(\theta+\theta\gamma+\gamma)$ . Similar calculations reveal that all

MA(k) processes with  $k < \tau$  have an intercept term and a constant multiplying  $\tau$ , but no terms multiplying higher powers of  $\tau$ . The slope term will be positive, but the sign of the intercept depends on the parameters of the process. For  $k > 1$ , we cannot identify the parameters of the MA process since we observe only a slope and intercept.

### *The AR Process*

Suppose instead that  $\varepsilon_t$ ,  $t = 1, \tau$  follows an AR(1) process,  $\varepsilon_t = \eta_t + \rho\varepsilon_{t-1}$ . In the general MA process, this is equivalent to  $k = \infty$  and  $\mu_i = \rho^i$  for  $i = 0, k$ . Using the fact that

$$E[\varepsilon_t \varepsilon_{t-k}] = \rho^k \frac{\sigma_\eta^2}{1-\rho^2}, \text{ we find that } E\left(\sum_1^\tau \varepsilon_t\right)^2 = \tau\sigma_\varepsilon^2 + 2\sigma_\varepsilon^2 \left[ \tau \left( \frac{\rho - \rho^\tau}{1-\rho} \right) - \frac{\rho - \rho^\tau (\tau-1)}{1-\rho} - \rho \left[ \frac{\rho - \rho^{\tau-1}}{(1-\rho)^2} \right] \right]. \text{ See Appendix A for a derivation.}$$

As  $\tau$  grows, for  $\rho > 0$ , this expression increases at an increasing rate and asymptotes to an increasing straight line. For  $\rho < 0$ , it increases at a decreasing rate. Thus, only negative first-order serial correlation is consistent with a positive coefficient on time between sales and a negative coefficient on its square, which is a common finding. Since negative autocorrelation is not common in economic data, it seems unlikely that an AR(1) error process explains the commonly observed pattern. Higher-order AR processes also

result in  $E\left(\sum_1^\tau \varepsilon_t\right)^2$  being a nonlinear function of the holding period.

### *The ARMA Process*

An ARMA(p, q) process has  $p^{\text{th}}$ -order autoregression and  $q^{\text{th}}$ -order moving average. For  $p = q = 1$ , we can write the process as  $\varepsilon_t = \phi\varepsilon_{t-1} + \eta_t + \theta\eta_{t-1}$ . In the general

MA process, this is equivalent to  $k = \infty$ ,  $\mu_0 = 1$ ,  $\mu_1 = \phi + \theta$ , and  $\mu_i = \phi^i + \phi^{i-1}\theta$  for  $i=2, k$ .

In this case,

$$E\left(\sum_{t=1}^{\tau} \varepsilon_t\right)^2 = \tau\sigma_{\eta}^2 \left[ \left( \frac{1+\theta^2+2\phi\theta}{1-\phi^2} \right) + 2 \left( \frac{\theta+\phi+\theta^2\phi+\theta\phi^2}{1-\phi^2} \right) \right] - 2\sigma_{\eta}^2 \left( \frac{\theta+\phi+\theta^2\phi+\theta\phi^2}{1-\phi^2} \right) + 2\sigma_{\eta}^2 \frac{(\phi^{\tau} - (\tau-1)\phi^2 + (\tau-2)\phi)}{(1-\phi)^2} \left( \frac{\theta+\phi+\theta^2\phi+\theta\phi^2}{1-\phi^2} \right).$$

See Appendix B for the derivation. As with the MA process, there is an intercept (which is easily negative) and a constant coefficient on the time horizon. As with the AR process, there is also a term which decays exponentially.

For  $\phi > 0$  (the “normal” case), the expectation of the square of the sum of the residuals increases with the holding period length. For  $\phi > |\theta|$ , it increases as an increasing rate, while for  $0 < \phi < |\theta|$ , it increases as at decreasing rate. This last possibility is consistent with a positive coefficient on the linear term and a negative coefficient on the quadratic term. It also seems to be the “minimal” assumption on the return error process to generate such concavity. As with AR processes, higher-order ARMA processes result in  $E\left(\sum_{t=1}^{\tau} \varepsilon_t\right)^2$  being a nonlinear function of the holding period.

One could fit a nonlinear regression in  $\tau$  to  $E\left(\sum_1^{\tau} \varepsilon_t\right)^2$ . As with the MA(1) process, one could only identify parameters of the stochastic process conditional on the assumption about the order of the ARMA process, but of course the order of the ARMA process cannot be recovered because we do not observe the errors in the individual asset returns, but only the summed residuals.

### 3.2 Flexible GLS

In order to correct for heteroskedasticity when the error term on the individual asset returns is correlated between periods, we propose a flexible approach in which 3<sup>rd</sup> stage weights are constructed by regressing the squared residuals from the first stage regression on dummy variables that represent the length of the holding period for each asset. This is a simple and useful approach that allows for autocorrelation even when the exact form of the correlation in the underlying assets cannot be identified.

## 4. An Application to Repeat Sales of Violins

Graddy and Margolis [2009] study returns to owning high-quality violins over a long time period dating back into the 19<sup>th</sup> Century. The data consists of 337 repeat sales of fine violins that took place between 1849 and 2009. The average holding period for each violin is 32 years. The shortest holding period is 5 years and the longest is 147 years.

Columns 1 and 2 of Table 1 report the coefficients on the OLS (first stage) of the repeat sales regressions, columns 3 and 4 report the coefficients using the Case and Shiller method for the 3<sup>rd</sup> stage regressions, and columns 5 and 6 report the coefficients using the flexible GLS estimator described above.<sup>5</sup> In Table 1 we also present the test results from the Koenker-Basset test for heteroskedasticity. In this test, the squared residuals from the regression model ( $\hat{u}_i^2$ ) are regressed on the squared estimated predicted values of the dependent variable ( $\hat{Y}_i^2$ ) and a constant:  $\hat{u}_i^2 = \alpha_1 + \alpha_2(\hat{Y}_i^2) + v_i$ .

---

<sup>5</sup> The actual returns in the OLS and standard Case and Shiller regressions are calculated and reported in Graddy and Margolis (2009).

The null hypothesis is that  $\alpha_2 = 0$ . If this is not rejected, then one could conclude that there is no heteroskedasticity. We also report the mean of the standard errors from these repeat sales regressions.

The results indicate that the null hypothesis for heteroskedasticity is rejected in both the OLS and the standard Case and Shiller regressions. As the standard Case and Shiller regression does not completely correct for heteroskedastic errors, non-i.i.d. errors are suspected. Only in the flexible GLS regression can we conclude that there is no remaining heteroskedasticity. Furthermore, the mean standard errors are lower with flexible GLS than in either of the other specifications.

To explore for evidence of non-i.i.d. errors, in Table 2 we report the results from various specifications of the second stage regressions. Column 1 reports the regressions from the standard Case and Shiller second stage, and column 5 reports the flexible GLS regression. We also consider other polynomials (with and without constants) and a logarithmic specification. Note that the standard Case and Shiller regressions appear to be dominated by the log specification using the measures of adjusted R-squared, AIC, and BIC. The flexible specification dominates all specifications, as indicated by adjusted R-squared, AIC, and BIC. As indicated in Table 1, this specification both corrects for heteroskedasticity and decreases the errors. In Graddy, Hamilton, and Campbell [2009], we test the flexible specification on two larger repeat sales datasets of historical house prices in the Herengracht district of Amsterdam, and on prices of art sold in Amsterdam, with very similar results.<sup>6</sup>

---

<sup>6</sup> If the errors in asset returns are i.i.d., the variance of the return errors for an individual property should grow linearly with time. The size of the Herengracht and Amsterdam art datasets allowed us to estimate 1 year, 2 year, 5 year, and 10 year returns. In the Amsterdam art dataset, we could clearly reject linear growth in returns when estimating the different return periods.

## 5. Implications and Conclusions

It is well-known that the logarithmic specification of the dependent variable results in the geometric mean across assets for each time period of the index. Goetzmann [1992] suggested that the coefficient on the time between sales should be used as an estimate of the cross section variance to give the following formula for the arithmetic mean,  $\mu^a \cong \exp\left(\mu^g + \frac{\sigma^2}{2}\right) - 1$ , where  $\mu^a$  and  $\mu^g$  are the arithmetic and geometric means and  $\sigma^2$  is the cross-section variance. This correction becomes problematic for non-i.i.d. errors.

Without a specific assumption on the errors in individual asset returns, the single period return variance in an asset cannot be identified from the second stage of the Case-Shiller regression results.<sup>7</sup> Calhoun [1996] proposes using  $\sigma_t^2 = At + Bt^2$  (where A and B are the linear and quadratic coefficients from the second stage—with no constant) as the variance in the geometric to arithmetic correction formula (in index form). There is a problem once the second stage includes more than a simple linear term—the estimated variance for any property becomes a function of the holding period. Even using the variance per period ( $A + Bt$ ) depends on the holding period. Any estimate of the arithmetic return depends on the planned holding period if the return errors are not i.i.d.

In further work, we plan to study two issues. One is the impact of serial dependence on the magnitude of standard errors. The other is how serial dependence affects the variance in revisions of coefficient estimates after re-estimation with additional periods of data.

---

<sup>7</sup> The S&P/Case-Shiller<sup>®</sup> price index directly estimates an arithmetic index to circumvent this problem.

## References

- Abraham, J., and W. Schauman, 1991, New Evidence on Home Prices from Freddie Mac Repeat Sales, *AREUEA Journal* 19: 333-352.
- Bailey, M., R. Muth, and H. Nourse, 1963, A Regression Method for Real Estate Price Index Construction, *Journal of the American Statistical Association* 58: 933-942.
- Calhoun, C., 1996, OFHEO House Price Indexes: HPI Technical Description, mimeo, Office of Federal Housing Enterprise Oversight.
- Case, K., and R. Shiller, 1987, Prices of Single Family Homes Since 1970: New Indexes for Four Cities, Cowles Foundation Discussion Paper No. 851, Yale University.
- Goetzmann, W., 1992, The Accuracy of Real Estate Indices: Repeat Sale Estimators, *Journal of Real Estate Finance and Economics* 5:5-53.
- Graddy, K., J. Hamilton, and R. Campbell, 2009, Repeat Sales Indexes: Estimation Without Assuming that Errors in Asset Returns Are Independently Distributed, CEPR Discussion Paper No. 7344.
- Graddy, K., and P. Margolis, 2009, Fiddling with Value: Violins as an Investment?, forthcoming, *Economic Inquiry*.
- Hwang, M., and J. Quigley, 2004, Selectivity, Quality Adjustment and Mean Reversion in the Measurement of House Values, *Journal of Real Estate Finance and Economics* 28: 161-178.
- Quigley, J., 1995, A Simple Hybrid Model for Estimating Real Estate Price Indexes, *Journal of Housing Economics* 4: 1-12.
- Shiller, R., 2007, Historic Turning Points in Real Estate, Cowles Foundation Discussion Paper No. 1610, Yale University (<http://ssrn.com/abstract=991107>).

Table 1  
Repeat Sales Regressions

period	<u>OLS</u>		<u>Case and Shiller</u>		<u>Flexible GLS</u>	
	coef	std error	coef	std error	coef	std error
1860	0.186	0.328	0.377	0.332	0.187	0.316
1870	0.237	0.252	0.204	0.247	0.299	0.221
1880	0.725	0.232	0.750	0.221	0.797	0.197
1890	0.347	0.211	0.325	0.201	0.247	0.182
1900	0.630	0.247	0.598	0.241	0.642	0.230
1910	0.236	0.228	0.265	0.219	0.217	0.203
1920	0.577	0.212	0.601	0.204	0.623	0.193
1930	1.103	0.213	1.065	0.204	0.993	0.190
1940	-0.347	0.204	-0.281	0.198	-0.166	0.191
1950	0.012	0.264	0.020	0.262	0.024	0.258
1960	0.708	0.249	0.651	0.246	0.504	0.233
1970	1.025	0.209	1.075	0.204	1.166	0.189
1980	1.783	0.153	1.777	0.145	1.698	0.133
1990	1.335	0.137	1.289	0.128	1.275	0.118
2000	0.353	0.138	0.361	0.128	0.388	0.120
2007	0.108	0.157	0.125	0.150	0.073	0.144
Koenker Basset $\alpha_2$	0.010	0.004	0.012	0.005	0.001	0.003
Average std error		0.215		0.208		0.195
Obs		337		337		337

Table 2  
Second Stage Regression Results

	1		2		3		4		5		6		7	
	coef	std error	coef	std error	coef	std error	coef	std error	coeff	std error	coef	std error	coef	std error
TBS	0.035	0.016	0.113	0.046	0.355	0.108					0.085	0.011	0.182	0.025
TBS <sup>2</sup>			-0.007	0.004	-0.059	0.021							-0.012	0.003
TBS <sup>3</sup>					0.003	0.001								
ln(TBS)							0.153	0.051						
Duration Dummies									13					
Cons	0.282	0.067	0.165	0.094	-0.067	0.132	0.271	0.060	0.295	0.478				
F-Stat	5.05		4.1		4.84		9.16		8.74		*		*	
Prob>F	0.025		0.017		0.003		0.003		0.000		*		*	
Adj R <sup>2</sup>	0.012		0.018		0.033		0.024		0.074		*		*	
AIC	810		809		805		806		706		826		810	
BIC	818		821		820		814		760		830		818	
Obs	337		337		337		337		337		337		337	

## Appendix A: Calculations for the AR Process

Using the fact that  $E[\varepsilon_t \varepsilon_{t-k}] = \rho^k \frac{\sigma_\varepsilon^2}{1-\rho^2}$ , we find that

$$\left( \sum_1^\tau \varepsilon_t \right)^2 = \sum_1^\tau (\varepsilon_t)^2 + 2 \sum_1^{\tau-1} \varepsilon_{t+1} \varepsilon_t + 2 \sum_1^{\tau-2} \varepsilon_{t+2} \varepsilon_t + \dots + 2 \sum_1^{\tau-(\tau-1)} \varepsilon_{t+\tau-1} \varepsilon_t.$$

$$\begin{aligned} \text{Thus, } E \left( \sum_1^\tau \varepsilon_t \right)^2 &= \tau \sigma_\varepsilon^2 + 2\rho(\tau-1)\sigma_\varepsilon^2 + 2\rho^2(\tau-2)\sigma_\varepsilon^2 + \dots + 2\rho^{\tau-1}\sigma_\varepsilon^2 \\ &= \tau \sigma_\varepsilon^2 + 2\sigma_\varepsilon^2 \left[ \rho(\tau-1) + 2\rho^2(\tau-2) + \dots + 2\rho^{\tau-1} \right] \end{aligned}$$

$$\text{where } \left[ \rho(\tau-1) + 2\rho^2(\tau-2) + \dots + \rho^{\tau-1} \right] = \sum_1^{\tau-1} \rho^k (\tau-k) = \tau \sum_1^{\tau-1} \rho^k - \sum_1^{\tau-1} \rho^k k.$$

The first term in this expression equals  $\tau \left( \frac{\rho - \rho^\tau}{1-\rho} \right)$ , using

$$Z = \rho + \rho^2 + \dots + \rho^{\tau-1} \text{ and } \rho Z = \rho^2 + \dots + \rho^\tau,$$

$$\text{while the second term equals } - \left[ \frac{\rho - \rho^\tau (\tau-1) + \rho \left[ \frac{\rho - \rho^{\tau-1}}{1-\rho} \right]}{1-\rho} \right], \text{ using}$$

$$Y = \rho + 2\rho^2 + (\tau-1)\rho^{\tau-1} \text{ and } \rho Y = \rho^2 + 2\rho^3 + (\tau-1)\rho^\tau + \tau\rho^\tau, \text{ and}$$

$$Y - \rho Y = (\rho - \rho^\tau (\tau-1)) + (\rho^2 + \rho^3 + \dots + \rho^{\tau-1})$$

$$= (\rho - \rho^\tau (\tau-1)) + \rho \left( \frac{\rho - \rho^{\tau-1}}{1-\rho} \right).$$

$$\text{Hence, } E \left( \sum_1^\tau \varepsilon_t \right)^2 = \tau \sigma_\varepsilon^2 + 2\sigma_\varepsilon^2 \left[ \tau \left( \frac{\rho - \rho^\tau}{1-\rho} \right) - \frac{\rho - \rho^\tau (\tau-1)}{1-\rho} - \rho \left[ \frac{\rho - \rho^{\tau-1}}{(1-\rho)^2} \right] \right].$$

## Appendix B: Calculations for the ARMA process

For  $p = q = 1$ , an ARMA(1, 1) process can be written as:

$$\varepsilon_t = \phi\varepsilon_{t-1} + \eta_t + \theta\eta_{t-1}, t = 1, \tau.$$

The expected values of the variances and covariances of errors equal:

$$E(\varepsilon_t^2) = \sigma_\eta^2 \left[ \frac{1 + \theta^2 + 2\phi\theta}{1 - \phi^2} \right], \quad E(\varepsilon_t \varepsilon_{t-1}) = \phi\sigma_\varepsilon^2 + \theta\sigma_\eta^2 = \frac{\theta + \phi + \theta^2\phi + \theta\phi^2}{1 - \phi^2} \sigma_\eta^2,$$

$$\text{and } E(\varepsilon_t \varepsilon_{t-k}) = \phi^k \hat{\beta} \text{ where } \hat{\beta} = \frac{\sigma_\eta^2}{\phi} \frac{\theta + \phi + \theta^2\phi + \theta\phi^2}{1 - \phi^2} \text{ for } k \geq 2.$$

Then, we can write the expected value of the square of the sum of the residuals as:

$$\begin{aligned} E\left(\sum_{t=1}^{\tau} \varepsilon_t\right)^2 &= E\left(\sum_{t=1}^{\tau} \varepsilon_t^2\right) + 2E\left(\sum_{t=1}^{\tau-1} \varepsilon_t \varepsilon_{t+1}\right) + 2E\left(\sum_{t=1}^{\tau-2} \varepsilon_t \varepsilon_{t+2}\right) + \dots + 2E\left(\sum_{t=1}^{\tau-(\tau-1)} \varepsilon_t \varepsilon_{t+(\tau-1)}\right) \\ &= E\left(\sum_{t=1}^{\tau} \varepsilon_t^2\right) + 2E\left(\sum_{t=1}^{\tau-1} \varepsilon_t \varepsilon_{t+1}\right) + 2\left\{\sum_{J=1}^{\tau-2} JE\left(\varepsilon_t \varepsilon_{t+(J)}\right)\right\}. \end{aligned}$$

Taking the expectations, we obtain:

$$E\left(\sum_{t=1}^{\tau} \varepsilon_t\right)^2 = \tau\sigma_\varepsilon^2 + 2(\tau-1)\frac{\theta + \phi + \theta^2\phi + \theta\phi^2}{1 - \phi^2} \sigma_\eta^2 + 2\hat{\beta}\phi^\tau \sum_{J=1}^{\tau-2} J\phi^{-J}.$$

For the last term, let  $K = \tau - 2$  and  $\omega = \frac{1}{\phi}$ . Then let  $Z = \sum_{J=1}^K J\omega^J$ .

Now  $Z = \omega + 2\omega^2 + 3\omega^3 + \dots + K\omega^K$  and  $\omega Z = \omega^2 + 2\omega^3 + \dots + K\omega^{K+1}$ , so

$$Z - \omega Z = \omega + \omega^2 + \omega^3 + \dots + \omega^K - K\omega^{K+1}.$$

Let  $Y = \omega + \omega^2 + \omega^3 + \dots + \omega^K$ . Then  $\omega Y = \omega^2 + \omega^3 + \dots + \omega^{K+1}$ , so  $Y - \omega Y = \omega - \omega^{K+1}$ , and

$$Y = \frac{\omega - \omega^{K+1}}{1 - \omega}. \text{ Substituting this into the earlier formula,}$$

$$Z - \omega Z = \frac{\omega - \omega^{K+1}}{1 - \omega} - K\omega^{K+1} = \frac{\omega - (K+1)\omega^{K+1} + K\omega^{K+2}}{1 - \omega}.$$

Hence,  $Z = \frac{\omega - (K+1)\omega^{K+1} + K\omega^{K+2}}{(1 - \omega)^2}$

Thus, we have  $E\left(\sum_{t=1}^{\tau} \varepsilon_t\right)^2 = \tau\sigma_\varepsilon^2 + 2(\tau-1)\left(\frac{\theta + \phi + \theta^2\phi + \theta\phi^2}{1 - \phi^2}\right)\sigma_\eta^2 + 2\hat{\beta}\phi^\tau\sigma_\eta^2.$

Substituting,  $E\left(\sum_{t=1}^{\tau} \varepsilon_t\right)^2 = \tau\left(\frac{1 + \theta^2 + 2\phi\theta}{1 - \phi^2}\right)\sigma_\eta^2 + 2(\tau-1)\left(\frac{\theta + \phi + \theta^2\phi + \theta\phi^2}{1 - \phi^2}\right)\sigma_\eta^2$

$$+ 2\sigma_\eta^2 \frac{(\phi^\tau - (\tau-1)\phi^2 + (\tau-2)\phi)}{(1-\phi)^2} \left(\frac{\theta + \phi + \theta^2\phi + \theta\phi^2}{1 - \phi^2}\right)$$

$$= \tau\sigma_\eta^2 \left[ \left(\frac{1 + \theta^2 + 2\phi\theta}{1 - \phi^2}\right) + 2\left(\frac{\theta + \phi + \theta^2\phi + \theta\phi^2}{1 - \phi^2}\right) \right] - 2\sigma_\eta^2 \left(\frac{\theta + \phi + \theta^2\phi + \theta\phi^2}{1 - \phi^2}\right)$$

$$+ 2\sigma_\eta^2 \frac{(\phi^\tau - (\tau-1)\phi^2 + (\tau-2)\phi)}{(1-\phi)^2} \left(\frac{\theta + \phi + \theta^2\phi + \theta\phi^2}{1 - \phi^2}\right).$$