Financial Econometrics Econ 40357
Regression review, Time-series regression
Some Necessary Matrix Algebra (sorry, can't avoid this)

N.C. Mark

University of Notre Dame and NBER

2020

Regression review

A time series is a sequence of observations over time. Let T be sample size. We write the sequence as

$$\{y_t\}_{t=1}^{T}$$

We use notation such as this

$$\begin{aligned} \mu_y &= \mathrm{E}(y_t) \\ \sigma_y^2 &= \mathrm{Var}(y_t) \end{aligned}$$

Regression in population

- We have in mind a joint distribution between two time series, y_t and x_t. This is our model.
- In finance, we are less concerned about exogeneity, instrumental variables, and establishing cause and effect.
- We are more concerned about understanding reduced form correlations. Understanding the statistical **dependence** across time series, and **dependence** of observations across time.
- The cross-moments of the joint distribution.

Regression in population

- Write the population regression as

$$y_t = \overbrace{\alpha + \beta x_t}^{E(y_t|x_t)} + \epsilon_t \quad (1)$$

  - The systematic part of **regression** is also called the **projection**. ϵ_t is projection **error**.
  - Assume error is iid but **not necessarily normal** (what does this mean?).

$$\epsilon_t \text{ is i.i.d. } \left(0, \sigma_\epsilon^2\right)$$

  - Think of fitted part of regression as **conditional expectation**.
  - Conditional expectation is the best predictor.
  - **Prediction** means the same thing as **forecast**
- We use regression for things like computing **betas**, which measure **exposure** of an asset to **risk factors**.

Regression in population

1. Take expectation of $y_t = \alpha + \beta x_t + \epsilon_t$

   $$E(y_t) = \alpha + \beta E(x_t) + E(\epsilon_t)$$

   Using the short-hand notation,

   $$\mu_y = \alpha + \beta \mu_x$$

   rearrange to get,

   $$\alpha = \mu_y - \beta \mu_x$$

2. Define $\tilde{y}_t \equiv y_t - \mu_y$ and $\tilde{x}_t \equiv x_t - \mu_x$. The 'tilde' represents the variable expressed as a deviation from its mean. Substitute this expression for α back into the regression (eq. 1). Doing so gives the regression in **deviations** from mean form.

   $$\tilde{y}_t = \beta \tilde{x}_t + \epsilon_t \tag{2}$$

Regression in population

- Multiply both sides of eq.(2) by $\tilde{x}_t$, then take expectations on both sides,

$$\begin{aligned} \tilde{y}_t\tilde{x}_t &= \beta\tilde{x}_t\tilde{x}_t + \epsilon_t\tilde{x}_t \\ E(\tilde{y}_t\tilde{x}_t) &= \beta E(\tilde{x}_t\tilde{x}_t) + E(\epsilon_t\tilde{x}_t) \end{aligned}$$

Solve for β

$$\beta = \frac{E(\tilde{y}_t\tilde{x}_t)}{E(\tilde{x}_t)^2} = \underbrace{\frac{\mathrm{Cov}(y_t, x_t)}{\mathrm{Var}(x_t)}}_{\text{Interpretation}} = \underbrace{\frac{\sigma_{y,x}}{\sigma_x^2}}_{\text{Notation}} = \underbrace{\frac{\sigma_{y,x}}{\sigma_y\sigma_x}\frac{\sigma_y}{\sigma_x}}_{\text{Algebra}} = \underbrace{\rho_{y,x}\frac{\sigma_y}{\sigma_x}}_{\text{Interpretation}} \qquad (3)$$

Regression in population

- Take conditional expectation on both sides of original regression conditional on x_t.

$$E(y_t|x_t) = \alpha + \beta x_t$$

  There's a theorem that says the conditional expectation is the **best** linear predictor of y_t conditional on x_t. The **best!**
- Since regression is conditional expectation, this motivates using it as the **forecast function.**

Estimation of α and β by least squares

- Eviews will do this work for you.
- Least squares estimates are the **sample** counterparts to the **population** parameters. What does this mean?
- In deviations from the mean form, solution to least squares problem is

$$\hat{\beta} = \frac{\sum_{t=1}^{T} \tilde{x}_t \tilde{y}_t}{\sum_{t=1}^{T} \tilde{x}_t^2} = \frac{\frac{1}{T}\sum_{t=1}^{T} \tilde{x}_t \tilde{y}_t}{\frac{1}{T}\sum_{t=1}^{T} \tilde{x}_t^2}$$

$$\tilde{y}_t = \hat{\beta}\tilde{x}_t + \hat{\epsilon}_t$$

$\hat{\epsilon}_t$ is **residual**, not error. $\hat{\beta}$ is a **random variable.** To see this, make substitutions,

$$\hat{\beta} = \frac{\sum \tilde{x}_t \left(\beta \tilde{x}_t + \epsilon_t\right)}{\sum \tilde{x}_t^2} = \beta + \frac{\sum \tilde{x}_t \epsilon_t}{\sum \tilde{x}_t^2}$$

$\hat{\beta}$ is a linear combination of $\epsilon_t' s$, which are random variables. Therefore $\hat{\beta}$ is a random variable, with a distribution.

Inference in Time Series Regression is a bit Different

- What is statistical inference?
- We want the **sample** standard deviation of $\hat{\beta}$. It is called the **standard error** of $\hat{\beta}$, because $\hat{\beta}$ is a statistic. What is a statistic?
- $\hat{\beta}$ divided by standard error is the **t-ratio**.

  We are mainly interested in t-ratios. Not interested in F-statistics. Nobody's opinion ever was changed by having a significant F-statistic and insignifcant t-ratios.
- What is statistical significance?
- In your previous **econometrics** class, you assumed x_t is **exogenous**. This alllows you to treat them as constants, and therefore, randomness in $\hat{\beta}$ is induced only by ϵ_t.
- In time-series, we can't do that (If y_t is Amazon returns and x_t is the market return, how can we say the market is exogenous? or if $x_t = y_{t-1}$, how can we treat x_t as constant? No way!).

Inference in Time Series Regression

Let's pretend the $\tilde{x}_t$ are exogenous. Even more inappropriately, let's pretend they are non-stochastic constants.

$$\begin{aligned} \text{Var}\left(\frac{\sum_{t=1}^{T}\tilde{x}_t\epsilon_t}{\sum_{t=1}^{T}\tilde{x}_t^2}\right) &= \frac{1}{\left(\sum_{t=1}^{T}\tilde{x}_t^2\right)^2}\sum_{t=1}^{T}\left(\tilde{x}_1^2\sigma_\epsilon^2+\tilde{x}_2^2\sigma_\epsilon^2+\cdots\tilde{x}_T^2\sigma_\epsilon^2\right) \\ &= \frac{\sum\tilde{x}_t^2}{\left(\sum_{t=1}^{T}\tilde{x}_t^2\right)^2}\sigma_\epsilon^2=\frac{\sigma_\epsilon^2}{\sum\tilde{x}_t^2} \end{aligned}$$

The standard deviation of the term is

$$\text{sd}\left(\frac{\sum_{t=1}^{T}\tilde{x}_t\epsilon_t}{\sum_{t=1}^{T}\tilde{x}_t^2}\right)=\frac{\sigma_\epsilon}{\sqrt{\sum\tilde{x}_t^2}}$$

The standard error of the term, and hence of $\hat{\beta}$ is

$$\text{se}\left(\hat{\beta}\right)=\frac{\hat{\sigma}_\epsilon}{\sqrt{\sum\tilde{x}_t^2}}$$

where we estimate $\hat{\sigma}_\epsilon$ with the sample standard deviation of the regression residuals $\hat{\epsilon}_t$. This particular formula is true only when the errors are *iid*, and for large (infinite) sample sizes. Why do we use? Because we can find the answer for large samples, and we hope that it is a good approximation to the exact true (but unknown) distribution

Inference in Time Series Regression

- So **time-series econometricians** do a thing called **asymptotic** theory. They ask how the numerator and denominator,

$$\text{numerator: } \frac{1}{T}\sum_{t=1}^{T} \tilde{x}_t \epsilon_t \rightarrow N\left(0, \sigma_\epsilon^2 Q\right)$$

$$\text{denominator:} \frac{1}{T}\sum_{t=1}^{T} \tilde{x}_t^2 \rightarrow Q$$

  behave as $T \rightarrow \infty$. It is very complicated business involves a lot of high-level math. Fortunately, at the end of the day, what comes out of all this is the same thing we learned in your first econometrics class!

Least squares estimation of α and β

- That is, we pretend we have an infinite sample size ($T = \infty$,) in which case,

$$t = \frac{(\hat{\beta} - \beta)}{s.e.\left(\hat{\beta}\right)} \sim N(0, 1)$$

$$se\left(\hat{\beta}\right) = \frac{\hat{\sigma}_\epsilon}{\sqrt{\sum \tilde{x}_t^2}} \tag{4}$$

$$\hat{\sigma}_\epsilon^2 = \frac{1}{T} \sum \hat{\epsilon}_t^2 \tag{5}$$

  The difference is we don't consult the t-table or worry about degrees of freedom. We consult the standard normal table.

- **The strategy:** The exact t-distribution is unknown (why?), so we use the asymptotic distribution (why?) and hope it is a good approximation to the unknown distribution.
- Finally, we are also interested in R^2, the measure of goodness of fit.

$$R^2 = \frac{SSR}{SST} = \frac{\sum \tilde{y}_t^2}{\sum \tilde{x}_t^2} = 1 - \frac{\sum \hat{\epsilon}_t^2}{\sum \tilde{x}_t^2} = 1 - \frac{SSE}{SST}$$

Story behind the t-test

In the 1890s, **William Gosset** was studying chemical properties of barley with small samples, for the Guinness company (yes, that Guinness). He showed his results to the great statistician Karl Pearson at University College London, who mentored him.

Gossett published his work in the journal *Biometrika*, using the pseudonym Student, because he would have gotten in trouble at Guinness if he used his real name.

t-test review: two sided test

For a one-sided test, the 5% rejection region is located solely in one tail of the distribution, as shown in Figures 3.15 and 3.16, for a test where the alternative is of the 'less than' form, and where the alternative is of the 'greater than' form, respectively.

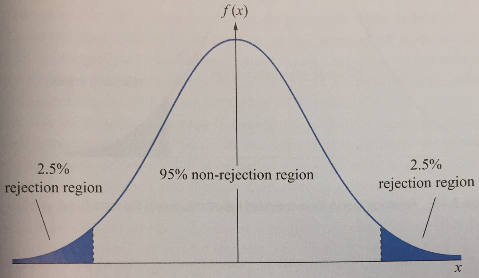


Figure 3.14 Rejection regions for a two-sided 5% hypothesis test

t-test review: one sided test

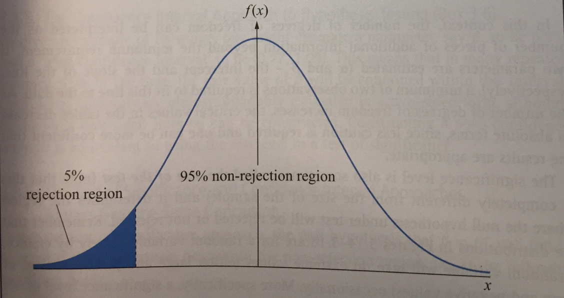


Figure 3.15 Rejection region for a one-sided hypothesis test of the form H_0: $\beta = \beta^*$, H_1: $\beta < \beta^*$

Some matrix algebra

- **Scalar**: a single number
- **Matrix:** a two-dimensional array. $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{23} \end{pmatrix}$ is a (3×2) matrix–that is, 3 rows and 2 columns. We say the number of rows then columns. a_{11} is the (1,1) element of A, and is a scalar. The **subscripts** of the elements tell us which row and column they are from.
- **Vector**: a one-dimensional array. If we take the first column of A and call it $A_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix}$, it is a (3×1) **column vector.** If we take the second row of A and call it $A_2 = \begin{pmatrix} a_{21} & a_{22} \end{pmatrix}$, it is a (1×2) **row vector.**

Square matrix: An $m \times n$ matrix is square if $m = n$.

$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$ is a square matrix. The **diagonal** of the matrix A are the elements $a_{11}, a_{22}, a_{33.}$

It only makes sense to talk about the diagonal of square matrices.

Symmetric matrix. For a square matrix, if the elements $a_{ij} = a_{ji}$, for $i \neq j$, then the matrix is symmetric. (notice the correspondence of the bold entries).

$$A = \begin{pmatrix} 2 & \mathbf{3} & \mathbf{4} \\ \mathbf{3} & 10 & \mathbf{6} \\ \mathbf{4} & \mathbf{6} & 11 \end{pmatrix}$$

Transpose of a matrix. The $i-th$ row becomes the $i-th$ column. The transpose of an $(m \times n)$ matrix is $(n \times \times m)$.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \end{pmatrix}, \text{ then } A' = A^T = \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix}.$$

$$A = \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix}, \text{ then } A' = A^T = \begin{pmatrix} a_{11} & a_{12} & a_{13} \end{pmatrix}.$$

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}, \text{ then } A' = A^T = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{pmatrix}$$

Zero matrix: All the entries are 0. $A = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$ is a zero matrix.

Identity matrix: A square matrix with 1s on the diagonal elements and 0 on the off-diagonal elements is called the identity matrix. $I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ is a (3×1) identity matrix. We always call an identity matrix I.

Matrix addition and subtraction

To add two matrices or to subtract one from the other, they must have the same dimensions. We do element by element addition or subtraction.

$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$,

Addition

$$C = A + B = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \underbrace{\begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \end{pmatrix}}_{C}$$

Subtraction

$$C = A - B = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} - \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}-b_{11} & a_{12}-b_{12} \\ a_{21}-b_{21} & a_{22}-b_{22} \end{pmatrix}$$

Scalar multiplication. Let $A = \left(\begin{array}{cc} a_{11} & a_{12} \\ a_{21} & a_{22} \end{array} \right)$, and c be a scalar.

$$cA = c \left(\begin{array}{cc} a_{11} & a_{12} \\ a_{21} & a_{22} \end{array} \right) = \left(\begin{array}{cc} ca_{11} & ca_{12} \\ ca_{21} & ca_{22} \end{array} \right) = Ac$$

Multiplying a matrix by a scalar means you multipliy every element by that scaler.

Matrix multiplication

If A is $(m \times n)$ and B is $(n \times k)$, they can be multiplied as AB, because the columns of A matches the rows of B. But you cannot multiply BA, because the columns of B doesn't match the rows of A. The result of multiplying an $(m \times n)$ matrix to an $(n \times k)$ matrix is $(m \times k)$.

Let A be (1×2) and B be (2×1) . A is a row vector and B is a column vector. $C = AB$ Multiplication is to do element by element multiplication, then sum the result.

$$A = \begin{pmatrix} a_{11} & a_{12} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} \\ b_{21} \end{pmatrix},$$

$$C = AB = \begin{pmatrix} a_{11} & a_{12} \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{21} \end{pmatrix} = (a_{11}b_{11} + a_{12}b_{21}) = C, \text{ (a scalar).}$$

Next, let's do it with actual matrices: Let

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}.$$

$C = AB$, is formed by $c_{ij} = \sum a_{ij} b_{ji}$. The i, j element of C is formed from multiplying row i of A and column j of B.

$$C = AB = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \\ b_{11}a_{31} + b_{21}a_{32} & a_{31}b_{12} + a_{32}b_{22} \end{pmatrix}$$

note: Even if A and B are both square matrices, the order matters. $AB \neq BA$.

Matrix Inverse

Determinant of a (2×2) matrix. Subtract the product of the off-diagonal elements from the product of the diagonal elements.

Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. $|A| = \text{det}(A) = ad - bc$.

Note: You can only get a determinant from **square** matrices. Calculating the determinant by hand from anything bigger than a (2×2) is beyond the scope of this class. But that's okay because we'll be doing it by computer.

Inverse of square matrix. Is a matrix when multiplied by itself gives the identity matrix. If $A^{-1}A = AA^{-1} = I$, then A^{-1} is the inverse of A. To get the inverse of a (2×2) matrix A (defined above), switch positions of the diagonal elements, multiply the off diagonal elements by -1, then divide everything by the determinant of A.

$A^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. Let's check: $\frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} =$

$\begin{pmatrix} \frac{ad}{ad-bc} - \frac{bc}{ad-bc} & 0 \\ 0 & \frac{ad}{ad-bc} - \frac{bc}{ad-bc} \end{pmatrix} = I$

Again, computing the inverse of anything bigger than a (2×2) matrix by hand is beyond the scope of this class. We just ask the computer to do it.

Regression in Matrix Form

Begin with

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

stack the dependent variable observations in a column vector and independent variables Independent variables: constant (a vector of 1s) and x_t

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}}_{y} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_T \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \alpha \\ \beta \end{pmatrix}}_{b} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{pmatrix}}_{\epsilon}$$

$$y = Xb + \epsilon$$

Multiply through by X'

$$
\begin{aligned}
X'y &= X'Xb + X'\epsilon \\
X'Xb &= X'\left(y - \epsilon\right) \\
b &= \left(X'X\right)^{-1} X'y - \left(X'X\right)^{-1} X'\epsilon
\end{aligned}
$$

Least squares forces the residuals $\hat{\epsilon}$ to be uncorrelated with the regressors. $X'\hat{\epsilon} = 0$. Hence, in matrix form, the least squares formula is

$$
\hat{b} = \left(X'X\right)^{-1} X'y
$$

Newey-West

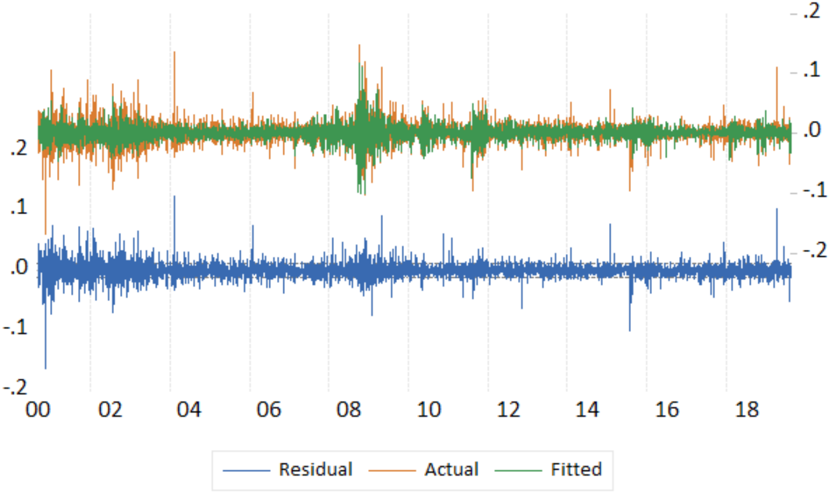
Motivate by runing the market regression and look at the error terms. See if there is volatility clustering. The market model (no excess return adjustment) for Disney, 8/20/2000 through 8/23/2019.

$$r_{i,t} = \alpha + \beta r_{m,t} + \epsilon_{i,t}$$

Dependent Variable: RET_DIS
Method: Least Squares
Date: 08/26/19 Time: 13:14
Sample (adjusted): 8/29/2000 8/23/2019
Included observations: 4603 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.000	0.000	1.090	0.276
RET_MKT	1.074	0.016	67.363	0.000
R-squared	0.497	Mean dependent var		0.000
Adjusted R-squared	0.496	S.D. dependent var		0.018
S.E. of regression	0.013	Akaike info criterion		-5.876
Sum squared resid	0.755	Schwarz criterion		-5.874
Log likelihood	13526.560	Hannan-Quinn criter.		-5.875
F-statistic	4537.714	Durbin-Watson stat		2.119
Prob(F-statistic)	0.000			

Plot residuals



Newey-West

- In basic econometrics, the reasoning was

$$\begin{aligned} Var\left(\hat{\beta}\right) &= Var\left(\frac{\sum \tilde{x}_t \epsilon_t}{\sum \tilde{x}_t^2}\right) = \frac{Var\left(\epsilon_t\right)\sum \tilde{x}_t^2}{\left(\sum \tilde{x}_t^2\right)^2} \\ &= \frac{\sigma_\epsilon^2}{\sum \tilde{x}_t^2} = \sigma_\epsilon^2 \left(X'X\right)^{-1} \end{aligned}$$

  numerator simplifies, variance of sum is sum of variances under independence. But now, Var(ϵ_t) isn't constant (due to conditional heteroskedasticity) and isn't independent (due to serial correlation).
- The variance after the first equals sign now has a bunch of covariance terms.

Newey-West

- The Newey-West covariance estimator does the correct calculations, taking into account these complicating factors. Instead of $\text{Var}(\hat{\beta}) = \sigma_\epsilon^2 \left(X'X\right)^{-1}$, it is

$$\text{Var}\left(\hat{\beta}\right) = \left(X'X\right)^{-1} S_T \left(X'X\right)^{-1} \tag{6}$$

$$S_T = S_0 + \frac{1}{T}\sum_{\ell=1}^{p} w(\ell) \sum_{t=\ell+1}^{T} \epsilon_t \epsilon_{t-\ell}\left(x_t x_{t-\ell} + x_{t-\ell} x_t\right) \tag{7}$$

$$w(\ell) = 1 - \frac{\ell}{p+1} \tag{8}$$

  Don't worry. There is an option in Eviews to do this computation and automatically get Newey-West t-ratios.
  Rule: In time-series regression **always** do Newey-West.

Newey-West

Method: Least Squares
Date: 08/26/19 Time: 14:06
Sample (adjusted): 8/29/2000 8/23/2019
HAC standard errors & covariance
Bartlett kernel, Newey-West fixed bandwidth = 10.00)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.000	0.000	1.206	0.227
RET_MKT	1.0739	0.0208	51.473	0.00
R-squared	0.496	Mean dependent var		0.000
Adjusted R-squared	0.496	S.D. dependent var		0.018
S.E. of regression	0.0128	Akaike info criterion		-5.876
Sum squared resid	0.755	Schwarz criterion		-5.873
Log likelihood	13526.56	Hannan-Quinn criter.		-5.875
F-statistic	4537.714	Durbin-Watson stat		2.118
Prob(F-statistic)	0.000	Wald F-statistic		2649.517
Prob(Wald F-statistic)	0.000			