Cedarville University Mathematics

Math 4800 Capstone Experience in Mathematics

Regression Analysis with Time Series Data

- By Logan O'Brien -

Date: April 21, 2020

Introduction

In this article, we will begin by giving a brief overview of econometrics as a whole and delve into its history. Then, we will look at how the topic of regression analysis with time series data fits within the broader category of econometrics by covering some of the characteristics of time series data, by looking at a popular model for analysis, and by examining the classic way to estimate the parameters for that model. With this theory in mind, we will then discuss the Gauss-Markov Theorem, a hugely important theorem foundational to classical linear parametric regression models. We will examine its importance, evaluate the assumptions upon which it relies, and prove a version of it. After that, we will look at a simple example in which the personal tax exemption's affect on the fertility rate is examined, to help clarify the concepts discussed in this paper before concluding the article.

Before we can get very far in the present discussion, we first need to know what econometrics is. It is generally believed that the origin of the name came from Ragnar Frisch (1895-1973), a Norwegian economist and a co-winner of the first Nobel Prize in the economic sciences [5]. Now, econometrics "involves the unified study of economics, economic data, mathematics, and statistical models" [5] and is "based upon the development of statistical methods for estimating economic relationships, testing economic theories, and evaluating and implementing government and business policy" [9]. Essentially, the goal of econometrics is to drive theoretical insights in the social sciences by studying and modeling economic principles. For this reason, the most common application of econometrics is the forecasting of important macroeconomic variables like interest rates, inflation rates, gross domestic product, etc. where whole economic systems are at play. However, econometrics can also be employed in other areas of the social sciences as well, e.g. market research and operations research, and the emphasis may be on examining individual aspects of a model with statistical inference rather than on generating forecasts [9].

Now, given all of this, the work cycle that an econometrician (someone who works in econometrics) performs, can be broken down into several steps that Wooldridge describes in [9]. After coming across a problem that needs to be solved, an analyst typically begins with defining an econometric model that describes the situation, and then makes hypotheses about the parameters (coefficients of the independent variables) in the model. After the model is created, data is collected and econometric methods are used to estimate the parameters. Then, the hypotheses of interest can be formally tested and forecasts can be made using the fitted model. A critical component to econometrics is regression analysis.

Regression analysis can be defined as "a statistical technique for modeling and investigating the relationships between an outcome or response variable and one or more predictor or

regressor variables" [5]. Regression is useful in multiple facets of econometrics. For instance, it can be used to predict future observations of the dependent variable when forecasting time series data, and for removing trends and seasonal effects in time series data prior to analysis [5].

History

Throughout history, the developments in time series analysis have always been influenced significantly by the applications for which they were used. One of, if not the earliest developments in regression analysis was the creation of the method of least squares, a technique for minimizing the error of fitting a model to a situation based on observable data. Adrien Legendre (1752-1833), a French mathematical scientist and contemporary of Laplace, was the first to publish an article containing the method, which was "the most widely used nontrivial technique of mathematical statistics" [6]. After its initial publication, the method quickly became a standard tool in the sciences and spread rapidly throughout Europe [6]. According to Stigler, "the rapid geographic diffusion of the method and its quick acceptance in [astronomy and geodesy]...is a success story that has few parallels in the history of scientific method" [6].

From this point in history, regression analysis continued to develop. In the 1920s and 1930s, the formal practice of time series forecasting via autoregressive models began through the work of Yule and Walker [10]. Then, after several decades, Box and Jenkins published their seminal work in 1970 entitled *Time Series Analysis* in which they presented a systematic process for the entire modeling procedure and unified the research surrounding time series analysis [7]. Since then, developments have continued in many nuanced areas of regression analysis. However, in the remainder of this paper, the discussion will be centered on the linear parametric models of regression, which Tsay deemed the "traditional analysis" methods [7]. Specifically, the multiple regression model will be closely examined since it is "still the most widely used vehicle for empirical analysis in economics and other social sciences" [9].

The Theory

Selecting a Model

Now, as explained above, an important first step in an econometrician's work cycle is model selection, and a popular choice is some variant of the multiple linear regression analysis (MRA) model. Why is this model so popular? In most tests of economic theory,

an important goal is to establish that one variable has a causal effect on another. In order to do this, all other factors in the model must be held constant, which is the notion of *ceteris paribus*. Now, in the realm of mathematical statistics, especially when applied in the physical sciences, this requirement is not too difficult to meet, since the work centers around carefully designed, reproducible experiments and studies that allow analysts to gather the data in a ceteris paribus fashion. However, econometrics, by nature, frequently deals with nonexperimental data in the social sciences. There, it is often unethical or even impossible to construe situations that meet the condition of ceteris paribus. The good news is that econometricians have some tools to help them deal with the additional difficulties of working with nonexperimental data. One such technique is the MRA model. As Wooldridge says, "The power of multiple regression analysis is that it provides this ceteris paribus interpretation even though the data have not been collected in a ceteris paribus fashion... the power of multiple regression analysis is that it allows us to do in nonexperimental environments what natural scientists are able to do in a controlled laboratory setting: keep other factors fixed" [9]. This is amazing! This means that even in the vast majority of cases where it is impossible to hold all else constant when gathering the data, the multiple regression analysis method can simulate a ceteris paribus experiment.

How can it do this? Consider the basic form of the MRA model for the population:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u,$$

where y is some quality that is explained by factors x_1 through x_k. The random variable u is the error term and represents all other factors not explicitly accounted for in the model that affects y. β_0 is the intercept in the model and β_1 through β_k are the parameters in the model. The presence of the parameters in this model is what allows it to simulate gathering the data in a ceteris paribus fashion, because, the coefficient β_i measures the ceteris paribus affect of x_i on y, where $i \in \{1, 2, 3..., k\}$. By including an explanatory variable explicitly in the model, we are able to control for its affects on y. It is also important to note that this model is linear, not because the explanatory variables are linear, but because the parameters are linear. As a result, this model is quite flexible and can be used in a variety of contexts.

The Nature of Time Series Data

Now that we have discussed the basic cross-sectional form of the MRA model, we can look at the version of the model that handles time series data. But first we need to discuss some of the characteristics of time series data. Time series data, stated simply, is data that is ordered temporally – it is the observation of one or more variables over time. This

type of data is important in a variety of applications because so often the situations and problems studied are impacted by time. Stock market trends, budget analyses, projections and forecasts are examples of applications involving time.

A term that frequently arises when discussing time series data is a *stochastic process*, which is a sequence of random variables indexed by time [9]. It is important to grasp intuitively why a stochastic process consists of random variables. The justification is this. Suppose we had a stochastic process that represented the observation of several variables over time. Now, it would be reasonable to expect that if some past event related to the time series process was rewound and played out differently, it would affect the resulting process described by the outcomes of the sequence of random variables. It is for this reason that time series data is considered random. Now, it is also important to note that when working with time series data, the statistical population is all realizations of a time series process and the sample size is the number of time periods over which the variables of interest are observed. Hence, a sample is the stochastic process observed over one time period.

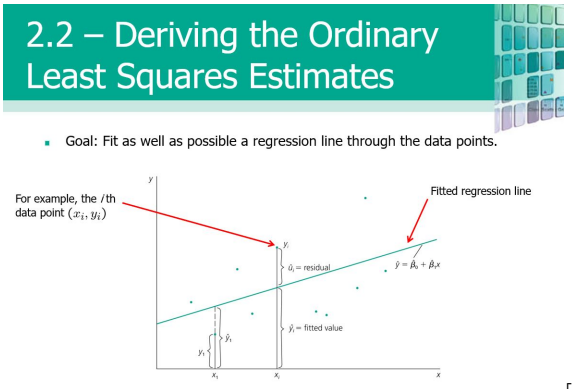
Now that we know a bit about time series data, we can consider the time series form of the MRA model. The stochastic process $\{(x_{t1}, x_{t2}, \cdots, x_{tk}, y_t) : t = 1, 2, \cdots, n\}$ follows the linear model

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t,$$

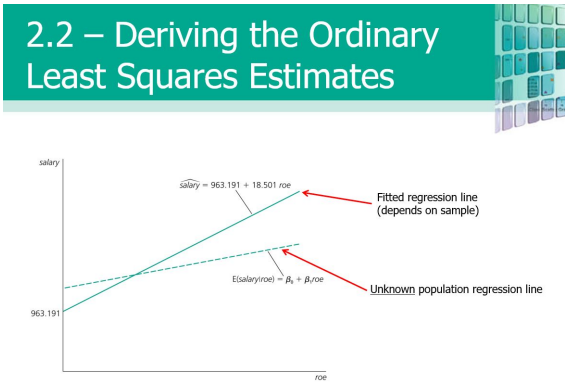
where $\{u_t : t = 1, 2, \cdots, n\}$ is the sequence of errors or disturbances and n is the sample size. The important difference between this model and the MRA model given earlier is the addition of the t subscripts for the time periods for which this population model holds.

The Ordinary Least Squares Method

After defining a model and making hypotheses, the next step for the econometrician is to estimate the parameters on each of the explanatory variables so that he or she can determine the estimated model and conduct further analysis. According to Wooldridge, one of the most popular methods for parameter estimation is the method of ordinary least squares (OLS) that was briefly introduced above [9]. OLS can be used in time series analysis to construct a model such that the sum of the squares of the residuals is minimized, where the residuals are the differences between the population equation and the estimated equation. For clarity, consider the images below, which come from Cengage's lecture slides on econometrics.



[4]



[4]

The first picture above handily illustrates the OLS method at work and also gives a nice visual explanation of what a residual is. In the second image, we see that the result of OLS parameter estimation is a fitted line (since we assumed the population fits the MRA model, which is linear). We further note that we seek an estimated line as close to the true model as possible. Now, it can be quite tedious to calculate the estimates for the betas by hand, especially when the MRA model contains 3 or more explanatory variables and requires multivariable calculus [9]. Thankfully, however, most modern statistical packages can determine these estimates with ease.

The Gauss-Markov Theorem

Theorem Statement

Now that we know a bit about the OLS method, why is it such a popular choice for parameter estimation, especially in the context of time series regression? The answer to that question lies with the Gauss-Markov Theorem, one of the fundamental results in the realm of econometrics. This theorem was developed by the work of German mathematician Carl Friedrich Gauss (1777-1855) and Russian mathematician Andrey Markov (1856-1922), pictured below from left to right. Though they were not contemporaries, their joint efforts produced what we now call The Gauss-Markov Theorem.


[2]


[1]

Here is the theorem:

Theorem (Gauss-Markov for time series data):
Under Assumptions TS.1 through TS.5, the OLS estimators are the best linear unbiased estimators (BLUE) conditional on $\boldsymbol{X}$.

In this theorem, Assumptions TS.1 through TS.5 (also called the Gauss-Markov Assumptions) are a set of 5 assumptions related to regression analysis with time series data, and $\boldsymbol{X}$ is the set of all independent variables for all time periods in the model. Let's take a look at the GM assumptions so we can better understand the theorem.

The Gauss-Markov Assumptions

It is important to note, before we begin, that several of these assumptions, especially TS.3, TS.4, and TS.5 can be difficult to meet when applied to the social sciences. Although this means that the GM Theorem cannot be applied in these contexts, the theorem and its associated assumptions are still important to study because they form the foundation for classical linear regression analysis, and work has been done to iron out some of these difficulties. For instance, Wooldridge outlines how a tweaked version of this theorem and its corresponding assumptions can be utilized to provide a basis for things like statistical inference when working with large sample sizes. So, let's begin.

The first of the GM assumptions, TS.1, essentially defines an MRA model that works for time series data and mandates that the time series process is linear in its parameters. The formal statement is below and a careful reader might notice that we already saw this model above when times series data was first introduced.

Assumption TS.1
The stochastic process $\{(x_{t1}, x_{t2}, ..., x_{tk}, y_t) : t = 1, 2, ..., n)$ follows the linear model

$$y_t = \beta_0 + \beta_1 x_{t1} + ... + \beta_k x_{tk} + u_t,$$

where $\{u_t : t = 1, 2, ..., n\}$ is the sequence of errors or disturbances. Here, n is the number of observations (time periods).

The second assumption is the No Perfect Collinearity Assumption and states
Assumption TS.2
No Perfect Collinearity: In the sample (and therefore in the underlying time series process), no independent variable is constant nor a perfect linear combination of the others.

Now, this assumption is important because if perfect collinearity occurs in a model, it cannot be estimated by OLS [9]. According to Wooldridge, there are several ways perfect collinearity can occur. If one is not careful when specifying a model, one variable may be expressed as a constant multiple of another, or a regressor might end up as an exact linear function of two or more of the other regressors. It can also fail if the sample size is too small, since the number of observations in the sample must be at least the number of parameters that we are seeking to estimate. Another way TS.2 can fail, on rare occasions, is due to bad luck. Consider this example from [9]. Suppose we created a model to study the effects of years of education and years of on-the-job training on salary, and come to discover that the sample values of the education variable in the data we gathered were always 3 times the value of the corresponding training variables. This would be perfect collinearity. Now, it is important to realize that this assumption does not rule out all collinearity between explanatory variables in the same sample, but only *perfect* collinearity. Also, this problem is not unique to time series applications since it crops up in cross-sectional applications as well.

The third assumption is
Assumption TS.3
Zero Conditional Mean: For each t, the expected value of the error u_t, given the explanatory variables for *all* time periods, is zero. Mathematically,

$$\mathbb{E}(u_t|\boldsymbol{X}) = 0, t = 1, 2, ..., n.$$

In other words, this assumption states that "The average value of u_i is unrelated to

the independent variables of all time periods" [9]. Note that if $\boldsymbol{X}$ and u_t are independent and $\mathbb{E}[u_t] = 0$, then TS.3 automatically holds [9]. The corresponding Gauss-Markov (GM) assumption for cross-sectional data is called MLR.4 and is worth a brief comparison. Assumption MLR.4 essentially states two things. One, that on average the value of the error term should be zero, and two, that on average it must be uncorrelated with the explanatory variables. That is, the explanatory variables should contain no information about the expected value of the error term.

Now, given this information about MLR.4, it is not surprising that we have a corresponding assumption for time series data, TS.3, that mandates that u_t be uncorrelated with any of the explanatory variables at time t, called *contemporaneous exogeneity*. However, TS.3 is stricter than MLR.4. Rather than being contemporaneously exogenous, u_t must be uncorrelated with x_{sj}, even if $s \neq t$, where $j \in \{1, 2, 3, \cdots, k\}$ and s, t are time periods over which the stochastic process is observed. When this happens, we say that the explanatory variables are *strictly exogenous*. The reason this is necessary for time series data but not cross-sectional data, is that when working with cross-sectional data we employ random sampling, which necessarily means that the error term for observation i is independent of the explanatory variables of any other observation. But, random sampling is almost never done in time series regression analysis because the data is temporarily ordered; we don't want it to be random, which is why we must explicitly assume TS.3.

Assumption TS.3 can fail in several ways. For instance, the model can be underspecified, where one or more important explanatory variables are omitted from the population model. Also, according to Wooldridge, measurement error in regressors can cause problems, and feedback in a model where the dependant variable affects later values of one or more independent variables also causes TS.3 to fail. That is, explanatory variables in a model cannot react to past values of the dependant variable [9]. Because of these issues, TS.3 can often be impractical to apply in the social sciences.

Interestingly, by the three assumptions above, we have the necessary material to support a theorem that states that the OLS estimators are unbiased. By adding two more assumptions, TS.4 and TS.5, however, we arrive at the full list of GM assumptions and meet the conditions for The Gauss-Markov Theorem, which says that the OLS estimators are BLUE, for time series data.

The fourth GM assumption says

Assumption TS.4

Homoskedasticity: Conditional on $\boldsymbol{X}$, the variance of u_t is the same for all t: $\mathbf{Var}(u_t|\boldsymbol{X}) = \mathbf{Var}(u_t) = \sigma^2, t = 1, 2, ..., n.$

This assumption is the time series version of MLR.5 and says that the variance of the unobservables (error term) cannot depend on $\boldsymbol{X}$ and that it must be constant. When this assumption does not apply, the errors are said to be heteroskedastic. Note that the size of the variance is important, because the larger the spread in the error term, the less precise our estimator is, which in turn means we have larger confidence intervals and less accurate hypothesis tests [9]. Although situations can arise in applied problems where the errors are heteroskedastic, for the sake of time we won't discuss them here.

Last but not least is the fifth Gauss-Markov assumption. It says,

Assumption TS.5

No Serial Correlation: Conditional on $\boldsymbol{X}$, the errors in two different time periods are uncorrelated: Corr$(u_t, u_s|\boldsymbol{X}) = 0$ for all $t \neq s$.

When this assumption fails, the error terms are said to suffer from serial correlation, or autocorrelation. Interestingly, serial correlation is not an issue for cross-sectional data because the random sampling assumption removes the possibility of it occurring, however as Wooldridge states in his textbook, serial correlation does characterize the error terms in many time series applications [9]. Possible solutions to TS.5 failure are discussed at length in [9].

The Proof

Now that we have discussed the importance of the Gauss-Markov theorem and outlined the five time series assumptions, let's prove it. However, in this discussion, we will actually prove the Gauss-Markov Theorem for cross-sectional data, rather than time series data. The reason for this is that Wooldridge's *Introductory Econometrics: A Modern Approach*, an authoritative source on econometrics, omits the proof for the time series version of the theorem because several of the GM assumptions and the justification for supporting theorems are so similar in both contexts. Hence, I will focus on the proof of the one he did cover, which was the cross-sectional version of the theorem. This version of the theorem states,

Theorem (Gauss-Markov for cross-sectional data):

Under Assumptions MLR.1 through MRL.5, $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ are the best linear unbiased estimators of $\beta_0, \beta_1, \cdots, \beta_k$, respectively.

Now, the key assumption under which this theorem operates is **MLR.1** which essentially

says that the true model fits the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u,$$

where $\beta_0, \beta_1, \cdots, \beta_k$ are the unknown parameters of interest and u is an unobserved random error or disturbance term. With this in mind, we can proceed to the proof, which is covered sparingly in [9].

Proof. Let us begin by defining $\hat{\beta}_j$ to be the OLS estimator for β_J, the corresponding parameter for the jth explanatory variable in the population model, and let $\tilde{\beta}_j$ be any other linear unbiased estimator. Then, we seek to show that $var(\tilde{\beta}_j) \geq var(\hat{\beta}_j)$. We know from **Assumption MLR.1** that

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u. \tag{1}$$

Since $\tilde{\beta}_j$ is a linear estimator, by definition

$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i, \tag{2}$$

where each w_{ij} can be a function of the sample values of all the independent variables. Without loss of generality, let $j = 1$ and plug (1) into (2). Then we have,

$$\tilde{\beta}_1 = \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \cdots + \beta_k \sum_{i=1}^{n} w_{i1} x_{ik} + \sum_{i=1}^{n} w_{i1} u_{i1}.$$

From here, we can take the expected value of both sides, yielding

$$\begin{aligned} \mathbb{E}[\tilde{\beta}_1 | \boldsymbol{X}] &= \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \cdots + \beta_k \sum_{i=1}^{n} w_{i1} x_{ik} + \sum_{i=1}^{n} w_{i1}\, \mathbb{E}[u_{i1} | \boldsymbol{X}] \\ &= \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \cdots \beta_k \sum_{i=1}^{n} w_{i1} x_{ik}, \end{aligned} \tag{3}$$

since the w_{i1} are functions of the x_{ij} and by the mean zero assumption MLR.4, $\mathbb{E}[u_i | \boldsymbol{X}] = 0$, for all $i = 1, 2, \cdots, n$. Now, since we assumed at the start of this proof that $\tilde{\beta}_1$ is unbiased, we know that $\mathbb{E}[\tilde{\beta}_1 | \boldsymbol{X}] = \beta_1$ for any values of the parameters. Hence,

$$\sum_{i=1}^{n} w_{i1} = 0, \sum_{i=1}^{n} w_{i1} x_{i1} = 1, \sum_{i=x}^{n} w_{i1} x_{ij} = 0, \text{for} j = 2, \cdots, k. \tag{4}$$

Next, let $\hat{r_{i1}}$ be the residuals from the regression of x_{i1} on $x_{i2}, x_{i3}, \cdots, x_{ik}$. Then, $\hat{x_{i1}} = \hat{\gamma_1} + \hat{\gamma_2}x_{i2} + \cdots + \hat{\gamma_k}x_{ik}$. So,

$$\begin{aligned}\sum_{i=1}^{n} w_{i1}\hat{x_{i1}} &= \sum_{i=1}^{n}(\hat{\gamma_1} + \hat{\gamma_2}x_{i2} + \cdots + \hat{\gamma}x_{ik}) \\ &= \hat{\gamma_1}\sum_{i=1}^{n} w_{i1} + \hat{\gamma_1}\sum_{i=1}^{n} w_{i1}x_{i2} + \cdots + \hat{\gamma_1}\sum_{i=1}^{n} w_{i1}x_{ik} \\ &= 0 \text{ (by (4))}. \end{aligned} \tag{5}$$

Because we regressed x_1 on the other explanatory variables, we know that $x_{i1} = \hat{x_{i1}} + \hat{r_{i1}}$. In light of this, we can rewrite (5) and determine that,

$$\begin{aligned}\sum_{i=1}^{n} w_{i1}\hat{x_{i1}} &= 0 \\ \sum_{i=1}^{n} w_{i1}(x_{i1} - \hat{r_{i1}}) &= 0 \\ \sum_{i=1}^{n} w_{i1}x_{i1} - \sum_{i=1}^{n} w_{i1}\hat{r_{i1}} &= 0 \\ 1 - \sum_{i=1}^{n} w_{i1}\hat{r_{i1}} &= 0 \text{ (by (4))}\end{aligned}$$

Hence,

$$\sum_{i=1}^{n} w_{i1}\hat{r_{i1}} = 1 \tag{6}$$

With this groundwork laid, we can now return to comparing the variance of both kinds of estimators under the 5 cross-sectional data Gauss-Markov assumptions. Now, I'm going to give expressions for the variance of both estimators without justification for the sake of time. However, if you wish to learn more you can examine [9] which presents justification for these expressions. So, from [9] we have that $Var(\tilde{\beta_1}|\boldsymbol{X}) = \sigma^2 \sum_{i=1}^{n} w_{i1}^2$ and $Var(\hat{\beta_1}|\boldsymbol{X}) = \frac{\sigma^2}{\sum_{i=1}^{n} \hat{r_{i1}}^2}$. Thus, the difference between the variances is

$$Var(\tilde{\beta_1}|\boldsymbol{X}) - Var(\hat{\beta_1}|\boldsymbol{X}) = \sigma^2 \sum_{i=1}^{n} w_{i1}^2 - \frac{\sigma^2}{\sum_{i=1}^{n} \hat{r_{i1}^2}}.$$

Now, we can drop σ^2 since it will not be needed in future steps and we can use (6) to write

the difference as

$$\sum_{i=1}^{n} w_{i1}^2 - \frac{\left(\sum_{i=1}^{n} w_{i1}\hat{r_{i1}}\right)^2}{\sum_{i=1}^{n} \hat{r_{i1}^2}}. \tag{7}$$

Surprisingly, (7) can be rewritten as

$$\sum_{i=1}^{n} (w_{i1} - \hat{\zeta_1}\hat{r_{i1}})^2, \tag{8}$$

where $\hat{\zeta_1} = \frac{\sum_{i=1}^{n} w_{i1}\hat{r_{i1}}}{\sum_{i=1}^{n} \hat{r_{i1}}^2}$. To see this, let's actually work in reverse, starting with (8). By squaring, summing, and then canceling terms, we have that

$$\begin{aligned}
\sum_{i=1}^{n} (w_{i1} - \hat{\zeta_1}\hat{r_{i1}})^2 &= \sum_{i=1}^{n} (w_{i1}^2 - 2\hat{\zeta_1} w_{i1}\hat{r_{i1}} + \hat{\zeta_1}^2 \hat{r_{i1}}^2) \\
&= \sum_{i=1}^{n} w_{i1}^2 - 2\hat{\zeta_1} \sum_{i=1}^{n} w_{i1}\hat{r_{i1}} + \hat{\zeta_1}^2 \sum_{i=1}^{n} \hat{r_{i1}}^2 \\
&= \sum_{i=1}^{n} w_{i1}^2 - 2\left[\frac{\sum_{i=1}^{n} w_{i1}\hat{r_{i1}}}{\sum_{i=1}^{n} \hat{r_{i1}}^2}\right] \sum_{i=1}^{n} w_{i1}\hat{r_{i1}} + \left[\frac{\sum_{i=1}^{n} w_{i1}\hat{r_{i1}}}{\sum_{i=1}^{n} \hat{r_{i1}}^2}\right]^2 \sum_{i=1}^{n} \hat{r_{i1}}^2 \\
&= \sum_{i=1}^{n} w_{i1}^2 - 2\frac{\left(\sum_{i=1}^{n} w_{i1}\hat{r_{i1}}\right)^2}{\sum_{i=1}^{n} \hat{r_{i1}}^2} + \frac{\left(\sum_{i=1}^{n} w_{i1}\hat{r_{i1}}\right)^2}{\sum_{i=1}^{n} \hat{r_{i1}}^2} \\
&= \sum_{i=1}^{n} w_{i1}^2 - \frac{\left(\sum_{i=1}^{n} w_{i1}\hat{r_{i1}}\right)^2}{\sum_{i=1}^{n} \hat{r_{i1}}^2}.
\end{aligned}$$

Since the last line above is (7), rewriting (7) as (8) is valid, and hence

$$Var(\tilde{\beta_1}|\boldsymbol{X}) - Var(\hat{\beta_1}|\boldsymbol{X}) = \sum_{i=1}^{n} (w_{i1} - \hat{\zeta_1}\hat{r_{i1}})^2.$$

Now, here is the climax of the proof. Notice that the expression on the right is always nonnegative. Therefore the difference between the variance of the two estimators must also be nonnative and so the variance of any other linear unbiased estimator is at least as large as the variance of the OLS estimator. Thus the OLS estimator is best. □

Application

Now that we have proved the Gauss-Markov Theorem, let's use it in a short example that was discussed in [9]. Suppose we wish to study the effects of the personal tax exemption on fertility rates. According to the work flow process explained earlier in this paper, we might

begin by selecting a model that describes the population. For this illustration, say we pick the following as our true model:

$$gfr_t = \beta_0 + \beta_1 pe_t + \beta_2 ww2_t + \beta_3 pill_t + u,$$

Where, gfr is the general fertility rate, the number of children born to every 1,000 women of childbearing age, pe is the personal tax exemption, and $ww2$ and $pill$ are dummy variables for the time period during World War 2 (1941-1945) and time period after birth control became available (1963), respectively. Dummy variables can actually be quite handy. Typically, a dummy variable is a binary variable (possesses only the value 1 or 0) and represents a qualitative not quantitative factor. So, in our example, the $ww2$ variable is 1 during World War 2, and 0 otherwise. This ensures that this variable can only impact gfr during the years in which World War 2 occurred. The variable $pill$ behaves in a similar manner.

After selecting a model, a typical next step would be to make hypotheses about how these factors might impact the dependent variable, or even make more involved predictions about specific qualities of this problem that fall under the category of statistical inference. Here, we will limit ourselves to two simple hypotheses. Looking at the model above, we would expect all three of the explanatory variables to have a significant impact on gfr, but that the dummy variables impact would be more significant than pe. Furthermore, it makes sense that β_3, the coefficient on pe would be positive, so that an increase in the personal tax exemption would increase the general fertility rate. Likewise, we would expect the two dummy variables to both have negative coefficients because we would guess that those time periods would reduce the general fertility rate.

Now that we have our model and hypotheses, the next step is to collect data and estimate the true model. Our data set comes from an article published in 1990 in the *American Economic Review* by Whittington, Alm, and Peters [8], but the data set was cleaned and organized by Cengage, and was grabbed from [3]. By the Gauss-Markov Theorem for time series data, we know that the OLS estimators are the best linear unbiased estimators for the parameters, assuming the 5 assumptions apply. Here, according to Wooldridge, they do apply, since the textbook utilized this method to estimate the model in this example. Now, thanks to modern technology, we can use a standard statistical program to calculate the estimated model using the OLS method. When we run such a program on our data, we

receive the following output:

$$\begin{gathered} \hat{gfr}_t = 98.68 + 0.83pe_t - 24.24ww2_t - 31.59pill_t \\ (3.21) \quad (.030) \qquad (7.46) \qquad (4.08) \\ n = 72, R^2 = .473, \ \overline{R^2} = .450. \end{gathered}$$

Here, we can see that the program provided numerical estimates for the parameters with the corresponding variance for each of those estimates in the second line directly below each variable. The last line reports that there were 72 samples in the data set and also gives two metrics for how well this estimated model fits the data.

Since we have an estimated model, we can now revisit our initial hypotheses. Recall that we expected the coefficients for $ww2$ and $pill$ to be negative, which is indeed what we see here. To grasp the significance of theses values, note that in the data set, gfr ranged from 65 to 127 births per 1000 women of childbearing age. Thus, according to this estimated model, during World War 2, about 24 less births occurred per year and after easy access to birth control, about 32 less children were born per year. Since the gfr values in the data set ranged from 65 to 127 per year, the impact of these dummy variables is quite significant, like we predicted. Also, we see by inspection that $\beta_1 = 0.83$. Now, according to the data set, the average value of pe was \$100.40. Based on the value for β_1, a \$12.00 increase in pe increases gfr by about one birth per year. Although this might not appear noteworthy, it is still a 0.7% increase in the general fertility rate, and hence it is significant, especially since there are a lot more than 1000 women of childbearing age. Thus, we see that our initial hypotheses were verified by harnessing the power of the OLS method in time series regression.

Conclusion

In this paper, we explored the topic of regression analysis with time series data. We began by defining several important terms in econometrics and explored the wide variety of applications for time series regression. Then, we followed those with a brief history of time series regression, back to the origin of the OLS method. After this, we dove into the theory of time series regression and looked at the MRA model, the nature of time series data, and the OLS method of parameter estimation. Next we examined the Gauss-Markov Theorem, its pivotal role in regression analysis, the assumptions it relies on, and went through a proof for the cross-sectional data version. Finally, we wrapped up the discussion with an example in studying the effects of personal tax exemption rates on fertility rates. I hope that the reader found this article both engaging and educational.

Acknowledgments

I would like to thank several individuals for their assistance in this research project. First, and foremost, I want to thank Dr. Lindsey McCarty for serving as my research advisor and instructor. Also, I want to credit Dr. Ashley Holland for aiding me with several steps in the proof of the Gauss-Markov Theorem and my research librarian, Nathanael Davis, for assisting me a bit with my initial research.

References

[1] *Andrey Markov*. Accessed: 04-13-2020. URL: https://en.wikipedia.org/wiki/Andrey_Markov.

[2] *Carl Friedrich Gauss*. Accessed: 04-13-2020. URL: https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss.

[3] Cengage. *Fertil3 Data Set*. Accessed: 04-12-2020. URL: https://www.cengage.com/cgi-wadsworth/course_products_wp.pl?fid=M20b&product_isbn_issn=9781111531041.

[4] Cengage. *The Simple Regression Model*. Slide presentation for Wooldridge: Introductory Econometrics 5e.

[5] Douglas C Montgomery, Cheryl L Jennings, and Murat Kulahci. *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.

[6] Stephen M Stigler. *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1986.

[7] Ruey S Tsay. "Time series and forecasting: Brief history and future research". In: *Journal of the American Statistical Association* 95.450 (2000), pp. 638–643.

[8] Leslie A Whittington, James Alm, and H Elizabeth Peters. "Fertility and the personal exemption: implicit pronatalist policy in the United States". In: *The American Economic Review* 80.3 (1990), pp. 545–556.

[9] Jeffrey M Wooldridge. *Introductory econometrics: A modern approach*. Nelson Education, 2016.

[10] Leila Zoubir. *A brief history of time series analysis*. URL: https://www.statistics.su.se/english/research/time-series-analysis/a-brief-history-of-time-series-analysis-1.259451.