SAS IN TAMING CLASSICAL VIOLATIONS

Kamal Shoukry, Loyola Marymount University

Introduction

Regression analysis lies at the heart of statistics and econometrics where the purpose is to estimate the true parameters and the ultimate goal is prediction. Reliable estimates are essential to accurate prediction. The purpose of this study is to ensure that parameters estimates are the best possible estimates of the true population values by attempting to isolate or minimize the distortion effects caused by some associated phenomena.

Estimation, or approximating populations' relationships using a sample data set is a key task of statistics and econometrics. Two statistical estimators of the population mean and variance are of particular importance in judging the desirable properties of estimators, namely unbiasedness, efficiency and minimum mean square error.

Bias refers to the distribution of sample estimates. A sample estimate is unbiased if its expected mean equals the true population value, $E(\hat{\beta}_i)=\beta_i$; and its variance equals zero $var(\hat{\theta})=0$. The estimator is unbiased when the distribution of the sample estimates is centered around the true population parameter. An efficient estimator has a relatively small variance, and thereby increases our confidence in the parameter estimates. The mean square error, $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$, provides a trade-off tool between unbiasedness and efficiency ib both cannot occur together. The estimator having the minimum MSE would be selected.

The Classical Assumptions

Ordinary Least Squares (OLS) parameters estimates will be Best Linear Unbiased Estimates (BLUE) if the following assumptions are satisfied, (Brown, 1991).

A1. The model is linear and "correctly specified" by the equation:

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$$

If A1 is violated, parameters estimates will be biased or inefficient.

A2. The independent variables are nonstochastic and thus noncorrelated with the random error, $\int(X,\epsilon)=0$. If A2 is violated, parameter estimates will be biased.

A3. The error term has a zero mean, $E(\epsilon) = 0$. If A3 is violated, $\hat{\beta}_0$ will be biased. To facilitate hypothesis testing, it is also usually assumed that the errors follow a normal distribution.

A4. Error terms are noncorrelated, $\int(\epsilon_i,\epsilon_j) = 0$. If A4 is violated, the error terms are said to be autocorrelated and parameter estimates will be inefficient.

A5. The errors have a constant variance,

$$\sigma_i 2 = \sigma_j^2 \text{ for all } i,j .$$

If A5 is violated, the error terms are said to be heteroscedastic and parameter estimates will be inefficient.

A6. No exact linear relationship can exist between any combination of the independent variables. If the independent variables are perfectly collinear, $X_1 = c\ X_2$, where c is a constant, perfect multicollinearity exists and it would be difficult to determine which regressor, X_1 or X_2 , is responsible for the change in the dependent variable Y, and parameter estimates cannot be trusted.

The Common Violations of Classical Assumptions : Causes and Effects

The most common violations of OLS classical assumptions are multicollinearity, autocorrelation, and heteroscedasticity. The following is a brief summary of their causes and effects.

MULTICOLLINEARITY is due to significant correlation between regressors. Its classic symptoms are low T ratios and inflated R^2. Parameters estimates signs could also be reversed. To avoid specification errors, all influential variables should be included as independent variables. The absence of such variables may cause bias to the model's parameters estimates. On the other hand, exact linear relationships between two or more of the independent variables lead to perfect multicollinearity. The severity or weakness of multicollinearity depends on the strength or the weakness of correlation between regressors.

Multicollinearity is a violation of the sixth assumption (A6) if it is due to correlation among regressors, and of the first assumption (A1) if it is due to specification errors.

AUTOCORRELATION is also called serial correlation. It is a violation of the fourth assumption (A4) which requires error terms to be noncorrelated, or that consecutive errors are statistically independent. Autocorrelation is extremely common in models using time series data. Autocorrelation biases R-squares and T ratios upward and makes parameter estimates inefficient due to large error variance and biased standard error of the estimates. Autocorrelation however does not bias the estimates. Serial correlation is most often caused by inertia or omitted variables. Forecasts made with models containing autocorrelated errors often miss turning points.

HETEROSCEDASTICITY is a violation of the fifth assumption (A5) which requires error variances to be constant. It leads to inefficient but unbiased parameters estimates. Heteroscedasticity is most common in cross sectional data, but can also be found in time series data, and is typically associated with one or more of the independent variables. The case of consumption / disposable income cross sectional study is one of the well known cases in macroeconomics literature, and is a case of heteroscedasticity.

Violations Detection and Remedies

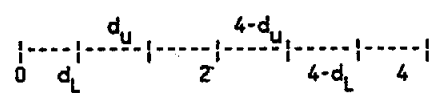
MULTICOLLINEARITY detection methods are based on correlation among the independent variables. A correlation matrix containing the correlation coefficients among all variables in the model is the easiest and most straight forward method. Strong correlation would indicate severe multicollinearity and the need to correct it.

A measure of severity of multicollinearity is the variance inflation factor (VIF). It is an estimate of how much multicollinearity has inflated the variance of an estimated coefficient. There is therefore, a VIF for each explanatory variable in the model. VIF ($\hat{\beta}_i$) = 1 / (1 - R^2_i) , where R^2_i is the unadjusted coefficient of determination of the auxiliary regression for each independent variable in the original model equation. There is an R^2_i and a VIF($\hat{\beta}_i$) for each regressor, (Studenmund, 1992).

Some statistical software programs use the Tolerance coefficient (TOL). The VIF and its alternative TOL are the most popular and comprehensive multicollinearity detection techniques. SAS computes the correlation matrix as an output for PROC CORR, and both VIF and TOL as options with PROC REG. Regarding the decision rule for the detection of severe multicollinearity many authors suggest VIF > 10 as a rule of thumb instead of VIF > 5. For example, in a sample where correlation coefficient between X_1 and X_2 is .88, the VIF is 4.4 and the TOL is approximately .226 . This is assuming that there are no other Xs in the model equation. Autocorrelation can be detected using graphical methods and or a more accurate computing method like the popular Durbin-Watson statistic, d.

A detectable pattern of a plot of residuals like upward or downward trend, or a wave-like pattern can indicate the existence of serial correlation.

The d statistic, $d = \Sigma^T_{t=2} (e_t - e_{t-1})^2 / \Sigma T_{t=1} e^2_t$, tests the null hypothesis of zero first order autocorrelation against the alternative, the presence of first order autocorrelation,(RHO)ρ. The DW method also generates an estimate of $\hat{\rho}$ the first order autocorrelation coefficient.
Given n, the sample size, and k, the number of independent variables in the model," d lower" (d_l) and "d upper" (d_u), can be obtained from the Durbin-Watson Statistic table. SAS computes both d and ρ . Then the decision rule is as follows.
No autocorrelation exists when d=2 and ρ=0 .
Perfect positive first order autocorrelation exists when d=0 and ρ=1 . Perfect negative first order autocorrelation exists when d=4 and ρ= - 1 .
On a scale from 0 to 4, the d_L and d_U obtained from the table are added to 0 in the case of positive autocorrelation, and subtracted from 4 in the case of negative autocorrelation. If d of SAS output is between d_l and d_u then autocorrelation is indeterminate. For the range around 2, from d_u to $4-d_u$, no autocorrelation exists.



The DW method is recommended only for large data sets. Generally, autocorrelation is corrected or eliminated using generalized least squares (GLS). Yet, when autocorrelation is caused by omitted variables, GLS may make the errors well behaved without correcting a likely bias resulting from the specification error. With small samples we would rather ignore autocorrelation, because even if ρ could be computed, it would be an incorrect estimation, (BROWN, 1994).
Heteroscedasticity is most common in cross sectional data, but can also be found in time series data, and is typically associated with one or more of the independent variables.
Graphical methods are quite helpful in detecting heteroscedasticity. A plot of residuals having a fanning out shape or a funnel shape indicate the existence of heteroscedasticity. These shapes suggest that errors variances are small at lower values of the independent variable and become larger as the values of the independent variable become larger. Running separate regressions for each independent variable and obtaining residual plots to examine their shapes can help to detect the suspicious independent variables responsible for heteroscedasticity. In addition to this graphical method the data set can be split into two subsets, then after running a separate regression for each subset, error variances for the two subsets are compared. If the two variances differ heteroscedasticity exists. There are also three more precise methods for detecting the existence of heteroscedasticity. Those are the Park Test, the Goldfeld-Quandt Test, and the Breusch-Pagan Test. There is no need nor space in this short study to describe those tests in detail. Suffice to say that the first test runs OLS, square the residuals, take their logs and those of suspect regressors, then in a model makes the former a function of the latter, and finally if the t-ratio is significant the null of homoscedasticity is rejected. The second test sorts the data by magnitude, removes about one fifth of the sample size from the middle, run two separate regressions for the remaining two subsets of data then obtain the ratio ESS_2 / ESS_1 , compare it to the tabular F for hypothesis testing; the null of homoscedasticity is rejected if the ESS ratio exceeds tabular F. The third test runs OLS, square the standardized residuals and make this variance a function of the independent variables, run OLS on this model and obtain the ratio RSS/2 which follows the χ^2 distribution, then test the null hypothesis of homoscedasticity using the Chi-square table.

SAS IN TAMING VIOLATIONS

It is important to note that violations' remedies, like medicine, have side effects. A stabilizing method like the square root can improve the serial correlation situation for a certain model at the expense of worsening the heteroscedasticity situation. In other words remedies involve trade-offs. Compromises have to be made according to the extent of multicollinearity, serial correlation, and of heteroscedasticity that can be tolerated for

a certain model. It is also worth noting that the application of remedies loosens the tightness of model fitness. R-squares become lower after differencing to reduce variance inflation caused by multicollinearity.

Appendix "A" contains the original data set followed by three panels. The first panel is for multicollinearity selected outputs, the second for serial correlation, and the third is for heteroscedasticity. Computer outputs of each panel contain parameters estimates and statistics before and after the application of remedies, prediction plots and residual plots. Consumption (CONS) has been selected as the dependent variable, and the following were chosen as independent variables because of their generally known influence on consumption in macroeconomics literature. This choice of variables for the purpose of model specification has been verified by SAS PROC STEPWISE. Correct model specification is a requirement of the first classical assumption (A). Those variables are: nominal gross domestic product (NGDP), the monetary base (BASE), employment (EMPL), investment (NVST), exports (XPRT), imports (MPRT), and GDP implicit price deflator (PRCE). According to SAS PROC STEPWISE, all those variables are significant at the 0.1500 level. Other variables have been used to create classical violations in order to provide for their remedies.

Treating Multicollinearity

In Panel 1, a comparison of the Variance Inflation Factor (VIF), R-square, and the T ratios, before and after differencing shows that VIF for NGDP has been reduced from 16275.8 to 51.8, and from 11439.7 to 20.5 for disposable income (DINC) or (DPNC), the symbol after differencing. The R-square was reduced from 1.0000 to .9999 . And among the T for Ho, it fell from 109.6 for NGDP to 87.4 and from 33.1 to 31.2 for imports.

As for the serial correlation situation, the Durbin-Watson d statistic was 2.073 with a negative first order autocorrelation of .041 before differencing and d= 2.506 with RHO= - .265 . The change in d and in rho is due to the side effects of differencing. The residual plot showed a slight indication of a faint sign of heteroscedasticity before differencing which has become clearer after differencing. This is another side effect. A second differencing (DIF2) was unable to deflate neither the VIF nor the T ratios. On the contrary, it has increased these quantities (desired) and also increased the R-square (undesired). For economic and business data there is no need to go beyond a second differencing. Given such situation, the only way to remedy what remains of multicollinearity is to increase the sample size because multicollinearity is a sample phenomenon not a population phenomenon.

Treating Autocorrelation

While generalized least squares method (GLS), is usually recommended for dealing with autocorrelation, yet there is a warning that with small samples, rho cannot be accurately estimated GLS may be inferior to OLS whose parameter estimates are inefficient but not biased. SAS AUTOREG procedure estimates parameters in regression time series models whose standard error estimates are biased. Parameter estimates produced by PROC AUTOREG are usually similar to OLS estimates, but the standard errors can be very different and therefore can distort significance tests. OLS estimates are shown in panel 2. It can be noted that AUTOREG procedure has reduced the SSE from 17.85 to 13.75, the MSE from .616 to .509, and the Root MSE from .785 to .714 (rounded to three decimal places). If our sample was for less than 30 years we could have simply ignored autocorrelation.

Treating Heteroscedasticity

In a smaller model where CONS is the dependent variable and NGDP, XPRT, MPRT, GPRT , and NVST are the independent variables the plot of residuals showed a fanning out pattern which suggest the existence of heteroscedasticity thus leading to inefficient parameter estimates though not biased. Panel 3 shows PROC AUTOREG output of parameter estimates, of OLS then Yule-Walker estimates. We note from the AUTOREG output that the DW d statistic is 2.02 indicating the absence of serial correlation. Weighted least squares (WLS), is the suggested method to deal with heteroscedasticity. The data is divided by a weight which is a function of the suspect independent variable. Another method suggested by SAS ETS is a standardized least squares in which the data is divided by its own individual standard deviation, obtained here using PROC MEANS. Parameter estimates are shown below the previous two outputs in Panel 3. Comparing the Root Mean Square Errors for the three situations, we find that it is .7846 for OLS, .7136 for Autoreg, and .0006 after performing the standardized least squares. We can conclude that if standardized least squares is applied to a case of heteroscedasticity, parameter estimates will be BLUE.

References:

Brown, William. Introducing Econometrics. 1991. St. Paul, West Publishing Company.

Gujarati, Damodar. Basic Econometrics. Second edition, McGraw-Hill, 1988.

SAS Institute Inc. SAS System for Elementary Statistical Analysis
1987.

SAS Institute Inc., SAS/ETS User's Guide. 1988 Version 6, First Edition,
Cary, NC: SAS Institute Inc., 1988.

SAS Institute Inc., SAS/ETS Software: Applications Guide 1, Version 6, First
Edition, 1991.

SAS Institute Inc., SAS/STAT User's Guide, Vols I and II, 1990 Version 6,
Fourth Edition. Cary, N.C.: SAS Institute Inc.