Economics Letters 36 (1991) 49–54
North-Holland

A note on computing *r*-squared and adjusted *r*-squared for trending and seasonal data

Jeffrey M. Wooldridge

Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received 8 November 1990
Accepted 7 January 1991

Modified *r*-squareds are offered to overcome the deficiencies of the usual and adjusted *r*-squareds in linear models with trending and seasonal data. These modified measures are shown to be consistent for the population *r*-squared when the data contain deterministic trends in the mean, or deterministic seasonal components in the mean, or both.

1. Introduction

A common feature of time series regressions using levels or log-levels is that the usual and adjusted *r*-squared tend to be large relative to how well one thinks the regressors actually explain the regressand. If the regressand y_g is an integrated process then a large *r*-squared might be attributed to the spurious regression problem, as in Granger and Newbold (1974) and Phillips (1986). However, spurious regression is not an issue if the regressand and regressors are trend-stationary. The question, then, is why are the *r*-squareds from regressions with trend-stationary data so large? In this paper I simply point out that, for trend-stationary data, the usual and adjusted *r*-squareds are inconsistent estimates of the population *r*-squareds, and I offer consistent estimates that can be obtained by very simple adjustments. I also consider adjustments when the data contain deterministic seasonal components.

2. Computing *r*-squared when the regressand is trending

To motivate the modified *r*-squareds for trending data, first recall the case where $\{(x_t, y_t): t = 1, 2, \ldots\}$ is strictly stationary sequence with finite second moments, where x_t is a $1 \times K$ vector and y_t is a scalar. The vector x_t can contain lagged values of y as well as other variables. Then the population *r*-squared is defined by

$$\rho^2 \equiv \frac{V[\mathrm{E}(y_t \mid x_t)]}{V(y_t)} = 1 - \frac{\mathrm{E}[V(y_t \mid x_t)]}{V(y_t)}, \tag{2.1}$$

where $E(y_t \mid x_t)$ is the expected value of y_t given x_t and $V(y_t \mid x_t)$ is the variance of y_t given x_t [see also Greene (1990, pp. 73–74]. In the context of the linear model

$$y_t = \alpha + x_t\beta + u_t, \qquad E(u_t \mid x_t) = 0, \tag{2.2}$$

ρ^2 can also be expressed as

$$\rho^2 = 1 - V(u_t)/V(y_t). \tag{2.3}$$

Note that this formula is valid whether or not $V(u_t \mid x_t)$ is constant and whether or not (u_t) is serially correlated. ρ^2 measures the amount of variation in y_t that is explained by x_t [or more precisely, explained by $E(y_t \mid x_t)$], regardless of whether here is serial correlation and/or heteroskedasticity in $\{u_t\}$.

With stationary data a consistent estimator of ρ^2 is the usual r-squared:

$$R^2 \equiv 1 - SSR/SST, \tag{2.4}$$

where SSR is the usual sum of square residuals and SST is the total sum of squares from the regression

$$y_t \text{ on } 1,\ x_t, \quad t = 1, \ldots, T. \tag{2.5}$$

Theil's adjusted r-squared, denoted $\bar{R}^2$, uses degrees of freedom adjustments in the numerator and denominator. With x_t $1 \times K$,

$$\bar{R}^2 \equiv 1 - \frac{SSR/(T-K-1)}{SST/(T-1)}. \tag{2.6}$$

$\bar{R}^2$ is also a consistent estimator of ρ^2.

When y_t is trend-stationary with a linear trend in mean, $E(y_t) = a_0 + a_1 t$, then $\eta_t \equiv y_t - E(y_t)$ is stationary and $E[V(y_t \mid x_t)]$ and $V(y_t)$ continue to be time invariant. Consequently, the population r-squared given by (2.1) is still well-defined and time invariant. However, neither R^2 or $\bar{R}^2$ consistently estimates ρ^2. The problem is that the usual total sum of squares,

$$SST \equiv \sum_{t=1}^{T} (y_t - \bar{y})^2,$$

when divided by either T or $T-1$ does not produce a consistent estimator of $V(y_t)$. This is simply because $E(y_t)$ is not constant, so that the sample average $\bar{y}$ is an inappropriate estimator of $E(y_t)$.

Instead, a consistent estimator of $V(y_t)$ is obtained by first computing the sum of squared residuals from the 'detrending' regression

$$y_t \text{ on } 1,\ t, \quad t = 1, \ldots, T. \tag{2.7}$$

Using the representation $y_t = E(y_t) + \eta_t$, denote the residuals from (2.7) by $\{\hat{\eta}_t : t = 1, 2, \ldots, T\}$ and the corresponding sum of squared residuals from (2.7) by SSR_η. Then a consistent estimator of $V(y_t)$ is simply SSR_η/T, while a degrees-of-freedom adjusted estimator is $SSR_\eta/(T-2)$ [two trend

parameters are estimated in (2.7)]. In the trend-stationary case, the sum of squared residuals used in estimating $E[V(y_t \mid x_t)]$, SSR, is obtained as usual from the regression

$$y_t \text{ on } 1,\ t,\ x_t, \quad t = 1, \ldots, T. \tag{2.8}$$

From regressions (2.7) and (2.8) we obtain a consistent estimator of ρ^2 when $E(y_t)$ is linear in time:

$$R_\eta^2 \equiv 1 - SSR/SSR_\eta. \tag{2.9}$$

The corresponding adjusted r-squared is

$$\bar{R}_\eta^2 - 1 - \frac{SSR/(T-K-2)}{SSR_\eta/(T-2)} = 1 - \frac{\hat{\sigma}_u^2}{\hat{\sigma}_\eta^2}, \tag{2.10}$$

where $\hat{\sigma}_u$ is the usual standard error of the regression (2.8) and $\hat{\sigma}_\eta$ is the standard error of the regression (2.7). Note that $R_\eta^2 \leq R^2$ but $\bar{R}_\eta^2$ need not be smaller than R^2 [because $SSR_\eta/(T-2)$ can be larger than $SST/(T-1)$].

There are some additional observations worth making about definitions (2.9) and (2.10). First, R_η^2 measures the amount of variation in $y_t - E(y_t 0$ explained by x_t, or equivalently by $x_t - E(x_t)$. R_η^2 is more relevant than R^2 as a goodness of fit measure because it nets out the trending nature of y_t. Second, if $E(y_t)$ is constant, the usual r-squared from (2.8) can be used whether or not x_t is trending because it consistently estimates ρ^2 in this case. Third, R_η^2 is numerically equal to the usual r-squared from the regression

$$\hat{\eta}_t \text{ on } 1,\ t,\ x_t, \quad t = 1, \ldots, T. \tag{2.11}$$

The adjusted measure $\bar{R}_\eta^2$ is most easily obtained from $\hat{\sigma}_u^2$ and $\hat{\sigma}_\eta^2$.

Suppose more generally that $E(y_t)$ is a Pth y-degree polynomial in t,

$$E(y_t) = a_0 + a_1 t + \ldots + a_p t^P, \tag{2.12}$$

and the model relating y_t and x_t is

$$y_t = \alpha + \gamma_1 t + \ldots + \gamma_Q t^Q + x_t \beta + u_t. \tag{2.13}$$

Generally, if an element of x_t has a polynomial trend with order higher than than P, then Q must be greater than P for the errors $\{u_t\}$ to the stationary with $E(u_t \mid x_t) = 0$. Under (2.12) and (2.13), SSR and $\hat{\sigma}_u^2$ are obtained from

$$y_t \text{ on } 1,\ t,\ t^2, \ldots, t^Q,\ x_t, \quad t = 1, \ldots, T, \tag{2.14}$$

and SSR_η and $\hat{\sigma}_\eta^2$ are obtained from

$$y_t \text{ on } 1,\ t,\ t^2, \ldots, t^P, \quad t = 1, \ldots, T. \tag{2.15}$$

The formula for R_η^2 is still given by (2.9), and $\hat{R}_\eta^2$ is given by (2.10). (Note that $\hat{\sigma}_u^2$ and $\hat{\sigma}_\eta^2$ always incorporate the appropriate degrees-of-freedom adjustments.) Again, it is the trend properties of y_t

alone that determine what appears in the denominator of R_η^2 and $\bar{R}_\eta^2$. However, the trend properties of x_t can be important for deciding on the value of Q in (2.13).

Letting $\{\hat{\eta}_t : t = 1, \ldots, T\}$ denote the residuals from (2.15) (the detrended y_t). R_η^2 is easily computed as the usual r-squared from the regression

$$\hat{\eta}_t \text{ on } 1,\ t,\ t^2, \ldots, t^Q,\ x_t, \quad t = 1, \ldots, T. \tag{2.16}$$

Even in cases where Q is less than P, e.g. $Q = 0$ and $P = 1$, the relevant r-squared measures are given by (2.9) and (2.10). Although the researcher is i mplicitly assuming that the trend and stochastic parts of y_t and x_t are related in exactly the same way, an estimate of the expected value of y_t must be subtracted off in estimating $V(y_t)$. This is achieved in the detrending regression (2.15).

3. Seasonal and trending data

Similar considerations arise with seasonal data. For example, suppose the y_t and x_t are monthly series with deterministic trends and deterministic seasonal components:

$$y_t = a_1 t + a_2 t^2 + \ldots + a_P t^P + c_1 \text{jan}_t + \ldots + c_{12}\text{dec}_t + \eta_t, \tag{3.1}$$

$$x_t = b_1 t + b_2 t^2 + \ldots + b_M t^m + d_1 \text{jan}_t + \ldots + d_{12}\text{dec}_t + \nu_t, \tag{3.2}$$

where $\{\eta_t\}$ is zero mean stationary time series and $\{\nu_t\}$ is a $1 \times K$ stationary sequence with zero mean. A general model is

$$y_t = \alpha + \gamma_1 t + \ldots + \gamma_Q t^Q + \delta_2 \text{feb}_t + \ldots + \delta_{12}\text{dec}_t + x_t\beta + u_t, \tag{3.3}$$

where, without further information, $Q = \max(P, M)$. The relevant OLS regression is

$$y_t \text{ on } 1,\ t, \ldots, t^Q, \text{feb}_t, \ldots, \text{dec}_t,\ x_t, \quad t = 1, \ldots, T. \tag{3.4}$$

Let SSR and $\hat{\sigma}_u$ be the sum of squared residuals and standard error of the regression from (3.4), and let SSR_η and $\hat{\sigma}_\eta$ be the corresponding quantities from

$$y_t \text{ on } 1,\ t, \ldots, t^P, \text{feb}_t, \ldots, \text{dec}_t, \quad t = 1, \ldots, T. \tag{3.5}$$

Then define R_η^2 as in (2.9) and $\bar{R}_\eta^2$ as in (2.10). These quantities measure the amount of variation in the detrended/deseasonalized y_t that is explained by x_t. The extension to data with frequencies other than monthly is immediate.

4. Empirical example

Data on real housing investment, population, and a housing price index for the U.S. are reported in McFadden (1990). Consider estimating a simple aggregate supply function for per capita housing

investment, using annual data for 1947–1988 (a total of 42 observations). A simple regression on a linear time trend yields

$$\begin{array}{lll} \log(inv_t/pop_t) = & -0.8413 + & 0.0081\ t + \hat{u}_t, \\ & (0.0447) & (0.0018) \\ & [0.0390] & [0.0024] \end{array} \tag{4.1}$$

$$R^2 = 0.3354; \quad R_\eta^2 = 0.0; \quad \bar{R}^2 = 0.3188; \quad \bar{R}_\eta^2 = 0.0; \quad SSR = 0.8112.$$

The quantities in $(\cdot)$ are the usual OLS standard errors, and those in $[\cdot]$ are heteroskedasticity/serial correlation robust standard errors computed as in Wooldridge (1989), allowing for two non-zero autocovariances. The robust t-statistic on the trend term is 3.38, suggesting that $\log(inv_t/pop_t)$ has a growing mean. By definition, the measures R_η^2 and $\bar{R}_\eta^2$ are zero in (4.1), just as R^2 and $\bar{R}^2$ are zero in a regression on a constant only.

The estimated supply function including a time trend and log price is

$$\begin{array}{lll} \log(inv_t/pop_t) = & -0.9131 + & 0.0098 - 0.3910\ \log(price_t) + \hat{u}_t, \\ & (0.1356) & (0.0035)\ \ (0.6788) \\ & [0.2003] & [0.0039]\ \ [0.9517] \end{array} \tag{4.2}$$

$$R^2 = 0.3408; \quad R_\eta^2 = 0.0080; \quad \bar{R}^2 = 0.3070; \quad \bar{R}_\eta^2 = -0.0174; \quad SSR = 0.8047.$$

Note that R_η^2 and $\bar{R}_\eta^2$ are much lower than their unmodified counterparts, reflecting that price does not explain much of the variation in $\log(inv_t/pop_t)$ about its mean. This is not surprising given the unexpected negative sign on $\log(price_t)$ and the insignificant t-statistic. However, even when right hand side variables are statistically significant, it is important to compute R_η^2 and $\bar{R}_\eta^2$. As an illustration, suppose that time is omitted from (4.2). This yields

$$\begin{array}{lll} \log(inv_t/pop_t) = & -0.5502 + & 1.2409\ \log(price_t) + \hat{u}_t, \\ & (0.0430) & (0.3824) \\ & [0.0839] & [0.6534] \end{array} \tag{4.3}$$

$$R^2 = 0.2084; \quad R_\eta^2 = 0.0022; \quad \bar{R}^2 = 0.1886; \quad \bar{R}_\eta^2 = -0.1912; \quad SSR = 0.9663.$$

Now $\log(price_t)$ has the economically correct sign and a robust t-statistic of about 1.9. But the very low R_η^2 and $\bar{R}_\eta^2$ from (4.3) reveal a problem, namely the omitted time trend. This is evident even without estimating (4.2) directly.

5. Concluding remarks

The modified r-squareds suggested in this paper are appropriate when the data have deterministic trends, deterministic seasonal components, or both. I covered the case of polynomial trends explicity, but other functions of time are straightforward to handle. The assumption of strict stationarity about the mean was made for convenience only. These measures are also useful for trending, heterogeneous processes, provided that the variances are uniformly bounded. Unfortunately, R_η^2 and $\bar{R}_\eta^2$ have no theoretical justification if y_t is integrated (note that x_t integrated does not cause problems). When y_t is integrated its variance is not bounded, and the modified total variation no longer consistently estimates $V(y_t)$. It is not clear how one should define goodness of fit when y_t is integrated.

References

Granger, C.W.J. and P. Newbold, 1974, Spurious regressions in econometrics, Journal of Econometrics 2, 111–120.

Greene, W.H., 1990, Econometric analysis (MacMillan, New York).

McFadden, D., 1990, Demographics, the housing market, and the welfare of the elderly, Mimeo. (Department of Economics, MIT, Cambridge, MA)

Phillips, P.C.B., 1986, Understanding spurious regressions in econometrics, Journal of Econometrics 33, 311–340.

Wooldridge, J.M., 1989, A computationally simple heteroskedasticity and serialo correlation robust standard error for the linear regression model, Economics Letters 31, 239–243.