26th Australasian Transport Research Forum
Wellington

USING PANEL DATA TO EVALUATE ROAD SAFETY COUNTERMEASURES

Nigel Rockliffe, Director, Economic Outcomes Pty Ltd

James Symons, Emeritus Reader in Economics, University College London

Dimitris Tsolakis, Coordinator, Economic Analysis, ARRB Transport Research

ABSTRACT

This paper studies the effectiveness of the Queensland School Transport Safety Program (SafeST). Our main general conclusion is that good statistical practice can improve the accuracy of forecasts of the effects of a countermeasure by an *order of magnitude*. Conversely, poor practice, in particular an inappropriate model specification coupled with a small data sample can produce imprecision. In particular, we suggest that unobserved demographics are generally best modelled by stochastic trends, as these allow for the demographic effect to wander away from a linear trend if this improves the fit. We show how the model can be estimated by GLS in these circumstances and demonstrate that efficiency gains can result.

1. INTRODUCTION

This paper describes our experience of evaluating the effectiveness of a road safety countermeasure, Queensland Transport's *SafeST Program*.[1] The data to be analysed present themselves as a *panel*, that is, the countermeasure acts through time at a number of different sites.

In the course of this work we have developed better statistical practice that will, in many instances, enable the analyst to gauge the true impact of countermeasures, and hence allocate scarce resources more efficiently. Our work is capable of wide application. Road safety countermeasures are notoriously hard to evaluate because their effects tend to be lost in the background 'noise' of road crashes. As a result, it is easy for resources to be devoted to countermeasures that offer little or no benefit (indeed, some may be counterproductive).

Our work demonstrates that it is possible to achieve gains in accuracy (that is, in econometric efficiency) from three sources: choosing the appropriate estimation procedure, parsimony in model selection, and including the intensity of the countermeasure as an explanatory variable.

1.1 CHOOSING THE RIGHT ESTIMATION PROCEDURE

Firstly, use of an appropriate estimation procedure can increase the efficiency of estimates by a considerable margin. By 'appropriate' here we mean a procedure that

1

[1] The SafeST Program aims to improve road safety for school children through improvements to pedestrian facilities, passenger set-down and pick-up areas, cycling paths and traffic management in the vicinity of schools. It is managed by Queensland Transport (QT) in partnership with the Queensland Department of Main Roads (DMR) and local authorities. Since it began in 1996, a total of over $30 million has been spent on over 600 SafeST projects affecting over 400 schools throughout Queensland. Annual spending is currently about $3.5 million.

takes into account the error structure of the underlying model. One might have information on this structure from theoretical considerations or from plausible assumptions about how the data arise. As an example of the former, the assumption that casualty data follow a Poisson process gives specific information about the magnitude of the realisation errors in observed data. As an example of the latter, unobserved demographics in the model can give rise to an error structure which, if allowed for in estimation, can yield useful efficiency gains.

1.2 PARSIMONY IN MODEL SELECTION

The second source of efficiency gains arises from parsimony in model selection. The amount of data to be exploited imposes limits on the number of variables one can include in the model without making the estimates of the parameter of interest hopelessly diffuse. For example, the policy variable to be analysed in this study is quite strongly trended. If one seeks to allow for unobserved demographics by the inclusion of linear trends, the correlation between these and the policy variable makes estimates of the policy parameter quite imprecise. Note the trade-off between bias (as would arise if trends were omitted when, in fact, they are genuinely present) and efficiency.

One possibility is to seek parsimony by testing for the exclusion of variables one suspects are causing problems; however, reductions in estimated standard errors obtained by this procedure will be partly illusory because of the ‘pre-testing’ problem. Another possibility is to transform the data so as to make unobserved demographics less of a problem. Below we shall adopt as dependent variable casualties at treated sites *relative to casualties at similar untreated sites*. This device may reduce or eliminate the need to model unobserved demographics (as well as offering other related econometric benefits).

1.3 COUNTERMEASURE INTENSITY

The third source of efficiency gains we consider here arises from measurement of countermeasure intensity. It will often be the case that a countermeasure acts with measurably different intensities at different sites. If the underlying intensity-response relationship is estimated, then the average response calculated from this will be more efficient than a procedure which acknowledges only the treated-untreated dichotomy: reduction in standard errors should be proportional to the coefficient of variation of treatment intensities.

1.4 STRUCTURE OF THIS PAPER

The paper is organised as follows. Section 2 contains the most novel material in the paper: in it we present our empirical results and show that unobserved demographics in the model are best modelled by the statistical process known as a ‘random walk’. In section 3 we use simulation methods to corroborate and amplify the conclusions from section 2. Section 4 concludes.

A technical appendix presents the statistical framework of the policy analysis, describes the model to be estimated, and analyses the statistical properties of the estimators we consider, particularly their efficiency. It also details the Monte Carlo simulation of estimation procedures referred to above.

2. EVALUATION OF THE SAFEST PROGRAM

In this section we describe our analyses of crash data relevant to the SafeST program. The *Technical appendix* details the analytical method.

2.1 DEPENDENT VARIABLE

In our analysis, the dependent variable for the analysis is formed from all pedestrian casualties occurring within one kilometre of Queensland schools between 1992 and 2001, disaggregated by treated and control (non-treated) schools (Figure 1). For analytical purposes, these were further disaggregated by 11 Statistical Divisions (SD).[2] We thus study 22 time-series—one treated and one control for each SD. The dependent variable is the difference between casualties per capita at treated and untreated sites. We thus have 110 annual[3] observations of the dependent variable (that is, 11 SDs for 10 years). The unit of measurement for casualties was chosen to be average per capita casualties at treated sites, 1992-96.

2.2 INDEPENDENT VARIABLES

We considered two variants of the forcing variable (that is, the variable that measures the effect of SafeST treatments in 'forcing' down the number of crashes). The first was a simple dummy variable for the years the SafeST projects were in place, 1997-2001. This is less than satisfactory given that projects are of different types and intensities at different sites.

Our second variable was cumulated expenditure per capita at treated sites. The unit of measurement was chosen to be average cumulated expenditure per capita over all treated sites, 1997-2001. This was done to ensure rough comparability between the parameter on the dummy variable (the average effect of SafeST treatment) and the parameter on the expenditure term (the effect on casualties of an average SafeST expenditure).

We also conducted some experiments allowing for a site-specific time trend in the estimating equation.

2.3 METHOD

Our data form a panel of 11 regions and 10 years. This could be estimated by ordinary least-squares regression (OLS). But because regional populations differ substantially, it can be shown (see *appendix section 3*) that weighted least-squares estimation (WLS) is appropriate. In support of this, we show by simulation (see *appendix section 4)* that WLS reduces standard errors by a factor of about two as compared to OLS. We can also allow for stochastic trends in the underlying

1

[2] Although a geographic smaller unit was perhaps desirable (*appendix section 3*) this level of aggregation was forced on us by data availability. This explains why the data were aggregated by SD, whereas in our simulation study, which was not data-constrained, we used a smaller geographic unit, the SSD. Since resources were unavailable to estimate population at treated and non-treated sites, SD populations are observed at the census years, but interpolated between census years and apportioned between treated and untreated sites in the proportion of average casualties at treated and untreated sites observed between 1992 and 1996 in each SD.

[3] There are small efficiency gains to be had from using monthly data but we did not regard them to be sufficiently large to warrant the added complexity of a monthly analysis.

demographic data by means of generalised least-squares (GLS) estimation (see *appendix section 5*).

2.4 RESULTS OF THE EVALUATION

We have analysed the data by means of six methods that differ in the definition of the forcing variable (continuous or dichotomous), in estimation method (OLS, WLS or GLS), and by type of trend (if any) (Table 1 and Figure 2).

Table 1 Estimating the effect of SafeST by different methods

Model	Forcing variable	Estimation Method	Trend	Estimate	Std Error
1	Expenditure per capita	WLS	No	-0.030	0.028
2	Expenditure per capita	OLS	No	-0.125	0.069
3	Intervention dummy	WLS	No	-0.048	0.032
4	Intervention dummy	OLS	No	-0.070	0.166
5	Intervention dummy	WLS	Yes	-0.149	0.095
6	Expenditure per capita	GLS	Stochastic	-0-023	0.015

Source: ARRB TR calculations. Note: None significant at the 5% level.

Model (1) has expenditure as the forcing variable, uses WLS, and does not fit a trend. The estimate of the expenditure effect is –0.030. This can be interpreted as the proportional reduction in casualties at treated sites arising from having in place an average SafeST level of per capita expenditure. The estimate is not significant at the 5% level, but the standard error is small. The implication is that this model *would have been capable* of detecting the impact of SafeST had SafeST been more effective. As it is, we conclude that SafeST is probably either ineffective or nearly so.

Models (1) to (4) demonstrate the efficiency gains of WLS over OLS, as supported by simulation using artificial data.

Comparing models (1) and (3) shows that using expenditure in place of the intervention dummy reduces standard errors by a small amount.

In model (5) we allow each site to have its own trend. The standard error increases by a large amount in this experiment, in line with the simulation results (*section 3*). The residual sum of squares is virtually the same between (1) and (5), so an *F*-test for the exclusion of trend variables is easily passed. We interpret this to mean that trends are unnecessary in the model, serving only to increase the standard error of the parameter of interest.

Finally in model (6) we allow for stochastic trends. The parameter estimate is not large, - 0.023, but the standard error is very small: the probability level of the estimate is 0.112, approaching significance at the 5% level. In terms of precision this is the best performing model, and demonstrates the advantage of using stochastic trends in model specification.

None of the 95% confidence intervals associated with these estimates excludes zero, but there is consistency in the results. Models (1) and (6) are most favoured by our theoretical analysis; both suggest SafeST reduces casualties by less than 10%—a valuable, if disappointing, conclusion.

3. SIMULATION STUDY

In this section we demonstrate the pros and cons of alternative model specifications by means of a simulation study. The *technical appendix* details the analytical method.

3.1 STUDY DESIGN AND DATA GENERATION

For simulation purposes we considered Queensland as consisting of 27 regions, notionally one for each Statistical Subdivision.[4] Each region is considered to be subjected to a treatment at the end of year 5, and this treatment is in force for the subsequent five years. For each region there is a corresponding region which is never treated (the control). Thus we have five years of untreated data (both treated sites and controls) followed by five years of treatments at the treated sites.

Crashes are generated by a Poisson process whose mean falls by 10% after the treatment at the end of year 5. A particular experiment consists of generating ten years of annual data of random variables, wherein the mean of the random variable falls by 10% at the treated sites at the end of the year 5, remaining constant elsewhere (and hitherto). There are no trends in mean.

These data describe a world in which, in truth, the treatment reduces casualties by 10% at each treated region. This known effect is then estimated by a variety of estimation techniques. The results from a number of such experiments allow comparison of their effectiveness.

3.2 ESTIMATION METHODS

We argued above that linear models offer some advantages. In this approach, the dependent variable is the *difference* between casualty rates in the treated and control sites. This is then modelled as linear function of a site-specific trend and a single dummy variable on the intervention periods.

This can be estimated by OLS, but the assumption that casualty data follow a Poisson distribution implies that the variance of the residuals will be inversely proportional to the population of the site. In this case, a version of WLS will be superior to OLS.

Finally, fitting a constant and trend will reduce efficiency if these variables are not needed to avoid misspecification. By construction, the artificial data have no constant or trend so it is possible to demonstrate by simulation the effect of their superfluous inclusion.

3.3 RESULTS OF THE SIMULATION

We have computed the mean and standard deviation from 1000 experiments (Table 2 and Figure 3). For the first three procedures, the 95% confidence intervals associated with the typical estimate all include zero. In particular, one would not to reject at the 5%level the null hypothesis that the policy had no effect on casualties. In contrast, the confidence interval for the fourth procedure excludes zero and is approaching a satisfactory level of precision. This indicates a substantial return to the employment of an efficient procedure.

1

[4] See note 2

In sum, Monte Carlo simulation analysis of the linear model suggests it performs well. Bias is small and the method seems robust to technical assumptions. The estimates of our linear models have smaller standard errors once heteroskedasticity is allowed for. WLS tends to halve the error variance; and removing the constant and trend reduces the error variance to about a third.

Table 2 Simulated estimates of treatment effect of 10% reduction in casualties

Estimated Method	Mean	Std Dev
Linear OLS with trend	- 0.103	0.145
Linear WLS with trend	- 0.102	0.073
Linear OLS no trend	- 0.100	0.051
Linear WLS no trend	- 0.101	0.025

Source: ARRB TR calculations.

4. CONCLUSIONS

This paper studies the effectiveness of the Queensland School Transport Safety Program (SafeST). Our main conclusion is that good statistical practice can improve the accuracy of forecasts of the effects of a countermeasure by an *order of magnitude* as compared with a more naïve procedure. Conversely, poor practice, in particular an inappropriate model specification coupled with a small data sample, can produce imprecision. In particular, we suggest that unobserved demographics are generally best modelled by stochastic trends, and we show how this can be done through the appropriate choice of estimation method.

The consequences of our findings are potentially far-reaching. Road safety countermeasures are notoriously hard to evaluate because their effects tend to be lost in the background ‘noise’ of road crashes. As a result, it is easy for resources to be devoted to countermeasures that offer little or no benefit (indeed, some may be counterproductive). In this paper we offer better statistical practice that will, in many instances, enable the analyst to gauge the true impact of countermeasures, and hence allocate scarce resources more efficiently.

TECHNICAL APPENDIX

1. MODEL SET-UP

GENERAL STATISTICAL FRAMEWORK

Let $z = z_{it}$ be casualties in region i at time t. We assume two variables act causally on z: d, which we shall call the policy variable or the variable of interest, and r, a collection of extraneous variables. A new countermeasure[5] is envisaged which will change the values of d; we wish to estimate the expected value of z in the new regime. We assume in the original regime the expected value of z conditional on d and r is linear:

(1) $$E(z \mid d, r) = bd + gr$$

The implicit assumption is that the relationship (1) remains the same in the new regime, and hence may be used to estimate the value of z in the new regime. For this assumption to be remotely plausible, it is necessary that r should include all causal variables that are correlated with d in the old regime, as otherwise the conditional expectation (1) would be likely to change over the regime change.

If it is assumed that the countermeasure will leave r unchanged, then the parameter g need not be calculated, since the object of interest is the difference between z in the two regimes, and r is common. In these circumstances one might describe r as a *nuisance variable.*

We write

(2) $$z = bd + gr + e$$

where

(3) $$e = z - E(z \mid d, r)$$

is the so-called regression error.

CASUALTY MODEL

Specifically we shall assume (2) takes the form

(4) $$z_{it} / P_{it} = bd_{it} + gr_{it} + e_{it}$$

1

[5] Note that the 'new countermeasure' will often be the absence of an existing countermeasure. Thus, to evaluate a countermeasure that has been active over some period, one will consider what would have happened if the countermeasure had not been implemented. The old regime is the countermeasure, the new regime is the absence of it.

where P_{it} is a measure of the appropriate population in i at t. This relationship is observed in the pre-change regime and the observed parameters are to be used to evaluate the new regime according to (1).

CONTROL REGIONS

The most obvious candidates for extraneous variables r are demographic trends, capturing perhaps changing propensities for per-capita casualties as the population structure changes, and state or national trends, arising perhaps from changing attitudes to safety (themselves possibly a response to public policy). As noted above, if these factors are not included as conditioning variables, it is unlikely that (1) will remain stable over the regime change if the policy variable itself is trended (as will be the case for an countermeasure that changes d after some date). It is inviting to measure these conditioning variables by deterministic linear trends but there is no guarantee that such variables will remove all of the correlation between the unobserved true trends and the policy variable.

A large part of this problem is evaded by the use of control regions. Assume for each region i there is another region i' where the countermeasure acts differently and let Δd etc. be the difference between d in i and i'. Then (4) becomes

(5) $$\Delta(z_{it} / P_{it}) = b\Delta d_{it} + g\Delta r_{it} + \Delta e_{it}$$

and may be used to infer the parameter of interest. If region and control are selected to be similar in the relevant characteristics, it is much more plausible to measure Δr by linear trends, or to omit it entirely.

2. REGRESSION ESTIMATION OF THE CASUALTY MODEL

Assume we have a sample of data generated by (4) or (5) written in the form

(6) $$z = Xd + e$$

where z is now a column vector of observations on z_{it}, X is the matrix of observations on the causal variables, d is the column vector of required parameters, and e is a column vector of regression errors. The ordinary least squares (OLS) estimate of d is given by the matrix product

(7) $$\hat{d}_{OLS} = (X'X)^{-1} X'z$$

The OLS estimator is unbiased in the sense that its expectation, conditional on the full set of observations on the causal variables, is d

(8) $$E(\hat{d}_{OLS} \mid X) = d$$

provided that the errors have zero conditional mean

(9) $$E(e \mid X) = 0$$

which we shall assume is the case.

3. EFFICIENCY ISSUES

A particular requirement of this study was to obtain accurate estimates from limited data, that is, we require our estimates to have minimal sampling error. If in addition to (9) a model such as (6) has the property that the errors are independent with common variance, compactly written in matrix form as

$$(10) \qquad E(ee' \mid X) = s^{2} I$$

then the Gauss-Markov theorem tells us that OLS is best linear unbiased (BLUE), that is, it has the least variance of all unbiased estimators that are linear functions of the data z. Thus if (10) holds, OLS would be just about the best one could do. As it happens, however, (10) is unlikely to hold for casualty data.

CASUALTY DATA AND THE POISSON PROCESS

It is commonly assumed that casualty data are generated by Poisson random variables. This is attractive because the underlying assumption of the Poisson – that occurrences within small intervals of time are independent events of low probability – seems suited to these data. Assume therefore that casualty data are Poisson and generated by (4). Then, exploiting the fact that Poisson random variables have equal mean and variance, one obtains

$$(11) \qquad Var(e_{it}) = g\, r_{it} / P_{it}$$

(if $b = 0$). Thus the e_{it} do not have common variance if the population varies over sites, that is, the residuals are not homoskedastic. If so, OLS is not the most efficient estimator. When one employs control regions, (11) takes the form

$$(12) \qquad Var(\Delta e_{it}) = g\, r_{it} / P_{it} + g\, r_{i't} / P_{i't}$$

Thus, if population varies over sites, the relationship (10) will fail and the Gauss-Markov theorem will no longer guarantee efficiency.

THE EXTENDED GAUSS-MARKOV THEOREM

Consider the problem of estimating (10) for a general covariance matrix, known up to a scale parameter s^{2}

$$(13) \qquad E(ee' \mid X) = s^{2} \Omega$$

The generalised least squares estimator (GLS) is given by

$$(14) \qquad \hat{d}_{GLS} = \left(X' \Omega^{-1} X \right)^{-1} \left(X' \Omega^{-1} z \right)$$

The extended Gauss-Markov theorem says that this estimator is BLUE. Thus, if we are able to nominate a covariance structure for the error process, we can claim estimates derived from (14) are efficient – the best that can be done.

GLS can be interpreted as follows. The data vectors are transformed by the matrix square root $\Omega^{-1/2}$, that is, we calculate $\tilde{X} = \Omega^{-1/2} X$ etc. One can apply this transformation to (6) in which case $E(\tilde{e}\tilde{e}') = I$. The transformed residuals now satisfy (10) so that the ordinary Gauss-Markov theorem applies to the transformed equation and one can deduce that the OLS estimate

$$(15) \qquad \hat{d}_{GLS} = \left(\tilde{X}\tilde{X}\right)^{-1}\left(\tilde{X}\tilde{z}\right)$$

is BLUE. It can easily be verified that formulae (14) and (15) are identical.

'WEIGHTED LEAST SQUARES' FOR CASUALTY MODELS

How might the covariance matrix look for a model such as(5)? It is reasonable to assume the off-diagonal entries are zero: within a given region, the errors can be thought of as random realisation errors; across regions, estimating casualties relative to a control should eliminate correlations due to common influences such as the weather. The diagonal of the covariance matrix consists of the variances of the e_{it}. If, as a first approximation, average casualty rates are constant across regions, then (12) implies these variances are proportional to $1/P_{it} + 1/P_{i't}$. Thus the structure matrix Ω can be taken to have terms such as this down the diagonal, and to be zero elsewhere. This matrix can then be used in (14).

It turns out that this procedure amounts to weighting each observation by the quantity $\left(1/P_{it} + 1/P_{i't}\right)^{-0.5}$ and then estimating by OLS. This estimator is called weighted least squares (WLS). The WLS procedure is sometimes called 'adjusting for heteroskedasticity'.

INEFFICIENCY FROM SUPERFLUOUS TRENDS

Fitting superfluous variables to a model can lead to losses in efficiency, in particular if region-specific constant and trend are fitted to the model. The point of these extra two variables is to pick up any systematic divergence between treated site and control over the period (demographics, say). If the controls are well selected, this may not be necessary and there are substantial efficiency gains to excluding these variables when they are not required. The reason is that a linear trend is often somewhat correlated with an countermeasure variable, which makes it harder to estimate its effect with precision.

The upshot is that excluding the constant and particularly the trend is most desirable if it can be supported. Essentially, the gain in efficiency on offer here is a premium for careful selection of controls.

INEFFICIENCY ARISING FROM AGGREGATION

In this section we investigate whether, in seeking to obtain regression estimates of some parameter of interest, it is better to work with aggregated or disaggregated data. The data in this study form a panel: observations are indexed by time and place, and clearly one can aggregate along both dimensions. For example, one could decide to work with annual casualty data rather than monthly or daily data; equally one might decide to work with Queensland-wide data rather than SSD or SLA data. There is a duality between aggregating over time and aggregating over region, and the properties of temporal aggregation will usually be reflected in those of spatial aggregation and vice-versa.

We assume the data are given by the model

(16) $$z = bd + e$$

We shall assume these data are arranged as column vectors. We assume that A is an *aggregator matrix,* that is, its rows consist of 1s and 0s so that Az etc. is the variable z aggregated in the desired way. These rows are essentially dummy variables on the items to be aggregated. The aggregated equation takes the form

(17) $$Az = bAd + Ae$$

Assume the errors in (16) satisfy (10) so that OLS is BLUE. The errors in (16) will not in general satisfy (10) but applying optimal GLS to (17) gives

(18) $$\hat{b}_A = \frac{z'Q_A d}{d'Q_A d}$$

where $Q_A = A'(AA')^{-1}A$. The relative efficiency of OLS applied to (16) versus GLS applied to (17) is given by

(19) $$\frac{Var(\hat{b})}{Var(\hat{b}_A)} = \frac{d'Q_A d}{d'd}$$

Equation (18) shows that GLS applied to the aggregated equation is a linear function of the data z. The Gauss-Markov theorem then immediately implies that OLS applied to the disaggregated equation must be more efficient. The extent of the efficiency gain is given by the RHS of (19), which can be interpreted as the R^2 from a regression of d on the indicator dummies making up the rows of A. The fitted values from such a regression will consist of the values of d replaced by their averages over the aggregated values. The R^2 is unity if and only if the values to be aggregated are constant.

The upshot of this discussion is that it is harmless to aggregate over values for which the policy variable shows no variance. Thus if the countermeasure were identical between regions, it would do no harm to consider them as a single region; equally, if a countermeasure acted constantly within years, one might as well consider annual data.[6]

1

[6] As an example, assume one wishes to test the hypothesis that more people go to the beach on hot days, and assume *daily* data are available for beach attendance. It is clear that one could establish a good relationship between temperature and beach-going with the data of even a single year. Aggregating to annual data means that the purported relationship would be between annual beach attendance and annual average temperature; clearly this would be more difficult to establish. Estimating in annual data will not produce bias, but will produce more diffuse estimates of the temperature effect. The intuition is that, in general, averaging loses information: the temporal connection between cause and effect. Now hypothesise a relationship between beach-going and leap years. One could test this in data of either frequency by regressing attendance on a dummy variable for leap years. Since this dummy is constant within a given year (for any particular beach), the above principle tells us that estimating in annual data loses no efficiency. Efficiency would be lost, however, if the model were

EFFICIENCY GAINS FROM MEASURING COUNTERMEASURE INTENSITY

Consider a countermeasure that varies in intensity, and assume that the average response is all that is required for the policy analysis. An unbiased estimate of the countermeasure can be obtained from the average response in treated areas relative to untreated areas. Alternatively, if the intensity itself can be measured, then the response-intensity parameter can be obtained by the methods discussed above and the average response to the countermeasure obtained accordingly.

The proportional reduction in the error-variance obtained by the intensity-response regression will, in this simple framework, be the square of the coefficient of variation of the policy variable. Thus there are efficiency gains to be had by measurement of countermeasure intensity if these vary over time and by site. If for example a countermeasure involves capital expenditure, it may be valuable to define countermeasure intensity as cumulated expenditure per head. In some circumstances, of course, differences in intensity may be difficult to measure.

4. SIMULATION STUDY OF THE METHODS

We demonstrated some of these efficiency gains and losses in a simulation study.

STUDY DESIGN

The variable z_{it} is some measure of casualties in region i at time $t, i = 1,..., N, t = 1,...$. Region i is subjected to a treatment at time $T, T+1,...$ We assume for each i there is a corresponding region i' which is never treated (the control). We assume we have five years of untreated data (treated sites and controls) followed by five years of treatment at the treated sites. The regions are notionally the 27 Queensland SSDs. It is assumed that $E(z_{it})$ falls by 10% at the treatment.

DATA GENERATION

Our basic analysis assumes that the z_{it} are generated by a Poisson process whose mean falls by 10% after the treatment at the end of the fifth year. Data are generated according to (4) and (5). Specifically, z_{it} was taken to be Poisson with mean $P_{it}\left(bd_t + r_{it}\right)$ where we take P_{it} as the average casualty rate at SSD i between 1992 and 2001, the casualty rate r_{it} is taken to be unity, $b = -0.1$, and d_t is a dummy variable for the treated years. (The implicit assumption is that the casualty rate does not vary over site; choosing population equal to average casualties then amounts to a choice of units of measurement for the population.)

A particular experiment consists of generating ten years of annual data of random variables, wherein the mean of the random variable falls by 10% at the treated sites at the end of the fifth year, remaining constant elsewhere (and hitherto). There are no trends in mean. These data describe a world in which, in truth, the treatment reduces casualties by 10% at each treated SSD. This known effect is then estimated by a variety of estimation techniques. The results from a number of such experiments allow comparison of the effectiveness of different techniques.

1

estimated in biannual data.

ESTIMATION METHODS

The dependent variable is the *difference* between casualty rates in the treated and control sites. This is modelled as linear function of a single dummy variable on the countermeasure periods. This can be estimated by OLS but, as discussed in section 3, the assumption that casualty data are Poisson implies that the variance of the residuals will be inversely proportional to the population of the site; in this case, a version of WLS will be superior to OLS. Finally, we pointed out above that fitting a constant and trend will reduce efficiency if these variables are not needed to avoid misspecification. By construction, the artificial data have no constant or trend so it is possible to demonstrate by simulation the effect of their superfluous inclusion.

RESULTS

Table 2 gives the mean and standard deviation of the results from 1000 applications of four estimation procedures (OLS and WLS by trend and no trend). Note that all four procedures are unbiased (less than two standard errors from the true value of 0.10). WLS tends to halve the error variance; removing the constant and trend reduces the error variance to about one third. There are thus substantial gains to WLS and to omitting the trend if this can be supported.

In a given empirical study, one will have a single estimate and a single estimated standard error. A typical estimate might be the mean reported in the table, with standard error given by the tabulated standard deviation. For the first three procedures, the 95% confidence intervals associated with the typical estimate all include zero. In particular, one would tend not to reject at the 5%level the null hypothesis that the countermeasure had no effect on casualties. In contrast, the confidence interval for the WLS-no trend procedure is approximately 0.10 ± 0.05, which excludes zero and is approaching a satisfactory level of precision. This indicates a substantial return to the employment of an efficient procedure.

5. STOCHASTIC OR DETERMINISTIC TRENDS?

STOCHASTIC TRENDS

We have argued that if the selection of the control-regions is good enough, it may not be necessary to include linear trends in the model, since both region and control will be subject to similar influences—except for the countermeasure. It is obviously useful in principle to allow for the possibility of separate developments at sites but, as argued in section 3 and demonstrated in section 4, this may come at the price of increased diffuseness of estimates.

A modern idea, derived from time series analysis of macroeconomic data, is that omitted demographic variables are unlikely to be well measured by deterministic linear trends (or higher-order polynomials). With a linear trend, the *change* in the demographic variable is modelled as a constant from period to period. It is a natural generalisation to assume that the change is *stochastic*. Consider therefore the model

(20) $$\Delta r_t = c + h_t$$

where Δr_t is the change in demographic variable r_t, c is a constant , and h_t is a random variable with zero mean. If it is further assumed that h_t is serially uncorrelated with constant variance—in which case h_t is referred to as *white noise*—then the process Δr_t is known as a *random walk with drift* c. Writing (20) in the form $r_t = r_{t-1} + c + h_t$ and back-substituting, one can recast the equation in the form

(21) $$r_t = r_0 + ct + h_1 + h_2 + ... + h_t$$

where r_0 is the pre-sample value of the demographic variable. Equation (21) expresses r_t as the sum of a linear trend and the *driftless* random walk

(22) $$w_t = h_1 + h_2 + ... + h_t$$

A driftless random walk is a *wandering* series: it is as likely to go up as down, irrespective of its current value. It will very often have apparent local trends, and will return to its starting value quite infrequently (indeed, it is more likely *never* to return than to spend half its time above, half below its initial value). Variables of the form (21) or (22) are called *stochastic* trends in contrast to the *deterministic* trend given by a linear function of time.

It is clear that omitted demographics will be better represented by stochastic trends. Our aim is to see how stochastic trends change estimation techniques and to quantify likely changes in the accuracy of estimates.

STOCHASTIC TRENDS AND THE GLS ESTIMATOR

Assume for the moment that we have data from a single site. The model we wish to estimate is, as before, of the form

(23) $$z_t = bd_t + r_t + e_t$$

where is z_t casualties in period *t*, d_t is the countermeasure variable and e_t is a stochastic error which we shall take to be white noise. We assume r_t is a stochastic trend given by (21). Substituting into (23) from (21) we find

(24) $$z_t = bd_t + r_0 + ct + w_t + e_t$$

We wish to estimate b. This model is identical to the models studied above except in one respect: the error process is now

(25) $$u_t = w_t + e_t$$

that is, the sum of a driftless random walk and a white noise process. Each u_t will have zero conditional mean, implying that an OLS estimate of b derived from (24) will be unbiased, but the errors are correlated at different dates so the Gauss-Markov theorem does not apply—OLS is not necessarily BLUE. Moreover the standard errors of the OLS estimates calculated in the usual way are no longer correct. In this case GLS is an appropriate estimation technique as discussed in section 3.

Efficiency gains from GLS when there are stochastic trends

We shall obtain some formulae for the relative efficiency of GLS over OLS. Write (24) in the stacked vector form

(26) $$z = bd + r_0 v_1 + c v_2 + u$$

where v_1 is the vector of ones , v_2 is the trend vector and $u_t = w_t + e_t$. We assume the covariance matrix of the error process *u* is known

(27) $$E(uu') = s^{2}\Omega$$

up to the scale parameter s^{2}. We are interested only in the parameter b so we purge (26) of the v vectors. Let V be the matrix whose columns are the two v vectors and let M_V be the projection that removes them, that is, $M_V x$ is the vector of residuals of a regression of any vector x on the v vectors. The two estimators are

(28) $$\hat{b}_{OLS} = zM_V d / d'M_V d$$

and

(29) $$\hat{b}_{GLS} = \tilde{z}'M_{\tilde{V}}\tilde{d} / \tilde{d}'M_{\tilde{V}}\tilde{d}$$

with variances given by

(30) $$Var\left(\hat{b}_{OLS}\right) = s^{2} z'M_V \Omega M_V d / \left(d'M_V d\right)^2$$

and

(31) $$Var\left(\hat{b}_{GLS}\right) = s^{2} / \tilde{d}'M_{\tilde{V}}\tilde{d}$$

Note that these expressions employ the ~-transformation introduced in section 3. To evaluate these expressions thus requires knowledge of the covariance matrix Ω. In the case where $u_t = w_t + e_t$ this matrix turns out to be

(32) $$\Omega = s_e^{2}\left(I + nJ\right)$$

where the matrix J , derived from the random walk component in u_t , is

(33) $$J = \begin{bmatrix} 1 & 1 & 1 & .. & 1 \\ 1 & 2 & 2 & .. & 2 \\ 1 & 2 & 3 & .. & 3 \\ : & : & : & : & : \\ 1 & 2 & 3 & .. & T \end{bmatrix}$$

and $n = s_h^2 / s_e^2$ is the variance of the h_t process relative to that of e_t. This quantity will be important in what follows. In the composite process $u_t = w_t + e_t$ we refer to h_t as a *permanent* shock and to e_t as a *transient* shock: the point is that, from period to period, h_t permanently alters the value of u_t whereas e_t disappears and is replaced by its next value. Thus n is the variance of the transient component of u_t relative to the permanent component. This needs to be calculated to perform GLS—it is needed to evaluate Ω—and we shall turn to this below. For the present, note that n will determine the relative efficiency of GLS to OLS. If n = 0, then OLS and GLS are the same so the relative efficiency is unity. When n > 0 there are gains to GLS. Assume we estimate in 15 years of annual data and that d_t is a countermeasure dummy taking effect over the last five years. Figure 4 gives the relative efficiency of OLS and GLS as a function of n , calculated from (30) and (31). Small values of n lead only to small efficiency gains, but once n is 0.25, GLS offers efficiency gains of about 10%; for n around 0.5, GLS is about 25% better than OLS.

GLS when there are many regions: the panel estimator

These results have been derived for the data from a single site. We now demonstrate how to extend GLS to the case where there are many sites $i = 1,...,N$. We assume the data in each region have been adjusted by the ~-transformation and normalise all data by the scale parameter s_i, which we allow to vary across regions. We write the complete model in matrix form

(34) $$\begin{bmatrix} \tilde{z}_1/s_1 \\ \tilde{z}_2/s_2 \\ \vdots \\ \tilde{z}_N/s_N \end{bmatrix} = \begin{bmatrix} \tilde{V}/s_1 & 0 & .. & 0 \\ 0 & \tilde{V}/s_2 & .. & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & .. & \tilde{V}/s_N \end{bmatrix} \begin{bmatrix} r_{01} \\ c_1 \\ \vdots \\ c_N \end{bmatrix} + b \begin{bmatrix} \tilde{d}_1/s_1 \\ \tilde{d}_2/s_2 \\ \vdots \\ \tilde{d}_N/s_N \end{bmatrix} + \begin{bmatrix} \tilde{u}_1/s_1 \\ \tilde{u}_2/s_2 \\ \vdots \\ \tilde{u}_N/s_N \end{bmatrix}$$

The stacked column vector on the left is $NT \times 1$; the first square matrix on the right is $NT \times 2N$ and consists of the (adjusted) vector of constant and trend written down the diagonal (with zeroes elsewhere); the second column matrix on the right consists of the $2N$ parameters associated with the region-specific constant and trend; the second column vector is the stacked countermeasure variables; the final column vector is the stacked residual vectors. If it is assumed that the composite errors are uncorrelated at different sites[7], the covariance matrix of this residual vector is the $NT \times NT$ identity matrix , I_{NT} . It follows that an OLS estimate of (34) is BLUE. One need not develop an estimate of b from (34) which would require a large matrix inversion): rather, one multiplies through by the projection matrix $M_{\tilde{V}}$ to eliminate the first term on the right and calculates the estimate of the parameter of interest by a univariate regression. It turns out that the panel estimator is just a weighted average of the GLS estimates in each region

(35) $$\hat{b}_{GLS,panel} = \sum_i w_i \hat{b}_i$$

1

[7] Plausible if the control regions are well-chosen.

where $w_i = \left(\tilde{d}_i' M_{\tilde{V}} \tilde{d}_i \big/ s_i^2\right) \Big/ \sum_j \tilde{d}_j' M_{\tilde{V}} \tilde{d}_j \big/ s_j^2$.

One might want to allow the Ω matrix to vary over sites, if one believed that the relative variance n varied over sites. Little is changed. In equation (34), the $\tilde{V}$ matrices now vary by site since the ~-transformation now varies by site. The estimator is given by a slightly modified version of (34) in which the $\tilde{V}$ matrices are indexed by i.

Below we report our experience with this estimator. It can be achieved by a sequence of simple data transformations. One begins with the primary equation (5). We assume the error variances s_i^2 are proportional to $(1/P_{i.} + 1/P_{i'.})$ (where $P_{i.}$ etc. denotes the average of P_{it} over time) and weight the variables by the factor $1/s_i$. The variables are then adjusted by the ~-transformation and the transformed constant and trend removed from the policy variable by regression . The final parameter estimate is then obtained by a simple univariate OLS regression.

CONSISTENCY ISSUES

The property of consistency of estimators, namely that the estimate approaches the true value in probability as the sample size grows[8], is desirable because it implies that, if the sample size is large, we can be confident that the estimate is close to the true value. It implies also that it is sensible to collect more data if greater precision is needed. In this section we shall consider consistency when the error process is a random walk.

A single region

Assume first that we study the data of a single region and that the model is given by

(36) $$x_t = b + w_t$$

where w_t is the random walk given by (22) and b is an unknown parameter which we want to estimate. In these circumstances the OLS estimator is just the sample mean,

(37) $$\bar{x}_t = b + h_1 + \frac{T-1}{T}h_2 + \frac{T-2}{T} + \ldots + \frac{1}{T}h_T$$

and the variance of the estimate about b is

(38) $$Var(\bar{x}_t) = (T+1)(2T+1)/6T$$

1

[8] In the ensuing discussion it will be convenient to employ the slightly stronger concept of consistency that the variance of the estimate about the true value approach zero.

This grows monotonically with T. Thus, even though the estimator is unbiased, it grows increasingly diffuse as the sample size grows. In terms of efficiency, the best OLS estimator of b is simply the *first observation*, $x_1 = b + h_1$. Clearly then, OLS is not a consistent estimator of b in (36). In certain circumstances, OLS will be consistent when there are random walk errors. It can be shown that, in the model

(39) $x_t = bd_t + w_t$,

OLS will be consistent if the sum of squares $d'd$ grows at a strictly faster rate than T^2 as T grows[9]. For the constant vector, $d'd$ grows at a rate T^2 and thus fails this test; the sum of squares of the trend vector, however, grows at T^3, implying that OLS will be consistent for a linear trend, even when the error is a random walk.

In the case of GLS, the extended Gauss-Markov theorem implies that the variance of the GLS estimator is less than or equal to that of OLS, so that, whenever OLS is consistent, so too will be GLS (since both are unbiased and linear and GLS is BLUE). The theorem implies as well that the variance of GLS cannot increase as the sample grows (since the estimator over a subset of the sample is also linear and unbiased). In the case (36) where a simple constant is fitted, the GLS estimator turns out to have variance s_h^2, which is independent of T. Thus, whereas the OLS estimator gets *worse* as the sample grows, GLS stays the same. Both, of course, are inconsistent. It can happen that OLS is not consistent but GLS is; an example is $d_t = t^{1/2}$, where the variance of OLS converges to a constant as T grows, while that of GLS converges to zero.

We summarise the discussion so far. For a single region, when the error process is a random walk, GLS will be better than OLS but both may be inconsistent. GLS never gets worse as the sample size grows; OLS may. A sufficient condition for the consistency of both is that the sum of squares of the dependent variable grows strictly faster than T^2.

Many regions

When there are many regions, there are two dimensions to which consistency can apply, N and T. We consider what happens as N grows. Assume that, as N grows, each new region is an independent draw from a fixed distribution. According to (35), the GLS estimate is an average of GLS estimates in single regions. The same applies to OLS. Since both are unbiased estimators, an appeal to the law of large numbers delivers the consistency of both.

Thus, when there are stochastic trends, increasing the number of regions N will always be a good idea, irrespective of estimation method. This is in contrast to increasing the number of time periods which, at worst, will not increase the variance of GLS, but which may well make OLS worse.

1

[9] The positive quantity f_T grows *strictly faster* than g_T if $g_T / f_T \to 0$ as T grows.

HOW IMPORTANT ARE STOCHASTIC TRENDS?

We have argued that stochastic trends may more plausibly represent omitted demographics in studies of the effects of interventions, and shown how their presence modifies the estimators. We have shown that substantial gains in efficiency are on offer if the permanent shocks are reasonably large relative to the transient shocks (Figure 4). We now propose to investigate how large the permanent shocks are in real casualty data, specifically total casualty monthly data for the 505 Queensland SLAs, from Jan. 1995 to Dec. 2001. The analysis is based on the first difference of monthly casualties. According to (24) this is, up to a constant,

(40) $\Delta z_t = e_t - e_{t-1} + h_t$ [10]

Seasonal adjustment

We wish to adjust (40) for seasonality. If seasonal effects are additive, this is done by subtracting the average of Δz_t for each month over the seven years of the sample[11]. Unfortunately, since we have so few years, this will change the variables on the right in (40). For example, e_t is replaced by

$$e_t - \sum_p e_{t+12p} / 7$$

where the index *p* ranges over the years of the sample. It is simple to see that seasonal adjustment reduces variance by a factor of 42/49 = 0.857. This is equally true for h_t and, since we are ultimately interested in the *ratio* of the variances of e_t and h_t, less harm is done than might be first thought. The serial correlation coefficients of the variables are unaffected except at lag 12, where the common seasonal adjustment will induce serial correlation, and lag 11, where the seasonal adjustment to e_t is common to that of e_{t-1} 12 months later. These considerations will arise from time to time below so we shall highlight them.

Seasonally adjusting Δz_t will

- reduce the variances of e_t and h_t by a factor of 0.857.
- leave the serial correlations of $\Delta z_t, e_t$ and h_t unchanged except at lags 11 and 12 (plus multiples of 12).

The correlogram of Δz_t

The first natural port-of-call is the correlogram of Δz_t. The process given by (40) is a first-order moving average (MA(1)) whose first-order serial correlation is given by

(41) $r_1(\Delta z_t) = -s_e^2 / (2s_e^2 + s_h^2) = -1/(2+n)$

1

[10] We assume the SafeST initiatives are unimportant at SLA level.

[11] Which will remove the omitted constant.

where n is the variance of the permanent shock relative to that of the transient shock. All other serial correlations are zero.

Figure 5 gives the average correlogram for the 505 SLAs. Standard errors for these correlations are of the order of 0.006. After the first correlation, all are less than one standard error from zero until the tenth, which is about three standard errors from zero. The correlations at lags 11 and 12 are marked, in accordance with the discussion above of seasonal-adjustment. This is about as perfect a correlogram of a seasonally-adjusted MA(1) as one could want. The (average) first-order correlation is estimated as –0.48864 with a sample standard error of 0.00408, which is about three standard errors from –0.5, the value that (41) would take if the permanent shocks were absent. This would imply a value for n of about 0.04650, that is the permanent shock is about 5% of the transitory shock in monthly data. But the problem of bias arises.

It is well known that estimates of the serial correlation coefficient are biased in small samples. To estimate the size of the bias in our context, we constructed 1000 random versions of $e_t - e_{t-1}$ for our sample size and used our Box-Jenkins package (TSP 4.5) to estimate the first-order serial correlation. We found an average estimate of –0.48582 (standard error 0.00276). This average is smaller in magnitude than that found in the SLAs—and thus implies a higher value of n —even though n is zero by construction in these data. This implies that we cannot base a reliable estimate of n on the first-order serial correlation.

Differencing at different horizons

An alternative approach is to study the effects of differencing at different horizons. Define

(42) $$\Delta_k z_t = z_t - z_{t-k} \qquad k = 1,2,...$$

Then it is easy to see that

(43) $$Var(\Delta_k z_t) = 2s_e^2 + ks_h^2$$

so that

(44) $$\frac{Var(\Delta_k z_t)}{Var(\Delta_1 z_t)} = 1 + \frac{n}{2n+1}(k-1)$$

The LHS is easily calculated from observed data; an estimate of $n/(2n+1)$ can be obtained from a regression of the LHS on $k-1$ and the corresponding value of n thus inferred. Seasonal-adjustment does not distort this estimate because (44) is essentially a ratio of variances and numerator and denominator are changed by the same factor. Problems arise at k = 11 and 12, however, because of seasonal-adjustment. In particular, at k = 12, the whole of the variance-reducing distortion is removed and we expect a jump of about 17% at that value of k—as is, in fact, observed. Figure 6 shows the SLA average relative variance for the first 10 values of k. The clear increasing trend is strikingly in accord with the theory developed above. Fitting a linear trend through these data (constrained to pass through (1, 1)) gives a quite precisely estimated slope parameter of 0.003488 (standard error = 0.00048).

This implies an average value for n of 0.007, that is, the permanent shock is about 0.7% of the transient shock in monthly data.

This number may seem small at first sight, but this would be a false impression as the shock is permanent. Assume we work with annualised data. Then the annual transient shock has 12 times the variance of the monthly transient shock. The permanent shock in annual data is

$$h_1 + (h_1 + h_2) + ... + (h_1 + h_2 + ... + h_{12}) = 12h_1 + 11h_2 + ... + h_{12}$$

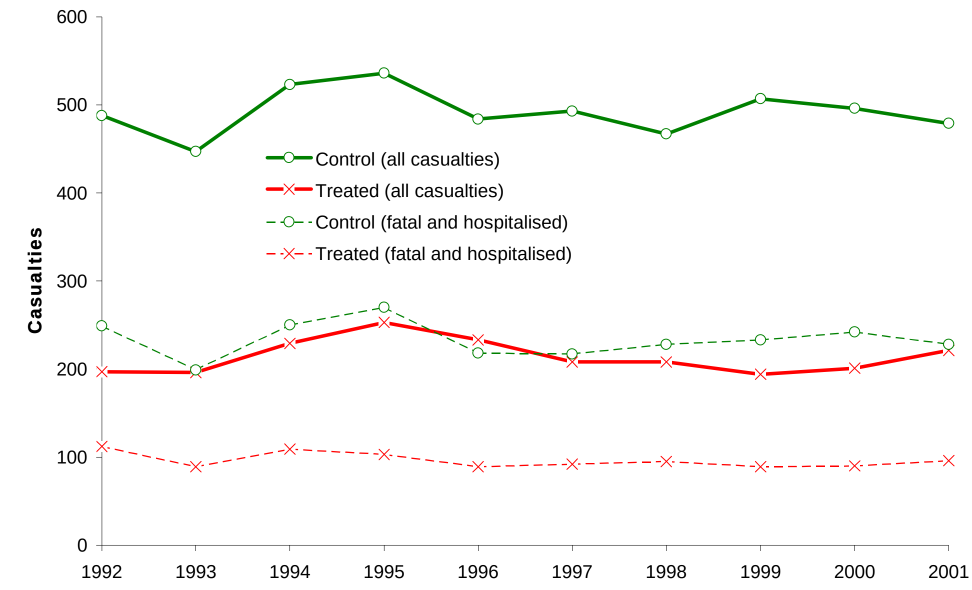
which has variance $650 s_h^2$. Thus n in annual data is 650/12 = 54.1 times as great as in monthly data: factoring up the monthly estimate gives an annual n of 0.38. Referring to Figure 4, one sees that OLS will be about 80% as efficient as GLS.

Practical problems of estimating the relative variance

We have calculated the slope of the regression (44) from the average SLA values of the RHS. Equally, this could be calculated by performing the regression at each SLA and averaging the resulting parameters. This is a useful way of thinking about it because, in some cases, one will have only a few sites, perhaps not SLAs, and will need to estimate values for n . It turns out that the standard deviation of the slope estimates over sites is very large relative to the mean—over eight times as large. Thus one will be nearly as likely, in a single SLA, to find a *negative n* as a positive one. How one proceeds in these circumstances deserves consideration. Bayesian methods are attractive.

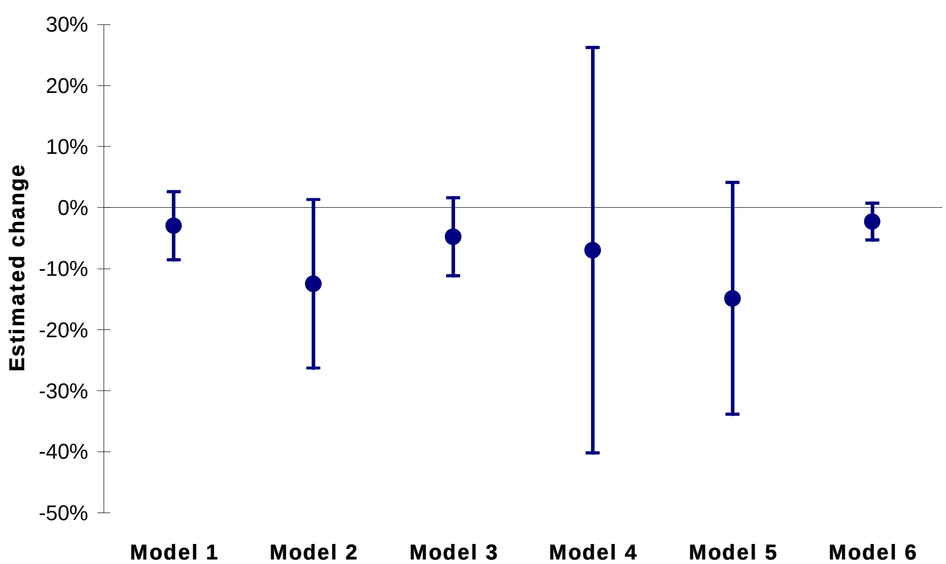
SUMMARY

We have argued that unobserved demographics will be better modelled by stochastic trends, as these allow for the demographic effect to wander away from a linear trend if this improves the fit. We have shown how the model can be estimated by GLS in these circumstances and have demonstrated that considerable efficiency gains can result if the error process driving the stochastic trend is large enough in variance relative to the equation error. We have studied casualty data for the 505 Queensland SLAs, 1995 to 2001, and have estimated the relative variance to be of the order of 0.38 on average in annual data, implying that, in annual data, OLS estimates will be about 80% as efficient as the proposed GLS.



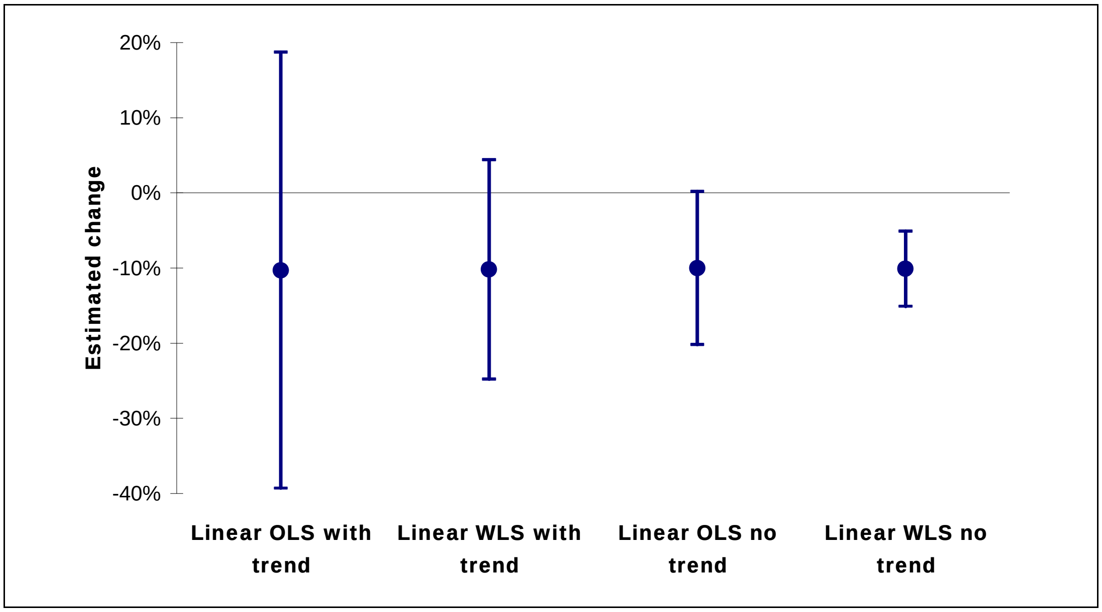
Source: QT crash data.

Figure 1 Pedestrian casualties at treated and control sites



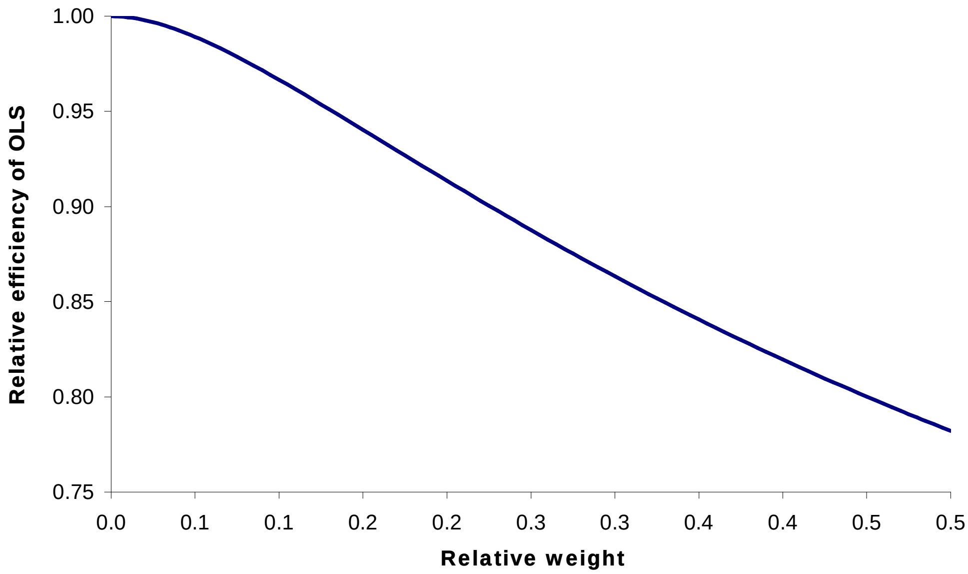
Source: ARRB TR calculations. Note: None significant at the 5% level.

Figure 2   95% confidence intervals for countermeasure effects



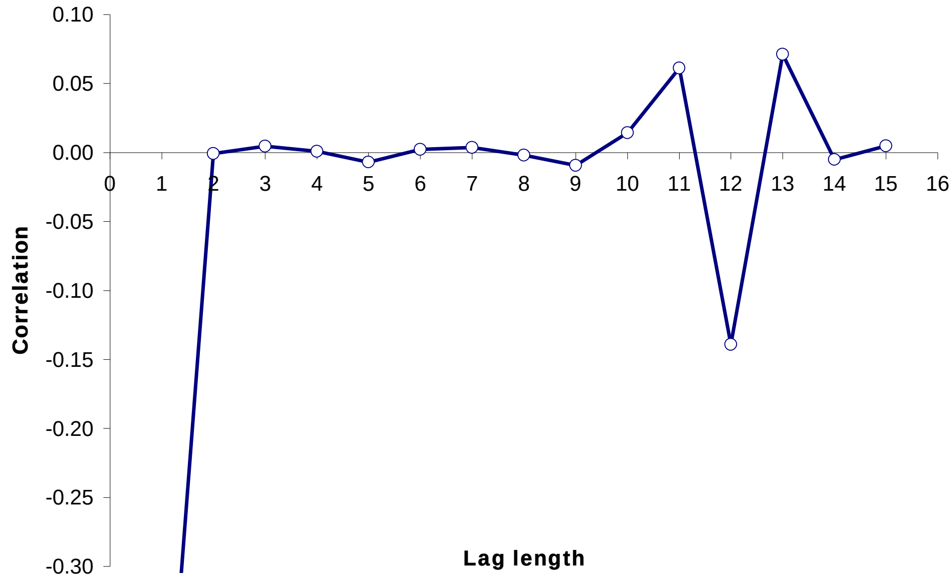
Source: ARRB TR calculations.

Figure 3 Simulated 95% confidence intervals for countermeasure effects



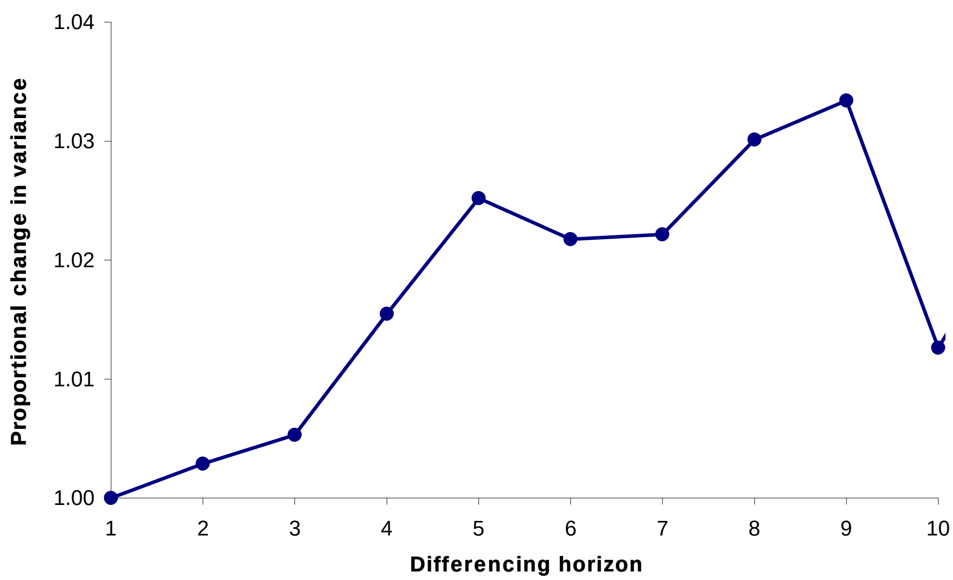
Source: Economic outcomes.

Figure 4   Relative efficiency of OLS and GLS against relative weight of permanent and transitory shocks



Source: Economic outcomes.

Figure 5 Correlogram of monthly changes in casualties: average of 505 SLAs



Source: Economic outcomes.

Figure 6 Effect of differencing horizon on variance of change in casualties