

Using Base Year Data in Scenarios

Eric Kemp-Benedict 13 January 2007 (last modified 15 February 2007)

When developing scenarios using an empirical model, there is an opportunity to make use of country-specific base-year data. Suppose that a scenario variable z_i for country i is explained by N variables x_{ji} :

$$z_i = f(x_{1j}, x_{2j}, \dots, x_{Nj}) + R_i. \quad (1)$$

In this equation, the values R_i represent the residuals that are the differences between the observed country-specific values z_i and the model values given by the function $f(\cdot)$. In the base year, the residuals are known, since the model has been specified and the initial values are known. In a future scenario year, the same equation is applied to a new set of explanatory variables x'_{ji} , with a new set of residuals R'_i ,

$$z'_i = f(x'_{1j}, x'_{2j}, \dots, x'_{Nj}) + R'_i. \quad (2)$$

In the scenario, the residuals are not known. However, the base-year residuals can be used to account for country-specific factors. An estimate \hat{z}'_i of the future value is constructed that includes the base-year residuals:

$$\hat{z}'_i = \theta R_i + f(x'_{1j}, x'_{2j}, \dots, x'_{Nj}). \quad (3)$$

The parameter θ can be any value from $\theta = 0$, in which case no base-year information is used, and the estimate is given entirely by the model, to $\theta = 1$, in which case it is assumed that the scenario residuals are the same as the base-year residuals. To show that the optimal value for θ probably lies somewhere between 0 and 1, assume for the moment that the scenario residuals R'_i can be found, and use Equation (2) to write Equation (3) as

$$\hat{z}'_i = \theta R_i + z'_i - R'_i. \quad (4)$$

Taking the difference between the scenario estimate and the actual values, squaring, and summing, gives

$$\sum_i (\hat{z}'_i - z'_i)^2 = \sum_i (\theta R_i - R'_i)^2 = \theta^2 \sum_i R_i^2 - 2\theta \sum_i R_i R'_i + \sum_i R_i'^2, \quad (5)$$

which is proportional to the mean squared error (MSE) of the scenario estimate. The optimal value for θ , θ^* , is found when the expression in Equation (5) is at its minimum.¹ This occurs when

$$2\theta^* \sum_i R_i^2 - 2 \sum_i R_i R'_i = 0. \quad (6)$$

Solving for θ^* shows the optimal value for θ to be

$$\theta^* = \frac{\sum_i R_i R'_i}{\sum_i R_i^2}. \quad (7)$$

If the base-year and scenario-year residuals are completely uncorrelated, then the numerator in Equation (7) will be zero, and the optimal value of θ will be $\theta^* = 0$. If the base-year and scenario-year residuals are identical, then the numerator and denominator in Equation (7) are identical, and the optimal value for θ will be $\theta^* = 1$.

In general, it can be expected that the farther separated in time the base year and scenario year are, the less correlated the residuals will be, so that θ^* is likely to start at 1 and decline toward 0 over time. The specific time dependence of θ^* depends on the course the scenario takes, but any given scenario is likely to be consistent with a number of different trajectories for θ^* . One criterion for assigning a simple functional form for a time-varying $\theta(t)$ is to ask that it be self-consistent, in that if a time interval is split into sub-intervals, and Equation (3) is

1 The idea of combining modeled values with location-specific values in order to minimize the MSE of the estimate is taken from the literature on Small Area Estimation, in particular from Rao, J.N.K. (2003), *Small Area Estimation*, Hoboken, New Jersey: John Wiley & Sons, Inc. As far as the author knows, the technique has not been applied to quantitative scenarios.

applied to each sub-interval, then the final result would be the same as if Equation (3) were applied to the whole interval. This will be true as long as

$$\theta(t)\theta(t')=\theta(t+t') . \quad (8)$$

The function that satisfies Equation (8) and the boundary condition that $\theta(t)$ go to zero as t gets large is an exponential decay:

$$\theta(t)=e^{-kt} . \quad (9)$$

If historical data are available, then using Equation (7) a historical trajectory for θ^* can be estimated. By fitting an exponential function to the trajectory, the parameter k in Equation (9) can be estimated and used when generating scenarios.