Response to Professor David Mackay's comments on my paper

Gordon Hughes

The mathematics of Professor Mackay's proposition

In his paper Professor Mackay claims that the model which I used to estimate the decline of performance of wind farms with age is not identified. He argues that it is possible to make an arbitrary change in the coefficient on age with no effect on the errors of the equation. While I have no quarrel with his mathematics, it appears that Professor Mackay has failed to recognise that he has imported some rather special assumptions about the nature of the data and the way in which my model has been estimated.

To explain my argument it is instructive to provide a slightly different version of his analysis designed to highlight the source of the problem of non-identification in this context. The key point is that under a particular set of assumptions the model can be rewritten so that the time variable occurs twice. This implies that the age variable and a time trend which forms part of the time effects are collinear. For simplicity, I will develop the argument for a version of the specification that Professor Mackay uses to estimate rates of decline in performance in his Section 3. This is:

$$y_{it} = \log(ma[lf_{it}, 3]) = \theta + \rho a_{it} + u_i + v_t + \varepsilon_{it} \,. \tag{1}$$

He assumes that the age of plant i in time period t is $a_{it} = t - b_i$. Without loss of generality we can rewrite the period effects to include an arbitrary time trend: $v_t = \sigma t + \tilde{v}_t = \sigma(a_{it} + b_i) + \tilde{v}_t$. Thus, equation (1) can be rewritten as

$$y_{it} = \theta + \rho a_{it} + u_i + \sigma(a_{it} + b_i) + \tilde{v}_t + \varepsilon_{it} \tag{2}$$

or

$$y_{it} = \theta + (\rho + \sigma) a_{it} + \tilde{u}_i + \tilde{v}_t + \varepsilon_{it} \tag{3}$$

where $\tilde{u}_i = u_i + \sigma b_i$. To maintain the assumption that the wind farm site effects sum to zero we need to define

$$\hat{u}_i = \tilde{u}_i - \left(\frac{1}{N}\right)\sum_{i=1}^{N} \tilde{u}_i = \tilde{u}_i - \sigma \overline{b} = u_i + \sigma(b_i - \overline{b}) \,. \tag{4}$$

where $\overline{b}$ is the average value of the b$_i$. Similarly, the period fixed effects are normalised as follows:

$$\hat{v}_t = \tilde{v}_t - \left(\frac{1}{T}\right)\sum_{t=1}^{T} \tilde{v}_t = \tilde{v}_t + \sigma \overline{t} = v_t + \sigma(t - \overline{t}) \ . \tag{5}$$

Then:

$$y_{it} = \left[\theta + \sigma(\overline{b} + \overline{t})\right] + (\rho + \sigma)a_{it} + \hat{u}_i + \hat{v}_t + \varepsilon_{it} \ . \tag{6}$$

In words, the period fixed effects have been expressed as the sum of a time trend (σt) and a different set of period fixed effects. Since age is a linear function of time, the time trend can be expressed as a function of age with an adjustment for the birth date of the wind farm which is collected into the site fixed effects. Then, an adjustment is made to the constant term to maintain the assumption that the site and period fixed effects are each normalised to sum to zero.

This is the logic which underpins Professor Mackay's claim that my model is not identified. Under his assumptions the coefficient ρ can be replaced by $(\rho + \sigma)$ for any arbitrary value of σ without altering the errors ε_{it} that are minimised in estimating the equation. However, the result is not general because his assumptions are stronger than he realises. What he has demonstrated is that my model might not be identified, but he has not shown that the variant of the model which I have actually estimated is not identified. In fact, as I will explain, non-identification is such a routine problem in statistical models of this kind that all statistical software in general use contains standard checks and adjustments to deal with it. It is difficult for an applied statistician to miss the symptoms of non-identification.

Identification

Professor Mackay suggests that the only way in which lack of identification can be or has been avoided is via the method of normalising the fixed effects, which he suggests will lead to arbitrary results. Further, he asserts that the problem of non-identification is independent of the method of estimation, because his view of the estimation is that it applies literally to the model written out in equations (1) or (6). Both of these assertions are incorrect because all efficient methods of estimating a model of this kind involve some kind of transformation designed partly to reduce the number of parameters to be estimated – e.g. by eliminating the site effects - and partly to avoid the danger of non-identification in general.

In practice, there are many ways in which statistical models are identified. These methods of identification do not, as a general rule, affect the parameters of interest. Instead, the adjustments required for identification are typically captured in the constant term, which is viewed as a nuisance parameter. In the case of my analysis, there are multiple potential sources of identification, though in practice the first is critical.

A. The analysis presented above relies entirely on a crucial substitution linking age and time: $t = a_{it} + b_i$. Given the structure of the data this assumes that monthly period effects are matched by measuring age in months. However, that is not the way in which my model is specified since I have measured age in years. It follows that age is a nonlinear function of time as in: $a_{it} = h_i(t)$. Now, (3) becomes

$$y_{it} = \theta + \rho a_{it} + \sigma h_i^{-1}(a_{it}) + u_i + \tilde{v}_t + \varepsilon_{it} \tag{7}$$

which means that the collinearity is removed provided that h_i^{-1} is not a single-valued linear function of age. With a little mathematical manipulation it is possible to generalise the argument developed above, but only for a specification of the form $a_{it} = h(t) - b_i$. This would be the case if all wind farms commenced their operations in the same month of the year. For practical purposes, the combination of measuring age in year and time in months will ensure that the model is identified.

Even when both age and time are measured in months, Professor Mackay's strictures will not always apply. The result requires that all of the period effects can be altered by different amounts. However, it is common to assume a period error structure that can be written as: $v_t = v_t^M + v_t^Y$ where the first component is a monthly effect over all years and the second component is a yearly effect over all months in the year. To illustrate the point, suppose that we were to estimate a model in which the period effects take the form of a monthly deviation from the annual average – or relative to the value for some base month – but with no variation across years. That specification is only consistent with equation (5) if $\sigma = 0$.

B. A standard method of identification in statistics is via what are called exclusion restrictions, i.e. by setting certain coefficients equal to zero. The transformation of the model so that it can be estimated without site effects is a form of exclusion restriction. An alternative exclusion restriction would be to drop the linear term from the model that is estimated.

To illustrate the point, consider the more general version of equation (1) with:

$$y_{it} = \theta + f(a_{it}) + u_i + v_t + \varepsilon_{it} \ . \tag{8}$$

As Professor Mackay argues, this can be rewritten as:

$$y_{it} = \left[\theta + \sigma(\bar{b} + \bar{t})\right] + f(a_{it}) + \sigma a_{it} + \hat{u}_i + \hat{v}_t + \varepsilon_{it} \tag{9}$$

Now suppose that f(a) is not linear and we estimate equation (9) without any linear term. This is an exclusion restriction which identifies the model by ensuring that $\sigma = 0$. Note that the previous point may be thought of as a variant of this if equation (7) is expressed in terms of time rather than age.

C. An alternative form of exclusion restriction is to transform the model so that it is estimated without a constant term. The classic methods of doing this are to estimate the model using either (a) first or annual differences of the data, or (b) differences relative to the mean over all sites for each time period (known as the between estimator). It is trivial to show that either of these methods requires that $\sigma = 0$.

There may be other sources of identification in other variants of the model. The general point is that the details of data transformations and methods of estimation matter. This is not a theorem that is independent of these details. Thus, whether or not a general model is identified when applied in specific circumstances can only be established by considering all of these details and/or by relying on standard numerical methods designed to flag when a model is not identified.

In the case of my analysis the key source of identification is A above, i.e. the fact that age is measured in years while all of the time effects are based on months. This can be confirmed by re-specifying age as being measured in months in which case the symptoms of non-identification appear immediately. These symptoms take the standard form of collinearity leading to one of the time dummies being dropped. That effectively fixes a value of $\sigma \neq 0$ and the value of $\rho + \sigma$ is reported as the coefficient on (monthly) age. It is easy to confirm that dropping different time dummies – e.g. for the first period or the last period – generates very different coefficients on age.

Reliance upon the timing of birth dates as a method of identification or as an instrumental variable is not unusual in other branches of economics, especially in labour economics. In this case, there is a strong seasonal element to the performance of wind farms so that it is entirely natural to measure age in years while including monthly period effects.

Statistical background

To understand what is meant by non-identification or lack of identification it is helpful to think of the classic linear model:

$$y_t = \alpha + \sum_{k=1}^{K} \beta_k x_{kt} + \varepsilon_t \qquad (10)$$

or in matrix notation

$$y = X\beta + \varepsilon . \qquad (11)$$

Intuitively, it should be obvious that it will be impossible to estimate distinct values for all of the β coefficients if two or more of the x variables are collinear, i.e. if there is some subsidiary relationship such that the variable x_i can be expressed as a linear combination of other x variables as in

$$x_i = \sum_{j \neq i} \gamma_j x_j . \qquad (12)$$

Under such circumstances it will only be possible to estimate a linear combination $C\beta$ of the β coefficients where one of the dimensions of C – and thus its rank – is less than K, the number of β coefficients. Normally, this is done by dropping the collinear variable from the estimation, i.e. by setting its coefficient to zero.

All of this is standard in applied statistics. The problem of accidental non-identification arises quite frequently in statistical models when dummy variables are used to represent fixed effects because the analyst may not be aware that the data embeds subsidiary correlations between these fixed effects. As an example, accidental non-identification may occur if an equation is estimated that includes a complete set of dummy variables for every month in the year or both genders along with a constant term α.

Since accidental non-identification is routine in applied statistics, every statistical package in general use will automatically test for collinear variables and non-identification before executing the calculations required to estimate the β coefficients. Such a test is easily carried out as it is sufficient to calculate the rank of the matrix X'X obtained from equation (11). Since the matrix X has K+1 columns (including a column of 1's for the constant term), the test is whether the rank of X'X is less than or equal to K+1. If the rank of X'X is less than K+1 then it is said to be a singular matrix. If X'X is singular, most software will calculate which variables are collinear, report the error and drop at least one of the collinear variables from the estimation.

As testing for collinearity is routine, it would be surprising if the symptoms of the non-identification claimed by Professor Mackay had not shown up in my empirical work. In any case, even if non-identification had not been explicitly flagged and allowed for, there are other symptoms which are hard to miss. Suppose that, as a result of rounding errors or other numerical inaccuracies, the rank of X'X is not clearly less than K+1. At a later stage in any least squares estimation the inverse of X'X is calculated – usually not directly but the procedure involves a series of calculations that are equivalent. This operation is numerically unstable if the model is not identified.[1] As a consequence the reported results of the estimation will change dramatically with small adjustments to the model or even when the estimation is run on different computers. Almost invariably, the reported standard errors of some of the coefficients take on extreme values.

So, if Professor Mackay's proposition that the effect of age on performance cannot be identified in this model were correct it would mean that everyone who has analysed the data has somehow missed or failed to report the symptoms of non-identification. Since they are so obvious, this seems unlikely but stranger things have happened. Hence, I have explicitly tested his assertion using the data which he examines, i.e. the log of the 3 month

[1] There are standard procedures for calculating what is known as the generalised inverse of a singular matrix. However, statistical software does not use this approach, precisely because it would mask the symptoms of collinearity. It is less efficient than routines that exploit the specific structure of the matrix X'X. In any case, even if the generalised inverse were to be used, the coefficient on the collinear variable would finish up being set to zero.

moving average of load factors for wind farms in existence in 2004. Empirically I find that it is not valid.[2]

Earlier in this note I have explained that the source of identification in this model lies in the fact that age has been measured in years whereas the period effects are monthly. I have repeated the test using age measured in months rather than years. As expected, the estimation procedure reports collinearity and sets one of the dummy variable to zero, thus generating an arbitrary solution. The test of the residual sums of squares for a range of age coefficients confirms that the model is not identified.

The results of this exercise are clear. The model which I have estimated with age measured in years does not suffer from a lack of identification. On the other hand, the same model with age measured in months is not identified and is properly flagged as such by the statistical procedures.

Conclusion

The difference between Professor Mackay and myself does not concern theorems, it is about facts. There is no disagreement that the general model which I specified and estimated to analyse the performance of wind farms as they age may not be identified. However, there is nothing unusual about this. There is a multitude of statistical models that, similarly, might not be identified. The issue is whether as a matter of fact, given the transformation of the general model into the version that I have estimated, the results that I have obtained have been affected by the potential lack of identification.

I have explained that identification in statistical models is routinely achieved by transforming variables or making assumptions about the structure of errors which remove collinearity. In this case the key element is to measure age as a discrete variable in years rather than as a quasi-continuous variable in months. There are strong seasonal patterns in wind availability and in the associated stresses on wind turbines, so this is an entirely natural transformation which reflects the physical reality of the factors likely to affect the performance of wind farms.

Because non-identification is a routine problem in statistical analysis, its symptoms are well known and statistical packages contain checks and adjustments to spot and deal with it. As is normal in any statistical study, I have carried out extensive tests of alternative

[2] The test is carried out by fitting the equation $(y_{it} - \rho a_{it}) = \theta + u_i + v_t + \varepsilon_{it}$ for different values of ρ and then comparing the residual sums of squares (RSS) – i.e. $\sum_{i,t} \varepsilon_{it}^2$. The values of RSS are not invariant with respect to changes in ρ. The changes in RSS due to changes in the age coefficient are not large, but that reflects the fact that the ratio of the age coefficient to its standard error is quite low, meaning that the coefficient is poorly determined. However, that is quite different from a situation in the errors are identical for different values of ρ.

specifications of my model using different methods of estimation. These tests have revealed no symptoms of non-identification, while a simple direct test confirms that the estimated residual errors are altered by changing the value of the coefficient on age. For these reasons, the possibility of non-identification highlighted by Professor Mackay does not, as a matter of fact, affect the results reported in my study.

We must be very clear that there is a fundamental distinction between statistical models which are not fully identified, i.e. there is no unique solution given the data, and those which are poorly determined. Lack of identification manifests itself as a pathological condition in which estimation is either not possible or generates extreme values for the standard errors. No statistician should miss the symptoms. Models which are poorly determined have large standard errors associated with the parameters of interest, so that there are wide confidence bands that describe the range of reasonable values for those parameters. In a different context, econometricians discuss models which are said to be "weakly identified". What this means is that the models are formally identified – i.e. there is a unique solution to the equations – but the statistical properties of the data mean that the standard errors associated with that unique solution are rather large, i.e. the model is poorly determined.

In Section 3 of his paper Professor Mackay offers some alternative estimates of the rate of decline in performance of wind farms based upon a subset of the data which I have compiled. As I have explained, his model is simply a variant of one of my basic models in which the log of the load factor for a wind farm is replaced by the log of a moving average of the load factor with separate rates of decline for each month. It is hard to understand why, as a matter of practicality or logic, one would wish to represent the decline in performance as a function of the month in the year. For purposes of policy or economic analysis the primary interest must lie in assessing the average rate of decline in total output from one year to the next.

For direct comparison with the results in my report I have estimated a version of Professor Mackay's model using the same data but with the same rate of decline for all months.[3] There is one issue with his specification that needs to be spelled out and dealt with. His

[3] There are some technical differences between Professor Mackay's analysis and my version of it. He says that he has used 99 wind farms in existence in 2004. My data, which is the source dataset, has only 94 wind farms that satisfy that criterion. It is not clear how his moving averages deal with the first and final months of data. He has imposed a constant annual rate of decline in performance over the lifetime of a plant, whereas my original study allows the rate of decline to vary with age. He reports rates of decline in the normalised load factors from age 1 to age 10. It is not clear whether he has excluded all data for wind farms of age 0 from the estimation of his fitted lines. Separately, it seems that his basic model measures age in months, so it is not clear how the age 1 estimates are calculated – are they for the 13th month of operation or are they an average of the values for months 13 to 24? To deal with these differences the comparisons reported use a common set of data and methods other than for the factors discussed in the text.

analysis takes no account of the fact that the same sites are observed at repeated intervals. This leads to serial correlation of the errors as a result of persistent site effects, which can cause the coefficient on age to be more or less severely biased. On testing it turns out that this bias is particularly large in months for which the estimated rate of decline is especially high, i.e. during the winter. Hence, it is essential to strip out site effects and I have done this by applying the usual within transformation.

The closest approximation that I can get to Professor Mackay's model yields an average annual rate of decline from year 0 onwards of 2.05% per year with a 95% confidence interval of 1.27% - 2.82%. Removing the duplication implied by use of moving averages rather than the simple monthly load factors increases the average annual rate of decline to 2.35%with a 95% confidence interval of 1.53% - 3.17%. As a comparison, the figures in my Figure 9B, which refer to the equivalent log-linear model, imply rates of decline of (i) 4.5% per year over 10 years and 5.1% per year over 15 years for the quadratic model, and (ii) 3.4% per year over 10 years and 6.1% per year over 15 years for the full age effects.

Clearly there are differences in our estimates of the rate of decline but there is no doubt that the performance of wind farms does decline with age. What is left is the task of assessing the average rate of decline and how this varies with other characteristics of wind farms. My analysis goes beyond Professor Mackay's in two respects.

First, there seems to be a clear indication in my analysis that the rate of decline in performance accelerates as wind farms get older, so that models which impose a constant rate of decline perform less well than ones which do not include this restriction.

Second, my analysis indicates that larger wind farms – and, probably, larger wind turbines – experience a more rapid decline in performance than smaller ones. This is tied up with the issue of capacity weighting. Professor Mackay's analysis can shed no light on this issue because there were only 3 wind farms with a capacity > 30 MW in existence in 2004 whereas there were 27 that commenced operations after 2004 including all wind farms of > 60 MW capacity. Large wind farms tend to rely upon large wind turbines (with a capacity of > 2 MW) which only came into widespread use after 2002. Work that I have carried out since my REF paper was written suggests that wind farms with large wind turbines have a distinctive pattern of performance. They have higher load factors than wind farms with smaller turbines up to age 4 and then their performance starts to decline much more rapidly so that by age 8 they are no better than the average and the rate of decline suggests that they will be significantly worse by age 10 and thereafter.

The idea that the performance of wind farms declines with age is regarded as perfectly normal by academic and other independent engineers. It is true for other electro-mechanical equipment subject to large stresses, so why would wind turbines be any different? There is an important question of whether the rate of decline can be slowed by adopting the best operational and maintenance practices. Since I have no access to data on O&M expenditures or regimes, I can only assess rates of declines in performance under the average O&M regime.

In my presentations and my paper I have consistently emphasised that my work should be the beginning of a research program. I believe that my results establish that there is a strong *a priori* case for believing that the decline in the performance of wind farms with age is sufficiently large to be a significant factor in the economics of wind generation. Professor Mackay's results seem to reinforce that conclusion. As more data on the output from wind farms is accumulated it should be possible to strengthen the empirical analysis of the performance of wind generation, especially from age 8 onwards, which is critical to assessing the longevity of and returns to investment in the sector. Technical issues of identification and estimation matter, but they are routine and easily dealt with. What we need to focus on is how we can take better account of factors which may affect performance but which could not be incorporated in my models. An example is variations in wind resources, which both I and others are attempting to examine in more detail.