Chapter 1

A dynamic test for misspecification of a linear model

Michael P. McAssey
University of California, Davis

Fushing Hsieh
University of California, Davis

Various linear models may be specified when attempting to estimate an apparent linear trend in real data sets, based on assumptions about the possibly unknown mechanism for generating the data. However, in many cases the best-fit line obtained from implementation of the specified model does not agree at all with the apparent trend in a scatterplot of the data, based on the evident geometric structure. We use both real and simulated data to illustrate the consequences of misspecifying a particular linear model for situations in which that model is inappropriate, and propose a dynamic test which may be used to determine when its specification is appropriate.

1.1. Introduction: Modeling the linear trend in real bivariate data

The World Health Organization (WHO) established the Multinational MONItoring of trends and determinants of CArdiovascular disease (MONICA) during the 1980s to study the association between known risk factors, like smoking and obesity, and trends in cardiovascular disease. The linear association between data on the average annual change in the observed risk score (X) and the average annual change in event rate (Y), both given as percentages, should be modeled using a measurement error structure, with the sampling error in the statistics taken as the heteroscedastic measurement error (Kulathinal, et al. (2002)). However, whether or not we account for the measurement error, we obtain about the same estimate of the slope of this linear trend when the specified model includes an addi-

tive equation error component, as in the simple linear model. In this case, the resulting best-fit line does not correspond with the perceived trend based on the geometric structure of the data cloud. Figure 1.1 displays the data scatterplots for males ($N = 38$) and for females ($N = 36$) separately, along with the best-fit lines derived under the ordinary least squares (OLS) method. The OLS method is based on the simple linear model (SLM), in which the ith response Y_i is associated with the covariate X_i by the relation $Y_i = \alpha + \beta X_i + \varepsilon_i$, for $i = 1, \ldots, n$. The additive equation error component ε_i is a (usually Gaussian) random variable with mean zero and variance σ^2, with $\varepsilon_1, \ldots, \varepsilon_n$ independent.

If one examines either data scatterplot and uses raw intuition—based on the apparent geometry of the data scatter—to project the location of the line which most suitably captures the linear association between the two variables, rather than using mathematical rigor, he would likely conclude that the slope of the plotted best-fit line based on the SLM is significantly smaller than the projected slope. A similar best-fit line is obtained when the measurement error is incorporated into the model using approaches whose structures are akin to that of the SLM. This apparent underestimation of the slope occurs because specification of the SLM for these data fails to capture the underlying mechanism which generated the data.

Since the data-generation mechanism which produced the geometric structure of the data cloud is often unknown, an investigator who wishes to derive reliable estimates for a linear trend in the data requires simple but effective tools to determine the most appropriate model for that trend among those available. In this paper, we propose such a tool and demonstrate its implementation for identifying when a proposed linear model is inappropriate. While the examples presented here involve misspecification of the SLM, this dynamic test is intended as a tool for identifying misspecification of any linear model.

1.2. Illustration: Slope underestimation with the SLM

To further illustrate the necessity for such a tool, we generate a random sample of 100 points scattered uniformly between -2 and 2 about the line $Y = 0$ in a thick cloud, so that the trendline has a slope of zero. In this case, the data-generation mechanism is inconsistent with the SLM. We rotate the trendline and each point in the data cloud about the origin through an angle greater than 45°, say 51.3°, so that the resulting data cloud should properly be perceived to follow a linear trend whose slope is

$\tan(51.3°) = 1.25$, based on the underlying data-generation mechanism we have specified. However, if OLS is applied to estimate the slope of the trend, the OLS slope estimate is only 0.98 — a significant underestimate. As we increase the angle of rotation, the underestimation of the slope by OLS becomes more pronounced. Figure 1.2 displays the data cloud along with its actual trendline and the OLS best-fit line for six different progressively steeper linear trends. Clearly, if a researcher attempted to estimate the linear association between the two variables in any of the displayed scenarios using OLS, the result would not guide the researcher to the truth about that association. While subsequent diagnostic tools might pick up the lack-of-fit in the most extreme cases, they may well miss it in more subtle cases. The data-generation mechanism which produces the geometric structure of a data cloud must not be ignored, and any model which does not properly capture the linear trend in light of that structure must be rejected.

In Figure 1.3, the slope of the data-generating trendline as it makes a complete rotation through 2π radians is shown, along with the corresponding estimated slope based on OLS regression on the generated data. The larger the magnitude of the slope of the actual trend, the greater the underestimation when OLS is applied. This illustrates a common outcome of misspecifying the SLM: the slope estimate is pulled toward zero as the trend becomes steeper. The behavior of the estimated slope under OLS as the actual trendline rotates depends on the extent to which the data-generation mechanism differes from the SLM. In a scenario in which the data are indeed generated from a SLM (with a reasonably small error variance) and the OLS slope estimate is computed as the plane rotates, the plot of the slope estimate follows that of the trendline slope very closely even when it becomes quite steep, as displayed in Figure 1.4. This is how the plot of slope estimates for rotating centered data should appear when the model is specified correctly—like the plot of the tangent function. Of course, as the error variance increases this ideal result will deteriorate even when the model is properly specified.

1.3. A test for misspecification of a linear model for real data

This illustration suggests a dynamic hypothesis test for misspecification of any linear model. Suppose we are presented with bivariate data, and we hypothesize that the association between the variables might be adequately explained by a particular linear model M_0 among linear models in a class $\mathcal{M}$.

- Center the data about zero for both variables, select one variable as the covariate, apply an estimation procedure consistent with M_0, and obtain a slope estimate $\widehat{\beta}$ and an estimate $\widehat{\sigma}^2$ for the error variance.
- Rotate the plane completely about the origin in small increments and compute a new slope estimate at each step. Plot the progression of the slope estimates against the angle of rotation, as in Figures 1.3 and 1.4.
- Meanwhile, generate a test data set from M, using the same centered values of the covariate, and the initial estimates $\widehat{\beta}$ for the slope and $\widehat{\sigma}^2$ for the error variance.
- Apply the same estimation procedure to compute slope estimates for the test data as the plane is rotated about the origin, and add the progression of these slope estimates to the previous plot.

If the original data were generated in a manner consistent with M, the two plots should be similar in amplitude, phase and shape.

Of course, variability is expected in the generated test data, especially when $\widehat{\sigma}^2$ is large. Thus we generate N test data sets in the same way, and repeat the rotating slope estimation procedure on each set. At each increment of the rotation we will have a distribution of N slope estimates. We may select an upper and lower quantile for this distribution at each angle and thereby obtain an empirical pointwise confidence band for the path of the slope estimate if the original data may indeed be regarded as having been generated from the M. If the path of slope estimates for the original data does not lie entirely within this band, we may reject the null hypothesis: that M is a suitable model for the generation mechanism of the original data, and conclude at the corresponding significance level the alternative hypothesis: that M is not a suitable model for the generation of the data. Note that, in the event the null hypothesis is rejected, this hypothesis test does not recommend an alternative model for consideration.

For instance, consider again the WHO MONICA data for males. When we implement the procedure described above, hypothesizing that M is the SLM, we obtain the results shown in the left panel of Figure 1.5. The progression of OLS slope estimates as the centered real data rotate about the origin is represented by the solid line. We generate $N = 1000$ data sets under the SLM, using the OLS slope estimate for the unrotated real data, and a robust estimate of the error variance based on the middle 50% of the ordered residuals. The pair of dashed lines represents the middle

50% of OLS slope estimates for the rotating generated data, and the pair of dotted lines represents the middle 90%. Clearly, the path for the real data does not fall within either band, since its amplitude and phase are quite different, so that we may conclude at the 0.1 significance level that the SLM would be a misspecification for these data. When we perform the same experiment with the centered WHO MONICA data for females, we find that the progression of OLS slope estimates also lies outside the confidence bands, as shown in the right panel of Figure 1.5. Hence the null hypothesis would also be rejected for the female data. We may therefore further suspect the suitability to these data of any measurement error model which involves an equation error component, as such models produce slope estimates very close to those obtained using OLS (as in Kulathinal, et al. 2002) and Patriota, et al. 2009).

Now consider an alternative data-generating mechanism for these data. Suppose we use orthogonal least squares to estimate the slopes for each data set, based on the linear model which regards the observations as perpendicular deviations from the trendline. In this model, $Y_i = \beta X_i + \nu_i$ for $i = 1, \ldots, n$, with independent errors $\nu_i = \varepsilon_i \sqrt{\beta^2 + 1}$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The progression of these slope estimates as the centered real data for males and for females rotate about the origin is represented by the solid line in each plot of Figure 1.6. We also generate $N = 1000$ data sets under this model, using the orthogonal least squares slope estimate for the unrotated real data, and an estimate of the error variance based on the entire collection of residuals. The pairs of dashed lines represents the middle 50% and 90% of slope estimates for the rotating generated data. But since there is almost no variability, both of these confidence bands are very narrow—almost superimposed. In this scenario, the path for the real data comes very close to falling within the bands, especially for the females, despite the narrowness of the bands. Moreover, the paths of the estimates have the tangent-function appearance that we expect when the model is specified well. Hence the slope estimate is slightly biased in each case, so that a bias correction of appropriate size should be added to the slope estimate to bring it within the confidence bands. This would produce a model that will pass the dynamic test proposed here.

Once we make the necessary bias corrections to the slope estimates, the amended model passes our hypothesis test. Hence when the measurement error for the observations on each variable is considered for these data, it would be wise to implement a model whose structure is more akin to this latter model, and less like the SLM, as presented in McAssey and Hsieh

(2010). In Figure 1.7, the trendlines for males and females based on the slope estimates derived using orthogonal least squares (OM) are added to the scatterplots given in Figure 1.1. It is intuitively apparent that the trajectories of these two trendlines correspond much more closely to the geometric structure of the respective data clouds.

1.4. Discussion

In many data analyses one has limited knowledge about the underlying data generation mechanism, so that the specification of a linear model may become an arbitrary choice. The above examples demonstrate that the cost of misspecification can be high. Hence researchers must be wary of specifying a particular linear model in the analysis of the linear association between two variables unless there is convincing evidence that the data were generated in a manner consistent with that model. We present a simple test that may be used to determine whether the specified model is a plausible representation of the underlying data generating mechanism for the observed data. A careful researcher may implement this simple test as a tool to assist in discriminating among proposed linear models and corresponding estimation procedures for analysis of real data.

Acknowledgements

This research is partially supported by the National Science Foundation under Grant No: HSD 0826844, and by the National Institutes of Health under Grant No: 1R01AG025218-01A2.

The authors thank Kari Kuulasmaa for making the WHO MONICA data available, and Alexandre Patriota for sharing the data.

References

Kulathinal, S. B., Kuulasmaa, K. and Gasbarra, D. (2002). Estimation of an errors-in-variables regression model when the variances of the measurement errors vary between the observations. *Statistics in Medicine*, **21**, 1089–1101.

McAssey, M. and Hsieh, F. (2010). Slope estimation in structural line-segment heteroscedastic measurement error models. *Statistics in Medicine*, **29**, 2631–2642

Patriota, A. G., Bolfarine, H. and de Castro, M. (2009). A heteroscedastic structural errors-in-variables model with equation error. *Statistical Methodology*, **6**, 408–423.

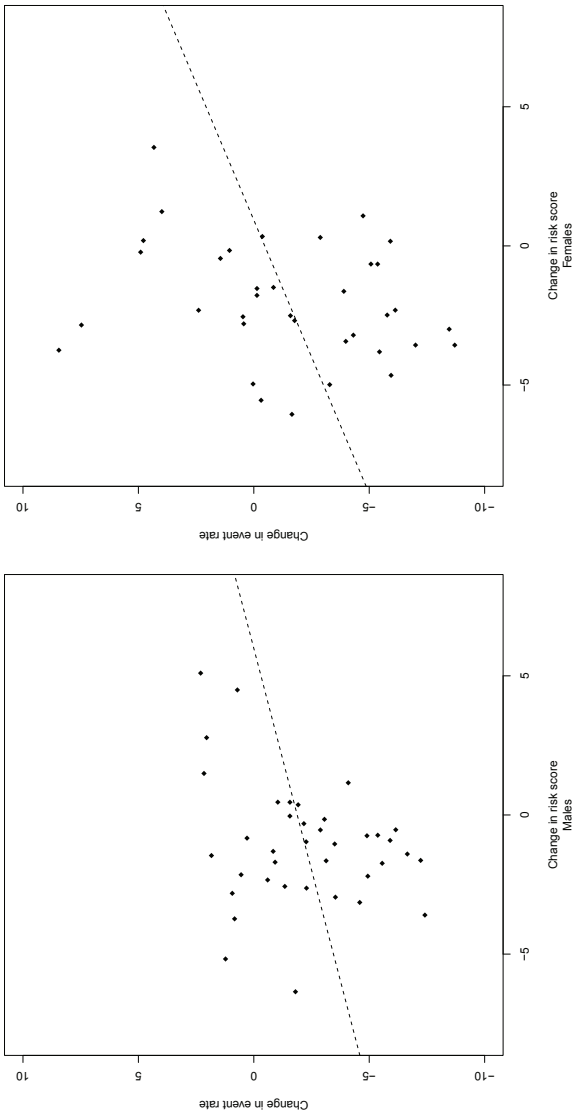

Fig. 1.1. Scatterplot of change in event rate versus change in risk score, from WHO MONICA project, and best-fit lines based on OLS regression, for males and females

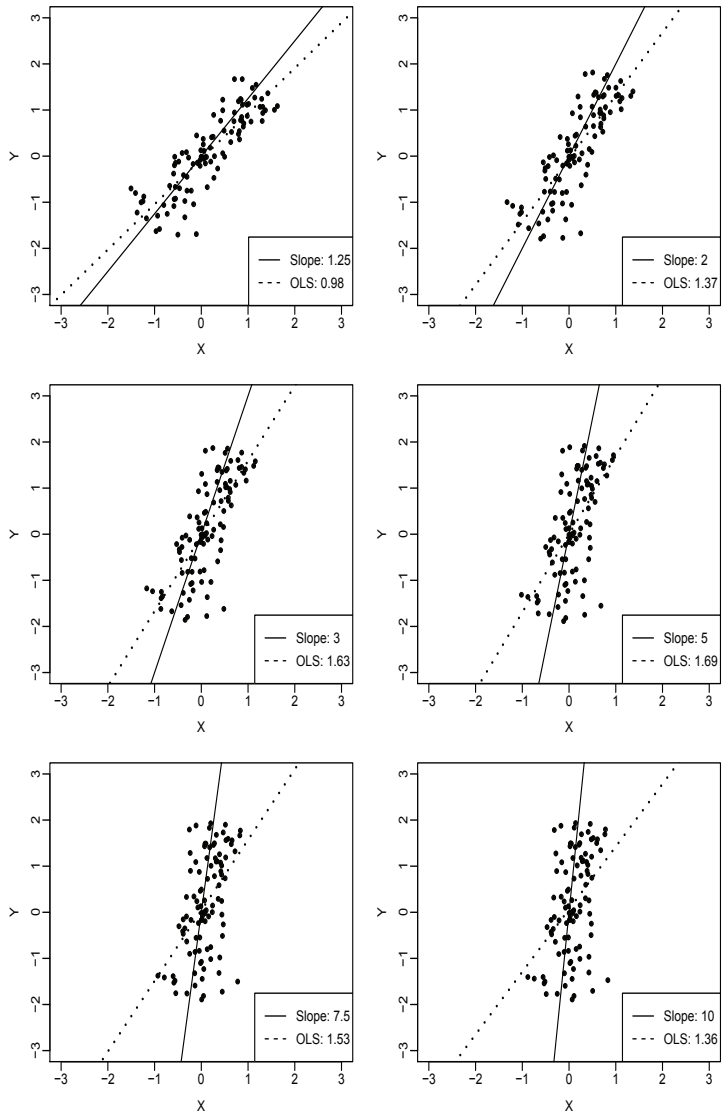

Fig. 1.2. Rotating data cloud, with the underlying linear trend and the estimated trend under OLS regression

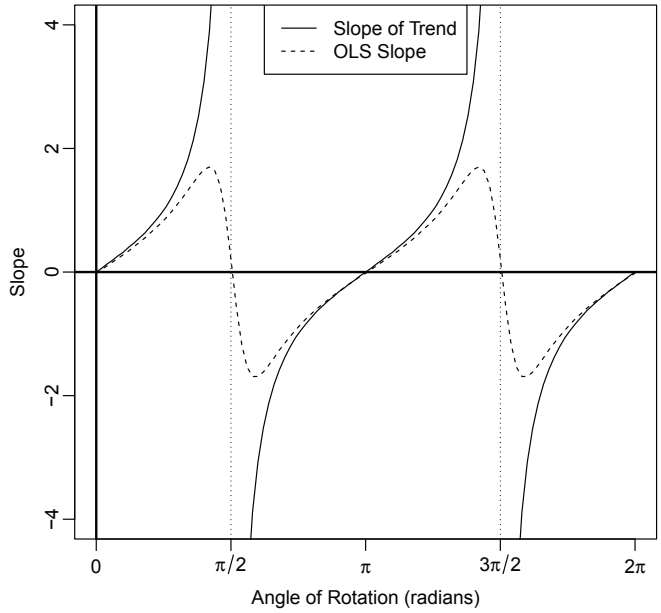

Fig. 1.3. Progression of the slope of the trendline for the rotating centered data cloud, and the corresponding estimated slope using OLS regression for data which are not generated from the SLM, as the plane rotates about the origin

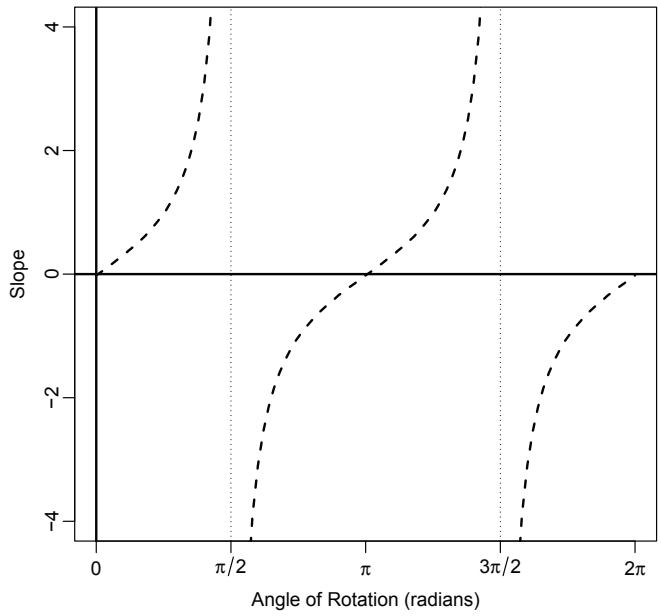

Fig. 1.4. Progression of the value of the slope estimate using OLS regression as the plane rotates about the origin when the centered data are generated from the SLM with a small error variance

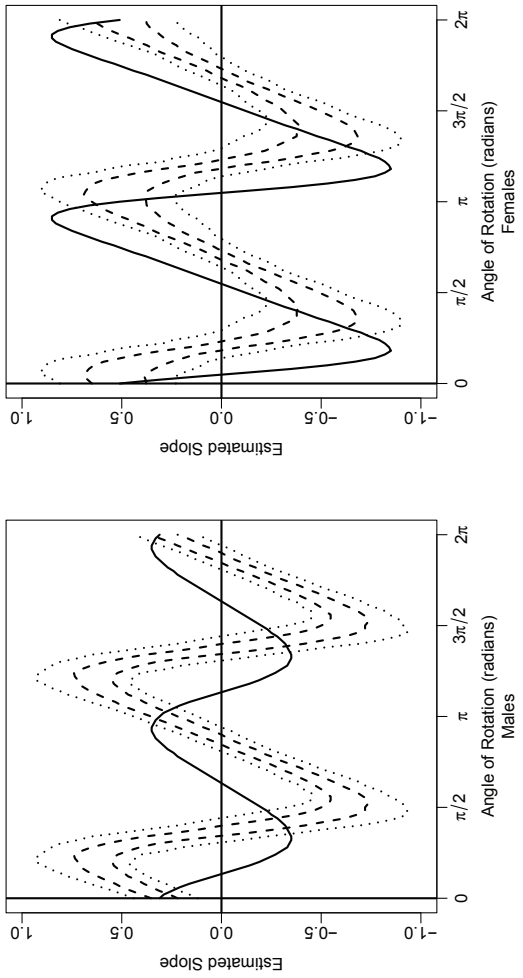

Fig. 1.5. Progression of the value of the slope estimate using OLS regression on the centered WHO MONICA data for males (left) and for females (right) as the plane rotates about the origin, along with 50% (dashed) and 90% (dotted) pointwise confidence bands for a specified SLM based on 1000 generated data sets

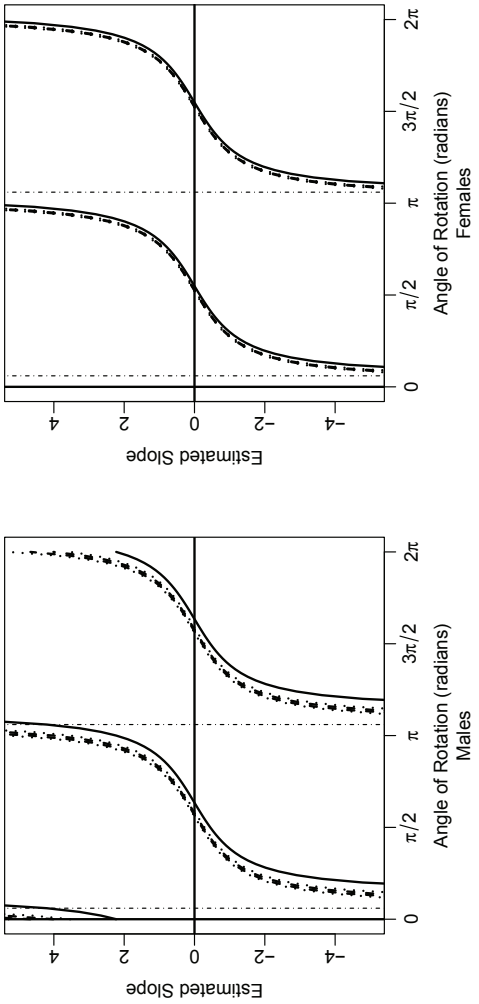

Fig. 1.6. Progression of the value of the slope estimate using orthogonal least-squares regression on the centered WHO MONICA data for males (left) and for females (right) as the plane rotates about the origin, along with 50% (dashed) and 90% (dotted) pointwise confidence bands for a specified perpendicular-deviation model based on 1000 generated data sets

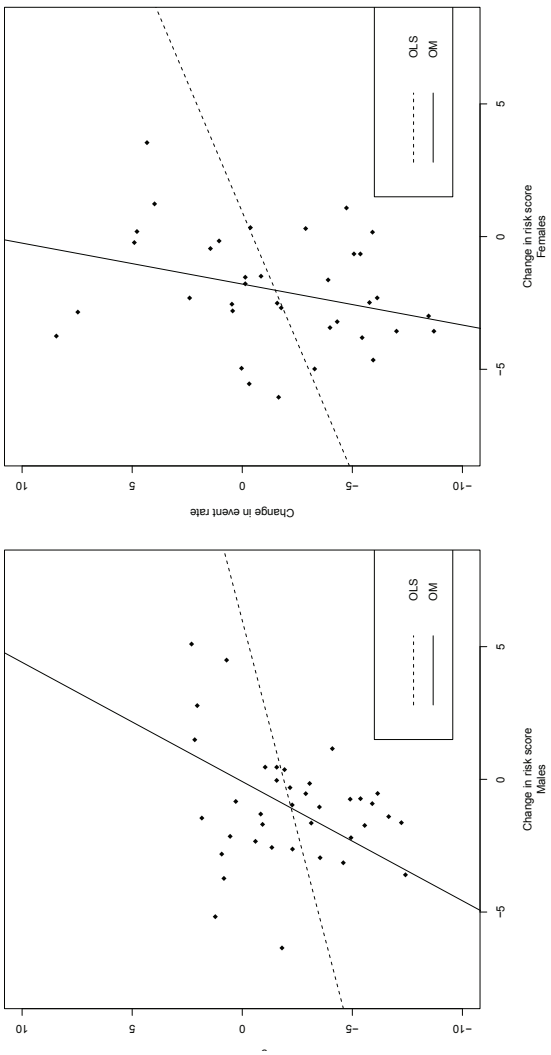

Fig. 1.7. Scatterplot of change in event rate versus change in risk score, from WHO MONICA project, and best-fit lines based on both orthogonal least-squares (OM) and ordinary least-squares (OLS) regression, for males and females