This is an Accepted Manuscript of an article accepted by Taylor & Francis for publication in Technometrics and posted online 19 July 2016, available online: http://www.tandfonline.com/10.1080/00401706.2016.1212737.

An exact test of fit for the Gaussian linear model using optimal nonbipartite matching

Samuel D. Pimentel*
Dylan S. Small
Paul R. Rosenbaum
Department of Statistics, Wharton School, University of Pennsylvania

May 23, 2016

Abstract

Fisher tested the fit of Gaussian linear models using replicated observations. We refine this method by (1) constructing near-replicates using an optimal nonbipartite matching and (2) defining a distance that focuses on predictors important to the model's predictions. Near-replicates may not exist unless the predictor set is low-dimensional; the test addresses dimensionality by betting that model failures involve a subset of predictors important in the old fit. Despite using the old fit to pair observations, the test has exactly its stated level under the null hypothesis. Simulations show the test has reasonable power even when many spurious predictors are present.

Keywords: Combinatorial optimization; network optimization; Tukey's test for nonadditivity.

* *Address for correspondence*: Samuel D. Pimentel, Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340 USA. E-mail: spi@wharton.upenn.edu. Supported by National Science Foundation Grant SES-1260782 from the Measurement, Methodology and Statistics Program of the NSF and by Fellowship FA9550-11-C-0028 from the Department of Defense, Army Research Office, National Defense Science and Engineering Graduate (NDSEG) Fellowship Program, 32 CFR 168a4.

1 Notation and review

1.1 The Gaussian linear model

The familiar Gaussian linear model assumes that an n-dimensional stochastic outcome $\mathbf{y}$ and an $n \times p$ dimensional fixed matrix $\mathbf{X}$, with $p < n$, are related by

$$\mathrm{E}\left(\mathbf{y} \middle| \mathbf{X}\right) = \mathbf{X}\boldsymbol{\beta}, \quad \boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \sim N_n\left(\mathbf{0},\, \sigma^2\mathbf{I}\right), \tag{1}$$

where $\boldsymbol{\beta}$ and σ^2 are unknown parameters, $\mathbf{0}$ and $\mathbf{I}$ are, respectively, the n-dimensional zero vector and identity matrix, and $N_n(\cdot, \cdot)$ is the n-dimensional multivariate Normal distribution. A test of fit of (1) is a test of the null hypothesis H_0 that (1) is true, and such a test is said to be exact, as opposed to asymptotic — that it, the test has exact level α — if the probability that the test rejects H_0 when it is true is $\leq \alpha$. Generally, we assume that $\mathbf{X}$ has full column rank p, so the least squares estimate of $\boldsymbol{\beta}$ under (1) is $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$, the fitted values are $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$ where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$, the residuals are $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$, and the unbiased estimate of σ^2 is $\widehat{\sigma}^2 = \mathbf{e}^T\mathbf{e}/\left(n - p\right)$.

1.2 Tests of fit based on replicates and near-replicates

Fisher (1922) proposed testing the Gaussian linear model in experiments by including replicates of design points, thereby providing an estimate of pure error unaffected by misspecification of the linear model, yielding an exact F-test of H_0. This device is commonly used in central composite designs in response surface experiments, with the center-point replicated several times, and the factorial and axial points appearing only once in isolation; see Box and Draper (1982) and Draper (1982).

Outside of designed experiments, exact replicates occur sporadically if at all. Several investigators have proposed an analogous test based on near-replicates; see, for instance, Christensen (1989; 1991; 2011, §6.6.2), Daniel and Wood (1971, §7.5), Green (1971), Joglekar et al. (1989), Neill and Johnson (1985), Shillington (1979), and Su and Yang (2006). This work emphasizes certain options in the choice of test statistic, whereas our contribution emphasizes the construction of the near-replicates. In particular, we use opti-

mal nonbipartite matching, reviewed in §1.3, and the device that Tukey (1949) introduced in constructing his "one degree of freedom for nonadditivity" test in the unreplicated row-by-column design. For different approaches to constructing near-replicates, see Miller, Neill and Sherfey (1998, 1999) and Miller and Neill (2008).

1.3 Optimal nonbipartite matching

Given L points with L even and an $L \times L$ symmetric matrix of nonnegative distances between pairs of points, an optimal nonbipartite match divides the L points into $L/2$ nonoverlapping pairs of two points so that the total of the $L/2$ within-pair distances is minimized. This combinatorial optimization problem may be solved in polynomial time by a suitable algorithm; see Jungnickel (2013, §14.4). In `R`, the `nbpMatching` package of Lu et al. (2011) makes available the algorithm of Derigs (1988), and we used it in the current paper.

Nonbipartite matching has been used to solve various problems in observational studies, including matching with time-dependent propensity scores (Lu 2005) and strengthening instrumental variables (Baiocchi et al. 2010, Zubizarreta et al. 2013). See Lu et al. (2011) for a survey of statistical applications of nonbipartite matching. Here, we use optimal nonbipartite matching as one aspect of constructing near replicates. For general discussion of optimal matching in observational studies, see Rosenbaum (2010, Part II) and Stuart (2010).

In the statistical applications described above, it is common to form pairs using only some of the available observations, with the algorithm itself deciding which observations to leave unpaired. This is done using so-called "sinks". Suppose that there are n observations with an $n \times n$ distance matrix and we want m pairs, with specified $m \leq n/2$. Then $n - 2m$ observations are not paired. If the $n \times n$ distance matrix contains any zeros off the diagonal, then we add a constant, say 1, to all of the off-diagonal entries, so they are all strictly positive. Introduce $n - 2m$ sinks that are at 0 distance to all observations and at infinite distance to one another. That is, expand the distance matrix with 3 blocks, a block of extra columns of 0's of dimension $n \times (n - 2m)$, a block of extra rows of 0's of dimension $(n - 2m) \times n$, and a square lower-right-corner block of ∞'s of dimension

$(n-2m)\times(n-2m)$. One then calculates an optimal nonbipartite match with this expanded distance matrix, regarding any observation paired with a sink as unpaired. This strategy forms m pairs of observations in such a way that the total of the m within pair distances is minimized over two choices: (i) which $n-2m$ observations to leave unpaired, and (ii) how to best pair the $2m$ observations that are paired.

In central composite experimental designs, only central points are replicated. In non-experimental data, there are often many points that have no near-replicate. Motivated by these considerations and some preliminary simulations, we leave approximately $(n-p)/3$ observations unpaired, pairing the rest. When n is large compared to p, nearly a third of the observations are unpaired and two thirds are paired, leaving nearly $n/3$ degrees of freedom within-pairs to estimate error from near replicates. Stated precisely, we form m pairs where m is $n/2-(n-p)/6$ rounded to the nearest integer, and we leave exactly $n-2m$ observations unpaired using $n-2m$ sinks. Here, $n-2m$ is approximately $(n-p)/3$.

1.4 Tukey's device and its extensions

A well-known problem with techniques that rely on near-neighbors or near-replicates is that, unless the number of predictors is very small, we will rarely see two individuals who are nearly the same on all of the predictors. In light of this, we need to define the distance with some guidance from the data about which predictors actually matter for prediction. At the same time, we need to prevent this double use of the $\mathbf{y}$'s from invalidating the test. For this purpose, a device introduced by Tukey (1949) is helpful.

Tukey (1949) proposed a test for interaction in the unreplicated row-by-column design using the following clever device. The device has been generalized several times, and we describe the generalized form for Gaussian linear models here; see, for instance, Mandel (1959), Scheffé (1959, Problem 4.19), Milliken and Graybill (1970), Andrews (1971), Rao (1973, §4e.1), St. Laurent (1990), Christensen and Utts (1992) and Christensen (2011, §9.5). A basic fact about the distribution of $\boldsymbol{\epsilon}$ in (1) is that projections of $\boldsymbol{\epsilon}$ onto orthogonal subspaces are independent; this fact is the key element in the Fisher-Cochran theorem. In particular, the fitted values $\widehat{\mathbf{y}}=\mathbf{H}\mathbf{y}$ and residuals $\mathbf{e}=\mathbf{y}-\widehat{\mathbf{y}}=(\mathbf{I}-\mathbf{H})\,\mathbf{y}$ are independent in (1). Write $\rho(\mathbf{M})$ for the rank of a matrix $\mathbf{M}$. Let $\mathbf{L}$ be any matrix with n rows that

is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$ such that $\rho([\mathbf{X}, \mathbf{L}]) < n$. It is easily seen that

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}^\mathbf{T}\mathbf{H}\mathbf{y} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\mathbf{y}}, \tag{2}$$

so that $\widehat{\boldsymbol{\beta}}$ is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$, and in particular, $\mathbf{L}$ can be a function of $\widehat{\boldsymbol{\beta}}$, as in Tukey (1949). Milliken and Graybill (1970, §2) observe that if (1) is true and $\mathbf{y}$ is regressed on $[\mathbf{X}, \mathbf{L}]$, then the usual F-test of the hypothesis that the coefficients of $\mathbf{L}$ are simultaneously zero has a central F-distribution with degrees of freedom $\rho([\mathbf{X}, \mathbf{L}]) - p$ and $n - \rho([\mathbf{X}, \mathbf{L}])$. Here, $[\mathbf{X}, \mathbf{L}]$ need not have full column rank, but must have rank less than n. As discussed by Milliken and Graybill (1970), the distribution of this F-statistic under the alternative that (1) is false is not, in general, a noncentral F-distribution and is typically intractable.

2 An exact test of fit for the Gaussian linear model

2.1 General procedure

Starting with a suitable distance matrix, we round $n/2 - (n-p)/6$ to the nearest integer to obtain m, as in §1.3, and we use optimal nonbipartite matching in §1.3 to build m pairs of two observations and $n - 2m$ unpaired observations so that the total distance within the m pairs is minimized. This match is intended to find the closest m pairs of near replicates and $n - 2m$ individuals who are further away, yielding roughly m degrees of freedom from near replicates to estimate an error variance less affected by any misspecification of model (1). Define $\mathbf{L}$ to be a matrix with $m + n - 2m = n - m$ columns, where the first m columns of $\mathbf{L}$ each contain exactly two ones and $n - 2$ zeros, the two ones in column k indicating the two individuals paired in pair k, $k = 1, \ldots, m$. The last $n - 2m$ columns of $\mathbf{L}$ each contain a one and $n - 1$ zeros, the 1 indicating the ℓth individual who was not paired, $\ell = 1, \ldots, n - 2m$. Notice that the $n - 2m$ unpaired individuals each have their own column and will be fitted exactly, somewhat in parallel with the proposal of Utts (1982); see also Christensen (2011, p. 153). The $n - m$ columns of $\mathbf{L}$ have rank $\rho(\mathbf{L}) = n - m - 1$ because each row of $\mathbf{L}$ sums to 1. In general, the rank of $\rho([\mathbf{X}, \mathbf{L}])$ will depend on $\mathbf{X}$.

The test of fit of (1) is simply an F-test of $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ in the Gaussian linear model

$$\mathbf{y} = [\mathbf{X},\, \mathbf{L}] \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} + \boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \sim N_n\left(\mathbf{0},\, \omega^2 \mathbf{I}\right) \tag{3}$$

with degrees of freedom $\rho\left([\mathbf{X},\, \mathbf{L}]\right) - p$ and $n - \rho\left([\mathbf{X},\, \mathbf{L}]\right)$. In our proposed test, the residual degrees of freedom, $n - \rho\left([\mathbf{X},\, \mathbf{L}]\right)$, will approach $n/3$ as $n \to \infty$ with p fixed. Because $[\mathbf{X},\, \mathbf{L}]$ is not of full column rank, a little care, of a conventional sort, is needed in computing the F-test.

Christensen (1991) proposed an alternative modified test, no longer the standard F-test of $H_0 : \boldsymbol{\gamma} = \mathbf{0}$, with a view to gains in power. Our limited simulation (not shown) comparing the standard F-test to this modified test suggests that the dimension reduction devices we describe later have large effects on power, while the choice of test statistic has a smaller effect, so we adhere to the standard F-test in our discussion here, rather than add an extra dimension to our simulated comparisons.

If the distances were based on $\mathbf{X}$ alone, then $[\mathbf{X},\, \mathbf{L}]$ would be a function or transformation of $\mathbf{X}$, and the test of (1) against (3) is simply a comparison of two nested Gaussian linear models. If the null hypothesis, namely (1), were true, then $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ is true in (3) and, in the standard way, the corresponding F-statistic has a central F-distribution; see, for instance, Rao (1973, §4b.2) or Christensen (2011, §3.2). In §2.2, we permit $\mathbf{L}$ to depend upon both $\mathbf{X}$ and $\widehat{\mathbf{y}}$; then, the corresponding F-statistic is no longer a standard test of a general linear hypothesis, but it still has a central F-distribution when the null hypothesis (1) is true using the generalization of Tukey's device; see Milliken and Graybill (1970), Rao (1973, pp. 251-252) or Christensen (2011, §9.5).

For instance, the distance matrix could be the Mahalanobis distance between pairs of rows of $\mathbf{X}$. The usual Mahalanobis distance can perform oddly when a column of $\mathbf{X}$ is either long tailed or a rare binary variable. An alternative robust Mahalanobis distance addresses both issues: it replaces the columns of $\mathbf{X}$ by column ranks before computing the distances, with average ranks for ties; however, it uses untied variances and covariance of ranks, thereby reducing the role of rare binary variables; see Rosenbaum (2010, §8). As is commonly done, we speak of the quadratic form as the Mahalanobis distance (or the robust Mahalanobis distance), whereas technically it is its square root that is a norm.

The estimate of ω^2 in (3) may be smaller than the estimate of σ^2 in (1) for two reasons. First, the estimate of ω^2 only reflects differences in y's between paired individuals, and paired individuals are as close as possible on the predictors. Second, the $n-2m$ individuals who were not paired do not contribute to the estimate of ω^2 because they are fitted exactly in (3). These $n-2m$ unpaired individuals are each far from all other individuals. If some of these $n-2m$ unpaired individuals are poorly fit by (1), eliminating them from the estimate of ω^2 may aid in recognizing this lack of fit.

Model (3) creates an estimate $\widehat{\omega}^2$ from neighbors that may be less affected by model misspecification than $\widehat{\sigma}^2$ obtained from fitting model (1). The parameter $\boldsymbol{\gamma}$ is of high dimension and is not typically of interest, so one may use computational simplifications with the highly structured matrix $\mathbf{L}$ to obtain the F-test without estimating $\boldsymbol{\gamma}$.

When the number of predictors is not small, close matches on all predictors will be rare. An alternative distance matrix is discussed in §2.2: it emphasizes the predictors that appear to matter in the fit of model (1), but avoids double use of the $\mathbf{y}$ by employing Tukey's device from §1.4.

2.2 Using y in the construction of the distance matrix

In principle, failures of model (1) could involve any of the predictors in the model. With just a few predictors, all of them could be used to define the distance. In other cases, it will often seem reasonable to bet that failures of model (1) involve predictors that exhibit some predictive power in the fit of model (1). For instance, this might be true if either $\mathbf{y}$ or a predictor requires a monotone increasing transformation, or if two important predictors require inclusion of their interaction.

Tukey's method in §1.4 permits the matrix $\mathbf{L}$ to be any function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$. In particular, Tukey's method yields a central F-distribution for the test statistic if $\mathbf{L}$ is built from an optimal nonbipartite match using a distance matrix that is itself a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$. One such very simple distance matrix has as a distance between individuals i and i' the absolute difference in their predicted values, $|\widehat{y}_i - \widehat{y}_{i'}|$. However, in addition to matching for $\widehat{\mathbf{y}}$, it makes sense to also match for several of the most important predictors.

Define d_j to be the square root of the jth diagonal element of $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$. The usual

t-statistic testing the hypothesis that the jth coordinate β_j of $\boldsymbol{\beta}$ in (1) is 0 is $\widehat{\beta}_j / (d_j \widehat{\sigma})$ where $\widehat{\beta}_j$ is the jth coordinate of $\widehat{\boldsymbol{\beta}}$. This t-statistic depends on $\widehat{\sigma}$, so it depends on the residuals $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$ and not just on the fitted values, $\widehat{\mathbf{y}}$, so the generalization of Tukey's method in §1.4 does not permit the use of this t-statistic. In contrast, the quantity $\widehat{\beta}_j / d_j$ is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$ because of (2). Because $\widehat{\sigma}$ in the t-statistic $\widehat{\beta}_j / (d_j \widehat{\sigma})$ is the same for all predictors, we may identify the r predictors with the largest absolute t-statistics as the r predictors with the largest $\left| \widehat{\beta}_j / d_j \right|$, which is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$. To emphasize, we can use the generalization of Tukey's method if we select a fixed number, r, of variables with the largest t-statistics because we can identify those variables using $\mathbf{X}$ and $\widehat{\mathbf{y}}$, but we cannot select all variables with, say, $\left| \widehat{\beta}_j \right| / (d_j \widehat{\sigma}) \geq 2$, because that makes use of $\widehat{\sigma}$.

The proposed test computes the robust Mahalanobis distance from $\widehat{\mathbf{y}}$ and the r predictors with the largest $\left| \widehat{\beta}_j / d_j \right|$. Here, $\widehat{\mathbf{y}}$ depends upon all predictors. Because this distance is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$, as noted above, $\mathbf{L}$ too is a function of $\mathbf{X}$ and $\widehat{\mathbf{y}}$, so the generalization of Tukey's (1949) method yields a null F-distribution for the F-statistic comparing models (1) and (3).

3 Simulation study of the power of the test

Tables 1 and 2 report simulated power of a 0.05-level test for five nonlinear functions with Gaussian errors. In additional simulations not shown in Tables 1 and 2, we found that a 0.05-level test did indeed reject a linear model in close to 5% of simulated samples. In the simulation, the model (1) is fit with a constant term included in $\mathbf{X}$, so there are $p' = p - 1$ predictors aside from the constant term. There are $p' = 10$, 30 or 50 predictors and $n = 100$ or 500 observations; however, we do not consider the combination of $p' = 50$ predictors and $n = 100$ observations. In all cases, a linear model with p' predictors is mistakenly fit to various nonlinear surfaces, and the question is whether the test can recognize this mistake.

Each sampling situation is replicated 3000 times, so the standard error of a simulated power is at most $\sqrt{0.25/3000} < 0.01$. For $p' = 30$ or $p' = 50$, many of the predictors x_j, $j = 1, \ldots, p'$, do not affect the response surface, but the investigator does not know this, so the fitted model mistakenly uses all p' predictors. When $p' = 30$ or $p' = 50$, the test is looking for genuine model failures involving a few predictors amid distraction from many

irrelevant predictors.

The test is performed in nine variations, $9 = 2 \times 5 - 1$. In five of the nine variations, the optimal nonbiparitite matching paired for $\widehat{y}$, and in four variations it did not. The optimal nonbiparitite matching paired for the r predictors with the largest absolute t-statistics, for $r = 0$, 3, 5, 10 and p'. One needs to pair for something, so the case of not pairing for $\widehat{y}$ and pairing for $r = 0$ predictors does not occur, making 9 variations in total. When $p' = 10$, the last two columns of Tables 1 and 2 are identical for $r = 10$ and $r = p'$. Two consecutive rows of Tables 1 and 2 — the first with 5 estimated powers, the second with 4 estimated powers — constitute one sampling situation in which 9 methods are competing to produce the largest power. In each sampling situation, the largest power or powers are in **bold**.

In matching, we use the robust Mahalanobis distance described in §2.1. As a consequence in Tables 1 and 2, matching for all predictors, $r = p'$, and matching for all predictors plus $\widehat{y}$ are slightly different. With the conventional Mahalanobis distance, $\widehat{y}$ would be linearly dependent on the constant plus p'-predictors and hence redundant, not affecting the distance.

The five nonlinear response surfaces will now be described. In Table 1, the true nonlinear regression is $y = x_1 + x_2 + x_3 + x_4 + x_3x_4 + x_4x_5 + x_5^2 + \epsilon$ and the predictors are multivariate Normal with covariances indicated in the table. Three of the response surfaces in Table 2 were discussed and depicted by Friedman (1991) and have been used in the literature before and after 1991 as test cases of nonlinear regression surfaces. In Table 2, the predictors are independent uniform random variables, with standard Normal errors, and response surfaces given by:

Exponential: $y = \exp\left(\sum_{j=1}^{10} x_j\right) + \epsilon$

Friedman (1991), equation (56): $y = 0.1e^{4x_1} + 4/\left\{1 + e^{-20(x_2 - 0.5)}\right\} + 3x_3 + 2x_4 + x_5 + \epsilon$

Friedman (1991), equation (61): $y = 10\sin(\pi x_1 x_2) + 20\left(x_3 - 0.5\right)^2 + 10x_4 + 5x_5 + \epsilon$

Friedman (1991), equation (66):

$$y = 40\frac{\exp\left\{8\left[(x_1 - 0.5)^2 + (x_2 - 0.5)^2\right]\right\}}{\exp\left\{8\left[(x_1 - 0.2)^2 + (x_2 - 0.7)^2\right]\right\}} + e^{8\left[(x_1 - 0.7)^2 + (x_2 - 0.2)^2\right]} + \epsilon.$$

All five models involve 10 or fewer variables, so that, when $p' > 10$, many of the variables are simply distractions, as noted above.

Consider, now, the estimated powers. The strongest pattern in Tables 1 and 2 is the least interesting: the power is higher when the sample size n is larger. Setting that aside, within a sampling situation or pair of rows, the power varies dramatically among nine methods.

Should one try to match for all p' variables? When $p' = 30$ or $p' = 50$, trying to match for all p' variables usually reduces power: it is better to use the mistaken linear fit to reduce the number of variables employed in the matching, even though the mistaken fit need not be a reliable guide to the importance or role of particular variables. Also, matching for $\widehat{y}$ alone with $r = 0$ is not the best procedure in any sampling situation, and it often has low power.

How many variables should be used in the matching? In Table 1, only 5 variables affect the response surface, and good power often occurs when matching for either the $r = 5$ variables or $r = 10$ variables with the largest t-statistics. The situation is slightly more complicated in Table 2, where the first response surface involves 10 variables and $r = 10$ is best, while the last response surface involves just 2 variables and $r = 3$ or $r = 5$ is better than $r = 10$.

Should one match just for the r variables with the largest t-statistics or should one additionally match for $\widehat{y}$? In each sampling situation, this is the comparison of two adjacent rows in the same column. There is no uniform winner here, but including $\widehat{y}$ in a match for $r = 5$ or $r = 10$ variables rarely does much harm and sometimes greatly increases power. For instance, consider in Table 2 the Exponential model with $r = p' = 10$ where including $\widehat{y}$ yields power 0.99 and excluding $\widehat{y}$ yields power 0.69. Also in Table 2, consider Friedman's equation (66) model with $r = 10$, where again including $\widehat{y}$ in the matching distance increases power. In Table 1, five variables affect the response surface, so matching for $r = 3$ variables must omit relevant variables: in the $r = 3$ column of Table 1, matching also for $\widehat{y}$ often yields meaningful gains in power.

The Exponential response surface in Table 2 is interesting. In this case, $\mathrm{E}(y|\mathbf{x})$ is an increasing but nonlinear function of $x_1, \ldots, x_{10}$, whereas $x_{11}, \ldots, x_p$ are irrelevant. As

Table 1: Simulated power of a 0.05-level test with $p' = p - 1$ predictors and n observations, $y = x_1 + x_2 + x_3 + x_4 + x_3x_4 + x_4x_5 + x_5^2 + \epsilon$, where $(x_1, \ldots, x_p, \epsilon)$ is multivariate Normal, with $\mathrm{E}(x_j) = 0$, $\mathrm{var}(x_j) = 1$ and, except as noted below, $\mathrm{cov}(x_j, x_{j'}) = 0$, and with $\mathrm{E}(\epsilon) = 0$ $\mathrm{var}(\epsilon) = 1$, $\mathrm{cov}(\epsilon, x_j) = 0$. The matching either matched for $\widehat{y}$, case 1, or did not, case 0, and it matched for $r = 0$, 3, 5, 10, or all p' predictors with the largest absolute t-statistics. Each situation was replicated 3000 times. A sampling situation is two consecutive rows, and the highest power in a sampling situation is in **bold**.

Nonzero				Matched for r predictors				
Covariances	p'	n	Matched for $\widehat{y}$	0	3	5	10	p'
None	10	100	1	0.10	0.25	0.48	**0.50**	**0.50**
	10	100	0		0.19	0.48	0.48	0.48
Predictors are	10	500	1	0.13	0.64	0.87	**1.00**	**1.00**
independent	10	500	0		0.40	0.81	**1.00**	**1.00**
	30	100	1	0.07	0.11	0.13	**0.14**	0.10
	30	100	0		0.09	**0.14**	**0.14**	0.08
	30	500	1	0.12	0.56	0.79	**0.81**	0.74
	30	500	0		0.40	0.72	0.75	0.68
	50	500	1	0.13	0.50	**0.74**	0.71	0.49
	50	500	0		0.38	0.65	0.68	0.42
$\mathrm{cov}(x_1, x_5) = 0.8$	10	100	1	0.20	0.49	**0.63**	0.55	0.55
	10	100	0		0.34	0.62	0.45	0.45
Nonlinear x_5	10	500	1	0.47	0.99	**1.00**	**1.00**	**1.00**
is highly correlated	10	500	0		0.73	**1.00**	**1.00**	**1.00**
with linear	30	100	1	0.09	0.16	**0.18**	0.14	0.10
predictor x_1	30	100	0		0.12	**0.18**	0.13	0.08
	30	500	1	0.42	0.96	**1.00**	0.98	0.83
	30	500	0		0.68	1.00	0.96	0.62
	50	500	1	0.36	0.93	**1.00**	0.95	0.55
	50	500	0		0.64	0.99	0.93	0.35
$\mathrm{cov}(x_5, x_6) = 0.8$	10	100	1	0.11	0.28	0.47	**0.53**	**0.53**
	10	100	0		0.23	0.41	0.51	0.51
Nonlinear x_5	10	500	1	0.17	0.75	0.99	**1.00**	**1.00**
is highly correlated	10	500	0		0.47	0.97	**1.00**	**1.00**
with irrelevant x_6	30	100	1	0.08	0.12	**0.15**	0.14	0.10
	30	100	0		0.10	**0.15**	0.13	0.09
	30	500	1	0.16	0.68	**0.98**	0.95	0.88
	30	500	0		0.44	0.96	0.91	0.79
	50	500	1	0.14	0.62	**0.96**	0.91	0.60
	50	500	0		0.41	0.94	0.87	0.45
$\mathrm{cov}(x_j, x_{j'}) = 0.5$	10	100	1	0.57	0.52	**0.59**	0.57	0.57
for all $j \neq j'$	10	100	0		0.52	0.57	0.49	0.49
	10	500	1	0.99	0.96	**1.00**	**1.00**	**1.00**
All predictors	10	500	0		0.93	0.99	**1.00**	**1.00**
are correlated.	30	100	1	0.20	**0.22**	0.21	0.16	0.12
	30	100	0		0.20	0.18	0.15	0.10
	30	500	1	0.97	0.94	**0.99**	0.95	0.91
	30	500	0		0.92	**0.99**	0.93	0.81
	50	500	1	0.94	0.93	**0.99**	0.93	0.70
	50	500	0		0.91	0.98	0.90	0.55

Table 2: Simulated power of a 0.05-level test with $p' = p - 1$ predictors and n observations, for four nonlinear functions. The matching either matched for $\widehat{y}$, case 1, or did not, case 0, and it matched for $r = 0$, 3, 5, 10, or p' predictors with the largest absolute t-statistics. Covariates are independent uniform random variables. Each situation was replicated 3000 times. A sampling situation is two consecutive rows, and the highest power in a sampling situation is in **bold**.

Function				Matched for r predictors				
	p'	n	Matched for $\widehat{y}$	0	3	5	10	p'
Exponential	10	100	1	0.60	0.71	0.65	**0.99**	**0.99**
	10	100	0		0.36	0.53	0.69	0.69
10 active	10	500	1	0.90	0.99	0.97	**1.00**	**1.00**
predictors	10	500	0		0.41	0.74	**1.00**	**1.00**
	30	100	1	0.38	0.48	0.40	**0.67**	0.56
	30	100	0		0.18	0.25	0.40	0.16
	30	500	1	0.93	**1.00**	0.99	**1.00**	**1.00**
	30	500	0		0.38	0.74	**1.00**	0.78
	50	500	1	0.93	**1.00**	0.99	**1.00**	**1.00**
	50	500	0		0.36	0.70	**1.00**	0.56
Friedman (1991, 56)	10	100	1	0.05	0.31	0.21	0.09	0.09
	10	100	0		**0.38**	0.22	0.09	0.09
5 active	10	500	1	0.05	**0.99**	0.95	0.56	0.56
predictors	10	500	0		0.98	0.96	0.48	0.48
	30	100	1	0.05	0.13	0.09	0.07	0.05
	30	100	0		**0.15**	0.10	0.07	0.05
	30	500	1	0.05	**0.98**	0.89	0.47	0.11
	30	500	0		0.97	0.94	0.42	0.09
	50	500	1	0.05	**0.97**	0.82	0.38	0.08
	50	500	0		0.94	0.89	0.37	0.07
Friedman (1991, 61)	10	100	1	0.06	0.46	0.47	0.34	0.34
	10	100	0		**0.50**	0.48	0.32	0.32
5 active	10	500	1	0.08	**1.00**	**1.00**	**1.00**	**1.00**
predictors	10	500	0		0.99	**1.00**	**1.00**	**1.00**
	30	100	1	0.06	0.16	0.15	0.10	0.06
	30	100	0		**0.18**	0.15	0.11	0.06
	30	500	1	0.06	**1.00**	**1.00**	0.90	0.51
	30	500	0		0.99	**1.00**	0.88	0.43
	50	500	1	0.07	**1.00**	0.99	0.82	0.25
	50	500	0		0.99	0.99	0.81	0.21
Friedman (1991, 66)	10	100	1	0.41	0.91	**0.93**	**0.93**	**0.93**
	10	100	0		0.80	0.74	0.52	0.52
2 active	10	500	1	0.54	**1.00**	**1.00**	**1.00**	**1.00**
predictors	10	500	0		0.97	0.96	0.85	0.85
	30	100	1	0.41	0.77	0.69	0.75	**0.79**
	30	100	0		0.62	0.59	0.60	0.22
	30	500	1	0.60	**1.00**	**1.00**	**1.00**	**1.00**
	30	500	0		0.97	0.97	0.91	0.61
	50	500	1	0.71	**1.00**	**1.00**	**1.00**	**1.00**
	50	500	0		0.97	0.97	0.93	0.50

might be expected, the highest powers occur with $r = 10$ including $\widehat{y}$ in the match. Even when $p' = 10$, so there are no irrelevant variables, it is still helpful to include $\widehat{y}$, presumably because a very high $\widehat{y}$ means most of $x_1, \ldots, x_{10}$ are high. For many values of r, omitting $\widehat{y}$ from the match for the Exponential surface can ruin the power. Matching for $\widehat{y}$ and $r = 5$ variables has lower power for the Exponential surface than matching for $\widehat{y}$ and $r = 3$ variables for reasons that are not completely clear, but perhaps because $\widehat{y}$ gets more attention in the Mahalanobis distance with $r = 3$ variables than with $r = 5$ variables.

How close are the "near replicates" produced by matching? Before matching, there were $\binom{n}{2}$ distances in the distance matrix, whereas after matching there were $m \doteq n/2 - (n-p)/6$ distances within m pairs. For $n = 500$ observations with $p' = 50$ predictors, $p = p' + 1$, there were initially $\binom{500}{2} = 124,750$ pairwise distances and $m = 175$ within pair distances. How does the average distance within m pairs compare to the average of $\binom{n}{2}$ distances before matching? We computed the two averages, averaging also over 3000 simulations, and took the ratio. If we match for $\widehat{y}$ and r predictors, then the distance is computed among n points in $r + 1$ dimensional space. Not surprisingly, if r is larger, the average distance after matching is a larger fraction of the average distance before matching: it is hard to find similar observations in high dimensions. Consider the case of $n = 500$ observations with $p' = 50$ predictors, matching for $\widehat{y}$ and r predictors in Tables 1 and 2. Among the eight such situations in Tables 1 and 2, the average distance within $m = 175$ pairs was never more than 2% of the average distance with $\binom{500}{2} = 124,750$ pairs if $r = 3$, was never more than 7% if $r = 5$, was never more than 19% if $r = 10$, and ranged from 54% to 57% for $r = p' = 50$. In other words, when trying to match for $r + 1 = 51$ variables, the matched pairs were closer than two observations picked at random, but the distance was reduced by less than half. This may partly explain why the power in Tables 1 and 2 is often higher when matching for $\widehat{y}$ and $r = 5$ predictors than when matching for $\widehat{y}$ and $r = 50$ predictors.

In brief, there is no uniformly best choice among our nine methods. We must choose a test in ignorance of the true response surface. For the admittedly limited situations we have considered, matching for $\widehat{y}$ plus the $r = 5$ variables with the largest t-statistics would have been a tolerable choice in most cases given our ignorance of the true response surface,

but $\widehat{y}$ plus the $r = 10$ variables is competitive, winning in many cases.

4 Example: testing fit without replicates in an experiment

Nelson (1981) discusses an experiment involving degradation of electrical insulation measured as y = dialectic strength in kV. There are two factors, duration of aging x_1 as 1, 2, 4, 8, 16, 32, 48, or 64 weeks, and the temperature x_2 as 180, 225, 250, 275 degrees Celsius. Nelson makes a physical argument for a particular nonlinear relationship, but for the purpose of illustrating our test of fit, we assume the investigator is unaware of this argument and ask whether our test will help the investigator discover this mistake. Each of the 8×4 factor combinations was replicated 4 times, making $8 \times 4 \times 4 = 128$ observations. Although the relationship between y and (x_1, x_2) is highly nonlinear, this is only very slightly apparent in the two bivariate plots of y versus x_1 and y versus x_2, so a careless investigator could fail to notice a serious problem. If one fits a Gaussian linear model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, and uses the four replicates to perform Fisher's test with exact replicates, then the linear model is rejected with a very small P-value. We adapted this example for illustration in two ways.

First, we created a smaller unreplicated design by randomly picking one replicate from each condition. This meant that each of the (x_1, x_2) combinations occurred once in an unreplicated design with $n = 8 \times 4 = 32$ observations. Here $\mathbf{X}$ has $n = 32$ rows and $p = 3$ columns, namely a constant and $p' = 2$ predictors . We then used an optimal nonbipartite matching based on $(\widehat{y}, x_1, x_2)$ to form $m = \lfloor n/2 - (n-p)/6 \rfloor = \lfloor 32/2 - (32-3)/6 \rfloor = 11$ pairs and $n - 2m = 10$ isolated observations, so that $21 = 11 + 10$ predictors in $\mathbf{L}$ were added to the linear model, and $[\mathbf{X}, \mathbf{L}]$ had $24 = 3 + 11 + 10$ columns. We did this 10 times, randomly picking one replicate from the 4 available each time. In 8 of the 10 tests, the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ was rejected at the 0.05 level, despite the reduction in sample size from 128 to 32 and the absence of exactly replicated observations.

Second, we added 10 independent Gaussian noise predictors to the original $n = 128$ observation design so that the revised design was now unreplicated in terms of all 12

predictors, and $\mathbf{X}$ had $n = 128$ rows and $p = 13$ columns. We then used an optimal nonbipartite matching based on $\widehat{y}$ and the five predictors with the largest t-statistics to create 45 pairs and 38 isolated observations, adding $83 = 45 + 38$ predictors in $\mathbf{L}$ to the model, so $[\mathbf{X}, \mathbf{L}]$ had $96 = 13 + 83$ columns. Again, we did this 10 times, creating 10 different sets of noise predictors. All ten tests rejected $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{12} x_{12} + \epsilon$ at the 0.05 level.

5 Discussion: Summary; Alternative methods for selecting variables

We have been testing the fit of the Gaussian linear model (1) with n observations and $p' = p - 1$ predictors by: (i) determining $\widehat{y}$ and the r predictors x_j with the largest $\left|\widehat{\beta}_j / d_j\right|$, (ii) creating a distance matrix using these variables, (iii) using optimal nonbipartite matching to form roughly $n/3$ pairs and $n/3$ isolated observations, and (iv) determining whether these $2n/3$ additional predictors enhance the fit of model (1). Here, the r predictors x_j with the largest $\left|\widehat{\beta}_j / d_j\right|$ are also the r predictors x_j with the largest t-statistics, but $\left|\widehat{\beta}_j / d_j\right|$ does not yield the numerical value of the t-statistic; that is, one cannot select the predictors with absolute t-statistics above 2. This is an exact test: when (1) is true, the probability of false rejection at nominal level α is $\leq \alpha$ (in fact, the test has size α). A key element is the generalization of Tukey's device in which functions of $\widehat{\mathbf{y}}$ and $\mathbf{X}$ may be used in test of fit of (1). We have been using functions of $\widehat{\mathbf{y}}$ and $\mathbf{X}$ to create near-replicate observations, yielding an estimate of error based on paired observations less affected by model misspecification. In simulations, matching for both $\widehat{\mathbf{y}}$ and $r = 5$ or $r = 10$ predictors with large $\left|\widehat{\beta}_j / d_j\right|$ gave good results for several nonlinear response surfaces with either few or many irrelevant predictors.

The proposed test of fit is not a substitute for other diagnostic checks of (1). In particular, one should check for outliers and for non-Gaussian errors. A single outlier, if sufficiently severe, can greatly reduce the power of an F-test, including specifically the test of (1) against (3).

There are many related methods that might be considered. For instance, if the p'

predictors in (1) are highly correlated, it might not be wise to select r predictors x_j for the distance using $\left|\widehat{\beta}_j/d_j\right|$ from the full p' variable model, because an important predictor might have a small value of $\left|\widehat{\beta}_j/d_j\right|$ due to its high correlation with other predictors. Could we, instead, use Mallows' C_P to select r variables x_j for the distance? As with t-statistics, the numerical value of C_P depends on $\widehat{\sigma}^2$, so one cannot use the numerical value of C_P, as one cannot use the numerical value of the t-statistic, if one is going to employ the Tukey-Milliken-Graybill device to obtain an exact test. Consider the $\binom{p'}{r}$ submodels $P \subseteq \{1,\ldots,p'\}$ of (1) that involve exactly r of the p' predictors. Write $\mathbf{X}_P$ for the $n \times (r+1)$ matrix obtained from $\mathbf{X}$ by retaining the constant and the r columns in P, and write $\mathbf{H}_P = \mathbf{X}_P \left(\mathbf{X}_P^T \mathbf{X}_P\right)^{-1} \mathbf{X}_P$ so the predicted values from model P are $\widehat{\mathbf{y}}_P = \mathbf{H}_P \mathbf{y}$. It is readily checked that $\widehat{\mathbf{y}}_P$ is a function of $\mathbf{X}$ and the predicted values $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ from the full model (1); specifically, $\widehat{\mathbf{y}}_P = \mathbf{H}_P \widehat{\mathbf{y}}$. Also, the usual C_P may be rewritten (Mallows 1973, §1) as:

$$C_P = \frac{\left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)^T \left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)}{\widehat{\sigma}^2} - p + 2\,(r+1)\,. \tag{4}$$

Now, (4) depends in a fundamental way upon $\widehat{\sigma}^2$ to compare models with different numbers of variables. However, if one restricts attention to models with exactly r predictors, then the model P with the smallest C_P and r predictors is the model with the smallest $\left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)^T \left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)$ and r predictors, so this model can be determined from $\widehat{\mathbf{y}}$ and $\mathbf{X}$ alone, so the Tukey-Milliken-Graybill device may be used. In brief, instead of selecting for the distance the r variables with the largest $\left|\widehat{\beta}_j/d_j\right|$, we may select the r variables in the r-variable model with the smallest $\left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)^T \left(\widehat{\mathbf{y}}_P - \widehat{\mathbf{y}}\right)$, or equivalently in the r-variable model with the smallest C_P. The identification of this model may be based on the algorithm of Furnival and Wilson (1974), as implemented in the `R` package `leaps`.

Tables 1 and 2 consider matching for $\widehat{\mathbf{y}}$ and/or r predictors. To avoid focusing on r individual predictors, one might match for a few functions of $\widehat{\mathbf{y}}$. For instance, let $P \subseteq \{1,\ldots,p'\}$ be the model in the previous paragraph determined using C_P for some fixed $r < p'$, and let $\overline{P} = \{1,\ldots,p'\} - P$ be the $p' - r$ variables left out of this model. One could match for three variables, namely $\widehat{\mathbf{y}}$, $\widehat{\mathbf{y}}_P = \mathbf{H}_P \widehat{\mathbf{y}}$, and $\widehat{\mathbf{y}}_{\overline{P}} = \mathbf{H}_{\overline{P}} \widehat{\mathbf{y}}$. This would avoid the impossible task of matching in high dimensions while permitting the ostensibly less important variables in $\overline{P}$ to contribute meaningfully to the match distance.

The proposed method forms roughly $n/3$ pairs and $n/3$ isolated observations, yielding roughly $n/3$ degrees of freedom for the within-pair estimate of error. With this structure, the simulation found good power when matching for $r = 5$ or $r = 10$ predictors with $n = 100$ or $n = 500$ observations. If n were much larger than 500, say $n = 30,000$ for data from an administrative database, then one might reconsider these choices. In particular, one might prefer fewer than $n/3$ pairs that match for more than $r = 10$ predictors. There is little value in having $30,000/3 = 10,000$ degrees of freedom for error, rather than a much smaller number of degrees of freedom; that is, we might be happy with fewer than $n/3$ pairs. On the other hand, think about cutting $r = 20$ predictors each at their median to form two categories per variable; then, there would be 2^{20} or about a million very coarse categories, and 30,000 people would be thinly spread among a million categories. In brief, having $n = 30,000$ rather than $n = 500$ has a big impact on degrees of freedom, but only a small impact on our ability to match for $r = 20$ predictors, so we might wish to have far fewer than $n/3$ pairs that are more closely matched for additional predictors. For instance, for large n, one might set a requirement for the distance, letting that requirement determine the number of pairs. If two multivariate observations are drawn independently from the same r-dimensional multivariate Gaussian distribution, then the Mahalanobis distance between them is distributed as two-times a chi-square random variable with r degrees of freedom; hence, the expected distance is $2r$. If κ is the ζ-quantile of the chi-square distribution on r degrees of freedom, then with probability ζ this Mahalanobis distance is less than 2κ. For $r = 20$ predictors and $\zeta = 0.05$, the expected Mahalanobis distance is $2r = 40$ and the 5% point is $2\kappa = 21.7$. One strategy for very large n would be to solve the maximum cardinality matching problem (Korte and Vygen 2008, §10.5): find the maximum number of disjoint pairs such that the Mahalanobis distance within every matched pair is at most 2κ.

It is virtually impossible to find many observations that are very close on many predictors; see Giraud (2015, §1.2). In light of this, the proposed procedure bets that lack of fit will involve predictors that appear to matter when the possibly mistaken model (1) is fitted, and, as the discussion above indicates, there are several if not many options for identifying these predictors. In all cases, the test has its nominal level α, because the level

is computed assuming model (1) is true. However, the power of the test is affected by whether the bet is correct. It is easy to construct examples in which the bet is mistaken, and the power is low, because subtle nonlinearity in some predictor gives the false impression that the predictor is unimportant, hence not included in the matching algorithm. In light of this, the test should be seen as a test of fit of (1) against alternatives in which the ostensibly active predictors enter the model in a misspecified form. This is a practical and interesting class of alternatives to (1), but it is far from exhaustive.

References

Andrews, D. F. (1971), "A note on the selection of data transformations," *Biometrika*, 58, 249-254.

Baiocchi, M., Small, D. S., Lorch, S. and Rosenbaum, P. R. (2010), "Building a stronger instrument in an observational study of perinatal care for premature infants," *Journal of the American Statistical Association*, 105, 1285-1296.

Box, G. E. P. and Draper, N. R. (1982), "Measures of lack of fit for response surface designs and predictor variable transformations," *Technometrics*, 24, 1-8.

Christensen, R. (1989), "Lack-of-fit tests based on near or exact replicates," *Annals of Statistics*, 17, 673-683.

Christensen, R. (1991), "Small-sample characterization of near replicate lack-of-fit tests," *Journal of the American Statistical Association*, 86, 752-756.

Christensen, R. (2011), *Plain Answers to Complex Questions: The Theory of Linear Models*, 4^{th} edition, New York: Springer.

Christensen, R. and Utts, J. (1992), "Testing for nonadditivity in log-linear and logit models," *Journal of Statistical Planning and Inference*, 33, 333-343.

Daniel, C. and Wood, F. S. (1971), *Fitting Equations to Data*, New York: Wiley.

Derigs, U. (1988), "Solving nonbipartite matching problems by shortest path techniques," *Annals of Operations Research*, 13, 225-261.

Draper, N. R. (1982), "Center points in second-order response surface designs," *Technometrics*, 24, 127-133.

Fisher, R. A. (1922), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

Friedman, J. H. (1991), "Multivariate adaptive regression splines," *Annals of Statistics*, 19, 1-67.

Furnival, G. M. and Wilson, R. W. (1974), "Regressions by leaps and bounds," *Technometrics*, 16, 499-511.

Giraud, C. (2015), *Introduction to High-Dimensional Statistics*, Boca Raton, FL: CRC Press.

Green, J. R. (1971), "Testing departure from a regression without using replication," *Technometrics*, 13, 609-615.

Joglekar, G., Schuenemeyer, J. H. and LaRiccia, V. (1989), "Lack of fit testing when replicates are not available," *American Statistician*, 43, 135-143.

Jungnickel, D. (2013), *Graphs, Networks and Algorithms* (4^{th} edition), New York: Springer.

Korte, B.H., and Vygen, J. (2008). *Combinatorial Optimization: Theory and Algorithms* (4^{th} edition), Heidelberg: Springer-Verlag.

Lu, B. (2005), "Propensity score matching with time-dependent covariates," *Biometrics*, 61, 721-728.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011), "Optimal nonbipartite matching and its statistical applications," *American Statistician*, 65, 21-30. (`R` package `nbpmatching`)

Mallows, C. L. (1973), "Some comments on C_P," *Technometrics*, 15, 661-675.

Mandel, J. (1959), "The measuring process," *Technometrics*, 1, 251-267,

Miller, F. R., Neill, J. W. and Sherfey, B. W. (1998), "Maximin clusters for near-replicate regression lack of fit tests," *Annals of Statistics*, 26, 1411-1433.

Miller, F. R., Neill, J. W. and Sherfey, B. W. (1999), "Implementation of a maximin power clustering criterion to select near replicates for regression lack-of-fit tests," *Journal of the American Statistical Association*, 94, 610-620.

Miller, F. R., and Neill, J. W. (2008), "General lack of fit tests based on families of groupings," *Journal of Statistical Planning and Inference*, 138, 2433-2449.

Milliken, G. A. and Graybill, F. A. (1970), "Extensions of the general linear hypothesis model," *Journal of the American Statistical Association*, 65, 797-807.

Neill, J. W. and Johnson, D. E. (1985), "Testing linear regression function adequacy without replication," *Annals of Statistics*, 13, 1482-1489.

Nelson, W. (1981), "Analysis of performance-degradation data," *IEEE Transactions on Reliability*, 2, 149-155. Data available at: http://www.itl.nist.gov/div898/strd/nls/data/nelson.shtml

Rao, C. R. (1973), *Linear Statistical Inference and its Applications*, New York: John Wiley.

Rosenbaum, P. R. (2010), *Design of Observational Studies*, New York: Springer.

Scheffé, H. (1959), *The Analysis of Variance*, New York: John Wiley.

Shillington, E. R. (1979), "Testing lack of fit in regression without replication," *Canadian Journal of Statistics*, 7, 137-146.

St. Laurent, R. T (1990), "The Equivalence of the Milliken—Graybill Procedure and the Score Test," *American Statistician*, 44, 36-37.

Stuart, E. A. (2010), "Matching methods for causal inference," *Statistical Science*, 25: 1-21.

Su, Z. and Yang, S-S. (2006), "A note on lack-of-fit tests for linear models without replication," *Journal of the American Statistical Association*, 101, 205-210.

Tukey, J. W. (1949), "One degree of freedom for non-additivity," *Biometrics*, 5, 232-242.

Utts, J. M. (1982), "The rainbow test for lack of fit in regression," *Communications in Statistics-Theory and Methods*, 11, 2801-2815.

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S., and Rosenbaum, P. R. (2013), "Stronger instruments via integer programming in an observational study of late preterm birth outcomes," *Annals of Applied Statistics*, 7, 25-50.