# Assessing Fit Quality and Testing for Misspecification in Binary Dependent Variable Models*

Justin Esarey† and Andrew Pierce‡

June 27, 2012

**Abstract**

In this paper, we present a technique and critical test statistic for assessing the fit of a binary dependent variable model (e.g., a logit or probit). We examine how closely a model's predicted probabilities match the observed frequency of events in the data set, and whether these deviations are systematic or merely noise. Our technique allows researchers to detect problems with a model's specification that obscure substantive understanding of the underlying data generating process, such as missing interaction terms or unmodeled non-linearities. We also show that these problems go undetected by the fit statistics most commonly used in political science.

## Introduction

All statistical models must make assumptions, simplifying the world they seek to describe in order to make it comprehensible and to focus on a particular relationship of substantive interest. Ideally, those assumptions allow us to glean substantively informative descriptions of the past or accurate predictions of the future. But inappropriate assumptions can also distort a model's ability to translate data into substantive insight. To take just one example, when researchers ignore non-linearities or interaction effects that are present in the DGP,

*We thank Drew Linzer, Mike Ward, Jacqueline H. R. Demeritt, Jeff Staton, John Freeman, Neal Beck, Patrick Brandt, Phil Schrodt, Teppei Yamamoto, Kevin Clarke, and Will H. Moore for their comments, suggestions, and conversations about previous iterations of the paper. Replication materials for all our simulations and data analysis can be found online at the Political Analysis dataverse: http://hdl.handle.net/1902.1/18399.

†Assistant Professor, Department of Political Science, Rice University. Corresponding author. E-mail: justin@justinesarey.com.

‡Department of Political Science, Emory University. E-mail: awpierc@emory.edu

their statistical models can indicate relationships that are too small, too large, or go in the wrong direction (Ai and Norton, 2003; Brambor, Clark and Golder, 2006; Franzese and Kam, 2007). How can researchers detect specification problems in the fit of a binary dependent variable model?

In this paper, we present a method to assess the fit of binary dependent variable models and compare this method to existing approaches. Drawing on the work of Greenhill, Ward and Sacks (2011), Azzalini, Bowman and Hardle (1989), and Hosmer and Lemeshow (1980),[1] we compare a model's predicted probability $\Pr(y = 1)$ to the observed frequency of $y$ in the data set. If the model is a good fit to the data, subsets of the data with $\Pr(y = 1) \approx m$ should have about $m$ proportion of cases for which $y = 1$ (and $1 - m$ proportion for which $y = 0$). The process is analogous to plotting fitted and observed values of the dependent variable against one another, a commonly used fit diagnostic for continuous dependent variable models. We use a local (nonparametric) linear regression to overcome data sparseness (i.e., a paucity of cases for all possible values of $m$) and parametric bootstrapping to determine whether prediction inaccuracies are attributable to sampling variation or misspecification. This information is presented in a visually impactful *heat map plot* that can quickly help a researcher visualize how well a statistical model's predictions match observed frequencies in the sample, whether inaccuracies are symptomatic of misspecification or are simply noise, and whether the model fits the data over the entire range of probabilities (e.g., whether it is equally capable of fitting high- and low-probability events). We also present a *heat map statistic*, a holistic evaluation of the heat map plot that provides a dispositive indicator of model misspecification.

Importantly, we also find that the fit diagnostics most commonly employed by political scientists do not detect specification problems, and indeed are not intended to do so. This includes classification-based approaches, such as Herron's (1999) expected percent correctly predicted (ePCP) and the receiver operating characteristic (ROC) curve. We demonstrate

[1] See also Lemeshow and Hosmer (1982); Copas (1983); Firth, Glosup and Hinkley (1991); Hosmer et al. (1997); Brown and Heathcote (2002); Gelman et al. (2004); and Ward, Greenhill and Bakke (2010).

that a model that near-perfectly predicts event probabilities can appear to be a poor fit using ROC or ePCP, and that a model that poorly predicts probabilities can appear to be a good fit with these approaches. Likelihood-based statistics (like the Akaike Information Criterion[2] (AIC), likelihood-ratio test, and Bayes factors) can only be used for relative comparison: if our statistical model has an $AIC = 879.83$, we know that the model fits the data better than another model with $AIC_2 = 882.3$, but not whether either model's assumptions distort our understanding of the data generating process. Heat mapping supplements these approaches without replacing them: we aim to weed deficient models out of the set of possibilities before comparing their relative quality (e.g., with AIC) or determining their power to classify cases (e.g., with ePCP).

The paper is structured as follows. First, we summarize the advantages of our approach and compare it to fit and model comparison statistics commonly used in political science for binary dependent variable models. Second, we lay out the technical details of the approach and the software package we use to implement it. We use Monte Carlo simulations to show that our approach allows researchers to detect a variety of misspecification problems that other approaches miss. Finally, we demonstrate the applied value of heat mapping by assessing two recent papers: an empirical model of interstate conflict, and a model of the relationship between democratization and foreign aid.

## Assessing the fit of binary DV models

To visually assess the fit quality of an OLS regression model and check for non-linearities or other specification problems, we commonly plot the model's predictions for the dependent variable ($\hat{y}$) against observed $y$ values. This allows a researcher to determine whether the structure of the model is a reasonable approximation of relationships in the underlying DGP; a linear model applied to a quadratic DGP, for example, will result in a characteristic pattern

[2]This statement includes close relations of the AIC, like the Bayesian Information Criterion and Deviance Information Criterion.

of misfit between $y$ and $\hat{y}$. For a binary dependent variable (or BDV) model, the analogous approach is to compare the model's predicted $\Pr(y = 1)$, which we abbreviate as $\hat{p}$, and the *true* probability $\Pr(y = 1)$, which we abbreviate as $p$. We base our statistic on this principle.

The difficulty is that (unlike $y$) the true $p$ is unobserved—and probably not even observable—for a BDV model. But if we collect all the observations with values of $\hat{p} \approx m$, approximately $m$ proportion of those cases should be $y = 1$ when $\hat{p}$ is a good estimate of $p$. That is, $\Pr(y = 1|\hat{p} = m) \approx m$ if our model is a good predictor of probability. Thus, instead of comparing a model's predicted $\hat{p}$ to the true $p$ directly, we compare $\hat{p}$ to an empirical estimate of $\Pr(y = 1|\hat{p})$, which we will call $R(\hat{p})$.

In brief, our approach is to estimate an empirical model, sort observations according to the predicted $\hat{p}$ of the model, and then determine whether this predicted probability is an accurate predictor of $R(\hat{p})$ by plotting them against each other. In a perfectly fitting model, $\hat{p} = R(\hat{p})$ for all values of $\hat{p}$ and the plot is a straight 45 degree line through the origin. If the statistical model tends to over-predicts an outcome, then the predicted probability is greater than the observed frequency: $\hat{p} > R(\hat{p})$. If the model under-predicts an outcome, then $\hat{p} < R(\hat{p})$. Note that a statistical model can do both at the same time: it might, for example, simultaneously over-predict rare events (where $\hat{p}$ is small) and under-predict common events (where $\hat{p}$ is large).

Though our approach builds on a very long tradition in statistics,[3] among political scientsts the dominant[4] mode of BDV model fit assessment is quite different. Specifically, fit assessments tend to focus on a model's ability to classify observations—that is, to predict *outcomes* $y = 1$ and $y = 0$ rather than *probabilities*—or compare the log-likelihood of one model to another. Before we expand on the technical details of our technique, we will briefly

[3]The literature on the nonparametric assessment of the goodness of fit of parametric models is very large; a helpful summary can be found in Hart (1997).

[4]Of the 18 binary dependent variable articles from the 2008-2010 issues of the *American Journal of Political Science* that assess model fit, 3 use the reciever operating characteristic (ROC) curve, 9 examine the percent correctly predicted (PCP), 3 calculate the Akaike Information Criterion (AIC), 4 use the Bayesian Information Criterion (BIC), 1 uses the Vuong test, 4 use the likelihood ratio test, and 6 measure the percent reduction in error.

describe the classification and likelihood approaches and why our approach provides a useful supplement; in brief, our approach is better-suited to the detection of specification problems and an assessment of how substantively relevant those problems are.

## Classification-based approaches

Classification-based approaches to model fitting ask how good a model is at accurately separating cases into two bins, one where the event occurs ($y = 1$) and one where it does not ($y = 0$), on the basis of observable aspects of the case. The resulting statistics are very substantively meaningful: it is easy to intuitively grasp the fit quality of a model whose classifications are 90% accurate compared to the 50% classification power of a coin flip. But, as we describe, this approach can give a misleading picture of a statistical model's fit quality in many political science applications, particularly in the rare events models that are prevalent in the analysis of civil and interstate conflicts.

The percent correctly predicted (PCP) sorts model predictions into two categories, $\hat{y} = 1$ and $\hat{y} = 0$, on the basis of $\hat{p}$. The analyst specifies a threshold $t \in [0, 1]$, that will be used to sort the cases. When $\hat{p} > t$, $\hat{y} = 1$; otherwise, $\hat{y} = 0$. The final PCP statistic, which is just the percentage of cases for which $y = \hat{y}$, gives the analyst a clear sense of whether a model is able to appropriately classify cases.[5]

Of course, the PCP statistic depends on a choice of $t$. Furthermore, when models are compared, different choices of $t$ may be optimal for different models. To overcome these issues, we could calculate the expected PCP (ePCP), which sets the threshold $t$ independently for each observation at the predicted $\hat{p}$ of the fitted empirical model (Herron, 1999). We could also examine the ROC curve, which generalizes the PCP by calculating the rate[6] of true positives ($y = \hat{y} = 1$) and false positives ($y = 0$, $\hat{y} = 1$) for all choices of $t$ and then plotting these pairs for all choices of $t$. The ROC assesses fit by examining the tradeoff

[5]This statistic is sometimes broken out to differentiate correct predictions of $y = 1$ and $y = 0$.

[6]The rate of true positives is the number of true positives divided by the total number of cases for which $y = 1$. The rate of false positives is the number of false positives divided by the total number of cases for which $y = 0$.

between true and false positives. A model that classified cases according to a randomly assigned $\hat{p}$ would tend to produce false and true positives at an equal rate, whereas a model that classified perfectly could achieve a 100% true positive rate for any value of the false positive rate. Thus, a "perfect" model produces an ROC curve with an area of 1, whereas a simple random classification produces an ROC curve with an area of $\frac{1}{2}$.

**Disadvantage: misleading indications of fit**

But a classification approach to model fit can sometimes be misleading. It can indicate that a model is a poor fit to data when the fit is nearly perfect. It can also indicate that a model is a good fit to data when the fit is poor. The reason is that this approach does not examine how well a model predicts probabilities, but how successfully it separates cases where $y = 1$ from those where $y = 0$. Thus, we expect classification statistics to be misleading in two cases. When the DGP does not strongly separate outcomes, classification statistics will indicate a poor fit even when the model is perfect. This can be a problem for researchers studying rare events, such as civil and interstate conflicts. When the DGP *does* strongly separate cases, classification statistics can indicate a good fit even when the model's predicted probabilities do not match the sample. This may present a problem for researchers studying contexts where incentives strongly determine behavior, such as in legislative voting.

Consider, for example, the model and associated ROC curve depicted in Figure 1. The true data generating process[7] (or DGP) in this case is

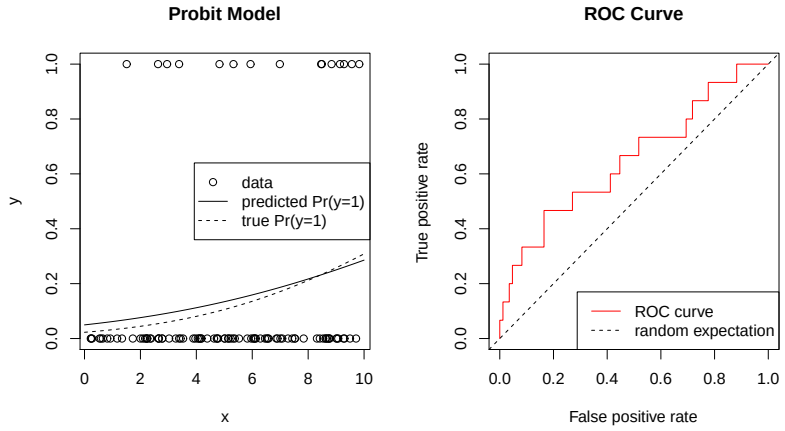
$$\Pr(y = 1|X\beta) = \Phi\left(0.15 * X - 2\right)$$

A simple probit model[8] does a good job of recovering the model in a sample size of $N = 100$: the fitted model is $\Phi\left(0.108 * X - 1.647\right)$. This model is a *nearly perfect* fit to the DGP, as the left panel of Figure 1 indicates: there is very little difference between the predicted and

[7] $X$ is drawn from the uniform distribution between 0 and 10.

[8] For this exposition, I presume a probit model with a normal link function $\Phi$, but the point applies to other binary DV models (such as the logit).

Figure 1: Error-Dominated Outcome Model and ROC Curve




true probabilities for any value of $X$.

However, the ROC curve indicates a poor fit to the data: true positives track false positives rather closely. The area under the curve is 0.652, which is considered less than an acceptable level of discrimination (Hosmer and Lemeshow, 2000, p. 162). As one might expect, the PCP is even worse: the optimal threshold is $t > 0.3$, which corresponds to simply classifying all cases as $\hat{y} = 0$ and accepting a 15% error rate (all false negatives). Any lower choice of threshold raises the aggregate error rate by adding false positives. The ePCP for this model is 0.755, but the ePCP for a constant-only model (i.e., setting $\hat{p} = \overline{y}$) is 0.745—a small difference despite the fact that our model is correctly specified.

The apparent poor fit indicated by the ROC curve (and PCP/ePCP) is misleading. The reason that the ROC curve from Figure 1 is bad is that $y$ is an *error dominated* outcome: causally related to observable factors like $X$, but also strongly determined by randomness. Any attempt to classify observations using $\Phi\left(X\hat{\beta}\right)$ will inevitably involve many false classifications, as demonstrated in Figure 1.[9] Unless $X\hat{\beta}$ can provide a basis for clean separation of $y = 1$ cases from $y = 0$ cases, classification-based fit statistics will indicate a poor fit—even when the model's fit is perfect.

[9]The example in Figure 1 is a low probability event, so most of the false classifications are false positives. But this problem is not peculiar to rare events: it will occur wherever $\Pr(y = 1|X\beta)$ does not change dramatically or quickly over the domain of $X$.

Error dominated outcomes are a special case, but one that often applies to substantive problems in political science. For example, civil and interstate wars are probably not entirely predictable in advance, even with a great deal of information (Gartzke, 1999). When conditions are right for an interstate war, an actual outbreak of hostilities requires a confluence of events that are partially random: the aggressiveness of a military commander, the resolve of negotiators and leaders, and so on. Events like these are *influenced* by causal factors, but are still partially random. Thus, researchers in this field need to be acutely aware of the potential challenges of fit assessment with a classification approach.

The ROC curve can also show an apparently good fit where the model is a poor predictor of probabilities. We created another model where the true DGP is

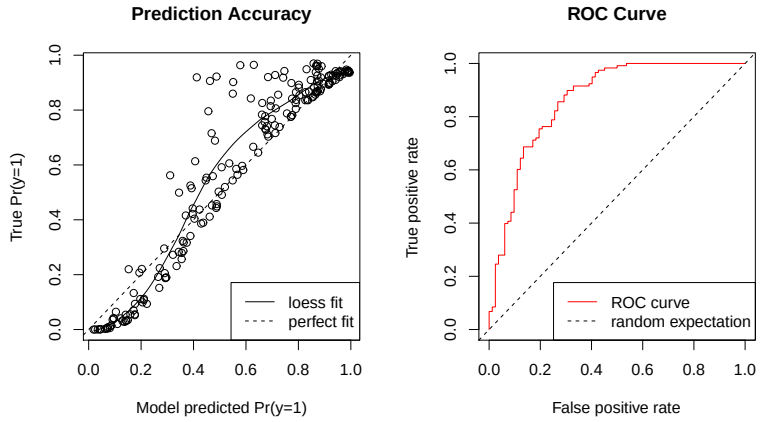
$$\Pr(y = 1|X\beta) = \Phi\left(0.05 * X + 0.05 * Z - 0.07 * X * Z + 1.5\right)$$

and generated a sample of $N = 200$.[10] We then fit a model to this data that omits an interaction term $X * Z$; the model is obviously misspecified. The left panel of Figure 2 plots the predicted probability of our misspecified model against its true probability for each observation in the data set; the model systematically overpredicts low probability events and under-predicts high probability events. But the ROC curve in the right panel indicates a good fit to the data set; the area under the ROC curve is $\approx 0.86$, considered "excellent" discrimination (Hosmer and Lemeshow, 2000, p. 162).

The ROC indicates a good fit for this model, despite its inability to correctly predict probabilities, because even this misspecified model does a reasonable job of classifying cases. The probabilities are not accurately predicted, but the zeroes and ones are mostly separated. We can see this from the PCP statistic: if we classify all observations with $\hat{p} > 0.4$ as $\hat{y} = 1$ and observations with $\hat{p} \leq 0.4$ as $\hat{y} = 0$, 81% of the cases are correctly predicted. But the model yields predicted probabilities that are too low for rare events, and too high for common events, in the sample. This could be problematic in models of legislative voting,

[10] $X$ and $Z$ are drawn from the uniform distribution between 0 and 10.

Figure 2: Misspecified Model and ROC Curve




where party identification and a legislator's own ideology are strong cutpoint determinants of behavior.

## Likelihood-based approaches

A second approach to model fit focuses on the log-likelihood of the sample for the model—that is, the likelihood of this data's appearance assuming that the model is correct. Statistics based on this approach are extremely adept at selecting the optimal specification of a model out of a set of alternatives. Unfortunately, they cannot say whether the best model out of a set of options yields accurate probability estimates, nor whether there are misspecification problems that interfere with the model's ability to accurately characterize the DGP.

Likelihood ratio tests, Vuong tests, the Akaike Information Criterion, Bayes factors, and other test statistics are all defined by transformations of the log-likelihood. For example, the AIC is defined by:

$$AIC = 2k - 2\ln(L)$$

Here, $k$ is the number of independent variables in the model and $\ln(L)$ is the log-likelihood

of the model. For a probit, the log likelihood would be:

$$\ln(L) = \sum_{i=1}^{N} y_i \ln\left[\Phi(X_i\hat{\beta})\right] + (1 - y_i) \ln\left(1 - \Phi(X_i\hat{\beta})\right)$$

Typically, fitting a probit model (or many other models) to data involves choosing $\hat{\beta}$ so as to maximize the log-likelihood of the data. This makes sense because we are trying to pick a model that is most consistent with a data set, i.e., a model under which the sample data were most likely to have occurred.

It is natural to wish to compare models on the basis of (some appropriate transformation of) their log-likelihood: a better model should (among other things) be more likely to have produced the sample data. Penalty factors (like the $2k$ factor imposed by the Akaike Information Criterion) can correct for the danger of model overfitting inherent in this approach.

**Disadvantage: no diagnosis of specification problems**

Selecting the best specification out of a set of possibilities is a very important problem, but not the only one. We must also decide whether a model's assumptions allow it to accurately capture relationships in the DGP. As just one example, there may be cases where the true DGP includes an interaction term or non-linearity that the analyst has no theoretical reason to suspect. It is therefore important for an analyst to have a way to empirically diagnose this sort of misspecification.

This assessment is quite difficult to make using likelihood-based fit statistics like the AIC. Unlike the PCP and ROC, likelihood-based fit statistics (like the AIC) are largely without a substantive interpretation. For example, the AIC of the probit model from Figure 1 is 81.106. We could use this statistic to conclude that the model is better than an alternative model that includes an irrelevant variable $Z$ (AIC = 82.252). But we cannot use this number to determine whether this model gives accurate estimates of $\Pr(y = 1)$. This makes it difficult to use the AIC to diagnose misspecification. It would be helpful to have a tool that could

assess whether a model can accurately recover event probabilities from the sample before comparing it to other alternative specifications on the basis of parsimony and fit. Our tool aims at this supplementary goal.

# Our approach: the heat map plot

As described above, our basic plan is to compare a model's predicted probability, or $\hat{p}$, to an in-sample empirical estimate of $\Pr(y = 1|\hat{p})$, or $R(\hat{p})$. If $\Pr(y = 1|\hat{p} = m) \approx m$ for every $\hat{p}$ predicted by the model, then the model is a good fit. In this section, we lay out the technical details of how we make this comparison. The first task is to empirically estimate $R(\hat{p})$, as the true probabilities $p$ are unobserved. Then, we lay out a visually compelling way of presenting how well the model fits—that is, how well $\hat{p}$ matches $R(\hat{p})$—and whether deviations are attributable to misspecification or are simply noise. We call this a *heat map plot*. The ideas behind the heat map plot inform the creation of a holistic statistical test for misspecification, which we call the *heat map statistic*; simulation evidence confirms that our heat mapping approach performs well when testing for model misspecification.

## Estimating the empirical frequency of $y = 1$, $R(\hat{p})$

How should we estimate $R(\hat{p})$? If we had $k$ many cases with $\hat{p} = m$, we could calculate an average of the dependent variable for these cases:

$$R(\hat{p}) = \frac{1}{k} \sum_{i=1}^{k} y_i \text{ if } \hat{p}_i = m$$

This average could then be compared to $\hat{p}$ to determine the quality of fit. But there are an infinite number of values of $\hat{p} \in [0, 1]$, and data are often sparsely distributed in this space. We must therefore find some way of overcoming the sparseness problem. Gelman et al. (2004, pp. 175-176) suggest ranking the observations according to $\hat{p}$, then dividing them into $J$ bins so that each bin contains an equal proportion of the sample. We could

then compare the average $\overline{y}_j$ in each bin to the average $\overline{\hat{p}_j}$ of the bin; the better the model fits, the closer these will be to one another. This is the basis for the Hosmer-Lemeshow goodness-of-fit statistic (Hosmer and Lemeshow, 1980); we compare our technique to this approach in a later subsection.[11]

Instead of binning, we use nonparametric smoothing (Hart, 1997), particularly first-degree local linear regression as implemented in the `loess` package in R. Instead of grouping observations into a set number of arbitrarily defined bins, each observation is (in essence) compared to its own bin that is nonparametrically constructed from surrounding observations. Nonparametric smoothing allows us to overcome the sparseness problem by including observations whose $\hat{p} \approx m$, but not exactly equal to $m$, in the estimate of $R(\hat{p} = m)$. We adopt nonparametric smoothing over binning for two reasons. First, in data sets where most of the observations are concentrated in one area of the predictor space (e.g., in rare events models where most observations have a near-zero probability of $y = 1$), binning may lump together most or all of the non-zero probability observations despite substantial variation in these observations' characteristics. Nonparametric smoothing allows us to assess how well the model fits these important observations despite their sparseness. Second, the continuous nature of a nonparametric smooth allows for a finer-grained visual assessment of fit compared to discrete binning. This may be relevant in cases where a coarse model is capable of classifying high- and low-probability observations but misses subtle differences inside these categories that indicate a potential specification problem. In exchange for these advantages, non-parametric smoothing is more computationally time consuming and complex than a binning approach, particularly in large data sets.[12] However, this disadvantage will decline with time as computing power becomes cheaper and more available.

[11] To preview, Monte Carlo simulations indicate that the Hosmer-Lemeshow statistic is slightly outperformed by heat mapping in three out of four cases, but that it sometimes correctly detects misspecfication when heat mapping does not. The converse is also true: heat mapping catches problems that the Hosmer-Lemeshow statistic misses.

[12] As an example, the heat map plot in Figure 7 takes about 223 seconds to generate on a Core 2 Q9550 with 4 GB of RAM running R 2.14.1 x64; this data set contains 8,608 observations. The Hosmer-Lemeshow statistic takes 0.03 seconds to compute on this same data set.

Observations $j$ are weighted by a function $K(m-\hat{p}_j)$ according to how close their value of $\hat{p}$ is to $m$; generally, closer observations are weighted more heavily and further observations are weighted less heavily.[13] Only a portion of the data nearest to $m$ is used for the fit; the proportion of data used is called the *bandwidth*. Our software package automatically chooses a bandwidth that minimizes a version of the AIC designed for smoothing parameter selection (Hurvich, Simonoff and Tsai, 1998), a selection procedure recommended by Hardle et al. (2004, pp. 113-118).[14] Note that the smooth avoids the "curse of dimensionality" (Cleveland and Devlin, 1988, p. 608): we smooth $R(\hat{p})$ using only $\hat{p}$, which is always one-dimensional regardless of the rank of $X$.[15]

Final smoothed estimates of $R(\hat{p})$ are created by feeding the weighted values of $y$ into an estimator. For a local estimator of the mean, the so-called Nadaraya-Watson estimator (Hardle et al., 2004, p. 89), one would compute a weighted average using the kernel weights:

$$R(\hat{p}=m) = \frac{\sum_{i=1}^{n} K(m-\hat{p}_i)y_j}{\sum_{i=1}^{n} K(m-\hat{p}_i)}$$

This is the approach of Azzalini, Bowman and Hardle (1989).[16] We opt, instead, to use a local linear regression. At each point $m$, a vector of coefficients $\gamma$ is chosen to minimize the weighted sum of squared errors:

$$\sum_{i=1}^{n} \left( \left[ y_i - (\gamma_0 + \gamma_1 \left[ m - \hat{p}_i \right]) \right]^2 K\left( m - \hat{p}_i \right) \right)$$

[13]The loess package uses the tricubic weight function $K = \left(1 - |u|^3\right)^3$, where the weight centers on the particular value of $m$: $u = (\hat{p}_j - m) / \max_{j \in 1 \ldots n} (\hat{p}_j - m)$, meaning that the weight for a particular observation $j$ is determined by its distance from $m$. Inference is not especially sensitive to the choice of weighting function (Cleveland and Loader, 1996, p. 10-11).

[14]See also Bowman and Azzalini (1997, pp. 77-79). More specifically, the program uses the criterion called $AIC_C$ by Hurvich, Simonoff and Tsai (1998). It can also employ the generalized cross-validation statistic (Craven and Wahba, 1979), if specified. $AIC_C$ and GCV statistics are calculated using a function written by Michael Friendly (available at `http://tolstoy.newcastle.edu.au/R/help/05/11/15899.html`).

[15]The procedure is similar to the way that a generalized additive model smooths using the one-dimensional index $X\hat{\beta}$ (Beck and Jackman, 1998).

[16]Additionally, Azzalini, Bowman and Hardle (1989) use $X\hat{\beta}$ as a predictor, rather than $\hat{p}$ as we do. This introduces additional non-linearities (caused by probability boundaries, and normally handled by the link function) into the estimate. By using $\hat{p}$, we can ensure that the relationship between the predictor and $R(\hat{p})$ is linear when the fit is good, enabling us to leverage the more efficient local linear regression.

and the prediction of $R(\hat{p} = m)$ is $\gamma_0^*$, the prediction for the fitted line at $\hat{p} = m$. Our reason for adopting a local linear regression is twofold. First, first degree approximations have been shown to be superior than zeroth degree (Nadaraya-Watson) approximations in many scenarios (Hardle et al., 2004, pp. 96-98; Hart, 1997, pp. 37-40). Second, as we show in the next subsection, $\hat{p}$ is linearly related to $R(\hat{p})$ when a model is a good fit. This consequently improves performance over a zeroth degree estimator.[17]

## A visual fit diagnostic based on $R(\hat{p})$: the heat map plot

If the model is a good fit to the data, then $\hat{p}$ should be a linear predictor of $R(\hat{p})$: that is, a plot of $\hat{p}$ against $R(\hat{p})$ should yield a straight 45 degree line.[18] Thus, the simplest model fit diagnostic that we propose is to display this plot, which we call the *heat map plot*. The heat map plot allows a researcher to determine whether a model's in-sample predicted probabilities are statistically distinguishable from empirical probabilities in the sample, and whether the differences are substantively meaningful. We implement this procedure, and all of the other statistics and diagnostics that we will describe, in an R package that we make freely available for public use.

An example is shown in Figure 3; in this plot, we re-examine the earlier data set from Figure 2. Recall that we drew a sample of $N = 200$ out of the data generating process

$$\Pr(y = 1) = \Phi\left(0.05 * X + 0.05 * Z - 0.07 * X * Z + 1.5\right)$$

and then estimated a misspecified probit model including only $X$ and $Z$ and omitting the interaction term. The plot of $R(\hat{p})$ against $\hat{p}$ (which we call the *heat map line*) is the colorful

[17] See footnote 30 for details.

[18] The idea is similar to one first advanced by Copas (1983), though he examines plots of $R(\hat{p})$ against individual covariates $X$ instead of $\hat{p}$, uses a Nadaraya-Watson smoother instead of a local linear regression, and does not attempt to assess uncertainty in the fit. Azzalini, Bowman and Hardle (1989) assesses fit uncertainty and suggests plotting $R(\hat{p})$ against $X\hat{\beta}$, but is otherwise similar to Copas. In a related and more recent vein, Greenhill, Ward and Sacks (2011) recommend plotting observed values of $y$ against each observation's fitted $\hat{p}$, but do not assess uncertainty in the fit or propose a formal test for fit quality.

solid line in the center of the plot; the dotted line is a 45 degree line indicating a perfect fit.[19] A rug of points at the bottom of the graph shows the location of observations in the data set.

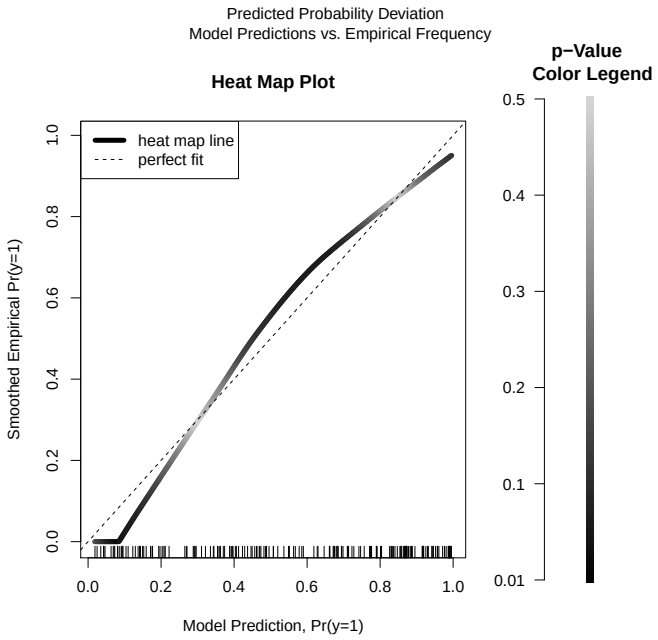
Figure 3 shows that the model tends to underpredict some events; observations with predicted probability $\hat{p} \in [0.4, 0.8]$ occur more frequently in the sample dataset then they are predicted to, as we can see from the fact that the smoothed empirical frequency $R(\hat{p})$ is above the perfect fit line. The model overpredicts other events; observations with $\hat{p} < 0.3$ and $\hat{p} > 0.9$ occur less frequently in the sample than predicted (for this region, $R(\hat{p})$ lies below the 45 degree line). The larger the gap between the 45 degree line and the heat map line, the larger (and more substantively meaningful) the under- or over-prediction. For example, for the portion of the sample with a $\approx 50\%$ predicted probability of $y = 1$ (0.5 on the $x$-axis of the figure), $\approx 60\%$ are observed as $y = 1$ (the corresponding point on the $y$-axis); the 10 percentage point gap is the substantive size of the model's prediction error.

These mispredictions may be evidence for misspecification (and indeed we know, in this case, that the model is misspecified), and at least it enables us to assess the strengths and weaknesses of this model's ability to predict the dependent variable in-sample. But mispredictions—even substantively meaningful ones—can also result from random variation in the data generating process, particularly in small samples. Simply eyeballing a plot makes it hard to tell whether deviations can be attributed to sampling variation or are evidence for misspecification.

How can we tell when the deviation in $R(\hat{p})$ from $\hat{p}$ is likely attributable to misspecification—that is, statistically significant? To determine the distribution of $R(\hat{p})$ under the assumption that the model is correctly specified, we simulate 1000 draws of the dependent variable $y$ from the fitted probit model—in this case, $\Phi\left(\hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Z\right)$–for each observation. We then draw the heat map line for each one of these 1000 data sets; in Figure 4, we plot these 1000 replicated lines in gray along with the original (shaded) heat map line. The resulting,

[19] The AICc-selected bandwidth for the $R(\hat{p})$ estimate in Figure 3 $\approx 0.93$.

Figure 3: $\hat{p}$ Plotted Against $R(\hat{p})$ for a Misspecified Model

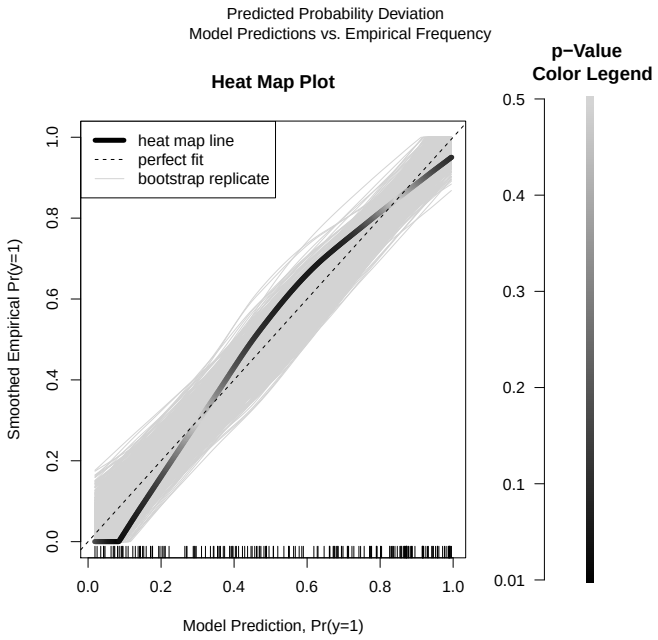



parametrically bootstrapped distribution of curves tells how much deviation from a perfect fit (the dashed line in Figure 4) is attributable to sampling variation rather than misspecification.[20] The purpose of the exercise is to determine whether the observed heat map line could have reasonably been produced by the fitted model if it is correctly specified.

Comparing the original heat map line to its parametrically bootstrapped distribution under a correct specification enables us to calculate a simulated one-tailed $p$-value for each predicted probability $\hat{p}$; the $p$-value is the proportion of bootstrapped lines whose observed empirical frequency $R(\hat{p})$ is further from the median bootstrap estimate than (and in the same direction as) the heat map model fitted on the full sample. Consider Figure 4 as an illustration. When the model predicts a probability of $\hat{p} = 0.10$, very few of the bootstrap replicates have an observed empirical frequency $R(\hat{p})$ at or below the heat map estimate of $\approx 0$, meaning that this deviation from perfect prediction is very unlikely to be attributable to sampling variation. On the other hand, for $\hat{p} = 0.25$, many of the bootstrap replicates have

[20]This bootstrapping approach is often used in the statistics literature; see Azzalini, Bowman and Hardle (1989); Firth, Glosup and Hinkley (1991); Brown and Heathcote (2002).

Figure 4: Plot from Figure 3 with Simulated Sampling Distribution

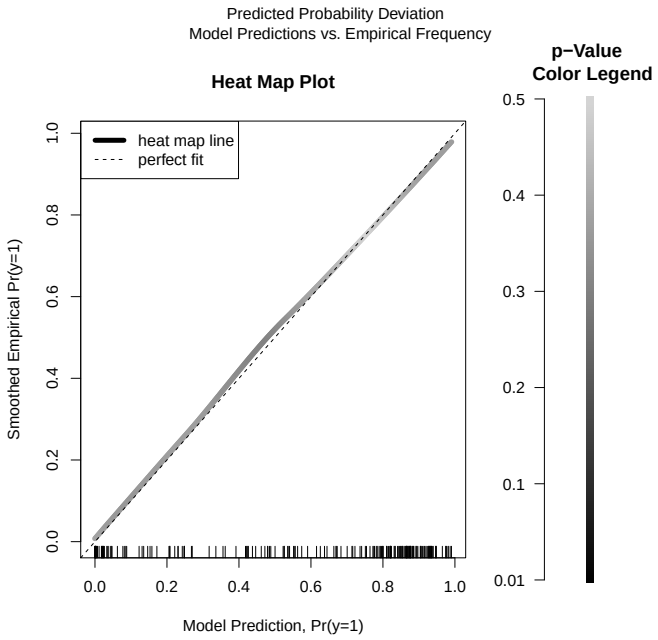



$R(\hat{p})$ further from the median; this deviation is easily attributable to sampling variation. The bootstrapped $p$-value determines the shading of the heat map line: "hot" (or dark black) points indicate deviations that are statistically distinguishable from what we would expect to see due to random variation (near the edges of the bootstrapped distribution plotted in gray), while "cold" (or light gray) points indicate deviations that are attributable to sampling variation (near the center of the bootstrapped distribution).

As we can see in Figures 3 and 4, the fitted model's prediction errors are well beyond what we would expect to see due to sampling variation. We know this because large swaths of the prediction space are "hot" (darkly colored); the dark coloration indicates that the difference between the predicted frequency of $y = 1$ and the observed frequency in the sample is considerably larger than the deviations produced by (bootstrap) sampling variation. For this reason, we suspect that our model might be a poor fit to the underlying data generating process (which, as we know, it is). By contrast, Figure 5 shows the same plot for a correctly specified probit model (one that includes $X$, $Z$, and $XZ$).[21] Though this model's predictions

[21]The AIC-selected bandwidth for the $R(\hat{p})$ estimate in Figure 5 $\approx 0.98$.

Figure 5: $\hat{p}$ Plotted Against $R(\hat{p})$ for a Correctly Specified Model



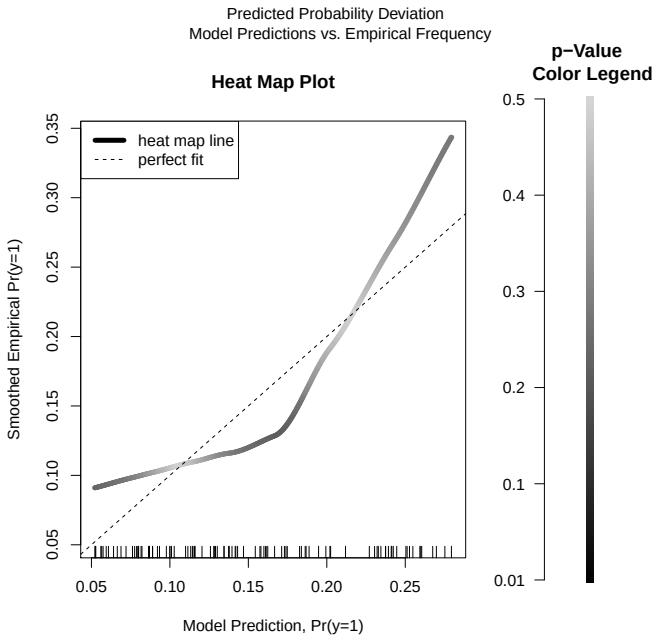

slightly deviate from perfection (there are small gaps between the heat map line and the "perfect fit" 45 degree line between $\hat{p} = 0.4$ and 0.6), the deviations are well within what we would expect from random variation in outcomes: almost all of the points on this line are "cold."

And what about our example from Figure 1? Recall that the ROC and PCP indicated that the correctly specified model from this example had a less than acceptable fit to the dataset. When we construct the heat map plot for this model, shown in Figure 6, we find that the model's fit to the data set cannot be statically distinguished from the data generating process.[22] The fit is substantively quite imperfect—for example, events with $\hat{p} \in [0.15, 0.20]$ are observed about 5-10 percentage points less frequently than we would expect in the sample dataset—but these deviations are within what we would expect from ordinary sampling variation according to the heat map. Thus, our approach (correctly) indicates that this model is a good fit to the data generating process.

Summarily, the heat map plot enables assessments of the substantive and statistical sig-

[22]The AIC-selected bandwidth for the $R(\hat{p})$ estimate in Figure 6 $\approx 0.97$.

Figure 6: $\hat{p}$ Plotted Against $R(\hat{p})$ for the Model from Figure 1




nificance of differences between a model's predicted probabilities and the sample's empirically observed probabilities. The size of the gap between the 45 degree perfect fit line and the observed heat map line shows a researcher to see the size of a model's prediction errors. The coloration of the line allows a researcher to determine whether that gap is statistically distinguishable from noise. As the examples show, deviations from perfect fit can be substantial and yet attributable to sampling variation, in data sets with few observations. Conversely, small deviations from perfect fit can be evidence of misspecification in very large data dets.

# Formalized fit diagnostic: the heat map statistic

The plots of Figures 3, 5, and 6 are designed to enable a researcher to qualitatively assess how well a model fits predicted probabilities wtithin a sample. A more holistic assessment of performance may be made using what we call the *heat map statistic*. We evaluate the ability of the heat map statistic to detect misspecification using simulation evidence. In particular, we will show that ROC curves (representing the classification approach to model fitting) are

not well-suited to detecting specification problems. By comparison, our heat map statistic and its close relative the Hosmer-Lemeshow statistic are good at this task.

## Heat map statistic

Our approach involves calculating the $p$-value on the heat map line for each data point in the sample (that is, at each point indicated in the rug on the figure). Sampling variation and random error would lead us to conclude that events with a $p$-value $\leq k$ will occur at most $2k$ proportion of the time (that is, $k$ proportion in each tail of the distribution). Thus, our diagnostic is to look at how often events with a $p$-value less than or equal to some threshold occur in the plot; if they occur too frequently, we can conclude that the prediction deviations in the plot are statistically distinguishable from random variation. In this way, the heat map statistic mirrors the way a reader visually assesses the statistical significance of a heat map plot: if a plot is too "hot," then the reader rejects the model specification.

Consider, as an example, the plots of Figures 3 and 5. We will look at the proportion of the time that points have a one-tailed $p$-value less than or equal to 0.1. In Figure 3, 48.5% of the observations have a one-tailed $p$-value of less than or equal to 0.1, substantially more than our maximum expectation of 20%. We therefore conclude that this model specification is not a good fit to the data set. In Figure 5, *none* of the observations have a one-tailed $p$-value of less than or equal to 0.1, less than our maximum expectation of 20%. We therefore conclude that this model specification is a good fit to the data set and that its predictions are good enough to be useful for inference.

The resulting fit diagnostic, which we call the *heat map statistic*, is simple: if more than 20% of the one-tailed $p$-values of observations on the heat map line are less than or equal to 0.1, reject the specification. Otherwise, accept the specification.

## Hosmer-Lemeshow statistic

The Hosmer-Lemeshow (or H-L) fit statistic (Hosmer and Lemeshow, 1980) is an earlier, conceptually related approach to which compare the heat map statistic.[23] As noted previously, Hosmer and Lemeshow suggested ranking the observations according to $\hat{p}$, then dividing them into $J$ bins so that each bin contains an equal proportion of the sample (each bin has $k/J = k_j$ observations). A two by $J$ contingency table is formed: the bins are in the rows, while the columns contain the observed frequencies of $y = 1$ and $y = 0$ respectively. The authors then compare the expected and observed frequencies in each bin to determine whether a model's expectations match empirical reality. Conventionally, $J = 10$.[24] The resulting Hosmer-Lemeshow (H-L) goodness-of-fit statistic is the chi-squared statistic for a contingency table:

$$\hat{C} = \sum_{i=0}^{1} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Here, $O_{ij}$ is the number of observed $y = i$ in the bin and $E_{ij}$ is $\overline{p}_j k_j$ (when $i = 1$) or $\left(1 - \overline{p}_j\right) k_j$ (when $i = 0$), where $\overline{p}_j$ is the average $\hat{p}$ in bin $j$. Hosmer and Lemeshow demonstrate (via simulation) that $\hat{C}$ follows the $\chi^2$ distribution with $J - 2$ degrees of freedom under the null hypothesis that the model is an accurate fit to the data. Thus, $\hat{C}$ can be compared to known $\chi^2$ tables to determine whether the deviation between $\hat{p}$ and $R(\hat{p})$ is statistically significant (typically using $\alpha = 0.05$).

## Comparative performance: simulation evidence

To test the performance of the heat map statistic, we perform large scale simulations to determine its size and power relative to the ROC curve and the Hosmer-Lemeshow statistic.[25]

[23] See also Lemeshow and Hosmer (1982) and Hosmer and Lemeshow (2000).

[24] There also exists a variant where residuals are calculated by subtracting each observation $y_i$ from a non-parametrically smoothed estimate of $\Pr(y_i = 1)$, where the sum of the squared residuals are then compared to its asymptotic chi-squared distribution as in the original H-L statistic (le Cessie and van Houwelingen, 1991).

[25] The code for implementing the Hosmer-Lemeshow statistic in R is written by Peter D. M. Macdonald (2011).

We say that the ROC rejects a model if the area under the curve is $< 0.7$, a less than acceptable level of discrimination (Hosmer and Lemeshow, 2000, p. 162).

Our Monte Carlo study examines the performance the heat map statistic in four environments: a DGP with interaction non-linearities,[26] quadratic non-linearities,[27] a structural break in the DGP,[28] and cyclical variation in probability.[29] We fit models with and without the necessary interaction or square term in each of 4000 data sets (with each data set having size $N = 1000$) and assess their fit with the heat map statistic and the H-L statistic (using $\alpha = 0.05$). The results are depicted in Table 1, which shows the percentage of of the time that correctly and incorrectly specified models were rejected in each environment.[30] For the quadratic and interaction simulations, the incorrectly specified model excluded the square/interaction term. For the structural break simulation, the incorrectly specified model assumed a single slope (ignoring the break). For the cyclical variation simulation, the periodic term ($\sin 2X$) was excluded.

This table shows that the heat map statistic performs comparably with the H-L statistic. In all cases, these two statistics have an almost identical false positive rate. The heat map statistic does slightly better at detecting neglected square terms, structural breaks, and cyclical/wave non-linearities in our simulation, while the H-L does slightly better at detecting neglected interaction terms. By comparison, the area under the ROC curve is much less likely

[26]The data generating process is $\Pr(y = 1) = \Phi(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ)$. $\beta_1 \sim U[-0.2, 0.2]$, $\beta_2 \sim U[-0.2, 0.2]$ and $\beta_3 \sim U[-0.055, 0.055]$. $\beta_0$ is set to have opposite sign of the maximum value of $(\beta_1 X + \beta_2 Z + \beta_3 XZ)$; its magnitude varies between $|1.5, 2.5|$. $X$ and $Z$ are drawn from the uniform distribution between 0 and 10.

[27]The data generating process is $\Pr(y = 1) = \Phi\left(\beta_0 + \beta_1 X + \beta_2 X^2\right)$. $\beta_1 \sim U[-0.5, 0.5]$ and $\beta_2 \sim U[-0.08, 0.08]$. $\beta_0$ is set to have opposite sign of the maximum value of $\left(\beta_1 X + \beta_2 X^2\right)$; its magnitude varies between $|1.5, 2.5|$. $X$ is drawn from the uniform distribution between 0 and 10.

[28]The data generating process is $\Pr(y = 1) = \Phi(\beta_0 + \beta_1 XZ + \beta_2 X(1 - Z))$. $\beta_1 \sim U[-0.4, 0.4]$ and $\beta_2 \sim U[-0.4, 0.4]$. $\beta_0 = 0$. $X$ is drawn from the uniform distribution between 0 and 10. $Z = 1$ if $X > 5$ and $= 0$ otherwise.

[29]The data generating process is $\Pr(y = 1) = \Phi(\beta_0 + \beta_1 X + \beta_2 \sin 2X)$. $\beta_1 \sim U[-0.4, 0.4]$ and $\beta_2 \sim U[-0.4, 0.4]$ but excluding values from $[-0.2, 0.2]$ to ensure a substantively notable level of cyclicity. $\beta_0$ is set to have opposite sign of the maximum value of $(\beta_1 X + \beta_2 \sin 2X)$; its magnitude varies between $|1.5, 2.5|$. $X$ is drawn from the uniform distribution between 0 and 10.

[30]We also repeated this simulation using a Nadaraya-Watson zeroth degree smoother instead of the local linear smoother our process uses by default. The results are depicted in an on-line only appendix as Table 3. To summarize, using the zeroth degree estimator results in diminished size for the heat map statistic.

Table 1: Fit Statistic Performance

| | quadratic simulation | | interaction simulation | |
|---|---|---|---|---|
| | % rejections, misspecified model | % rejections, correct model | % rejections, misspecified model | % rejections, correct model |
| heat map statistic | 50.2% | 2.5% | 32.2% | 3.4% |
| H-L, $\alpha = 0.05$ | 48.0% | 2.3% | 36.5% | 3.4% |
| ROC, area$< 0.7$ | 13.4% | 11.8% | 16.1% | 13.5% |
| | structural break simulation | | sine wave simulation | |
| | % rejections, misspecified model | % rejections, correct model | % rejections, misspecified model | % rejections, correct model |
| heat map statistic | 75.1% | 1.2% | 56.4% | 3.0% |
| H-L, $\alpha = 0.05$ | 71.9% | 1.4% | 56.1% | 3.5% |
| ROC, area $< 0.7$ | 29.9% | 23.5% | 34.0% | 22.5% |

to reject misspecified models, and rejects correctly specified models at a roughly equal rate, in all simulations. If a researcher had no reason to suspect misspecification, the ROC curve would not alert him/her to its presence. We take this as evidence that our approach and its conceptual relatives are answering a different question than the one asked by the ROC, one that we believe it is important for political scientists to ask.

Whether the heat map or H-L statistic will perform better seems to be highly specific to the underlying DGP. Consider, for example, the structural break simulations. Of the 3,003 misspecified models rejected by the heat map statistic, 255 (or 8.5%) were *not* rejected by the H-L statistic. Conversely, of the 2,876 misspecified models rejected by the H-L statistic, 128 (or 4.5%) were *not* rejected by the heat map statistic. Similar results come out of the other simulations. Consequently, we cannot unambiguously recommend one approach over another. The primary advantage of heat mapping over the Hosmer-Lemeshow statistic is in the fine-grained and informative plot that it produces, while the primary disadvantage is in the computational time required to generate that plot. What these simulations demonstrate is that both approaches yield similar answers when asked to render an up-or-down, overall evaluation of model fit.

# Application: a model of interstate rivalries

To show our statistic in action, we apply the heat mapping approach to a recent model of international conflict published in the 2011 *American Journal of Political Science*. Our statistic uncovers that this model has some difficulty predicting in-sample probabilities, and that improvements to the model change the inferences we derive from it.

In this article, Daniel Morey (2011) argues that international rivalries are only terminated when they become intense enough to create domestic opposition to further conflict inside both of the rivals. Simulations from his Conflict and Rivalry (CAR) model show "that if a conflict generates a large number of casualties over a short period of time, termed concentration, for

both states, domestic support for further conflict will decrease. This loss of public support restrains states from initiating new conflicts and leads to rivalry termination. This deduction resolves the conflict-rivalry termination puzzle by showing that the higher the concentration from a conflict, the greater the odds a rivalry will terminate" (Morey, 2011, p. 264).

To test the prediction of his model, Morey examines 1,205 rivalries between 1816 and 2001 identified by Klein, Goetz and Diehl (2006). The dependent variable is rivalry *termination*, which "occurs when a dyad surpasses 10–15 years without a dispute" (Morey, 2011, p. 269). Morey models rivalry termination with a logit event history, where the dependent variable (or DV) = 0 until the rivalry terminates, at which point the DV = 1 and the case leaves the data set. The key independent variable is *Concentration*, which is measured by dividing the total number of battle deaths by conflict duration; when the two states in a rivalry have different values, the lowest value for the dyad is used. The tested hypothesis is that "as the lowest level of concentration from conflict increases, the probability of rivalry termination increases" (Morey, 2011, p. 269); that is, *Concentration* should be positively associated with the probability of rivalry *termination*. Numerous control variables (including a baseline hazard function estimated with the Taylor series approximation technique of Carter and Signorino (2010)) are also included in both the original model and our replication; details on these controls are available in the article (Morey, 2011, p. 271). By using data generously provided by Morey, we were able to exactly replicate his key result (shown in column 1 in Table 2). The area under the ROC curve = 0.8172, considered "excellent" discrimination (Hosmer and Lemeshow, 2000, p. 162). The AIC is 4813, though we have no way of using this number to determine whether there are problems with the model's specification.

According to this model, "moving *Concentration* from its minimum to maximum value increases the odds of rivalry termination by 91%, controlling for the influence of the other variables" (Morey, 2011, p. 270). But a heat map plot for this model, shown in Figure 7, illustrates significant fit issues: in particular, events with a low predicted probability happen more often than the model predicts, and events with a high predicted probability

Table 2: Rivalry Model (Table 1, Column 1 from Morey (2011))

| | 1 | | 2 | | 3 | |
|---|---|---|---|---|---|---|
| Variable | $\beta$ | se | $\beta$ | se | $\beta$ | se |
| Concentration | 0.006 | .00182 | .00335 | .000959 | .00310 | .000943 |
| Concentration> 0 | | | 1.574 | .129 | 1.601 | .130 |
| Hostility | -.763 | .124 | -.949 | .129 | -.405 | .148 |
| Major Power Dyad | -.456 | .185 | -.397 | .196 | -.388 | .195 |
| Joint Democracy | .523 | .171 | .524 | .180 | .486 | .173 |
| Territory | -.254 | .111 | -.359 | .117 | -.415 | .116 |
| Negotiated | .882 | .125 | .737 | .130 | .644 | .129 |
| Imposed | 1.434 | .122 | 1.091 | .130 | 1.004 | .130 |
| Stalemate | 1.252 | .0934 | 1.138 | .0968 | .992 | .0983 |
| World War I | 1.301 | .185 | 1.367 | .192 | 1.356 | .180 |
| World War II | 4.038 | .286 | 3.400 | .274 | 3.471 | .260 |
| Power Ratio | -.560 | .150 | -.605 | .159 | -.592 | .156 |
| Distance | .039 | .0160 | .0557 | .0168 | .0579 | .0165 |
| Rivalry Years | -.089 | .00660 | -.0915 | .00679 | -.5580022 | .0752 |
| Rivalry Years$^2$ | .001 | .0000794 | .000679 | .0000792 | .0546 | .0110 |
| Rivalry Years$^3$ | | | | | -.00262 | .000662 |
| Rivalry Years$^4$ | | | | | .0000643 | .0000191 |
| Rivalry Years$^5$ | | | | | -8.25e-07 | 2.80e-07 |
| Rivalry Years$^6$ | | | | | 5.27e-09 | 1.98e-09 |
| Rivalry Years$^7$ | | | | | -1.32e-11 | 5.41e-12 |
| Constant | -1.940 | .106 | -1.962 | .108 | -1.184 | .148 |
| AIC | 4813 | | 4637 | | 4577 | |

$N = 8606$, DV = termination. S.E.s clustered on rivalry.

happen much less often than the model predicts.[31] The deviations from a perfect fit are substantively large, as can be seen in the size of the gap between the perfect fit line and the heat map line; the deviation is as large as 20 percentage points too low for events with a predicted probability $\approx 50\%$. Additionally, many of these deviations are larger than we would expect from sampling deviation in a data set of this size. In particular, the model seems to overestimate the probability of the least likely events (below about 10% predicted probability) and the most likely events (above about 35% probability), while underestimating the probability of mid-likelihood events (between roughly 10% and 35% probability).

Calculating the heat map statistic for this model, we find that about 54% of the observations have one-tailed $p$-values of less than 0.1, substantially greater than our maximum expectation of 20% when the model is a good fit. The Hosmer-Lemeshow statistic agrees with the diagnosis of poor fit: with $J = 10$, the $p$-value associated with the H-L $\chi^2$ statistic is 0.048, allowing us to reject the null hypothesis of acceptable fit using a conventional $\alpha = 0.05$. The disagreement between our fit statistic and the ROC curve makes clear the different fit criteria each uses to assess a model's quality—and, we believe, highlights the need for political scientists to adopt a statistic like ours.

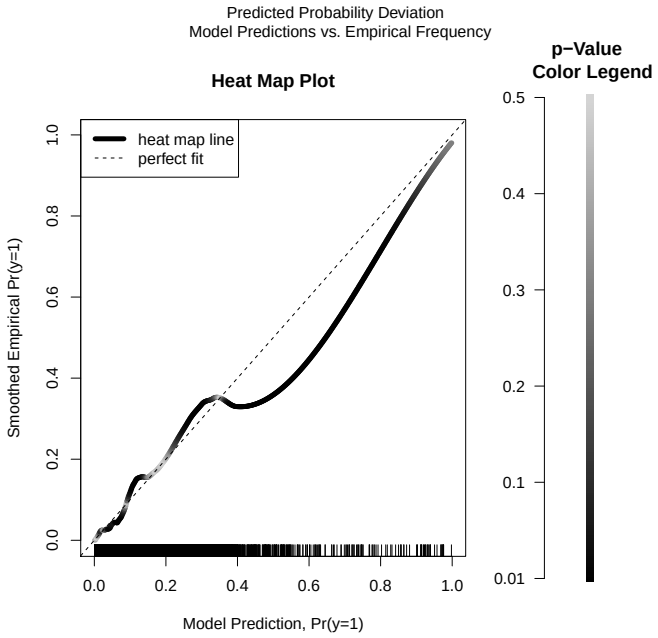
Given the number of variables in the model and the apparent complexity of the data generating process, it is difficult to propose a specification that resolves all of the fit issues in this model.[32] However, we can improve on Morey's specification and demonstrate that these improvements change the substantive understanding of rivalry termination that emerges from the model. In particular, the relationship between the key independent variable *Concentration* variable and rivalry *termination* seems to be more complex than the original model might indicate.

Consider Figure 8a, which compares a bivariate logit fit of the *termination* variable using *Concentration* to a loess fit (span = 0.1); the data points are also plotted as points (with jitter to indicate the density of overlapping distributions). First, note that the vast

[31] The AIC-selected bandwidth for the $R(\hat{p})$ estimate in Figure 7 $\approx 0.20$.

[32] This is one important reason why Achen (2002) argues for the parsimonious specification of models.

Figure 7: Heat map plot for the Morey (2011) model of rivalry termination (Column 1 of Table 2)




majority of the data is concentrated at low values of *Concentration*: 92.5% of the observations have *Concentration* = 0, and 97.0% of the data has *Concentration* < 1. Second, the spike in the lowess plot at *Concentration* = 0 seems to indicate that the greatest increase in rivalry termination probability occurs when any battle deaths occur, a ≈ 40 percentage point increase; a greater concentration of deaths does increase the termination probability, but only by ≈ 20 percentage points. Third, the logit model seems to mismatch the structure of the data in the sample—at least without some kind of adjustments to account for non-linearities in the data-generating process.

Similar issues seem to exist in modeling the relationship between rivalry *termination* and *Rivalry Years*, as indicated in Figure 8b. Morey's original specification used a logit model with Rivalry Years and Rivalry Years$^2$ terms, a specification plotted in the figure along with a loess fit (span = 0.3) . The loess plot indicates a generally declining hazard of rivalry termination with duration, with some heretofore unexplained rises and falls in termination probability with time. The logit model seems to underestimate rivalry termination proba-

bility for very short duration rivalries (2 years or less), overestimate it for medium duration rivalries (between 2 and 20 years), underestimate it for long duration rivalries (between 20 and 110 years), and drastically overestimate it for the very longest rivalries ($> 110$ years).

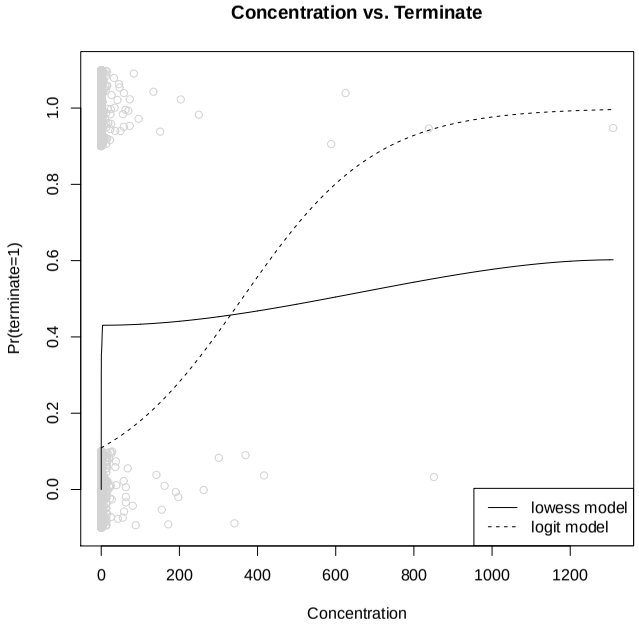
Model 2 in Table 2 adds a binary variable to Morey's specification (*Concentration* $> 0$) that equals 1 if *Concentration* is greater than zero, and equals zero otherwise. As the table shows, adding this variable to the model substantially improves its fit as assessed by the AIC. The substantive interpretation of the model also changes substantially: as *Concentration* goes from 0 to 1, Pr(terminate) rises by .04 percentage points in Model 1 but by 15.8 percentage points in Model 2. Model 3 adds more polynomial powers of Rivalry Years to Model 2, all of which are statistically significant and which result in a much better fit according to the AIC.

The heat map plot of Model 3 (shown in Figure 9) shows that it substantively improves the fit between predicted and observed in-sample probabilities when compared to Model 1: its heat map line is closer to the perfect line over most of the domain. For example, while in the original model events with a predicted probability $\approx 50\%$ occur 20 percentage points too infrequently, in the revised model they prefer 10 percentage points too infrequently. However, our specification tests indicate that the deviations are still statistically distinguishable from those we would expect to occur due to sampling variation. The heat map statistic indicates that about 52% of in-sample predictions have a bootstrapped $p < 0.1$, and the Hosmer-Lemeshow statistic yields a $p \approx 0.018$. Thus, while our model is a substantive improvement over the one from Morey (2011), neither is fully satisfying.

To us, this example demonstrates how assessing a statistical model's ability to predict in-sample probabilities can help to drive improvements to substantive work in political science. The fit discrepancies detected via heat mapping in the model of Morey (2011) suggest an opportunity for future work to improve upon the theory connecting rivalry termination to rivalry intensity (and to the other independent variables in the model). Morey's theory correctly anticipates a positive relationship between the two, but does not anticipate apparent

Figure 8: Possible non-linearities in the rivalry termination DGP

(a) Concentration model




(b) Rivalry years model

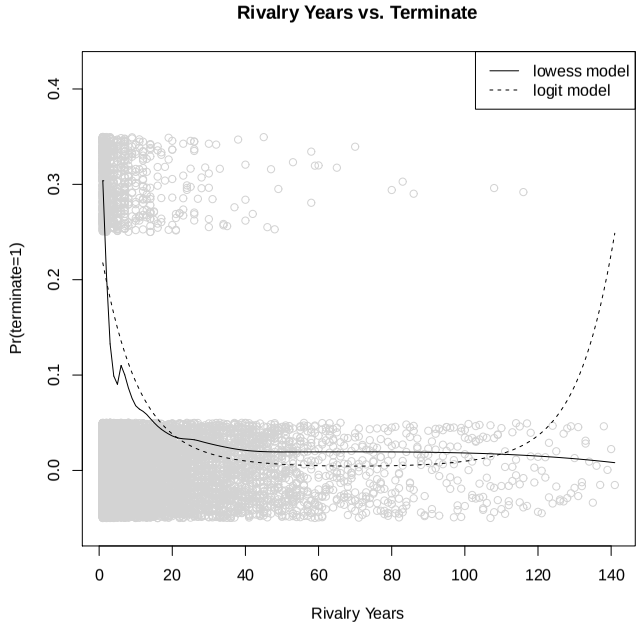
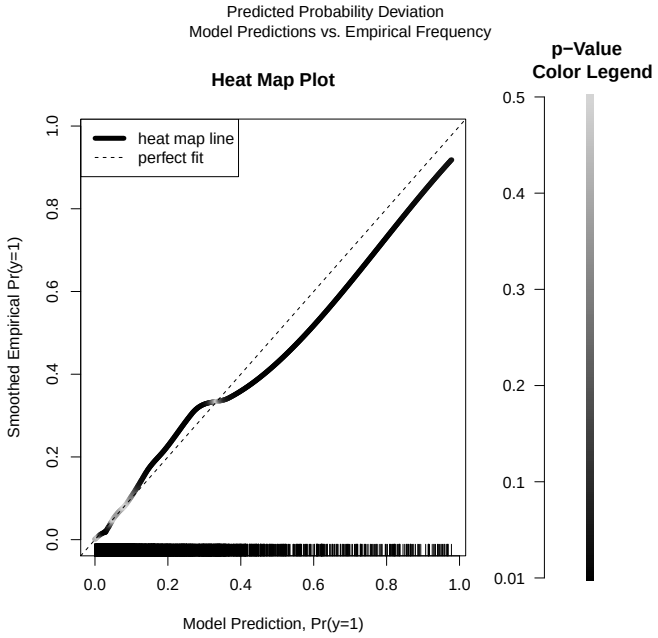


Figure 9: Heat map plot for our revised model of rivalry termination (Column 3 of Table 2)




non-linearities and structural breaks in that relationship. We might be able to improve the fit of the model by adding polynomial terms or creating additional structural breaks for the independent variables in the model, but these changes would be brute force curve-fitting well beyond the theory advanced by (Morey, 2011). A better approach—though one beyond the scope of this paper—would be to rethink the causal process through which rivalry intensity influences public opinion and propose an alternative specification on the basis of that theory. For example, *Concentration* divides total battle deaths by the total duration of rivalry. But if military casualties and the news coverage that they generate have a cumulative effect on public support for a war, then 100 casualties a year for 10 years might be more impactful than 1000 casualties in the first year and none for the next nine—despite the fact that in the tenth year *Concentration* would be equal for both cases.

# Application: democratization, civil conflict, and foreign aid

In the application to rivalry termination, both heat mapping and the Hosmer-Lemeshow statistic find evidence of misspecification and possible room for refinement of the theory behind the model. But heat mapping can also indicate that a model is generally capable of accurately predicting in-sample probabilities while pointing out areas of strength and weakness in fit. This is the case for a recently published study of the relationship between foreign aid and democratization.

Savun and Tirone (2011) argue that foreign aid from democratic countries can aid democratization in transitional regimes while avoiding the "dark side of democratization," which includes an increased propensity for civil and interstate conflict. The authors theorize that democracy assistance programs help fledgling regimes solve commitment problems, which, in turn, lowers the propensity for civil and interstate conflict. Consequently, the authors hypothesize that "democratizing states that receive high levels of external democracy aid are less prone to civil wars than democratizing states that receive no or low levels of democracy aid, holding everything else constant" (Savun and Tirone, 2011, p. 237).

In order to test this theory, the authors analyze a sample composed of Official Development Aid (ODA) eligible countries between 1990 and 2003 (Savun and Tirone, 2011, p. 237). The authors use country-year as the unit of analysis, modeling a binary measure of conflict initiation defined as 1 if "a domestic conflict with at least 25 battle deaths begins after at least two years without an initiation" and 0 otherwise (Savun and Tirone, 2011, p. 237). The authors' key independent variable is the level of democracy aid based on the OECD's categorization of aid as intended for "Government and Civil Society." The authors also include independent variables measuring the current state and rate of change of democracy within a country-year, real GDP and real GDP growth, population, the number of prior peace years, and a measure of conflict in the prior year.

Based on a logit model, the authors find support for their hypothesis that democracy aid tends to alleviate conflict due to democratization. Furthermore, this result appears to robust to alternative specifications accounting for potential endogeneity problems. From these results, the author predict around an 8 percentage point decrease in the probability of conflict initiation when moving from the 1st to the 40th percentile in democracy aid (Savun and Tirone, 2011, Figure 1 p. 242). We assess the fit of the authors' model using an ROC plot and our heat mapping approach, both of which are shown in Figure 10.

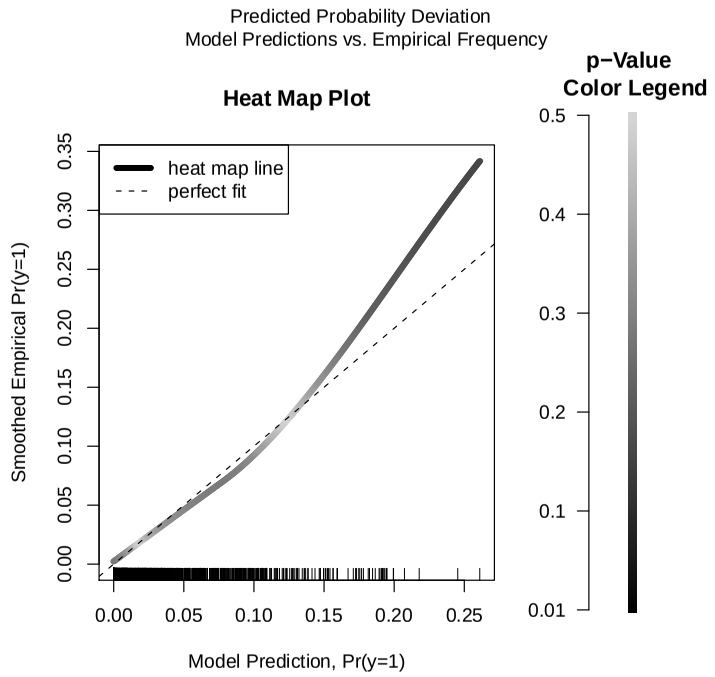
The heat map plot helps us to understand the strengths and weaknesses of the model in a way that the ROC is not designed to do. In particular, we see that the model predicts outcomes with low probabilities well and outcomes with high probabilities less accurately (tending to underestimate them). As denoted in the rug plot on the heat map, most of the data occurs on the low probability end of the scale, which suggests a very good model fit overall. But we might be wary of the model's ability to predict civil conflict outbreaks: it apparently tends to favor the prediction of common (non-conflict) events at the expense of predicting rare (conflict) events, a natural consequence of the fact that non-conflict events are numerous and thus dominate the likelihood function. If we wanted to use such a model for out-of-sample prediction of future civil conflicts, we might consider refitting the model after adjusting the characteristics of the likelihood function to value correct predictions of conflict outbreak more highly than correct predictions of peace, so as to trade off false negative predictions against false positives.

We also note that the ROC curve seems to understate the predictive power of the model. The heat map shows that the model fits well when predicting outcomes that occur less than 15% percent of the time; indeed, no observations have a one-tailed $p < 0.1$. By contrast, the area under the ROC curve is .797,[33] classified as "acceptable discrimination" (and on the borderline of "excellent discrimination") according to Hosmer and Lemeshow (2000). Thus,

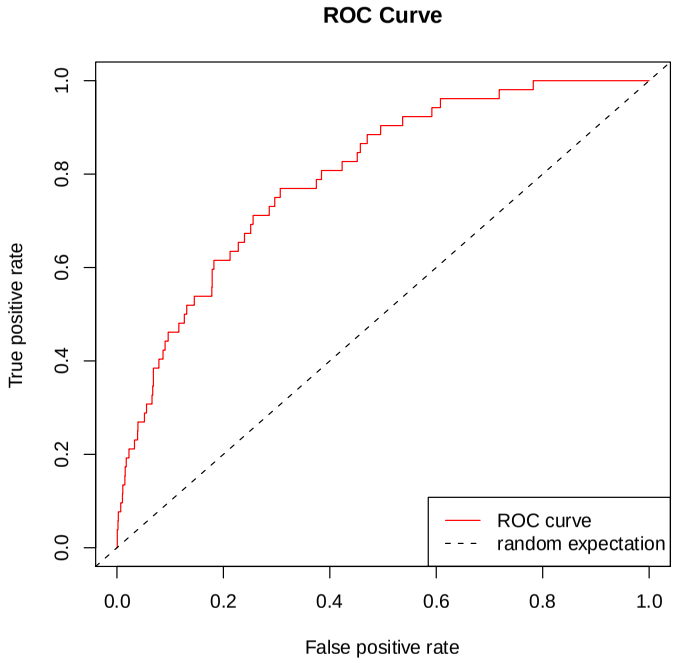
[33]We calculated this number using the ROCR package in R. Calculating the area under the ROC in STATA 11 returns slightly different values depending on the inclusion of different observations when calculating predicted probabilities. These values are statistically indistinguishable from the number presented here.

Figure 10: Fit Assessment Alternatives of Model 1 in Savun and Tirone (2011)

(a) Heat Map Fit Assessment




(b) ROC Curve Fit Assessment



this model represents an applied example of the phenomenon we noted before: classification-based measures of fit can understate the predictive power of rare events models.

# Conclusion

We believe that it is important for empirical researchers to ensure that their statistical models' assumptions—including and especially these models' specifications—do not interfere with our understanding of the underlying data generating process. The ePCP and ROC can help researchers understand how well their models will be able to make definitive predictions about binary outcomes. The AIC, Bayes factors, and similar statistics allow us to compare the relative fit of models while avoiding overfitting. The heat mapping approach adds another capability to a researcher's arsenal. With it, s/he can determine whether the model's predicted probability $\hat{p}$ is an accurate predictor of empirically observed probabilities $p$ in the sample. This capability allows a researcher to determine whether the model's assumptions are mild enough to allow the model to approximate the structure of the data. As political scientists have become more sensitive to the presence of interaction and non-linearity in their models (Franzese and Kam, 2007) and more concerned with assessing the predictive power of their models (Ward, Greenhill and Bakke, 2010) and their ability to fit the sample (Greenhill, Ward and Sacks, 2011), we believe this tool will be increasingly important to substantive researchers in international conflict, civil war, voting behavior, and anywhere else binary dependent variables are important explananda.

We recommend that empirical researchers with a binary dependent variable use goodness-of-fit statistics, like our heat map statistic and the Hosmer-Lesmeshow statistic, in concert with classification and likelihood-based fit statistics to assess the quality of their statistical models. As we have shown, the heat map statistic can be very useful in helping a researcher determine whether there are unmodeled non-linearities or other problems in the structure or specification of an empirical model. Likelihood and classification-based fit statistics, while

valuable on their own terms, are not as helpful for this purpose. We believe this assessment will be particularly important and valuable for scholars examining rare events data, such as those working in international conflict, where classification fit statstics can be misleading. The R software that we provide makes it easy for an applied researcher to begin using the heat map statistic immediately with minimal start-up cost.

# References


Achen, Christopher H. 2002. "Toward a New Methodology: Microfoundations and ART." *Annual Reviews of Political Science* 5:423–450.

Ai, Chunrong and Edward C. Norton. 2003. "Interaction Terms in Logit and Probit Models." *Economics Letters* 80:123–129.

Azzalini, A., A. W. Bowman and W. Hardle. 1989. "On the use of nonparametric regression for model checking." *Biometrika* 76:1–11.

Beck, Nathaniel and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42:596–627.

Bowman, Adrian W. and Adelchi Azzalini. 1997. *Applied Smoothing Techniques for Data Analysis*. Oxford University Press.

Brambor, Thomas, William Clark and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14:63–82.

Brown, Scott and Andrew Heathcote. 2002. "On the Use of Nonparametric Regression in Assessing Parametric Regression Models." *Journal of Mathematical Psychology* 46:716–730.

Carter, David B. and Curtis S. Signorino. 2010. "Back to the Future: Modeling Time Dependence in Binary Data." *Political Analysis* 18:271–292.

Cleveland, William S. and Clive Loader. 1996. Smoothing by Local Regression: Principles and Methods. In *Statistical Theory and Computational Aspects of Smoothing*, ed. W. Hardle and M. G. Schimek. Springer pp. 10–49. `http://www.stat.ucla.edu/~cocteau/stat204/readings/cleveland.pdf`.

Cleveland, William S. and Susan J. Devlin. 1988. "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." *Journal of the American Statistical Association* 83:596–610.

Copas, J. B. 1983. "Plotting p against x." *Journal of the Royal Statistical Society, Series C* 32:25–31.

Craven, Peter and Grace Wahba. 1979. "Smoothing noisy data with spline functions." *Numerische Mathematik* 31:377–403.

Firth, D., J. Glosup and D. V. Hinkley. 1991. "Model Checking with Nonparametric Curves." *Biometrika* 78:245–52.

Franzese, Robert J. and Cindy D. Kam. 2007. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. University of Michigan Press.

Gartzke, Eric. 1999. "War Is In The Error Term." *International Organization* 53:567–587.

Gelman, Andrew, John B. Carlin, Hal Stern and Donald B. Rubin. 2004. *Bayesian Data Analysis*. 2nd ed. ed. Chapman and Hill/CRC.

Greenhill, Brian, Michael D. Ward and Audrey Sacks. 2011. "The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models." *American Journal of Political Science* Forthcoming:1–13.

Hardle, Wolfgang, Marlene Muller, Stefan Sperlich and Alex Werwatz. 2004. *Nonparametric and Semiparametric Models*. Springer.

Hart, Jeffrey D. 1997. *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer.

Herron, Michael. 1999. "Postestimation Uncertainty in Limited Dependent Variable Models." *Political Analysis* 8:83–98.

Hosmer, D. W., T. Hosmer, S. Le Cessie and S. Lemeshow. 1997. "A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model." *Statistics in Medicine* 16:965–980.

Hosmer, David W. and Stanley Lemeshow. 1980. "A goodness-of-fit test for the multiple logistic regression model." *Communications in Statistics* A10:1043–1069.

Hosmer, David W. and Stanley Lemeshow. 2000. *Applied Logistic Regression*. Wiley Interscience.

Hurvich, Clifford M., Jeffrey S. Simonoff and Chih-Ling Tsai. 1998. "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion." *Journal of the Royal Statistical Society, Series B* 60:271–293.

Klein, James P., Gary Goetz and Paul F. Diehl. 2006. "The New Rivalry Dataset: Procedures and Patterns." *Journal of Peace Research* 43:331–348.

le Cessie, S. and J. C. van Houwelingen. 1991. "A Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods." *Biometrics* 47:1267–1282.

Lemeshow, Stanley and David W. Hosmer. 1982. "The use of goodness-of-fit statistics in the development of logistic regression models." *American Journal of Epidemiology* 115:92–106.

Macdonald, Peter D. M. 2011. "R Functions for ROC Curves and the Hosmer-Lemeshow Test." Online.
**URL:** *http://www.math.mcmaster.ca/peter/s4f03/s4f03_0607/rochl.html*

Morey, Daniel. 2011. "WhenWar Brings Peace: A Dynamic Model of the Rivalry Process." *American Journal of Political Science* 55:263–275.

Savun, Burcu and Daniel C. Tirone. 2011. "Foreign Aid, Democratization, and Civil Conflict: How Does Democracy Aid Civil Conflict?" *American Journal of Political Science* 55:233–246.

Ward, Michael D., Brian Greenhill and Kristin Bakke. 2010. "The Perils of Policy by p-value: Predicting Civil Conflicts." *Journal of Peace Research* 46:363–375.

# Online-only Appendix: Performance with zeroth degree smoothers

To obtain a non-parametric estimate of $R(\hat{p})$, we leverage the fact that $\hat{p} = R(\hat{p})$ when the model fit is good. This means that the plot of $R(\hat{p})$ against $\hat{p}$ should yield a straight line, and hence a local linear estimator of $R(\hat{p})$—one that assumes a linear relationship with $\hat{p}$—should outperform a Nadaraya-Watson (N-W) smoother that does not make that assumption. In particular, the local linear smoother should reject fewer correct models (as the assumption of linearity is correct when the model is appropriately specified). To test this idea, we replicated the analysis of Table 1 with a N-W smoother for our heat map curve. The results are depicted in Table 3. The table indicates a substantial decrease in size for the N-W based heat map statistic: the rate of rejection of correct models substantially increases in all environments.

Table 3: Fit Statistic Performance with Different Smoother Settings

| | quadratic simulation | | interaction simulation | |
|---|---|---|---|---|
| | % rejections, misspecified model | % rejections, correct model | % rejections, misspecified model | % rejections, correct model |
| heat map, local linear smooth | 50.2% | 2.5% | 32.2% | 3.4% |
| heat map, N-W smooth | 57.9% | 5.8% | 46.7% | 10.8% |

| | structural break simulation | | sine wave simulation | |
|---|---|---|---|---|
| | % rejections, misspecified model | % rejections, correct model | % rejections, misspecified model | % rejections, correct model |
| heat map, local linear smooth | 75.1% | 1.2% | 56.4% | 3.0% |
| heat map, N-W smooth | 82.5% | 3.0% | 57.6% | 8.1% |