

A NEW VISUAL METHOD FOR EVALUATING THE PREDICTIVE POWER OF BINARY MODELS

BRIAN D. GREENHILL, MICHAEL D. WARD, AND AUDREY SACKS

ABSTRACT. We present a new visual method for assessing the predictive power of models with binary outcomes. This technique allows the analyst to quickly and easily choose among alternative model specifications based upon the models' ability to consistently match high-probability predictions to actual occurrences of the event of interest, and low-probability predictions to non-occurrences of the event of interest. Unlike existing methods for assessing predictive power for logit and probit models such as the use of "percent correctly predicted" statistics, Brier scores and the ROC plot, our "separation plot" has the advantage of producing a visual display that is more informative and easier to explain to a general audience than a ROC plot, while also remaining insensitive to the user's often arbitrary choice of threshold for distinguishing between events and non-events. We show how to implement this technique in R and demonstrate its effectiveness in building predictive models in four different areas of political research.

1. STATE OF THE ART

Binary data are widespread in political science, and political scientists have contributed tremendously to methods of the systematic study of dichotomous variables. Aldrich and Nelson (1984) provided a didactic introduction to logit and probit regression. Political Science settled on logit, as opposed to probit, largely because of the availability of computer software for the former, and that has become the de facto norm. However, until recently regression results from these binary, discrete regression models were mainly presented as tables of coefficients and associated measures of precision. Because these numerical results are so difficult to interpret, and so prone to misinterpretation, King, Wittenberg, and Tomz (2000) introduced a more graphical way of presenting the results in terms of calculating expected values conditional on the estimated model. This paper was aptly titled "Enough with the Logit Coefficients, Already!", but that title sadly did not survive the scrutiny of the editorial process. The basic ideas presented therein serve as the basis for the general approach to presenting results found in the Omnibus software package known as *Zelig*.

Despite these improvements, very scant attention has been paid to the notion of the fit of the model from a more modern and visual perspective. Until recently, most inferential problems in political science were entirely theoretical and retrospective: looking at statistical

Prepared for presentation at the 2009 Annual Meeting of the American Political Science Association, held September 3-6, Toronto, Canada. We appreciate the feedback from colleagues associated with the ICEWS project, especially, Sean O'Brien, Philippe Loustanau, & Laura Stuart. Colleagues at the 10th Anniversary Conference on Statistics and the Social Sciences, held 4-5 June, 2009 at the University of Washington, Seattle, WA, also provided useful input, especially Andrew Gelman and Steve Fienberg. We also presented a version of these ideas at the 2009 Summer Political Methodology Conference, held 23-25 July at Yale University and received valuable reactions from a variety of colleagues. Despite all this good advice, we remain responsible for results presented herein.

significance and looking at observed data. However, today there is considerably greater interest in cross-validation and out of sample prediction than ever before. As a result, the die-hards from the classical perspective who wanted “some measure of fit” have been joined by a wide variety of other interests who require greater information about the quality of empirical models than is provided by tables of numbers representing the means and variances of the estimated parameters. There is now greater interest in ROC plots and tables of specificity and sensitivity as a way of gauging the validity of the estimated model, in empirical terms.

When it comes to assessing the predictive power of logit or probit models, ultimately one wants to be able to distinguish good models from bad models on the basis of their ability to generate “correct” predictions. The difficulty, however, lies in distinguishing “correct” from “incorrect” predictions: logit and probit models generate fitted values that lie somewhere along a continuous 0 – 1 scale (e.g., $\hat{y} = 0.68$) whereas the actual values of the dependent variable are dichotomous (for all observations, $y = 1$ or $y = 0$). To put it in more concrete terms, if one has a logit model of US presidential election outcomes that generated predicted probabilities of 0.68 and 0.32 for Obama and McCain victories in the 2008, how “correct” was the model? In this paper we present a new visual method – what we call the “separation plot” – for dealing with this problem. First, however, we shall briefly review the solutions that are most frequently used to assess the predictive power of logit or probit models.

1.1. Option 1: Dichotomize Everything. Consider the hypothetical data on war and peace for a sample of six countries shown in Table 1. Our dependent variable, y , is coded such that each instance of war is assigned a value of 1 and each instance of peace a value of 0. The fitted values, \hat{y} , obtained from our logit or probit model are shown in the third column.

TABLE 1. Sample Data

| Country | Actual Outcome (y) | Fitted Value (\hat{y}) |
|---------|------------------------|----------------------------|
| A | 0 | 0.774 |
| B | 0 | 0.364 |
| C | 1 | 0.997 |
| D | 0 | 0.728 |
| E | 1 | 0.961 |
| F | 1 | 0.422 |

To determine how frequently the model makes the “correct” predictions, we can simply dichotomize the fitted values. For example, we could establish a rule whereby every fitted value less than, say, 0.5 is considered to be an instance of peace, whereas every value greater than or equal to 0.5 is considered an instance of war. In this case we can generate the following contingency table:

| | Predicted War | Predicted Peace |
|--------------|---------------|-----------------|
| Actual War | {C, E} | {F} |
| Actual Peace | {A, D} | {B} |

We find that three countries were correctly predicted – C, E and B – whereas three other were incorrectly predicted – A, D and F. But these results, of course, depend upon the chosen

threshold; 0.5 is an entirely arbitrary cut-point and were one to choose a cut-point of, say, 0.3, the results would look rather different:

| | Predicted War | Predicted Peace |
|--------------|---------------|-----------------|
| Actual War | {C, E, F} | {} |
| Actual Peace | {A, B, D} | {} |

In this case we are able predict war in three of the cases where there actually was a war – C, E and F – but now we don’t predict any instances of peace. Instead, we incorrectly predict war in the three actual cases of peace – A, B and D. By lowering the threshold from 0.5 to 0.3, we have increased our number of true positives at the cost of a higher number of false positives.

While these 2x2 tables are easy for readers to understand, the problem of course is that the results are entirely dependent on the author’s (often arbitrary) choice of threshold.¹

1.2. Option 2: The ROC Curve. Receiver Operating Characteristic (ROC) curves provide a visual summary of the tradeoff between false positives and false negatives as the threshold is varied. Usually these are presented in the form of a plot of the false positive rate against the true positive rate obtained for each possible threshold τ . The false positive rate (FPR) is defined as the number of false positives divided by the sum of the false positives and true negatives (in other words, all the incorrectly identified negatives divided by all the actual negatives), whereas the true positive rate (TPR) is defined as the number of true positives divided by the sum of the true positives and false negatives (again, all the correctly identified positives divided by all the actual positives).²

The ROC curve for our six-row dataset on war and peace is shown in Figure 1 below. In this case, the six discrete \hat{y} values give rise to seven possible combinations of FPR/TPR values, and therefore seven points on the curve. The calculation of all seven points on the curve are shown in Table 2 below.

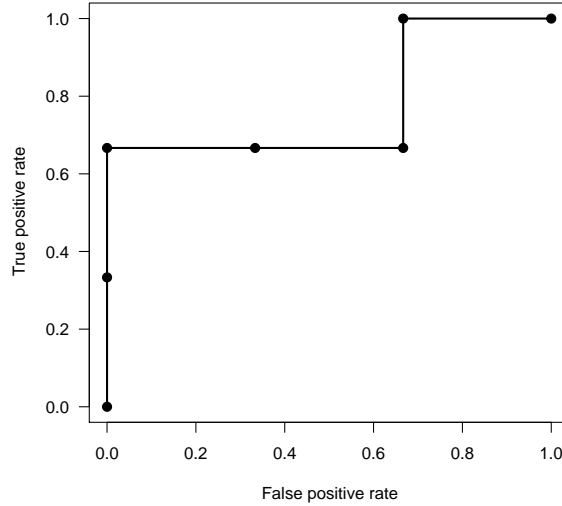
TABLE 2. Calculation of the points on the ROC curve.

| Threshold | TP | FP | FN | TN | FPR | TPR |
|------------------------|-----------|-----------|-----------|-----------|------------------------|------------------------|
| $0 < \tau < 0.364$ | {C, E, F} | {A, B, D} | {} | {} | $\frac{3}{3+0} = 1$ | $\frac{3}{3+0} = 1$ |
| $0.364 < \tau < 0.422$ | {C, E, F} | {A, D} | {} | {B} | $\frac{2}{2+1} = 0.67$ | $\frac{3}{3+0} = 1$ |
| $0.422 < \tau < 0.728$ | {C, E} | {A, D} | {F} | {B} | $\frac{2}{2+1} = 0.67$ | $\frac{2}{2+1} = 0.67$ |
| $0.728 < \tau < 0.774$ | {C, E} | {A} | {F} | {B, D} | $\frac{1}{1+2} = 0.33$ | $\frac{2}{2+1} = 0.67$ |
| $0.774 < \tau < 0.961$ | {C, E} | {} | {F} | {A, B, D} | $\frac{0}{0+3} = 0$ | $\frac{2}{2+1} = 0.67$ |
| $0.761 < \tau < 0.997$ | {C} | {} | {E, F} | {A, B, D} | $\frac{0}{0+3} = 0$ | $\frac{1}{1+2} = 0.33$ |
| $0.997 < \tau < 1$ | {} | {} | {C, E, F} | {A, B, D} | $\frac{0}{0+3} = 0$ | $\frac{0}{0+3} = 0$ |

¹There may be some applications where the threshold is not arbitrary. For example, if we were to decide that when it comes to predicting wars, a false negative is ten times more costly than a false positive, then a threshold of 0.1 might be deemed appropriate.

²The True Positive Rate is also referred to as the *sensitivity* of a classifier. The False Positive Rate is equivalent to $1 - \text{specificity}$, where specificity is defined as $\frac{TN}{FP+TN}$.

FIGURE 1. ROC plot for the data shown in Table 1.



ROC curves have the advantage of providing a visual description of the predictive power of the model over all possible thresholds. Models with high levels of predictive power will tend to have true positive rates that are consistently higher than the corresponding false positive rates, giving rise to curves that have the appearance of being pulled towards the upper-left corner of the plot. As a result, the overall predictive power of the model (across all possible thresholds) can be summarized in terms of the area under the ROC curve, since the ROC is defined on the unit square (the so-called “AUC score”).³

1.3. Option 3: Other Summary Statistics.

1.3.1. *Brier Scores.* A common choice is the Brier Score, which is the mean value of the squared difference between the fitted and actual values of the dependent variable. The Brier score was developed in the early 1950s to provide a way to grade probabilistic weather forecasts. The formula for this metric is quite simple:

$$B = (p - X)^2$$

where p is the probabilistic forecast and X is the dichotomous, binary variable of whether or not the forecast event occurred (1=yes; 0=no). The closer to zero the Brier score, the better the forecast. When many forecasts are made, as for example in a single model applied to many cases, the mean Brier score is often reported.

In this case the mean value of the Brier score across all six observations is 0.266.

³AUC scores are bounded between 0 and 1. A model that is no better than a simple coin-toss should have an AUC of 0.5; lower scores are occasionally observed when, on average, the false positive rate *exceeds* the true positive rate.

TABLE 3. Calculation of Brier Scores.

| Country | Actual Outcome (y) | Fitted Value (\hat{p}) | Brier Score $(\hat{p} - p)^2$ |
|---------|------------------------|----------------------------|-------------------------------|
| A | 0 | 0.774 | 0.599 |
| B | 0 | 0.364 | 0.132 |
| C | 1 | 0.997 | 0.000 |
| D | 0 | 0.728 | 0.530 |
| E | 1 | 0.961 | 0.002 |
| F | 1 | 0.422 | 0.334 |

1.3.2. *Pseudo R^2* . Sometimes scholars (and software) will report a *Pseudo R^2* which is typically one minus the ratio of the likelihood for a null model to the likelihood for the estimated model, so that if the null model and the estimated model have about the same likelihood, then the pseudo score is close to zero. Like the R^2 this measure has many flaws as a measure of fit.

To help address the issue of providing a nuanced, visual yardstick for the performance of such models, we develop in the following section a new approach, called the separation plot.

2. THE SEPARATION PLOT

2.1. **The Concept.** The alternative method of evaluating predictive power that we present here involves a simple re-arrangement of the data presented in Table 1 such that the fitted values are presented in ascending order. We then note whether each of these corresponds to an actual instance of the event (war) or a non-event (peace).

TABLE 4. Rearrangement of the data presented in Table 1 for use in the separation plot.

| Country | Fitted Value (\hat{y}) | Actual Outcome (y) |
|---------|----------------------------|------------------------|
| B | 0.364 | 0 |
| F | 0.422 | 1 |
| D | 0.728 | 0 |
| A | 0.774 | 0 |
| E | 0.961 | 1 |
| C | 0.997 | 1 |

The predictive power can now be evaluated by simply gauging the extent to which the actual instances of the event are concentrated at the bottom end of the table, and the non-events at the top end of the table. A model with no predictive power – i.e., one whose outcomes can be approximated by a random coin-toss – would generate an even distribution of 0s and 1s along the column on the right-hand side. On the other hand, a model with perfect predictive power will produce a complete separation of the 0s and 1s in the right-hand column: low fitted values would always turn out to be associated with actual instances of

FIGURE 2. Separation plot representing the data presented in Table 1.



peace (0s), whereas high fitted values would always be associated with actual instances of war (1s).

It turns out to be very easy to discern these differences using a simple graphical representation that we call the “separation plot” (see Figure 2 below). In this graph, the dark and light panels correspond to the actual instances of the events and non-events respectively, ordered such that the corresponding \hat{y} values increase from left to right.

As the graph shows, our “model” does a reasonably good job of describing the data. Clearly, a perfect model would produce a plot where all the events are clustered at the right-hand pole and all the non-events at the left-hand pole, i.e., . A completely ineffective model that shows no such separation taking place might look more like .⁴

2.2. Model Fitting. The main advantage of the separation plot lies in its ability to provide a clear visual description of the fit of a model (whether in-sample or out-of-sample). This can be especially helpful at the model selection stage, where single-number summaries of predictive power may lack nuance and ROC curves are more difficult to compare.

Consider the following example of an attempt to build a model that describes the incidence of political insurgencies among a group of 29 countries in the Asia-Pacific region over the 1998-2004 period ($n=812$). The project and data are described in more detail in O’Brien (2010 forthcoming). Figure 3 shows the results of using separation plots at successive stages of the model selection process. In the first plot, we show the results of fitting a model that has an intercept and no covariates. In this case, the fitted values are identical for every observation in the dataset and are simply the mean of the insurgency variable. The model clearly does a poor job of separating events from non-events, and the black line representing the corresponding values of \hat{y} remains at a constant low level across all observations.

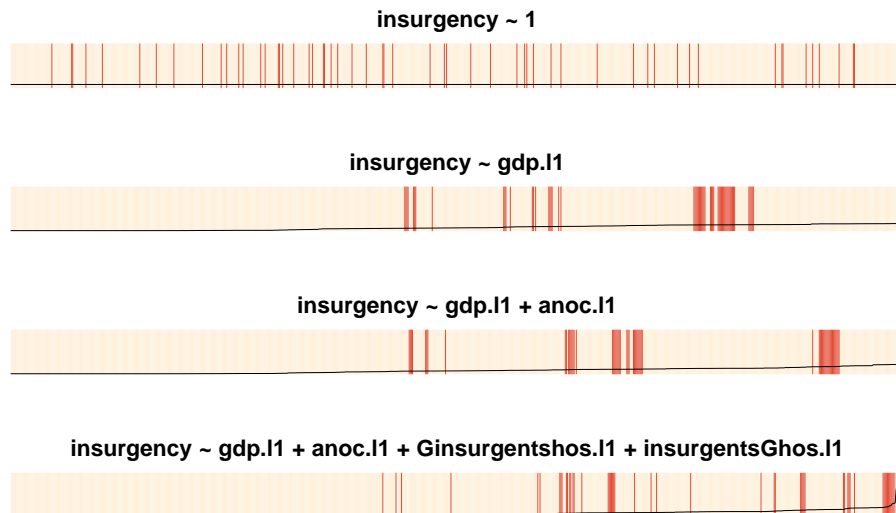
In the second plot we add a single covariate, GDP per capita (lagged by one year). This clearly improves the model’s ability to distinguish events from non-events. We now see three distinct clusters of events arranged towards the right-hand side of the plot space (corresponding to the higher values of \hat{y}). Meanwhile, no events are detected in the regions corresponding to the lower $\sim 40\%$ of observations.

The third graph shows the results of adding a second covariate, anocracy, which can be thought of as an indicator of the extent to which a state has unstable political institutions (Marshall and Jaggers, 2003). Adding this variable clearly improves the predictive power of the model: the clusters of red lines indicating actual instances of the event are shifted further to the right, and no events appear to the left of the leftmost insurgency event in the graph above.

Finally, the graph at the bottom of Figure 3 shows the effect of including two further variables representing detailed event stream data on the number of instances of hostility between the government and the insurgent groups in the preceding period. In this case we

⁴Worse still, a model that consistently makes the *wrong* predictions would produce a plot looking like – i.e., one in which the events and non-events have been sorted in the wrong direction.

FIGURE 3. Separation plots used in the development of a model of insurgency in the Asia-Pacific region, 1998-2004. The text above each separation plot shows the variables included in the model.



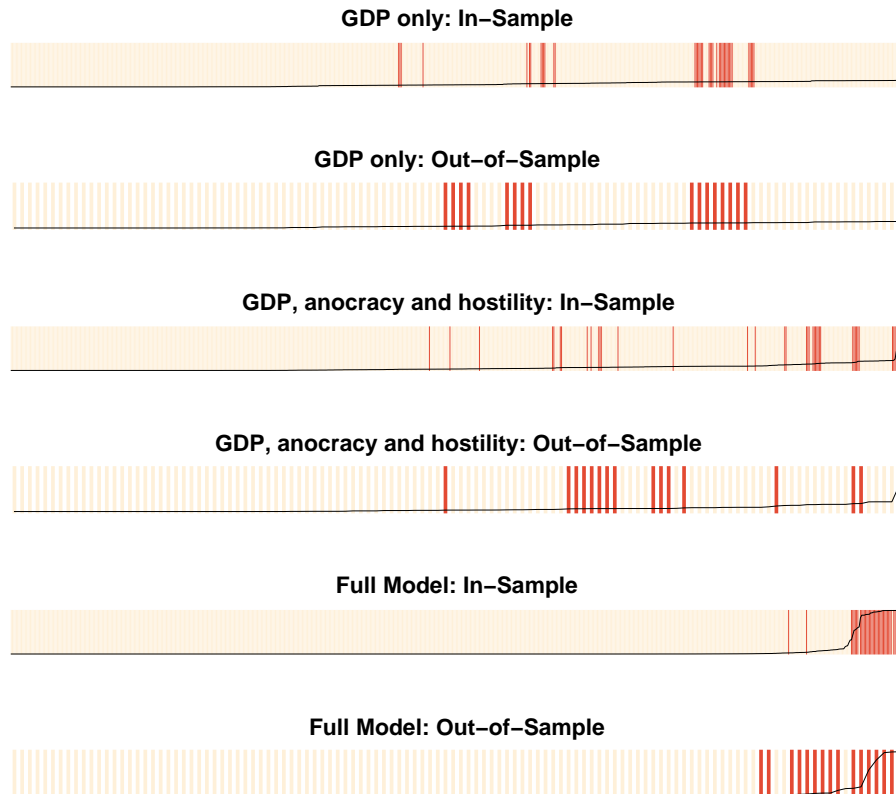
see a distinctive clustering of observations corresponding to insurgency events on the far right-hand side of the graph, and we can also see that these are associated with significantly higher values of \hat{y} (as shown by the vertical position of the black line).

These results show how the separation plots can be used to gain a more fine-grained understanding of the way in which the predictive power of the model improves as one changes the model specification. Its key advantage over a single statistic such as area under the ROC curve is that it allows the user to consider gains and losses in predictive power within different regions of the data. For example, it may be the case that the user is especially interested in developing a model that maximizes the number of correctly-predicted events among the highest values of \hat{y} , in which he or she could focus only on the region on the far right-hand side while ignoring other parts of the graph. On the other hand, certain applications may call for the user to focus on minimizing the number of events among lower values of \hat{y} , in which case the third model in Figure 3 would actually be considered superior to the fourth.⁵

Of course, the separation plot need not only be used to measure in-sample fit. It can just as easily be used to evaluate out-of-sample predictive power in cross-validation exercises. This is especially useful when the user is interested in maximizing predictive success within a particular “test” phase. Figure 4 shows three pairs of separation plots obtained for both the in-sample and out-of-sample evaluation of models of insurgency in the Asia-Pacific region. The upper plot in each pair represents the in-sample fit of a model fitted using the 1998-2003 period as a training set, while the lower plot represents the out-of-sample predictive power in the 2004 period.

⁵In the R package for the separation plot, we also provide a facility to highlight individual observations of interest. This allows the user to see how the predicted probability assigned to one or more critical cases changes under different specifications of the model.

FIGURE 4. Separation plots used in the development of a model of insurgency in the Asia-Pacific region, 1998-2004. The text above each separation plot shows the variables included in the model.



The first pair of plots in Figure 4 shows the corresponding in and out-of-sample results for a regression of insurgency on levels of GDP per capita in the prior period. (Note that the in-sample graph shown here is slightly different from the corresponding graph in Figure 3 because this one shows the performance for a model fitted using only the 1998-2003 training set, as opposed to the full 1998-2004 period.)

The second pair of plots shows the results for a slightly more complex model that includes GDP per capita, anocracy (a measure of the absence of any effective government) and the government to insurgents and insurgents to government hostility scores. While this results in a clear improvement in in-sample fit, it is not obvious that this more complex model performs better in the 2004 test period than the simple GDP-only model.

Finally, the third pair of plots in Figure 4 shows the results of a more complex model that includes the following covariates: GDP per capita, anocracy, competitiveness of political participation, number of minority groups, government-insurgent and insurgent-government hostility scores, a year counter variable and a spatial lag that reflects the levels of insurgency

among each country’s “similar” peers in the preceding period.⁶ As is clear from the separation plots, this model provides an excellent in-sample fit to the 1998-2003 data and almost perfectly predicts the instances of insurgency in the 2004 test period.

3. EXAMPLES

We turn now to four examples from wide ranging parts of the discipline to illustrate this approach.



3.1. Political Campaigns: Hillygus and Jackman (2003).

Hillygus and Jackman (2003) develop a model of voting intentions that shows not only that campaign events such as the party conventions and presidential debates lead to changes in voter preferences, but that the effect that these events have on voter preferences varies according to the preferences that the voters expressed in an earlier survey. They interpret this finding as evidence that voters assimilate new information about the candidates in a way that is conditional upon their earlier preferences (Hillygus and Jackman, 2003: 590).

The authors employ a logit model of voter preference. In addition to discussing the estimates of the coefficients in the model, they also comment on the fit of the model in the following way:

“In gross terms, the estimated transition models fit the data very well, as does any model with a lagged dependent variable in the equation. The convention model correctly predicts 91% of vote preferences, and the debate model correctly predicts 94% of vote preferences, using $p = .5$ as the classification threshold. Moreover, each of the models has an area under the ROC curves anywhere from .79 to .89. These numbers indicate that each of our models does an acceptable to excellent job of discriminating those who transitioned from those who were stable.” (Hillygus and Jackman, 2003: 592-593)

We believe that the separation plot provides a more informative yet compact tool for conveying the same point. As is shown in Figure 5 below, the models, as illustrated in the separation plot, appear to make an excellent fit to the data.

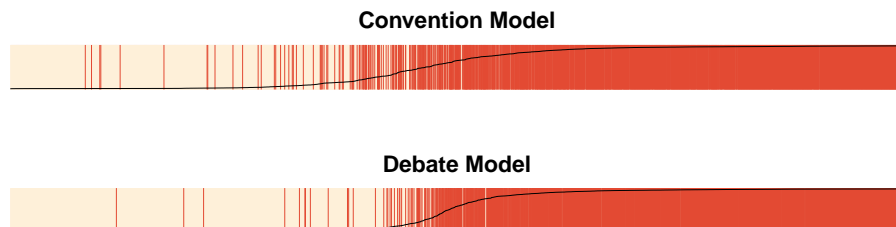
Moreover, one might choose to simplify the presentation further by displaying the separation plots as an in-line graphic akin to the concept of the “sparkline” developed by Tufte (2006). In this way, we might say that Hillygus and Jackman model of voter preference following the party conventions fits the data like this , whereas the corresponding model for the post-debate period fits the data like this .

What is abundantly clear from this pair of graphics is that there is a fairly large number of events in these data, i.e., statements of intentions to vote, and further that most of the highest predicted probabilities are those individuals who stated an intention to vote.

3.2. Civil War: Fearon and Laitin (2003).

⁶The connectivity matrix is based on Gower’s distances computed using a number of different variables designed to capture the extent to which pairs of states experience similar political events.

FIGURE 5. Separation plots for the Hillygus and Jackman models of voting intentions in the 2000 presidential election. The upper plot shows the results of the survey conducted in the period following the party conventions, while the lower plot shows the results of the survey conducted after the presidential debates. Both models make an excellent fit to the data.



Fearon and Laitin (2003) developed a highly influential model of civil war onset in the post-WWII period. Based upon the statistical significance of the variables included in the logistic regression model, the authors claim that various economic and geographical factors that favor insurgencies (e.g., poverty, large populations and mountainous terrain) tend to be associated with the onset of civil war. Most importantly, they find that cultural factors such as the extent of ethnic heterogeneity are not associated with higher probabilities of civil war onset.

However, when the in-sample predictive power is visualized using a separation plot, the model appears to make a relatively poor fit to the data. Indeed, as we show in Figure 6 below, this model fits the data only marginally better than a far more parsimonious model that includes GDP per capita (logged) as its only covariate, despite the presence of a large number of statistically-significant variables. This is a point that we develop in greater detail in Ward, Greenhill and Bakke (forthcoming).

FIGURE 6. Comparison of the separation plots produced by replicating Model 1 of Fearon and Laitin (2003), and by re-estimating the model with logged GDP per capita as the only covariate.

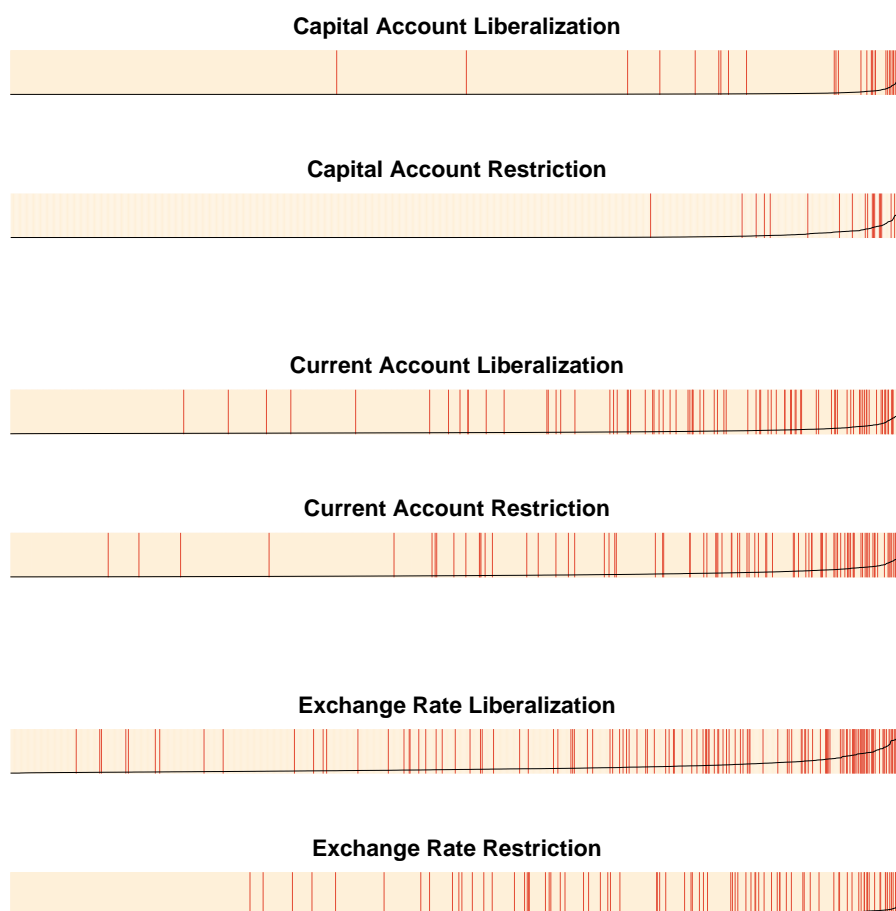


3.3. Policy Diffusion: Simmons and Elkins (2004).

Simmons and Elkins (2004) argue that a country's decision to adopt liberal economic policies cannot be explained by domestic conditions alone, but instead depends to a significant

extent on the decisions made by other influential countries. One of their more notable findings is that a country's decision to adopt liberal policies with respect to its capital account, current account and/or control of its exchange rate is closely correlated with the decisions made by other countries that share a common religion with the country of interest. They find that a spatially-lagged variable that indicates the liberal/illiberal nature of the economic policies found among a country's co-religionists has an estimated effect that is positive and statistically significant ($p < 0.05$) in all models. This holds in all three of the dimensions of economic policy covered in their study, as well as for transitions in both directions – i.e., transitions to a more liberal policy and transitions to a more restrictive policy. The authors suggest that this finding reflects the fact that, when faced with incomplete information about the costs and benefits of adopting a particular policy, states tend to mimic the policies of culturally-similar others (Simmons and Elkins, 2004:187).

FIGURE 7. Separation plots for the Simmons and Elkins models.



Simmons and Elkins use a Weibull survival model to estimate the time taken for each country to undergo a transition to (or from) a liberal economic policy. In Figure 7 we show the separation plots obtained by re-estimating the model using a simpler logit specification. This specification is a common choice for handling a duration variable that occurs within

discrete units of time (e.g., when the data only contains information on the year of adoption of policy) and when the levels of the covariates for each country are not constant over time (Box-Steffensmeier and Jones, 2004:108). Using a logit model therefore allows us to generate a predicted probability of transition for each country-year up to and including the year in which the transition occurred. These values can then be compared with the values of the dichotomous variable indicating whether or not the transition occurred in each country-year case. In this sense the separation plots can be interpreted in the same way as before: to the extent that the model fits the data well, higher probabilities of transition should correspond to the actual occurrences of transition for each country in the sample.

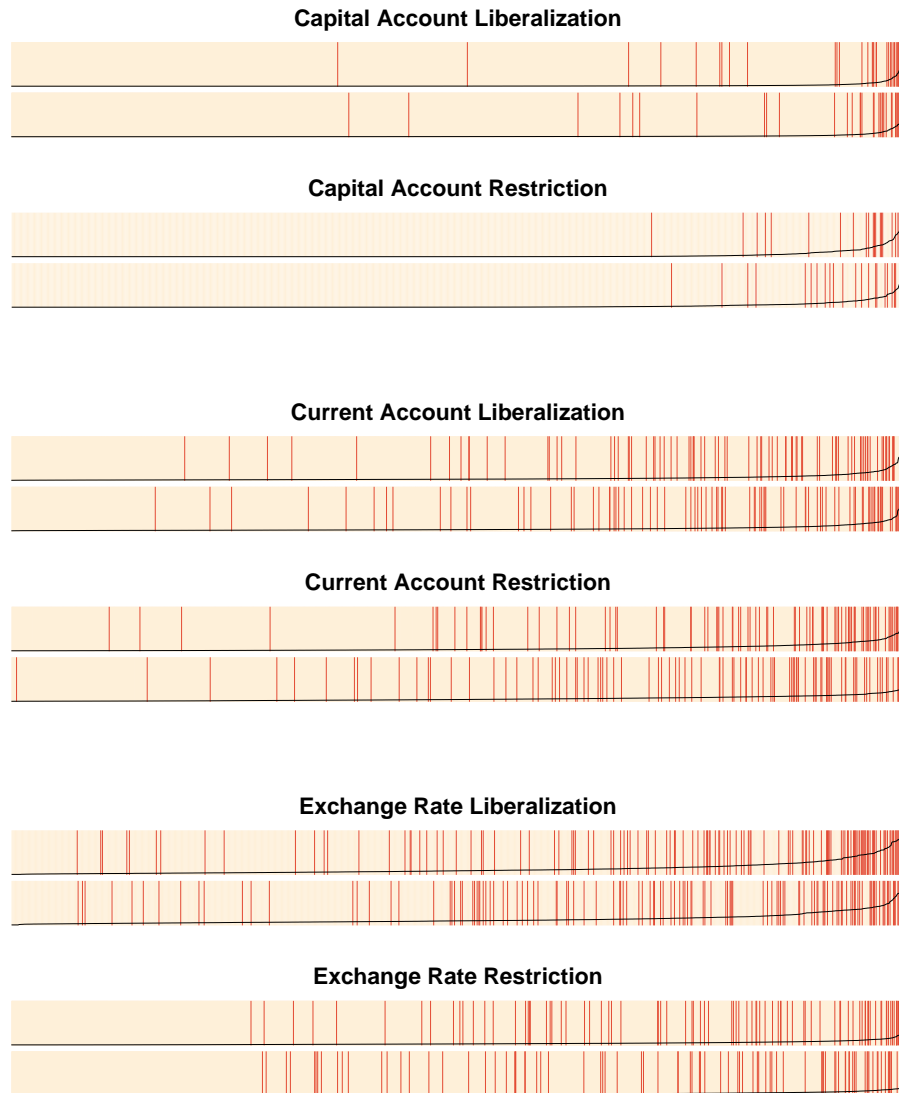
The separation plots show that the model of capital account openness does a reasonably good job of predicting the country-year cases at which transitions occur. This is true for the model of adoption of liberal policies (the uppermost plot in Figure 7) and the corresponding model for transition to restrictive policies (the second plot). The results of the models of current account and exchange rate liberalization are, however, less impressive; these models assign relatively low probabilities to many of the actual cases of transition.

In their article, Simmons and Elkins present the results of a series of likelihood ratio tests to show that the joint inclusion of the three measures of cultural diffusion used in their study (specifically, the spatial lags calculated using data on common religions, common languages and common colonial histories) significantly improve the fit of the models. They report that the inclusion of these three variables leads to an improvement in fit that is statistically significant at the 0.05 level for five out of the six models. The only one that is not is the model of transitions to more restrictive capital account policies (Simmons and Elkins, 2004: Table 3).

In Figure 8 below we demonstrate how this analysis can be re-performed using a series of separation plots. Figure 8 is essentially the same as Figure 7 except that we have now added an additional separation plot that allows us to compare the results of the models in Figure 7 with equivalent models estimated without the three cultural diffusion variables. In each pair of plots the upper plot shows the results of the fully-specified model, whereas the lower plot represents the same model without the cultural diffusion variables. If the inclusion of these variables do in fact lead to a substantial improvement in overall model fit, we should expect to see a greater degree of separation in the upper plots when compared to the lower ones.

However, the results for the pairs of separation plots shown in Figure 8 do not suggest that the inclusion of the cultural diffusion variables leads to a significant improvement in model fit. Instead it appears that the in-sample predictive power is only improved in two of the six models – specifically, the models of current account restriction and exchange rate liberalization. In both of these cases we see a tighter clustering of red lines (representing country-years where the policy is adopted) on the right-hand side of the plot. In the other four models there is little if any evidence of improved predictive power, and in the case of the capital account models one might argue that the inclusion of these variables actually *reduces* the predictive power of the model.

FIGURE 8. Separation plots for the Simmons and Elkins models, estimated with and without the cultural diffusion variables. The upper decks of the pairs of plots represent the in-sample fit of the fully-specified models (and are therefore identical to the plots in Figure 7 above), whereas the lower decks represent the in-sample fit of the corresponding model re-estimated without the cultural diffusion variables.



4. WHAT CAN GO WRONG?

In cases where the sample size is very large ($N \approx > 10,000$), it becomes difficult to see the individual lines that corresponds to individual cases in the separation plot. This makes interpretation difficult when the amount of separation in the plot is low and the predicted values for events and non-events fall across the 0.0 to 1.0 range. In some scenarios, the white lines become virtually impossible to see and the plot looks like a solid red rectangle. We address this problem by creating an alternative version of the plot that separates the events and non-events into two separate plots and uses a color spectrum with each color corresponding to a range of fitted values. Lighter shades of red denote smaller probabilities and darker shades denote larger probabilities. In addition to making the visualization of the amount of separation in the plot easier, it allows us to compare the number of events and non-events that we predicted with low and high probabilities. As the model fit improves, we should observe fewer shades of red in each of the plots.

4.1. Example: Voting Behavior (Rosenstone and Hansen, 1993).

As an example, we draw on Rosenstone and Hansen (1993) and King, Tomz and Wittenberg (2000). Rosenstone and Hansen (1993) employ a logit model to explain why some individuals are more likely than others to vote in U.S. presidential elections. King, Tomz and Wittenberg (2000) then used this model as a means to illustrate how to improve the statistical interpretation of regression analyses. For expository purposes, King, Tomz and Wittenberg (2000) focused only on the following demographic variables that Rosenstone and Hansen (1993) emphasized for explaining voter turnout in Presidential elections from the years 1960 to 1996: Education; Income; Age; and Race (whites and non-whites). King, Tomz and Wittenberg (2000) point out that after estimating a logit model, many scholars like Rosenstone and Hansen (1993) present only the coefficients, standard errors, and statistical significance. Rather than focusing on parameter estimates and t-statistics, they suggest that researchers should present quantities of interest using, for example, simulated first differences or expected values.

We recommend that researchers should go one step further beyond calculating quantities of interest and the degree of certainty about these quantities to evaluating model fit. Neither King, Tomz and Wittenberg (2000) nor Rosenstone and Hansen (1993) comment on the model's fit. We replicate the model of voter turnout described above and use the predicted and actual values to create a separation plot. As illustrated in Figure 9 below, the individual white and red lines become impossible to distinguish at this scale because of the large N ($N=15,837$) in this study.

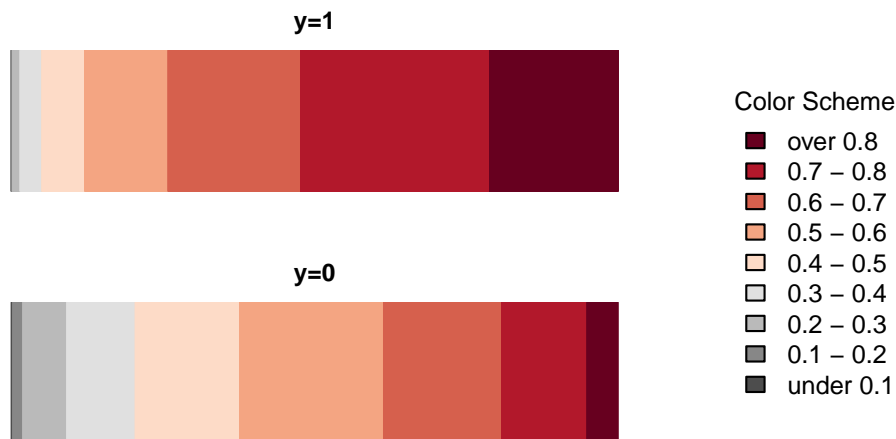
FIGURE 9. Separation plots for the Rosenstone and Hansen (1993) and King, Tomz and Wittenberg (2000) model of voter turnout in presidential elections between the years 1960 and 1996. The individual red and white lines are impossible to distinguish at this scale when N is so large.



Rather than expanding the size of the plot, we use the alternative version of the separation plot that groups probabilities into discrete bands as described above. In Figure 10 below, the colors correspond to ranges of probabilities; the darkest shade of red corresponds to a probability of 0.8 and higher and the lightest shade of red corresponds to a probability of 0.1 and lower.

The separation plot suggests that the model performs reasonably well in assigning high probabilities to actual cases of voter turnout. The darker red bands representing the high probabilities of the event occurring are considerably wider on the upper deck of the plot (which consists of the actual events) than in the lower deck (which consists of the non-events). Likewise, the grey bands corresponding to low probabilities of the event are wider in the lower deck than they are in the upper deck.

FIGURE 10. An alternative version of the separation plots for the Rosenstone and Hansen (1993) and King, Tomz and Wittenberg (2000) model of voter turnout in presidential elections between the years 1960 and 1996. The upper plot shows the results for cases of actual turnout, while the lower plot shows the results for cases of absenteeism. The upper plot illustrates a stronger fit to the data than the lower plot.



5. FUTURE AVENUES

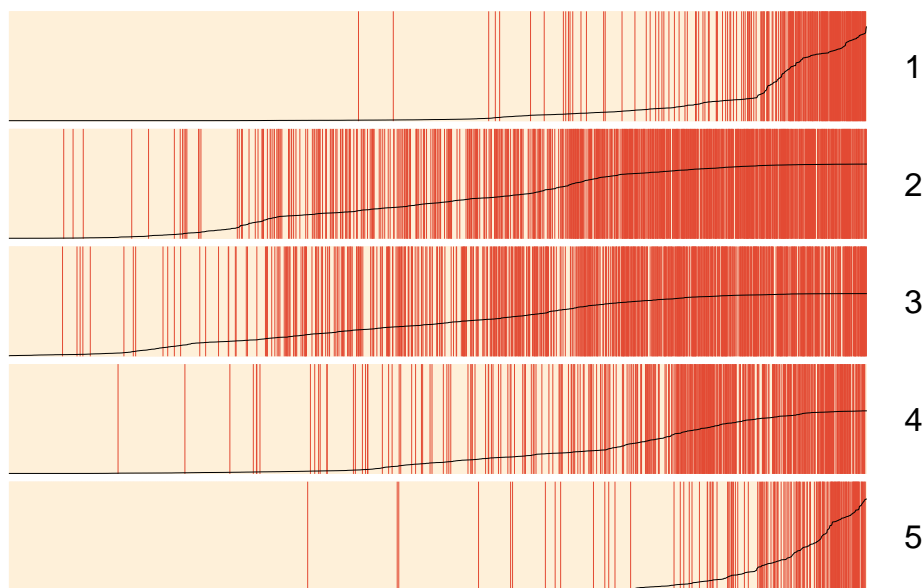
5.1. Polytomous Dependent Variables.

The concept of the separation plot can be extended to assist in the analysis of ordered probit/logit or multinomial models that have more than two categorical outcomes. In these cases, an individual separation plot is generated for each category of outcome. Each plot therefore compares the probabilities of obtaining one particular outcome against a dichotomous indicator of whether or not each case is in fact associated with that particular category

of outcome. The user can then review an array of separation plots that represent the predictive power of the model across all possible categories of outcome.

In Figure 11 we show an example of how separation plots can be used to assess the predictive power of an ordered probit model. In this figure we replicate the results of Neumayer (2005: Table 2, Model 6), which examines the effect that ratification of the International Covenant on Civil and Political Rights has on states' human rights practices. The dependent variable is the five-point "Political Terror Scale" which measures the extent to which governments engage in human rights violations aimed at disrupting non-violent political activities. Higher scores on the Political Terror Scale represent greater levels of repression.

FIGURE 11. Separation plots for each of the five levels of personal integrity rights estimated in Neumayer (2005: Table 2, Model 6).



As the graphs show, on the whole Neumayer's model does a good job of matching high probabilities of each outcome to actual occurrences of each outcome. For example, the first separation plot in Figure 11 shows that the highest predicted probabilities of generating an outcome of 1 on the Political Terror Scale do in fact correspond to actual outcomes of 1. For the intermediate categories 2, 3 and 4, the model appears to be somewhat less discriminating than it is for categories 1 and 5.

When evaluating the predictive power of models for polytomous dependent variables, it is important to keep in mind that what matters is not the proportion of dark and light lines in each plot (which will of course depend upon how the outcome variable is distributed between the possible categories), but rather the overall extent to which the dark lines are separated from the light lines. For example, in the case of the Neumayer model in Figure 11, many more country-year cases in that dataset have Political Terror Scale values of 2, 3 and 4 than the extremes of 1 and 5.

6. CONCLUSION

The separation plot allows one to assess the fit of a (logit or probit) regression model for which the dependent variable is binary. It adds information about the fit of discrete regression models, and can enhance our ability to understand the uncertainty of the predictions which are made by such statistical models. The separation plot has the following characteristics:

- It provides a quick visual summary of the distribution of events and nonevents in the data. Thus, are the events rare or frequent?
- It illustrates whether observations with high predicted probabilities actually experienced the event. If so, the model has a high degree of predictive fit, which can be assessed visually, without predetermining the threshold for binning predictions.
- It illustrates the existence of clusters of false positives and false negatives, if either or both exist.
- It permits a fairly direct comparison between different models, both for the same data and for different data.
- It is relatively easy to implement, explain, and present. Since it is visual, it is grasped quickly.
- Just to be sure, it does not substitute for the use of simulation to compare expected values under different scenarios.

What is the downside of the separation plot? The basic downside is its upside. It is a visual presentation, one that has a certain subjective element to it. There are many single number summaries that could be employed to assess the fit of these kinds of models. Some of these are also arbitrary. What is the value for the area under the curve that indicates a model has a high degree of fit? Others, e.g., likelihood ratios, have a basis in statistical theory but generally require some comparison with other (often imaginary) models. The separation plot does not provide a single number summary, but rather provides a visual presentation of the fit. In this way it goes along with presentations of distributions of expected values under different scenarios, in which specific numerical tests are rarely presented, and the extent of differences is portrayed graphically. However, it would be possible to calculate a variety of interesting statistics on the separation plot, including but not limited to the Mann-Whitney U test. We prefer to leave the development of this to scholars who may find them necessary. Rather, we think that the use of separation plots to present the fit of logistic and probit models will serve the needs of many social scientists.

APPENDIX A. R PACKAGE

An R package containing the functions required to generate the various types of separation plot discussed in this paper will soon be available through the Comprehensive R Archive Network. In the meantime, copies of these functions can be obtained by contacting Brian Greenhill (bdgreen@u.washington.edu).

The main function in the package is called `separationplot`. It requires only two arguments: a vector of predicted probabilities (`pred`), and a vector of the corresponding outcomes (`actual`). The generation of the separation plot is therefore completely independent of the model estimation procedure. However, when using the function in model-fitting procedures

such as that described in Section 2.2 above, we would recommend incorporating the function within a cross-validation routine. This allows the user to easily compare alternative specifications of a model based upon a quick glance at their out-of-sample separation plots.

Optional arguments to the `separationplot` function include a facility to interactively identify one or more individual cases on the plot via mouse clicks (e.g., `locate=2`), and to flag a particular observation prior to generating the plot (e.g., `flag=1234`, `flagcol="blue"`). Other arguments provide fine-grained control over the form of the graphic output.

The two related functions included in the package are `sp.categorical` – a wrapper for `separationplot` that generates arrays of plots for models with polytomous dependent variables (see the example in Section 5.1 above), and `sp.largeN` – a function for generating the separation plot with binned probabilities as demonstrated in Section 9 above.

REFERENCES

- Aldrich, John and Forrest Nelson. 1984. *Analysis with a Limited Dependent Variable: Linear Probability, Logit, and Probit Models*. Beverly Hills, CA: Sage Publishers.
- Box-Steffensmeier, J.M. and B.S. Jones. 2004. *Event history modeling: A guide for social scientists*. Cambridge University Press.
- Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probabilities." *Bulletin American Meteorological Society* 78:1–3.
- Fearon, J.D. and D.D. Laitin. 2003. "Ethnicity, insurgency, and civil war." *American Political Science Review* 97(01):75–90.
- Hillygus, D.S. and S. Jackman. 2003. "Voter decision making in election 2000: campaign effects, partisan activation, and the Clinton legacy." *American Journal of Political Science* 47(4):583–596.
- Imai, Kosuke, Gary King and Olivia Lau. 2008. "Toward A Common Framework for Statistical Analysis and Development." *Journal of Computational and Graphical Statistics* 17(4):892–913.
- King, G., M. Tomz and J. Wittenberg. 2000. "Making the most of statistical analyses: Improving interpretation and presentation." *American Journal of Political Science* pp. 347–361.
- Marshall, M. and K. Jaggers. 2003. "Political Regime Characteristics and Transitions 1800–2003." *Polity IV Project*.
- Neumayer, E. 2005. "Do international human rights treaties improve respect for human rights?" *Journal of conflict resolution* 49(6):925–953.
- O'Brien, Sean P. 2010 forthcoming. "Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research." *International Studies Review* 6(4):tba.
- Rosenstone, S.J. and J.M. Hansen. 1993. *Mobilization, participation, and democracy in America*. Macmillan Pub Co.
- Simmons, B.A. and Z. Elkins. 2004. "The globalization of liberalization: Policy diffusion in the international political economy." *American Political Science Review* 98(1):171–189.
- Tufte, E.R. 2006. *Beautiful evidence*. Graphics Press Cheshire, Conn.
- Ward, M.D., B.D. Greenhill and K. Bakke. forthcoming. "The Perils of Policy by P-Value: Predicting Civil Conflict." *Journal of Peace Research*.

DEPARTMENT OF POLITICAL SCIENCE, UNIVERSITY OF WASHINGTON, SEATTLE, WA, 98195
E-mail address: bdgreen@u.washington.edu

DEPARTMENT OF POLITICAL SCIENCE, DUKE UNIVERSITY, DURHAM, NC, 27707
E-mail address: mw160@duke.edu

DEPARTMENT OF SOCIOLOGY, UNIVERSITY OF WASHINGTON, SEATTLE, WA, 98195
E-mail address: sacks@u.washington.edu