

Formal Tests of Substantive Significance for Linear and Non-Linear Models*

Justin Esarey[†] and Nathan Danneman[‡]

January 23, 2011

Abstract

We present a critical statistic c^* for determining the substantive significance of an empirical result, which we define as the degree to which the result justifies a particular decision (such as the decision to accept or reject a theoretical hypothesis). Our procedure, which is built on ideas from Bayesian statistical decision theory, helps researchers improve the objectivity, transparency, and consistency of their assessments of substantive significance. We provide software tools for calculating c^* for a wide variety of models.

Key Words: substantive significance; statistical decision theory

JEL Classification: C11, C12, C44

Introduction

How should economists come to conclusions on the basis of empirical evidence? Even in academic research, researchers are expected to go beyond presenting results (e.g., coefficients, marginal effects, and out-of-sample predictions) and synthesize these results into a judgment about some scientific or policy question. In the context of theory testing, for example, authors and readers are expected to decide whether empirical evidence is consistent with a hypothesis (and therefore supports a theory). Economists may also be expected to comment on whether various policies are justified by the results. These judgments, which we will call judgments of the *substantive significance* of a finding, cannot be read directly off of a table,

*This research was supported by a University Research Committee grant from Emory University.

[†]Assistant Professor, Department of Political Science, Emory University. Corresponding author (jesarey@emory.edu).

[‡]Department of Political Science, Emory University. E-mail: (ndannem@emory.edu).

plot, or confidence interval: interpretation is required. Furthermore, statistical significance tests are not dispositive tests of substantive significance because they do not highlight the substantive magnitude of a result along with its uncertainty (Cohen, 1994; Schmidt, 1996; McCloskey and Ziliak, 1996; Gill, 1999; Ziliak and McCloskey, 2004; Zellner, 2004; Ziliak and McCloskey, 2008).

Consider, for example, the decision to reject or accept (fail to reject) a theoretical hypothesis based on quantitative evidence. The scientific consequences of acceptance (compared to rejection) are contingent on the true state of the world. Posterior probability distributions derived from the econometric analysis of data tell researchers the rational beliefs they should hold about the state of the world based on prior beliefs and present evidence. If we have complete and transitive preferences over consequences, we should in principle be able to formally structure our decision about the substantive significance of a result (in this case, whether the data supports accepting the hypothesis) using expected utility theory (Wald, 1950; DeGroot, 2004 {1970}; Pratt, Raiffa and Schlaifer, 1996; Manski, 2007, ch. 12). But we rarely bring this procedure to bear when making decisions using statistical results. The result is a lack of transparent and consistent standards for substantive significance that make it unclear how the magnitude and uncertainty of a result combine with the researcher's preferences to produce the resulting judgment (Horowitz, 2004; Lunt, 2004).

In this paper, we present a formalized process for making judgments of substantive significance using Bayesian statistical decision theory¹ and present a new application to non-linear econometric models and linear models with interaction terms. The process uses a critical statistic, c^* , that allows researchers to easily summarize and communicate the substantive significance of their results. We believe that using this procedure offers three benefits to the scientific community. First, c^* helps researchers make decisions about substantive significance that are as objective, transparent, and consistent as judgments of statistical significance. Second, it makes executing a statistical decision theoretic analysis considerably easier, removing

¹The basic ideas for the process we present were first expressed in Esarey (2010).

a major barrier to more widespread use of formalized judgments of substantive significance (Altman, 2004, 654-656). Finally, our software is flexible enough to accommodate a variety of preference structures without reprogramming, allowing researchers to set different standards for different decisions and easily communicate these standards to others in an objective and transparent way.

Why formalize substantive significance?

As we alluded in the introduction, a great deal of work has criticized the use of statistical significance tests as indicators of substantive significance. However, although this work argues that the size of the estimated effect is as important as its uncertainty, no formal measure or test statistic is offered that would help regularize these decisions the way that a *t*-test regularizes judgments of statistical significance.

For example, suppose a theory predicts that economic growth should be positively associated with incumbent re-election in democracies. If an appropriate econometric model indicated that a 1% increase in growth is associated with a 0.01% increased probability of re-election, would it matter whether the variance on that re-election probability was small enough to definitively exclude zero? In either case, the practical consequences of the theory's predictions are probably too small for the evidence to be called even weakly supportive of the theory, no matter the associated *p*-value. Certainly, few policy recommendations could be built on such a finding even if $p < 0.001$. But just how large (and certain) would the effect have to be before a researcher would deem it substantively meaningful? What criteria underlie this judgment? The formal framework that we present provides a means of making these decisions and making their criteria clear.

Perhaps when an estimated effect is tiny, as in the growth-election example above, we can all agree on its substantive insignificance. But without clear, communicable, and objective criteria, it is much more difficult to understand more marginal decisions—say, where a 1%

increase in growth is associated with a 10% increased probability of re-election, with a 95% confidence interval² of $\pm 13\%$ (two-tailed $p = 0.132$). Such a result might be substantively significant—it might provide a sound basis for some conclusions or decisions—but without a formal framework it is hard to decide whether the large scale of the result can offset its comparatively large variance.

A framework for formalizing substantive significance

Rational choice theory can inform a test statistic for substantive significance that helps researchers make decisions about substantive significance in a rational, consistent, and transparent way. We begin with a reasonably simple framework for rational choice that goes back to the foundations of economic theory; although the application to statistical inference dates back at least to Wald (1950; see also Zellner and Chetty, 1965), the ideas have never entered the mainstream of scientific statistical practice despite some important follow-on work among Bayesian statistical decision theorists (DeGroot, 2004 {1970}; Pratt, Raiffa and Schlaifer, 1996; Zellner, 2002; Manski, 2007, ch. 12). Consider a binary decision, such as the decision to accept or reject a bet with a price p and a probability π of paying off winnings w . If an individual has a utility function u mapping payoffs into utility space and the status quo utility is normalized to 0, then s/he should accept the bet whenever:

$$\pi u(w - p) + (1 - \pi)u(-p) \geq 0$$

That is, a rational individual should accept the bet whenever the probability-weighted utility of acceptance is greater than the certain utility of turning down the bet. If there were a continuum of possible payoffs (and losses) $w \in \mathbb{R}$ each associated with a probability density

²It is worth noting that these confidence intervals are usually still reported at the 95% or 90% levels that correspond to the $\alpha = 0.05$ standard in a two or one-tailed test, respectively.

$f(w)$, then acceptance is contingent on:

$$\int u(w - p)f(w)dw \geq 0$$

Our approach to substantive significance is to apply this rational choice framework to statistical inference as prescribed by statistical decision theory and described in Esarey (2010). The key methodological innovation is to distill the process into an easy-to-interpret critical statistic, c^* , that researchers can use to determine the substantive significance of their result.

Rational choice and substantive inference with c^*

For statistical inference, cash winnings are replaced by states of the world that bear on some decision, such as the decision to accept or reject a hypothesis predicted by a theory. Each state of the world has an associated probability density that comes out of the posterior belief distribution of an empirical model. A utility function specified by the researcher maps these states of the world into consequences contingent on the decision. Making a rational decision involves combining the empirical results (the posterior probability density) with the utility function to find the choice that maximizes expected utility.

How do empirical results translate into consequences in the context of academic research? Consider the example of economic growth and incumbent re-election. If we have a theory that predicts that positive economic growth leads to increases in re-election probability, the consequences of accepting that theory are contingent on the true state of the world. If the true relationship is *negative*, presumably accepting the theory leads to worse consequences (misinformed future research and mistaken policy advice) compared to rejecting the theory. If the true relationship is positive but very small—too small to be meaningful for understanding political economy or formulating policy advice—accepting the theory still has negative consequences compared to rejecting it, for the same reasons as above. If the true relationship is positive and sufficiently large, better consequences come from accepting the theory

compared to rejecting it: our future research and policy advice becomes better-informed via using the theory.

As noted, we supposed that relationships must be of a certain size before they can be considered substantively meaningful. Presumably, this threshold might be different for different sorts of decisions. Call the threshold for substantive importance for a particular decision c : a relationship must be of at least size c before it makes sense to make a certain decision (accept a theory, change policy advice, etc.) on the basis of that relationship.

Continuing the applied example above, let an empirical estimate of the marginal effect of economic growth on re-election probability be designated by k . The critical condition³ for accepting the substantive significance of a result is:

$$\int [u(\text{accept}|k - c) - u(\text{reject}|k - c)] f(k)dk \geq 0 \tag{1}$$

We can numerically solve equation 1 for a root in c , calling the root c^* , and use this solution as a statistic to measure the substantive significance of a result on a continuous scale. That is, c^* divides the space of decisions in two: decisions with $c > c^*$ cannot be justified by the empirical evidence at hand, while decisions with $c \leq c^*$ can be justified.⁴ Quadrature techniques can be used to calculate the integral for any particular value of c , while maximization algorithms (such as Newton-Raphson) allow the finding of c^* ; in both cases, readily-available libraries can perform the necessary calculations for most statistical software packages.

The c^* statistic helps a researcher quickly and intuitively determine whether his/her results are substantively significant. The analyst simply asks whether a relationship of size c^* is large enough to be substantively meaningful in the context of the decision at hand. If so, then the effect should be considered substantively significant.

³When $k < 0$, and the predicted effect is negative, the $k - c$ term should be replaced by $-(k - c)$.

⁴When $k < 0$, these inequalities are reversed.

Utility functions

Of course, the inference that a researcher draws about the substantive significance of a result is contingent on his/her choice of utility function, and there is no universally correct function to employ (just as there is no universally correct α -level for a t -test). But formalizing substantive significance neither creates nor exacerbates this problem: one reason for formalization is to clarify the criteria for these decisions and to help researchers make them consistently.

Our formal test of substantive significance is designed to help researchers make reasonable choices about the utility function u under the most common empirical situations, including significant flexibility to adjust this utility function to fit the criteria of a particular decision, and to help them clearly communicate the basis for their decisions about substantive significance to other researchers. Importantly, the goal is not to choose u to match the way that people behavioralistically make choices under uncertainty. Rather, our goal is to allow a researcher to consciously and methodically articulate goals, and to help that researcher rationally and consistently make decisions by strict adherence to those stated goals; indeed, we are trying to *avoid* decision paradoxes that might otherwise find their way into these decisions under a behaviorally accurate model.

Two types of utility functions are built into our software tools by default, both designed for testing of directional hypotheses (e.g., that the marginal effect of X on y is positive or negative): a kinked linear loss averse utility function, and the constant absolute risk aversion (CARA) function. Many other functions can be added with minimal programming.

Kinked linear loss aversion

A kinked linear utility function represents simple loss-averse preferences:

$$u(\text{accept}|k, \gamma, c) - u(\text{reject}|k, \gamma, c) = \gamma^{-\text{sign}(k-c)} [k - c]$$

Here, $k > 0$ represents a finding in the predicted direction and $k \leq 0$ represents a finding in the opposite direction.⁵ Thus, an inference about substantive significance is made by numerically determining a root c^* that solves:

$$\int \left[\gamma^{-\text{sign}(k-c)} [k - c] \right] f(k) dk = 0 \quad (2)$$

The parameter γ measures the researcher's degree of loss aversion.

The kinked linear utility function mathematically embodies several qualities, which we believe are a good match for many researchers' preferences:

1. Accepting a correct prediction of substantively meaningful size ($> c$) yields positive utility; accepting a prediction of meaningless size, or in the wrong direction, yields negative utility.
2. False positives (accepting incorrect hypotheses) are treated differently than false negatives (rejecting correct hypotheses). If $\gamma > 1$, false positives are treated as being more harmful than false negatives. If $\gamma \in (0, 1)$, false negatives are treated as more harmful. For example, if false positives are four times as damaging as false negatives, then $\gamma = \sqrt{4} = 2$. The γ parameter can be set by the researcher to fit the decision at hand.
3. Within the gains frame or losses frame, utility is proportionate to the size of a finding on an interval scale and relative to the scale of the underlying variables in the data set. This implies that, holding the scale of the data constant, a correct prediction of magnitude k is half as important as a correct prediction of magnitude $2k$, but only $\frac{1}{\gamma}$ as important as an *incorrect* prediction of magnitude k .

The key advantages of the kinked linear loss aversion function is its simplicity and flexibility. The γ parameter is easy to explain and choose: γ should be set so that false positives

⁵As indicated above, replace $k - c$ with $-(k - c)$ when the predicted direction is negative.

are γ^2 times as important or damaging to scientific progress or policy outcomes as false negatives. Setting this parameter permits the researcher a considerable amount of discretion in setting the criteria for the particular decision, but makes the precise nature of that discretion completely clear and communicable to others interpreting the results. Finally, inferences do not change as a function of the scale of the dataset, so long as the resulting c^* is interpreted relative to these scales.

Constant absolute risk aversion (CARA)

A concave-downward utility function represents constant absolute risk averse preferences:

$$u(\text{accept}|k, \delta, c) - u(\text{reject}|k, \delta, c) = \frac{1}{\delta} (1 - \exp(-\delta [k - c]))$$

An inference about substantive significance is made by determining a root c^* that solves:

$$\int \left[\frac{1}{\delta} (1 - \exp(-\delta [k - c])) \right] f(k) dk = 0 \tag{3}$$

The δ parameter is the Arrow-Pratt coefficient, a measure of the concavity of the utility function. Values of $\delta > 0$ represent risk averse preferences (false positives are treated as being more harmful than false negatives), while values of $\delta < 0$ represent risk seeking preferences (false negatives are treated as being more harmful than false positives).

The CARA function shares some qualities with the kinked linear utility function above. As with loss averse preferences, positive utility comes from accepting a prediction of substantively meaningful size ($k > c$) and negative utility comes from accepting any other prediction ($k \leq c$). False positives are still treated differently than false negatives; the degree to which one is overvalued is set by the parameter δ . However, there are key differences:

1. When using the CARA function, gains have marginally diminishing positive utility, while losses have marginally increasing negative utility. In the kinked linear utility function, gains and losses have constant marginal utility.

2. As a corollary to item (1) above, the distinction between the gains and losses frame is greater when compared to the kinked linear utility function. More specifically, gains tend to be more strongly undervalued relative to losses in the CARA function.
3. For the CARA function, the scale of the underlying independent variables is relevant to inference. The reason is similar to the one laid out in Rabin and Thaler (2001): doubling the scale of a bet (in this case, an uncertain empirical finding) changes its certainty equivalent because of decreasing marginal utility in gains and increasing marginal utility in losses. Thus, choices of δ may be not only specific to a problem, but to the *scale* of a problem.
4. The δ parameter has a somewhat technical interpretation as the degree of relative decline in marginal utilities:

$$\delta = -\frac{u_i''(w)}{u_i'(w)}$$

This makes δ more challenging than γ to choose or interpret, especially in light of item (3).

For all these reasons, the CARA function may be more challenging to use and interpret than kinked linear loss aversion. Nevertheless, some researchers may believe that concave risk aversion (or convex risk-seeking) preferences are the best match for their situation, and our software enables them to act rationally on this choice and communicate it clearly to the outside world. Procedures originally developed for experimental studies exist to help people determine the δ appropriate for a given problem (Holt and Laury, 2002). For inferences of substantive significance, it might be simpler to have a researcher answer the following question suggested by Esarey (2010):

Suppose that $c = 0$ (any relationship is substantively meaningful) and evidence that you gathered presented you with two⁶ possibilities: there is a 95% chance

⁶The posterior is presented as a distribution with two point masses to simplify calculations and make the situation easy to understand.

that the true standardized coefficient β describing the relationship between X and Y is $\beta_H = 1$, and a 5% chance that the true relationship is some value of $\beta_L < 0$. How low would β_L have to be before you refused to conclude that $\beta > 0$?

With a reported β_L , δ can be determined by solving:

$$0.95 \frac{1}{\delta} (1 - \exp(-\delta)) + 0.05 \frac{1}{\delta} (1 - \exp(-\beta_L \delta)) = 0$$

For example, a $\beta_L = -2$ implies a $\delta = 1.36$, while a $\beta_L = -4$ implies a $\delta = 0.551$.

Parameter Choice and Sensitivity Analysis

In both the default utility functions that we offer above, and in many others besides, inference depends on a choice of some parameter (γ or δ in our defaults). In most situations, e.g. for the testing of hypotheses derived from theories, there is no universally correct choice for these parameters: the goal of our procedure is to make the criteria for substantive significance transparent and consistent, not to create universal criteria.

Nevertheless, it may be difficult to determine the right parameter value that corresponds to a researcher's qualitative level of risk or loss aversion. As a result, a researcher may reasonably inquire about how sensitive conclusions about substantive significance are to differences in γ or δ . It is relatively easy to make this determination: a researcher can simply determine c^* using equation 2 (or 3) for a sequence of γ (or δ) values, then plot the resulting vector of c^* against γ (or δ). We will show examples of sensitivity analysis in the next section.

Tools for determining substantive significance

Performing and interpreting a statistical decision theoretic analysis for every empirical model might be onerous without simple and flexible tools to help a researcher make choices for $u()$ and translate these choices into inferences about substantive significance. This is especially

true for empirical models whose key results must be extracted from the model via simulation rather than read directly from a coefficient table. We make it comparatively simple to perform analysis with brand new tools for a wide variety of statistical models, which we have implemented in the R and Stata statistical environments.

As indicated by equation 1, we need three elements to calculate c^* for a judgment of substantive significance:

1. a utility function
2. an estimate of k , the statistic of interest
3. an estimate of $f(k)$, the posterior density of the statistic of interest

In most contexts, the consequences of a decision are contingent on a marginal effect or a point prediction. For example, comparative static predictions of a theory are typically tested by comparing them to estimated marginal effects from an empirical model. Thus, our tools focus on testing for substantive significance of predicted marginal effects.

We have implemented our tools for a wide variety of models that encompass most tools of everyday econometric analysis:

1. simple linear regressions, where marginal effects are typically identical to estimated coefficients
2. linear regressions with interaction terms, where the distribution of marginal effects is an analytical function of estimated coefficients
3. non-linear models where the distribution of predicted marginal effects are derived from estimated coefficients via simulation

Calculating and interpreting c^* in simple linear regressions was first implemented in Esarey (2010). But the tools proposed there will not work for linear models with interaction terms and non-linear models, where simulation methods must be used to calculate most marginal effects of interest and their distributions.

Simple linear models

For simple linear regressions, marginal effects are typically identical to estimated coefficients. Because of this identity, testing for substantive significance using statistical decision theory is straightforward in many instances. Let the utility function u be a function of the size of the marginal effect, β , and the minimum size threshold for substantive significance for a particular decision, c . Thus, we should accept the theory whenever:

$$\int [u(\text{accept}|\beta, c) - u(\text{reject}|\beta, c)] f(\beta|\text{data})d\beta > 0 \quad (4)$$

When using the kinked linear utility function,⁷ this leads us to seek the root c^* of the equation:

$$\int [\gamma^{-\text{sign}(\beta-c)} [\beta - c]] f(\beta|\text{data})d\beta = 0$$

Because marginal effects are usually identical to estimated coefficients for linear regressions, $f(\beta|\text{data})$ gives us the posterior belief distribution of marginal effects that we need to assess how consistent this evidence is with the theory being tested. The formula for the posterior distribution of β given a dataset is given by Bayes' rule (Zellner, 2007):

$$f(\beta|\text{data}) = \frac{f(\text{data}|\beta)f(\beta)}{\int f(\text{data}|\beta)f(\beta)d\beta} \quad (5)$$

This is the “Bayesian” part of Bayesian statistical decision theory. Our software tools default to an uninformative prior distribution, a uniform distribution on the interval $\hat{\beta} \pm 8\hat{\sigma}_\beta$, which allows us to interpret frequentist results in a Bayesian fashion. Informative prior distributions can be used with additional programming.

⁷As above, $\beta - c$ is replaced by $-(\beta - c)$ when $\beta < 0$.

Table 1: The impact of anti-sweatshop protests on real wages in Indonesia (Table 2, Column 5 from Harrison and Scorse, 2010)

Variable	Beta	SE
Foreign-owned or export oriented industry	-0.097	0.025
Textiles, footwear, and apparel industries	-0.031	0.032
Foreign/export X TAP industry	0.202	0.036
Controls	–	–

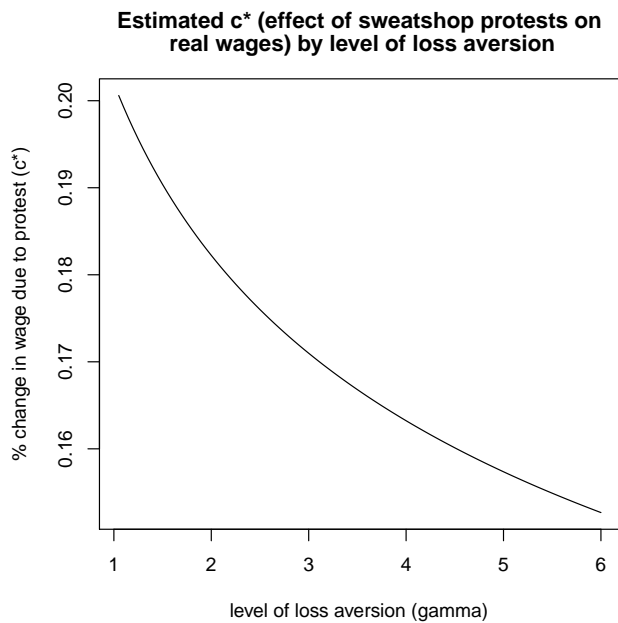
OLS model of the logged change in plant unskilled wage rates between 1996 and 1990. $R^2 = 0.13$, $N = 5,920$.

Applied Example: Harrison and Scorse (2010)

To illustrate how our technique can be used to communicate the substantive significance of empirical results, we re-examine results reported by Harrison and Scorse (2010) in the *American Economic Review*. Harrison and Scorse examine whether anti-sweatshop activism in the 1990s targeted at Nike (and other shoe and apparel manufacturers) increased the real wages of textile, footwear, and apparel (TFA) workers in Indonesia. To do so, the authors use OLS to compare logged growth in the TFA industries that are foreign-owned or export oriented to growth in other foreign-owned or export oriented sectors. One of their key results is depicted in Table 1. According to their results, Indonesian plants in the exporting or foreign owned TFA sectors targeted by activists increased their real wages 20.2 percent compared to foreign owned or export oriented plants in non-targeted industries. This finding, the apparent effect of anti-sweatshop protests on wages, is highly statistically significant ($p < 0.001$). But is it substantively significant?

The answer is depicted in Figure 1. To produce this figure, we adopt the loss aversion framework and calculate c^* for values of γ between 1 and 6; these values of γ represent valuing losses between being equally important compared to gains ($\gamma = 1$) to being 36 times more important than gains ($\gamma = 6$). As the figure shows, even under extreme loss aversion, the data are consistent with a substantively strong and positive effect of protest on wages. That is, the effect of protests seems to be strongly substantively significant for reasonable

Figure 1: Sensitivity Analysis



ranges of aversion to mistaken decisions.

Interacted Linear Models

In many cases, such as for linear models with interaction effects, marginal effects are not identical to estimated coefficients, but a function of these coefficients. Consider a linear regression model with an interaction term:

$$y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_p XZ \quad (6)$$

The marginal effect of X on y is:

$$\frac{\partial y}{\partial X} = y'_X = \beta_1 + \beta_p Z \quad (7)$$

Our estimate of the variance of this effect is:

$$\text{var}(y'_X) = \text{var}(\beta_1) + Z^2 \text{var}(\beta_p) + 2Z \text{cov}(\beta_1, \beta_p) \quad (8)$$

Thus, to test a theoretical prediction about the marginal effect of X on y in an interacted model, we must look at the distribution of y'_X , $f(y'_X|f(\beta|\text{data}), Z)$ for multiple values of Z . This task is fairly straightforward with access to simulated draws from the distribution of coefficients β , $f(\beta|\text{data})$, which are easily calculated via Markov Chain Monte Carlo analysis using equation 5 or by drawing from the asymptotically normal distribution of $f(\beta|\text{data})$ using the VCV of an ordinary least squares estimate. With these draws from $f(\beta|\text{data})$ in hand, one need only calculate y'_X for each draw at values of Z and then examine the resulting empirical distribution.

Inference about the substantive significance of such a marginal effect can make use of samples from the distribution $f(y'_X|f(\beta|\text{data}), Z)$ to calculate the root in c of:

$$\int [u(\text{accept}|y'_X(\beta, Z), c) - u(\text{reject}|y'_X(\beta, Z), c)] * f(y'_X|f(\beta|\text{data}), Z) dy'_X = 0 \quad (9)$$

With a sufficient number of samples from $f(y'_X|f(\beta|\text{data}), Z)$ to allow for accurate kernel density estimation of the underlying distribution, the integral on the right hand side of equation 9 can be numerically calculated as easily as that of equation 4 using standard numerical integration packages. Of course, the root c^* will be different for every value of Z : there may be values of Z for which the marginal effect of X on y is very strong and certain (and substantively significant) and others for which it is less strong or certain. We should therefore be able to plot calculated values of c^* against Z to show the substantive significance of this marginal effect for different values of the interacted variable. Our software implements this process, as demonstrated in the following example.

Applied Example: Brambor, Clark and Golder (2006)

In an influential paper, (Brambor, Clark and Golder, 2006, hereafter BCG) describe the importance of interpreting interaction terms substantively, and provide tools and techniques

Table 2: The impact of presidential elections on the effective number of electoral parties (Table 1 from Brambor, Clark and Golder, 2006)

Regressor	β	s.e.
Election Proximity	-3.44	0.49
Presidential Candidates	0.29	0.07
Proximity * Pres. Cand.	0.82	0.22
Controls	–	–
Constant	3.01	0.33

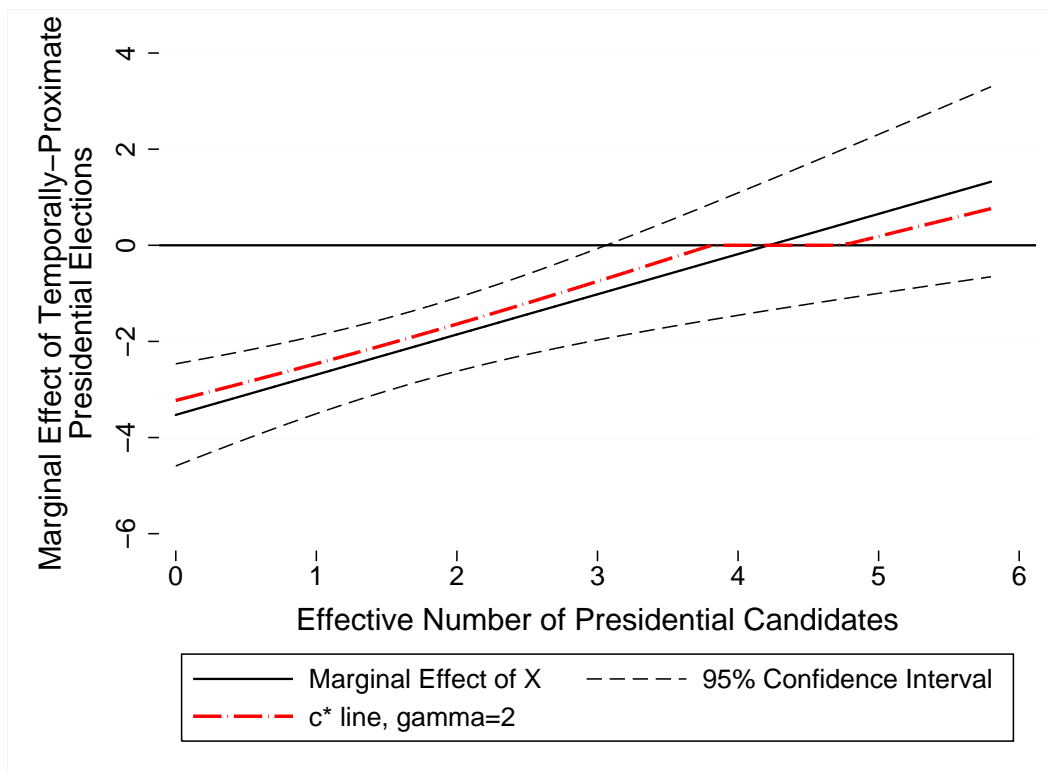
OLS model of the number of electoral parties in the legislature. $R^2 = 0.34$, $N = 522$.

for informative interpretation. These tools and techniques are readily adapted to substantive inference using c^* . In one example, BCG present a graphical method of presenting marginal effects from interacted linear models, like those of equation 7. Their example comes from Golder (2006), a model of the relationship between presidential elections and legislative fragmentation (the number of parties in the legislature). The model is listed in Table 2; note that the relationship between the proximity of presidential elections and the number of electoral parties in the legislature is contingent on the number of presidential candidates.

To test hypotheses about the marginal effect of presidential elections on legislative fragmentation, BCG recommend constructing a plot like the one in Figure 2, which shows this marginal effect and its distribution for different numbers of presidential candidates. We have augmented the code provided by BCG to construct this plot to add a c^* line that shows the root of equation 9 for each number of presidential candidates using the kinked linear loss averse utility function from equation 2; we set $\gamma = 2$ for this plot.

As the figure shows, there is a range of presidential candidates (between 3.9 and 4.7) for which there is *no* substantively significant relationship between presidential elections and legislative fragmentation; that is, $c^* = 0$ for these values. But above 4.7 presidential candidates, a substantively significant relationship can be supported for decisions with a sufficiently small threshold. For example, when the number of presidential candidates equals

Figure 2: Marginal Effect of Proximate Presidential Elections on Number of Electoral Parties in the Legislature (Figure 3 from Brambor, Clark and Golder, 2006)



5.8, $c^* = 0.76$; that is, if an average increase in $\frac{3}{4}$ of a party in the legislature is substantively meaningful, then these results support a substantively significant relationship between presidential elections and legislative fragmentation.

Note that c^* is essentially the same ($\approx \frac{3}{4}$) when the number of presidential candidates = 3, the point at which this relationship becomes statistically significant, as it is when the number of presidential candidates is 5.8. This is one example of how c^* helps analysts to make consistent decisions once they have specified their preferences: if we believe that a meaningful negative relationship between presidential elections and legislative fragmentation exists when the number of presidential candidates = 3, one must rationally conclude that a *positive* and substantively significant relationship exists for a very large number of presidential candidates (= 6).

We can also adjust γ to perform a sensitivity analysis on the results. For example,

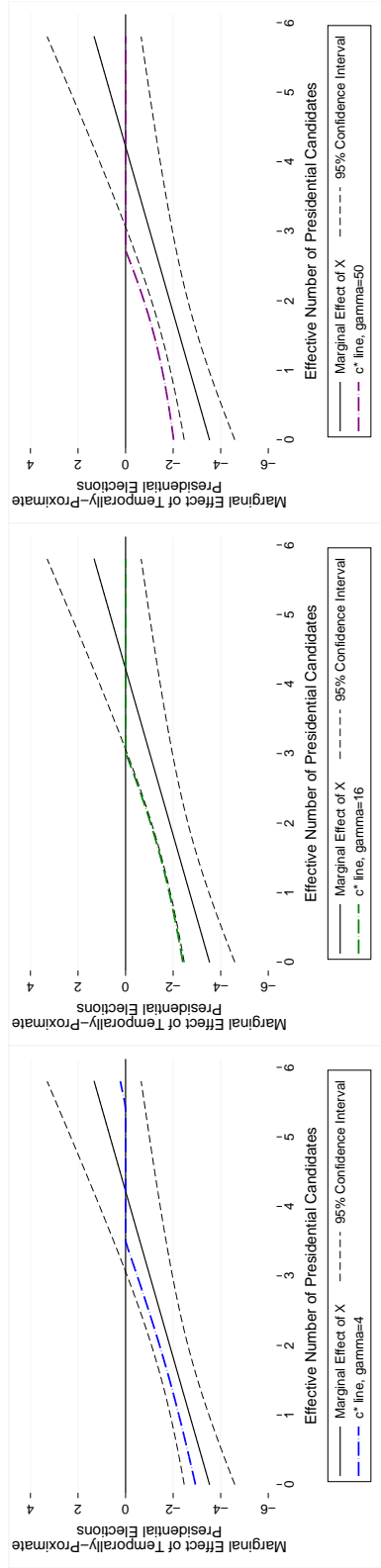
Figure 3 shows the analysis from Figure 2 repeated for three different values of γ : 4, 16, and 50 respectively. The higher that γ is set—that is, the more that an analyst values false positives more than false negatives—the more conservative the results become: when $\gamma = 4$, for example, the number of presidential candidates for which there is no substantively significant relationship between presidential elections and legislative fragmentation expands to [3.5, 5.4], compared to [3.9, 5.7] when $\gamma = 2$. Consider another illustration: when $\gamma = 2$ (false positives are valued 4 times as much as false negatives), the c^* for this relationship when the number of presidential candidates is one is $c^* = -2.46$, while when $\gamma = 4$ (false positives are valued 16 times as much as false negatives), $c^* = -2.24$. Thus, as γ rises, an analyst must have a lower threshold for substantive significance in order to continue believing that the result is substantively meaningful. Stated differently, the substantive significance of a result falls as an analyst becomes more averse to false positives compared to false negatives.

Another interesting result of the sensitivity analysis can be seen in the center panel of Figure 3: when $\gamma = 16$, the c^* line is nearly identical to the 95% confidence interval threshold. Typically, an analyst uses this threshold to test hypotheses: when a marginal effect is not statistically significant, it is equivalent to zero effect and therefore substantively insignificant for any purpose. Yet as we can see, this behavior is consistent with a preference ordering that values false positives $16^2 = 256$ times more than false negatives, a degree of loss aversion probably well beyond what many analysts would embrace. To put the preference into perspective, a person with these preferences would refuse to pay \$4.00 for a lottery ticket with a 50% chance of paying off \$1000!

Non-Linear Models

In many cases, non-linear models will present a disjunct between fitted coefficients and the effects of substantive interest similar to the case of interacted linear models above. Consider

Figure 3: Sensitivity Analysis for Figure 2



a fairly straightforward probit model:

$$\Pr(y = 1) = \Phi(\beta_0 + \beta_1 X + \beta_2 Z) \quad (10)$$

The marginal effect of X on $\Pr(y = 1)$ is not the coefficient β_1 , but:

$$\frac{\partial \Pr(y = 1)}{\partial X} = \phi(\beta_0 + \beta_1 X + \beta_2 Z) \beta_1 \quad (11)$$

It is typical to calculate and showcase changes in $\Pr(y = 1)$ predicted by discrete changes in a particular X holding other independent variables constant at some central value. This process is facilitated by software libraries such as Clarify (King, Tomz and Wittenberg, 2000). A researcher substitutes the appropriate value for Z , such as its mean, then calculates the value of equation 10 for two different values of X ; the difference between them is the discrete effect of X on $\Pr(y = 1)$. The distribution of this marginal effect can be calculated by drawing from the distribution of β , then repeating the difference calculation for each draw. Thus, the statistic of interest is:

$$d\Pr(y = 1) = p'_X = \Pr(y = 1|X = X_{hi}, Z = \bar{Z}) - \Pr(y = 1|X = X_{lo}, Z = \bar{Z})$$

A calculation of c^* for this discrete change in $\Pr(y = 1)$ involves finding the root of:

$$\int [u(\text{accept}|p'_X(\beta, X_{lo}, X_{hi}, Z), c) - u(\text{reject}|p'_X(\beta, X_{lo}, X_{hi}, Z), c)] * \quad (12)$$

$$f(p'_X|f(\beta|\text{data}), X_{lo}, X_{hi}, Z) dy'_X = 0$$

As with the case of interacted linear models, the distribution of the discrete change p'_X has to be calculated via simulation. Also as before, with a sufficient number of draws from $f(p'_X|f(\beta|\text{data}))$ we can use a kernel density estimate of the underlying distribution to numerically integrate equation 12 and use standard techniques to find a root c^* . Our

software modifies the Clarify package to calculate a c^* for generalized linear models, such as the probit, and we demonstrate the software below.

Applied Example: Brender and Drazen (2008)

In their 2008 piece, (Brender and Drazen, 2008, hereafter B&D) examine the effects of economic growth and fiscal policy on re-election prospects in a broad panel of democracies. Their stated aims are to examine whether loose fiscal policies and increased deficit spending during election years, and growth generally over the course of an incumbent's term, raises the probability that incumbents will be re-elected. B&D further explore whether these relationships differ across democracies at different levels of economic development.

The results of their core model are shown in Table 3. In developed countries, B&D find that expansionary fiscal policies (specifically, falling budget surplus-to-GDP ratios in the period preceding the election) hurt incumbents' chances of re-election. Economic growth during the incumbents' term, however, has no effect on re-election prospects in developed countries. In developing nations, by contrast, expansionary fiscal policy is negatively related to re-election prospects, but economic growth is associated with a higher probability of incumbent re-election.

B&D produce statistics to show that the effects of macroeconomic policy and growth reported in Table 3 are substantively meaningful, not just statistically significant; we can use our c^* software for probit models to formally examine the substantive significance of these results. For example, in developed countries, a one percent increase in the budget surplus to GDP ratio over an incumbent's term (that is, more austere fiscal policy) raises the probability of incumbent re-election by 3.28% (95% CI:[-0.450%, 7.27%]) when the values of all variables are set at their mean. Using a kinked linear loss averse utility function with $\gamma = 2$, this result translates to a c^* of 2.19%. Loss-averse analysts (who believe that false positives are four times more damaging than false negatives) can use this statistic—which encapsulates the size and uncertainty of the estimate, plus their aversion to falsely positive

Table 3: Fiscal policy, growth, and incumbent re-election (Table 2, Equations 6 and 7 from Brender and Drazen, 2008)

	Developed		Less Developed	
	β	s.e.	β	s.e.
Surplus/GDP, term	13.225*	7.938	13.483*	7.171
Surplus/GDP, election year	35.188***	11.037	1.210	9.814
Real GDP Growth Rate, term	-0.755	9.495	34.468***	8.354
New Democracy	1.266**	0.594	0.191	0.356
Majoritarian Election System	0.586	0.399	0.703*	0.372
Constant	-0.182	0.309	-1.739***	0.354

Probit model of whether incumbent party is re-elected to executive office ($=1$), election years between 1960-2003. Developed (OECD Countries): $N = 180$, pseudo- $R^2 = 0.071$. Less Developed (non-OECD Countries): $N = 167$, $R^2 = 0.112$.

results—to make decisions about substantive significance. They need only decide whether a 2.19% increase in re-election probability is large enough to justify the conclusion that there is a “political business cycle” of sorts—that is, that governments have incentives for austerity before an election.

By contrast, a 1% increase in the real GDP growth rate (holding all variables at their mean) is associated with a 0.296% decrease in probability of incumbent re-election in developed countries ($c^* = 0$, s.e.= 2.34%) and an 8.01% increase in probability of incumbent re-election in non-developed countries ($c^* = 6.80\%$, s.e.= 2.02%). In both of these cases, the decision is much more clear cut, as revealed by c^* : an analyst with $\gamma = 2$ can safely conclude that there is no substantively significant relationship between growth and incumbent re-election in developed countries, and that there is a very substantively significant and positive relationship between the two in non-developed countries.

Conclusion

We believe that most researchers agree that showcasing the substantive significance of a result—the degree to which it supports scientifically and politically relevant decisions, such as the decision to accept or reject a hypothesis—is at least as important as showcasing its statistical significance. Unlike statistical significance, however, most researchers have not adopted a common and consistent procedure for assessing substantive significance, probably because software tools for incorporating this assessment did not exist. Our tools provide transparent, consistent, yet flexible means to make these assessments.

Our procedure is rational and consistent because it is built on rational choice theory: it helps an analyst who specifies his/her preferences to make decisions (e.g., about theoretical hypotheses) that are always consonant with these preferences. It is transparent because it obligates a researcher to specify the standards by which they will judge empirical evidence (viz., the utility function). It is flexible because both default utility functions allow the choice of parameters that correspond to the researcher’s aversion to false positives (compared to false negatives), and because additional utility functions can be added with little reprogramming. Finally, it is convenient because the process is pre-programmed for use in a wide variety of models and can be implemented with only a few keystrokes. We hope that our software tools will allow researchers to place assessments of substantive significance at the core of their statistical analyses.

References

- Altman, Morris. 2004. “Statistical Significance, Path Dependency, and the Culture of Journal Publication.” *Journal of Socio-Economics* 33:651–556.
- Brambor, Thomas, William Clark and Matt Golder. 2006. “Understanding Interaction Models: Improving Empirical Analyses.” *Political Analysis* 14:63–82.

- Brender, Adi and Allan Drazen. 2008. "How Do Budget Deficits and Economic Growth Affect Reelection Prospects? Evidence from a Large Panel of Countries." *American Economic Review* 98:2203–2220.
- Cohen, Jacob. 1994. "The Earth is Round ($p < .05$)." *American Psychologist* 49:997–1003.
- DeGroot, Morris. 2004 {1970}. *Optimal Statistical Decisions*. Wiley Interscience.
- Esarey, Justin. 2010. "A Formal Test for Substantive Significance." Available on-line. URL: <http://userwww.service.emory.edu/~jesarey/riskstats.pdf>.
- Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3):647–674.
- Golder, Matt. 2006. "Presidential Coattails and Legislative Fragmentation." *American Journal of Political Science* 50:34–48.
- Harrison, Ann and Jason Scorse. 2010. "Multinationals and Sweatshop Activism." *American Economic Review* 100:247–273.
- Holt, Charles A. and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92:1644–1655.
- Horowitz, Joel. 2004. "Comments on 'Size Matters'." *Journal of Socio-Economics* 33:551–554.
- King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44:347–361.
- Lunt, Peter. 2004. "The Significance of the Significance Test Controversy: Comments on 'Size Matters'." *The Journal of Socio-Economics* 33:559–564.
- Manski, Charles F. 2007. *Identification for Prediction and Decision*. Harvard University Press.

- McCloskey, Deirdre and Stephen Ziliak. 1996. "The Standard Error of Regressions." *Journal of Economic Literature* 34:97–114.
- Pratt, John W., Howard Raiffa and Robert Schlaifer. 1996. *Introduction to Statistical Decision Theory*. MIT Press.
- Rabin, Matthew and Richard Thaler. 2001. "Anomalies: Risk Aversion." *Journal of Economic Perspectives* 15:219–232.
- Schmidt, Frank. 1996. "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers." *Psychological Methods* 1:115–129.
- Wald, Abraham. 1950. *Statistical Decision Functions*. Wiley.
- Zellner, Arnold. 2002. "Information Processing and Bayesian Analysis." *Journal of Econometrics* 107:41–50.
- Zellner, Arnold. 2004. "To Test or Not to Test and if So, How? Comments on 'Size Matters'." *Journal of Socio-Economics* 33:581–586.
- Zellner, Arnold. 2007. "Generalizing the Standard Product Rule of Probability Theory and Bayes's Theorem." *Journal of Econometrics* 138:14–23.
- Zellner, Arnold and V. Karuppan Chetty. 1965. "Prediction and Decision Problems in Regression Models from the Bayesian Point of View." *Journal of the American Statistical Association* 60:608–616.
- Ziliak, Stephen and Deirdre McCloskey. 2004. "Size Matters: The Standard Error of Regressions in the American Economic Review." *The Journal of Socio-Economics* 33:527–546.
- Ziliak, Stephen and Deirdre McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press.