6

Inference and Interpretation

After settling on a set of defensible model specifications we can undertake the process of interpretation. Conventional practice emphasizes statistical inference using null hypotheses, point estimates, and *p*-values for specific parameters. With the possible exception of simple linear models, such practice communicates relatively little. In more complicated, nonlinear models of the sort described in this book, basing inference on null hypothesis testing for particular parameters can be misleading. In this chapter, we focus on model interpretation rather than parameter inference. Model interpretation involves describing a model's implications – and our uncertainty about them – using directly interpretable and substantively important quantities of interest. We derive these implications by constructing plausible and substantively relevant scenarios and then using the model, along with likelihood theory, to generate simulations of the data-generating process.

6.1 THE MECHANICS OF INFERENCE

At a general level, *statistical inference* is the process of making knowledge claims based on the analysis of observed data. Part of the power of statistical inference is the ability to quantify, communicate, and interpret our uncertainty about these claims. To do this, many scholars examine the Big Ugly Table of Numbers (BUTON) produced by their statistical program. If the scholar's null hypothesis is that the parameter of interest, β_f, equals zero, and if the ratio of a point estimate to its standard error is greater than 1.96, then the scholar claims a "statistically significant" finding. That is, the scholar makes a knowledge claim in which the hypothesis that $\beta_f = 0.0$ has been rejected in favor of some other, such as $\beta_f > 0$. The scholar then proceeds to adorn his/her BUTON with ***.

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press

The claim of statistical significance is shorthand for saying "if $\beta_f = 0.0$, and we were to repeatedly generate new independent, random samples and calculate a $\hat{\beta}_f$ for each, then we should see $\hat{\beta}_f$ values at least as large as the one we just calculated less than 5% of the time." You can see why we have developed a shorthand phrase. Alternatively, the author might construct a 95% *confidence interval* around $\hat{\beta}_f$, which has a similar interpretation.

In case you were wondering … 6.1 Frequentist confidence intervals

Given observed data and a statistical model with parameter $\boldsymbol{\theta}$, we construct an estimate, $\hat{\boldsymbol{\theta}}$. The $100(1-\alpha)\%$ *confidence interval* (CI) around $\hat{\theta}_j$ is the set of values, C, such that, for any $x \in C$ the p-value of $|\hat{\theta}_j - x|$ is $\geq \alpha$.

The MLE is asymptotically normally distributed, so the asymptotic 95% confidence interval around the MLE is

$$\left\{x : \hat{\theta}_j - 1.96se(\hat{\theta}_j) \leq x \leq \hat{\theta}_j + 1.96se(\hat{\theta}_j)\right\}.$$

Confidence intervals are frequently misinterpreted. A 95% frequentist CI does *not* mean there is a 95% probability that true θ lies in the interval.[a]

[a] Bayesian credible regions, however, do have such an interpretation, although the Bayesian understanding of "probability" differs.

Importantly, however, this line of inference only works if there is a sense in which our data can be considered a random sample of cases. One way this might hold is when we have such a large population that we can actually randomly draw a set of cases to analyze. Another way to achieve this is to randomly assign cases different values of the covariate, X_f, so the observed data are but one realization of many possible random assignments.[1] However, many studies are observational, and the statistical inference framework developed for experiments and samples from larger populations is not entirely satisfactory.

An Example: World Trade and Democratization I

Ahlquist and Wibbels (2012) use world trade volumes as a tool for examining the relationship between income inequality and regime transitions. Rather than a random sample, Ahlquist and Wibbels gather data for all available country-years from 1875–2001, excluding the periods of World War I and II. Existing theoretical arguments implied that the relationship between trade and

[1] Permutation tests and randomization inference, discussed in Chapter 5, exploit this exact property.

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press

TABLE 6.1 *Probit regression of democratic transitions 1875–2001, a reestimation of Ahlquist and Wibbels (2012).*

	(1)	(2)
Labor endowment	−0.116	−0.500
	(0.175)	(0.175)
World trade	−0.020	−0.027
	(0.012)	(0.012)
Labor endowment × world trade		0.014
		(0.006)
Global % democracies	0.020	0.018
	(0.004)	(0.004)
Neighborhood % democracies	0.470	0.467
	(0.202)	(0.202)
Prior democratic failure	0.324	0.312
	(0.061)	(0.061)
Communist	−0.620	−0.616
	(0.227)	(0.227)
Gold Standard	0.171	0.105
	(0.146)	(0.146)
Interwar	−0.163	−0.241
	(0.235)	(0.235)
Post–Bretton Woods	0.394	0.419
	(0.216)	(0.216)
Neighborhood democratic transition	0.386	0.373
	(0.133)	(0.133)
n	8,347	8,347
$\log \mathcal{L}$	−621	−619
AIC	1,279	1,278
BIC	1,406	1,419

Note: Estimated model is a dynamic probit; interaction terms with lagged dependent variable omitted for simplicity, as is the constant term. Robust standard errors are reported, following Ahlquist and Wibbels (2012).

the transition from an autocratic to a democratic government should depend on a country's labor endowment, so Ahlquist and Wibbels fit a variety of probit models that include a multiplicative interaction term between labor endowment and trade. Table 6.1 is a streamlined presentation of their theoretically preferred model alongside a simpler alternative that omits the interaction term.[2]

[2] Ahlquist and Wibbels estimate a dynamic probit model that accounts for transitions out of democracy as well. We omit those terms from the table here for simplicity; they are readily available in the original paper or with the data and code accompanying this volume.

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press

Looking at in-sample performance using the AIC and BIC, they find that the models with and without the interaction term perform almost identically, providing little little reason to prefer the more-complicated version. Nevertheless, a standard form of inference might proceed to look at the point estimate and standard error of the endowment $\times$ trade term and notice that $0.014/0.006 \approx 2.3$, which implies a (two-sided) p-value of about 0.03. What knowledge claims are we to make with this result?

6.2 INTERPRETING MODELS AND PRODUCING SCENARIOS OF INTEREST

To make knowledge claims based on a model, we need to interpret the model in its domain of intended use. This means describing and evaluating the model's *implications*, not simply looking at stars next to particular parameters. Model implications, then, are statements about how we expect the outcome variable to behave under different conditions, along with statements of our uncertainty about those outcomes.

Our uncertainty comes from three sources. *Fundamental uncertainty* is the idea that the world, or at least our perceptions and measurements of it, have a fundamentally stochastic character. This is captured in the model's stochastic term, $Y \sim f$, in which we assume that the stochastic nature of our measurement and observation can be described using a particular probability model, such as the normal or Bernoulli distribution. *Estimation uncertainty* stems from the fact that we have limited data and, as such, our parameter estimates are subject to variability or revision upon observing more data. Estimation uncertainty in the likelihood framework is usually described using the large sample theory outlined in Chapter 2 and contained in the covariance matrix that we obtain when fitting models to actual data. Finally, there is *model uncertainty*, something we discussed in the previous chapter, when we emphasized that model evaluation and selection must occur prior to any interpretation.

6.2.1 Quantities of Interest

Quantity of interest is the generic name for the value we are going to use to interpret our model. As we learned in Chapter 3, the parameters of statistical models frequently fail to have direct and useful interpretations. It is the researcher's job to transform an estimated model into something informative to both ourselves and audiences. It is almost always easiest to understand a model when its implications are presented on the scale of the dependent variable. Predicted probabilities are more transparent than log-odds or odds ratios. Wages or incomes are easier to understand and communicate than elasticities.

In generating implications on the scale of the response, there are two primary quantities of interest: the *expected value* and the *predicted value*, where both are

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press

conditional on fixed covariate values and the estimated model parameters. The difference between the two is that the predicted value incorporates fundamental uncertainty, whereas the expected value does not.

Examples will help. Recall the linear regression of log CO_2 on log per capita GDP, reported in Table 1.2. In this model we estimated that

$$Y_i \sim f_{\mathcal{N}}(\mu_i, 2.112^2)$$
$$\mu_i = -0.08 + 1.04X_i.$$

Suppose we are interested in the model's predictions of India's CO_2 emissions if its per capita GDP were 10% larger than in 2012, i.e., if $x_i = \log(1.1 * 4{,}921.84)$. The expected amount of CO_2 is then $\exp(1.04*\log(1.1*4{,}921.84) - 0.08) = \exp(8.86) = 6{,}915.3$kT. The predicted amount, however, will take account of the fundamental uncertainty, as represented by the normal distribution with variance estimated at 2.112^2. To do this, we can take a draw from $\mathcal{N}(8.86, 2.112^2)$. One such draw is 6.71; $\exp(6.71) = 821$, which represents one predicted value.

As a second example, recall the exercise in Chapter 3 in which we estimated the predicted probability of a woman being in the paid labor force when she had a college degree, holding the other covariates at their central tendencies. Based on a logistic regression model we calculated this expected value as $\text{logit}^{-1}\left(\bar{\mathbf{x}}_{\neg\text{coll}}\hat{\boldsymbol{\beta}}\right)\Big|_{\text{coll}=1} = 0.75$. A predicted value in this case must be either 0 or 1. To generate a predicted value that accounts for fundamental uncertainty, we would sample from a Bernoulli distribution with $\theta = 0.75$. Or flip an appropriately weighted coin.

Exploring how expected values change under different values of covariates helps interpret the systematic component of the model we fit. This is frequently of primary interest. And, in general, we will not generate a single predicted value; we will take advantage of our computing power and generate many, the average of which will necessarily converge on the expected value. Nevertheless, the predicted values are useful for understanding how much fundamental uncertainty remains after we have fit our model. Does fundamental uncertainty still swamp whatever systematic relationship we have? If so, this is usually reason for both some humility in our claims and a need for more research, including gathering more data.

6.2.2 Scenarios

In calculating our expected and predicted quantities, we must choose specific values for the covariates in the model. Since our model posits a description of the data-generating process, these vectors of covariate values represent scenarios, or possible situations that we are interested in. These scenarios are sometimes referred to as *counterfactuals*, since they provide an answer to the question: "What would we expect to happen if an independent variable took

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press

on a value different from what we observe in the data?" Counterfactuals are ultimately at the heart of causal inference in randomized experiments in which we interpret the (average) effect of the treatment as what we would expect to happen to the control group were we to administer the treatment to them.

In the context of model interpretaton, the question remains, "What is a good scenario?" Good and useful scenarios should:

- Address a concrete and important substantive issue that motivated the research. If we are interested in the effect of natural disasters on support for democracy in poor countries versus rich ones, then we should construct a scenario that holds national income at a low level and at a high level, not necessarily near the mean of the data set.
- Be compared to other scenarios. Construct several scenarios of theoretical or policy relevance to provide a picture of the model's implications in its intended domain of application. One scenario is almost never enough.
- Be simple enough for meaningful comparisons to be made. This usually means only varying one thing at a time.
- Include descriptions of our uncertainty about the model's implications. Remember that we have estimation uncertainty (the variance-covariance matrix) and fundamental uncertainty as reflected in the stochastic component of our models. Both of these are directly accessible from the model in our computers. Uncertainty over the models themselves is the most difficult to quantify and express.
- Keep the scenarios "close to" our experience (more later in this chapter). A useful baseline typically involves estimating what happens near the center of our data, i.e., holding variables at their central tendencies.
- Respect the structure of the data. If a variable of interest is only observed or available in discrete chunks, then the scenarios of interest should reflect this. For example, in Figure 3.7 the scenarios we report reflect the fact that senators are up for election every other year, inducing inherent discreteness in the time-to-election variable.
- Avoid extreme changes – unless that's what we're interested in. Unless we are toggling indicator variables, it usually does not make much sense to take the variable of interest from its minimum value to its maximum value. For example, if we're interested in the relationship between income per capita on regime transitions, using the GDP per capita of the Democratic Republic of Congo and Luxembourg is unlikely to provide much insight for a world in which income per capita changes relatively smoothly and slowly.

6.2.3 Interpolation, Extrapolation, and Convex Hulls

Some of the advice above can be restated in a more formal fashion. King and Zeng (2006a,b) make the important point that the further our counterfactual

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press

scenarios are from the support of the data used to estimate the model, the more *model dependent* the implied predictions become. As our modeling assumptions increasingly determine our predictions, our uncertainty about these "extreme counterfactuals" is understated since the calculated confidence bands fail to incorporate our uncertainty about the model itself. As model uncertainty comes to dominate, our inferences about model predictions become increasingly tenuous.

The difference between interpolation and extrapolation provides a relatively intuitive criterion for whether our counterfactuals deviate too far from our experience. To define these terms we must first understand the *convex hull* of our data. The convex hull is the smallest convex set that can contain all the data points. The easiest way to get a feel for the hull is to visualize it, as in Figure 6.1. Statisticians have long defined *interpolation* as a way of constructing new data points that are inside the convex hull of the data, wheras *extrapolation* involves going outside the hull and necessarily further from our recorded experience.

Once we move out of two dimensions, calculating and visualizing convex hulls becomes difficult, but software exists to assist us.

Our perspective is that predictions involving scenarios within the convex hull of the data, i.e., interpolation, are better for describing a model's implications. But pressing policy problems may require that we pose questions beyond the realm of our recorded experience, i.e., our data. Should we disregard the data we do have and the models we can fit in these situations? Probably not. Nevertheless, an important first step in making such

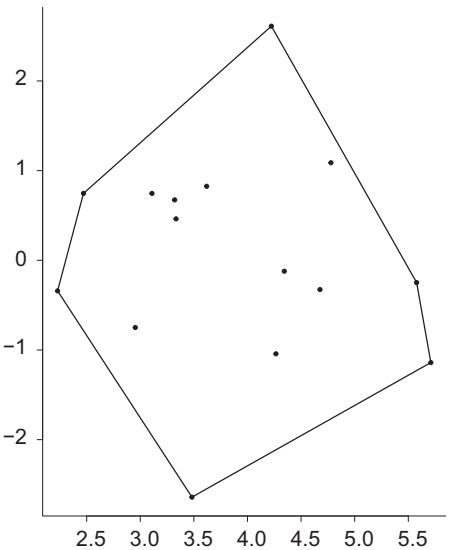

FIGURE 6.1 Visualizing the convex hull in 2 dimensions.

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press

predictions is recognizing that the scenario of interest is in fact quite far from our observed experience and, as such, predictions are more likely to depend on modeling decisions. A second step, then, would be to conduct *sensitivity analysis*, generating predictions from a variety of models that perform well in the observed data set to produce some sense of the prediction uncertainty.

An Example: World Trade and Democratization II

Ahlquist and Wibbels chose to interpret their model by comparing a labor-scarce autocracy to a labor-abundant one at different levels of world trade. To set appropriate values, they chose the labor endowment values for Argentina (scarce) and China (abundant) in 1980. They hold other covariates at their mean or modal values. They then consider 150 different world trade values within the historical interquartile range, 15%–30% of world GDP.

We compare Ahlquist and Wibbels' counterfactual scenarios to their data to see which are similar to observed experience and which are outside the convex hull. Summary results appear in Table 6.2. In both the labor-scarce and the labor-abundant examples, all 150 scenarios fall inside the convex hull. We also see that the labor-scarce scenarios based on Argentina's labor endowment are, on average, closer to the observed data than the labor-abundant scenarios based on China. This is unsurprising since China's labor endowment value in 1980 of 4.2 put it at about the 87th percentile of all observed values in the data set, closer to the hull. Argentina's value of 0.5 was almost exactly the 25th percentile.

6.2.4 Simulating Estimation Uncertainty

Interpreting models in terms of scenarios and expected (or predicted) values requires that we incorporate our estimation uncertainty. Sometimes these values can be derived analytically, but there is no need to restrict either our modeling decisions or our interpretation to simple or easy-to-calculate objects. With existing computing power we can easily generate many samples of our model

TABLE 6.2 *Model dependence and the Ahlquist and Wibbels interpretation scenarios.*

	Labor Scarce	Labor Abundant
Labor endowment value	0.5	4.2
Number of scenarios	150	150
% scenarios in convex hull	100	100
Avg. % observed data "nearby"	29	18

Note: Data points "nearby" are within one geometric variance of the scenario point. The distance measure is Gower's G^2.

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press

parameters from distributions reflecting our estimation uncertainty. We can then combine each of these samples with our scenarios and models to build up distributions for any quantity we may be interested in.

In regular likelihood problems we can use the theory from Chapter 2 to sample from the multivariate normal distribution, just as we did in Chapter 3. To review, we do the following:

1. Fit the model of interest and store the estimated parameters, $\hat{\boldsymbol{\theta}}$, along with the covariance matrix, $\text{cov}(\hat{\boldsymbol{\theta}})$;
2. Construct a scenario embodied in the vector of covariates, $\mathbf{x}_s$;
3. Draw a parameter vector, $\tilde{\boldsymbol{\theta}}$, from $\mathcal{N}\left(\hat{\boldsymbol{\theta}}, \text{cov}(\hat{\boldsymbol{\theta}})\right)$;
4. Calculate the quantity of interest based on $\mathbf{x}_s$ and $\tilde{\boldsymbol{\theta}}$. In the models explored in this book this means calculating the linear predictor, $\mathbf{x}_s^{\mathsf{T}}\hat{\boldsymbol{\beta}}$, and then mapping that quantity into $\text{E}[Y]$. In the logit case this was $\text{logit}^{-1}(\mathbf{x}_s^{\mathsf{T}}\hat{\boldsymbol{\beta}})$;
5. Repeat steps 3–4 many times to build up the distribution for the quantity of interest under this scenario;
6. Calculate important quantiles (e.g., 0.025 and 0.975) from this distribution or display the distribution;
7. Return to step 2 and construct a different scenario;
8. Once finished, we can display the model implications, along with the estimation uncertainty around them.

In nonstandard or more customized applications when large-sample likelihood theory does not apply, we can apply the logic of bootstrap resampling to repeatedly resample from our existing data to fit models and calculate quantities of interest.

Interpreting Interactive and Conditional Relationships

Much has been written on the use and interpretation of models containing multiplicative (or other nonlinear) terms. This literature rightly points out that direct examination of coefficient point estimates is insufficient for interpreting conditional relationships. The basis for this discussion is couched in terms of marginal effect calculations and standard errors in linear regression. If $\text{E}[Y_i] = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$, then

$$\frac{\partial \text{E}[Y_i]}{\partial x_{i1}} = \beta_1 + \beta_3 x_{i2}.$$

It follows that

$$\begin{aligned}\text{var}\left(\frac{\partial \text{E}[Y_i]}{\partial x_{i1}}\right) &= \text{var}\left(\beta_1 + \beta_3 x_{i2}\right) \\ &= \text{var}\left(\beta_1\right) + x_{i2}^2 \text{var}\left(\beta_3\right) + 2x_{i2}\text{cov}\left(\beta_1, \beta_3\right),\end{aligned}$$

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press

from which we can easily calculate the standard error of the marginal effect. Several observations emerge from these expressions:

- Both the relationship between X_1 and Y and our uncertainty about it depends on X_2. It therefore makes no sense to declare the conditional relationship to be "significant," since our confidence intervals may contain 0 at some values of X_2 but not others.
- It often makes no sense to talk about "main effects" and "interaction effects." There is simply the marginal or predicted effect; any instance in which $x_{i2} = 0$ is simply a special case. The one exception here is when X_1 or X_2 are binary variables encoded as 0 and 1.
- There is almost never an instance in which it makes sense to include an interactive term without the constitutive terms.

These calculations for the standard error of the marginal effect only apply to linear regression. But the observations echo points raised in Chapter 3 about marginal effects in a logit context. This means that our strategy of specifying scenarios of interest, simulating from the model's sampling distribution, and then constructing displays of the model's predicted consequences encompasses the interpretation of interactive, conditional, and other relationships that are nonlinear in the covariates. We do not need to derive marginal effects (and variance) expressions for each specific class of models; we can simulate them directly. But our discussion of model validation and selection adds another important observation: before engaging in model interpretation and hypothesis testing, analysts proposing models with conditional, nonlinear, or other complicated functional relationships have a special burden to show that their more-complicated models out-perform their model's simpler cousins.

An Example: World Trade and Democratization III

Ahlquist and Wibbels (2012) estimated models with a "significant" interaction term between world trade and labor endowment. To interpret this model they constructed a series of counterfactual scenarios, allowing us to compare the model's predicted probability of democratic transition as a function of world trade in labor-abundant and labor-scarce autocracies. In these scenarios, they varied world trade across its interquartile range. "Labor scarce" was defined as the value observed in Argentina in 1980 while "labor abundant" was the value observed in China in the same year. All these scenarios were found to be in the convex hull of the data.

Ahlquist and Wibbels then proceed to calculate predicted probabilities – expected values – for each of these 300 scenarios and simulate from the sampling distribution to describe their estimation uncertainty. Figure 6.2 reproduces their key interpretive results, with the solid line representing the

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press

labor-scarce and the dashed line depicting the labor-abundant autocracy. The shaded regions represent the 95% confidence intervals around the predictions. If the data were consistent with the literature's predictions, the plot would look like an "X," with the risk of transition increasing in world trade for the labor-abundant scenario, and the reverse for the labor-scarce. While the risk of a transition decreases for labor-scarce autocracies as trade increases, there is no discernible relationship between trade and democratization in labor-abundant autocracies. The confidence intervals overlap substantially for all values of trade, making it difficult to sustain the claim that there is an important conditional relationship at work here, notwithstanding the small *p*-value on the interaction term. Had the authors simply declared victory based on theory and a small *p*-value on an interaction term, erroneous inference could have occurred.

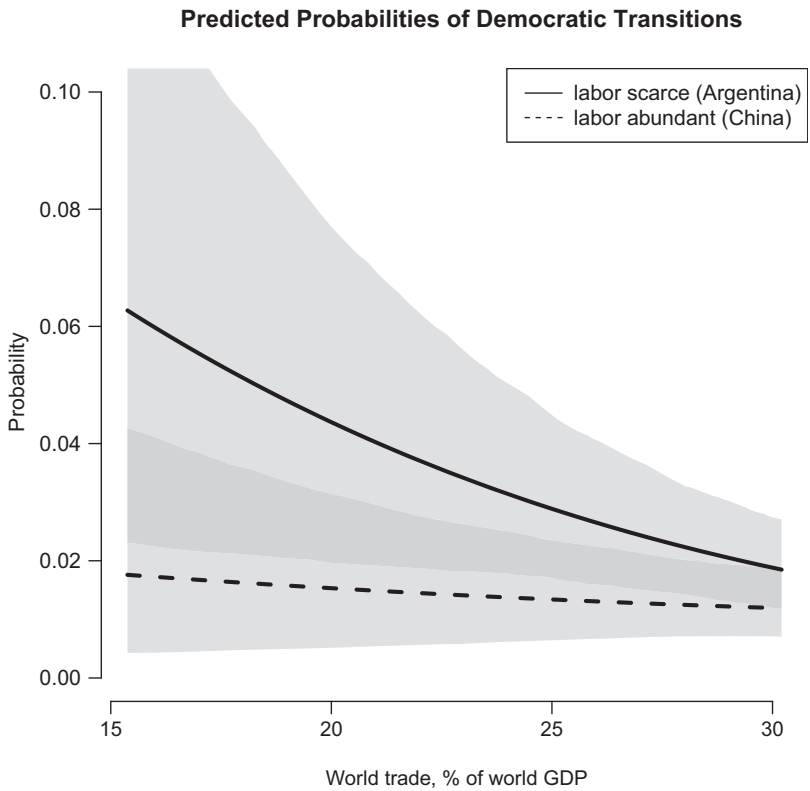

FIGURE 6.2 The relationship between world trade and the probability of a democratic transition in labor-abundant and labor-scarce autocracies. Reproduced from Ahlquist and Wibbels (2012), figure 4.

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press

6.3 CONCLUSION

Conventional statistical inference focuses on deciding whether particular model parameters are sufficiently far away from an arbitrary "null" value, based on estimation uncertainty and a particular model of data sampling. While useful, this mode of reasoning can be misleading or miss the point entirely, especially when the data do not conform to an actual sample from a population or a researcher-controlled experiment. Moreover, in many models the parameters do not have ready interpretations applicable to the substantive question at hand, meaning statements of statistical significance, as the primary mode of inference, do not have immediate meaning to audiences.

A more readily accessible form of inference explores a model's implications by using substantively relevant scenarios to compare predicted outcomes under different conditions. Model interpretation is almost always most effective when communicated on the scale of the dependent variable, whether as an expected or predicted value. By incorporating (and displaying) estimation uncertainty around these implications we can develop a richer understanding of both the size of a model's predicted "effects" as well as the level of uncertainty around them.

6.4 FURTHER READING

Applications

Shih et al. (2012) provide an excellent interpretation and presentation of conditional effects from interaction terms in a custom-designed model for rank data. They explicitly consider the convex hull in their interpretive scenarios.

Previous Work

King et al. (2000) discuss the construction and display of meaningful quantities of interest for model interpretation.

Ai and Norton (2003); Berry et al. (2010, 2012); Brambor et al. (2006); Braumoeller (2004); and Kam and Franzese (2007) all discuss the inclusion and interpretation of multiplicative interaction terms in the linear predictor.

Advanced Study

Cox (2006) provides a detailed discussion of the theory, practice, and history of statistical inference.

Software Notes

Tomz et al. (2001) developed the stand-alone CLARIFY package for calculating quantities of interest and estimation uncertainty around them. This functionality has been incorporated into $\mathcal{R}$'s `Zelig` library.

There are several $\mathcal{R}$ packages for calculating convex hulls. The basic `chull` function only handles two-dimensional data. Stoll et al. (2014) developed `WhatIf` that allows us to compare proposed scenarios to the convex hull of our data. The `geometry` package (Habel et al., 2015) also contains functionality for calculating the convex hull in more than two dimensions.

https://doi.org/10.1017/9781316888544.009 Published online by Cambridge University Press