Reasonable Measures of Uncertainty Under Separation*

Carlisle Rainey†

ABSTRACT

When facing data sets with small numbers of observations or "rare events," political scientists often encounter important explanatory variables that perfectly predict binary events or non-events. In this situation, maximum likelihood provides implausible estimates and the researcher must incorporate some form of prior information in the estimation. The most sophisticated research uses Jeffreys' invariant prior to stabilize the estimates. While Jeffreys' prior has the advantage of being automatic, I show that, in many cases, it offers too much prior information, providing confidence intervals that are much too narrow. I show that the choice of a more reasonable prior can lead to different substantive conclusions about the likely magnitude of an effect and I offer practice advice for choosing a prior distribution that represents actual prior information.

*I thank [many people]. Thanks to Mark Bell and Nicholas Miller for making their data available to me. The analyses presented here were conducted with R 3.1.0 and JAGS 3.3.0. All data and computer code necessary for replication are available at github.com/carlislerainey/priors-for-separation.

†Carlisle Rainey is Assistant Professor of Political Science, University at Buffalo, SUNY, 520 Park Hall, Buffalo, NY 14260 (rcrainey@buffalo.edu).

The Logistic Regression Model

Political scientists commonly use logistic regression to model the probability of an event of interest. In the typical situation, the researcher uses an $n \times k + 1$ design matrix X consisting of an intercept and k covariates to model a vector n of binary outcomes y, where $y_i \in \{0, 1\}$ using the model $\Pr(y_i) = \Pr(y_i = 1|X) = \frac{1}{1 + e^{-X_i \beta}}$, where β is a parameter vector of length $k + 1$.

Using this model, it is straightforward to calculate the likelihood function

$$\Pr(y|\beta) = L(\beta|y) = \prod_{i=1}^{n} \left[\left(\frac{1}{1 + e^{-X_i \beta}} \right)^{y_i} + \left(\frac{1}{1 + e^{-X_i \beta}} \right)^{1-y_i} \right].$$

As usual, one can take the natural logarithm of both sides to calculate the log-likelihood function

$$\log L(\beta|y) = \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1 + e^{-X_i \beta}} \right) + (1 - y_i) \log \left(\frac{1}{1 + e^{-X_i \beta}} \right) \right]$$

and take the derivatives of the log-likelihood function to obtain the score function

$$\frac{\partial \log L(\beta|y)}{\partial \beta} = \sum_{i=1}^{n} \left(y_i - \frac{1}{1 + e^{-X_i \beta}} \right) X_i.$$

Researchers routinely obtain estimates $\hat{\beta}$ of the model parameters β by setting the score function equal to zero and solving for β (i.e., maximizing the likelihood of the observed data) and estimate the standard errors are by calculating the square root of the diagonal of the inverse of Fisher's information matrix evaluated at $\hat{\beta}$ (i.e., calculate the curvature around the maximum of the likelihood function to obtain an estimate of the uncertainty of the estimate). While this approach works quite well in most applications, it fails in a situation known as separation (Zorn 2005).

Separation

Separation occurs in models of binary outcome data when one explanatory variable (or perhaps a combination of explanatory variables, see **?**) perfectly predicts zeros, ones, or both. *Complete separation* occurs when the "problematic" explanatory variable s (for *s*eparating explanatory variable) perfectly predicts both zeros and ones and *quasicomplete separation* occurs when s perfectly predicts either zeros or ones, but not both (Albert and Anderson 1984; Zorn 2005). *Overlap*, the ideal case, occurs when no explanatory variable (or combination) perfectly predicts zeros or ones. In this situation, the usual maximum likelihood estimates exist and provide reasonable estimates of parameters. However, under complete or quasicomplete separation, maximum likelihood estimates do not exist (Albert and Anderson 1984; Zorn 2005).

Complete separation occurs when a covariate perfectly predicts zeros and ones. For example, suppose an explanatory variable s , such that $y = 0$ for $s > 0.5$ and $y = 0$ for $s \leq 0.5$. This corresponds to the middle panel of Figure 1. To maximize the likelihood of the observed data, the "S"-shaped logistic regression curve must assign probabilities of zero when $s \leq 0.5$ and probabilities of one when $s > 0.5$. Since the logistic regression curve lies strictly between zero and one, this likelihood cannot be achieved. However, it can be approached asymptotically as the coefficient for s approaches infinity. Thus, the likelihood function under complete separation is monotonic (has no maximum) and a finite maximum likelihood estimate does not exist.

Quasicomplete separation occurs when a covariate perfectly predicts zeros *or* ones. Figure 1 shows an example pattern in the right panel, where y always equals zero when s equals zero. This situation occurs often in applied political science research with binary inputs. For example, Gelman et al. (2008) find no African-American respondents in their data support Barry Goldwater in 1964, leading to a maximum likelihood estimate of negative infinity for the coefficients for the indicator of African-American respondents. Similarly, **?** (see **?**) finds no instances of states with nuclear weapons engaging in war with each other. In this case, the estimated coefficient for the variable indicating symmetric nuclear dyads (in which both states possess nuclear weapons) equals negative infinity. To maximize the likelihood in this situations, the model must assign probabilities of zero to observations for which $s = 1$ (African-American respondents or symmetric nuclear dyads, in these examples). Again, because the logistic regression curve lies strictly above zero, this cannot happen, though it can be approached asymptotically as the coefficient for s goes to negative infinity. As with complete separation, the likelihood function under quasicomplete separation is monotonic (has no maximum) and a finite maximum likelihood estimate does not exist.

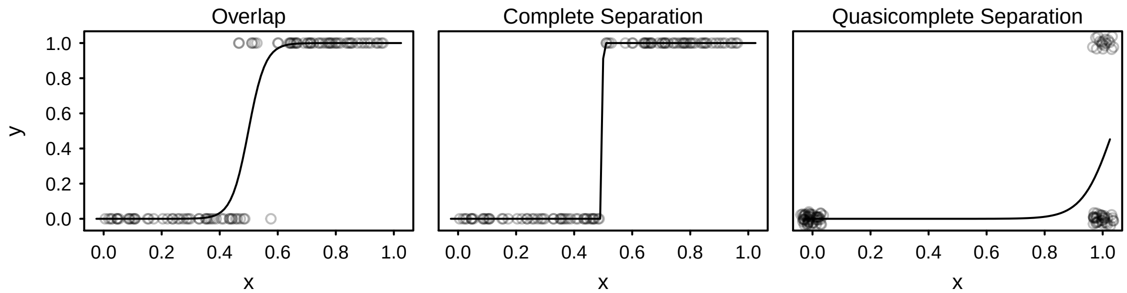

FIGURE 1: This figure illustrates overlap, complete separation, and quasicomplete separation as defined by Albert and Anderson (1984). The maximum likelihood estimates only exist under overlap. Under complete and quasicomplete separation, maximum likelihood fails, returning infinite estimates and standard errors.

Along coefficient estimates under separation are infinite in theory, the hill-climbing algorithms approximate the infinite estimates with large, finite values that increase with the precision of the algorithm. Table 1 shows the estimates from R's `glm()` function for each of the hypothetical data sets in Figure 1. To illustrate the problem, I vary the convergence tolerance between $\epsilon = 10^{-8}$ (the default) and $\epsilon = 10^{-16}$. With the default tolerance, `glm()` returns very large estimates and standard errors. There is no finite maximum, but the likelihood is flat enough around the large estimates to satisfy the algorithm, given the default tolerance. With the more sensitive tolerance of $\epsilon = 10^{-16}$, the algorithm converges even closer to infinity, but still returns finite estimates. This is a failure of maximum likelihood, because estimates of infinity are usually implausible (i.e., there is always some, though perhaps tiny, probability of a zero or one in any given situation, see Heinze and Schemper 2002).

	Overlap		Complete Separation		Quasicomplete Separation	
	$\epsilon = 10^{-8}$	$\epsilon = 10^{-16}$	$\epsilon = 10^{-8}$	$\epsilon = 10^{-16}$	$\epsilon = 10^{-8}$	$\epsilon = 10^{-16}$
constant	-17.05	-17.05	-739.27	-739.27	-19.57	-26.57
	(5.79)	(5.79)	(139744.58)	(139744.58)	(1520.85)	(50363.70)
x	34.19	34.19	1483.21	1483.21	18.90	25.90
	(11.92)	(11.92)	(280801.40)	(280801.40)	(1520.85)	(50363.70)
Log Likelihood	-9.74	-9.74	0.00	0.00	-32.05	-32.05
Num. obs.	100	100	100	100	100	100

Standard errors in parentheses.

TABLE 1: A table providing estimates based on the data shown in Figure 1 using the R function `glm()` varying the convergence tolerance under overlap, complete separation, and quasicomplete separation. Notice that the estimation algorithm returns estimates and standard errors arbitrarily close to infinity under both types of separation as the tolerances shrinks to zero.

Perhaps more starkly, notice that the strong pattern in the middle panel of Figure 1 does not produce statistically significant result. This is because the likelihood is essentially flat around the "maximum" found by the hill-climbing algorithm. As the region around the maximum flattens, the estimates of the standard errors increases. Thus, separation leads to implausible large estimates *and* standard errors. Notice, for example, that while the data in the middle and right panel would almost never occur under the null hypothesis of no relationship, none of the estimates are statistically significant.

The Importance of the Prior

Choosing a reasonable prior distribution is crucial for dealing with separation in a substantively meaningful manner. In many cases, the data (though the likelihood) swamp the contribution of the prior. However, in the case of separation such that s_i perfectly predicts events, the likelihood determines the shape of the left-hand side of the posterior distribution and the prior (symmetric about zero) determines the shape of the right hand side of the posterior.

The likelihood has an “S”-shape that approaches a limit of one as the parameter β_s for the separating variable s approaches infinity. Thus, for large values of β_s, the likelihood is essentially flat, which allows the prior distribution to drive the inferences. Thus the prior distribution is not an arbitrary choice made for computational convenience–but a choice that affects the inferences.

The Impact of the Prior in Theory

Suppose that an explanatory variable s_i perfectly predicts a binary outcome variable $y_i = 1$, such that whenever $s_i = 1$, $y_i = 1$, but when $s_i = 0$, y_i might equal zero or one. Albert and Anderson (1984) refer to this situation as quasicomplete separation. Suppose further an additional set of covariates X_i and the analyst wishes to obtain plausible estimates of coefficients the model $Pr(y_i = 1) = \text{logit}^{-1}(\alpha + \beta s_i + X_i \gamma)$. It is easy to find plausible estimates of γ using the techniques discussed above (even maximum likelihood usually provides reasonable estimates of these parameters), but finding plausible estimates of α and β proves more difficult because maximum likelihood suggests estimates of $-\infty$ and $+\infty$, respectively. In order to obtain a plausible estimate of β (which will, in turn, provide a plausible estimate of α), the researcher must introduce prior information into the model. My purpose here is to characterize how this prior information impacts the posterior distribution.

In the general situation, the analyst is interested in computing and characterizing the posterior distribution of the coefficient for s_i given the data. Using Bayes’ Rule, this posterior depends on the likelihood and the prior, so that $p(\beta|y) = p(y|\beta)p(\beta)$. In particular, the analyst might have in mind a family of priors centered at and monotonically decreasing away from zero with varying scale σ, so that $p(\beta) = p(\beta|\sigma)$. Suppose that for a particular $\beta^* \geq 0$ the prior distribution is decreasing in β at a decreasing rate. Intuitively, this assumption of a β^* allows the result to generalize to many common distributions.[1] Finally, suppose that the informativeness of the prior distribution depends on scale parameter σ that is chosen by the researcher and "flattens" the prior $p(\beta) = p(\beta|\sigma)$, such that as σ increases, the rate a which the prior descends to zero decreases.

Theorem 1. *The impact of the researchers choice of σ on the posterior distribution $p(\beta|y)$ is increasing in β for $\beta > \beta^*$.*

In many cases, researchers summarize the posterior distribution by providing the 5th and 95th percentiles and a measure of centrality, such as the median.

> PRACTICAL IMPLICATION OF THEOREM 1: Under quasicomplete separation where x_i perfectly predicts $y_i = 1$, the prior has a small impact on the lower bound of the 90% credible interval, a moderate impact on the measures of the location of the posterior (i.e., mean, median, and mode), and a large impact on the upper-bound of the credible interval.

The Impact of the Prior in Practice

To illustrate the impact of the prior on inferences when facing separation, I replicate a results from Barrilleaux and Rainey (2014), who are interesting in the effect of partisanship on governors' decisions to oppose the Medicaid expansion in their states under the Patient Protection and Affordable Care Act (ACA).[2] As the authors note, no Democrats opposed the expansion

[1] In particular, if the prior distribution is in the form of a double-exponential, which lacks "shoulders," then $\beta^* = 0$. However, the most common prior distributions used in applied work, such as the normal, t, and Jeffreys', have "shoulders" such that $\beta^* > 0$. In this case, the exact curvature of the distribution in the region $[0, \beta^*]$ affects the relative impact of the prior.

[2] Barrilleaux and Rainey (2014) use a logistic regression modeling the probability that a governor opposes the expansion using the following explanatory variables: the partisanship of the governor, the percent of the state's residents who are favorable toward the ACA, whether Republicans control the state legislature, the percent of the state that is uninsured, a measure of the fiscal health of the state, the Medicaid multiplier for the state, the percent of the state that is nonwhite, and the percent of the state that resides in a metropolitan area. See their paper for more details.

leading to a problem of separation. I use MCMC to simulate from the posterior using several different prior distributions, including Jeffreys' prior (Zorn 2005) and the Cauchy prior with scales of 1, 2.5, and 5 (Gelman et al. 2008). While the choice of prior does not affect the conclusion about the *direction* of the effect, it has a large impact on the conclusion about the *magnitude* of the effect. This can be especially important when researchers are making claims about the substantive importance of their estimated effects (see Rainey 2014, Gross 2014, and McCaskey and Rainey 2014).

Figure 2 shows the posterior distribution for the coefficient for the indicator of Republican governors. Notice that the different priors lead to different posterior distributions. Notice, in particular, that the choice of the prior has a large impact on the right-hand side of the posterior. More informative priors (e.g., Jeffrey's prior) lead to a more peaked posterior distribution that rules out very large effects. Less informative priors (e.g., Cauchy(2.5)) lead to the conclusion that even large effects are plausible. These differences affect the conclusions that the researchers draw about the likely magnitude of the effect.

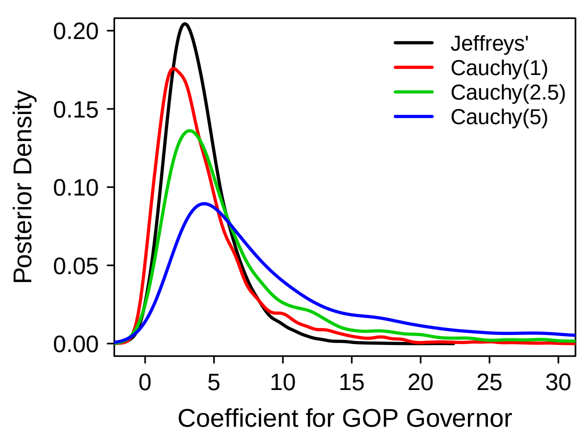

FIGURE 2: This figure provides the posterior distribution for the coefficient of the indicator for GOP governors in the model offered by Barrilleaux and Rainey (2014). Notice that the location and the spread of the posterior depend on the prior chosen, especially the right-hand side of the distribution.

Figure 3 shows how the choice of prior impacts the 90% credible interval. Notice that different prior distributions lead to different conclusions about the plausible values of the effect. In particular, different priors lead to different conclusions about the upper-bound on the plausible effect sizes. For example, Jeffreys' prior, the default proposed by Zorn (2005) and Heinze and Schemper (2002), suggests the effect lies in the range $\beta_{\text{GOP Gov.}} \in [0.9, 8.4]$, with a posterior mean of 3.9. On the other hand, the less informative Cauchy(2.5) prior, the default proposed by Gelman et al. (2008), suggests the effect lies in the range $\beta_{\text{GOP Gov.}} \in [1.0, 22.5]$, with a posterior mean of 7.3. A simple change from one proposed default to another more than doubles the upper bound on the 90% credible interval and almost doubles the posterior mean. Further, the Cauchy(5) prior, a plausible prior if one believes the effect might be large, produces the upper-bound on the 90% credible interval from is more than four times larger than the upper-bound produced by Jeffrey's prior. The posterior mean from the Cauchy(5) prior is larger falls above the upper-bound from the 90% credible interval from Jeffrey's prior.

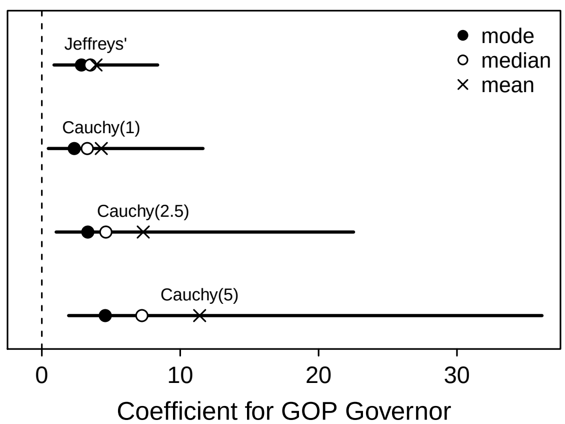

FIGURE 3: This figure provides the (equal-tailed) 90% credible intervals for the coefficient of the indicator for GOP governors in the model offered by Barrilleaux and Rainey (2014). Notice that the location and the spread of the posterior depend on the prior chosen, especially the right-hand side of the distribution. Note that Jeffrey's prior, suggested by Zorn (2005), is the most informative of these priors, suggesting that a coefficient smaller than about 10 is quite unlikely. On the other hand, credible interval using the Cauchy(2.5) prior, as suggested by Gelman et al. (2008), is about *twice* as wide as the credible interval from Jeffreys' prior. Finally, notice that the Cauchy(5) prior–a plausible prior if the researcher believes the effect might be large–produces a posterior mean larger than the upper bound of the 90% credible interval using Jeffrey's prior.

This leads us to conclude that the choice of prior matters–it affects the inferences that we draw from the data. It is not sufficient to rely on the prior distribution designed as a default for other purposes. Instead, we must rely on prior distributions that represent actual prior information about the likely magnitude of the coefficients.

References

Albert, A., and J. A. Anderson. 1984. "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika* 71(1):1–10.

Barrilleaux, Charles, and Carlisle Rainey. 2014. "The Politics of Need: Examining Governors' Decisions to Oppose the 'Obamacare' Medicaid Expansion." Forthcoming in *State Politics and Policy Quarterly*. Manuscript at http://www.carlislerainey.com/files/need.pdf.

Bell, Mark S., and Nicholas L. Miller. 2014. "Questioning the Effect of Nuclear Weapons on Conflict." *Journal of Conflict Resolution* .

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics* 2(4):1360–1383.

Gross, Justin H. 2014. "Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance." Forthcoming in *American Journal of Political Science*.

Heinze, Georg, and Michael Schemper. 2002. "A Solution to the Problem of Separation in Logistic Regression." *Statistics in Medicine* 21(16):2409–2419.

Lesaffre, E., and A. Albert. 1989. "Partial Separation in Logistic Discrimination." *Journal of the Royal Statistical Society. Series B (Methodological)*. 51(1):109–116.

McCaskey, Kelly, and Carlisle Rainey. 2014. "Meaningful Inferences: The Importance of Explicit Statistical Arguments for Substantive Significance." Working paper. Latest version at https://github.com/carlislerainey/meaningful-inferences.

Rainey, Carlisle. 2014. "Aruging for a Negligible Effect." Forthcoming in *American Journal of Political Science*.

Rauchhaus, Robert. 2009. "Evaluating the Nuclear Peace Hypothesis: A Quantitative Approach." *Journal of Conflict Resolution* 53(2):258–278.

Zorn, Christopher. 2005. "A Solution to Separation in Binary Response Models." *Political Analysis* 13(2):157–170.

Online Appendix

Proof of Theorem 1

Assumption 1 (Separation). *Suppose quasicomplete separation such that s_i perfectly predicts $y_i = 1$.*

Assumption 2 (Prior Shape). *Suppose that the researcher computes the posterior distribution $p(\beta|y) = p(y|\beta)p(\beta)$ such that for a particular $\beta^* \geq 0$ the prior distribution is decreasing at a decreasing rate.*

Intuitively, this assumption of a β^* allows the result to generalize to a range of common distributions. In particular, if the prior distribution is in the form of a double-exponential, which lacks "shoulders," then $\beta^* = 0$. However, the most common prior distributions used in applied work, such as the normal, t, and Jeffreys', have "shoulders" such that $\beta^* > 0$. In this case, the exact curvature of the distribution in the region $[0, \beta^*]$ affects the relative impact of the prior.

Assumption 3 (Scale Parameter). *Suppose finally the that informativeness of the prior distribution depends on scale parameter σ "flattens" the prior $p(\beta) = p(\beta|\sigma)$, such that as σ increases, the rate a which the prior descends to zero decreases.*

σ is chosen by the researcher based on prior information about the likely values of the coefficients.

Before proving Theorem 1, it is helpful to show several initial results.

Lemma 1. $\dfrac{\partial p(y|\beta)}{\partial \beta} > 0$ *for all* β.

Proof of Lemma 1. The quantity $p(y|\beta)$ is the probability of observing y (i.e., an outcome variable separated by s). Increasing values of β make this separation increasingly likely. Thus, $p(y|\beta)$ is increasing in β so that $\dfrac{\partial p(y|\beta)}{\partial \beta} > 0$. □

Lemma 2. *$p(\beta|\sigma) > 0$ for all β.*

Proof of Lemma 2. The quantity $p(\beta|\sigma)$ is a probability distribution defined to have support over the real line and thus $p(\beta|\sigma) > 0$ for all β. □

Lemma 3. *$p(y|\beta) > 0$ for all β.*

Proof of Lemma 3. The quantity $p(y|\beta)$ is a probability and thus bounded between zero and one. As long as data lie within the support of the probability model, this quantity lies strictly above zero. Since the theorem defines the data as such, $p(y|\beta) > 0$. □

Lemma 4. *$\dfrac{\partial^2 p(\beta|\sigma)}{\partial\beta\partial\sigma}$ for $\beta > \beta^*$.*

Proof of Lemma 4. By assumption, the prior density is decreasing at a decreasing rate in β for all $\beta > \beta^*$. Also by assumption, the scale parameter σ controls the rate at which β decreases such that increasing σ leads to a slower rate of decrease. These two assumptions together imply that $\dfrac{\partial^2 p(\beta|\sigma)}{\partial\beta\partial\sigma}$ for $\beta > \beta^*$. □

Now recall Theorem 1:

Theorem 1. *The impact of the researchers choice of σ on the posterior distribution $p(\beta|y)$ is increasing in β for $\beta > \beta^*$.*

Proof of Theorem 1. To show that the effect of σ is increasing in β, I simply need to show that $\dfrac{\partial^2 p(\beta|y)}{\partial\beta\partial\sigma} > 0$ for $\beta > \beta^*$.

Recall that the posterior $p(\beta|y)$ is proportional to the likelihood $p(y|\beta)$ times the prior $\beta(\beta|\sigma)$, so that $p(\beta|y) \propto p(y|\beta)p(\beta|\sigma)$. First, we can use the product rule to obtain the derivative of $p(\beta|y)$ so that

$$\frac{\partial p(\beta|y)}{\partial\beta} \propto \frac{\partial p(y|\beta)}{\partial\beta} p(\beta|\sigma) + p(y|\beta)\frac{\partial p(\beta|\sigma)}{\partial\beta\partial}.$$

Only the final term involves σ, so we can easily obtain the desired derivative

$$\frac{\partial^2 p(\beta|y)}{\partial\beta\partial\sigma} \propto \overbrace{\frac{\partial p(y|\beta)}{\partial\beta}}^{\text{Lemma 1: }+} \overbrace{p(\beta|\sigma)}^{\text{Lemma 2: }+} + \overbrace{p(y|\beta)}^{\text{Lemma 3: }+} \overbrace{\frac{\partial^2 p(\beta|\sigma)}{\partial\beta\partial\sigma}}^{\text{Lemma 4: }+}. \tag{1}$$

Each term on the right-hand side of Equation 1 is positive for $\beta > \beta^*$ (Lemmas 1-4), so that $\frac{\partial^2 p(\beta|y)}{\partial\beta\partial\sigma} > 0$ for $\beta > \beta^*$.

$\square$