Learning & Behavior (2021) 49:307–320
https://doi.org/10.3758/s13420-020-00457-y

Mixed effects modeling of Morris water maze data revisited: Bayesian censored regression

Michael E. Young [1] · Michael R. Hoane [2]

Accepted: 10 December 2020 / Published online: 22 February 2021
© The Psychonomic Society, Inc. 2021

Abstract

Young, Clark, Goffus, and Hoane (*Learning and Motivation, 40*(2), 160–177, 2009) documented significant advantages of linear and nonlinear mixed-effects modeling in the analysis of Morris water maze data. However, they also noted a caution regarding the impact of the common practice of ending a trial when the rat had not reached the platform by a preestablished deadline. The present study revisits their conclusions by considering a new approach that involves multilevel (i.e., mixed effects) censored generalized linear regression using Bayesian analysis. A censored regression explicitly models the censoring created by prematurely ending a trial, and the use of generalized linear regression incorporates the skewed distribution of latency data as well as the nonlinear relationships this can produce. This approach is contrasted with a standard multilevel linear and nonlinear regression using two case studies. The censored generalized linear regression better models the observed relationships, but the linear regression created mixed results and clearly resulted in model misspecification.

Keywords Memory · Morris water maze · Data analysis · Censored regression · Bayesian analysis

A medical researcher conducts a 10-year prospective study of the effectiveness of surgical versus nonsurgical interventions on surviving a heart attack. The main variable of interest is the number of years the patient survives after the intervention. After 10 years, 60% of the patients studied are still alive, which is a good outcome, but one that presents problems for data analysis: Does the researcher categorize these people as surviving for 10 years (but no longer), omit them from the analysis because the number of years survived is unknown, or find a statistical technique that treats these data appropriately as being at least 10 years, but perhaps much longer? Another scientist is measuring how long it takes various lizard species to escape a puzzle box (Cooper et al., 2019) and encounters a problem: Some subjects take so long to complete the task that it is necessary to impose a 3-minute time limit for solving the puzzle. When analyzing the data, running into the time limit presents a problem because the scientist recognizes that it is not possible to safely conclude that the solution time was exactly 3 minutes—it could have been much longer. These examples describe situations involving *censoring*, in which a subject's or trial's observed value represents a minimum but otherwise unknown estimate (e.g., at least 10 years or 3 minutes).

This paper will focus on censoring in the Morris water maze because it presents multiple challenges to the data analyst. First, the data have a multilevel structure that creates data dependencies. Conventionally, each rat is assessed multiple times in a day and across multiple days. Thus, data within a day for the same rat and data from the same rat across days are more highly correlated. A proper analysis must incorporate these dependences (Aarts, Verhage, Veenvliet, Dolan, & van der Sluis, 2014). Second, there is a natural floor for performance that creates curvature in the relationship between latency and trial or day (see Fig. 1, rats 43 and 48). Thus, standard linear regression is not appropriate for many subjects. However, most researchers do not try to model the functional relationship and instead analyze day as if it were an unordered categorical variable, which complicates presentation and interpretation and can result in an underpowered analysis (Young et al., 2009). Third, rats are routinely removed from the maze if they have not found the hidden platform after some predetermined deadline (commonly 60 s for mice and 90 s for

✉ Michael E. Young
michaelyoung@ksu.edu

[1] Department of Psychological Sciences, Kansas State University, 492 Bluemont Hall, Manhattan, KS 66506-5302, USA

[2] Augusta University, Augusta, GA, USA

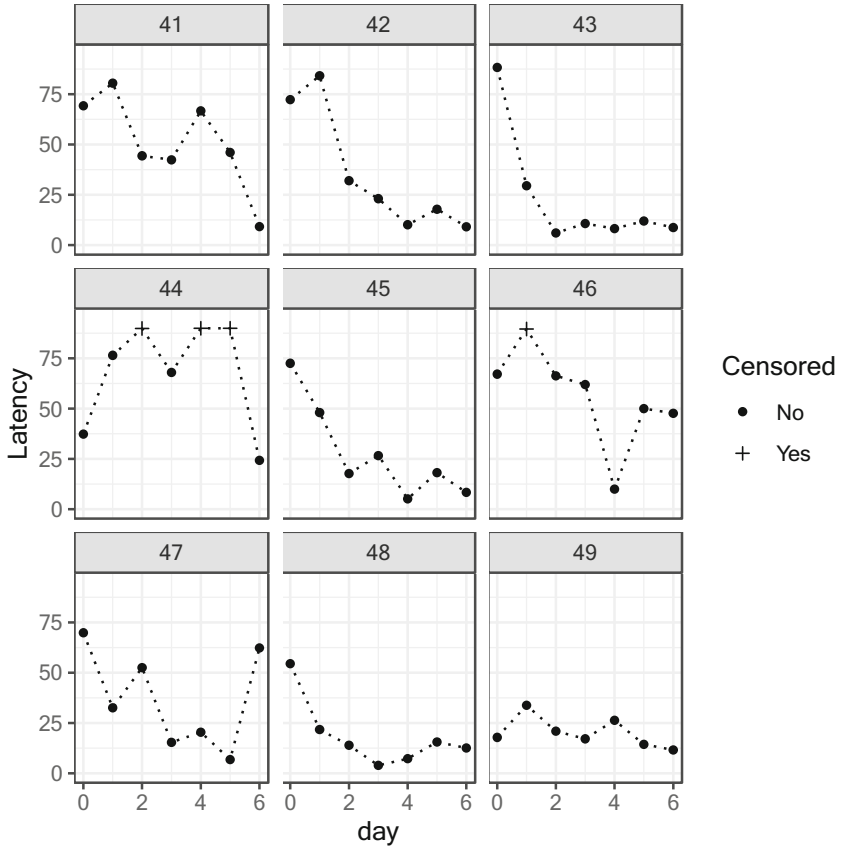

Fig. 1 Examples of real rat latencies in a Morris water maze. The data points designated with a cross are the result of removing a rat from the maze that had not found the platform by the 90-s deadline (i.e., these data points were censored)

rats; see the data points designated by a cross in the sampling of rats shown in Fig. 1—in this unpublished study, 14% of the 1,540 observations were censored). Researchers score these trials as having the deadline value (60 s or 90 s) for the analysis. This censoring of the data (Tobin, 1958) artificially decreases the variability in latencies on these trials and biases the estimated latencies downwards. On these trials, researchers know that the rat did not find the platform by the deadline, but they do not know how much longer it would have taken.

To address the first two challenges, modeling the data dependencies and the curvilinear relationship, Young et al. (2009) suggested the use of multilevel modeling. They compared the traditional repeated-measures approach to Morris water maze data to linear and nonlinear multilevel modeling. All three approaches model at least some of the data dependencies, but the repeated-measures analysis of variance (ANOVA) is restricted to within-subjects predictors being categorical. One of the primary advantages of multilevel modeling is its ability to conduct repeated-measures analysis in which one or more of the within-subjects predictors is continuous (most commonly for Morris water maze data, this continuous variable would be day and/or trial). Young et al. documented that both linear and nonlinear multilevel modeling greatly increased the ability to accurately identify condition differences relative to approaches that model each day as categorically different from another. Although the nonlinear version produced better fits, the linear version fared just as well in identifying group differences (at least for the variety of situations tested in their simulations).

To address the third challenge—censoring—the present paper will evaluate the use of an easy tool to perform Bayesian censored multilevel generalized regression, the *brms* package in R (Bürkner, 2017, 2018), and how this approach fares relative to the use of standard linear multilevel regression and generalized linear multilevel regression. Although censored regression will produce a better model of the process, we will explore the practical implications of its use relative to analyses that do not attempt to account for the censoring.

Bayesian approaches to censored regression are not new (e.g., Gelman, 2004; Kruschke, 2014; Ntzoufras, 2011), but the use of censored regression in Morris water maze data is extremely rare. When we examined the 10 most cited rat/mice Morris water maze papers published since 2016, only one had conducted a censored regression. Furthermore, only one of these papers reported the percentage of censoring that they observed (15% of values in a drug group vs. 7% in another drug group vs. 1% in a saline control), although at least one-third of the latency graphs in these 10 articles documented

mean latencies near the cutoff in at least one group on a day or session. When a mean latency is near the cutoff value, there is evidence of censoring approaching 50% for that condition on that day or session. Note that by only focusing our informal survey on recent highly cited papers, these numbers approximate the best analytical approaches that we would expect to find in the literature. Extending our search to older and less cited papers would document even lower rates, given that censored regression of repeated-measures data has only become tractable relatively recently.

Censoring

One of the cautionary issues raised by Young et al. (2019) was the potential consequences of censoring in the Morris water maze. Although they classified the issue as one of truncation (in which values greater than a threshold are removed) rather than the more accurate term, censoring (in which values greater than a threshold are assigned the threshold value), the challenges of analyzing censored data have been known since Tobin (1958). In the Morris water maze, predicting latencies greater than the deadline is problematic and can only emerge due to linear or nonlinear extrapolation. Statistical analysis of subjects or conditions in which more than a few values are censored could result in severe underestimation of both the mean and variance for those subjects and for those conditions in which data were more likely to be censored. This artificial decrease of the observed variance in latencies near the deadline also violates the assumption of homogeneity of variance for ANOVA and linear-regression analyses. Although Jahn-Eimermacher, Lasarzik, and Raber (2011) recognized the utility of censored regression to address censoring in the Morris water maze and related tasks, the absence of readily available repeated-measures versions of censored regression meant that it was not possible to use their technique for many situations (but see Andersen, Wolf, Jennings, Prough, & Hawkins, in press; Faes, Aerts, Geys, & De Schaepdrijver, 2010). The challenge of a proper censored analysis is further increased by the skewed distributions commonly observed in latency measures that have a hard minimum of zero and a long upper tail.

The emergence of practical Bayesian analysis permits the use of generalized multilevel censored regression of latency data for preparations like the Morris water maze. An introduction to Bayesian analysis is beyond the current paper (see Kruschke & Liddell, 2018), but the approach is able to address issues like censoring because it relies on Monte Carlo simulation to create data with particular properties. This simulation-based approach is also the reason why Bayesian analysis was not practical until recently—computational power and memory demands restricted its use to either simple analyses or to researchers with considerable computational resources (and patience!) at their disposal. Although early implementations of tools like WinBUGS could run such an analysis, they were too demanding of computational resources to prompt broad use. As a result of it becoming more feasible to use multilevel censored regression, the current paper revisits the conclusions of Young et al. (2009) and how an analysis that properly considers censoring might change the conclusions derived from a Morris water maze study. Analyses that do not consider censoring might underestimate means, misestimate the rate of learning, and overestimate the degree of certainty in those estimates (considered in our first case study), but they may also misclassify the presence or absence of interactions (considered in our second case study). We begin with a brief introduction to how censored regression works, and then present two case studies to demonstrate the differences between the results of a standard multilevel regression of censored data versus a censored regression. Because of the computational demands of Bayesian analysis, we will eschew the approach of Young et al. in which thousands of analyses were run for various conditions. The outcomes are apparent enough to be revealed by the case studies presented here.

Censored regression

Data points that are beyond the point of censoring are known to lie in the tail, but where they lie is unknown. Censoring can occur in either tail, but for the purposes of the present discussion involving the Morris water maze, we restrict our discussion to the upper tail (known as "right censoring"), where censoring is known to occur. Censored regression can predict latencies on those trials in which the subject was removed from the maze at the deadline as being in the tail of the distribution, but with uncertainty regarding its position within the tail. The particulars involve estimating the value of a latent or unobserved latency variable, $latency_{Unobs}$, that has a known relationship with the observed latency, $latency_{Obs}$. For an experiment using a deadline of 90 s:

$$latency_{Obs} = \begin{cases} latency_{Unobs}, & latency_{Unobs} < 90 \\ 90, & latency_{Unobs} \geq 90 \end{cases} \quad (1)$$

All other variables in the analysis are then modeled as affecting the latent unobserved latency rather than the observed latency. To ensure that the model behaves well, it is important that the assumed distribution of the latencies is well specified. Latencies can be modeled with a number of different distributions, and the upper tail of each of these will have different probabilities associated with the likelihood of these values being beyond the cutoff.

Given that latencies have a natural floor of 0 s, the appropriate choice of distribution should have a minimum of 0 and a long upper tail. Possibilities include the lognormal, gamma,

Weibull, and exponential distributions. The fit of each of these to a particular data set can be examined, but our observation is that the choice among the lognormal, gamma, and Weibull is not likely to have a major impact on an analysis when the amount of censoring in any particular data condition is relatively low (i.e., 10% or less). However, the Monte Carlo simulations we conducted in the first case study did reveal differences between the two distributions that we examined (gamma and Weibull) when there was substantial censoring with small sample sizes.

The latency variable will comprise both censored and uncensored latencies (i.e., $latency_{Obs}$). To set up data for a censored regression, it is necessary to include an extra variable for each trial with a value that indicates the presence and type of censoring. In the R computing environment and *brms* library that we used, right-censoring is designated as a 1 for this extra variable, and no censoring is designated with a 0. This extra variable signaling censoring differentiates a trial in which the rat found the hidden platform at the 90-s deadline versus being removed at the deadline due to a failure to find the platform.

To perform multilevel censored regression, we used the Bayesian analysis package, *brms*, in the R platform (Bürkner, 2017) for all analyses (censored, not-censored, and across all distributional specifications). The *brms* package is an R front-end for Stan, a widely used Bayesian analysis package. By using *brms*, the analyst can write an analysis using a more familiar syntax (that used in the *lme4* package) and let the *brms* package compile the Stan code to be used in the computation. The general form for all of the analyses was:

$$\begin{aligned}&\text{brm(Latency|cens(Censored)}\sim\text{Day}\times\text{Group}+(\text{Day}|\text{Subject}-\text{ID}),\\&\text{family}=\text{gamma(link}=\text{“log”})\ldots.).\end{aligned}\quad(2)$$

This example represents a full-factorial of Day (as a continuous variable) by Group as predictors of Latency; Day was a within-subjects variable with a slope that can vary across subjects, and the outcome variable, Latency, was modeled as having a gamma distribution with an inverse-logarithmic (i.e., exponential) relationship with the predictors. For details of the analyses, see the Supplemental Materials in OSF (https://osf.io/ntgxu/).

Because we are using a Bayesian analysis, there are no *p* values produced by the analysis. Instead, parameter estimates are coupled with estimated error (the standard deviation of the estimated means) and 95% credible intervals (analogous to 95% confidence intervals; here, they represent the 2.5% quantile and the 97.5% quantile). For more information on the interpretation of the results of a Bayesian analysis, there are numerous resources (e.g., Franck, Koffarnus, McKerchar, & Bickel, 2019; Kruschke, 2014; Young, 2019). However, to connect the current work to the decisions common in many laboratories, we will consider credible intervals that do not include zero as equivalent to rejecting the null hypothesis, despite the drawbacks to such an approach.

Case study 1

To illustrate the results that emerge from analyses that do and do not model censoring, we begin by using Monte Carlo simulation to create a sample data set and then assess how each analytical approach recovers the original population values used to generate this sample. The hypothetical experiment involved five groups (a between-subjects variable) in which the animals were tested across 10 days (designated Day 0 through 9) and a 90 s deadline was used before removal from the maze. Each group contained 10 rats, and each group was simulated to either have population Day 0 latency that was 100 s, thus creating a moderate level of censoring, or 50 s, thus creating little or no censoring. Each group was simulated to learn the task by producing shorter latencies with each subsequent day of testing with the learning rate being faster in some groups than in others; the rate change was logarithmic to ensure that latencies could never be less than zero and to create typical learning curves.

More precisely, the Day 0 population mean (original scale and log transformed) and daily change on a logarithmic scale are shown in Table 1. The 10 rats in each group were simulated to have different population mean performance on Day 0 (these values were drawn from a normal distribution around the rat's group population value in Table 1), and different learning rates (again drawn from a normal distribution around the rat's group population value). Finally, gamma-distributed noise was added to the specific rat's population latency; a gamma distribution was used to ensure that latencies could not be negative (both gamma and Weibull distributions fit well for the Morris water maze rat data sets we had available). For details of the Monte Carlo simulation, see the OSF repository for this manuscript (https://osf.io/ntgxu/).

Note that there are many different situations that can be modeled using this approach. We informally examined situations with longer initial latencies than those tested here, thus creating even more censoring, slower learning rates, more and less variability, and smaller and bigger differences between

Table 1 Population latency values for Day 0 (original and natural log transformed) and the slope of latency change for each subsequent day (natural log scale)

Group	Day 0 (seconds)	Day 0 (log seconds)	Slope (log scale)
1	100	4.60	−0.074
2	50	3.91	−0.137
3	100	4.60	−0.200
4	50	3.91	−0.263
5	100	4.60	−0.326

Note. These values were the basis for creating simulated subjects with Day 0 latencies and latency slopes that varied around these values. Additional measurement error was also incorporated

groups. We settled on the current parameters because they approximated a real data set that could not be used due to averaging across trials, thus losing information on which trial data had been censored. This real data set involved five conditions each, with different degrees of brain injury both with and without treatments intended to improve performance. Thus, a brain injury could have produced much poorer initial performance in two groups, but faster learning in a treated group; the treatment could have improved initial performance, but had no impact on learning; or any of a host of other possible outcomes.

Figure 2 contains violin plots of the generated sample data in which the distribution of latencies on each day in each group are shown (a violin plot shows the complete distribution of observed values). It is apparent that the groups and days differed in the likelihood with which the rats could find the hidden platform before the deadline for removal from the maze. Censoring on the first 2 days of testing is evident for Groups 1, 3, and 5, for which the population mean performance on Day 0 was 100 s.

To analyze these data, we used two censored regression techniques that allow for logarithmic changes in latencies (censored gamma and Weibull regressions), a technique that allowed for logarithmic change but did not recognize the censoring (gamma regression), and a standard linear regression that also did not recognize the censoring. All four analyses were conducted using Bayesian multilevel (i.e., repeated measures) linear regression. The goal was to depict the inference of latencies beyond the 90-s deadline as well as any differences in each technique's inferences regarding group parameter estimates and group differences.

The first censored regression technique involves the specification of a gamma distribution of the outcome variable to capture the obvious skew in latencies that is especially evident in Groups 4 and 5; thus, this analysis was a generalized linear multilevel regression with censoring, with the assumption that the unobserved latencies followed a gamma distribution (with a log-link function to predict the mean). The second censored regression technique specified a Weibull distribution, but was otherwise identical to the first; this approach was explored because initial fits with the gamma consistently overestimated censored values. The third and fourth techniques ignored the censoring and analyzed the data using a standard generalized (gamma) linear model or a standard multilevel linear regression. Although these last techniques were fully expected to underestimate the unobserved latencies that would have occurred on censored trials, they provide a benchmark of comparison because the vast majority of analyses of Morris water maze data do not use censored regression.

A failure to recognize censoring thus creates clear model misspecification when censoring is present, and the linear approach is further misspecified by assuming a linear relationship between day and latency when this relationship cannot be linear (otherwise, negative latencies could be predicted). A further question was whether ignoring the censoring would lead to different conclusions regarding group differences than would the two censored regressions. To ensure direct comparison, all four models were conducted using Bayesian modeling in R's *brms* package and specified weakly informed priors for the parameters. For analysis details and *brms* code, see the OSF repository for this manuscript (https://osf.io/ntgxu/).

The four panels of Fig. 3 show example individual simulated rats and depicts the uncensored data ($latency_{Unobs}$) as well as the censored data ($latency_{Obs}$) with the best fit function derived using each analytical technique. The censored regression techniques are predicting unobserved latencies that are larger than the 90-s deadline, as expected, and both tended to overestimate the actual (unobserved) latencies. However, the gamma regression is overestimating the actual (unobserved) latencies more than the Weibull regression. It should be noted, however, that the uncertainty of these predictions is extremely high, and thus their influence on each group's estimates would be correspondingly much weaker than predictions obtained from latencies that occurred before the deadline (for more on this concept of shrinkage, see Young, 2017). Regardless, the result prompted us to test the scope of this overestimation, and analyses involving significantly larger samples (e.g., 100 simulated rats per condition) produced much better estimation of the censored latencies. Given that the

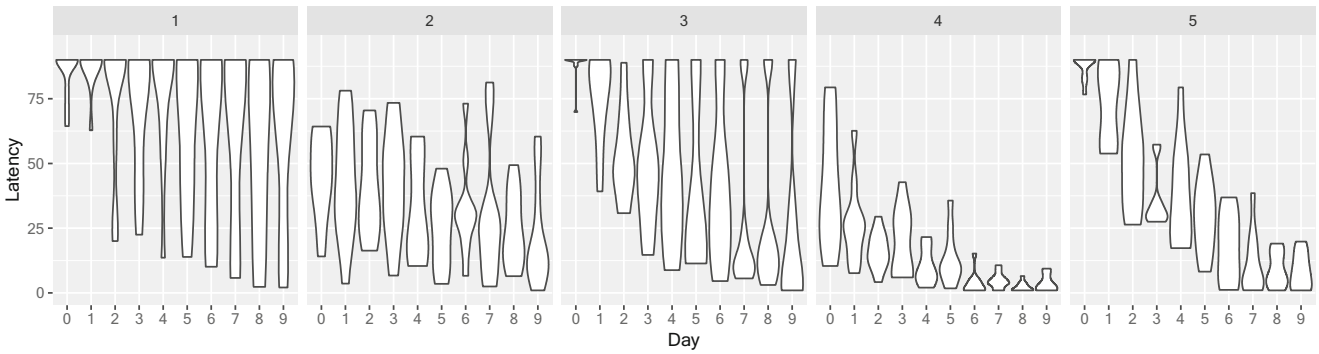

Fig. 2 Violin plots showing the distribution of observed latencies for all of the simulated rats on each day of testing in each of the five groups. Censoring at 90 s is evident in Groups 1, 3, and 5

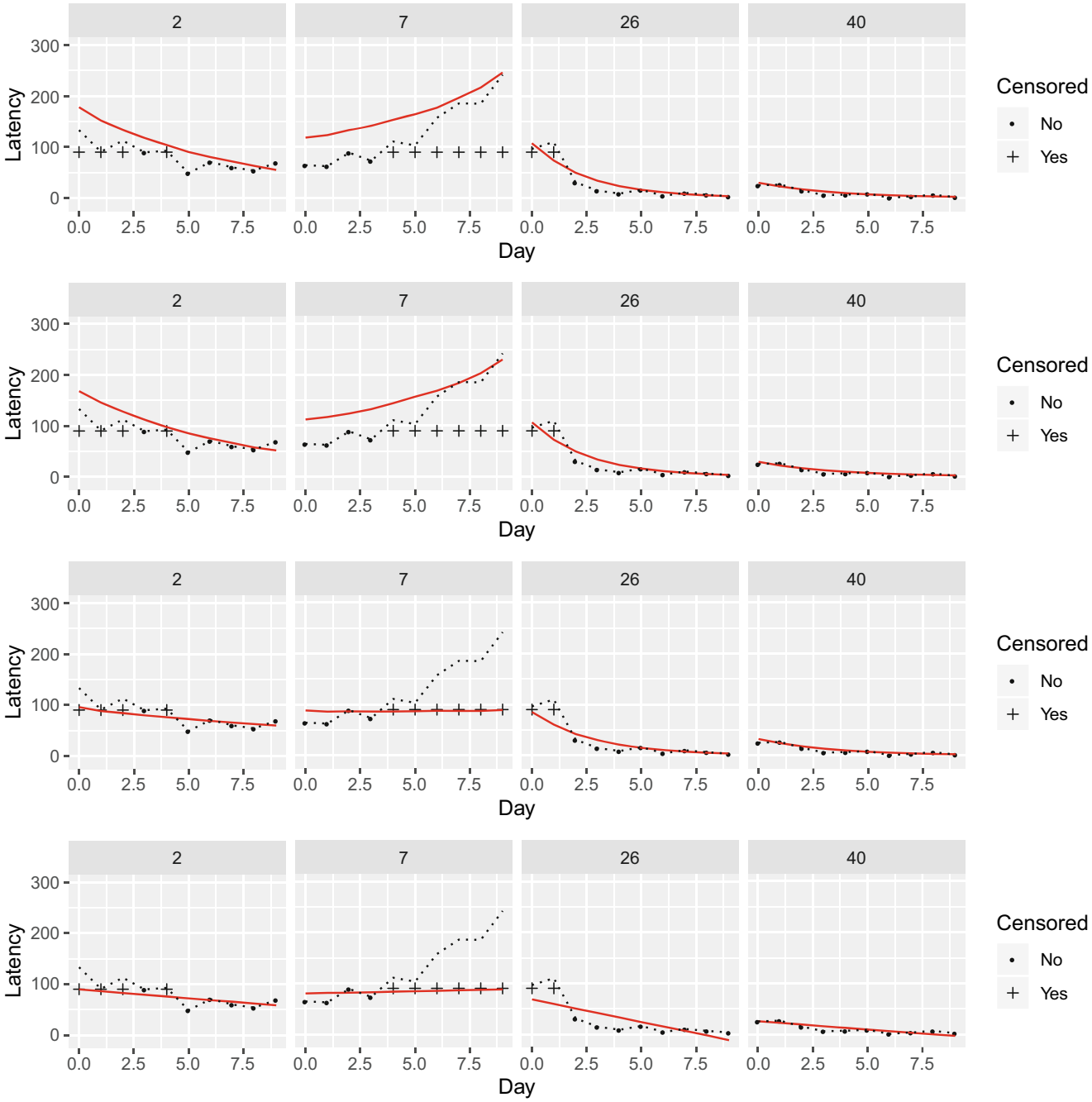

Fig. 3 Superimposed fits for example rats derived from the four modeling approaches. The dashed line shows uncensored data. Each fit was generated using a single multilevel model and not by fitting each individual rat independently from the other

Weibull regression had a smaller tendency to overestimate with small samples, we deemed the Weibull to be more appropriate for the small sample sizes typical in Morris water maze studies.

There are two additional issues of importance in interpreting these graphs. First, the linear regression both underestimates the initial and final performance and overestimates the performance in the middle (see Subject 26). This result is caused by trying to fit a straight line to curved data. Second, the two models (bottom two rows) that do not recognize that censoring has occurred will nearly always underestimate the latencies for trials on which it did occur.

More importantly for scientific inference, we must consider the estimates and fits for the groups as a whole. In multilevel modeling, unusual data or subjects have less influence on the model fit, but they are retained in the analysis rather than using an often-arbitrary exclusion criterion. Furthermore, group estimates involving a lot of censored data should be more uncertain (i.e., have wider error bars) due to the necessary

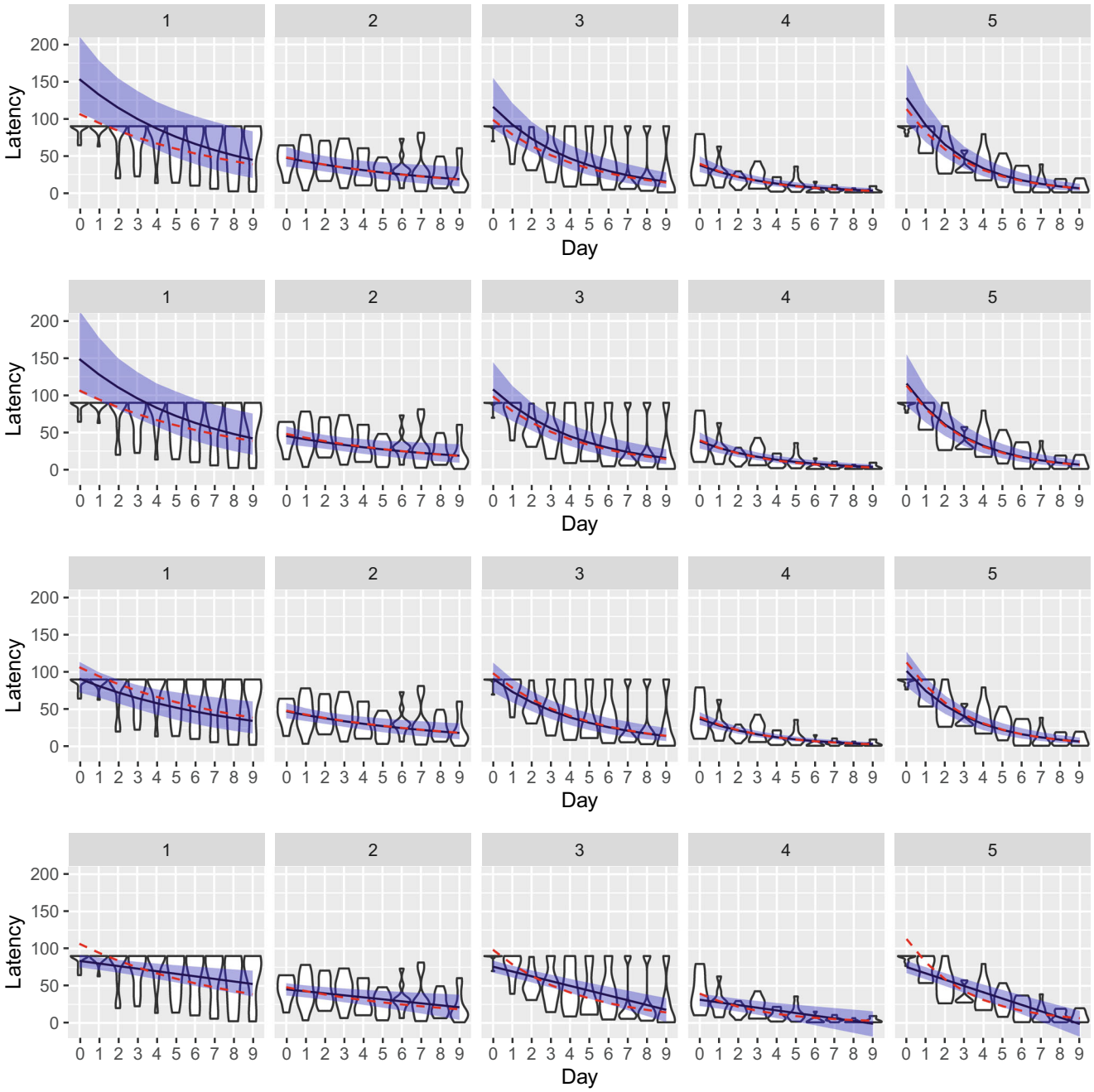

Fig. 4 Violin plots with superimposed fits from the four modeling approaches for the five groups. The blue error region is a 95% credible interval. The red dashed line is the best fit from a model of the uncensored data. (Color figure online)

inference of unobserved latencies. The results are shown graphically in Fig. 4 supplemented by a dashed line indicating the fit from an omniscient gamma regression fit to the uncensored ($latency_{Unobs}$) data.

A few aspects are immediately apparent. First, the estimates for the standard multilevel regressions that do not recognize censoring (bottom two rows) are much lower, as expected. Second, for the censored regressions, the width of the error ribbons (95% credible intervals) are much broader for the conditions that involved a lot of censored data (Groups 1 and 3 and, to a lesser extent, Group 5). Third, the omniscient fit (red dotted line) was within the 95% CI for all of the analyses except for the linear regression. Fourth, the censored regressions for groups involving significant censoring (Groups 1 and 3) tended to overestimate the unobserved latencies. Finally, the linear assumption for the standard linear regression deviated systematically from the omniscient fit.

Figure 5 plots the 66% and 95% credible intervals for the intercepts and slopes for each condition as computed by the three analyses being compared. The figures include a violin

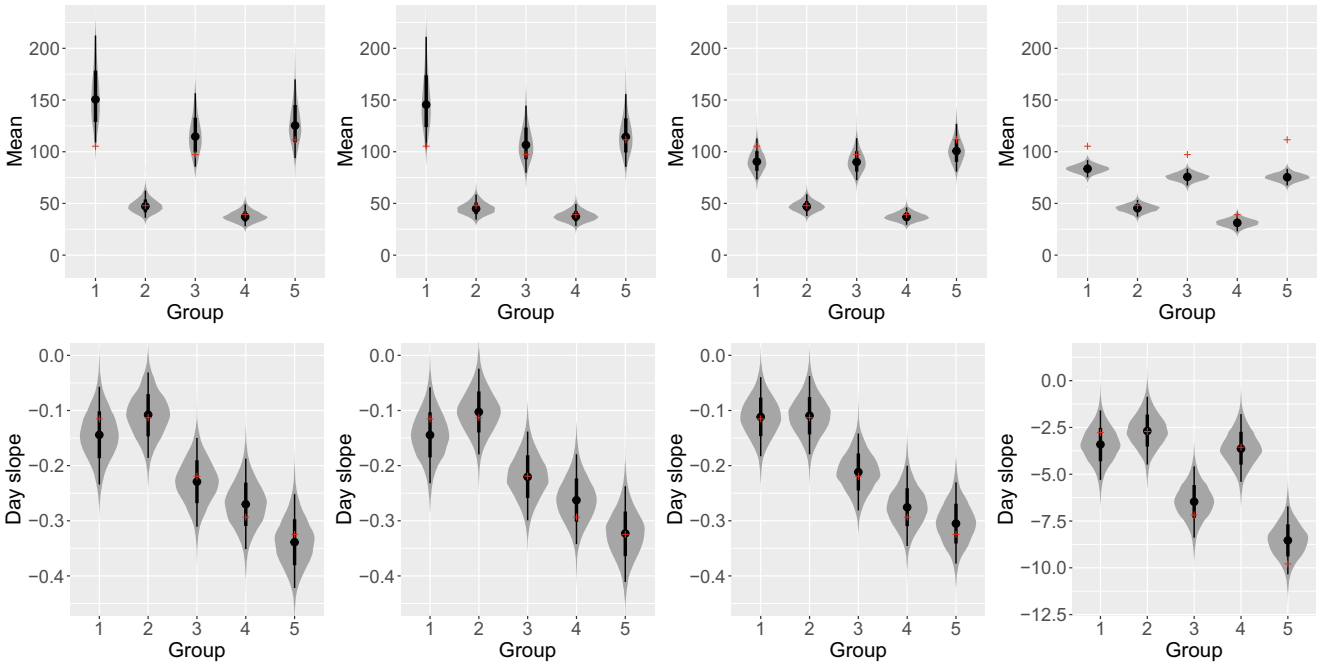

Fig. 5 Violin plots of the full posterior distributions of the intercept (top) and slope (bottom) estimates from the three modeling approaches. The dot represents the median value, the wide line the 66% CI, and the thin line the 95% CI. The red crosses indicate the best fitting values for the corresponding regression that was computed on the uncensored latencies (i.e., an omniscient analysis). The mean intercepts for the first three columns were back-transformed from the log scale to the original scale to ease comparison. The day slopes for the first three columns are on a log scale, whereas the day slope for the linear analysis was on a linear scale. (Color figure online)

plot that illustrates the empirically derived complete posterior marginal distribution of plausible values of each parameter. Unlike the 95% confidence interval, which cannot make any claims about the relative likelihood of each value within the interval, the 95% credible interval (CI) provides what we intuitively believe a confidence interval represents.

Intercepts

The intercept for each group (i.e., Day 0 estimated mean latency) represents initial performance in the Morris water maze on the first day of testing. Thus, this performance should not reflect any significant learning by the subjects. In Case Study 1, three of the groups had much worse initial performance ($latency_{Unobs}$ = 100 s; see Table 1) thus simulating that they had been subjected to an intervention like a brain injury. The other two groups had much better initial performance with a mean of 50 s that was much shorter than the censoring point of 90 s, although some of the simulated rats in these groups could still have produced censoring due to simulated variability.

The first row of Fig. 5 reveals that the predicted Day 0 performance (intercepts) successfully identified that the means for Groups 1, 3, and 5, are similar and higher than those for Groups 2 and 4. The red cross in each group indicates the best fitting value from a gamma regression based on the uncensored data (the omniscient analysis). Three results stand out in these graphs. First, the CIs are much larger for Groups 1, 3, and 5 for the censored analyses (first two columns) than the uncensored ones. Given that these three groups have encountered much more censoring that the other two groups, especially on Day 0 (see Fig. 2), it is appropriate that these estimates generated greater uncertainty in this initial performance. Second, the censored regressions tended to overestimate Day 0 performance in Groups 1, 3, and 5, whereas the standard gamma and linear regressions tended to underestimate them. The degree of deviation was largest for the standard linear regression and fell well outside the 95% CI. The standard gamma regression fared well for all five groups, but we must caution that its ability to extrapolate beyond the 90-s interval will be very limited. Due to the censoring in Groups 1, 3, and 5, the standard gamma regression also produced artificially narrow CIs. Third, the CIs for the standard linear regression were much too narrow, especially for the groups containing censored data. This finding is expected to be quite robust and can lead to an inflation of Type II error.

Day slopes

The day slope for each group represents the rate of learning and should be negative if the latencies are growing shorter with each day. All five groups were simulated to show some degree of learning (see Table 1). In the population being simulated, the slopes should have systemically differed with the first group, showing the shallowest slope and each subsequent group producing increasingly more negative slopes at equally

spaced intervals. The particular random sample presented here (red crosses) had very similar day slopes for Groups 1 and 2, even when analyzed using an omniscient analysis, but the other groups showed the expected increases in slope negativity that are anticipated given the population characteristics.

The second row of Fig. 5 shows the day slopes. Three results stand out in these graphs. First, the first three analyses all generated very similar estimates and uncertainties. The CIs for the standard gamma regression (third analysis) are slightly smaller than the first two, but the slopes derived from the uncensored gamma regression (the red crosses) were well within the 66% CI for all groups for these three analyses. Second, the slopes for the standard linear regression showed a deviation from the expected order, with Group 4 having a much flatter slope than anticipated. This outcome was caused by the inability of a linear regression to adequately capture the curvature in that group—even an omniscient linear analysis that had access to the uncensored latencies showed the same tendency. Third, for the final graph, an omniscient linear analysis based on the uncensored latencies revealed that the linear analysis was underestimating the slopes in Groups 3 and 5 in a way that will be a common outcome for situations in which learning is rapid and performance runs into a floor effect, thus flattening the slope unless properly analyzed using a logarithmic model. Relatedly, Group 4 also has the same underestimation of this learning slope, but this underestimation is only apparent when contrasted with the estimated Group 4 slopes in the logarithmic relationships modeled in the other three analyses.

To reassure the reader, the findings reported here were typical across multiple runs of our simulation; for simplicity, we only show the results from one of those runs. The censored gamma and Weibull regressions performed similarly well, but they overestimated means for those conditions involving significant censoring relative to a gamma regression with full access to the uncensored data (see Figs. 3, 4, and 5). This overestimation decreases with the use of Weibull regression and, not shown here, with the use of larger sample sizes. The standard gamma regression performed admirably in estimating the means and slopes for the five conditions tested, with only small underestimation of means involving censored data. Although we caution that the degree of underestimation will be much higher when there is more censoring (e.g., see Subject 7 in Fig. 3), this nonlinear approach appears to work well with modest degrees of censoring, thus confirming the findings of Young et al. (2009).

In contrast, fitting standard linear regressions to censored data was problematic. First, there was significant underestimation of initial Day 0 performance due to both the curvature common in learning data as well as the censoring (see Figs. 3, 4, and 5); the analysis also computed overly narrow CIs for the intercepts. The inability of a standard linear regression to capture the curvature (and thus the anticipated floor effects as performance rapidly improved) resulted in a misordering of the estimated slopes across the five conditions (see Fig. 5). Thus, standard multilevel linear regression cannot be recommended given these systematic deviations in estimating means and slopes as well as the size of the CIs for means.

There is one analytic approach that we did not test here that was included in Young et al. (2009)—treating the day variable as categorical. We chose not to include this analysis for two reasons. First, when testing this approach Young et al. found much poorer sensitivity—an analysis treating day as an unordered categorical variable was much poorer at correctly identifying real group differences. Second, presenting the results of this analysis would be much more complex because the outcome of a 5 × 10 (Group × Day) ANOVA would require a long series of post hocs or planned comparisons in which the five groups were compared for each of the 10 days of testing (for further discussion of this issue, see Young, 2016).

Implications for researchers

Incorrect inferences will arise when censored data are treated inappropriately. We consider three situations that illustrate the consequences of inappropriate analysis of censored Morris water maze data.

Censored regression models provide a natural way to model the uncertainty when making inferences about observations beyond the deadline. For example, a researcher interested in the consequences of extending the response deadline from 60 s to 90 s would be poorly served by models that both underestimate these inferred latencies and do so with exaggerated certainty. An examination of the censored regressions shown in Figs. 3 and 4 suggests that doubling the cutoff from 90 s to 180 s would capture most of the behavioral profiles exhibited for subjects like Subjects 2 and 7, but the fits suggest that this doubling is unlikely to eliminate all censoring. In contrast, using the standard regressions for such inference would not be trusted by a seasoned researcher because these regressions suggest that even a small increase in the deadline from 90 s to 100 s may suffice to eliminate censoring (see Figs. 3 and 4).

Secondly, imagine a group or rat for which a bulk of the data were censored (e.g., Subjects 2 and 7 in Fig. 3). Any inferences regarding slope differences for similar subjects or for entire conditions with this profile of behavior will be fundamentally flawed. Rather than having a high confidence that there is no learning for Subject 7 (as suggested by the standard gamma and linear multilevel regressions), the censored regressions of this subject suggest that performance may be improving or it could be getting substantially worse over time.

Finally, the ceiling effect created by inappropriate handling of censored data as well as the floor effect natural to latency data both raise the possibility that interactions might be inferred or missed. This issue is explained and explored in our second case study.

Case study 2

In this case study involving Morris water maze data, we examined the ability of these approaches to identify the presence or absence of an interaction. In a previous study involving accuracy data, Dixon (2008) noted that a standard multilevel linear regression can infer the presence of an interaction where none is present and miss an interaction that is indeed present. This occurs due to the presence of ceiling and floor effects when accuracy (0 to 100%) is the outcome. Given the ceiling effect created by censoring of Morris water maze data and the floor effect present with latency data, we hypothesized that standard multilevel linear regression would similarly fail, but that the use of censored regression may eliminate the impact of the latency ceiling at 90 s, and that the use of gamma or Weibull regression would eliminate the impact of the latency floor at 0 s.

To illustrate why floor and ceiling effects are problematic for linear regression, Fig. 6 shows four panels. In the first panel in the top row, we see a situation involving two groups learning at the same rate (these lines are parallel when plotted in log-transformed space). In the second panel, censoring is applied, which greatly reduces the judged slope of one of the groups, thus creating the illusion of a Group × Day interaction when this censoring is ignored. In the third panel (first panel in the second row), there is a situation in which there is a clear interaction between group and day. But the final panel illustrates how censoring flattens the performance of one of the groups, thus underestimating the degree of interaction.

We simulated two data sets to illustrate each of the situations shown in Fig. 6. In the first, no interaction was present between day and group, whereas in the second, an interaction was present. Each data set was analyzed five ways: multilevel Weibull and standard linear regression of the uncensored data (omniscient analyses); Weibull censored regression; Weibull standard regression; and standard linear regression of the censored data. All fits were derived using Bayesian analysis. The omniscient analyses provide the benchmarks against which the other analyses are measured; however, the standard linear regression of the uncensored data is flawed because it cannot fit the learning curves common to Morris water maze data. The modeling details are provided in the OSF materials (https://osf.io/ntgxu/).

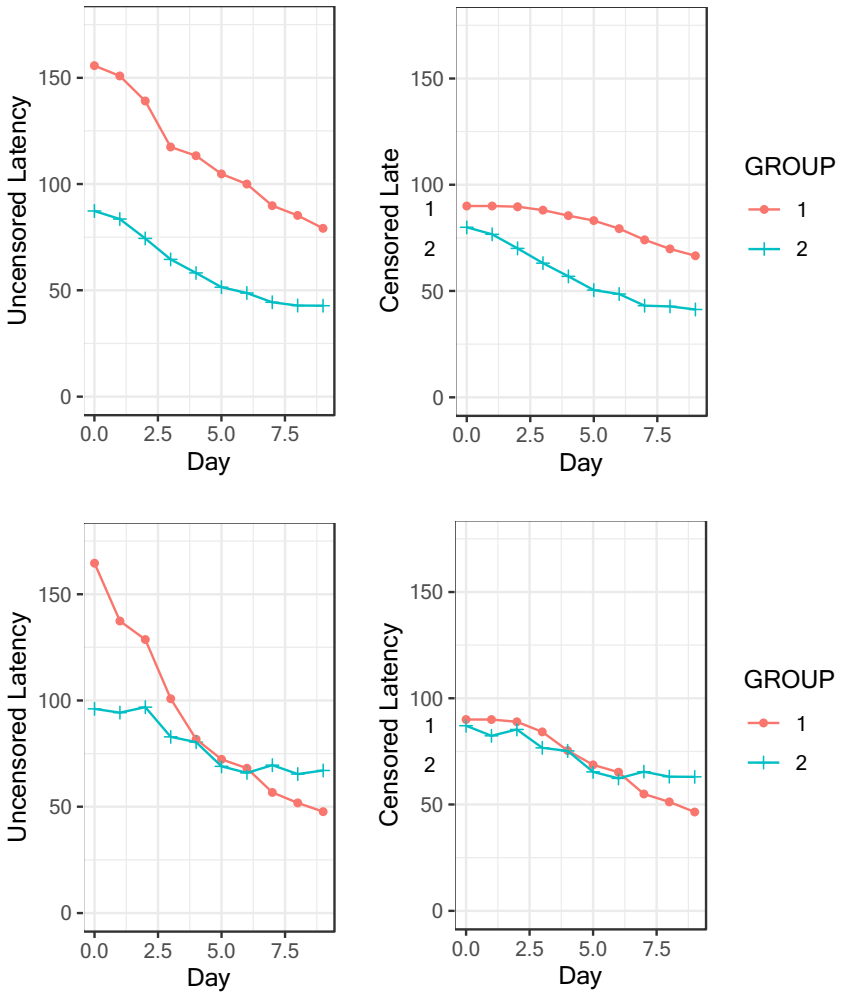

Fig. 6 Plots illustrating how censoring can create a spurious interaction (top two panels) or make a real interaction disappear (bottom two panels)

Spurious interaction

In the situation illustrated at the top of Fig. 6, censoring can make it appear as if an interaction is present when it is not. This will occur when both groups produce similar learning rates, but one of the groups undergoes more censoring than the other due to a main effect of group (overall poorer performance). A secondary factor can offset this effect when the rats learn the tasks quickly; the linear regression for a group with shorter overall latencies can produce flatter learning rates due to the floor effect not being adequately modeled. In the present simulation, we intentionally avoided significant floor effects because it complicates interpretations of the effect of censoring.

The top of Fig. 7 compares the model fits for the three Weibull regressions for the uncensored data (first panel) and censored data (next two panels; first fit is from a censored regression), and the standard linear regression of the censored data (last panel). The censored Weibull overestimated the latencies for the group subject to significant censoring (this overestimation also occurred in the first case study), but the analysis documented the high degree of uncertainty for a group with this level of censoring. In contrast, the standard regressions produced strong underestimation of the latencies in this group and suggest a strong interaction in which the two groups' learning curves diverge with additional training.

To move beyond a merely visual comparison, statistical inferences regarding the presence of an interaction hinge on the estimated regression weight for the Group × Day term. The bottom of Fig. 7 illustrates the full posterior distribution of the interaction regression weight for each of the four fits shown in Fig. 7. The median best fitting value derived from the omniscient fits are shown in each panel using a cross symbol.

Given that the true population value is zero, both the omniscient and censored regressions produced posterior distributions in which zero is within the 95% CI, thus supporting the correct conclusion that there is no interaction. In contrast, the standard regressions produced posterior distributions in which zero is not in the 95% CI supporting the incorrect conclusion that an interaction exists. Because the three Weibull estimates can be directly compared (they all entail estimating slopes in a log-transformed space), it is noteworthy that the predicted interaction regression weight for the standard Weibull was nearly three times larger than that for the omniscient Weibull regression.

Missed or weak interaction

In the situation illustrated at the bottom of Fig. 6, censoring can make it appear as if an interaction is absent when it is present. This will occur when both groups produce very

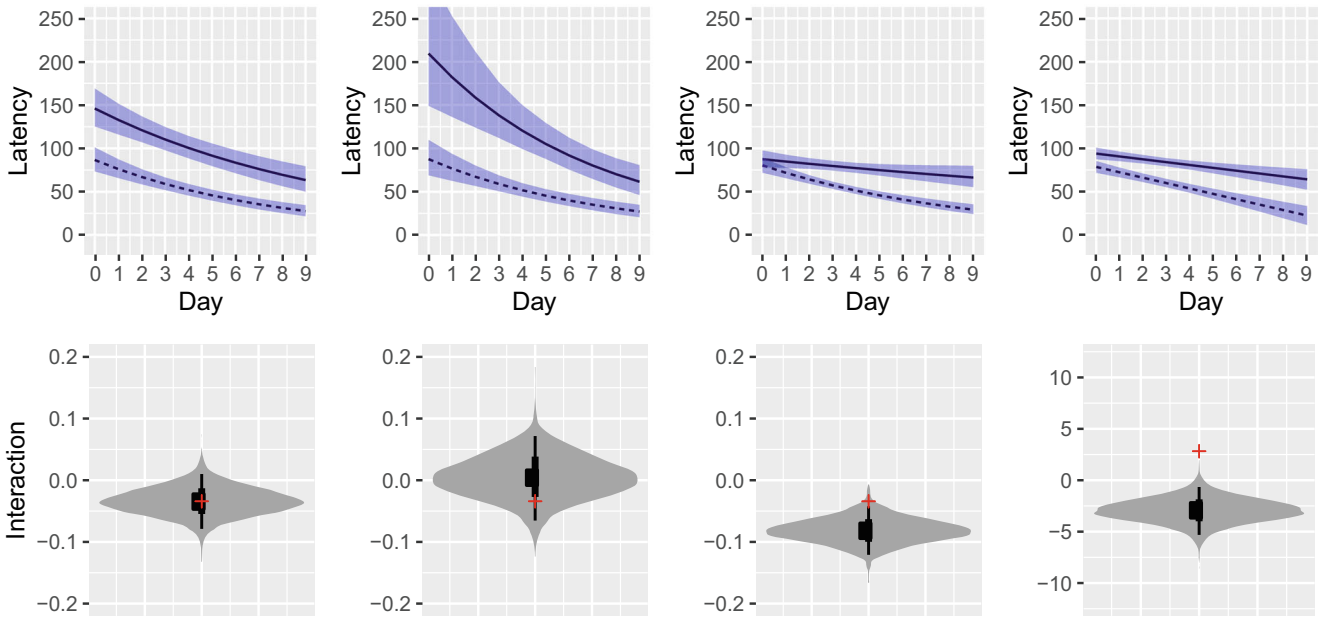

Fig. 7 Top: Plots of the best fitting line from the Bayesian multilevel analyses of data sampled from a population without an interaction in the log-transformed space before censoring. The plot on the left represents an omniscient analysis with full knowledge of the uncensored sample values. The next three plots show analyses of data censored at 90 s. Bottom: Violin plots of the posterior distributions of the regression weight for the Day × Group interaction. The dot represents the median value, the wide line the 66% CI, and the thin line the 95% CI. The population value for this regression weight was zero; the red crosses in each diagram represent the most likely value derived from a corresponding analysis of the uncensored sample. (Color figure online)

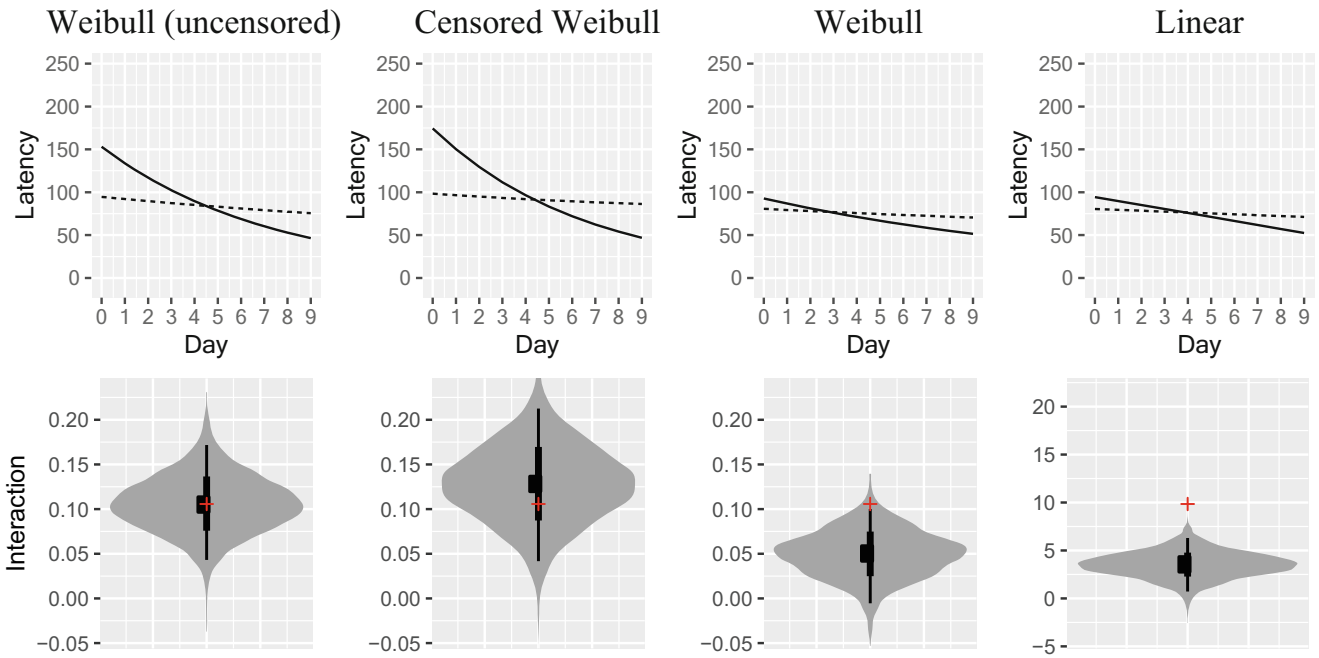

Fig. 8 Top: Plots of the best fitting lines from the Bayesian multilevel analyses of data sampled from a population with an interaction in the log-transformed space before censoring. The plot on the left represents an omniscient analysis with full knowledge of the uncensored sample values. The next three plots show analyses of data censored at 90 s. Bottom: Violin plots of the posterior distributions of the regression weight for the Day × Group interaction. The dot represents the median value, the wide line the 66% CI, and the thin line the 95% CI. The population value for this regression weight was 0.10; the red crosses in each diagram represent the most likely value derived from a corresponding analysis of the uncensored sample. (Color figure online)

different learning rates, but one of the groups undergoes more censoring than the other thus flattening its estimated slope due to the ceiling effect.

The top of Fig. 8 compares the model fits for the three Weibull regressions for the uncensored data (first panel) and censored data (next two panels), and the standard linear regression of the censored data (last panel). The censored Weibull again overestimated the latencies for the group subject to significant censoring, but the analysis documented the high degree of uncertainty for a group with this level of censoring. In contrast, the standard regressions again produced strong underestimation of the latencies in this group and suggest much weaker evidence of an interaction.

We again examined the estimated regression weight for the Group × Day term. Here, the population value was 0.10 in the log-transformed space. The bottom of Fig. 8 illustrates the full posterior distribution of the interaction regression weight for each of the four fits shown in Fig. 8. The median best fitting value derived from the corresponding omniscient fits are shown in each panel, using a cross symbol.

Given that the true population value is 0.10 for the first three analyses, both the omniscient and censored regressions produced posterior distributions in which 0.10 is well within the 66% and 95% CIs, thus producing very good estimates of the magnitude of the interaction. Furthermore, the censored Weibull judged a zero slope, indicating no interaction to be highly implausible (zero was well outside the 95% CI), thus supporting the correct conclusion that there as an interaction. In contrast, the standard regressions produced posterior distributions in which zero was at the edge of the 95% CI ([0.00, 0.10] for the Weibull) or outside the 95% CI ([0.77, 6.44] for the linear), supporting the incorrect conclusion that an interaction was absent. Even though the median interaction regression weight obtained by the standard Weibull and linear regressions was higher than zero (0.05 and 3.57, respectively), these median regression weights were at best half that derived from the corresponding omniscient analysis (0.10 and 9.75, respectively).

Conclusions

Using two case studies, we demonstrated the behavior of censored regression as applied to Morris water maze performance in a repeated-measures design. The censored regressions made statistically sound inferences regarding unobserved latencies beyond the deadline, modeled the greater uncertainty in estimated latencies that had been censored, and better judged the magnitude of an interaction.

When there was limited censoring, a standard multilevel gamma regression did well at estimating mean latencies, how those latencies changed across time, and the uncertainties in those estimates. In contrast, a multilevel linear regression did not perform well even for conditions that involved little censoring because it fails to capture the curvature common in experiments in which learning occurs. Furthermore, analyzing data that has significant censoring but using analyses that do not incorporate that censoring produced significant underestimation of latencies and artificially reduced uncertainty in the estimates. Finally, we examined situations in which the ceiling effect caused by censoring resulted in misestimation of interactions between day of testing and an independent variable.

Censored data like those routinely encountered in Morris water maze studies present challenges that are nicely addressed by the use of censored regression. Although the censored analysis must guess the nature of unobserved data (latencies greater than the deadline), these guesses are not arbitrary but are rather the result of explicit assumptions about the distributional characteristics of the latencies, the behavior of other rats, and experience (i.e., prior experiments) concerning the plausible range of intercepts and slopes generated in Morris water maze experiments.

Censored multilevel regressions in which the distributional properties of the dependent variable can be specified was only possible through the use of Bayesian statistical methods. Bayesian analysis allows researchers to conduct much more sophisticated analyses than have been historically possible. Although there are costs involving the need to learn an unfamiliar approach to data analysis and the greater computational power that is necessary, the increased flexibility of the approach has many benefits beyond those considered here for analyzing censored Morris water maze data.

One such benefit is the ability to specify a wider range of outcome distributions. Here, we examined gamma, Weibull, and Gaussian (normal) distributions, but there are other distributions that possess the long upper tail typical of latency data, including the exGaussian, lognormal, and shifted lognormal. In many cases, distinguishing among these long-tailed distributions may not be possible because of the limited amount of data available. It is best, however, to choose a distribution with characteristics consistent with the measure to ensure that the model cannot predict impossible values (e.g., latencies less than zero).

Another advantage of the Bayesian approach is the ability to perform robust regression in which unusual subject values are not as influential as they would be in a standard regression (Gelman et al., 2013), thus reducing the pressure to remove "outliers." Unusual values or subjects can be true outliers and thus represent the operation of a different process or population (e.g., the rat that cannot swim due to illness, the subject who presses the same key on every trial out of boredom, the nonnative English speaker in a reading study). However, unusual values can also be occurring at the rate at which you would expect given the assumed distribution of individual differences or behavior across time; removing these values will then distort an analysis by asymmetrically truncating a distribution and/or artificially decreasing the observed variability (Ulrich & Miller, 1994). In the case studies considered here, we used *t* distributions with three degrees of freedom for the regression weight priors because *t* distributions have longer tails than normal distributions, thus making unusual observations in the tails less unusual and thus less influential in estimating the parameters. This approach allows the retention of unusual data while reducing their influence on group estimates in an explicit and systematic way.

Given the ability to conduct repeated-measures censored regression using Bayesian approaches in R, these techniques are more accessible than previously possible. By placing the R code used in the present paper in a public OSF repository, it becomes easier for others to apply the approach to their own data. However, although the analyses are relatively easy to run and plots are easy to generate, the statistical output will be unfamiliar because it lacks the ubiquitous degrees of freedom, *t* values, *F* values, and *p* values that are commonly reported in published studies. Researchers, reviewers, and editors will need to gain greater comfort with the rich reporting of posterior distributions, 95% credible intervals, and R-hat values. The ability to address longstanding challenges like censored Morris water maze data makes the effort worthwhile.

Acknowledgements Research reported in this publication was supported by the Cognitive and Neurobiological Approaches to Plasticity (CNAP) Center of Biomedical Research Excellence (COBRE) of the National Institutes of Health under Grant Number P20GM113109.

References

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & van der Sluis, S. (2014). A solution to dependency: Using multilevel analysis to accommodate nested data. *Nature Neuroscience, 17*(4), 491–496. https://doi.org/10.1038/nn.3648

Andersen, C. R., Wolf, J., Jennings, K., Prough, D. S., & Hawkins, B. E. (in press). Accelerated failure time survival model to analyze Morris water maze latency data. *Journal of Neurotrauma*. https://doi.org/10.1089/neu.2020.7089.

Bürkner, P. C. (2017). brms: An R package for Bayesian generalized linear mixed models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal, 10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Cooper, T., Liew, A., Andrle, G., Cafritz, T., Dallas, H., Niesen, T., Slater, E., … Mendelson, J., III. (2019). Latency in problem solving as evidence for learning in varanid and helodermatid lizards, with comments on foraging techniques. *Copeia, 107*(1), 78–84. https://doi.org/10.1643/CH-18-119

Dixon, P. (2008). Models of accuracy in repeated-measures design. *Journal of Memory and Language, 59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

Faes, C., Aerts, M., Geys, H., & De Schaepdrijver, L. (2010). Modeling spatial learning in rats based on Morris water maze experiments. *Pharmaceutical Statistics, 9*(1), 10–20. https://doi.org/10.1002/pst.361

Franck, C. T., Koffarnus, M. N., McKerchar, T. L., & Bickel, W. K. (2019). An overview of Bayesian reasoning in the analysis of delay-discounting data. *Journal of the Experimental Analysis of Behavior, 111*(2), 239–251. https://doi.org/10.1002/jeab.504

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association, 99*(466), 537–545. https://doi.org/10.1198/016214504000000458

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton: CRC.

Jahn-Eimermacher, A., Lasarzik, I., & Raber, J. (2011). Statistical analysis of latency outcomes in behavioral experiments. *Behavioral Brain Research, 221*(1), 271–275. https://doi.org/10.1016/j.bbr.2011.03.007

Kruschke, J. K. (2014). *Doing Bayesian data analysis*. Elsevier.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 25*(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. Hoboken, NJ.: Wiley. https://doi.org/10.1002/9780470434567

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26*, 24–36.

Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General, 123*(1), 34–80. https://doi.org/10.1037/0096-3445.123.1.34.

Young, M. E. (2016). The problem with categorical thinking by psychologists. *Behavioural Processes, 123*, 43–53. https://doi.org/10.1016/j.beproc.2015.09.009

Young, M. E. (2017). Discounting: A practical guide to multilevel analysis of indifference data. *Journal of the Experimental Analysis of Behavior, 108*(1), 97–112. https://doi.org/10.1002/jeab.265

Young, M. E. (2019). Bayesian data analysis as a tool for behavior analysts. *Journal of the Experimental Analysis of Behavior, 111*(2), 225–238. https://doi.org/10.1002/jeab.512

Young, M. E., Clark, M. H., Goffus, A., & Hoane, M. R. (2009). Mixed effects modeling of Morris water maze data: Advantages and cautionary notes. *Learning and Motivation, 40*(2), 160–177. https://doi.org/10.1016/j.lmot.2008.10.004

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.