Supplementary Materials for "A Time To Event Framework For Multi-touch Attribution"

Dinah Shender, Ali Nasiri Amini, Xinlong Bao, Mert Dikmen, Jing Wang, Amy Richardson Fricke

In Appendix A we prove a result about $E[NormalizedCredit(j)]$ mentioned in section 3.2.1 of the main text. In Appendix B we give a detailed discussion of the similarities and differences between Backwards Elimination (our proposed attribution method) and Shapley Values, another commonly used method. In Appendix C, we present additional simulation scenarios and their results.

The supplement also contains an additional file, data.csv, with sample data used in our simulations; see the README file for details. We do not include code to replicate the simulation results. While our method can be applied using standard Poisson regression methods, our simulation framework is tightly integrated with our production environment and our proprietary data format. This makes sharing the code impractical.

A Appendix: Expected Value of $C_{\textbf{norm}}(j)$

In section 3.2.1 we claim that assuming that our estimated intensity is correct, we have

$$E[C_{\text{norm}}(j)] = \int \hat{\lambda}(t, \mathcal{A}(j)) - \hat{\lambda}(t, \mathcal{A}(j-1))\mathrm{d}t,$$

where the expectation is over all conversion occurrences in the path. Recall that

$$C_{\text{norm}}(j) = \frac{\hat{\lambda}(t^*, \mathcal{A}(j)) - \hat{\lambda}(t^*, \mathcal{A}(j-1))}{\hat{\lambda}(t^*, \mathcal{A}(n))},$$

where without loss of generality, n was taken to be the number of ads at or before time t^*. Then we have

$$\begin{aligned} E\left[C_{\text{norm}}(j)\right] &= E\left[\frac{\hat{\lambda}(t^*, \mathcal{A}(j)) - \hat{\lambda}(t^*, \mathcal{A}(j-1))}{\hat{\lambda}(t^*, \mathcal{A}(n))}\right] \\ &= \int_0^\infty \frac{\lambda(t, \mathcal{A}(j)) - \lambda(t, \mathcal{A}(j-1))}{\lambda(t)}\mathrm{d}Y(t), \end{aligned}$$

where $Y(t)$ is the measure for the occurrences in the Poisson process (i.e. the conversion occurrences). Then we have

$$\begin{aligned} &= \int_0^\infty \frac{\lambda(t, \mathcal{A}(j)) - \lambda(t, \mathcal{A}(j-1))}{\lambda(t)}\lambda(t)\mathrm{d}t \\ &= \int_{t_j}^\infty \lambda(t, \mathcal{A}(j)) - \lambda(t, \mathcal{A}(j-1))\mathrm{d}t \end{aligned}$$

as claimed. The last equality follows because the intensities of a path with the first $j-1$ ads only differs from the intensity of a path with the first j ads after the j^{th} ad has occurred.

For non-incrementality models, the integral above can be computed (replacing λ with its estimate $\hat{\lambda}$) as long as the features and times of the first j ads are known, since $\lambda(t, \mathcal{A}(j))$ does not depend on events after t_j. However, as discussed in Section 3.2.2, for incrementality models,

$\lambda(t, \mathcal{A}(j))$ can depend on all query (i.e. non-ad) events between t_j and t, in addition to all ad events up until t_j. Therefore the integral can only be computed if all query events (but not necessarily ad events) that ever occur are known. More practically, for t much larger than t_j, we would expect the contribution from the j^{th} ad to be effectively 0. In practice, it should suffice to assume that query events are known as long as $\lambda(t, \mathcal{A}(j)) - \lambda(t, \mathcal{A}(j-1))$ is non-negligible, which may be only a few days, depending on the data.

B Appendix: Ad Synergies in Backwards Elimination Attribution and Comparison to Shapley Values

We mentioned previously that backwards elimination assigns credit for conversions requiring multiple ads to the last ad in the group. In this section, we make that more precise and compare how backwards elimination and Shapley values divide this credit. We start by defining what we mean by an ad's marginal effect as well as "conversions requiring multiple ads."

Define the marginal credit of a set of one or more ads $\mathcal{A}$ to be $m(\mathcal{A}) = \hat{\lambda}(\mathcal{A}) - \hat{\lambda}(\emptyset)$, where we have dropped the t^* for brevity. This marginal credit is exactly the difference in conversion intensity at conversion time with exactly these ads versus without any ads. When our algorithm starts, it has $m(\mathcal{A}(n))$ units of raw credit to distribute and when it gets to the j^{th} ad it has $m(\mathcal{A}(j))$ units of credit left to distribute, by definition. If we think of the algorithm as splitting credit between two groups of ads, $\mathcal{A}(j-1)$ and the singleton j^{th} ad, $\{A_j\} = \{(t_j, X_j)\}$, then the credit given to the earlier ads, $\mathcal{A}(j-1)$, is exactly their marginal credit. Meanwhile, the credit assigned by the algorithm for the later ad, $\{A_j\}$, is $m(\mathcal{A}(j)) - m(\mathcal{A}(j-1))$.

Then we can define the synergy due to a user seeing both $\mathcal{A}(j-1)$ and $\{A_j\}$ as

$$S(\mathcal{A}(j-1), A_j) = m(\mathcal{A}(j)) - m(\mathcal{A}(j-1)) - m(A_j),$$

i.e. the difference between the credit received by A_j when seen after the ads in $\mathcal{A}(j-1)$ and the marginal ad credit for a path containing only $\{A_j\}$[2]. When the conversion intensity is super-additive, so that showing multiple ads leads to more conversions than the sum of the marginal increase in conversions from showing each ad individually, $S(\mathcal{A}(j-1), A_j)$ is positive and backwards elimination distributes the extra synergy credit to the later ad. When the conversion intensity is subadditive, so that showing multiple ads leads to fewer conversions than the sum of the marginal increase in conversions from showing ad individually, $S(\mathcal{A}(j-1), A_j)$ is negative, and the later ad gets credit smaller than its marginal credit.

Example: Suppose that a user path has 2 ads, A_1, A_2, prior to a conversion and suppose that $\hat{\lambda}(t^*, \{A_1, A_2\}) = \exp(f_1(t^* - t_1) + f_2(t^* - t_2))$, i.e. with no other interactions and with the baseline rate equal to 1. Suppose also that $\hat{\lambda}(t^*, \{A_1\}) = \exp(f_1(t^* - t_1)) = 2$, $\hat{\lambda}(t^*, \{A_2\}) = \exp(f_2(t^* - t_2)) = 3$, in other words the contribution from each ad is the same as if it were the only ad in the path. Then the overall raw ad credit for the 2-ad path will be $m(\{A_1, A_2\}) = \hat{\lambda}(t^*) - \hat{\lambda}(\emptyset) = 2 \times 3 - 1 = 5$. The raw marginal contribution of these ads is $m(A_1) = 1$, $m(A_2) = 2$. The raw credit of these ads is

$$C_{\text{raw}}(A_2) = 2 \times 3 - 2 = 4$$
$$C_{\text{raw}}(A_1) = 2 - 1 = 1.$$

[2]In fact, the synergy can be defined more generally for an ad A and a set of ads $\mathcal{A}$, independent of their order, as $S(\mathcal{A}, \{A\}) = m(\mathcal{A} \cup A) - m(\mathcal{A}) - m(A)$, but we will not require this more general definition.

We can see that the first ad receives credit equal to its marginal effect. The second ad receives credit equal to its marginal effect, ($m(A_2) = 2$), plus the synergy between the first and second ads, $S(\{A_1\}, A_2) = m(\{A_1, A_2\}) - m(A_1) - m(A_2) = 5 - 1 - 2 = 2$.

We make two additional remarks regarding this example:

1. In general, the synergy between ads will decrease as the time between ads increases. More precisely, if we assume the parameterization in the example, some straightforward algebra shows that
$$S(\{A_1\}, A_2) = (\exp(f_1(t^* - t_1)) - 1)\,(\exp(f_2(t^* - t_2)) - 1)\,.$$
Assuming the f_i are decreasing functions, and fixing both the conversion time, t^*, as well as the time between the last ad and the conversion, $t^* - t_2$, then as the time between the ads, $t_2 - t_1$, increases, the synergy will generally decrease, since $t^* - t_1$ will increase, causing the first term to decrease. In other words, if the ads are far apart, their synergy will be smaller.
2. While the synergy in the example is positive, negative synergy can occur when there is diminishing or even zero marginal benefit to showing multiple ads, perhaps due to user ad fatigue. As a numerical example, consider the case where $\hat{\lambda}(t^*, \{A_1\}) = 2$ and $\hat{\lambda}(t^*, \{A_2\}) = 3$ as before, but $\hat{\lambda}(t^*, \{A_1, A_2\}) = 3$. In other words, the fact that the user has seen multiple ads doesn't increase the intensity; only the timing of the last ad matters. This can be parameterized by allowing a dependence between ads, as described in Equation (3). In this case $C_{\text{raw}}(A_1) = 1$ as before, but $C_{\text{raw}}(A_2) = 3 - 2 = 1$, which is less than $m(A_2)$, implying that the synergy is negative and that we would be better off showing the ads to different users rather than the same user.

Giving all of this synergy or interaction credit to the last ad may seem unfair, since the additional increase or decrease in the conversion intensity only happens if **all** of the ads occur. However, recall that one potential use case for attribution is as an input to bidding, leading to higher bids on ads that drive more conversions. Giving the synergy to the last ad rather than the earlier ad reflects our knowledge when we are bidding on the earlier ad. At that time, without either additional assumptions or modeling, we do not know if this user will have future ad impressions for us to bid on, therefore one reasonable approach is to bid based on the marginal effect of this current ad, together with its synergies with past ads, while ignoring any synergies between the ad being bid on and future ads (which may or may not occur). Thus, in this application, backwards elimination might be desirable precisely because it gives all the synergy credit to the last ad.

Other methodologies, such as the popular Shapley value method used in Dalessandro et al. (2012), Shao and Li (2011), Du et al. (2019), split this synergy evenly amongst the ads involved. Shapley values come from game theory and try to fairly divide the payoff (increase in intensity) for a coalition (group of ads) amongst the players (individual ads). It does this by giving each player credit proportional to its average marginal contribution to all possible coalitions (subsets of ads). More precisely, in a game with player set $\Omega, (|\Omega| = N)$ and value function v, the payoff to player i is given by:

$$\phi_j(v) = \sum_{O \subseteq \Omega \setminus \{j\}} \frac{1}{N!} |O|!(N - |O| - 1)!\,[v(O \cup \{j\}) - v(O)]\,.$$

In the case of ad attribution, the value function is $v(\cdot) = \hat{\lambda}(t^*, \cdot)$ (or equivalently, $v(\mathcal{A}) = \hat{\lambda}(t^*, \mathcal{A}) - \hat{\lambda}(t^*, \emptyset)$ for any set $\mathcal{A}$), while the set of players equals the set of all ads before the conversions, i.e. $\Omega = \mathcal{A}(n)$. j is the index of the ad receiving credit.

For a path with only a single ad, both backwards elimination and Shapley values will result in the same credit assignment. Similarly, the total credit given to all ads (as opposed to the baseline), is also the same for both methods. However, if there are 2 or more ads with nonzero synergy, the credit assignment will differ.

Example: Continuing with our previous example, it is straightforward that the raw ad credit for A_1 using the Shapley value method will be

$$\phi_{A_1}(\hat{\lambda}(t^*, \cdot)) = \frac{1}{2}\left(\hat{\lambda}(t^*, \{A_1\}) - \hat{\lambda}(t^*, \emptyset)\right) + \frac{1}{2}\left(\hat{\lambda}(t^*, \{A_1, A_2\}) - \hat{\lambda}(t^*, \{A_2\})\right) = 2.$$

Similarly, we will have

$$\phi_{A_2}(\hat{\lambda}(t^*, \cdot)) = 3.$$

This gives more credit to the first ad compared to the Backwards Elimination method. In fact, its credit equals its marginal effect, 1, plus half of the synergy between the ads, $1/2 \times 2$. Similarly, the credit for the second ad equals its marginal effect plus half the synergy.

More generally, it can be shown that for synergy requiring k ads, Shapley values gives each ad additional credit equal to $1/k$ of that synergy. The proof is similar to Theorem 1 in Zhao et al. (2018). For some applications or media types, this strategy of dividing the synergy may be preferred. For example, when retrospectively examining data for a group of ad campaigns splitting the synergy may be a more desirable way to compare the relative contributions of the campaigns.

While backwards elimination and Shapley values divide the synergy credit differently, both depend on an ad's contribution to $\lambda(t^*)$. In particular, this implies that given a suitable model for $\lambda(t)$, both methods will satisfy Requirement #2 from Section 2.2: conversion credit for an ad will decrease the further it is from conversion time, and the credit given to ads depends not just on their order, but on their spacing in time[3]. In that sense, both are reasonable choices. Given a model for $\lambda(t)$, the two methods can be compared based on the specific goals of the application.

C Appendix: Additional Simulations

In this section we consider some additional simulations beyond Scenario 1 from the main text. Scenario 2 is a simpler 1 ad version of Scenario 1, which we include for completeness. Scenario 3 is similar to Scenario 1, but the second ad type has no effect. Scenario 4 is an extension of Scenario 1, where we have the same two types of ads, but the number of ads per user is random. Scenarios 5-8 use the same simulated dataset (an extension of Scenario 2), but consider different ways the intervals might be misspecified in the model.

As before, in each scenario, we simulate 500 distinct data sets, each with 1 million users and 30 days of data per user. Given the intensity for each user, we follow Algorithm 5 in Pasupathy (2010) to simulate the corresponding Poisson process. For each dataset, we also create 1000 bootstrap replicates and fit the model on both the original and replicated datasets. We compute the "basic" or reverse percentile bootstrap interval (Davison and Hinkley, 1997) for the model coefficients:

$$(2\hat{\theta} - \hat{\theta}^*_{1-\alpha/2}, 2\hat{\theta} - \hat{\theta}^*_{\alpha/2}),$$

[3]Requirement #1, that we handle incomplete or censored data, depends on only on $\lambda(t)$, not the attribution methodology.

Table 3: Ground truth and estimated model coefficients for Scenario 2.

	Ground Truth	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
Baseline (per day) [$\exp(\alpha_0)$]	0.0333	0.0333 [0.0332, 0.0334]	95.2%
Short term [$\exp(\beta_1)$]	2.0	2.009 [1.994, 2.025]	95.4%
Medium term [$\exp(\beta_2)$]	1.5	1.509 [1.495, 1.523]	93.8%
Long term [$\exp(\beta_3)$]	1.2	1.201 [1.196, 1.206]	95.4%

where $\hat{\theta}$ is the estimate of the parameter using the original dataset and $\hat{\theta}^*_\alpha$ is the α quantile of the bootstrap estimates for the parameter. We take $\alpha = 0.95$ and also compute the coverage of the resulting CIs over the 500 distinct datasets.

As discussed in Scenario 1 in the main text, we can also consider the additive error (AICPE$-$ICPE) and the relative error (AIPCE/ICPE $-$ 1) and whether they cover 0. See the discussion in Scenario 1 for more detail.

C.1 Scenario 2

Here we simulate exactly one ad per user, with the ad occurrence time being uniform over our 30 day observation window, i.e. $[0, 30.0]$. We simulate a pre-query conversion rate of 1 conversion / 30 days and an ad effect that doubles the conversion rate on the first day after the ad, increases it by a factor of 1.5 on the second day, and increases it by a factor of 1.2 after that. We refer to these as the short, medium, and long term effects, respectively. This scenario fits the model in Equation (1):

$$\log(\lambda(t)) = \alpha_0 + \beta_1 I\{t - t_1 \leqslant 1\} + \beta_2 I\{1 < t - t_1 \leqslant 2\} + \beta_3 I\{2 < t - t_1 \leqslant 30\},$$

where we have rescaled time to be measured in days rather than hours.

Table 3 shows the ground truth parameters, an example estimate and bootstrap 95% CI for a single advertiser (to help with the reader's intuition about the typical width of the CI's), as well as the coverage percentage of the 95% CI's over the 500 datasets. As we can see, we get reasonably close to the nominal 95% coverage.

With exactly one ad per user, it is straightforward to see that $E[\text{ICPE}] = 11.95$. Leveraging that $E[\text{AICPE}] = E[\text{ICPE}]$, we can look at the coverage for ICPE, AICPE, and the additive and relative errors. As we see in Table 4, the coverage is close to the nominal coverage.

C.2 Scenario 3

In this scenario we build off of Scenario 1 in the main text and simulate the second ad type as having no effect on user's conversion rates. This could occur for public service announcements or similar types of ads. The other simulation details and model formulation are as in Scenario 1. The table below shows the true values used in our simulations, together with the parameter

Table 4: Coverage for AICPE, ICPE and their errors for Scenario 2.

	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
ICPE $\times$ 1$e-2$	12.00 [11.75, 12.23]	94.2%
AICPE $\times$ 1$e-2$	12.03 [11.85, 12.21]	95.0%
Additive Error $\times$ 1$e-2$	0.03 $[-0.26, 0.33]$	94.8%
Relative Error $\times$ 1$e-2$	0.22 $[-2.17, 2.74]$	94.8%

estimates and CI's, derived as before. Here we see that the intervals tend to overcover, i.e. are slightly too wide.

Since ad type 2 has no effect on conversions, it is straightforward to see that $E[\text{ICPE}] = 11.95$, just as for Scenario 2. Leveraging that $E[\text{AICPE}] = E[ICPE]$, we can look at the coverage for ICPE, AICPE, and the additive and relative errors. As we see in Table 6, the coverage is close to the nominal coverage.

C.3 Scenario 4

In this scenario we use the same two types of ads as in Scenario 1, but allow the number of ads per user to vary. We also allow more freedom in the model we fit than is necessary to describe the data generating process for the simulations, giving insight as to how the model performs when it has extra degrees of freedom. The details are below.

We allow the number of ads per user to be either 1, 2, or 3. The probability of each is proportional to the probability of the corresponding number of events for Poisson random variable with mean 2. Equivalently, we can think of the number of events for each user as being a Poisson(2) random variable that is clipped to be between 1 and 3. On average, 40.6% of the users have a single ad in the path, 27.1% have 2 ads, and 32.3% have 3 ads. Each ad is equally likely to be of either type. So conditional on a user seeing 3 ads total, they may see 0, 1, 2 or 3 ads of each type, as long as the overall number of ads is 3. The ad occurrence times are still independent and uniform on our 30 day observation window, i.e. on [0, 30].

Recall that ad type 1 increased the conversion intensity by a factor of 2x on the first day, 1.5x on the second day, and 1.2x after that, while ad type 2 increased the conversion intensity by a factor of 1.5x on the first day, 1.2x on the second day and 1.0x (no increase) after that. We assume that the effect of a second or third ad of the same type is the same as the effect of the first ad. For example, if a user sees an ad of type 1 at time 0 and a second ad of type 1 one hour later, then the user's ad intensity will be $2 \times 2 = 4$x higher than if they had seen no ads at all. More generally, we simulate the intensity as

$$\begin{aligned}\log(\lambda(t)) = \alpha_0 &+ \beta_1\#\{i : 0 < t - t_i \leqslant 1, \text{ad type}(i) = 1\} + \beta_2\#\{i : 1 < t - t_i \leqslant 2, \text{ad type}(i) = 1\} \\ &+ \beta_3\#\{i : 2 < t - t_i \leqslant 30, \text{ad type}(i) = 1\} \\ &+ \beta_4\#\{i : 0 < t - t_i \leqslant 1, \text{ad type}(i) = 2\} + \beta_5\#\{i : 1 < t - t_i \leqslant 2, \text{ad type}(i) = 2\} \\ &+ \beta_6\#\{i : 2 < t - t_i \leqslant 30, \text{ad type}(i) = 2\},\end{aligned}$$

Table 5: Ground truth and estimated model coefficients for Scenario 3.

		Ground Truth	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
Baseline (per day) [$\exp(\alpha_0)$]		0.0333	0.0333 [0.0332, 0.0334]	93.4%
Ad Type 1	Short term [$\exp(\beta_1)$]	2.0	1.997 [1.981, 2.015]	94.2%
	Medium term [$\exp(\beta_2)$]	1.5	1.502 [1.487, 1.516]	95.6%
	Long term [$\exp(\beta_3)$]	1.2	1.203 [1.198, 1.207]	94.4%
Ad Type 2	Short term [$\exp(\beta_4)$]	1.0	1.000 [0.998, 1.011]	97.8%
	Medium term [$\exp(\beta_5)$]	1.0	1.000 [0.997, 1.011]	97.6%
	Long term [$\exp(\beta_6)$]	1.0	1.000 [1.000, 1.000]	94.8%

Table 6: Coverage for AICPE, ICPE and their errors for Scenario 3.

	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
ICPE $\times 1e-2$	11.94 [11.71, 12.19]	94.2%
AICPE $\times 1e-2$	12.04 [11.87, 12.23]	94.2%
Additive Error $\times 1e-2$	0.10 $[-0.21, 0.41]$	94.2%
Relative Error $\times 1e-2$	0.80 $[-1.85, 3.42]$	93.8%

where $\#\{i : \textit{conditions}\}$ denotes the number of ads i satisfying the condition. Equivalently we could write

$$\begin{aligned}\log(\lambda(t)) = \alpha_0 &+ \sum_{i:\text{ad type}(i)=1} \left[\beta_1 I\{0 < t - t_i \leqslant 1\} + \beta_2 I\{1 < t - t_i \leqslant 2\} + \beta_3 I\{2 < t - t_i \leqslant 30\}\right] \\ &+ \sum_{i:\text{ad type}(i)=2} \left[\beta_4 I\{0 < t - t_i \leqslant 1\} + \beta_5 I\{1 < t - t_i \leqslant 2\} + \beta_6 I\{2 < t - t_i \leqslant 30\}\right].\end{aligned}$$

However, when we fit the model, we do not assume that the second and third ad of the same type have the same effect as the first. Instead we allow for more degrees of freedom. For this section we will use γ to denote the coefficients estimated by the model, while continuing to use β for the true coefficients used to generate the simulation data. Then the model we fit is

$$\begin{aligned}\log(\lambda(t)) = \alpha_0 &+ \sum_{k=1}^{3} \Big[\gamma_{11k} I\{\text{exactly } k \text{ type 1 ads with } 0 < t - t_i \leqslant 1\} \\ &\qquad + \gamma_{12k} I\{\text{exactly } k \text{ type 1 ads with } 1 < t - t_i \leqslant 2\} \\ &\qquad + \gamma_{13k} I\{\text{exactly } k \text{ type 1 ads with } 2 < t - t_i \leqslant 30\}\Big] \\ &+ \sum_{k=1}^{3} \Big[\gamma_{21k} I\{\text{exactly } k \text{ type 2 ads with } 0 < t - t_i \leqslant 1\} \\ &\qquad + \gamma_{22k} I\{\text{exactly } k \text{ type 2 ads with } 1 < t - t_i \leqslant 2\} \\ &\qquad + \gamma_{23k} I\{\text{exactly } k \text{ type 2 ads with } 2 < t - t_i \leqslant 30\}\Big].\end{aligned}$$

For each γ_{ijk}, i indexes the ad type (1 or 2), j indexes the time interval or term (1 for (0, 1] (short term), 2 for (1, 2] (medium term), and 3 for (2, 30] (long term)), while k indexes the exact number of ads of type i in interval type j that is under consideration. In particular, it is straightforward to see that the ground truth value for γ_{11k} is $k \times \beta_1$ or on the original scale $\exp(\gamma_{11k})$ should equal $\exp(k\beta_1)$. Similarly, γ_{12k} corresponds to $k \times \beta_2$ or on the original scale $\exp(\gamma_{12k})$ corresponds to $\exp(k\beta_2)$.

As before, we simulate 500 independent data sets, each with 1 million users, and fit models independently on each data set. We use 1000 bootstrap replicates for each dataset to create a 95% CI for each coefficient, and calculate the coverage of the CI's across the 500 datasets. The results are in Table 7.

We notice that for the short and medium term effects in the 3 ad case (3 ads in a term), coverage is much worse. The sample point estimate and CI suggest that the point estimates are less accurate and the CI's are wider. The coverage for the remaining coefficients remains reasonable. These results can be explained by the relative sums of the interval lengths for which the ad effects are active. In the remainder of this section we will give numerical results from the simulation as well as intuition.

The model fit in this scenario does not assume that the effect of the second or third ad is the same as the first. As a result, each γ_{ijk} is estimated based only on data from the sections of user paths where it is active (i.e. where the corresponding indicator function is not 0). The amount of data can be quantified by summing up the interval lengths (i.e. summing the offsets where the feature is active). Table 8 gives the average (across the 500 data sets) interval length, rounded to the nearest day, that each γ_{ijk} is active. Since the two ad types are symmetric in this respect, their results are quite similar - we show them side by side to save space.

Table 7: Ground truth and estimated model coefficients for Scenario 4.

Parameter to Estimate					
Ad type	Time Term	# ads in term	Ground Truth	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
	Baseline (per day) [$\exp(\alpha_0)$]		0.0333	0.0334 [0.0333, 0.0335]	93.4%
Ad Type 1	Short Term	1 ad [$\exp(\gamma_{111})$]	2.0 [$\exp(\beta_1)$]	1.986 [1.969, 1.999]	93.6%
		2 ads [$\exp(\gamma_{112})$]	4.0 [$\exp(2\beta_1)$]	4.127 [3.913, 4.328]	95.0%
		3 ads [$\exp(\gamma_{113})$]	8.0 [$\exp(3\beta_1)$]	7.458 [4.004, 10.912]	34.6%
	Medium Term	1 ad [$\exp(\gamma_{121})$]	1.5 [$\exp(\beta_2)$]	1.500 [1.487, 1.514]	94.0%
		2 ads [$\exp(\gamma_{122})$]	2.25 [$\exp(2\beta_2)$]	2.167 [2.006, 2.322]	91.8%
		3 ads [$\exp(\gamma_{123})$]	3.375 [$\exp(3\beta_2)$]	1.194 [-0.415, 1.975]	33.0%
	Long Term	1 ad [$\exp(\gamma_{131})$]	1.2 [$\exp(\beta_3)$]	1.199 [1.194, 1.204]	95.2%
		2 ads [$\exp(\gamma_{132})$]	1.44 [$\exp(2\beta_3)$]	1.433 [1.424, 1.444]	91.6%
		3 ads [$\exp(\gamma_{133})$]	1.728 [$\exp(3\beta_3)$]	1.680 [1.651, 1.709]	92.8%
Ad Type 2	Short Term	1 ad [$\exp(\gamma_{211})$]	1.5 [$\exp(\beta_4)$]	1.496 [1.482, 1.508]	93.0%
		2 ads [$\exp(\gamma_{212})$]	2.25 [$\exp(2\beta_4)$]	2.343 [2.170, 2.502]	94.0%
		3 ads [$\exp(\gamma_{213})$]	3.375 [$\exp(3\beta_4)$]	1.212 [-0.512, 2.137]	45.8%
	Medium Term	1 ad [$\exp(\gamma_{221})$]	1.2 [$\exp(\beta_5)$]	1.207 [1.194, 1.219]	96.0%
		2 ads [$\exp(\gamma_{222})$]	1.44 [$\exp(2\beta_5)$]	1.456 [1.317, 1.581]	94.8%
		3 ads [$\exp(\gamma_{223})$]	1.728 [$\exp(3\beta_5)$]	1.232 [-1.048, 1.560]	9.0%
	Long Term	1 ad [$\exp(\gamma_{231})$]	1.0 [$\exp(\beta_6)$]	1.000 [0.995, 1.000]	99.2%
		2 ads [$\exp(\gamma_{232})$]	1.0 [$\exp(2\beta_6)$]	1.000 [0.993, 1.002]	99.4%
		3 ads [$\exp(\gamma_{233})$]	1.0 [$\exp(3\beta_6)$]	1.000 [0.999, 1.032]	99.8%

Table 8: Average observed offset length for model coefficients in Scenario 4.

Parameter to Estimate		Ad Type 1 Offsets (days)	Ad Type 2 Offsets (days)
Time Term	# of ads in term		
Short Term	1 ad [$\exp(\gamma_{i11})$]	922,555	922,526
	2 ads [$\exp(\gamma_{i12})$]	9,972	9,977
	3 ads [$\exp(\gamma_{i13})$]	44	44
Medium Term	1 ad [$\exp(\gamma_{i21})$]	891,288	891,255
	2 ads [$\exp(\gamma_{i22})$]	9,633	9,636
	3 ads [$\exp(\gamma_{i23})$]	42	42
Long Term	1 ad [$\exp(\gamma_{i31})$]	8,173,706	8,172,859
	2 ads [$\exp(\gamma_{i32})$]	1,831,234	1,831,550
	3 ads [$\exp(\gamma_{i33})$]	229,927	230,008

The γ_{ijk} with the smallest average offset are the ones with the worse coverage: there simply isn't enough data to estimate those parameters well, and there isn't enough data for the asymptotics for the bootstrap CI's to kick in: we would need to instead use a bootstrap CI more appropriate for small samples. Indeed, when we increase the number of users by 10x but otherwise keep the simulation the same, the CI coverage improves markedly.

To understand why some interval lengths are much shorter, we can think of these interval lengths as depending on two factors: frequency (number of users) for which γ_{ijk} is active and the length of time per user that it is active. In general, effects that require more ads of a single type will be active less often, and short and medium term effects are active for less time than long term effects.

First consider the frequency, or the number of users for which γ_{ijk} is active and $k > 1$ (# of ads >1). Recall that users have a 32% chance of having 3 ads and a 27% chance of having 2 ads total. Moreover, given that users have 3 (resp 2) ads total, the chance that they are all of the same type is only $1/4$ (resp $1/2$). Consider first the long-term effects of having multiple ads of the same type. We focus on the effect of having 3 ads of the same type: it is straightforward to see that this is the rarest of the three, and the computations for the case of 2 ads of the same type are more involved, since they must consider paths with both 2 ads total and 3 ads total. For the 3 ad case, we might expect all of the users with 3 ads of the same type (or approximately 1 million $\times\, 0.32 \times 1/4 = 80$K users) to eventually have the long-term effect of 3 ads of the same type be active. In fact, the average number of users with either γ_{133} or γ_{233} active is 65K, since we must account for the cases where the third ad occurs in the last two days of the observation period, in which case there is no point in the user path where the long-term effect of three ads is active. Nevertheless, the take-away is that a fairly large number of users will at some point in their path have active the long term effects of having multiple ads.

By contrast, the frequency of short and medium term effects with multiple ads active is much lower. Not only does the user need to have 2 or 3 ads of the same type, but they must occur within a 24 hour period. This turns out to be fairly rare: on average, out of 1 million users there are approximately 262 users with 3 ads of one type within 24 hours and approximately 80K users with 2 ads of one type within 24 hours of each other, again split evenly amongst the two types[4].

Now consider the length of time for which the ad effects are active. The short and medium term effects of ads, regardless of the number of ads, can last at most 24 hours per ad, by definition. Depending on the number of ads in a user's path, they can have at most 3 x 24hour intervals where the short or medium term effects of a single ad are active or 1 x 24 hour interval where the short or medium term effects of 2 or 3 ads are active. However in the 2 (resp 3) ad cases, the interval will generally be shorter than 24 hours, since for e.g., the short term, it will start at the time of the second (resp third) ad and end 24 hours after the first ad occurred[5]. The effect of having multiple ads in the medium term will similarly be based on the same interval but shifted forward by 24 hours - although it can be shorter if it is cut off by the end of the 30 day observation window. By contrast, the intervals for the long term effects can be much longer, up to 28 days long, with the length depending on the position of the ads in the path. While we

[4] In general, users with 3 ads of the same type within 24 hours will also have a brief window where the short or medium term effect of having 2 ads of the same type will also be active. However, at 262 such users, this does not affect the frequency much

[5] Unless a third ad of the same type occurs within 24 hours of the first, in which case the corresponding interval for the effect of two ads of the same type will be shorter.

still expect intervals for the long-term effects of 2 or 3 ads to potentially be shorter than those for just 1 ad, the intervals overall are generally much longer than for short and medium term effects. So all else equal, we should expect to have more data for estimating the long term effects of ads than for short or medium term effects, while also expecting the intervals for the effect of multiple ads to be shorter than those for a single ad.

While these explanations are heuristic, this intuition matches the results shown in Table 8 and can in principle be made more precise. They also give a sense of the amount of data needed to estimate the different parameters in the model. While this also depends on the model fitting procedure used (e.g., how much, if any, regularization is applied), we can see how the accuracy of the estimates change with the amount of data we have for a parameter.

We can confirm this by repeating the simulations with 10x as many users. The coverage can be see in Table 9 and as expected, the coverage is closer to the nominal level, although it doesn't quite achieve it. For example, the coverage for $\exp(\gamma_{213}$ goes from 45.8% to 93.0%.

We've focused so far on the issues around the model coefficients. The AICPE errors suffer no such issues, as seen in Table 10, presumably because they don't depend on any particular narrow slice of data.

C.4 Scenarios 5-8

For the next several scenarios we use a single dataset and consider the performance of several misspecified models. As in Scenario 2, we simulate two possible types of ads, with each user seeing exactly one ad of each type and each ad type having a different effect on conversions. Also as in Scenario 2, the ad occurrence times are independent and uniform on $[0, 30]$. The first ad type increases the conversion rate by a factor of 3 on day 1, a factor of 2.5 on day 2, a factor of 2 on day 3, a factor of 1.5 on day 4, and a factor of 1.2 thereafter. The second ad type increases the conversion rate by a factor of 2.5 on day 1, a factor of 2 on day 2, a factor of 1.5 on day 3, a factor of 1.2 on day 4, and a factor of 1.0 (no increase) thereafter. More formally, we simulate the intensity as

$$\begin{aligned}\log(\lambda(t)) = \alpha_0 &+ \beta_1 I\{0 < t - t_{\text{ad type 1}} \leqslant 1\} + \beta_2 I\{1 < t - t_{\text{ad type 1}} \leqslant 2\} + \beta_3 I\{2 < t - t_{\text{ad type 1}} \leqslant 3\}\\ &+ \beta_4 I\{3 < t - t_{\text{ad type 1}} \leqslant 4\} + \beta_5 I\{4 < t - t_{\text{ad type 1}} \leqslant 30\}\\ &+ \beta_6 I\{0 < t - t_{\text{ad type 2}} \leqslant 1\} + \beta_7 I\{1 < t - t_{\text{ad type 2}} \leqslant 2\} + \beta_8 I\{2 < t - t_{\text{ad type 3}} \leqslant 2\}\\ &+ \beta_9 I\{3 < t - t_{\text{ad type 2}} \leqslant 4\} + \beta_{10} I\{4 < t - t_{\text{ad type 2}} \leqslant 30\},\end{aligned}$$

with the ground truth values for the parameters as in Table 11.

We consider 4 different misspecified models. As before, each model fits a piecewise constant intensity function, but the pieces no longer match the ground truth. For two of the models we eliminate some of the breakpoints, resulting in fewer pieces, but still use the correct breakpoints for the remaining pieces. The remaining two models use incorrect breakpoints with varying numbers of pieces. The exact details are in Table 12. We focus on models where the misspecification is in the shape of the ad effect, and thus the intensity is misspecified, because using a Poisson process and its intensity to model user conversions is one of our key contributions. Other types of misspecification, such as the data not following a Poisson distribution, are also possible of course, but they can be understood by considering the effects of misspecification on Poisson regression generally.

Analyzing the performance of the misspecified models is not straightforward. We can fit them and get confidence intervals for the coefficient estimates as before, however, it's no longer

Table 9: Ground truth and estimated model coefficients for Scenario 4 with 10x increase in users.

Parameter to Estimate					
Ad type	Time Term	# ads in term	Ground Truth	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
	Baseline (per day) [exp(α_0)]		0.0333	0.0334 [0.0333, 0.0334]	95.2%
Ad Type 1	Short Term	1 ad [exp(γ_{111})]	2.0 [exp(β_1)]	2.003 [1.998, 2.008]	95.2%
		2 ads [exp(γ_{112})]	4.0 [exp($2\beta_1$)]	3.945 [3.878, 4.007]	95.4%
		3 ads [exp(γ_{113})]	8.0 [exp($3\beta_1$)]	6.044 [4.809, 7.369]	80.6%
	Medium Term	1 ad [exp(γ_{121})]	1.5 [exp(β_2)]	1.500 [1.496, 1.505]	94.0%
		2 ads [exp(γ_{122})]	2.25 [exp($2\beta_2$)]	2.198 [2.151, 2.245]	92.2%
		3 ads [exp(γ_{123})]	3.375 [exp($3\beta_2$)]	3.380 [2.284, 4.411]	88.6%
	Long Term	1 ad [exp(γ_{131})]	1.2 [exp(β_3)]	1.199 [1.198, 1.201]	96.6%
		2 ads [exp(γ_{132})]	1.44 [exp($2\beta_3$)]	1.443 [1.440, 1.446]	92.6%
		3 ads [exp(γ_{133})]	1.728 [exp($3\beta_3$)]	1.722 [1.712, 1.731]	94.8%
Ad Type 2	Short Term	1 ad [exp(γ_{211})]	1.5 [exp(β_4)]	1.500 [1.496, 1.505]	91.8%
		2 ads [exp(γ_{212})]	2.25 [exp($2\beta_4$)]	2.250 [2.201, 2.300]	93.6%
		3 ads [exp(γ_{213})]	3.375 [exp($3\beta_4$)]	3.604 [2.670, 4.557]	93.0%
	Medium Term	1 ad [exp(γ_{221})]	1.2 [exp(β_5)]	1.200 [1.196, 1.204]	94.6%
		2 ads [exp(γ_{222})]	1.44 [exp($2\beta_5$)]	1.465 [1.426, 1.502]	92.4%
		3 ads [exp(γ_{223})]	1.728 [exp($3\beta_5$)]	1.279 [0.688, 1.630]	62.6%
	Long Term	1 ad [exp(γ_{231})]	1.0 [exp(β_6)]	1.000 [0.999, 1.000]	96.6%
		2 ads [exp(γ_{232})]	1.0 [exp($2\beta_6$)]	1.000 [0.999, 1.001]	99.6%
		3 ads [exp(γ_{233})]	1.0 [exp($3\beta_6$)]	1.000 [0.999, 1.002]	100%

Table 10: Additive and relative error of AICPE for Scenario 1.

	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
Additive Error $\times 1e-2$	-0.11 $[-0.43, 0.17]$	94.6%
Relative Error $\times 1e-2$	-0.80 $[-3.09, 1.15]$	94.8%

Table 11: Ground truth coefficients for Scenarios 5-8.

		Ground Truth	Duration
Baseline (per day) [exp(α_0)]		0.0333	N/A
Ad Type 1	Short term 1 [exp(β_1)]	3.0	[0, 1]
	Short term 2 [exp(β_2)]	2.5	(1, 2]
	Short term 3 [exp(β_3)]	2.0	(2, 3]
	Short term 4 [exp(β_4)]	1.5	(3, 4]
	Long term [exp(β_5)]	1.2	(4, 30]
Ad Type 2	Short term 1 [exp(β_6)]	2.5	[0, 1]
	Short term 2 [exp(β_7)]	2.0	(1, 2]
	Short term 3 [exp(β_8)]	1.5	(2, 3]
	Short term 4 [exp(β_9)]	1.2	(3, 4]
	Long term [exp(β_{10})]	1.0	(4, 30]

Table 12: Intervals for Piecewise Constant Intensity Function, Scenarios 5-8

	Correct Breakpoints		Incorrect Breakpoints	
Ground Truth	Scenario 5	Scenario 6	Scenario 7	Scenario 8
$[0, 1], (1, 2), (2, 3], (3, 4], (4, 30]$	$[0, 4], (4, 30]$	$[0, 2], (2, 4], (4, 30]$	$[0, 0.5], (0.5, 4.5], (4.5, 30]$	$[0, 0.5], (0.5, 2.5], (2.5, 5.5], (5.5, 30]$

clear what the ground truth or target for these estimates is. For example, for Scenario 5, what "should" the coefficient be for the interval corresponding to the 4 days after the ad, $[0, 4]$? One reasonable choice would be the average ground truth, weighted by time. So for example, for an effect corresponding to $[0, 4]$ for ad type 1, we'd target

$$\frac{1}{4}(3 + 2.5 + 2 + 1.5) = 2.25. \tag{6}$$

In Scenario 8, for the effect corresponding to the interval $(0.5, 2.5]$ for ad type 1, we'd target

$$\frac{1}{2}\left(\frac{1}{2} \times 3 + 1 \times 2.5 + \frac{1}{2} \times 2.0\right) = 2.5.$$

For a model with coefficients that match these time-weighted averages and an interval with a single active ad effect, the predicted number of conversions will equal the expected number of conversions based on the ground truth. However, due to the misspecification of the structure of the ad effect, this is not enough to predict well for intervals with multiple active ad effects. The data from such intervals pushes the model estimates away from the naive time-weighted averages. The worse the misspecification, the more they are pushed away. We illustrate this with a detailed example.

Example: As an extreme example, consider a path where one ad of each type occurs at $t = t_0$ and consider the interval $[t_0, t_0 + 4]$. Of course, two ads happening simultaneously is highly unlikely in our simulation set-up, but this keeps the calculations simple; a similar point holds for any interval with multiple active ad effects. The ground truth for the number of expected conversions in this interval is

$$\frac{1}{4}(3 \times 2.5 + 2.5 \times 2 + 2 \times 1.5 + 1.5 \times 1.2) = 4.325.$$

In Scenario 5, the time-weighted average for the short-term effect (lasting for the first 4 days after the ad) for ad type 1 is 2.25 (as in equation (6)), while a similar calculation shows that the time-weighted average for the short-term effect for ad type 2 is 1.8. If these were the model coefficients, then the predicted number of conversions in $[t_0, t_0 + 4]$ would be $2.25 \times 1.8 = 4.05$, underpredicting relative to the ground truth. Thus, as we will see in the results for Scenario 5, the model will instead estimate slightly higher coefficients, so that it is less biased on the intervals with multiple ad effects.

A similar computation for Scenario 6 shows that if the model coefficients matched the naive time-weighted averages, the predicted number of conversions in $[t_0, t_0 + 4]$ would be 4.275. That is, since the model has more intervals (and is therefore less misspecified), a model with coefficients that predict perfectly (in expectation) on intervals with a single ad effect, is able to do better, in the form of less severe underprediction, on intervals with multiple ad effects. As a result, we would expect the estimated coefficients in Scenario 6 to be higher than the naive averages, but lower than for Scenario 5. We would also expect better performance in Scenario 6 than in Scenario 5, since in the former scenario the tradeoff between predicting well on intervals with a single active ad effect and predicting well on intervals with multiple active ad effects isn't as steep.

Repeating the computation for Scenario 7, we see that if the model coefficients matched the naive time-weighted averages, the predicted number of conversions in $[t_0, t_0 + 4]$ would be 3.79. For Scenario 8 it would be 4.06. The comparison between Scenarios 7 and 8 parallels that between Scenarios 5 and 6. As we increase the number of intervals, a model with coeffcients matching the naive targets underpredicts less on intervals with multiple active ad effects. However, Scenario 7 still underpredicts more severely than Scenario 5, despite having more intervals. This is likely due to the misspecification of the breakpoints. In particular, the ground truth long-term effect (active in the interval $(4, 30]$) lowers the naive average for the effect corresponding to $(0.5, 4.5]$ for both ad types, which subsequently lowers the predicted conversions in $[t_0, t_0 + 4]$ for a model that matches the naive averages.[6] Scenario 8 underpredicts more severely than Scenario 6 for analogous reasons. ■

As alluded to in the example, the model will try to predict well on all types of intervals, pushing the coefficient estimates away from the naive time-weighted averages. There is also a dependence on the data: if the data has more intervals with multiple active ad effects (or more ad effects active at once), this effect will be stronger. Thus we don't expect the estimated model coefficients to match the time-weighted averages, but we do expect them to get closer the less misspecified the model is. This makes the frequency with which the CI's for the model coefficients cover the naive time-weighted averages an interesting metric to gain insight the effect of misspecification on model performance.

Of course, this is not of practical use for identifying model misspecification on real data. Instead, as in the other scenarios, we also include the AICPE additive and relative error. As discussed for previous simulations, a correctly specified model will have $E[\text{AICPE}] = E[\text{ICPE}]$. Indeed, in the results below, the more misspecified the model, the greater the AICPE error. This suggests that the AICPE error is a useful metric for identifying misspecification. While, as discussed in Section 4, we cannot slice by the number of ads on a user path, there are some other features (discussed there) that we can slice on. The resulting sliced AICPE can also be useful for identifying model misspecification, particularly when we allow the ad effect to vary with user or ad features.

Estimating the AICPE error does require data from an experiment where ads are artificially suppressed for some users, which can be costly. Once the overall model structure is validated, the frequency with which experiments need to be repeated depends on how stable the overall model structure is.

C.4.1 Result Summary

Overall, as the number of piece-wise constant intervals used for the ad effect increases, coverage for both the AICPE error and the target model coefficients improves. We also see that models with misspecified breakpoints do worse than models with the correct breakpoints, even when the total number of intervals is similar. These results support the conclusion that AICPE becomes a worse estimate of ICPE as the severity of model misspecification increases.

[6] As we will see in the detailed simulation results, this does not necessarily translate into a uniformly worse model. Improved estimation of the ad effect in $[0, 0.5]$ and well as the occurrence frequency of intervals with multiple active ad effects will also affect the comparison.

Table 13: Estimated model coefficients and naive targets for Scenario 5.

		Target	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
Baseline (per day) $[\exp(\alpha_0)]$		0.0333	0.0335 [0.0334, 0.0336]	41.6%
Ad Type 1	Short term $[\exp(\gamma_1)]$	2.25	2.271 [2.261, 2.280]	31.8%
	Long term $[\exp(\gamma_2)]$	1.2	1.189 [1.184, 1.193]	38.6%
Ad Type 2	Short term $[\exp(\gamma_3)]$	1.8	1.811 [1.803, 1.820]	14.6%
	Long term $[\exp(\gamma_4)]$	1.0	0.998 [0.996, 1.003]	81.4%

Table 14: Coverage for AICPE, ICPE and their errors for Scenario 5.

	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
Additive Error $\times\, 1e-2$	-0.55 $[-0.80, -0.28]$	57.4%
Relative Error $\times\, 1e-2$	-2.06 $[-3.01, -1.06]$	56.8%

C.4.2 Scenario 5 Results

In this scenario, we fit a short-term effect that lasts 4 days and a long-term effect that lasts for the remaining 26 days. Formally, the model is

$$\begin{aligned}\log(\lambda(t)) = \alpha_0 &+ \gamma_1 I\{0 < t - t_{\text{ad type 1}} \leqslant 4\} + \gamma_2 I\{4 < t - t_{\text{ad type 1}} \leqslant 30\} \\ &+ \gamma_3 I\{0 < t - t_{\text{ad type 2}} \leqslant 4\} + \gamma_4 I\{4 < t - t_{\text{ad type 2}} \leqslant 30\}.\end{aligned}$$

The coeffcient estimates are in Table 13. As discussed at the beginning of the section, the coverage refers to the coverage for the time-weighted averages of the simulated ground truth ad effects ("targets"). Table 14 gives the results for the AICPE error as well. The coverage is quite poor for both sets of results. That said, while the coverage is poor, the absolute difference between the example estimates and the truth is not huge.

C.4.3 Scenario 6 Results

In this scenario, we fit a short-term effect that lasts 2 days, a medium-term effect that lasts 2 days, and a long-term effect that lasts for the remaining 26 days. Formally, the model is

$$\begin{aligned}\log(\lambda(t)) = \alpha_0 &+ \gamma_1 I\{0 < t - t_{\text{ad type 1}} \leqslant 2\} + \gamma_2 I\{2 < t - t_{\text{ad type 1}} \leqslant 4\} + \gamma_3 I\{4 < t - t_{\text{ad type 1}} \leqslant 30\} \\ &+ \gamma_4 I\{0 < t - t_{\text{ad type 2}} \leqslant 2\} + \gamma_5 I\{2 < t - t_{\text{ad type 2}} \leqslant 4\} + \gamma_6 I\{4 < t - t_{\text{ad type 2}} \leqslant 30\}.\end{aligned}$$

Table 15: Estimated model coefficients and naive targets for Scenario 6.

		Target	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
Baseline (per day) [$\exp(\alpha_0)$]		0.0333	0.0333 [0.0333, 0.0335]	92.0%
Ad Type 1	Short term [$\exp(\gamma_1)$]	2.75	2.761 [2.747, 2.775]	94.6%
	Medium term [$\exp(\gamma_2)$]	1.75	1.758 [1.747, 1.768]	92.0%
	Long term [$\exp(\gamma_3)$]	1.2	1.193 [1.189, 1.198]	92.6%
Ad Type 2	Short term [$\exp(\gamma_4)$]	2.25	2.252 [2.242, 2.264]	93.6%
	Medium term [$\exp(\gamma_5)$]	1.35	1.346 [1.338, 1.355]	93.4%
	Long term [$\exp(\gamma_6)$]	1.0	1.000 [1.000, 1.000]	95.6%

Table 16: Coverage for AICPE, ICPE and their errors for Scenario 6.

	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
Additive Error $\times 1e-2$	-0.34 [$-0.59, -0.11$]	93.2%
Relative Error $\times 1e-2$	-1.29 [$-2.20, -0.40$]	93.0%

The coeffcient estimates are in Table 15 and the attribution results are in Table 16. As discussed at the beginning of the section, the coverage refers to the coverage for the time-weighted averages of the simulated ground truth ad effects ("targets"). The coverage, while not quite at 95%, is fairly close and is much improved relative to Scenario 5. This suggests that the model can give reasonable results when the misspecification is not too severe.

C.4.4 Scenario 7 Results

In this scenario, we fit a short-term effect that lasts 0.5 days, a medium-term effect that lasts 4 days, and a long-term effect that lasts for the remaining 25.5 days. Formally, the model is

$$
\begin{aligned}
\log(\lambda(t)) = \alpha_0 &+ \gamma_1 I\{0 < t - t_{\text{ad type 1}} \leqslant 0.5\} + \gamma_2 I\{0.5 < t - t_{\text{ad type 1}} \leqslant 4.5\} \\
&+ \gamma_3 I\{4.5 < t - t_{\text{ad type 1}} \leqslant 30\} \\
&+ \gamma_4 I\{0 < t - t_{\text{ad type 2}} \leqslant 0.5\} + \gamma_5 I\{0.5 < t - t_{\text{ad type 2}} \leqslant 4.5\} \\
&+ \gamma_6 I\{4.5 < t - t_{\text{ad type 2}} \leqslant 30\}.
\end{aligned}
$$

Table 17: Estimated model coefficients and naive targets for Scenario 7.

		Target	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
Baseline (per day) [exp(α_0)]		0.0333	0.0335 [0.0334, 0.0336]	52.4%
Ad Type 1	Short term [exp(γ_1)]	3	2.997 [2.969, 3.023]	88.6%
	Medium term [exp(γ_2)]	2.025	2.046 [2.036, 2.055]	12.0%
	Long term [exp(γ_3)]	1.2	1.190 [1.184, 1.194]	48.0%
Ad Type 2	Short term [exp(γ_4)]	2.5	2.509 [2.486, 2.531]	92.2%
	Medium term [exp(γ_5)]	1.6125	1.623 [1.616, 1.632]	6.2%
	Long term [exp(γ_6)]	1.0	1.000 [1.000, 1.007]	88.2%

Table 18: Coverage for AICPE, ICPE and their errors for Scenario 7.

	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
Additive Error $\times\ 1e-2$	-0.49 [$-0.71, -0.19$]	67.4%
Relative Error $\times\ 1e-2$	-1.83 [$-2.68, -0.74$]	66.4%

The coeffcient estimates are in Table 17 and the attribution results are in Table 18. As discussed at the beginning of the section, the coverage refers to the coverage for the time-weighted averages of the simulated ground truth ad effects ("targets").

The ground truth ad effect is constant in $[0, 0.5]$ and so it is unsurprising that the model generally does a good job of estimating γ_1 and γ_4. Most of the error due to the misspecification is in the estimate of the medium term effects, γ_2 and γ_5. The AICPE error covers 0 at a slightly higher rate than in Scenario 5, presumably due to the improved estimate of the short-term effect, as well as the fact that here we divide the ad effect into 3 pieces, while Scenario 5 only divides it into 2 pieces.

C.4.5 Scenario 8 Results

In this scenario, we fit a short-term effect that lasts 0.5 days, a first medium-term effect that lasts 2 days, a second medium-term effect that lasts 3 days, and a long-term effect that lasts for

Table 19: Estimated model coefficients and naive targets for Scenario 8.

		Target	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
Baseline (per day) [$\exp(\alpha_0)$]		0.0333	0.0334 [0.0333, 0.0335]	91.0%
Ad Type 1	Short term [$\exp(\gamma_1)$]	3	3.003 [2.975, 3.029]	93.4%
	Medium term 1 [$\exp(\gamma_2)$]	2.5	2.515 [2.501, 2.527]	90.2%
	Medium term 2 [$\exp(\gamma_3)$]	1.433	1.440 [1.432, 1.449]	67.8%
	Long term [$\exp(\gamma_4)$]	1.2	1.193 [1.184, 1.194]	89.0%
Ad Type 2	Short term [$\exp(\gamma_5)$]	2.5	2.508 [2.485, 2.529]	93.4%
	Medium term 1 [$\exp(\gamma_6)$]	2	1.998 [1.988, 2.009]	86.0%
	Medium term 2 [$\exp(\gamma_7)$]	1.15	1.151 [1.145, 1.158]	83.4%
	Long term [$\exp(\gamma_8)$]	1.0	1.000 [1.000, 1.005]	95.4%

the remaining 24.5 days. Formally, the model is

$$\begin{aligned}\log(\lambda(t)) = \alpha_0 &+ \gamma_1 I\{0 < t - t_{\text{ad type 1}} \leqslant 0.5\} + \gamma_2 I\{0.5 < t - t_{\text{ad type 1}} \leqslant 2.5\} \\ &+ \gamma_3 I\{2.5 < t - t_{\text{ad type 1}} \leqslant 5.5\} + \gamma_4 I\{5.5 < t - t_{\text{ad type 1}} \leqslant 30\} \\ &+ \gamma_5 I\{0 < t - t_{\text{ad type 2}} \leqslant 0.5\} + \gamma_6 I\{0.5 < t - t_{\text{ad type 2}} \leqslant 2.5\} \\ &+ \gamma_7 I\{2.5 < t - t_{\text{ad type 2}} \leqslant 5.5\} + \gamma_8 I\{5.5 < t - t_{\text{ad type 2}} \leqslant 30\}.\end{aligned}$$

The coeffcient estimates are in Table 19 and the attribution results are in Table 20. As discussed at the beginning of the section, the coverage refers to the coverage for the time-weighted averages of the simulated ground truth ad effects ("targets").

As in Scenario 7, the ground truth ad effect is constant in $[0, 0.5]$ and so the model does a good job of estimating γ_1 and γ_5. Most of the error due to the misspecification is in the estimate of the four medium term effects, $\gamma_2, \gamma_3, \gamma_6$ and γ_7. The AICPE error covers 0 at a much higher rate than in Scenario 7, but a slightly lower rate than Scenario 6. This suggests that, as in Scenario 6, the increased number of pieces for the ad effect improves the estimation (by decreasing the severity of the misspecification), but that the incorrect breakpoints do have an effect.

Table 20: Coverage for AICPE, ICPE and their errors for Scenario 8.

	Example Estimate [Example CI]	95% CI Coverage (over 500 datasets)
Additive Error $\times 1e-2$	-0.37 $[-0.60, -0.10]$	92.2%
Relative Error $\times 1e-2$	-1.37 $[-2.25, -0.39]$	91.4%