Chasing the Golden Forecast: Seasonality and Social Media in Forecasting Daily Movie Ticket Sales

By Thomas Li *

We use continuous-time duration models to produce a data-driven analysis of moviegoer behavior surrounding the 2023 film Wonka. *In particular, we examine the plausibility of duration dependence, zero-inflation, and heterogeneity in consumers' propensities to watch the film. We find that a zero-inflated exponential model provides the best base model, upon which we further incorporate various calendar and social media effects. Our findings indicate that including calendar effects significantly improves the model's predictive accuracy, providing valuable insights for marketing strategies in the film industry. The relevance of social media on ticket sales, however, remains an open question.*

Understanding the factors that drive daily movie ticket sales is crucial for strategic managerial decision making in the film industry. This paper seeks to address how calendar effects and social media buzz can be used to model moviegoer behavior. We explore several research questions:

- How do calendar effects such as the day of week and holidays impact ticket sales?
- To what extent can social media activity predict daily ticket sales?

Our analysis utilizes a dataset on ticket sales from December 15, 2023 to March 15, 2024 for the 2023 film *Wonka* taken from a representative sample of 50,000 active moviegoers. This dataset allows us to construct a nuanced model that integrates these covariates, providing a deeper understanding of the dynamics at play in movie ticket purchasing behavior.

I. Base Model Selection

A good base model provides the foundation of our story of consumer behavior. Accordingly, we explore several families of continuous-time duration models before incorporating covariates. In particular, we consider the Pareto II (EG), Burr XII (WG), Exponential (E), and Weibull (W), as well as a few zero-inflated (ZI) and 2-segment latent class (2-seg) variations on these classes of models. The exponential class (EG, E) models the time of initial adoption as an Exponential(λ) distribution whereas the Weibull class (WG, W) allows for the possibility of duration dependence in the time of adoption. Heterogeneity is examined through either a Gamma mixing distribution or latent-class segmentation, and the possibility of "hardcore-never-Wonkas" (HCNWs) who will never watch the movie is ascertained through zero-inflation. By assessing all these models, we can deduce whether

* University of Pennsylvania (email: thomli@sas.upenn.edu).

there is duration dependence and HCNWs in addition to the degree of heterogeneity in the behavior of moviegoers.

To determine the most appropriate model, we consider model fit as described by the Bayesian Information Criteria (BIC), mean absolute percent error (MAPE), and root mean square error (RMSE). BIC appropriately balances fit and parsimony with penalties for excessive parameters while MAPE and RMSE provide interpretable measures of model fit. Given that the number of incremental adopters in later periods become small, percent error may be high even if the fit is decent (e.g., expecting 3 when there are 4 actual adopters yields a percent error of 25%). As such, RMSE may be a more useful metric. To assess the forecasting abilities of the models, we fit to data for the first seven weeks (December 15, 2023 to February 1, 2024), thus creating a holdout set of the last six weeks (February 2, 2024 to March 15, 2024). Table 1 presents the parameter estimates and goodness-of-fit metrics for the eight base models we consider. Figure 1 displays compares tracking plots for the EG and zero-inflated exponential (ZIE).

Table 1—Estimated parameters and fit metrics: Base Models

	EG	ZIEG	WG	ZIWG	2-seg E	ZIE	ZI 2-seg E	ZIW
π_{ZI}	–	0.684	–	0.685	–	0.684	0.684	0.685
r	0.183	29667	0.167	4552	–	–	–	–
α	8.233	517280	8.007	81822	–	–	–	–
c	–	–	1.046	1.013	–	–	–	1.013
λ_1	–	–	–	–	< 0.001	0.057	0.057	0.056
λ_2	–	–	–	–	0.057	–	0.057	–
π_1	–	–	–	–	0.684	–	0.600	–
LL	−84528	−84126	−84524	−84126	−84126	−84126	−84126	−84125
BIC	169077	168285	169081	168294	168285	168274	168296	168283
MAPE-IS	113.47%	81.63%	112.45%	80.93%	81.63%	81.63%	81.63%	80.92%
MAPE-OOS	439.61%	54.28%	428.34%	52.40%	54.28%	54.27%	54.27%	52.37%
RMSE-IS	162.10	156.08	163.57	156.84	156.08	156.08	156.08	156.83
RMSE-OOS	53.12	37.77	51.87	38.35	37.77	37.77	37.77	38.36
parameter		π_{ZI}	c	c				c
compared to		EG	EG	ZI EG				ZI E
χ^2 LRT		803.00	7.63	1.79				1.897
p-value		< 0.001	0.006	0.181				0.168

Note: The following abbreviations are used: E (exponential), W (Weibull), G (gamma), ZI (zero-inflated), 2-seg (latent class with 2 segments). Table 1 reports maximum likelihood estimates of parameters for each of eight base model candidates. For zero-inflated models, π_{ZI} is the proportion who are hardcore-never-Wonkas. For gamma mixtures (EG, WG), r, α respectively refer to the shape and scale parameters. For Weibull models, c refers to the duration dependence shape parameter. For pure exponential and Weibull models, λ_i refers to the rate parameter of segment i (where $i = 1$ for single-segment models). For 2-segment models, π_1 refers to the proportion with λ_1. Log likelihood (LL) and the Bayesian Information Criteria (BIC) are reported for all models alongside in-sample (IS) and out-of-sample (OOS) MAPE and RMSE. Likelihood ratio tests (LRT) are performed for select parameters as indicated (df=1).

The parameter estimates offer three takeaways: there is (1) clear evidence for zero-

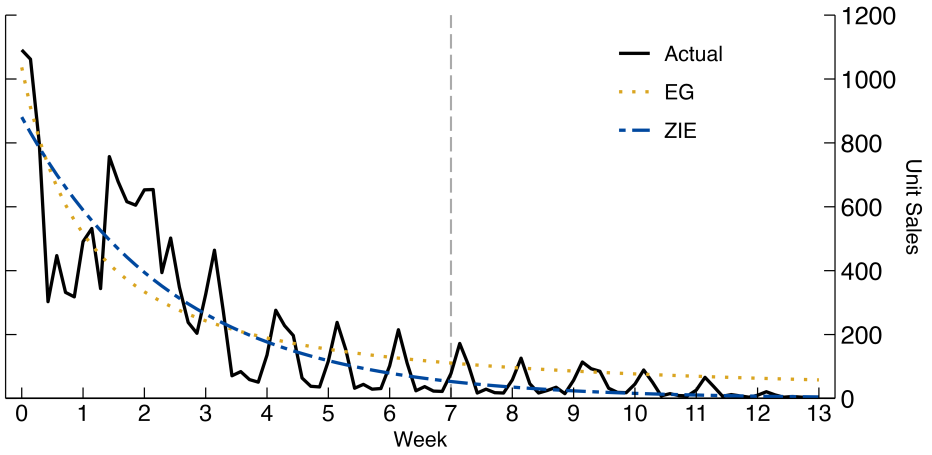

FIGURE 1. TRACKING PLOTS: EG AND ZIE

Note: Figure 1 displays the tracking plots for the EG and ZIE models compared to the actual data. The vertical grey dashed line at Week 7 denotes the split between training data (Weeks 0–7) and holdout data (Weeks 7–13).

inflation, (2) minimal—if not a lack of—duration dependence, and (3) no heterogeneity after accounting for HCNWs. The zero-inflated models all have the lowest BIC, MAPE, and RMSE, and an LRT between ZIEG and EG confirms a drastically better fit from including a spike at zero. Figure 1 visually confirms the improvement. Even the 2-segment exponential drives λ_1 to 0 with a segment proportion π_1 nearly identical to other estimates of the proportion π_{ZI} of HCNWs. The robustness of the estimate of π_{ZI} to the model specification provides strong evidence that 68.4% of moviegoers are HCNWs in our training set. In assessing the plausibility of duration dependence, we note that LRTs of Weibull models with their exponential counterparts yield insignificant p-values (the significance in the WG-EG comparison disappears once HCNWs are factored in) and the BIC penalizes the additional parameter. Turning to the nature of heterogeneity, ZIEG and ZIWG yield exceedingly large estimates of r, α, suggesting homogeneity to the point where λ is almost concentrated at a point. In fact, these estimates respectively yield mean lambda values of 0.057 and 0.056 with variances < 0.001—precisely the estimates in the ZIE and ZIW models, with the negligible variance reflected in the fact that $\lambda_1 = \lambda_2$ in the ZI 2-seg E. The similar fit produced by the ZIE and ZIW with fewer parameters is confirmed by the lower BICs and unchanged MAPEs and RMSEs. Altogether, the zero-inflation with no duration dependence and homogeneity is best captured by the zero-inflated exponential (ZIE), which will serve as the base model for the remainder of our analysis.

Before any covariates have been considered, we already have a partially-formed story of consumer behavior. According to the ZIE, π_{ZI} of moviegoers will never watch *Wonka*—a reasonable assumption given genre preferences and film taste. The remaining $1 - \pi_{ZI}$ watch the movie for the first time as given by an exponential distribution with parameter λ, suggesting a mean watch date of $\frac{1}{\lambda}$ days after release. The estimated values of π_{ZI} and

λ will be nailed down in Section 4, once covariates are incorporated.

II. Calendar Effects

With a base model in hand, we turn to incorporating the first set of covariates: calendar effects. Intuition and a visual analysis of the data (see Figure 1) suggest a weekly cyclicality with additional effects around holidays. Experience indicates that people are more likely to go to movie theaters on weekends and holidays. Accordingly, we assess the inclusion of (1) weekly seasonality via dummy variables for the day of week, (2) holiday effects via dummy variables for periods around federal holidays, (3) interactions between weekends and holidays, and (4) a one-time effect on the premiere date.

Each covariate is supported by a reasonable story. People are more free on weekends and thus more likely to go to a theater on those days—such a phenomenon would be reflected in larger positive values on ω_6, ω_7, the coefficients on dummy variables for Saturday and Sunday. To avoid multicollinearity, we drop Monday dummies such that the other ω_k's are expressed relative to a baseline set by Monday. Differential effects on holidays are similarly plausible due to spare time or—in the case of Valentine's Day—a special occasion, as corroborated by box office data showing higher revenues around holidays (Box Office Mojo, 2023). To that end, we include a dummy variable indicating Christmas and New Year's break (December 23 to January 1[1]), MLK weekend (January 13–15), Valentine's Day (February 14), and President's Day Weekend (February 17–19). To assess the possibility that the days around Christmas and New Year's exhibit different effects, we also consider separate indicator variables for Christmas/New Year's and all other holidays. Due to longer breaks, reunited families, and the air of holiday festivities, moviegoers may behave differently around Christmas than a typical long weekend (Whitten, 2019). In theory, both Christmas (η_1) and holiday (η_2) effects should be positive, with η_1 predicted to exhibit a larger effect. An additional concern is that the effect of holidays may overlap with the effect of weekends in a non-linear manner—namely, the effect of a holiday may differ for weekends than weekdays. We address this complication by adding an interaction term η_3 between weekends and holidays, expecting η_3 to be negative: since people already tend to go to the theater on regular weekends, the incremental effect of a holiday should be smaller for a holiday weekend than a holiday weekday. Finally, we consider the distinctiveness of the premiere date (December 15). The first showing of the film may draw additional attention that we would expect to manifest in a positive coefficient α. To assess all these factors, we fit Models 1–5 to consider different combinations of these calendar effects and present results in Table 2 and Figure 2.

We have evidence for the significance of each of the factors discussed above. Both the BIC and LRT confirm that the addition of weekly seasonality, separate holiday and Christmas effects, holiday-weekend interactions, and the premiere each improve model fit more than additional parameters are penalized. Moreover, the addition of these factors improve both in-sample and out-of-sample RMSE and MAPE, bringing the OOS RMSE (respectively

[1]We defined this period ex ante (prior to fitting models) considering that December 23 is a Saturday whereas January 2 is a Tuesday when many may return to work.

TABLE 2—ESTIMATED PARAMETERS AND FIT METRICS: ZIE WITH CALENDAR EFFECTS

Model	1	2	3	4	5
π_{ZI}	0.681	0.676	0.674	0.676	0.672
λ	0.050	0.039	0.029	0.029	0.027
ω_2 (Tue)	−0.053	0.037	0.149	0.176	0.192
ω_3 (Wed)	−0.253	−0.162	−0.046	−0.019	−0.004
ω_4 (Thur)	−0.266	−0.173	−0.054	−0.025	−0.011
ω_5 (Fri)	0.239	0.378	0.550	0.571	0.441
ω_6 (Sat)	0.477	0.481	1.082	1.102	1.153
ω_7 (Sun)	0.145	0.146	0.748	0.769	0.817
η_1 (Holiday)	–	0.386	0.842	1.102	1.129
η_2 (Christmas)	–	–	–	0.824	0.907
η_3 (Weekend×Holiday)	–	–	−1.081	−1.115	−1.181
α (Premiere)	–	–	–	–	0.508
LL	−83666	−83432	−82945	−82926	−82852
BIC	167418	166962	165998	165971	165834
MAPE-IS	61.44%	55.85%	38.99%	36.89%	36.45%
MAPE-OOS	46.41%	53.33%	43.70%	37.94%	46.08%
RMSE-IS	135.19	129.33	92.50	91.81	74.11
RMSE-OOS	31.01	28.52	22.99	24.25	21.15
parameter	$\omega_2, \ldots, \omega_7$	η_1	η_3		α
compared to	ZIE	1	2		4
χ^2 LRT	920.92	467.04	974.70		147.82
p-value	< 0.001	< 0.001	< 0.001		< 0.001

Note: Table 2 reports maximum likelihood estimates of parameters and goodness-of-fit statistics for five different specifications of calendar effects. $\omega_2, \ldots, \omega_7$ capture weekly cyclicality. η_1 captures holiday effects, where holidays are defined as all holidays in Models 1–3 and only non-Christmas/New Year holidays in Models 4–5. η_2 captures Christmas/New Year holiday effects. η_3 captures an interaction effect between weekends and any holiday. α captures the effect of the premiere day. The LRTs are performed with df=1 for Models 2,3,5 and df=6 for Model 1.

MAPE) down from 37.77 (54.27%) to 21.15 (46.08%)—a substantial improvement that is visually confirmed by Figure 2. We further note that the estimate of π_{ZI} has remained relatively consistent around 0.68, bolstering the robustness of evidence for HCNWs. The estimate of λ decreases somewhat as more calendar effects are added, suggesting that the base model flattens somewhat after adjusting for calendar covariates. On a technical note, some ω_k's are near zero and may not be individually significant, but we keep all ω_k's together due to the robustness of the story of weekly cyclicality and the joint significance of the coefficients. Keeping the entire set of ω_k's also allows us to isolate the weekly cyclicality, as in Figure 6 (see Appendix for details).

For the most part, our predictions match up with the empirical findings. There is a weekly surge in ticket sales from Friday to Sunday with a spike on Saturday (see Section 5 for a graph and more detailed analysis of the weekly seasonal structure). There is also a large positive effect on holidays with a negative attenuation on holiday weekends relative to holiday weekdays, as expected and explained previously. Moreover, the premiere exhibits

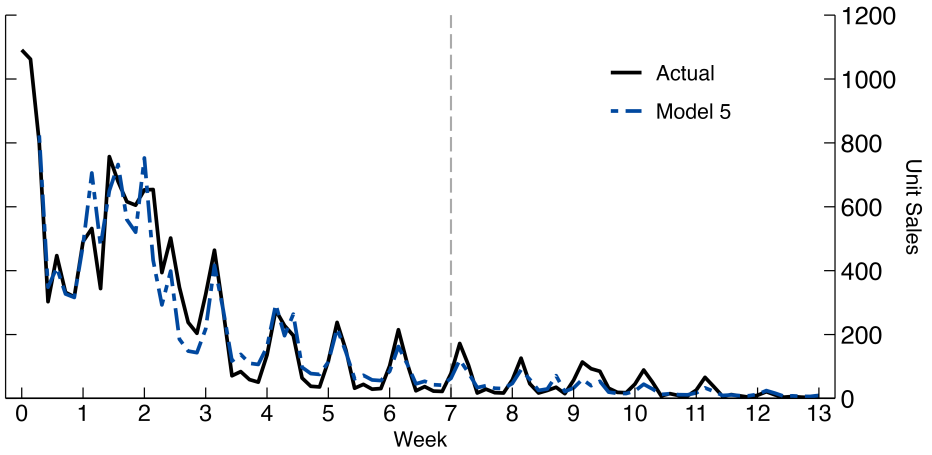

FIGURE 2. TRACKING PLOT: MODEL 5 (ZIE WITH ALL CALENDAR EFFECTS)

Note: Figure 2 displays the tracking plot for the ZIE with all calendar effects (Model 5) compared to the actual data. The vertical grey dashed line at Week 7 denotes the split between training data (Weeks 0–7) and holdout data (Weeks 7–13).

a positive effect on sales, in accordance with our previously stated belief in the appeal of being among the first to watch a movie. One unexpected result is that the Christmas season exhibits a somewhat smaller effect than other holidays. This could be explained by either the longer duration of the Christmas-New Year break wherein additional ticket sales are more spread out over the break (rather than concentrated in one or two days as may be the case on a long weekend) or people opting to spend the break in other ways such as traveling. Nevertheless, we have evidence that the effect of Christmas is different than that of regular holidays.

III. Social Media Effects

Previous literature in box office predictions has explored the predictive power of word-of-mouth or social media hype surrounding a film (Liu, 2006; Lehrer and Xie, 2022; Liao et al., 2022). These studies conclude that social media can improve predictions; they, however, seek to predict overall box office success rather than granular daily ticket sales. To examine the plausibility that these results may extend to daily ticket sales, we assess the incorporation of several measures of social media volume surrounding the *Wonka* film.

We obtain data from Keyhole, a social media analytics firm, on the daily number of social media posts using the hashtag "#WonkaMovie" across platforms including Instagram and Twitter/X from December 8, 2023 to March 15, 2024. The data may not be of the best quality, given that counts are rounded to the nearest 100 and some days have counts of 0 (perhaps due to a limited domain from which Keyhole scrapes these data). Nevertheless, it may serve as an adequate proxy for social media volume around the film. A sample of the data is provided in Table 3 and a plot is presented in Figure 3.

TABLE 3—SAMPLE OF SOCIAL MEDIA DATA

Date	12/8/23	12/9/23	12/10/23	...	3/13/24	3/14/24	3/15/24
Number of Posts	800	1700	1200	...	0	100	100

Source: Keyhole.

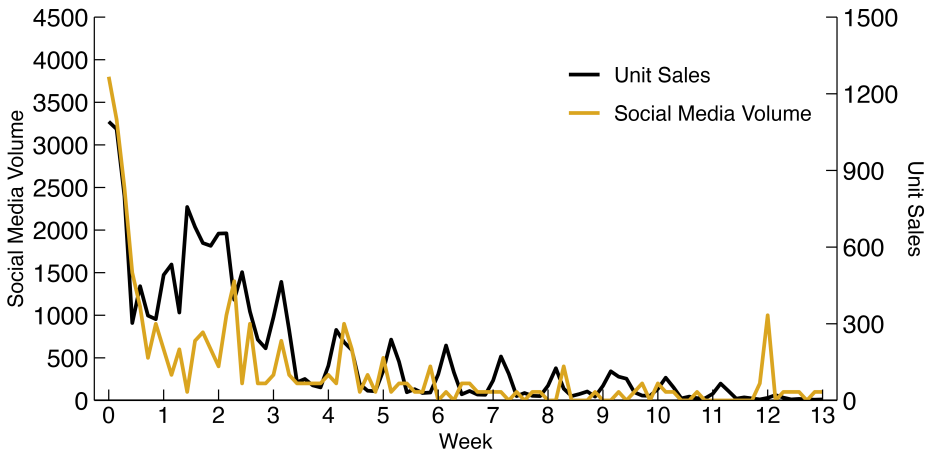

FIGURE 3. SOCIAL MEDIA DATA AND UNIT TICKET SALES

Note: Figure 3 displays the time series of *Wonka* ticket sales and social media volume. The former is plotted in black and uses the right axis scale. The latter is plotted in yellow and uses the left axis scale.

From these data, we produce two measures of social media volume for each day t with social media count s_t: (1) a rolling sum of posts over the previous week ($x_t = \sum_{k=t-7}^{t-1} s_k$) and (2) a lagged count $x_t = s_{t-1}$. The former operates under the story that social media posts can influence moviegoer behavior for up to seven days (plausible for "planners" who make plans a few days in advance) whereas the latter only lets posts affect sales for the next day (more spontaneous behavior). Given the generally positive reviews of *Wonka*, greater social media volume should in theory increase unit sales. By only using lags rather than contemporaneous social media effects, we attempt to sidestep potential concerns of endogeneity: a ticket sale in time t may have reverse causality on social media volume in time t as people may post after watching the movie, but it is less likely that a ticket sale in time t affects social media volume in time $t-1$. This strategy of using lagged explanatory variables to avoid endogeneity has been previously used in econometric literature (Green, Malpezzi and Mayo, 2005; Aschhoff and Schmidt, 2008), but we also note that some scholars are concerned with whether it truly removes endogeneity (Bellemare, Masaki and Pepinsky, 2017). On top of the two suggested measures, we further explore quadratic terms for each variable definition in case the effect is not linear. Results for these four models (Models 6–9) are presented in Table 4 and Figure 4.

Table 4—Estimated parameters and fit metrics: ZIE with calendar and social media effects

Model	6	7	8	9
π_{ZI}	0.690	0.691	0.684	0.681
λ	0.044	0.044	0.034	0.030
ω_2 (Tue)	0.127	0.119	0.102	0.190
ω_3 (Wed)	−0.072	−0.080	−0.062	−0.034
ω_4 (Thur)	−0.086	−0.088	−0.089	−0.029
ω_5 (Fri)	0.346	0.345	0.364	0.418
ω_6 (Sat)	1.104	1.108	1.263	1.398
ω_7 (Sun)	0.783	0.783	0.921	1.012
η_1 (Holiday)	0.918	0.897	1.150	1.164
η_2 (Christmas)	0.629	0.634	0.822	0.843
η_3 (Weekend×Holiday)	−1.114	−1.108	−1.362	−1.417
α (Premiere)	0.466	0.492	0.626	0.534
Social Media Predictor (X)	Past Week	Past Week	Past Day	Past Day
β_1 (X)	−0.148	−0.201	−0.080	0.067
β_2 (X^2)	–	0.016	–	−0.031
LL	−82787	−82786	−82814	−82793
BIC	165714	165724	165768	165737
MAPE-IS	33.20%	33.28%	31.35%	29.19%
MAPE-OOS	57.74%	60.60%	40.03%	33.45%
RMSE-IS	66.25	65.48	71.16	69.59
RMSE-OOS	35.55	36.25	28.29	25.59
parameter	β_1	β_2	β_1	β_2
compared to	5	6	5	8
χ^2 LRT	130.55	1.43	77.10	41.90
p-value	< 0.001	0.232	< 0.001	< 0.001

Note: Table 4 reports maximum likelihood estimates of parameters and goodness-of-fit statistics for four different specifications of social media effects added onto Model 5. Models 6 and 7 consider the volume of social media in the past week whereas Models 8 and 9 consider the volume in the previous day. A linear and quadratic effect is examined for each. All LRTs are performed with df=1.

First assessing the quadratic term, we observe that β_2 is insignificant in Model 7 but significant in Model 9 according to both BIC and LRT, so we compare Models 6 and 9. BIC suggests Model 6 is better whereas out-of-sample MAPE and RMSE provide more support for Model 9. The parameter estimates, however, raise some questions. The β_1 for Model 6 is negative, suggesting that greater social media volume decreases unit sales. Likewise, the coefficients on Model 9 suggest an positive effect up to a point, after which more posts hurt ticket sales. Given the positive reviews (e.g. 82% on Rotten Tomatoes), this result seems to run counter to the story of social media effects. Potential explanations could be the aforementioned poor data quality and a potential failure to eliminate endogeneity. Further studies could seek better data on social media volume to reassess the effect of social media. For this paper, considering the totality of evaluation metrics, we choose to discard social media effects in favor of a ZIE with all calendar covariates (Model 5), which has a well-founded story, a smaller out-of-sample RMSE, and better parsimony.

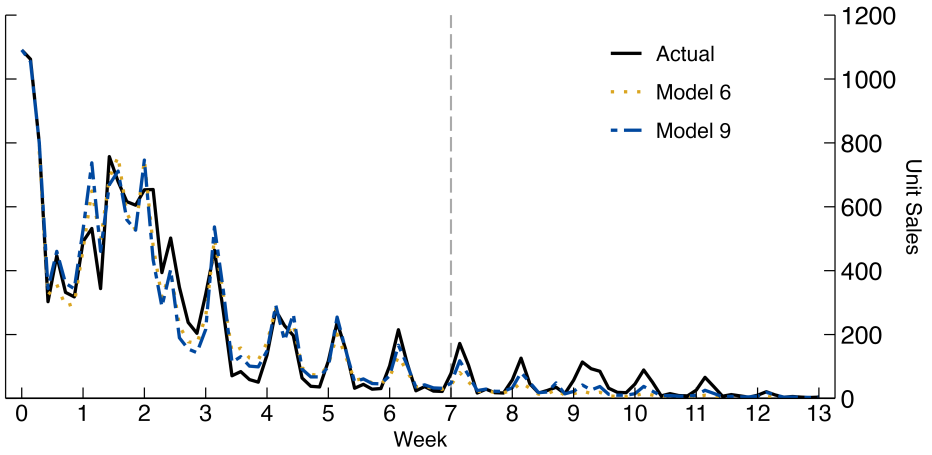

FIGURE 4. TRACKING PLOTS: MODELS 6 AND 9

Note: Figure 4 displays the tracking plot for Models 6 and 9 compared to the actual data. The vertical grey dashed line at Week 7 denotes the split between training data (Weeks 0–7) and holdout data (Weeks 7–13).

IV. Robustness

Building on Model 5, we briefly assess the robustness of parameter estimates by varying the estimation sample, as presented in Table 5.

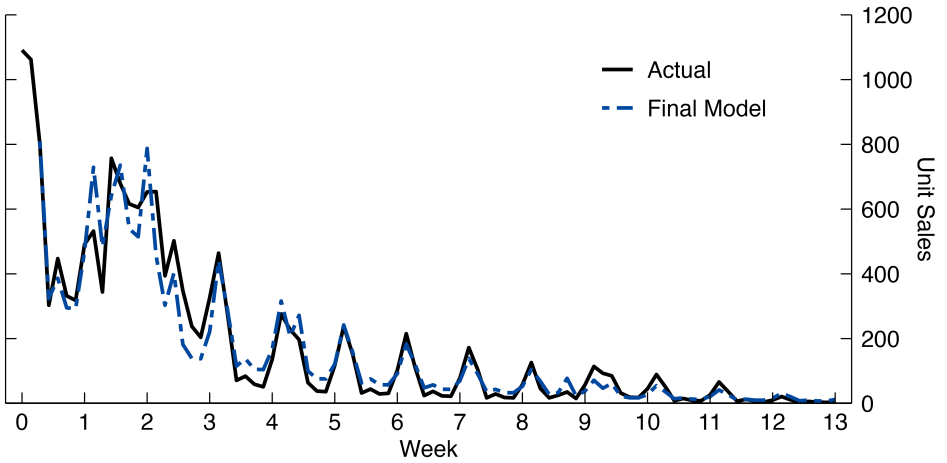

FIGURE 5. TRACKING PLOT: ZIE WITH CALENDAR EFFECTS, FITTED ON FULL SAMPLE

Note: Figure 5 displays the tracking plot for the final model, the ZIE with all calendar effects fitted on the full sample, compared to the actual data.

Table 5—Robustness Checks: ZIE with calendar effects

Estimation Sample	5 weeks	7 weeks	9 weeks	Full sample
π_{ZI}	0.660	0.672	0.671	0.665
λ	0.026	0.027	0.026	0.024
ω_2 (Tue)	0.177	0.192	0.192	0.204
ω_3 (Wed)	−0.011	−0.004	−0.028	−0.039
ω_4 (Thur)	−0.023	−0.011	−0.020	−0.025
ω_5 (Fri)	0.382	0.441	0.462	0.481
ω_6 (Sat)	1.068	1.153	1.200	1.232
ω_7 (Sun)	0.746	0.817	0.846	0.877
η_1 (Holiday)	0.984	1.129	1.084	1.155
η_2 (Christmas)	0.877	0.907	0.914	0.952
η_3 (Weekend×Holiday)	−1.112	−1.181	−1.208	−1.214
α (Premiere)	0.553	0.508	0.500	0.549
MAPE-IS	34.39%	36.45%	37.64%	45.26%
MAPE-OOS	67.85%	46.08%	45.93%	–
RMSE-IS	79.53	74.11	68.57	56.92
RMSE-OOS	21.50	21.15	20.24	–

Note: Table 4 reports maximum likelihood estimates of parameters and goodness-of-fit statistics for the ZIE with all calendar effects fitted on four different estimation samples.

All parameter estimates are stable across the estimation samples, strengthening the evidence for the robustness of parameter estimates in the full model. The final model fitted on the full sample is plotted in Figure 5.

V. Discussion & Conclusion

We conclude with a managerial analysis of the final model: the ZIE with all calendar effect fitted on the full sample. Given the representative sample, we can generalize conclusions to the broader moviegoer population.

As noted in Section 2, $\pi_{ZI} = 0.665$ suggests that 66.5% of active moviegoers will never watch *Wonka*, and the remaining 33.5% have a homogeneous propensity to watch the film given by an exponential distribution with parameter $\lambda = 0.024$. The mean time to watch is 41.7 days after release (around January 25). We also observe a weekly cycle driving ticket sales, which is plotted in Figure 6. The derivations of these day-of-week effects from the ω_k estimates is provided in the Appendix. There is low activity from Monday to Thursday but sales picking up on Friday and the weekend, with Saturday being the most popular day. There is a slight pick-up on Tuesdays, possibly due to many cinemas having "discount Tuesday" deals. Long weekends and holiday breaks are also popular days for watching movies. For a theater manager, these insights into consumer demand may prove valuable for ticket pricing and theater allocation.

Using our model, we can forecast sale in the future (Figure 7). It seems as if sales will be dying down in the coming weeks, with fewer than 5 sales expected from the initial sample of 50,000 consumers by the end of April. This kind of model can thus provide a

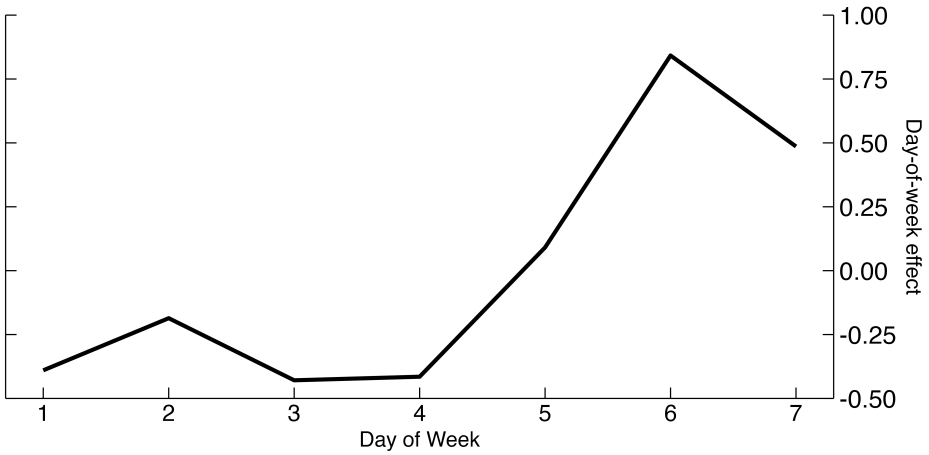

FIGURE 6. WEEKLY SEASONAL EFFECT

theater manager will key information on when to stop showing the film. Overall, this paper highlights the significant role that calendar effects play in shaping movie ticket sales, offering actionable insights for managers to optimize film marketing and distribution strategies effectively.

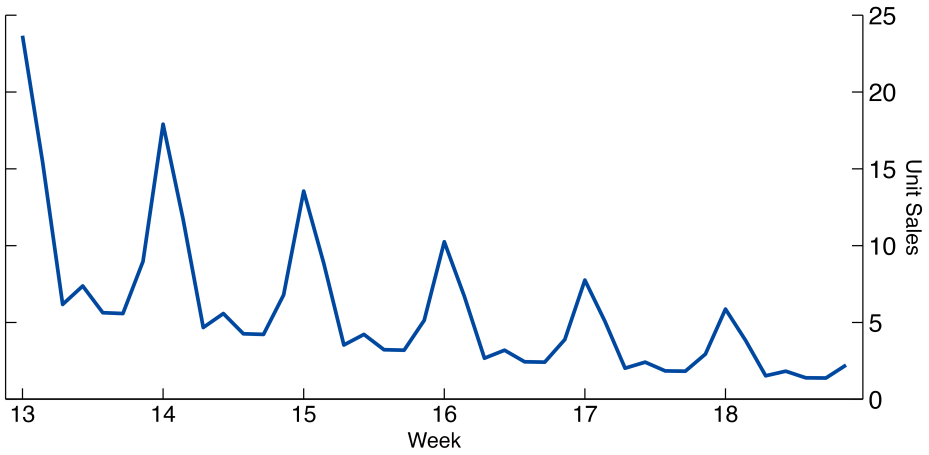

FIGURE 7. FORECASTED SALES: MARCH 16 TO APRIL 26

REFERENCES

Aschhoff, Birgit, and Tobias Schmidt. 2008. "Empirical evidence on the success of R&D cooperation—happy together?" *Review of Industrial Organization*, 33: 41–62.

Bellemare, Marc F, Takaaki Masaki, and Thomas B Pepinsky. 2017. "Lagged explanatory variables and the estimation of causal effect." *The Journal of Politics*, 79(3): 949–963.

Box Office Mojo. 2023. "2023 Holiday Box Office." IMDbPro.

Green, Richard K, Stephen Malpezzi, and Stephen K Mayo. 2005. "Metropolitan-specific estimates of the price elasticity of supply of housing, and their sources." *American Economic Review*, 95(2): 334–339.

Lehrer, Steven F, and Tian Xie. 2022. "The bigger picture: Combining econometrics with analytics improves forecasts of movie success." *Management Science*, 68(1): 189–210.

Liao, Yi, Yuxuan Peng, Songlin Shi, Victor Shi, and Xiaohong Yu. 2022. "Early box office prediction in China's film market based on a stacking fusion model." *Annals of Operations Research*, 1–18.

Liu, Yong. 2006. "Word of mouth for movies: Its dynamics and impact on box office revenue." *Journal of marketing*, 70(3): 74–89.

Whitten, Sarah. 2019. "This is the most important week of the year for movie theater owners." *CNBC*.

Appendix

A1. Derivation of day-of-week effects

Appendix A1 discusses how we back out the day-of-week effects in Figure 6 from the dummy coefficients $\omega_2, \ldots, \omega_7$. We have estimates of the coefficients on dummy variables for Tuesday (ω_2) to Sunday (ω_7). Recall that we exclude the dummy variable ω_1 for Monday to avoid multicollinearity. Hence, the effect of Monday is absorbed into the estimated baseline intercept term β_0 and all other dummy coefficients are relative to the Monday effect. In our proportional hazards regression framework, β_0 is used to establish the baseline propensity: $\lambda = \exp(\beta_0)$. Hence, $\beta_0 = \ln(\hat{\lambda})$, where $\hat{\lambda}$ is our estimate of the baseline exponential rate ($\hat{\lambda} = 0.024$ in the final model).

Let $\Omega_1, \ldots, \Omega_7$ respectively denote the day-of-week effects for Monday to Sunday, and let β_0^* denote the "true" intercept term less the effect of ω_1. Then, we have the system

$$\begin{cases} \beta_0 = \beta_0^* + \Omega_1 \\ \omega_2 = \Omega_2 - \Omega_1 \\ \omega_3 = \Omega_3 - \Omega_1 \\ \qquad \vdots \\ \omega_7 = \Omega_7 - \Omega_1 \end{cases}$$

Note that the left-hand side of the system equations are observed while the right-hand side contains the unobserved terms we wish to deduce. In our additive model of seasonality, $\Omega_1 + \Omega_2 + \cdots + \Omega_7 = 0$, which gives us 8 equations and 8 unknowns. To solve this system, we first add β_0 to $\omega_2, \ldots, \omega_7$ to obtain $\omega_k + \beta_0 = \beta_0^* + \Omega_k$ for $k = 2, \ldots, 7$. Taking the mean of $\beta_0, \omega_2 + \beta_0, \ldots, \omega_7 + \beta_0$ and applying $\Omega_1 + \Omega_2 + \cdots + \Omega_7 = 0$, we get

$$\frac{1}{7}\left[\beta_0 + \sum_{k=2}^{7}(\omega_k + \beta_0)\right] = \frac{\Omega_1 + \Omega_2 + \cdots + \Omega_7 + 7\beta_0^*}{7} = \frac{0 + 7\beta_0^*}{7} = \beta_0^*$$

All that remains is to the subtract β_0^* from $\beta_0, \omega_2 + \beta_0, \ldots, \omega_7 + \beta_0$. Our estimates of the day-of-week effects are thus given by

$$\begin{cases} \Omega_1 = \beta_0 - \beta_0^* \\ \Omega_2 = \omega_2 - \beta_0^* \\ \Omega_3 = \omega_3 - \beta_0^* \\ \qquad \vdots \\ \Omega_7 = \omega_7 - \beta_0^* \end{cases}$$