Analyzing Album Sales Using a Segmented Weibull Distribution with Leading Airplay Data

Executive Summary

The trend in sales of the Radiohead's album, *The Bends*, is strongly modeled by a segmented Weibull distribution with two covariate effects: the 5-week prior airplay data, and a Christmas seasonal effect. Additionally, the parameters derived from the model tells us the following story of the two customer groups in the population: Customer 1 is the higher-propensity buyer that is impacted somewhat heavily by the Christmas season; Customer 2 is the near non-zero buyer who is more impacted by radio airplay relative to Customer 1, and also exhibits positive duration dependence (the chance they buy an album increases the longer they have not bought it yet). We will discuss below if this story makes sense.

Background

We are analyzing the sales of Radiohead's breakthrough album, *The Bends*, given data on its domestic sales and radio airplay. *The Bends* is considered one of the greatest albums of all time, influencing many British bands after it and being named as Rolling Stone's 110th greatest album of all time. We want to forecast the domestic sales of the album for five weeks in March-April 1996 – seems simple enough; one clarification we must make is that for this analysis, we assume the airplay figures of the album were **preset** before the album's release, such that how the album is performing/its popularity does not impact how much it is played on the radio.

Covariate Talk

Immediately after looking at the data, it was somewhat apparent to me that the sales did not look very “exponential” like to me, as there are multiple weeks where sales are sharply increasing; I knew this behavior had to be captured by some combination of increasing duration dependence and covariate effects. The next thing I noticed was a surge in sales around Christmas season – I covered up this “peak” with my finger, noting that the sales curve seems to continue decently well through this section. This naturally brought up the question: “why do sales keep on increasing, even after Christmas?” Lastly, something I noticed was almost a lag effect, where for much of the year the behavior of sales closely mirrors the behavior of radio airplay **from weeks before**. Could that help explain why sales were increasing week to week?

The working story I came up was: at the core, the number of weeks it takes a customer to buy the album is exponentially distributed. Depending on the results of the models, there could be possible heterogeneity/duration dependence effects included to help define each individual’s lambda; however, the **crux** of our model lie in two covariate effects.

From above, I fleshed out the “Christmas effect” as a 4-week period from 12/17-1/7 where sales are positively impacted, and afterwards, sales return to pre-Christmas behavior; this is a clearly time-varying seasonal effect that happens every year. The story behind this is that consumers, for whatever reason, buy albums as gifts, inflating sales numbers for just the holiday season and leaving no long-term effect; I somewhat arbitrarily defined the period for the Xmas effect as the four weeks in which sales seemed to peak. Then, I defined the 5-week leading airplay to be the second covariate effect. The story here is that radio airplay, which is *preset* and independent of album sales, exposes listeners to the music and makes them want to buy the album. Why a 5-week lag, or any lag at all? I argue that intuitively, it takes time for songs people hear to “grow on them”, and that it does require weeks for listeners to catch on and get familiar with songs on the radio; I again somewhat arbitrarily chose 5 weeks, mostly by observing that the many of the peaks/troughs in the sales and airplay were separated by around five weeks. I also could make an argument for choosing 2 or 3 weeks, but at least for this album it does not seem to appear in the data as such. Here I again make the important clarification that sales and airplay are **not endogenous**, as airplay is fixed and is not at all causally impacted by sales. Finally, using the 5 weeks of leading airplay begs the question: what airplay data do you use for the first five weeks? I settled on using the average airplay data from our dataset (0 after standardization) as opposed to fitting the numbers – this seemed more reasonable to me than assuming/conjuring data that better suited my story.

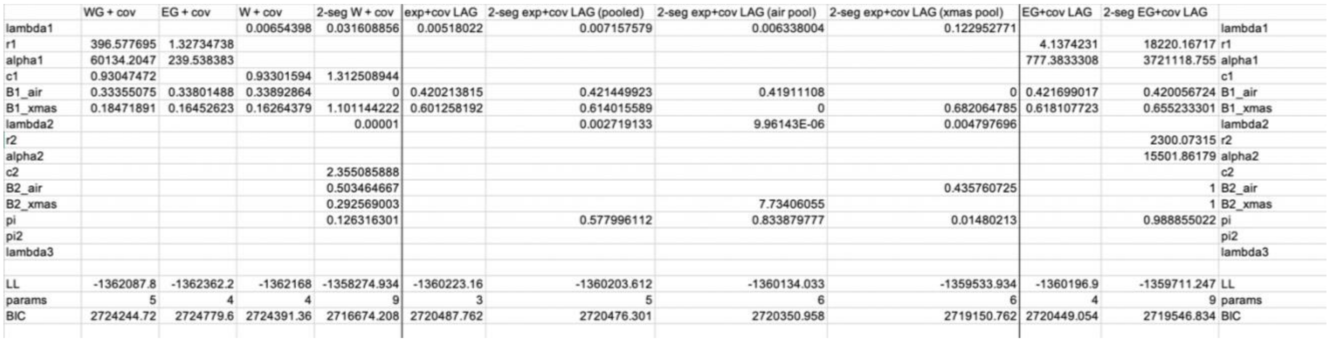
Both these covariate effects would be incorporated into our model using the concept of proportional hazards, with the airplay data standardized and the Christmas covariate set to 1 during the 4-week period. As a side note, I ultimately decided against bringing in external covariates, largely in part due to how difficult it would be to claim that outside effects are endogenous. This was however a last resort if simply using the provided data was insufficient to adequately fit our model.

Parameter Intuition

Before moving into fitting the models, we consider what we expect the derived parameters to look like. First, out of the 1 million households we assume to be in our population, only 220287 have bought the album through nearly a year. This implies a **low average lambda**, as we would expect the average household to take a long time to buy; additionally, when we run segmented models, we might expect one large segment of extremely low, near-zero lambdas (0 if we incorporate never-triers), and a smaller segment of moderate lambdas (around 0.05). It is hard to form the intuition for duration dependence, especially if the airplay data might be the factor causing increasing sales from week-to-week, but intuitively I predict $c > 1$, or negative duration dependence – it could be that if you have not bought it, you are even less likely to get it as time goes on, especially as it is played so much on radio. Finally, for our covariate beta parameters, it is almost a given that these parameters should be **positive**; I would also expect both parameters to be relatively significant. Let’s run our models!

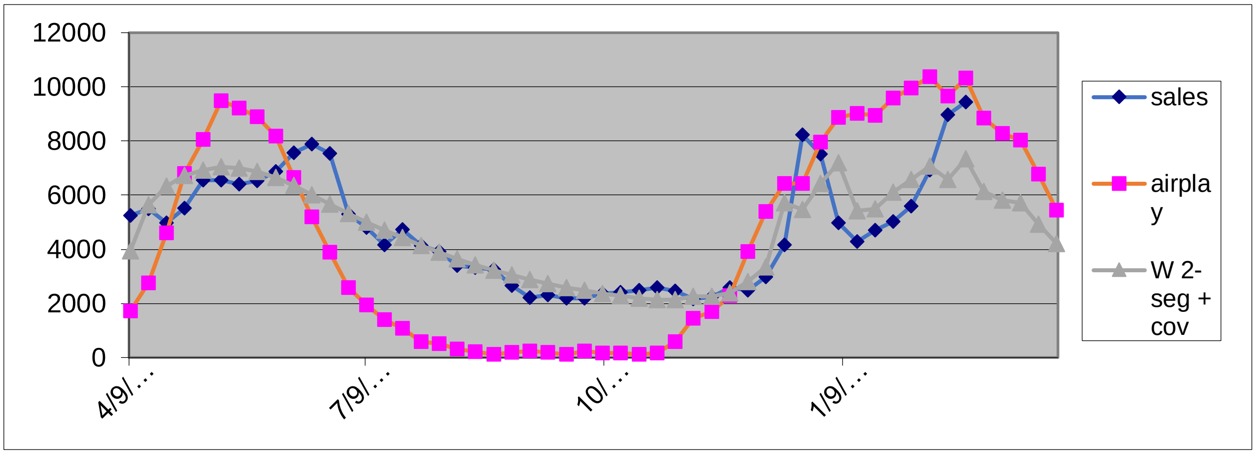
Models + Implications

Here is a summary of all the first half of models ran, with all parameter estimates, log-likelihoods, and BIC score.



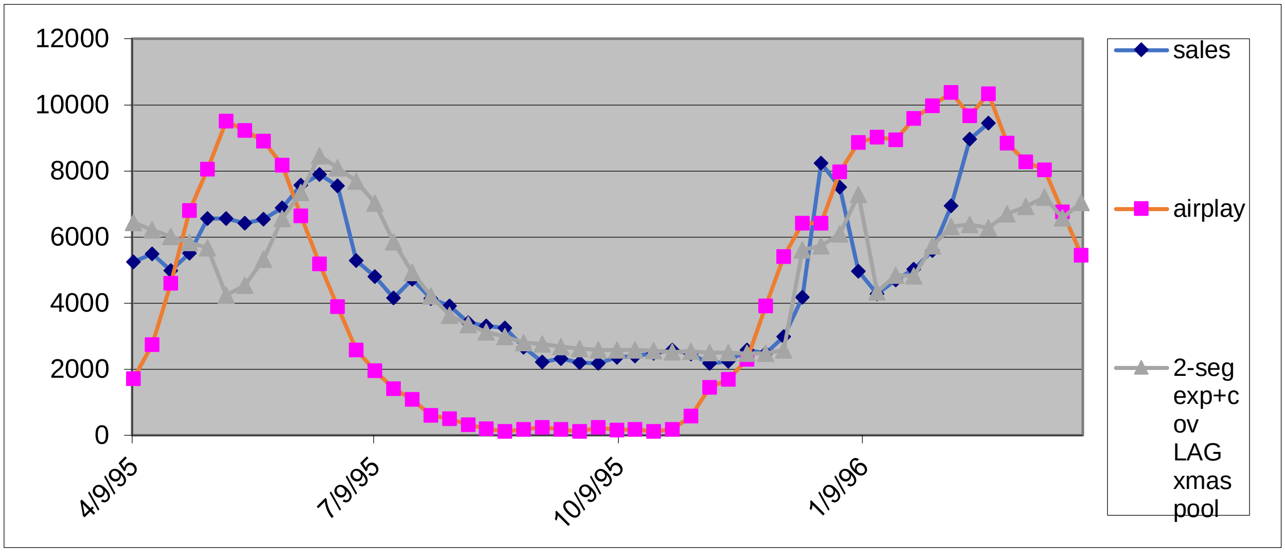
	WG + cov	EG + cov	W + cov	2-seg W + cov	exp+cov LAG	2-seg exp+cov LAG (pooled)	2-seg exp+cov LAG (air pool)	2-seg exp+cov LAG (xmas pool)	EG+cov LAG	2-seg EG+cov LAG	
lambda1			0.00654398	0.031608856	0.00518022	0.007157579	0.006338004	0.122952771			lambda1
r1	396.577695	1.32734738							4.1374231	18220.16717	r1
alpha1	60134.2047	239.538383							777.3833308	3721118.755	alpha1
c1	0.93047472		0.93301594	1.312508944							c1
B1_air	0.33355075	0.33801488	0.33892864	0	0.420213815	0.421449923	0.41911108	0	0.421699017	0.420056724	B1_air
B1_xmas	0.18471891	0.16452623	0.16264379	1.101144222	0.601258192	0.614015589	0	0.682064785	0.618107723	0.655233301	B1_xmas
lambda2				0.00001		0.002719133	9.96143E-06	0.004797696			lambda2
r2										2300.07315	r2
alpha2										15501.86179	alpha2
c2				2.355085888							c2
B2_air				0.503464667				0.435760725		1	B2_air
B2_xmas				0.292569003			7.73406055			1	B2_xmas
pi				0.126316301		0.577996112	0.833879777	0.01480213		0.988855022	pi
pi2											pi2
lambda3											lambda3
LL	-1362087.8	-1362362.2	-1362168	-1358274.934	-1360223.16	-1360203.612	-1360134.033	-1359533.934	-1360196.9	-1359711.247	LL
params	5	4	4	9	3	5	6	6	4	9	params
BIC	2724244.72	2724779.6	2724391.36	2716674.208	2720487.762	2720476.301	2720350.958	2719150.762	2720449.054	2719546.834	BIC

The first section of models are run without incorporating the 5-week lag in sales numbers, and as expected, they mostly fail to capture the pattern of the data, simply mimicking the trend of the airplay instead. As an example, the graph for the 2-segment Weibull:

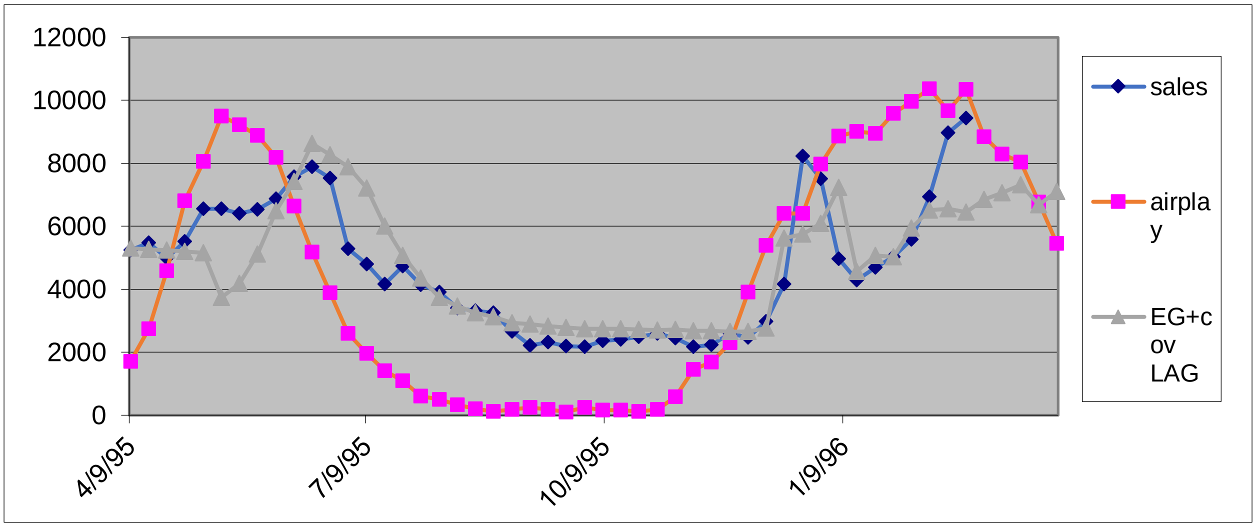


So far, our parameter estimates seem to check out – c is less than 1, and all the betas are positive!

We move on to our next section of models: the exponential models incorporating the lagged sales effect. Out of these, it seems the most effective is the 2-segment exponential model, with pooled Xmas coefficients. Here is the graph:

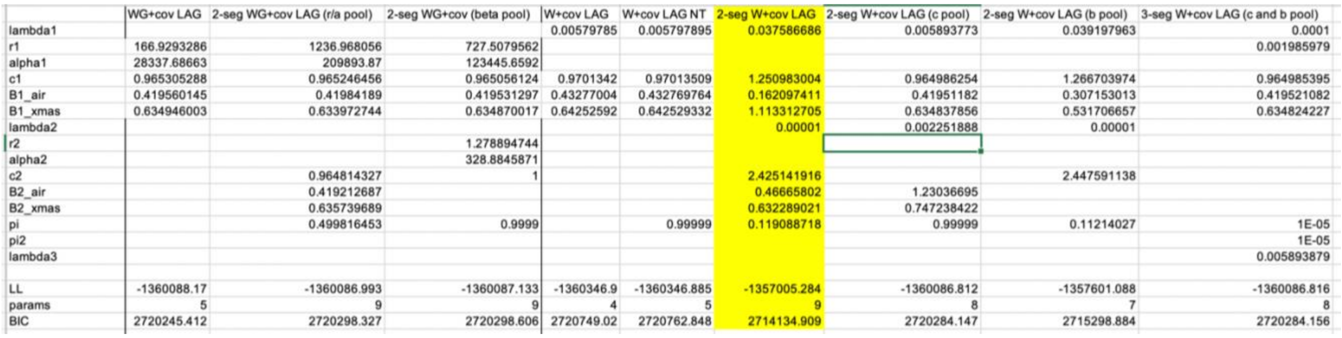


Our last section in this half are the exponential-gamma models; interestingly, when we segment the EG model, the r and alpha parameters blow up! At that point, much of the heterogeneity is explained by the segmenting (pi parameter). Here is the graph for the EG model with lagged sales included:



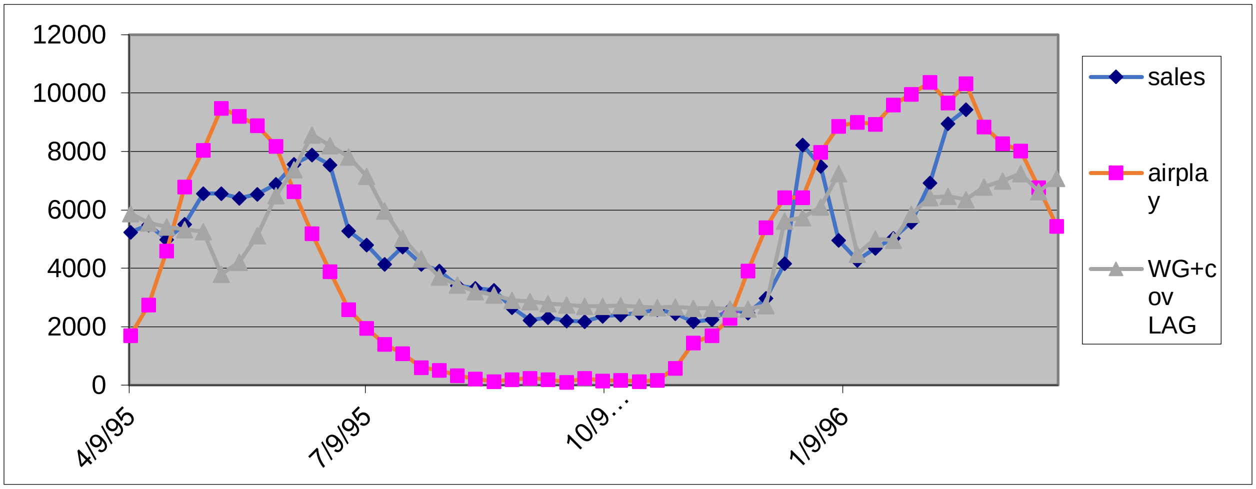
To this point, all the models so far have fit the data decently well, but all fail to capture the behavior after Christmas! We keep an eye on this as we run our remaining models.

A summary of the second half of our models:

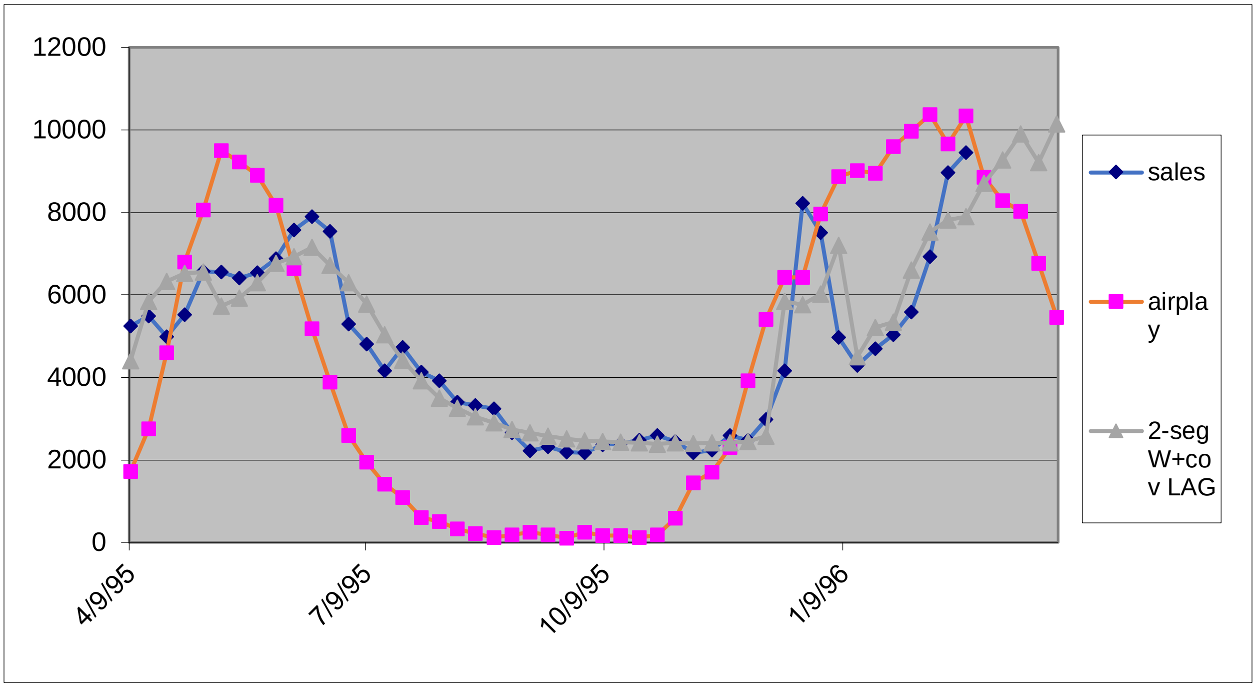


	WG+cov LAG	2-seg WG+cov LAG (r/a pool)	2-seg WG+cov (beta pool)	W+cov LAG	W+cov LAG NT	2-seg W+cov LAG	2-seg W+cov LAG (c pool)	2-seg W+cov LAG (b pool)	3-seg W+cov LAG (c and b pool)
lambda1				0.00579785	0.005797895	0.037586686	0.005893773	0.039197963	0.0001
r1	166.9293286	1236.968056	727.5079562						0.001985979
alpha1	28337.68663	209893.87	123445.6592						
c1	0.965305288	0.965246456	0.965056124	0.9701342	0.97013509	1.250983004	0.964986254	1.266703974	0.964985395
B1_air	0.419560145	0.41984189	0.419531297	0.43277004	0.432769764	0.162097411	0.41951182	0.307153013	0.419521082
B1_xmas	0.634946003	0.633972744	0.634870017	0.64252592	0.642529332	1.113312705	0.634837856	0.531706657	0.634824227
lambda2						0.00001	0.002251888	0.00001	
r2			1.278894744						
alpha2			328.8845871						
c2		0.964814327	1			2.425141916		2.447591138	
B2_air		0.419212687				0.46665802	1.23036695		
B2_xmas		0.635739689				0.632289021	0.747238422		
pi		0.499816453	0.9999		0.99999	0.119088718	0.99999	0.11214027	1E-05
pi2									1E-05
lambda3									0.005893879
LL	-1360088.17	-1360086.993	-1360087.133	-1360346.9	-1360346.885	-1357005.284	-1360086.812	-1357601.088	-1360086.816
params	5	9	9	4	5	9	8	7	8
BIC	2720245.412	2720298.327	2720298.606	2720749.02	2720762.848	2714134.909	2720284.147	2715298.884	2720284.156

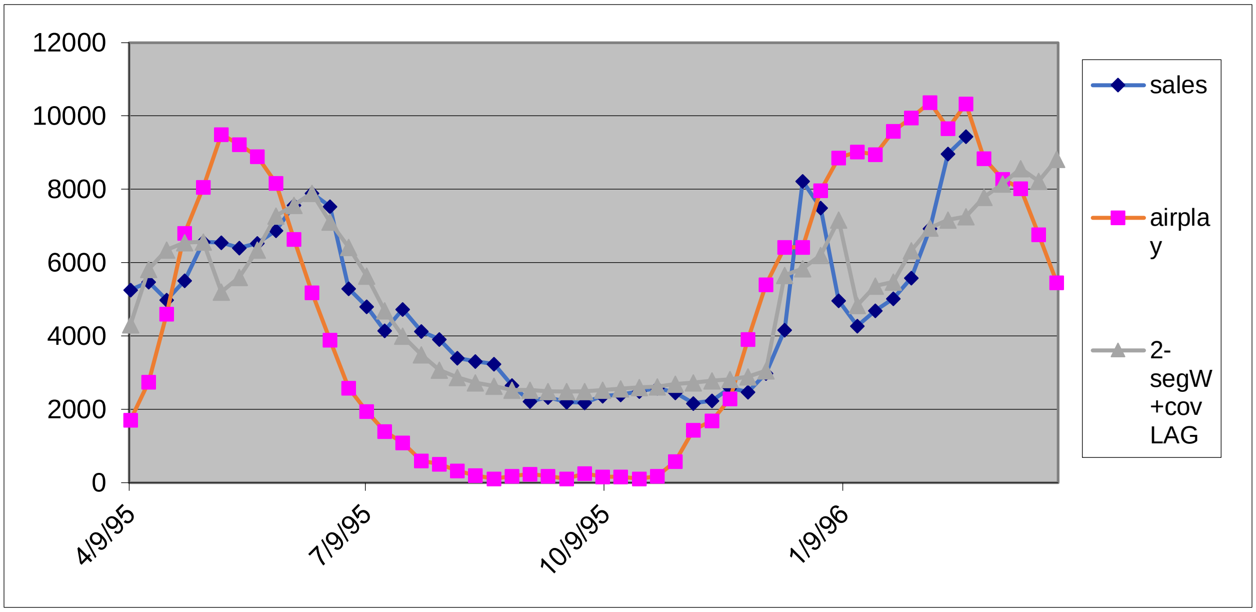
We begin by running Weibull-Gamma models incorporating the lag in sales, adding segmentation and pooling different covariates. The graph for the non-segmented WG model:



This still does not accurately predict the in-sample trend after Christmas; we move onto our final section, the Weibull models. Turns out, when we segment the data and do not pool our covariates, the model fits pretty nicely!



Although it is not perfect, the behavior after Christmas is a lot closer than our other models. We compare this to another 2-segment Weibull, but with the covariate betas pooled:

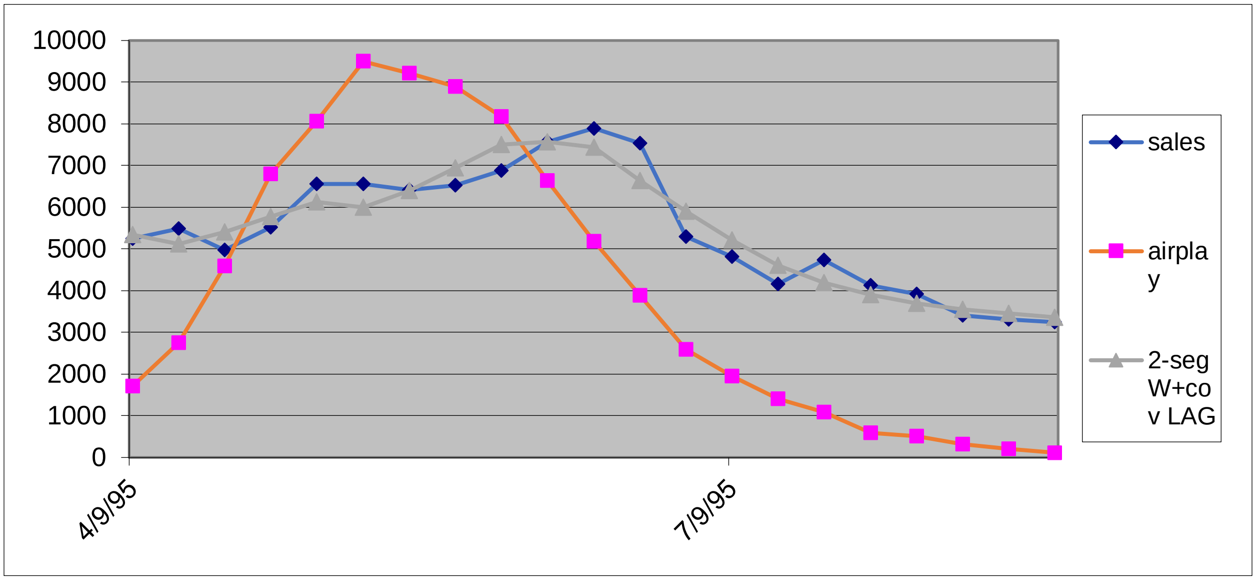


Here, the pooled version has a lower BIC score, and less accurately forecasts the in-sample post-Xmas data.

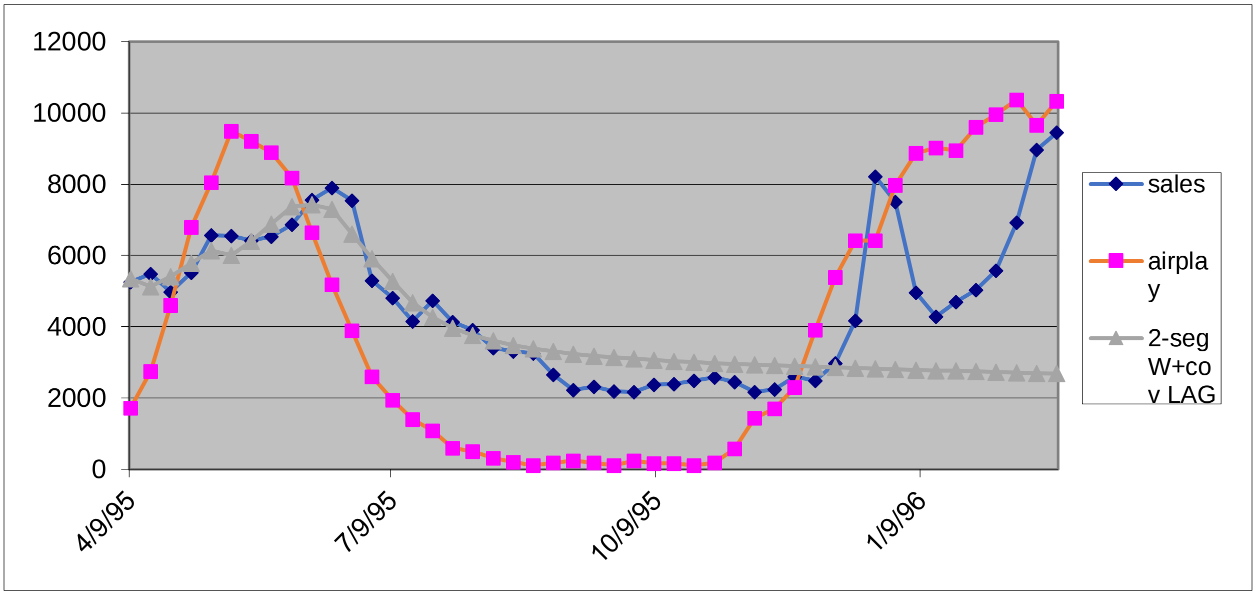
It seems right to proceed with the 2-segment Weibull model; first we carefully think about the story it is telling. From our parameters, we are essentially segmenting the population into 2 groups, one that is higher propensity and relatively more impacted by Xmas, and the almost non-zero buyer who is more impacted by airplay and shows positive duration dependence. Much of this is in line with our *a priori* estimates, and could even explain the growth in sales after Christmas – could it be that the almost never-buyer with $c > 1$ is causing the increase in sales post Xmas? Although this slightly conflicts with our earlier guess ($c > 1$), one plausible explanation is that there "distance makes the heart grow fond" effect finally starts to kick into effect now, nearly 8 months later, since these customers have such low lambdas; this also could possibly be an outside seasonal effect, something we leave to future research. For now, since the story does not seem wildly unreasonable, we proceed with this model.

Robustness and validity

Right off the bat, our model does have 9 parameters. How do we know we are not overfitting? We use a holdout period of the first 21 weeks, about 40% of our data, to test the robustness; it is important to note that such a holdout period does not test the Xmas covariate effect. Here is the fit of the 2-segment Weibull on just the holdout data:

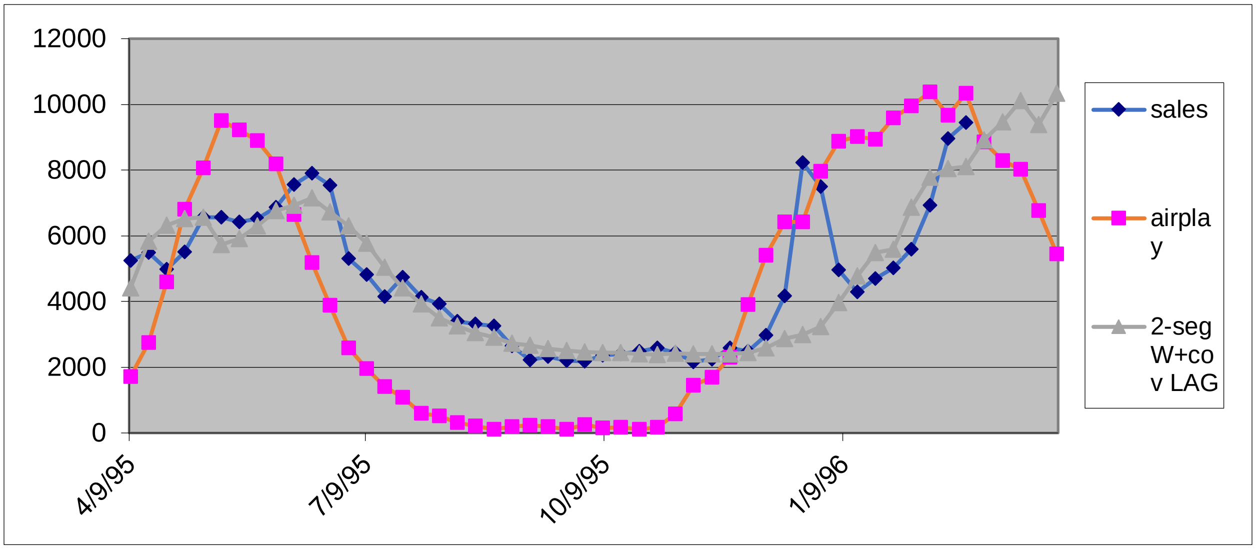


Looks pretty good! Using those same parameters on our entire dataset (and turning off the Xmas covariate):



Looks less great. From the cutoff of the holdout period up to Christmas (9/3 - 12/17), at the very least it does accurately show a slight decrease in sales. I will admit that the model might not be the most robust, but could also argue that not having the Xmas covariate data hurts its ability to accurately fit.

To see if our intuition holds (if Christmas did not exist, the trend would continue through it), we turn off the Xmas covariate effect on our whole data set:



Looks about right! In the absence of Christmas, the sales trend would have continued upward, without suddenly spiking, as it did before. Since the model did not crash or burn, we cautiously proceed to forecasts.

Forecasts

Using our 2-segmented Weibull model that incorporates a Christmas covariate effect and the 5-week leading airplay data, our forecasts are:

3/3/96	8711.898
3/10/96	9268.792
3/17/96	9903.469
3/24/96	9205.509
3/31/96	10150.2

Further Study

It would be interesting to see if the model could be improved with 3, 4, or 5 segments, something I do not think I could carry out with Excel on my laptop. Intuitively though, it seems as if segments would end up being pushed together, something that is already nearly happening in a lot of models with just two segments. Additionally, I would likely conduct further research on potential external covariates that could explain why sales increase after Christmas, whether seasonal or just random noise.