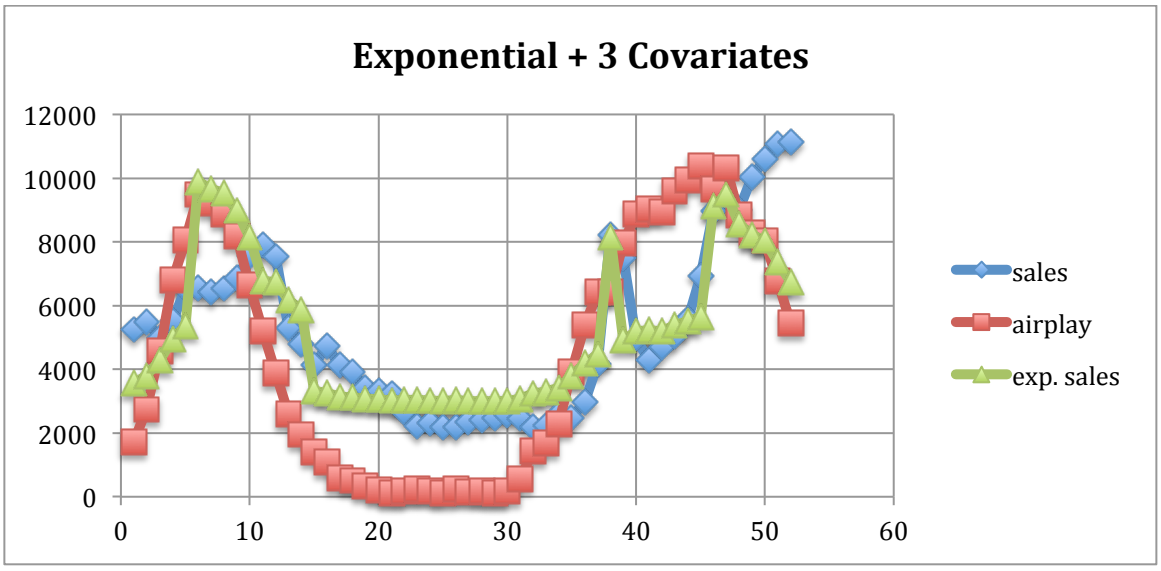
**Executive Summary**

The probability model of Radiohead's album *The Bends* that ultimately had the best fit for estimating the number of sales of the album in any given week in the 4/9/1995 – 3/31/1996 range is an Exponential model that incorporates three covariates: airplay, media rankings, and Christmas.



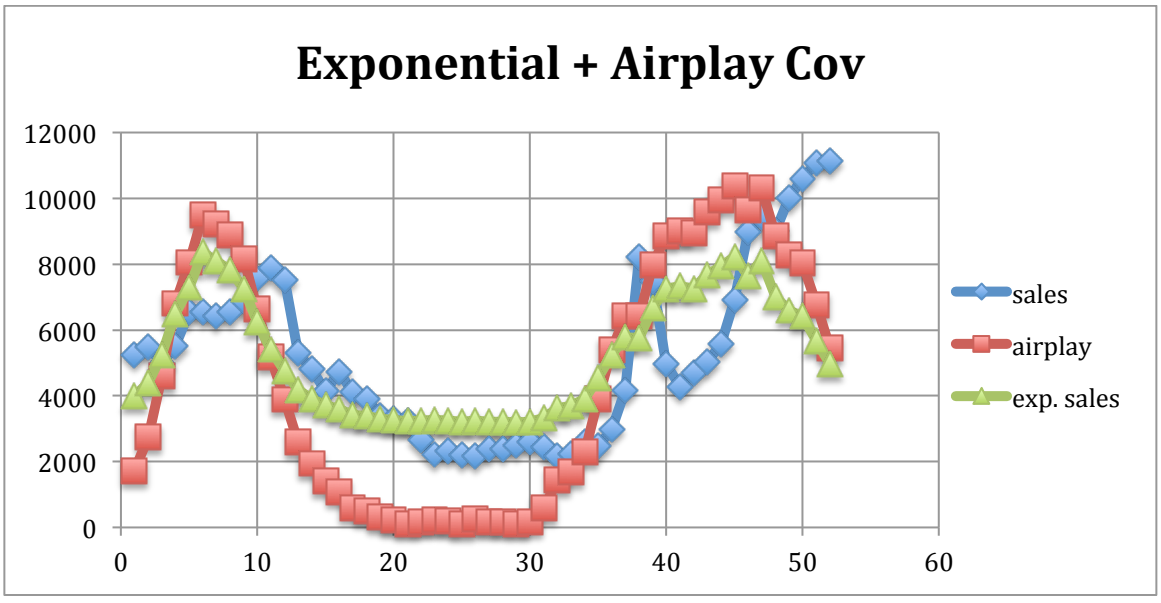

MAPE = 23.01%
LL = -1,847,548
BIC = 3,695,126
Lambda = 0.0016
b_airplay = 0.066
b_xmas = 0.608
b_top100 = 0.533
gamma_1$^{st}$ = 41
delta_1$^{st}$ = 0
gamma_2$^{nd}$ = 0
delta_2$^{nd}$ = 13

Of all the models, the Exponential with 3 variables has the least MAPE, and also visually the closest fit. As predicted, all three of the covariates are positive because they all increase the expected sales. Interestingly, there is a large time segment 7/16/95 – 12/10/95, where there is almost no airplay, sales are at their lowest, and none of

covariates play a role. The one clear imperfection with this model is the behavior in the last 5 time periods: despite actual sales (and popularity ranking) aggressively rising, the expected sales are falling (most likely due to the influence of the decreasing airplay covariate.)

**Benchmark**

The simplest model that attempts to model this behavior in sales is an Exponential with only the airplay covariate (shown below). Another model that could also have been the benchmark model is a Weibull with the airplay covariate (in appendix), but because it's MAPE is almost identical to the Exponential model's and it's c value is 1.003 it becomes evident that there is little duration dependence. While the Weibull is generally more dynamic, for this dataset specifically, it wasn't worth adding the extra parameter because it added negligible fit and sacrificed some of the model's simplicity.




MAPE = 31.81%
LL = -1855321
BIC = 3710672
Lambda = 0.0017
b_airplay = 0.097

## Description of Covariates

The addition of covariates into this model substantially improves the fit and simplifies the mechanics behind the model. Because the nature of the music industry was drastically different in 1995-1996, the covariates of airplay and popularity manifested themselves a lot more rigorously; specifically, airplay was completely determined by radio stations and concert venues, which are both heavily reliant on their ability to maintain popularity and thus money. Popularity was affected by media rankings such as those on MTV, Billboard, and Rolling Stone. The current era of YouTube and Spotify allows users to pick exactly what song they want to listen to, so exterior influences like money hold less influence today. Because of the large influence that airplay and magazine rankings had, we incorporate both into the model, as well as a Christmas covariate.

### Airplay

Because the volume of airplay is overwhelmingly large compared to the quantifiable influence of the other two covariates, I rescaled airplay to be 1000x less in order to avoid problems with solver. My hypothesis is that airplay will be a positive covariate because increased airplay means that more people hear the songs and thus more people would be willing to buy the album.

### Media Rankings

Because the world didn't have access to unlimited free music back in '95, the rankings that "experts" i.e. music-focused magazines and radio shows gave songs have a substantial impact on their sales. Specifically, Billboard Magazine releases their top 200 song rankings in every week's edition and has a searchable archive of all the top 200 lists

they've ever published. *The Bends* appeared on Billboard's top 200 in the t=6 to t=14 range, disappeared in the t=15 to t=45 range, and reappeared in t=46. It is important to note that once a song is published in the top 200 at least once, it has a sort of permanent positive impact on the songs popularity because of the fact that it "made it". In order to capture this effect I used a formula similar to the one Ryan Dismukes used to model the effect of 9/11 on airplane sales:

$$\beta_{top200}\,[1 - \gamma(1 - e^{-\delta|\tau - T})]$$

where $\gamma$ is the normal sales level (pre Billboard listing), $\delta$ is the rate of change for sales going back to their new (post listing) level, and T is the peak of popularity (t=12 for the first segment, t=52 for the second segment.) Because there are two disjoint time periods when *The Bends* appeared in the top 200 (t=6 to t=14 and t = 46 onwards) this effect happened twice, and thus needs to be modeled with a different gamma and delta for each time period. My prediction is that b_top200 is also positive because the appearance of a song in the top 200 will make more people aware of it, thus raising people's desire to listen to and purchase it.

Christmas

Christmas is a covariate that is relevant for almost any sales model, because customer's propensity to purchase anything skyrockets the week of Christmas. To account for this I made the Christmas covariate equal to zero for all weeks except the week of 12/24/1995 where it was equal to one. I expect b_xmas to be positive because Christmas will surely increase the number of sales.

**Examining the Effect of Heterogeneity**

Incorporating the three above covariates into the benchmark Exponential + airplay covariate model substantially improves its fit as MAPE goes from 32% to 23%. The next step is examining the potential impacts of heterogeneity on the dataset, so I incorporated a gamma mixture model in order to convert the Exponential + 3Cov into an Exponential Gamma (Pareto II) model. The EG (in appendix) appears to have a virtually identical MAPE, BIC, and LL, as well as virtually identical covariates. It is noteworthy at the EG had an alpha of 20,000, meaning that introducing heterogeneity was not particularly beneficial for the fit of the model. This result isn't too surprising because ultimately almost nobody purchases multiple copies of the same album, so there theoretically shouldn't be much of a source of heterogeneity anyways.

## Examining the effect of Latent Classes

While we concluded upon examination of the Exponential Gamma that there isn't much observable heterogeneity once covariates are in the model, it is possible that latent classes can reveal some of the heterogeneity that the previous models didn't capture. To check this, I ran the Exponential + 3 covariates model with two latent classes (in appendix.) This model had a MAPE of 23%, BIC of 3,695,126, and LL of -1847548, which is essentially identical to the Exponential + 3 covariates model without latent classes. Perhaps even more evidently, lambda_1 and lamda_2 had identical values of .00159, and pi_1 = .49, pi_2 = .51 showing that there really is no noteworthy heterogeneity among the classes.

## Conclusion

Neither the introduction of a mixture model nor the incorporation of latent classes seems to have any positive impact on the fit of the model. Most of the improvement from

the benchmark model came from the addition of the covariates. This shouldn't be too surprising because the nature of album sales is that there is not much reason to purchase more than one (unless it is gifted or broken) because a 2nd CD provides no utility to the user. Thus I would expect heterogeneity to not have much impact, since ultimately everyone ends up being a once-buyer or a never-buyer.

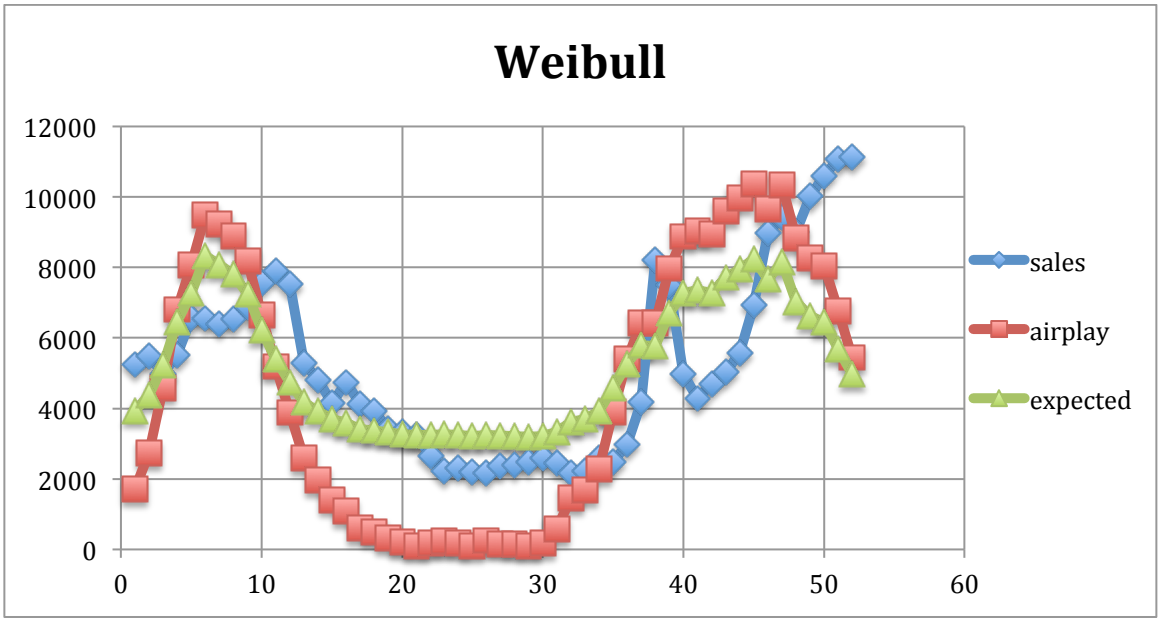
On the flipside, what is remarkably important is the covariates. This is empirically evidenced by the fact that the addition of Christmas and magazine rankings decreased the MAPE from 32% to 23%. Logically this shouldn't be surprising: increasing the amount that an album is played on the radio will increase people's exposure to it, and thus increase their propensity to buy the album. Getting listed on (and climbing up) the Billboard Magazine top 200 would also increase exposure of the album and thus increase sales. Christmas too undoubtedly has a strong positive impact on the number of sales. This all explains why the Exponential + 3 covariates is the best-fitting model that was run, even despite it's striking simplicity when compared to the other models.

Those other models that account for heterogeneity could theoretically be used and wouldn't provide any substantially bad results, but ultimately there negligible beneficial impact is not enough to justify the drastic increase in complexity that they inevitably bring. The same effect is noticed when examining the incorporation of duration dependence with the Weibull Gamma model–although the model's fit can't possibly be made worse with the addition of an extra parameter, the impact is so negligible that it is simply not worth the complexity. The Exponential + 3 covariates model is by far the most succinct model when it comes to achieving the best fit, so it is the one I ultimately chose. Surely, this fit could be a lot stronger because there are plenty of other factors that

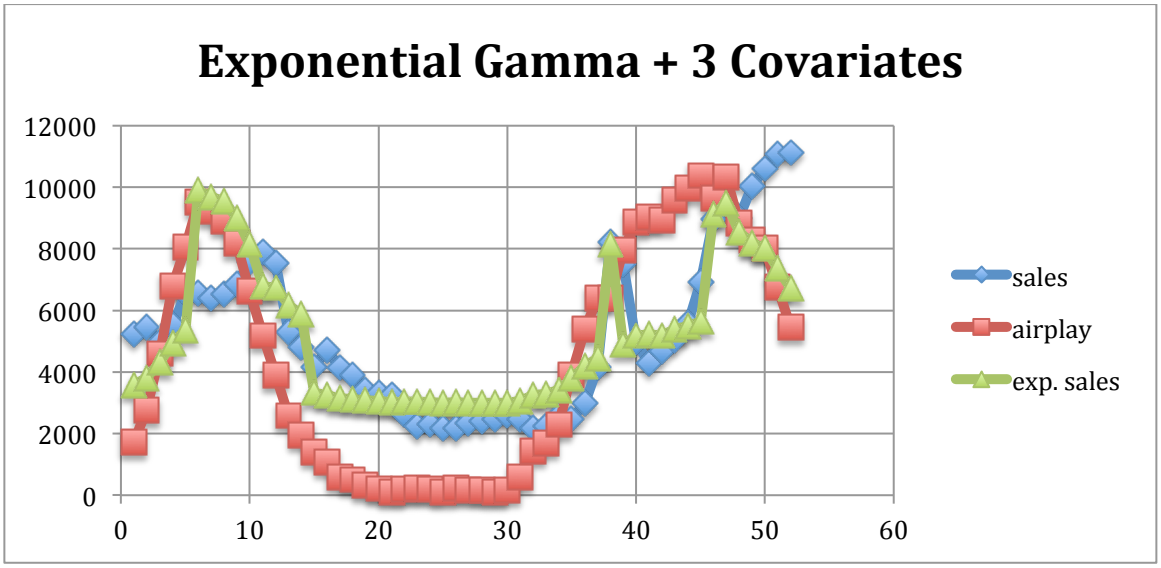
impacted Radiohead sales; for example, the previously mentioned magazines, MTV and Rolling Stone, undoubtedly wrote about and ranked *The Bends*, but unfortunately I was unable to find their archives. Similarly, there must have been TV shows, movies, and concerts that all featured *The Bends* and would have impacted the popularity and thus the sales of this album.

This model doesn't hold much relevance from a managerial perspective because the nature of both the airplay and the magazine rankings covariates has changed drastically over the last 20 years due to increased access and cheaper pricing within the music industry. The high values of the covariates show exactly this effect–that airplay and magazine rankings actually had a huge impact on the popularity and thus sales of Radiohead albums. For example, the lack of any substantial airplay and corresponding low sales in the t=15 through t=45 period demonstrates exactly this effect. The strength of these factors 20 years ago goes to show some interesting changes in the music industry. Now, because everyone can (for free) listen to whatever songs they want as many times as they want, the popularization of music has become a lot more democratized (through means such as Spotify and YouTube, which show the number of plays for each song) and less dependent on money or negotiations (which could surely impact the airplay of a song or it's placement in the top 200.)

**Appendix**

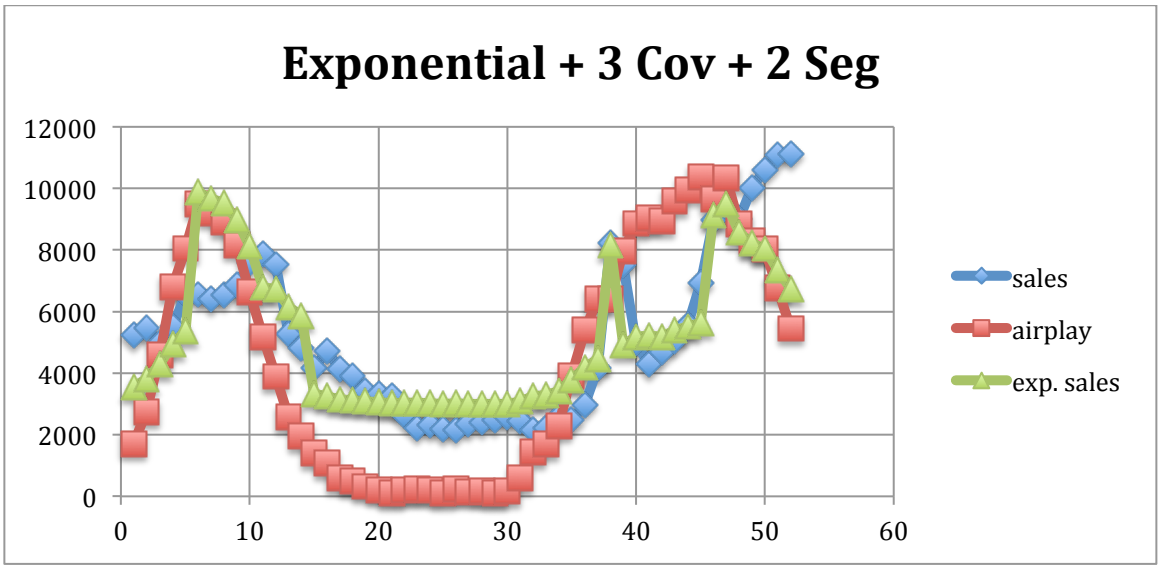



MAPE = 31.85%
LL = -1,855,321
BIC = 3,710,685
Lambda = 0.0017
b_airplay = 0.097




MAPE = 23.01%
LL = -1,847,579

BIC = 3,695,188
alpha = 19,997
r = 31.96
b_airplay = 0.066
b_xmas = 0.608
b_top100 = 0.533
gamma_1st = 41
delta_1st = 0
gamma_2nd = 0
delta_2nd = 13




MAPE = 23.01%
LL = -1,847,548
BIC = 3,695,126
Lambda_1 = 0.0016
Lamda_2 = 0.0016
Pi_1=.49
Pi_2=.51
b_airplay = 0.066
b_xmas = 0.608
b_top100 = 0.533
gamma_1st = 41
delta_1st = 0
gamma_2nd = 0
delta_2nd = 13