# Understanding How Covariates Perform Across Different HB Packages

*Jeffrey Dumont*
*Jeff Keller*
*Nelson Whipple*
*Albert Boateng*

RSG the science of insight

## MANAGERIAL IMPLICATIONS

Recently, such as at the 2014 ART Forum, modelers have debated how to get the most value out of the upper-level of a Hierarchical Bayesian choice model. This debate includes the pros and cons of using covariates in the upper-level structure.

Our main objective is to further the discussion regarding the benefits of including covariates in the mixed logit model and add to the set of empirical evidence. In practice, modelers have limited time to estimate models, so it is important to understand whether the value of modeling covariates justifies the time investment to do so.

Our poster examines the potential for covariate modeling to improve the following:

- explaining the observed choices
- predicting future choices
- explaining preference heterogeneity

In addition to the main objective, we also address concerns about how covariates are implemented in the popular software packages for HB model estimation. This helps researchers to be more informed about the implications arising from similarities and differences across the packages.

## 1 THE COVARIATE MODEL

Popular model estimation applications such as Sawtooth CBCHB and R packages bayesm and RSGHB allow covariates to be included in the upper-level structure of the Hierarchical Bayesian model.

In the basic HB mixed logit model, the population hyper-parameters are assumed to be distributed multivariate normal with means, A, and covariance matrix, D:

$$C \sim N(A, D)$$

Introducing covariates to the basic model relaxes the assumption of unimodality, allowing the population means to vary by demographic variables (Z). This results in an upper-level structure that can be multimodal in nature.

$$C \sim N(A + F * Z, D)$$

The estimation of the parameters, F, requires the addition of another layer to the Gibbs sampler used in the parameter estimation. Although this additional layer enables a multimodal distribution to be modeled, it can cause models to take significantly longer to estimate.

## 2 THE METHODOLOGY

**Data description:**

- Data come from a stated preference study fielded in 2015
- 449 respondents completed 8 stated choice tasks
- Binary choice with alternatives defined by 8 multi-level attributes
- Demographic data was also collected to serve as potential covariates in the model
  - Covariates were identified using MNL

**Holdout sample selection:**

- **External** holdout: 10% of respondents (n = 45) were held out from model estimation
- **Internal** holdout: The remaining 90% had 1 choice task held out from estimation

**Models estimated:**

- RSGHB, bayesm and Sawtooth CBCHB
- Basic and covariate models
- For RSGHB, we modified the default priors to better match those used by CBCHB and bayesm

**Statistics measured:**

*To understand performance on estimation dataset*

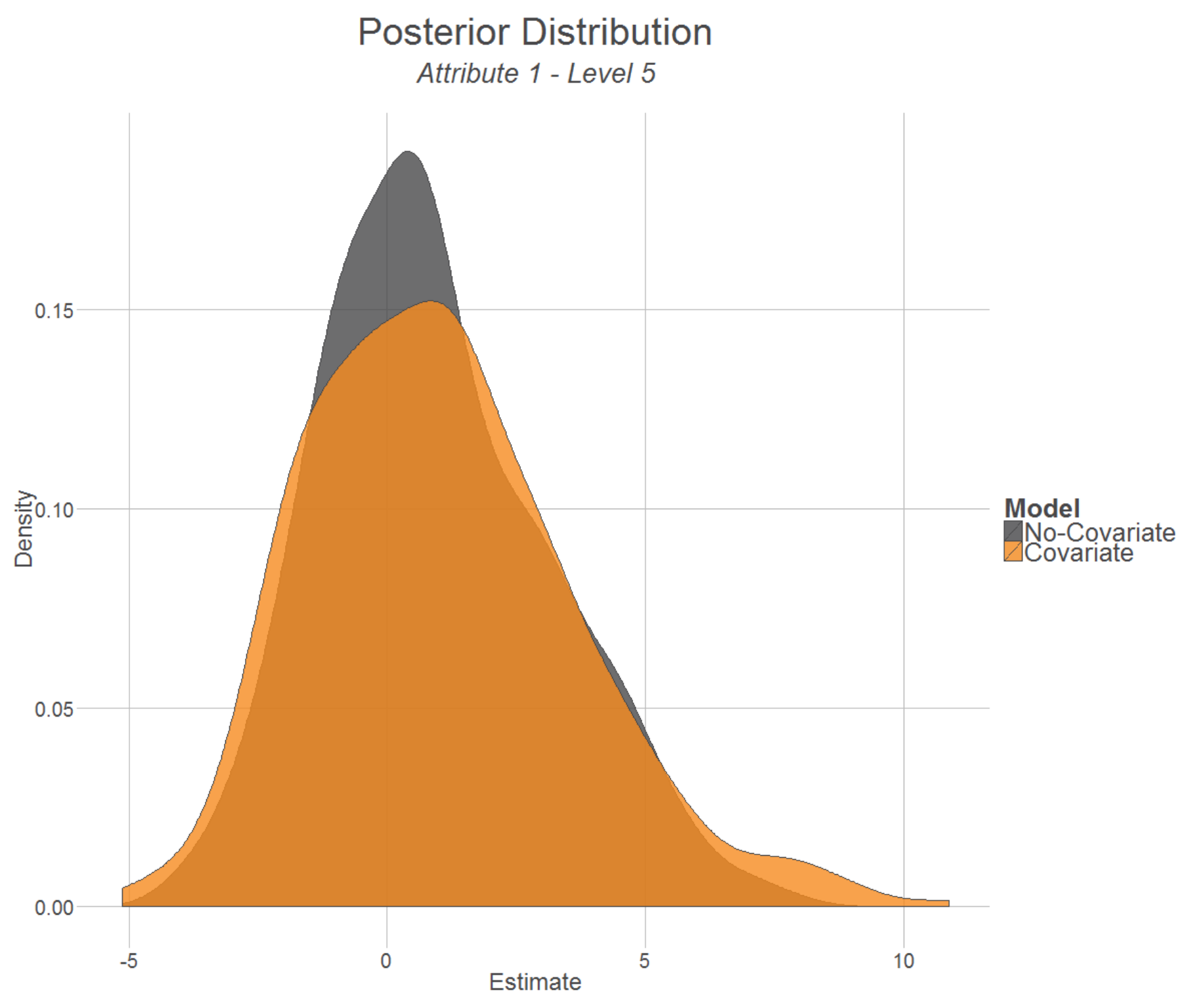
- Mean root-likelihood (RLH) from the lower-level of the HB model
- Simulated log-likelihood from the upper-level of the HB model

*To understand predictive performance*

- Hit rate on external and internal holdout

*To understand explanations of preference heterogeneity*

- Kolmogorov-Smirnov Tests



## 3 RESULTS

**Performance on estimation data:**

| Model | RLH (Lower Level) | Log-Likelihood (Upper Level) |
|---|---|---|
| Sawtooth (Basic) | 0.912 | -1489.41 |
| Sawtooth (Covariate) | 0.925 | -1453.43 |
| bayesm (Basic) | 0.854 | -1488.13 |
| bayesm (Covariate) | 0.866 | -1538.07 |
| RSGHB Default (Basic) | 0.709 | -1495.83 |
| RSGHB Default (Covariate) | 0.730 | -1466.79 |
| RSGHB Modified (Basic) | 0.893 | -1469.24 |
| RSGHB Modified (Covariate) | 0.924 | -1441.82 |

- Introduction of the covariates improves the model fit on the estimation data both in the lower- and upper-levels of the HB model for all tests
- General improvement in fit suggests that the covariate adds additional information to the model
- The default settings (reduced prior variances) in RSGHB result in worse fit on the estimation data

**Predictive performance:**

| Model | Internal Holdout Hit Rate | External Holdout Hit Rate |
|---|---|---|
| Sawtooth (Basic) | 0.658 | 0.628 |
| Sawtooth (Covariate) | 0.678 | 0.626 |
| bayesm (Basic) | 0.693 | 0.626 |
| bayesm (Covariate) | 0.651 | 0.631 |
| RSGHB Default (Basic) | 0.693 | 0.657 |
| RSGHB Default (Covariate) | 0.691 | 0.647 |
| RSGHB Modified (Basic) | 0.681 | 0.629 |
| RSGHB Modified (Covariate) | 0.683 | 0.631 |

- The models all do generally better with predicting the internal holdout sample compared to the external
- The introduction of the covariate to the model does not improve the performance on the external holdout sample
- Better estimation fit does not translate into better predictive performance on holdout samples (classic example of over-fitting)
- Reducing the prior variance results in better holdout performance (reduced issues with over-fitting)

**Differences in preference heterogeneity:**

- Kolmogorov-Smirnov tests provide evidence that the addition of covariates result in different distributions of heterogeneity
  - For Sawtooth: 74% of the distributions are different
  - For bayesm: 78% of the distributions are different
  - For RSGHB Default: 93% of the distributions are different
  - For RSGHB Modified: 96% of the distributions are different
- These results suggest that information from covariates can help place each individual in a more appropriate population distribution

## 4 SUMMARY

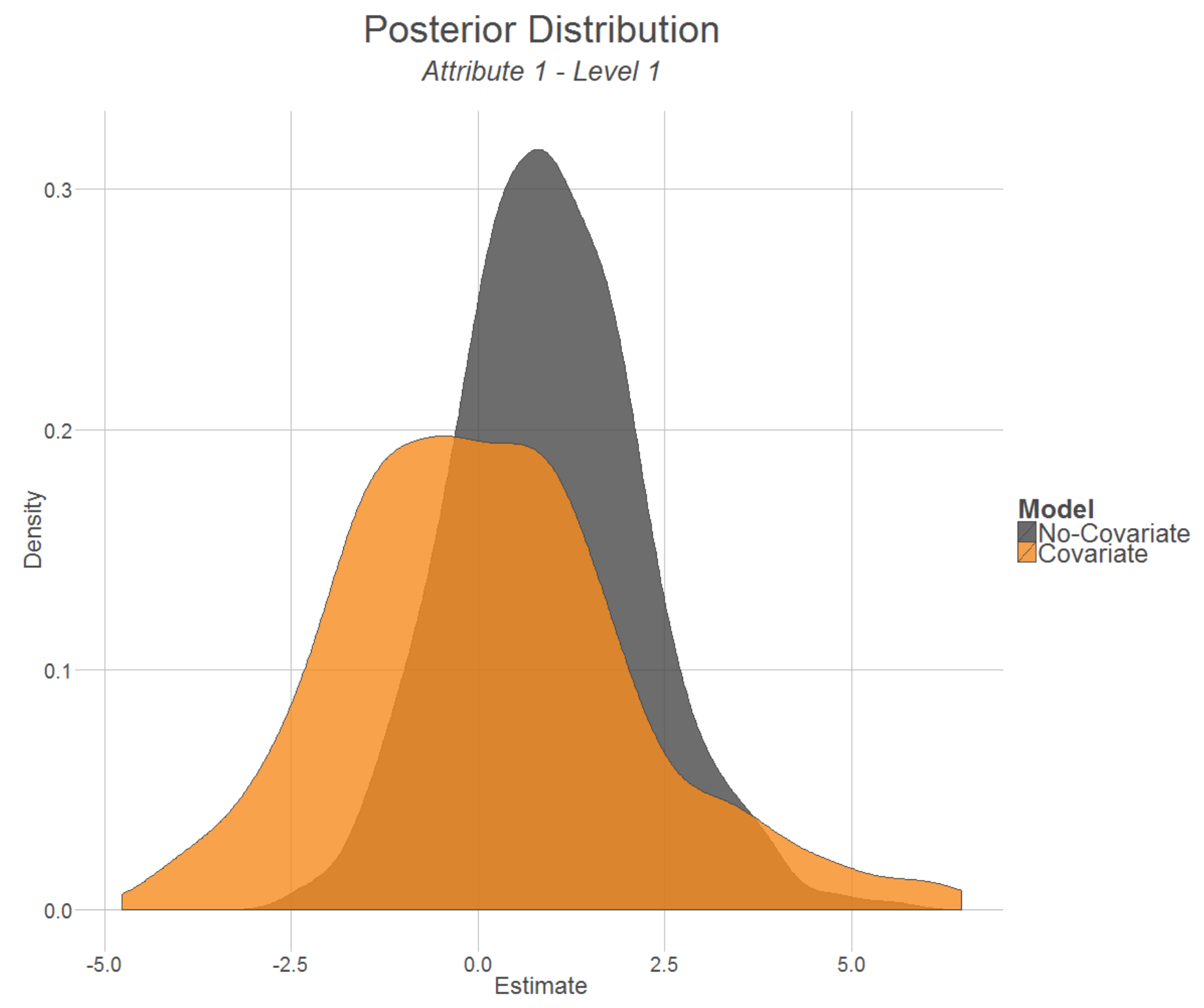
This poster presents evidence that the addition of covariates can improve choice models in terms of:

- Better model fit on the estimation data
- Individualized distributions of preference heterogeneity instead of one-size-fits-all
- Improvement in internal holdout sample prediction

However, this better performance does not necessarily translate into better performance in predicting external holdout sample.

These findings reinforce the need to keep the intended use of the model in mind when making evaluations. If the goal is to understand the distribution of preferences across the population, incorporating covariates into mixed logit models appears to add some value. However, if the end goal is to predict choices, this research suggests that covariates might provide limited additional value. This finding might vary by situation. In fact, simply reducing the prior variance improves the predictive performance of the model beyond the improvement we observe from adding covariates.

Additionally, these findings appear to be independent of the software used to estimate the models.



## REFERENCES

1. Sawtooth Software, Inc. *CBC Hierarchical Bayes Module*. Version 5.5.2. http://www.sawtoothsoftware.com/products/conjoint-choice-analysis
2. Sawtooth Software, Inc. Application of Covariates within Sawtooth Software's CBC/HB Program: Theory and Practical Example. http://www.sawtoothsoftware.com/support/technical-papers/hierarchical-bayes-estimation/application-of-covariates-within-sawtooth-software-s-cbc-hb-program-theory-and-practical-example-2009
3. Peter Rossi. (2012). *bayesM: Bayesian Inference for Marketing/Micro-econometrics.* R package version 2.2-5 http://CRAN.R-project.org/package=bayesm
4. Kenneth Train. (2009). *Discrete Choice Methods with Simulation, 2nd edition.*
5. J. Dumont, J. Keller, and C. Carpenter. (2015). *RSGHB: Functions for Hierarchical Bayesian Estimation*. R package version 1.1.0. http://cran.r-project.org/web/packages/RSGHB/index.html
6. J. Keller, J. Dumont, N. Whipple. (2014). *RSGHB: Using R to Expand the Capabilities for Hierarchical Bayesian Model Estimation (Case Study)* https://github.com/jeffdumont/RSGHB/tree/master/2014_ART_Forum_Poster
7. J. Dumont, M. Giergiczny, S. Hess (2015). Individual Level Models vs. Sample Level Models: Contrasts and Mutual Benefits. Transportmetrica A: Transport Science, Vol. 11, Issue 6. http://www.tandfonline.com/doi/abs/10.1080/23249935.2015.1018681