Recast Applied Statisitcal Modeller Take Home

Preliminaries

We suggest you have an installation of R on your computer. We also recommend using Rstudio, but you're welcome to use whatever IDE you prefer.

We recommend you use `cmdstanr` for this problem. You can read more about this library at https://mc-stan.org/cmdstanr/ but you won't need all the functionality. You should be familiar with the `cmdstan_model` function and the `draws` and `sample` methods.

If you prefer to use `rstan` or any other interfaces to Stan, that is also fine.

If you use other libraries in the process of completing this take home (e.g. `tidybayes`, `posterior`, or `bayesplot`) then that is also fine.

We expect this assignment to take no more than 4 hours. You have 5 days to finish this assignment upon receipt.

Introduction

In the package you've received, you will find three files:

- `channel-spend.rds` – This is an RDS file containing a dataframe of spend within marketing channel plus the dependent variable. For the purposes of this assignment, you can think of these data as being measured in hundreds of thousands of dollars.
- `simple-model.stan` – This is a highly simplified version of the model we use in production. There is no need to edit this file to successfully complete this homework (but feel free to play around with it after your submission).
- `analysis.R` – An R script where you will be working.

In what follows, you will be editing some existing R code. You will be required to make at the very minimum one plot. You will also be asked a few questions regarding code output and your thoughts on the output.

Purpose of the model

Imagine the dependent variable for this model (`depvar` in `channel-spend.rds`) is daily revenue (in hundreds of thousands of dollars) for a company. Each day, we expect some base amount of revenue to come in. For example, people will buy cheeseburgers regardless of whether they saw an ad for the restaurant recently. On top of that, the rest of the revenue was driven by marketing of various types - on television, search engines, social media, etc.

The purpose of this model is to estimate how much of the dependent variable each day was driven by each advertising channel, and how much was not.

In `simple-model.stan`, the `predicted` variable contains predictions for revenue on each day in our dataset. This variable and its prior/posterior draws will be the focus of this take home assignment.

What To Do In `analysis.R`

Step 0

Read this entire document before going forward. Once you've finished this document, return to Step 1 and proceed as necessary.

Step 1

Upon opening `analysis.R` you will find an incomplete function called `load_in_channel_data`. Write a function to load `channel-spend.rds` into the R session, then call that function assigning the output to the variable `channels`.

If done correctly, the `channels` variable should be a 100×8 dataframe with the following columns

- `influencers`
- `linear_tv`
- `mailers`
- `meta_prospecting`
- `meta_retargeting`
- `podcast`
- `search_non_branded`
- `depvar`

The `depvar` column contains the dependent variable (i.e. observed revenue). Each row is one day and hence represents one step in time.

Step 2

Sample from the prior predictive by setting `prior_only=1` in `dat`. Extract the draws for the `predicted` variable. Use these draws to visualize the prior predictive distribution anyway you see fit.

In no more than 5 sentences per question, answer the following:

- Do you think the priors we have set are reasonable? Yes, or No?
- What is your justification for your answer above?

Step 3

Regardless of your answer above, sample from the model by setting `prior_only=0` in `dat`.

In no more than 5 sentences per question, answer the following:

- What is the output of `fit$cmdstan_diagnose()`? What does this information tell you about the model you just sampled?
- If this was a real model for a real client, would you feel comfortable proceeding with this model? Why or why not?

Step 4

If you think the model is fine, and you would feel comfortable proceeding with this model, then great! Please submit your solution indicating so.

If you think this model is deficient in some manner, try editing the priors in order to improve the model. There is no need to edit the Stan model itself. If you have edited the priors, then in no more than 5 sentences per question, answer the following:

- What did you change?
- Why did you change the priors you selected?
- How do you know the model is better than before?

Submission

We suggest you write the answers to these questions in a clearly formatted and organized document (PDF is fine, google docs is fine, whatever works best for you).

If you have made plots to supplement your answers, insert then as necessary with a caption to provide context.