Data Generation Diagnostics

William Ruth

Introduction

The goal of this document is to check the validity of a Monte Carlo sample for my Trust Study Analysis. To this end, I check various conditional and marginal moments of the response variables by comparing them against group-wise and global sample means.

In each section, I will identify an expected value (conditional or marginal), and derive a theoretical expression for this expected value. I then identify an estimator of this expectation by averaging suitable quantities from a typical Trust Study dataset. We treat covariates as random, so typically we can treat these averages using standard approaches for iid samples. We then compare our estimator with the theoretical quantity to verify that our simulated data don't obviously violate any assumptions. We start with expressions for M, the mediatior, then move to Y, the outcome. This is natural, since the dependence of Y on M makes theoretical calculations a bit more intricate.

Many of the theoretical quantities can be expressed in terms of the so-called "logit-normal' ' integral. That is, the mean of a random variable whose logit is normally distributed. We open this document with a brief exploration of these logit-normal integrals and some corresponding notation. We then move on to the various means, their theoretical analysis, and comparison of theory with numerics.

The Logit-Normal Integral

Let X be normally distributed and $Y = expit(X) = exp(X)/(1 + exp(X)) = 1/(1 + exp(-X))$. In the analysis that follows, we will often encounter quantities of the form $\mathbb{E}Y$. There is no closed-form expression for this expected value, but we can write it as

$$\begin{aligned}\mathbb{E}Y &= \int_{-\infty}^{\infty} \frac{\phi(x;\mu,\sigma)}{1+exp(-x)}dx \\ &= \int_{-\infty}^{\infty} \frac{\phi(z;0,1)}{1+exp(-\mu-\sigma z)}dz\end{aligned}$$

This integral can be approximated using numerical quadrature. Specifically, we use the `integrate` function in `R`. Note: When implementing this expression, be sure to use `-Int` and `Int` as the range for `integrate`, as just choosing large finite values can cause inaccuracies.

More generally than the above analysis, we will often need to apply the logit-normal integral to a vector. To this end, we apply the expit, or inverse logit function elementwise on vectors, and define the following logit-normal integral function:

$$\varphi(\mu,\sigma) = \int_{\mathbb{R}^p} \text{expit}(\mu+\sigma z)\phi(z;0_p,I_p)dz$$

The function φ can be evaluated by repeated application of numerical quadrature via the `integrate` function. I have implemented this function in my `MultiMedUQ` package as `phi` in the file `Exact_Asymptotic_Helpers.r`.

Mediator

We start granular, conditioning on everything relevant to the mean of M, then gradually marginalize out these dependencies until we get an expression for the marginal mean of M.

First, we set some notation. Let X, Z be covariates, with $X \in \mathbb{R}^{n\times p}$ and $Z \in \mathbb{R}^{n\times q}$. We take all covariates' entries to be iid Bernoulli(0.5). Write $D = \{X, Z\}$ for the data. Next, let $\alpha \in \mathbb{R}^p$ be the vector of fixed effects and $U \sim N(0, \Sigma) \in \mathbb{R}^q$ be the random effects. Taken together, we write $g[\mathbb{E}(M|U, D)] = X^T\alpha + Z^TU$, where $g = \text{logit}$ is the link function. Note that we have specified this relationship for a single observation. Since covariates are taken to be random, each observation within a group is iid, given the random effect for that group.

$\mathbb{E}(M|U, D)$

This quantity corresponds to the mean of a particular observed M. There's nothing really to average here, so we don't have an empirical reference to compare against our theory. Nevertheless, we will present the theory as a foundation for more interesting comparisons (also, because it's fast).

Applying the inverse link function to both sides, we get a simple expression for this section's conditional expectation:

$$\begin{aligned}\mathbb{E}(M|U, D) &= g^{-1}(X^T\alpha + Z^TU)\\ &=: g^{-1}[\eta(U, D)]\end{aligned}$$

where η is the linear predictor, $\eta(U, D) = X^T\alpha + Z^TU$.

$\mathbb{E}(M|U)$

This is where things start to get interesting. Marginalizing over the data while conditioning on the random effect corresponds to averaging over values of M within a single group.

Theory

This theoretical representation takes a bit more work. We will need to take expectation over all covariates. Fortunately, these covariates are all binary, so each only takes two possible values. We represent this averaging process by a new operator. I'm going to be a bit informal here to avoid having to be too pedantic about the presence of an intercept or possible overlap between X and Z in D.

First, let $\mathcal{C}$ be the corners of the unit cube with dimension equal to the number of unique random covariates in D. I.e. The set of all vectors of zeros and ones of the appropriate length. For some conformal function, f, define the operators T and $\bar{T}$ as

$$\begin{aligned}Tf &= \sum_{d\in\mathcal{C}} f(d)\\ \bar{T}f &= |\mathcal{C}|^{-1}\, Tf.\end{aligned}$$

When applying $\bar{T}$ to functions with multiple arguments, we will use subscripts to denote which variables are being summed over.

The operator $\bar{T}$ corresponds exactly to taking the expected value over D. The conditional expectation in this section can thus be easily represented using this operator.

$$\begin{aligned}\mathbb{E}(M|U) &= \mathbb{E}_D\mathbb{E}_M(M|U, D)\\ &= \bar{T}_D\mathbb{E}_M(M|U, D)\\ &= \bar{T}_D g^{-1}[\eta(U, D)]\end{aligned}$$

Numerics

We estimate the conditional expectation in this section by averaging over values of M within the same group. Note that, in order to compute the theoretical value of this conditional expectation, we need to know the value of the random effect, U. This is easily done by making a minor modification to my sampling function. In fact, I even included an argument to return each group's random effect.

```
#! WARNING: The order of groups in data is not the same as the order of groups in all_REs

all_E_Ms = c()
all_M_bars = c()

for(i in seq_along(all_REs)){
    # if (i %% 50 == 0) print(paste0("i = ", i, " of ", length(all_REs)))

    # i=1
    this_group = all_REs[[i]][["data"]]
    # this_group = filter(data, group == paste0("G", i))
    this_REs = all_REs[[i]][["M"]]

    this_data_M = this_group %>%
                    select(-Y, -M) %>%
                    cbind(1, .)

    lin_preds_M = lin_preds_from_vecs(this_data_M, b_M, seq_len(ncol(this_data_M)), this_REs, c(1, 2))

    #* Compute conditional expectation of M given U

    data_all_combs = cbind(1, expand.grid(c(0,1), c(0,1), c(0,1)))
    lin_pred_all_combs = lin_preds_from_vecs(data_all_combs, b_M, seq_len(ncol(data_all_combs)), this_RE
    probs_all_combs = expit(lin_pred_all_combs)

    this_E_M = mean(probs_all_combs)
    all_E_Ms[i] = this_E_M

    #* Estimate conditional expectation of M given U using MC sample
    this_M_bar = mean(this_group$M)
    all_M_bars[i] = this_M_bar
    # this_M_SD = sd(this_group$M)
}

# Compare analytical and numerical values of E(M | U)
data_M_given_U = data.frame(theory = all_E_Ms, numerical = all_M_bars)
print(head(data_M_given_U))
```

```
##      theory numerical
## 1 0.6079750     0.609
## 2 0.5327142     0.525
## 3 0.5346503     0.536
## 4 0.5736016     0.590
## 5 0.5860608     0.585
## 6 0.4423146     0.450
```

```
# Mean and SD of relative error (normalized by theoretical value)
rel_errs = abs(all_M_bars - all_E_Ms)/ all_E_Ms
print(mean(rel_errs))
```

```
## [1] 0.02593547
```

```
print(sd(rel_errs))
```

```
## [1] 0.02109653
```

$\mathbb{E}(M)$

We now proceed to the marginal mean of M. This one is estimated by averaging the groups' means. Here we see our first example of a logit-normal integral.

Theory

We obtain this conditional expectation by taking the expected value with respect to U of $\mathbb{E}(M|U)$. To save me a bit of typing, I'm going to omit the bounds of integration. They're all over an appropriate dimensional real space.

$$\begin{aligned}
\mathbb{E}(M) &= \mathbb{E}_U \mathbb{E}_M(M|U) && (1)\\
&= \int \mathbb{E}_M(M|U=u)\phi(u;0,\Sigma) && (2)\\
&= \int \bar{T}_D g^{-1}[\eta(U,D)]\phi(u;0,\Sigma) && (3)\\
&= \bar{T}_D \int g^{-1}[\eta(U,D)]\phi(u;0,\Sigma) && (4)\\
&= \bar{T}_D \int g^{-1}[X\alpha + Zu]\phi(u;0,\Sigma) && (5)\\
&= \bar{T}_D \int g^{-1}[X\alpha + Z\Gamma z]\phi(z;0,I_q) && (6)\\
&= \bar{T}_D \varphi(X\alpha, Z\Sigma Z^T), && (7)
\end{aligned}$$

where Γ is any[1] square root of Σ (i.e. $\Gamma\Gamma^T = \Sigma$). While Equation (7) looks intimidating, in our numerical example there are only three random covariates, so the operator $\bar{T}_D$ only involves summing over $2^3 = 8$ terms. This does still require us to compute 8 logit-normal integrals, but that shouldn't be too hard.

Numerics

[1] We may prefer to require that Γ be symmetric and positive definite. There is only one such square root matrix for a given Σ.