Regress adaptation index on a feature

YC, LL, LG

December 12, 2022

We observe data points (a_i, b_i, x_i), for $i = 1, \ldots, n$. a_i is the control response, b_i is the test response, x_i is a feature extracted from the input image presented (e.g. contrast, orientation, etc.). For simplicity, we assume x_i is univariate, even though the theory can be extended to the multivariate situation. We are interested in estimating the effect of x on the adaptation index. In the noiseless setting, it is just finding a functional relationship

vector output from ANN

$$\frac{b_i}{a_i} = f(x_i).$$

If we restrict ourselves to linear function, $f(x) = x \cdot \beta + c$ and β can be estimated via linear regression of $\frac{b_i}{a_i}$ on x_i. In the presence of noise, the situation is a bit more complicated. There are two natural estimators.

1. **Estimator 1: take ratio then regress.** We calculate the ratio $\frac{b_i}{a_i}$ and the regress $\frac{b_i}{a_i}$ on x_i to find β.

$$\hat{\beta}_{1,\_} = \arg\min_{\beta,\beta_0} \sum_{i=1}^{n} \left(\frac{b_i}{a_i} - x_i \cdot \beta - \beta_0 \right)^2 .$$

2. **Estimator 2: regress b_i.** Regress b_i on $a_i \cdot x_i$ and a_i, get the coefficient in front $a_i \cdot x_i$ as an estimator of β.

$$\hat{\beta}_{2,\_,\_} = \arg \min_{\beta,\gamma,\gamma_0,} \sum_{i=1}^{n} (b_i - (a_i x_i) \cdot \beta - a_i\gamma - \gamma_0)^2$$

or predict R1 from stim feature via ANN

3. **Estimator 3: two-stage estimator.** First, regress a_i on x_i to obtain α, α_0. Then regress b_i on $((\alpha x_i + \alpha_0)x_i)$ and a_i as follows to obtain β:

$$\hat{\beta}_{3,\_,\_} = \arg \min_{\beta,\gamma,\gamma_0,} \sum_{i=1}^{n} (b_i - ((\alpha x_i + \alpha_0)x_i) \cdot \beta - a_i\gamma - \gamma_0)^2$$

The goal of this write-up is show that the third estimator is less sensitive to noise in a_i, through a short math derivation and some simulations.

We make two main assumptions.

Assumption 1. *For each $i \in \{1, \ldots, n\}$, (a_i, b_i) is independent and identically distributed (i.i.d.).*

INITIAL

Assumption 2. *For each $i \in \{1, \ldots, n\}$, $\mu_{A,i}$ is generated*

assume R1 depends on stim
assume adaptation depends on stim INITIAL

$$\mu_{A,i} = x_i \cdot \alpha^* + \alpha_0 + \epsilon_{\alpha,i}.$$

θ_i is generated as

$$\theta_i = x_i \cdot \beta^* + \beta_0^* + \epsilon_{\theta,i},$$

where $\epsilon_{\alpha,i}, \epsilon_{\theta,i}$ are small Gaussian noise $\mathcal{N}(0, \sigma_\theta^2)$. Given θ_i, (a_i, b_i) is a random sample drawn from the joint Gaussian distribution

$$\begin{bmatrix} A \\ B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{A,i} \\ \theta_i \mu_{A,i} \end{bmatrix}, \begin{bmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_B^2 \end{bmatrix}\right),$$

with parameters $\mu_{A,i}, \theta_i, \sigma_A^2$ and σ_B^2. In other words, $a_i = \mu_{A,i} + \epsilon_A, b_i = \theta_i \mu_{A,i} + \epsilon_B$, (ϵ_A, ϵ_B) has a joint Gaussian distribution.

The above assumptions can be quite restrictive in practice, they are open for discussion.

INITIAL

Note that unlike in the previous draft, we made ρ the correlation between the noise in A and B be zero here!!! It is an assumption that makes the third estimator much better, but it is not clear whether can assume this in practice. It would be nice to test.

1 The distribution of residuals for true β^* in two estimators

Under the above assumptions, for the first estimator, if we plugin the true β^*, we have

$$\begin{aligned}
\frac{b_i}{a_i} - x_i \cdot \beta^* - \beta_0^* - 0 &= \frac{\theta_i \mu_A + \epsilon_B}{\mu_A + \epsilon_A} - x_i \beta^* - \beta_0^* - 0 \\
&= \frac{\theta_i(\mu_A + \epsilon_A) - \theta_i \epsilon_A + \epsilon_B}{\mu_A + \epsilon_A} - x_i \beta^* - \beta_0^* - 0 \\
&= \epsilon_{\theta,i} + \frac{-\theta_i \epsilon_A + \epsilon_B}{\mu_A + \epsilon_A} \\
&= \underbrace{\epsilon_{\theta,i} + \frac{-(x_i \beta^* + \epsilon_{\theta,i})\epsilon_A + \epsilon_B}{\mu_A + \epsilon_A} x_i\beta^*}_{E_1}
\end{aligned}$$

The main concern is that the division by $\mu_A + \epsilon_A$ might cause the residual to have high variance. Also because of that division, the best estimator might be biased for estimating β^*.

For the second estimator, if we plugin the true β^*, we have

$$\begin{aligned}
b_i - (a_i x_i) \cdot \beta^* - a_i \beta_0^* + 0 &= (\theta_i \mu_A + \epsilon_B) - ((\mu_A + \epsilon_A) x_i)\, \beta^* - (\mu_A + \epsilon_A)\, \beta_0^* \\
&= (x_i \beta^* + \beta_0^* + \epsilon_{\theta,i})\mu_A + \epsilon_B - (\mu_A + \epsilon_A)(x_i \beta^* + \beta_0^*) \\
&= \underbrace{\mu_A \epsilon_{\theta,i} + \epsilon_B + (x_i \beta^* + \beta_0^*)\epsilon_A}_{E_2}.
\end{aligned}$$

The residual for true β^* in the second estimator should have smaller variance that that in the previous one, especially in the case where $\mu_A + \epsilon_A$ can be close to 0. **The main concern** here is that the error E_2 is correlated with $a_i x_i$ and a_i.

In the third estimator, the two-stage estimator, the first-stage of estimator decorrelates

?

2 The distribution of estimated β's.

In multivariate linear regression, for response $\mathbf{y} \in \mathbb{R}^n$ and design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, the least squares solution has a generic form

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}.$$

Consider the design matrix $\begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$ as fixed and the condition number is not too bad, then the variance of $\hat{\beta}_1$ is approximately proportional to the variance of E_1. We are skipping the exact calculation of the variance, which is feasible but too tedious.

Similarly, if the condition number of the design matrix $\begin{bmatrix} a_1 x_1 & a_1 & 1 \\ \vdots & \vdots & \vdots \\ a_n x_n & a_n & 1 \end{bmatrix}$ is not too bad, the variance of $\hat{\beta}_2$ is approximately proportional to the variance of E_2.

3 Simulations

In this section, we carry out several simulations under the two assumptions 1 and 2 to support/visualize the claims in the previous sections.

3.1 Case 1

We simulate $N = 10000$ data points of (a_i, b_i, x_i) i.i.d. from Assumption 2 with parameter choices

$$\begin{aligned} x_i &\sim \text{Uniform}[0.4, 1.2] \\ \beta^* &= 1 \\ \alpha^* &= 2 \\ \mu_{A,i} &= \beta^* * x_i + 0.01\mathcal{N}(0,1) \\ \theta_i &= \alpha^* * x_i + 0.01\mathcal{N}(0,1) \\ \rho &= 0 \\ \sigma_A = \sigma_B &= 0.3 \end{aligned}$$

We use the β estimate ± 1.96 standard error as 95% confidence interval (CI), and obtain

$$\begin{aligned} \hat{\beta}_1 &= 0.9041, \quad 95\% CI = [0.829, 0.979] \\ \hat{\beta}_2 &= 1.6683, \quad 95\% CI = [1.640, 1.696] \\ \hat{\beta}_3 &= 0.9989, \quad 95\% CI = [0.985, 1.013] \end{aligned}$$

Obervations:

- $\hat{\beta}_2$ is clearly wrong, because the residual is correlated with the covariates as we explained above.

- $\hat{\beta}_1$ is OK but the CI is slightly off, because the actually tail is much heavier than Gaussian tail. The problem will become more serious in the next experiment.
- $\hat{\beta}_3$ is good.

3.2 Case 2

We increase the variance. We simulate $N = 10000$ data points of (a_i, b_i, x_i) i.i.d. from Assumption 2 with parameter choices

$$\begin{aligned} x_i &\sim \text{Uniform}[0.4, 1.2] \\ \beta^* &= 1 \\ \alpha^* &= 2 \\ \mu_{A,i} &= \beta^* * x_i + 0.01\mathcal{N}(0,1) \\ \theta_i &= \alpha^* * x_i + 0.01\mathcal{N}(0,1) \\ \rho &= 0 \\ \sigma_A = \sigma_B &= 0.6 \end{aligned}$$

We use the β estimate ± 1.96 standard error as 95% confidence interval (CI), and obtain

$$\begin{aligned} \hat{\beta}_1 &= 1.2201, \quad 95\% CI = [0.692, 1.748] \\ \hat{\beta}_2 &= 1.5791, \quad 95\% CI = [1.541, 1.617] \\ \hat{\beta}_3 &= 0.9982, \quad 95\% CI = [0.978, 1.018] \end{aligned}$$

Obervations:

- Again, $\hat{\beta}_2$ is clearly wrong, same reason as above.
- Now $\hat{\beta}_1$ has very large standard error, because the error distribution is not Gaussian (because of division by 0) so the error variance is large.
- $\hat{\beta}_3$ is good.

4 Takeaways

If the two assumptions are satisfied, clearly Estimator 3 (the two-stage estimator) is a better estimator for estimating β. However, one needs to be careful about the assumptions and check whether these assumptions are reasonable in real data.

1. b_i, a_i are all approximately Gaussian distributed.
2. There are no correlation in variance for b_i and a_i: they are correlated because their mean is related. it could be a strong assumption
3. Both relationships are approximately linear:

$$\mu_{A,i} = x_i \cdot \alpha^* + \alpha_0 + \epsilon_{\alpha,i}.$$

$$\theta_i = x_i \cdot \beta^* + \beta_0^* + \epsilon_{\theta,i},$$

   If a nonlinear model is known to be better, this part may be extended to nonlinear models.