Model Selection Review Sheet

Ec143, Spring 2023

Bryan S. Graham

In preparing for the exam review your lecture notes, assigned course readings, problem sets and prepare answers to the questions on review sheets (such as this one). You may bring to the exam a single 8.5×11 inch sheet of paper with notes on it. No calculation aides are allowed in the exam (e.g., calculators, phones, laptops etc.). You may work in pencil or pen, however I advise the use of pencil. Please be sure to bring sufficient blue books with you to the exam if possible. Scratch paper will be provided.

[1] You have been hired by UNICEF to estimate the prevalence of childhood stunting (low height-for-age) across municipalities in a country where childhood malnutrition is commonplace. Let Y_{it} be the height-for-age Z score of individual $t = 1, \ldots, T$ in municipality $i = 1, \ldots, N$. In each municipality you draw T children at random and compute the average height-for-age Z score

$$\bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it}.$$

You assume that $Y_{it} | \theta_i \sim \mathcal{N}(\theta_i, \sigma^2)$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. In this model the expected height-for-age Z score, θ_i, varies across municipalities. Your goal is to estimate the municipality (population) means $\theta_1, \theta_2, \ldots, \theta_N$. Municipalities with low θ_i estimates will be slated to receive new anti-hunger and nutrition programs. Initially you may assume that σ^2 is known (in a healthy population of children $\sigma^2 \approx 1$ since height-for-age Z scores are calibrated to have unit variance in such a setting).

[a] Explain why, if $f(y_{it} | \theta_i)$ is Gaussian, the municipality mean is also Gaussian:

$$\bar{Y}_i \big| \theta_i \sim \mathcal{N}\left(\theta_i, \frac{\sigma^2}{T}\right).$$

[b] Let $\|\mathbf{m}\| = \left[\sum_{i=1}^{N} m_i^2\right]^{1/2}$ denote the Euclidean norm of a vector. Let $\theta = (\theta_1, \ldots, \theta_N)'$. Show that

$$\mathbb{E}\left[\left\|\hat{\theta} - \theta\right\|^2\right] = \sum_{i=1}^{N} \mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right],$$

with $\hat{\theta}$ some estimate – based upon the sample data $\mathbf{Y} = (Y_{11}, \ldots, Y_{1T}, \ldots, Y_{N1}, \ldots, Y_{NT})'$ – of θ. Explain why this measures *expected* estimation accuracy or *risk*? What is being averaged in the expectation?

[c] Further show that

$$\mathbb{E}\left[\left\|\hat{\theta}-\theta\right\|^2\right]=\sum_{i=1}^{N}\mathbb{V}\left(\hat{\theta}_i\right)+\sum_{i=1}^{N}\left(\mathbb{E}\left[\hat{\theta}_i\right]-\theta_i\right)^2.$$

Interpret this expression.

[d] Consider the following family of estimators for θ_i (for $i=1,\ldots,N$):

$$\hat{\theta}_i=(1-\lambda)\,\bar{Y}_i+\lambda\mu,$$

with μ the country-wide mean of Y_{it} (i.e., the expected height-for-age Z score of a randomly sampled child from the full country-wide population). You may assume that μ is known (perhaps from prior research). Assume that $0\leq\lambda\leq 1$. Interpret this estimator? Why might the estimator with $\lambda=0$ be sensible? How might you justify the estimator when $\lambda>0$.

[e] Show, for the family of estimates introduced in part [d], that

$$\mathbb{E}\left[\left\|\hat{\theta}-\theta\right\|^2\right]=(1-\lambda)^2\frac{N}{T}\sigma^2+\lambda^2\sum_{i=1}^{N}\left(\theta_i-\mu\right)^2.$$

You hear, in the hallways of Evans, that "small λ means small bias" and "big λ means low variance". Explain?

[f] Show that the risk-minimizing choice of λ, say λ^*, is

$$\lambda^*=\frac{N\sigma^2}{N\sigma^2+\sum_{i=1}^{N}T\left(\theta_i-\mu\right)^2}.$$

Is an estimator based upon λ^* feasible? Why or why not? What is the optimal choice of λ^* as $T\to\infty$? Provide some intuition for your answer. What happens to the optimal choice of λ^* as σ^2 becomes large? Provide some intuition for your answer.

[g] Show that

$$\sum_{i=1}^{N}\mathbb{E}\left[\left(\bar{Y}_i-\hat{\theta}_i\right)^2\right]=\sum_{i=1}^{N}\mathbb{E}\left[\left(\bar{Y}_i-\theta_i\right)^2\right]+\sum_{i=1}^{N}\mathbb{E}\left[\left(\hat{\theta}_i-\theta_i\right)^2\right]-\frac{2\sigma^2}{T}\mathrm{df}\left(\hat{\theta}\right)$$

with the degree-of-freedom of $\hat{\theta}$ (or *model complexity*) equal to

$$\mathrm{df}\left(\hat{\theta}\right)=\sum_{i=1}^{N}\frac{T}{\sigma^2}\mathbb{C}\left(\bar{Y}_i,\hat{\theta}_i\right).$$

We call the term to the left of the first equality above *apparent error*.

[h] Show that

$$\sum_{i=1}^{N} \mathbb{E}\left[\left(\bar{Y}_i - \theta_i\right)^2\right] = \frac{N}{T}\sigma^2$$

and also, for the family of estimates indexed by λ introduced in part [d] above, that

$$\mathrm{df}\left(\hat{\theta}\right) = N\left(1-\lambda\right).$$

[i] Calculate apparent error and model complexity for $\hat{\theta}$ when $\lambda = 0$. Explain?

[j] Calculate apparent error and model complexity for $\hat{\theta}$ when $\lambda = 1$. Explain?

[k] You are roaming around Evans Hall looking for Professor Graham's office. You can't find his office because of the confusing floor plan. However, after a few hours of wandering around aimlessly, you bump into someone who introduces himself as Chuck Stein. He rearranges the expression you derived in part [g] above to get

$$\sum_{i=1}^{N} \mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right] = -\sum_{i=1}^{N} \mathbb{E}\left[\left(\bar{Y}_i - \theta_i\right)^2\right] + \sum_{i=1}^{N} \mathbb{E}\left[\left(\bar{Y}_i - \hat{\theta}_i\right)^2\right] + \frac{2\sigma^2}{T}\mathrm{df}\left(\hat{\theta}\right).$$

He then notices your results from part [h] further imply that

$$\sum_{i=1}^{N} \mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right] = -\frac{N}{T}\sigma^2 + \sum_{i=1}^{N} \mathbb{E}\left[\left(\bar{Y}_i - \hat{\theta}_i\right)^2\right] + 2N\frac{\sigma^2}{T}\left(1-\lambda\right).$$

Finally he says therefore an unbiased estimate of risk is:

$$\mathrm{SURE}\left(\bar{Y}, \lambda\right) = -\frac{N}{T}\sigma^2 + \sum_{i=1}^{N} \lambda^2 \left(\bar{Y}_i - \mu\right)^2 + 2\frac{N}{T}\sigma^2\left(1-\lambda\right).$$

Provide an explanation for Chuck Stein's claim. Next show that

$$\mathbb{E}\left[\left(\bar{Y}_i - \mu\right)^2\right] = \frac{\sigma^2}{T} + \left(\theta_i - \mu\right)^2$$

and hence that

$$\mathbb{E}\left[\mathrm{SURE}\left(\bar{Y}, \lambda\right)\right] = \sum_{i=1}^{N} \mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right]$$

as implied by Chuck's unbiasedness claim. <u>HINT</u>: Don't forget the work you've done in part [e] above (and you may need to factor a quadratic equation in λ). Why is this result awesome?

[l] Let $\hat{\lambda}^*$ be the value of λ which minimizes SURE $\left(\bar{Y}, \lambda\right)$. Show that

$$\hat{\lambda}^* = \frac{\frac{N}{T}\sigma^2}{\sum_{i=1}^{N}\left(\bar{Y}_i - \mu\right)^2}.$$

Relate this feasible estimator to the infeasible oracle estimator based upon λ^* defined in part [f] above. HINT: use the expression for $\mathbb{E}\left[\left(\bar{Y}_i - \mu\right)^2\right]$ derived in part [k] to argue that when N is large enough $\hat{\lambda}^* \approx \lambda^*$. Discuss.

[m] You decide to use the estimator based upon $\hat{\lambda}^*$. For each municipality you calculate

$$\hat{\theta}_i = \left(1 - \hat{\lambda}^*\right)\bar{Y}_i + \hat{\lambda}^*\mu,$$

and then report when $\hat{\theta}_i < -1$. You tell the Minister of Health that those municipalities where $\hat{\theta}_i < -1$ should be targeted for supplemental child nutrition programs to combat stunting. A few month's later the mayor of village $i = 19$ comes to the capital as says: "You really screwed up. In my village *every single one* of the T sampled children had a height-for-age Z score, Y_{19t}, less than negative -1. My villages' mean was $\bar{Y}_{19} = -5/4$, but because your goofy econometrician decided to shrink all the village means toward the country-wide mean of $\mu = 0$ (with $\hat{\lambda}^* = 1/5$), they reported $\hat{\theta}_{19} = -1$ to you. As a result I have a bunch of hungry kids not getting the help they need. Your estimator is biased!" Write a response to this Mayor's concern.