Exaggerated Certainty: One Statistician's Take on EoE

Richard Lockhart

PHYSTAT

October 30, 2024

Summary, not outline

- Uncertainty is routinely underestimated.
- EoE is a very nice idea.
- Statisticians usually analyze one experiment at a time.
- Exceptions are in Meta Analysis and in auxiliary data and paradata in Survey Sampling.
- Wilks theorem hinges on having "asymptotically unbiased" estimating equations.
- When adding terms to log-likelihoods focus on bias of the resulting estimating equations

Summary concluded

- A toy example shows importance of independence assumption.
- If you insist on frequency guarantees do you average over both data sets or condition on auxiliary data?
- Should all systematic errors be treated as if error rates could have frequentist control?
- Bayesian posteriors smoothly incorporate auxiliary information about common parameters into new analysis.
- Lots of study needed to deal with degrees of freedom.
- Some work on degrees of freedom in smoothing problems.

Simple vision

- Two experiments: main and auxiliary
- Two models: primary has parameters θ and γ.
- Auxiliary had parameters γ and say τ.
- In primary experiment θ, γ is not identifiable or poorly identified
- Or just "borrow strength" from the auxiliary experiment to reduce overall uncertainty about γ and therefore θ.
- Natural but usually impractical frequentist strategy: put the two data sets together and fit a model to the combined data.
- Log-likelihood is, assuming independence of experiments, (m, a for *main* and *auxiliary*)

$$\ell_{\text{comb}}(\theta, \gamma, \tau) = \ell_m(\theta, \gamma) + \ell_a(\gamma, \tau)$$

- Same marginal models for the two experiments, more parameters to fit.

Standard HEP alternative

- Find normal approximation to estimation error for γ in auxiliary experiment.
- Model $\hat{\gamma}_a$ as Normal(γ, σ^2) with σ known.
- Actually σ^2 estimated usually via standard MLE with inverse of negative Hessian .
- Call this estimate V.
- Now do "natural", less "impractical" thing and treat the data set as the primary data together with $\hat{\gamma}_a$.

Standard HEP alternative 2

- Joint model is original model for main expt with independent Gaussian model for $\hat{\gamma}_a$.
- Log-likelihood is replaced by

$$\ell_1(\theta, \gamma) = \ell_m(\theta, \gamma) - \frac{(\gamma - \hat{\gamma}_a)^2}{2V}$$

- Estimate parameters by MLE – solve system of *likelihood equations*:

$$\frac{\partial}{\partial \theta} \ell_1(\theta, \gamma) = 0$$

and

$$\frac{\partial}{\partial \gamma} \ell_1(\theta, \gamma) - \frac{\partial}{\partial \gamma} \frac{(\gamma - \hat{\gamma}_a)^2}{2V} = 0$$

- Key point, fixed by Glen and Enzo, this system of equations is biased.

Estimating equations, Wilks Theorem

- Real likelihood equations are

$$\text{Score} = 0$$

- Basic ingredient of all MLE theory: these equations are unbiased

$$\mathrm{E}_{\theta,\gamma}\left\{\nabla_{\theta,\gamma}\ell(\theta,\gamma)\right\} = 0$$

- Wilks theorem needs this to centre the χ^2 approximation.
- So you want for HEP strategy:

$$\mathrm{E}_{\theta,\gamma,\tau}\left\{\ell_1(\theta,\gamma)\right\} = 0$$

  which means we need

$$\mathrm{E}_{\gamma,\tau}\left\{\frac{\gamma - \hat{\gamma}_a}{V}\right\} = 0$$

- The presence of V makes this unlikely to be right.

Errors on Errors fixes the problem

- Most classical case for statisticians: γ is a mean parameter in a Gaussian linear model .
- Simplest: $Y_1, \ldots, Y_n$ independent $\mathrm{N}(\gamma, \tau)$. $(\tau = \sigma^2)$
- Auxiliary likelihood ($\tau = \sigma^2$, $S = \sum (Y_i - \bar{Y})^2$)

$$\ell_a(\gamma, \tau) = -\frac{n}{2}\log(\tau) - \frac{n(\bar{Y} - \gamma)^2 + S}{2\tau}.$$

- The MLE of γ, τ is then

$$\hat{\gamma}_a = \bar{Y} \quad \hat{\tau} = S/n.$$

- Wilks theorem: for true value of γ

$$\Lambda(\gamma) = 2\left\{\ell_a(\hat{\gamma}, \hat{\tau}) - \ell(\gamma, \hat{\tau}(\gamma))\right\} \approx \chi^2_1$$

where $\hat{\tau}(\gamma) = \sum (Y_i - \gamma)^2/n$.

Enzo and Glen summary for me not audience

- Glen and Enzo treat the auxiliary data not as $Y_1, \ldots, Y_n$ but as $\hat{\gamma}, \hat{\tau}$ and write down the exact likelihood for this data.
- The actual distribution of $\hat{\tau}$ is given by

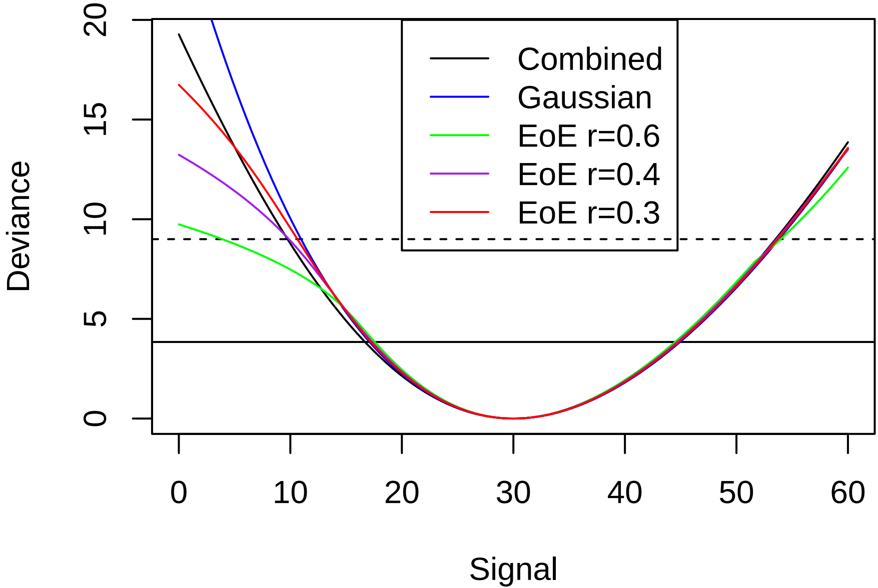
$$\frac{n\hat{\tau}}{\tau} \sim \chi^2_\nu$$

  where $\nu = n - 1$ is the degrees of freedom.
- And $\hat{\tau}$ is independent of $\hat{\gamma}_a = \bar{Y}$
- Degrees of freedom $n - 1$; uses unbiased variance estimate.
- Exact auxiliary likelihood is as in Enzo and Glen's work: normal$(\gamma, \tau)\times$ *independent* chi-squared.
- Corresponding r has $\nu = n - 1 = \frac{1}{2r^2} = \frac{1}{2\epsilon^2}$
- Largest r, smallest df (for linear model) are df $= 1$ and $r^2 = \frac{1}{2}$.

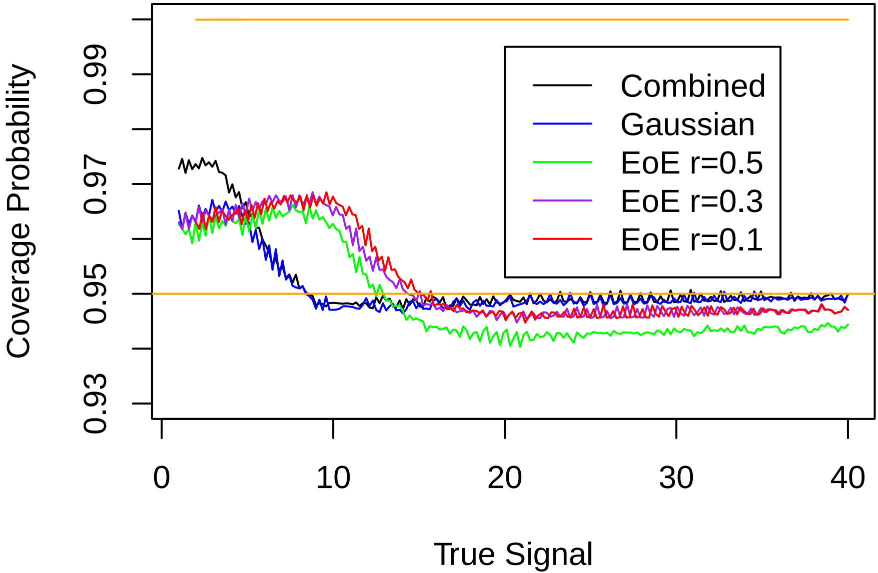
A very small example

- Goal: to illustrate extent to which independence assumption and gamma assumption could be flawed.
- To show these assumptions may not be really crucial.
- On-off. Main experiment observe $Y \sim$Poisson$(s + b)$. Unidentified model for On.
- Auxiliary experiment observe $X \sim$Poisson(b).
- We have $\hat{b}_a = X$. Estimated variance of $\hat{b}$ is $V = X$ from neg Hessian.
- Three likelihoods to compare: ℓ_{comb}, ℓ_{HEP} and ℓ_{EoE}.
- Do low statistics strong (apparent) signal example: $Y = 40$, $X = 10$.

Deviance Drop



Coverage probabilities



Slides commentary

- Even in this small example impacts not huge at coverage 95%.
- Choice of r matters to coverage.
- Discrete data gives unsmooth coverage probs
- The profile score for the background can have two local maxima in the permissible parameter space.

Meta Analysis and Auxiliary Information

- Statisticians have a long history of focusing on analyzing each individual data set as if it were the only data to be had.
- In Meta Analysis several experiments / studies are analyzed together parameters of interest in common but different nuisance parameters in each.
- Typical problem: different biases for the estimates of the common parameters of interest. Handled in a variety of somewhat unsatisfactory ways.
- Reminds me of problem of nearly 20 years ago: parton distribution function estimates differ between experimental groups by huge number of standard errors.
- In survey samples it is common to have control measurements produced from census data or to have paradata about respondents (e.g. how hard you had to work to get a response)
- In the medical literature use of empirical likelihood.

What needs doing; things to investigate maybe

- Models with many parameters: risk of accumulating bias.
- Maximum likelihood underestimates variance parameters.
  - In linear regression MLEs of variances are adjusted to remove bias.
  - Variance estimates are independent of mean estimates, have gamma distributions, and known degrees of freedom.
- In other models the location estimates are skewed and not independent of the variability estimates. We need more study of the impact of multiple slightly skewed estimates.
- There is work by, for example Zou, Hastie, and Tibshirani on degrees of freedom in variable selection.

Bayes has its advantages

- I am skeptical of the wisdom of treating theory uncertainty via frequentist methods.
- That said it seems the theory uncertainties are not cleanly handled any other way.
- The posterior from the auxiliary experiment captures any nuance in the distribution of the uncertainty in γ.
- So routine use of Bayesian methods would give sensible combinations of evidence.
- Bayes susceptible, in complex models, to hidden strengths of assumptions encoded in priors.
- Computer experiment literature full of uncertainty assignments to things like theory calculations where there is no frequentist uncertainty.

Three References

- H Zou, T Hastie, R. Tibshirani (2007) On the “degrees of freedom” of the LASSO. *Annals of Statistics*, **25**,2173-2192.
- C Chen, P Han, F He (2022) Improving main analysis by borrowing information from auxiliary data. *Statistics in Medicine*, **41**, 567-579.
- M Borenstein, L Hedges, J Higgins, H Rothstein (2021) *Introduction to Meta-Analysis*, 2nd ed. Wiley,