Efficiency of design-based inference

Thomas Lumley

Biostatistics
University of Washington

Copenhagen — 2009–4–3

Simulations

An advantage of using R for survey analysis is easier simulations.

A case–control study could be analyzed as a stratified random sample (stratified on case status), as survey statisticians do, or by maximum likelihood, as biostatisticians do.

We can compare these analyses by simulation [cf Scott & Wild, JRSSB 2000].

```
population<-data.frame(x=rnorm(10000))
expit <- function(eta) exp(eta)/(1+exp(eta))
population$y<- rbinom(10000, 1, expit(population$x-4))
population$wt<-ifelse(population$y==1, 1,(10000-sum(population$y))/600)

one.sample<-function(){
   cases<-which(population$y==1)
   controls<-sample(which(population$y==0),600)
   population[c(cases,controls),]
}
```

Simulations

```
mle<-function(sample) coef(glm(y~x, data=sample, family=binomial()))
svy<-function(sample){
   design<-svydesign(id=~1,strat=~y, weight=~wt, data=sample)
   coef(svyglm(y~x, design=design, family=quasibinomial()))
}
```

Testing shows that it seems to work, and 3 replicates takes about 1 second. 1000 replicates will take about 5 minutes.

```
> replicate(3, {d<-one.sample(); c(mle(d), svy(d))})
                    [,1]       [,2]       [,3]
(Intercept) -1.1575190 -1.1570389 -1.1347550
x            0.9621509  0.9834298  0.9640187
(Intercept) -3.9351437 -3.9335068 -3.8950293
x            0.9514965  0.9725670  0.9194183
```

Simulations

The true parameter values are $(-4, 1)$. Based on 1000 simulations the bias and variability of the estimators are

```
> results<-replicate(1000, {d<-one.sample(); c(mle(d), svy(d))})
> round(apply(results,1,median)-c(-4,1),2)
(Intercept)           x (Intercept)           x
       2.85        0.01        0.08       -0.02
> round(apply(results,1,mad),3)
(Intercept)           x (Intercept)           x
      0.017       0.060       0.035       0.080
```

So the MLE has a relative efficiency of nearly 2 for the slope estimate and is biased for the intercept, as expected.

Misspecified model

Try the analysis when the true model is not linear

$$\text{logit}P[Y = 1|X = x] = e^x - 4$$

but still fit

$$\text{logit}P[Y = 1|X = x] = \alpha + \beta x$$

'True' answer fitting logistic model to population is $(-3.66, 0.75)$

```
> round(apply(results,1,median)-c(-3.6625,0.7518),2)
(Intercept)           x (Intercept)           x
       2.84       -0.16        0.00        0.00
> round(apply(results,1,mad),3)
(Intercept)           x (Intercept)           x
      0.005       0.033       0.013       0.049
```

MLE is biased for population result but has lower variance.

Case-control example

`data(esoph)` in R are data from a case–control study of esophageal cancer in Ile-et-Vilaine, in northern France. The study examined smoking and alcohol consumption as risk factors.

Alcohol consumption is coded into categories 0-39, 40-79, 80-119, 120+ grams/day, smoking into 0-9, 10-19, 20-29, 30+ grams/day.

Age is an extremely strong risk factor, and so is quite likely to be a confounder. It is coded in 10-year categories.

Case-control example

```
> summary(model1)
Call:
glm(formula = cbind(ncases, ncontrols) ~ agegp + tobgp + alcgp,
    family = binomial(), data = esoph)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6891  -0.5618  -0.2168   0.2314   2.0643

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9108     1.0302  -5.737 9.61e-09 ***
agegp35-44    1.6095     1.0676   1.508 0.131652
agegp45-54    2.9752     1.0242   2.905 0.003675 **
agegp55-64    3.3584     1.0198   3.293 0.000991 ***
agegp65-74    3.7270     1.0253   3.635 0.000278 ***
agegp75+      3.6818     1.0645   3.459 0.000543 ***
tobgp10-19    0.3407     0.2054   1.659 0.097159 .
tobgp20-29    0.3962     0.2456   1.613 0.106708
tobgp30+      0.8677     0.2765   3.138 0.001701 **
alcgp40-79    1.1216     0.2384   4.704 2.55e-06 ***
alcgp80-119   1.4471     0.2628   5.506 3.68e-08 ***
alcgp120+     2.1154     0.2876   7.356 1.90e-13 ***
```

Case-control example

Alcohol and tobacco both show increasing association with increasing dose. We need tests for the set of coefficients for each of these variables.

Likelihood ratio tests are natural for maximum likelihood

```
>  anova(update(model1,.~.-tobgp), model1, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ agegp + alcgp
Model 2: cbind(ncases, ncontrols) ~ agegp + tobgp + alcgp
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        79     64.572
2        76     53.973  3   10.599     0.014
```

Case-control example

```
>  anova(update(model1,.~.-alcgp), model1, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ agegp + tobgp
Model 2: cbind(ncases, ncontrols) ~ agegp + tobgp + alcgp
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        79    120.028
2        76     53.973  3   66.054 2.984e-14
```

Case-control example

Wald tests can be extended more easily to probability-weighted models (working LRT is coming soon).

```
> library(survey)
> regTermTest(model1,~alcgp)
Wald test for alcgp
 in glm(formula = cbind(ncases, ncontrols) ~ agegp + tobgp + alcgp,
    family = binomial(), data = esoph)
Chisq =  57.89887  on  3  df: p= 1.652e-12
> regTermTest(model1,~tobgp)
Wald test for tobgp
 in glm(formula = cbind(ncases, ncontrols) ~ agegp + tobgp + alcgp,
    family = binomial(), data = esoph)
Chisq =  10.76880  on  3  df: p= 0.013044
```

Case-control example

We do not have the sampling fractions for controls, but the incidence of esophageal cancer is about 5/100,000 per year. We do not know how many years of cases were recruited, but it is reasonable to assume the sampling fraction is no smaller than 1/2000 for controls.

Using the `survey` package we can fit the same model with probability weights. It is necessary first to expand the data to one record per person, then to declare the sampling design, then use `svyglm` to fit the model.

The estimated odds ratios based on a sampling fraction of 1/2000 are very similar to the unweighted analysis, and the intercept is different, as expected.

Case-control example

```
> desoph<-svydesign(id=~1, weight=~prob, data=expanded)
> pmodel1<-svyglm(cancer~agegp+tobgp+alcgp, design=desoph, family=binomial)

> round(rbind(unweighted=coef(model1), weighted=coef(pmodel1)),3)

           (Intercept) agegp35-44 agegp45-54 agegp55-64 agegp65-74 agegp75+
unweighted      -5.911      1.610      2.975      3.358      3.727    3.682
weighted       -13.422      1.607      2.995      3.321      3.644    3.657

           tobgp10-19 tobgp20-29 tobgp30+ alcgp40-79 alcgp80-119 alcgp120+
unweighted      0.341      0.396    0.868      1.122       1.447     2.115
weighted        0.275      0.340    0.782      1.104       1.437     2.004
```

Case-control example

In this example the standard errors are also similar, so the design-based estimate is nearly efficient

```
> round(rbind(unweighted=SE(model1), weighted=SE(pmodel1)),2)

            (Intercept) agegp35-44 agegp45-54 agegp55-64 agegp65-74 agegp75+
unweighted        1.03       1.07       1.02       1.02       1.03     1.06
weighted          1.01       1.06       1.02       1.01       1.02     1.06

            tobgp10-19 tobgp20-29 tobgp30+ alcgp40-79 alcgp80-119 alcgp120+
unweighted        0.21       0.25     0.28       0.24        0.26      0.29
weighted          0.21       0.25     0.29       0.24        0.27      0.30
```

Efficiency?

Statisticians tend to do the simulations for Normal x, where the efficiency turns out to be low.

Design-based estimator is fully efficient at $\beta = 0$, and is fully efficient for saturated models.

Efficiency is quite high for categorical predictors where cell size isn't too small.

Distribution	$\beta = 0$	$\beta = 0.5$	$\beta = 1$	$\beta = 1.5$	$\beta = 2$
Normal	100	85	48	20	22
Square	100	96	94	92	97
Triangular	100	94	84	76	77

Is the gain in efficiency real?

Gene–environment independence

Consider a case–control study of drug–gene interaction

$$\text{logit } P[Y = 1] = \alpha + \beta_g I(gene) + \beta_d I(drug) + \gamma I(drug \cap gene)$$

Often plausible that drug and genetic variant are independent in the population.

The interaction parameter e^{γ} is called the synergy index.

Finding $\gamma \neq 0$ is useful if there are alternative drugs with similar average effectiveness (eg blood pressure)

Example: diuretics and α-adducin

A variant in the α-adducin protein in rats causes salt-sensitive hypertension that responds particularly well to thiazide diuretics.

We looked at the same variant in people, to see if thiazides protect against heart attack and stroke more effectively in the presence of this variant [Psaty et al, JAMA, 2000].

Everyone has treated hypertension so $D = 1$ is thiazides, $D = 0$ is all other blood pressure drugs.

		G	
	D	0	1
Case	0	103	85
	1	94	41
Control	0	248	131
	1	208	128

Example: diuretics and α-adducin

For $2 \times 2 \times 2$ table and rare events, case-only estimation exploits the independence [Piegorsch et al, 1994]

		G	
	E	0	1
Case	0	a	b
	1	c	d
Control	0	e	f
	1	g	h

case-control estimator:

$$\frac{ag/ce}{bh/df} = \frac{ad/bc}{eh/gf}$$

case-only estimator:

$$\frac{ad}{bc}$$

For thiazides and α-adducin:

Case-only: 0.45 (0.33–0.84) Case–control 0.53 (0.26–0.79).

Efficiency

- The case-only estimator is the MLE under the assumption of gene–drug independence in controls.

- The variance of the case-only estimator is the same as for a case–control estimator with infinitely many controls, or half the variance of a case–control estimator with 1:1 sampling.

- The case-only estimator is biased if gene and drug are not independent in controls

If we know a priori that independence holds, we can use the case–only estimator.

Can we ask the data?

We can estimate the gene–drug log odds ratio in controls from the data.

$$\hat{\psi} = \log \frac{eg}{gf}$$

Its estimated variance is

$$\widehat{\mathrm{var}}[\hat{\psi}] = \frac{1}{e} + \frac{1}{f} + \frac{1}{g} + \frac{1}{h}$$

So we can reliably detect bias when

$$\left| \log \frac{eg}{gf} \right|^2 \gg \frac{1}{e} + \frac{1}{f} + \frac{1}{g} + \frac{1}{h}$$

and in the absence of a priori knowledge we should expect bias of about

$$\left| \log \frac{eg}{gf} \right|^2 \sim \frac{1}{e} + \frac{1}{f} + \frac{1}{g} + \frac{1}{h}$$

Efficiency again

Suppose we have no a priori reason to believe that bias is much smaller than the detectable level.

The efficiency gains from the case–only analysis are that the variance is reduced from

$$\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}+\frac{1}{e}+\frac{1}{f}+\frac{1}{g}+\frac{1}{h}$$

to

$$\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}$$

a reduction in MSE of

$$\frac{1}{e}+\frac{1}{f}+\frac{1}{g}+\frac{1}{h}$$

Efficiency again

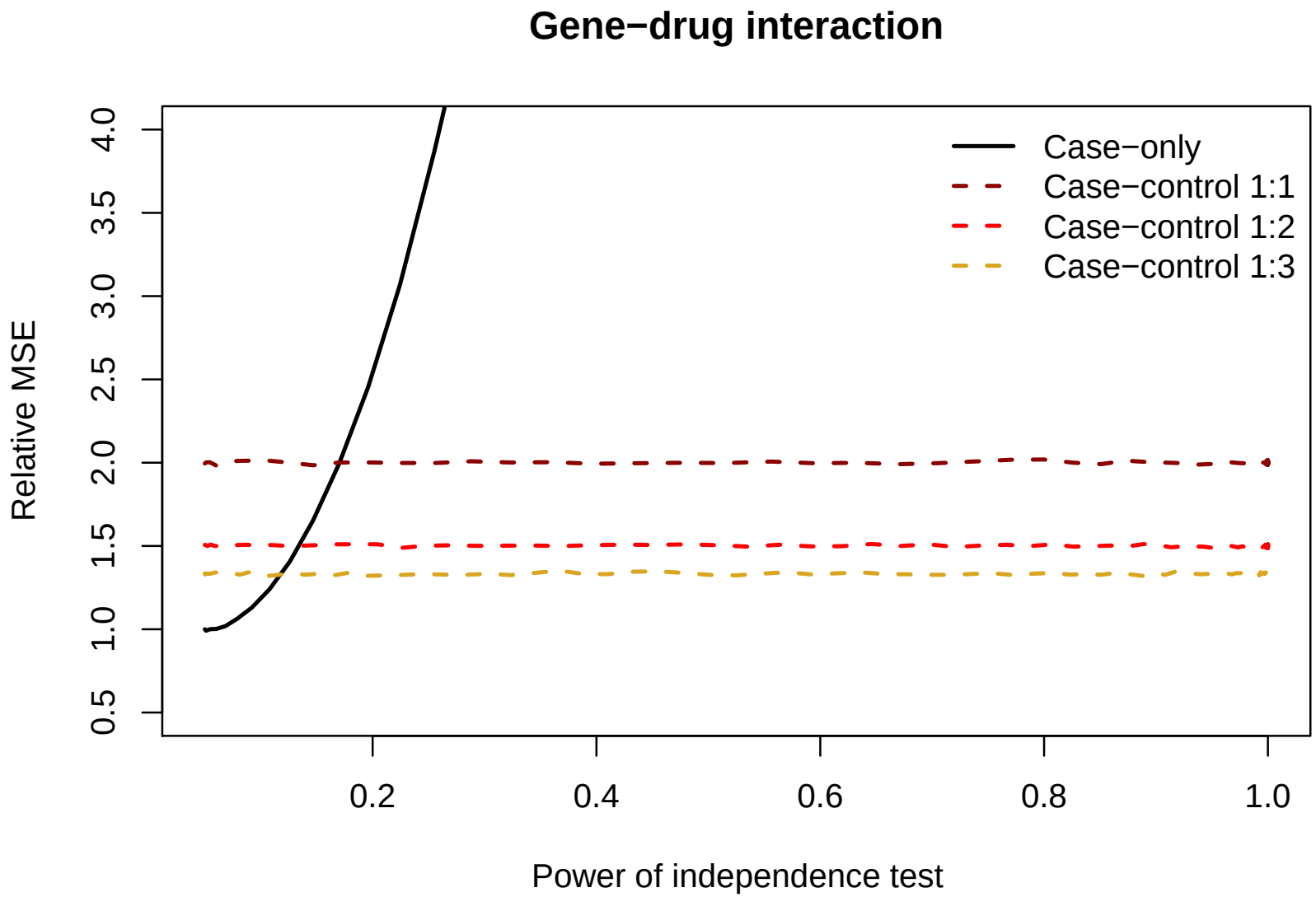
If we cannot detect bias, we still expect the squared bias to be around

$$\frac{1}{e} + \frac{1}{f} + \frac{1}{g} + \frac{1}{h}$$

so the added bias term in the MSE of the case–only estimator will typically be about the same as the variance reduction.

Case–only estimator should have similar MSE to the case–control estimator (and worse interval coverage).

Efficiency again



Messages

- A large efficiency gain is still only an $O(n^{-1/2})$ change in the estimator, so $O(n^{-1/2})$ biases can counteract it.

- The bias increase and variance reduction terms in the MSE are linked

- To trust the case-only estimator we need good a priori reasons to believe that the biases are much smaller than could be detected.

Shrinkage

A weighted combination of the case-only and case–control estimators is possible (Mukherjee & Chatterjee, Biometrics)

- Better worst–case than case–only
- Better best-case than case–control
- Does not dominate either estimator — **it can't**: they are both MLEs and thus efficient
- Not a regular estimator: null distribution is not quite Normal – can matter for whole-genome studies

Asymptotics

More complicated models than $2 \times 2 \times 2$ table need asymptotic approximations.

We are interested in 'nearly true' models, where the misspecification can't be reliably detected

Asymptotics for a fixed data-generating distribution are not useful: we can always distinguish correct from incorrect models with enough data.

If $P_n \in \mathcal{P}$ is a sequence of distributions exactly satisfying a model $\mathcal{P}$, say that the model is **nearly true** in a sequence Q_n that is **mutually contiguous** with P_n.

Contiguity

Contiguity means (equivalently)

- For any sequence of events A_n: $P_n(A_n) \to 0$ if and only if $Q_n(A_n) \to 0$

- The likelihood ratio Q_n/P_n is bounded in probability under Q_n.

In particular, even if we knew P_n and Q_n, the sequence of Neyman–Pearson tests for whether the data come from P_n or Q_n would not be consistent.

In the $2 \times 2 \times 2$ table, contiguity means exactly that $\psi^2/\text{var}[\widehat{\psi}]$ is bounded.

Two-phase studies

Sample a cohort of N people from population and measure some variables then subsample n of them and measure more variables [genes, biomarkers, coding of open-text data, copies of original medical records]

We have some semiparametric model (eg Cox, glm) that we would know how to fit with complete data, the **outcome model**

A simple estimator is the solution to

$$\sum_{i:R_i=1} \frac{1}{\pi_i} U_i(\theta) = 0$$

where $U_i(\theta)$ is any complete-data estimating function for the parameter we are estimating.

Calibration and AIPW estimators

Robins, Rotnitzky & Zhao defined Augmented IPW estimators for two-phase designs

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i} U_i(\theta) + \sum_{i=1}^{N} \left(1 - \frac{R_i}{\pi_i}\right) A_i(\theta) = 0$$

where $U_i(\theta)$ is the complete-data efficient score function and $A_i()$ can be any function of phase-1 data.

The optimal A_i is

$$A_i(\theta) = E[U_i(\theta)|\text{observed data}]$$

and using regression to estimate the expectations is easy and works reasonably well.

Calibration and AIPW estimators

All AIPW estimators are consistent for the same limit as if we had complete data, whether or not the outcome model is correct.

They are the only such estimators. The AIPW class includes the calibration estimators, and for any AIPW estimator there is a calibration estimator that is at least as efficient.

Calibration estimators are typically not fully semiparametric-efficient if we assume the model is correct.

The actual loss of efficiency can be substantial (eg 30–40% in some real examples)

Efficient estimator

RRZ also defined efficient estimators under the assumption that the outcome model is true.

The efficient estimator uses a different influence function $V_i(\theta)$ rather than the complete-data influence function $U_i(\theta)$, and the optimal A_i for that influence function.

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i} V_i(\theta) + \sum_{i=1}^{N} \left(1 - \frac{R_i}{\pi_i}\right) A_i^{(\text{opt})}(\theta) = 0$$

Calculating $V(\,)$ from its definition can be hard, profile likelihood is often used instead.

What is truth?

When the model is misspecified we need to define the target of estimation in order to talk about efficiency.

One reasonable general definition is the quantity that would be estimated if we had complete data. This has the advantage that our 'error' always decreases as n increases for fixed N.

Define θ^* as the limit of the estimator of θ from complete data, as $N \to \infty$.

[In particular situations we might prefer another target quantity (eg Scott & Wild 2002)]

We want to compare $\sqrt{n}\left(\widehat{\theta}_{\mathsf{AIPW}} - \theta^*\right)$ and $\sqrt{n}\left(\widehat{\theta}_{\mathsf{eff}} - \theta^*\right)$

Convolution theorem

When the model is correctly specified, the Convolution theorem tells us that $\widehat{\theta}_{\mathsf{cal}} - \widehat{\theta}_{\mathsf{eff}}$ is asymptotically independent of $\widehat{\theta}_{\mathsf{eff}}$

Suppose

$$\sqrt{n}\left(\widehat{\theta}_{\mathsf{eff}} - \theta^*\right) \stackrel{P_n}{\to} N(0, \sigma^2)$$

and

$$\sqrt{n}\left(\widehat{\theta}_{\mathsf{cal}} - \theta^*\right) \stackrel{P_n}{\to} N(0, \omega^2 + \sigma^2)$$

Then

$$\sqrt{n}\left(\widehat{\theta}_{\mathsf{eff}} - \theta_{\mathsf{cal}}\right) \stackrel{P_n}{\to} N(0, \omega^2)$$

We want to know what happens under the misspecified model Q_n.

LeCam

LeCam's Third Lemma relates the joint limiting distribution of a statistic X and the N–P test statistic $\log L$ under Q_n to their joint distribution under P_{θ_n}

$$X_n \overset{P_n}{\to} N(\mu, \tau^2)$$

then

$$X_n \overset{Q_n}{\to} N(\mu - 2\sqrt{k}\rho\tau, \tau^2)$$

where ρ is the correlation between X_n and $\log L$ under P_{θ_n} and k is the expected value of $\log L$.

k measures how misspecified the model is: we cannot reliably distinguish $k \sim 1$ from $k = 0$

LeCam

We will take $X_n = \sqrt{n}(\widehat{\theta}_{\text{eff}} - \widehat{\theta}_{\text{AIPW}})$ to get $\mu = 0$, $\tau = \omega$.

MSE of $\widehat{\theta}_{\text{eff}} = 4k\rho^2\omega^2 + \sigma^2$

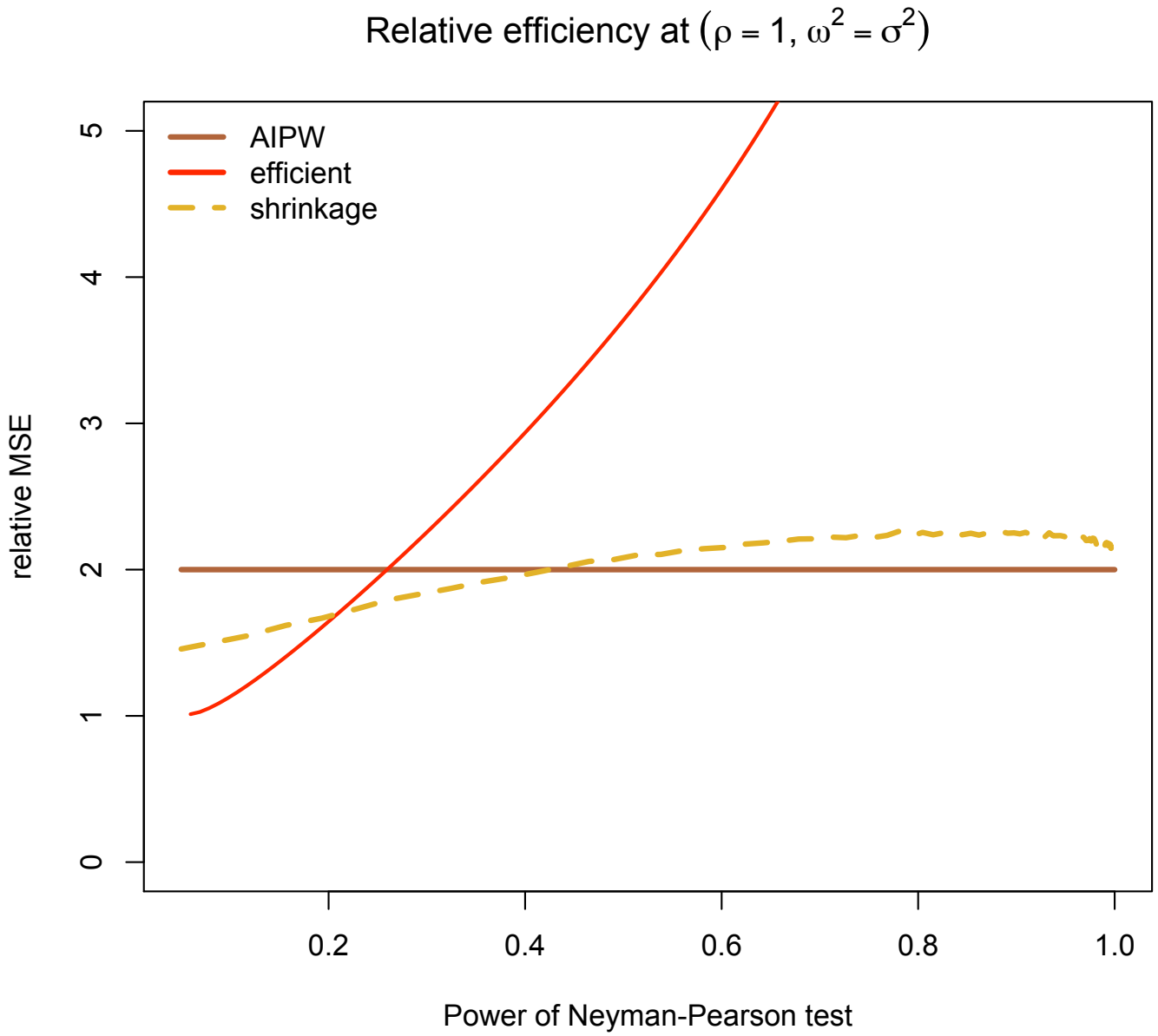
MSE of $\widehat{\theta}_{\text{cal}} = 0 + \left(\omega^2 + \sigma^2\right)$

Calibration estimator is more efficient if $4k\rho^2 > 1$

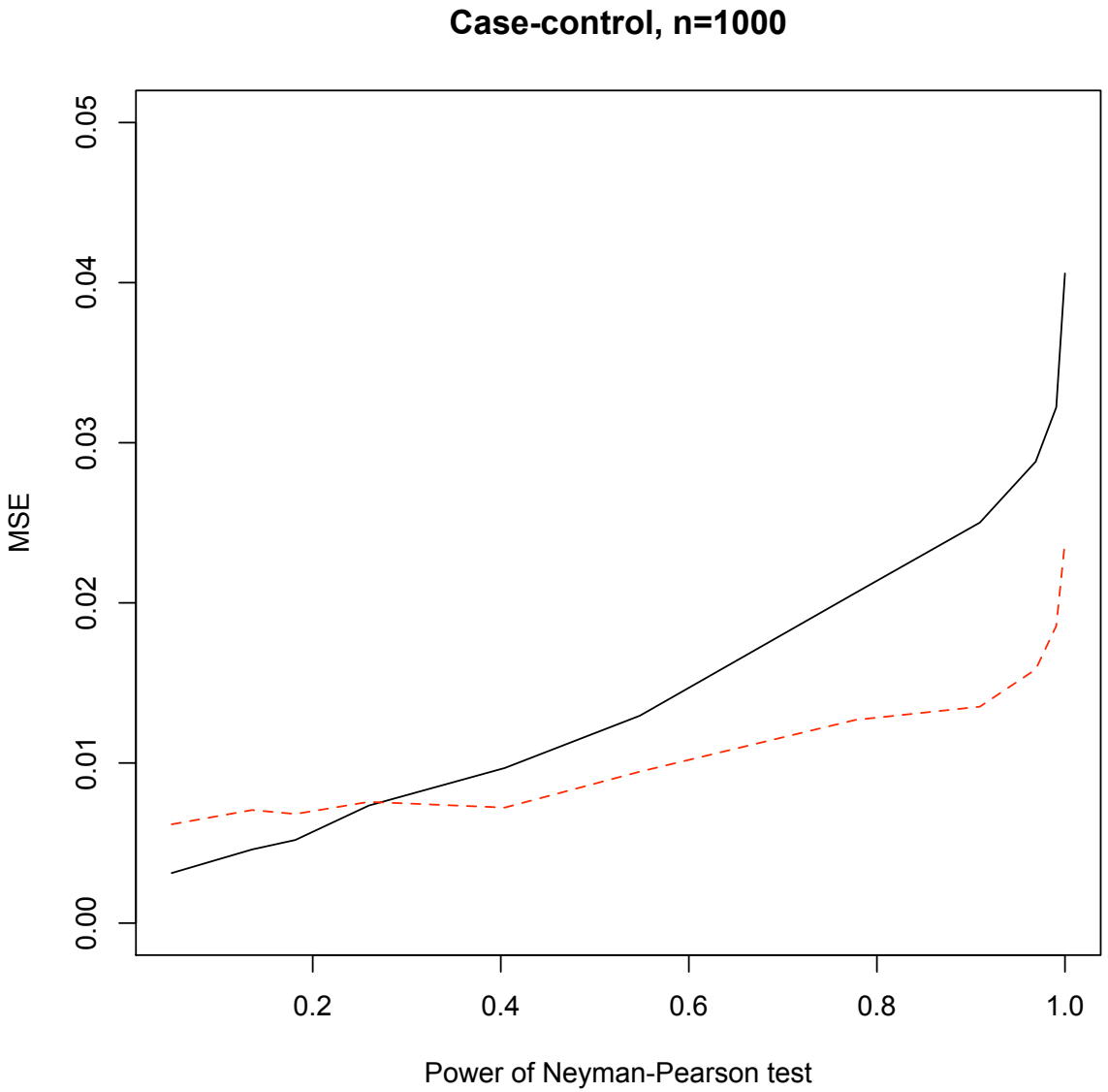
Theorem: if $\widehat{\theta}_{\text{cal}}$ is the best calibration estimator and is not semiparametric efficient we can get ρ arbitrarily close to 1 (by using $U_i(\theta^*) - V_i(\theta^*)$ to define the direction of misspecification).

For the Horvitz–Thompson estimator ρ may be bounded well away from 1.

Efficiency: asymptotic



Efficiency: empirical



Model robustness

Gaining information from assuming gene:environment independence or Hardy–Weinburg Equilibrium makes people nervous

Gaining information from assuming a perfectly specified outcome model does not appear to make (most) statisticians nervous.(!)

Issue can't be evaded by talking about diagnostics, goodness-of-fit, careful model specification:

- the information bound is strictly worse if you don't a priori know that the model is correct.

- Non-magical procedures cannot improve on the information bound in large samples (Convolution theorem/LAM)

Open questions

Where does the information come from: what parts of the assumptions can be weakened without losing the precision?

When are there good reasons to prefer $\lim_{n,N\to\infty} \widehat{\theta}_{\text{eff}}$ over θ^* as the target of estimation?

What actually happens in finite samples?

Final notes

- Caring about efficiency commits you to caring about $O(n^{-1/2})$ biases

- If there is substantial extra work in constructing the efficient estimator it may not be justified

- LeCam's Third Lemma is extremely cool

If you liked today's talks and want to read more

or

If you didn't like today's talks and want a book instead

Complex Surveys: a guide to analysis using R

coming later this year from Wiley .