Using the whole cohort in analysis of subsampled data.

Thomas Lumley

Dept of Biostatistics
University of Washington.

Program in Computational Biology
Fred Hutchinson Cancer Research Center

Biometrics on the Lake — 2009–12–2

Outline

Weighted estimators in two-phase designs are not semiparametric-efficient. In some models the efficient estimator is known. Should we be using it?

- Two-phase designs and estimators: using the whole cohort

- Sensitivity to model misspecification: nearly-true models

Theme: where does the information come from?

Two-phase studies

Sample a cohort of N people from population and measure some variables (Z, Y) then subsample n of them and measure more variables X.

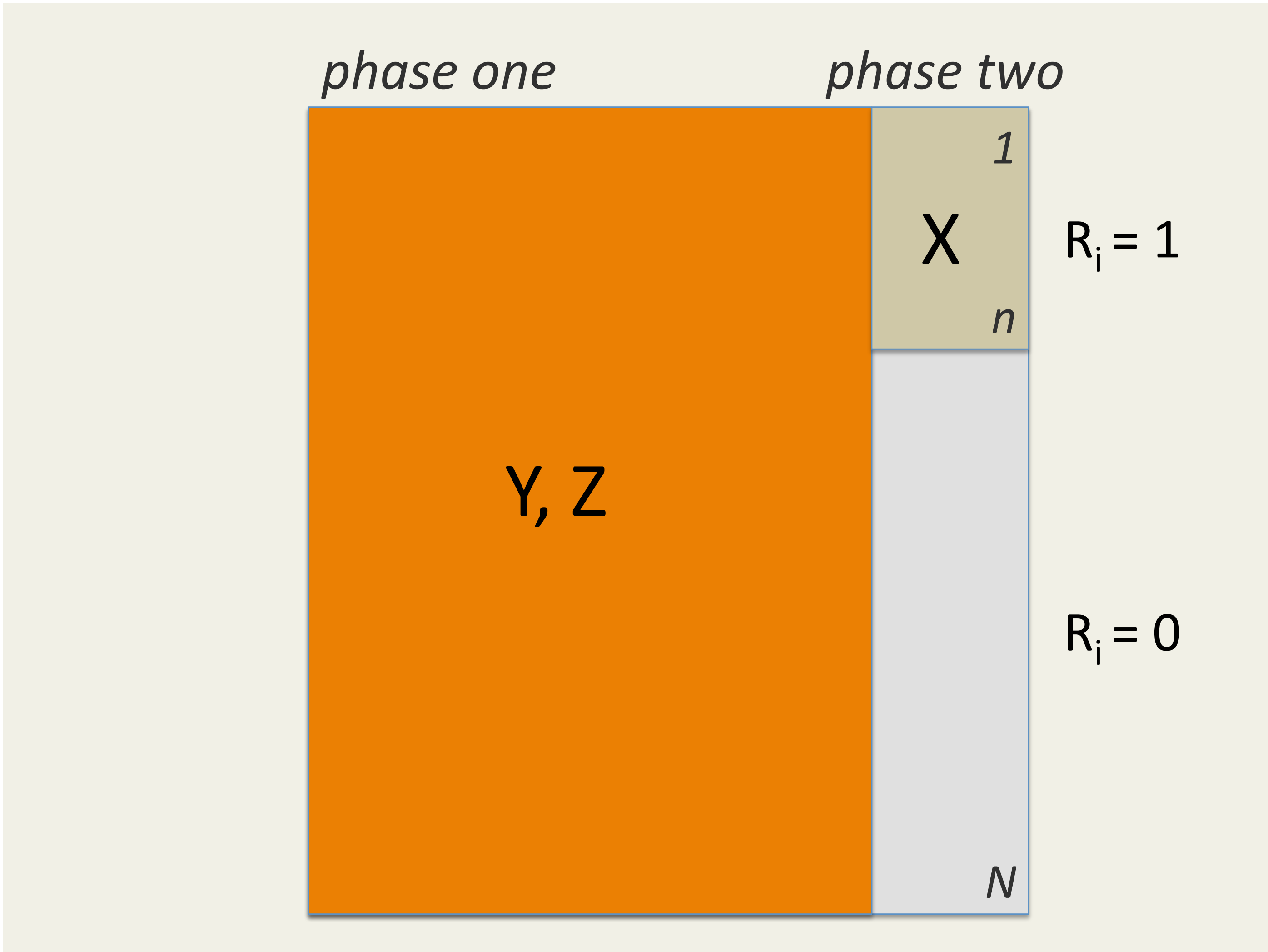
$R_i = 1$ indicates that person i is sampled. The sampling probabilities $E[R_i|Z, Y] = \pi_i$ are known for everyone in the sample.

Examples:

- New measurements: new assays on frozen blood samples
- Coding of free-text responses
- Validation of self-report, discharge diagnosis, billing code
- Validation of cheap assay with accurate assay

[sampling should be stratified as finely as possible: not covered]

Two-phase studies



Modelling

We have some semiparametric **outcome model** (eg logistic, Cox) for Y that we would know how to fit with complete data, by solving estimating equations.

$$\sum_{i=1}^{N} U_i(\theta) = 0$$

for example,

$$U_i = x_i(y_i - \mu_i)$$

for logistic or linear regression.

Modelling

A simple estimator in subsampled data is the 'Horvitz-Thompson' estimator, the solution to

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

Since $R_i \perp U_i \,|\, Y, Z$:

$$E\left[\sum_{i=1}^{N} U_i(\theta)\right] = E\left[\sum_{i=1}^{N} \frac{R_i}{\pi_i} U_i(\theta)\right]$$

whether or not Y really follows the outcome model.

Better estimators

Robins, Rotnitzky & Zhao (JASA, 1995) defined Augmented IPW estimators for two-phase designs.

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

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

The AIPW estimator can be rewritten as

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

using A_i from the whole cohort, residuals $U_i - A_i$ from the subsample.

Better estimators

(Asymptotically) we can guarantee the AIPW is at least as good at the HT estimator by adding parameters

Write $\widehat{U}_i(\theta) = \alpha_0 + \alpha_1 A_i(\theta)$, use linear regression to estimate α

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i} \left[U_i(\theta) - \widehat{U}_i(\theta) \right] + \sum_{i=1}^{N} \widehat{U}_i(\theta) = 0$$

A survey regression estimator of the population total of $U_i(\theta)$, using $\widehat{U}_i(\theta)$ **from the whole cohort.**

The same as **survey calibration estimators**, implemented in the R survey package.

[Deville & Särndal, 1994 JASA]

Which auxiliary variables?

$A_i(\theta)$ needs to be **linearly** correlated with $U_i(\theta)$, so it needs to be an estimating function for a similar model.

Strategy

- Impute X using Y, Z
- Fit a model to phase-one data using **imputed** X
- Extract score or influence functions from the model, use as $A_i(\theta)$
- Refit model to subsample using A_i as calibration variables.

Similar **efficiency** to just relying on imputation, but **validity** does not rely on valid imputation model.

Example: Wilms' Tumor

Rare childhood kidney tumour, 90% curable.

National Wilms' Tumor Study Group clinical trials recruit nearly all cases.

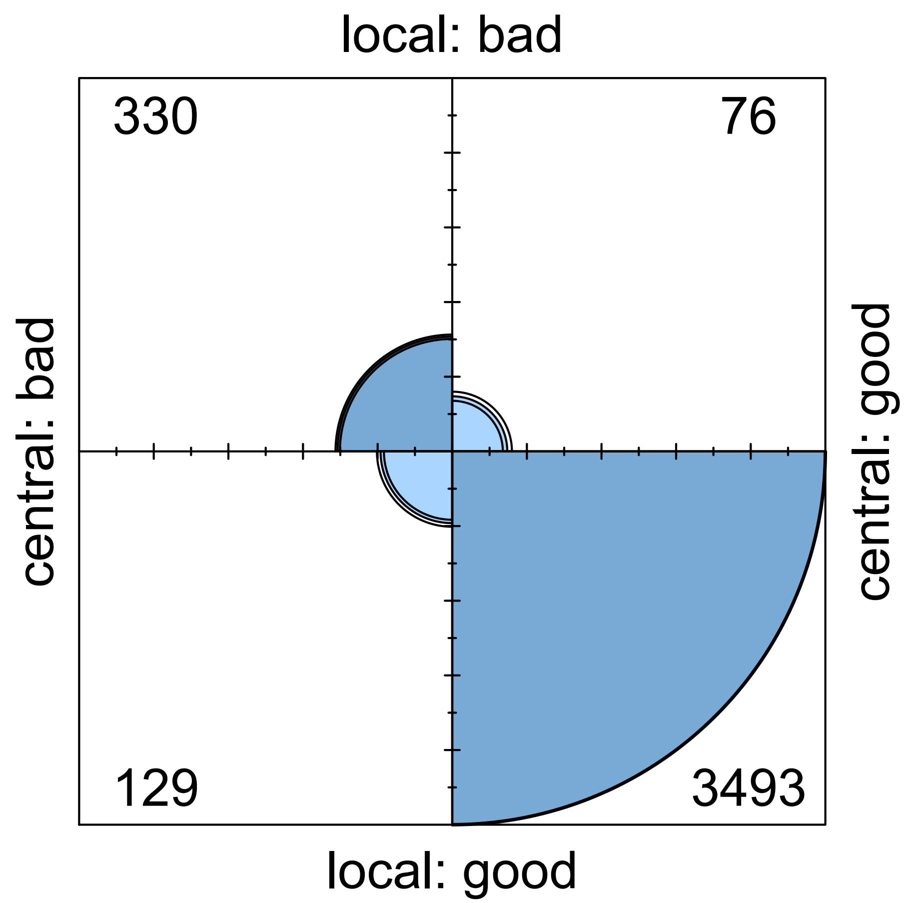
Interest is in long-term outcomes: identifying lower-risk patients for less aggressive treatment.

NWTS central pathologist invented the histologic classifications, is more accurate than anyone else.

Unfavorable (high-risk) histology is rare (about 10%). Other hospital pathologists miss about 25% of the high-risk samples.

[Breslow, Lumley et al., Am J Epi 2009; Breslow, Lumley, et al., Stats in Biosciences 2009; Kulich & Lin JASA 2004]

Example: Wilms' Tumor



Example: Wilms' Tumor

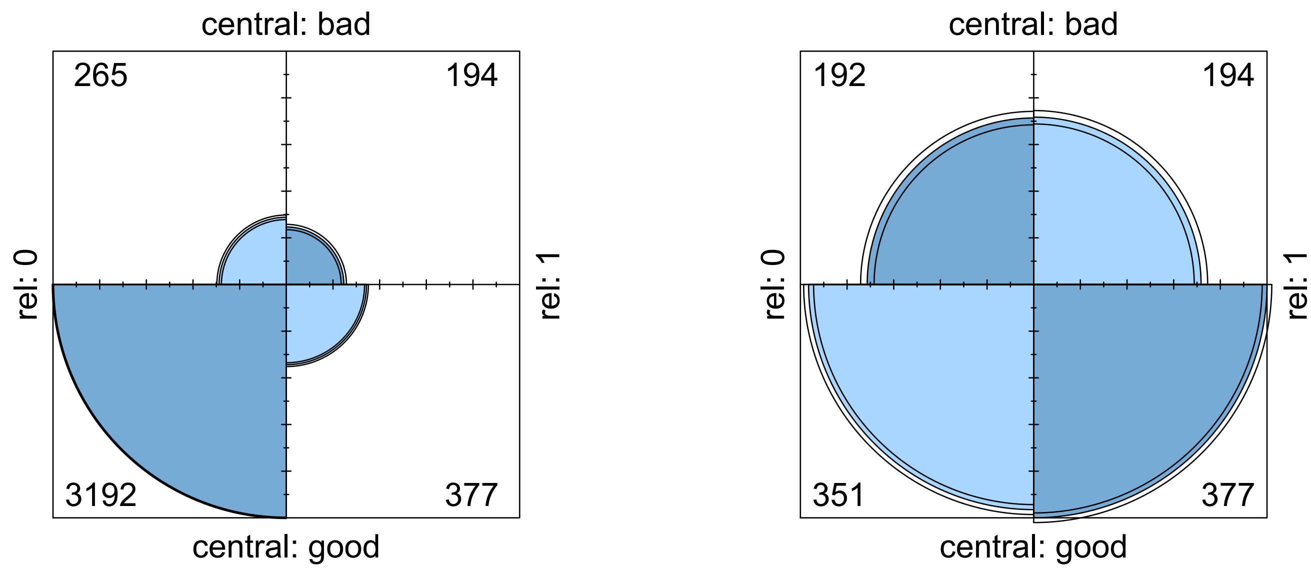
Imaginary two-phase design

- Classify on outcome (relapse) and local-hospital histology (favorable/unfavorable)

- Sample all relapses, all unfavorable histology according to local lab, 10% of remainder (total about 25%)

- At phase two, determine central-lab histology on subsample.

Example: Wilms' Tumor



Balance in histology $\times$ response categories better than simple random sample or case-control sample.

Example: Wilms' Tumor

Analysis

- Impute central-lab histology using outcome, local-hospital histology, other covariates
- Fit logistic model to phase-one data using imputed central-lab histology, to estimate effect of histology on relapse
- Extract influence functions from the model, use as $A_i(\theta)$
- Fit the same logistic model to phase-two data using measured central-lab histology.

Compare: HT estimator, calibrated estimator, imputation estimator, full data.

Example: Wilms' Tumor

	Two-phase sample			full data
	HT	calibration	imputation	
Coefficient estimate				
histology	1.808	2.113	2.108	1.932
age	0.055	0.101	0.101	0.096
stage > 2	1.411	1.435	1.432	1.389
tumor diameter	0.043	0.061	0.061	0.058
histology:age	-0.116	-0.159	-0.159	-0.144
stage$>$ 2:diameter	-0.074	-0.084	-0.083	-0.079
Standard error				
histology	0.221	0.171	0.174	0.157
age	0.023	0.014	0.016	0.016
stage > 2	0.361	0.276	0.249	0.250
tumor diameter	0.021	0.016	0.014	0.014
histology:age	0.054	0.039	0.040	0.035
stage $>$ 2:diameter	0.030	0.022	0.020	0.020

Example: Wilms' Tumor

- Imputation uses data from whole cohort, biased if imputation model is wrong.
- Calibration uses data from whole cohort, loss of efficiency but no bias if imputation model is wrong
- Using data from the whole cohort increases efficiency for all coefficients
- Gain is larger for variables available on whole cohort.

Validity, Efficiency

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.

If we **assume** the outcome model is correct there are usually more efficient estimators.

The actual loss of efficiency when the outcome model is known to be correct can be substantial.

Efficient estimator

Robins, Rotnitzky, and Zhao (JASA, 1995) also characterized efficient estimators **assuming exactly correct outcome model.**

Calculating the estimator from this definition can be hard: profile likelihood approaches are more practical.

Methods and software available for generalized linear models with discrete data at phase one. (Scott, Wild, co-workers)

Loss of efficiency

The efficient estimators can be usefully more efficient in some examples.

Simplest example is classical case–control design:

- unweighted logistic regression is efficient
- logistic regression using sampling weights is best AIPW estimator.

Efficiency can be as low as 50% in realistic simulations.

Where does the extra information come from?

We already used $E[U_i(\theta)|\textit{whole cohort}]$

Loss of efficiency

Loss of efficiency in weighted logistic regression often attributed to variability in sampling weights.

Explanation doesn't fit facts. Regardless of variation in sampling weights, weighted logistic regression:

- is fully efficient at $\beta = 0$
- is fully efficient for saturated models
- is close to fully efficient with small numbers of discrete covariates.

Efficiency gain is actually related to power for detecting model misspecification: relying on the model helps most when it is hard to validate.

Asymptotics for misspecification

Need asymptotic approximations because exact distributions are too complicated

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

Among 'nearly true' models we will look at the worst-case misspecification (more later).

We **stipulate** that the efficient estimator is asymptotically well-behaved when the model is correct.

Asymptotics for misspecification

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 if

- 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.

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.

What is truth?

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

One reasonable definition is the quantity that would be estimated **if we had complete data**: that is how two-phase sampling is motivated, and is certainly the target of inference with missing data.

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

Theorem

Suppose under P_n

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

and

$$\sqrt{n}\left(\widehat{\theta}_{\text{AIPW}} - \theta^*\right) \xrightarrow{P_n} N(0, \omega^2 + \sigma^2)$$

Then under Q_n

$$\sqrt{n}\left(\widehat{\theta}_{\text{eff}} - \theta^*\right) \xrightarrow{P_n} N(-2k\rho\omega, \sigma^2)$$

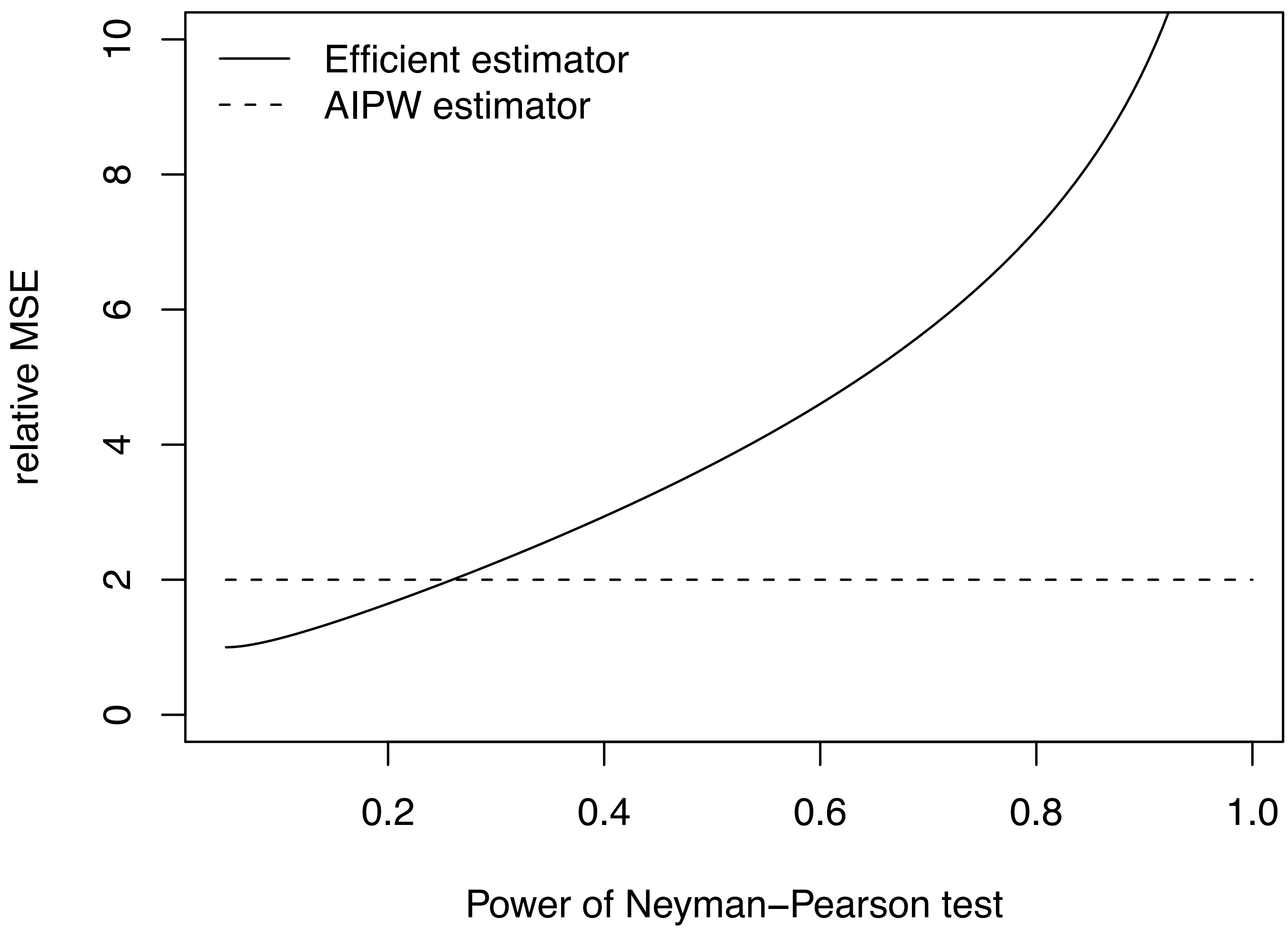
and

$$\sqrt{n}\left(\widehat{\theta}_{\text{AIPW}} - \theta^*\right) \xrightarrow{P_n} N(0, \omega^2 + \sigma^2)$$

where k measures the size of misspecification and ρ measures the direction

[Convolution theorem plus LeCam's third lemma, applied to $\sqrt{n}(\widehat{\theta}_{\text{AIPW}} - \widehat{\theta}_{\text{eff}})$]

Efficiency: asymptotic



Efficiency: asymptotic

Simulations

SImplest two-phase model: case–control

$X \sim N(0,1)$, outcome model based on logistic model

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

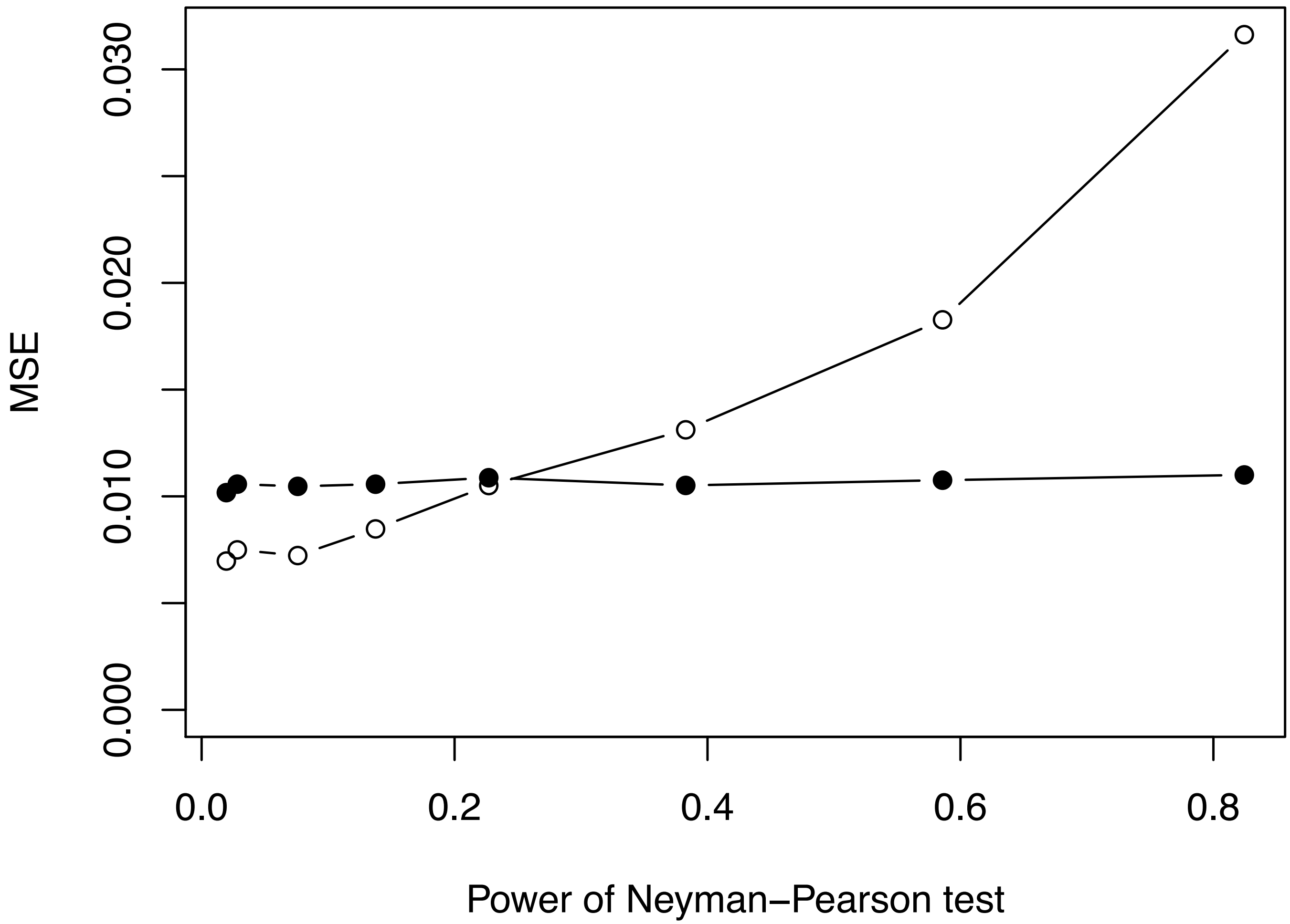
with $(\alpha, \beta) = (-3.5, 1)$, distorted in the most unfavorable direction.

Measure Y at phase 1 in population $N = 10000$, subsample all (≈ 500) cases and same number of controls, and measure X.

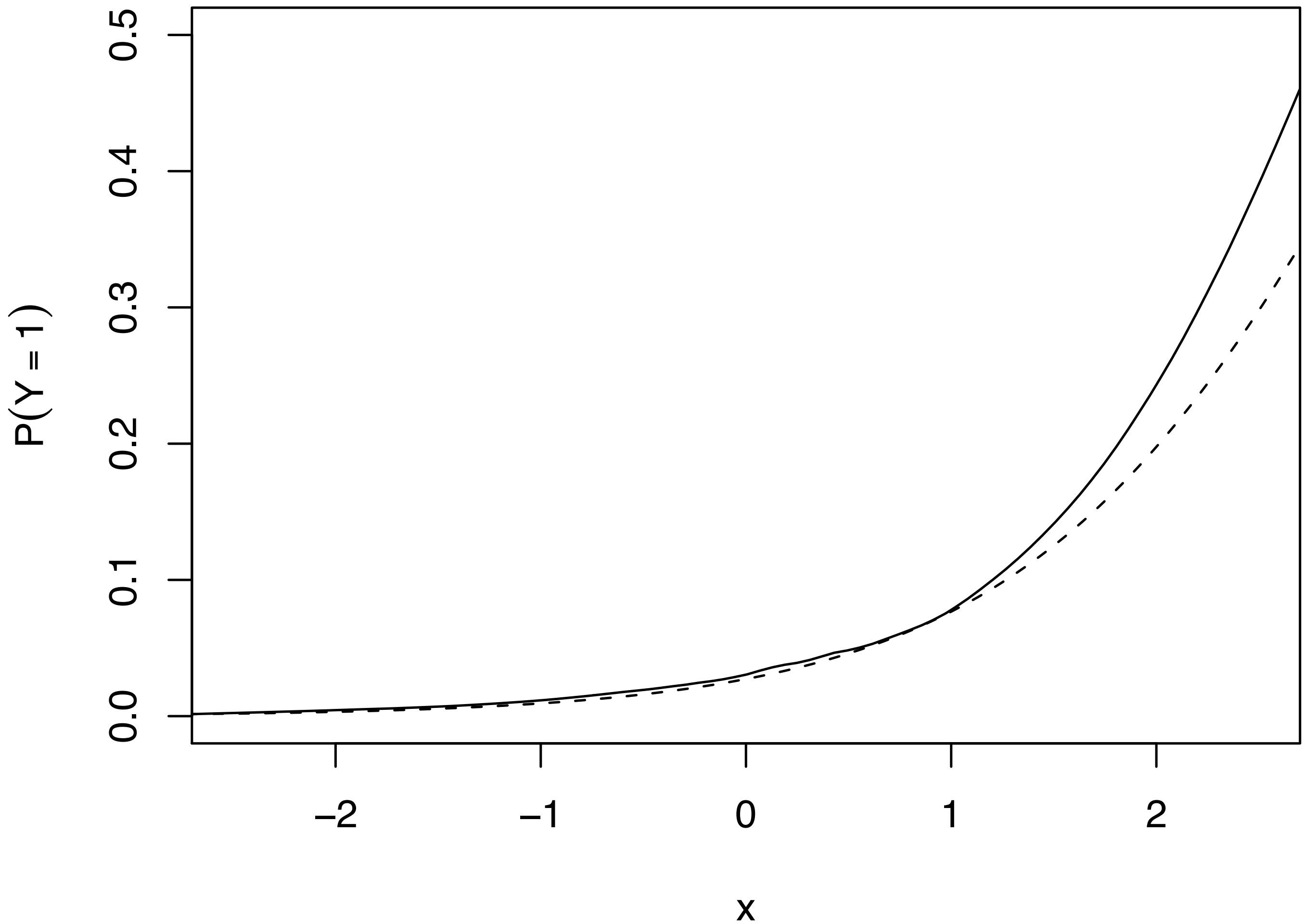
Compare maximum likelihood and weighted likelihood, fitting the misspecified model

$$\text{logit}\, E[Y=1] = \alpha + X\beta$$

Efficiency: empirical



Small bias (is enough)



Model robustness

Extra precision is available, but only if the model is **known to be correct.**

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)

Tradeoffs

If the model is correct, $(\widehat{\theta}_{\text{eff}} - \widehat{\theta}_{\text{AIPW}})$ is the **gain** from using the efficient estimator.

If the model is not correct, $E\left[\widehat{\theta}_{\text{eff}} - \widehat{\theta}_{\text{AIPW}}\right]$ is the **bias**

- If ω^2 is small the threshold for undetectable bias is strict, but the gain in precision from the efficient estimator is small

- If ω^2 is large, the gain in precision from the efficient estimator is large, but undetectable biases can be quite large.

A useful efficiency gain is only available if you really know the model is true. . .

...or in other words

http://www.flickr.com/photos/eric/8850/

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
- Behavior under contiguous model misspecification ('nearly true', 'local asymptotic minimax') is a useful way to think about estimators.
- Constructing a reasonably good AIPW estimator is worthwhile (and not that hard).

Technical report available from `http://www.bepress.com/uwbiostat/`

Slides from `http://faculty.washington.edu/tlumley/taupo.pdf`