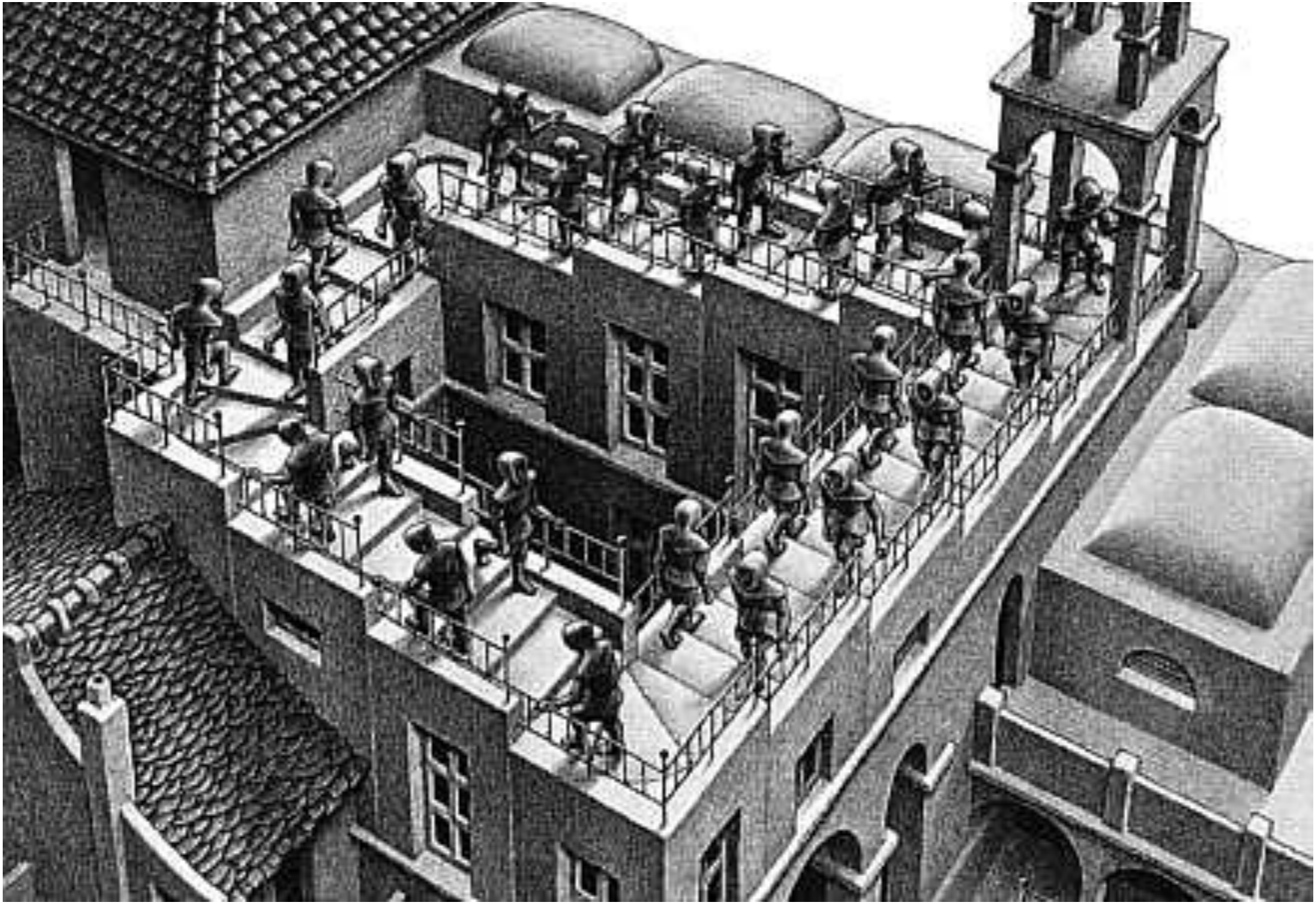
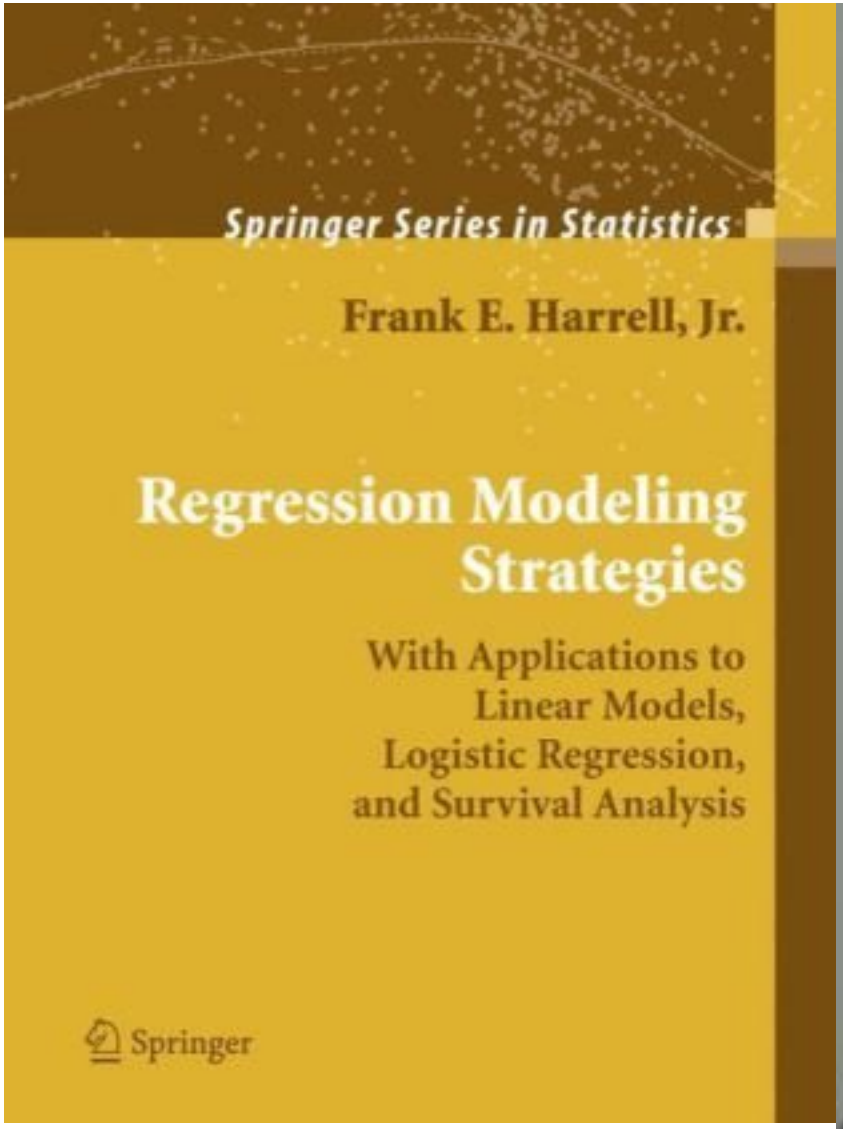


Why stepwise isn't so wise – Daniel Ezra Johnson



Frank Harrell
Chair of Biostatistics, Vanderbilt
Regression Modeling Strategies
first edition 2001, revised 2011
Design package in R

stepwise variable selection is bad!

- one of the most widely used and abused of all data analysis techniques
- very commonly employed for reasons of developing a concise model or of a false belief that it is not legitimate to include "insignificant" regression coefficients when presenting results to the intended audience - Frank Harrell
- a very popular technique for many years, but if it had just been proposed as a statistical method, it would most likely be rejected because it violates every principle of statistical estimation and hypothesis testing - Frank Harrell
- causes severe biases in the resulting multivariable model fits while losing valuable predictive information from deleting marginally significant variables. - Frank Harrell

- personally, I would no more let an automatic routine select my model than I would let some best-fit procedure pack my suitcase. - Ronan Conroy, biostatistician, RCS (Ireland)
- treat all claims based on stepwise algorithms as if they were made by Saddam Hussein on a bad day with a headache having a friendly chat with George Bush.
  - Steve Blinkhorn, psychometrician, author of one of *Nature*'s "magnificent seven" in 2003
- I don't know what knowledge we would lose if all papers using stepwise regression were to vanish from journals at the same time as programs providing their use were to become terminally virus-laden. - Ira Bernstein, professor of Clinical Sciences, UTSW

what is the big deal?

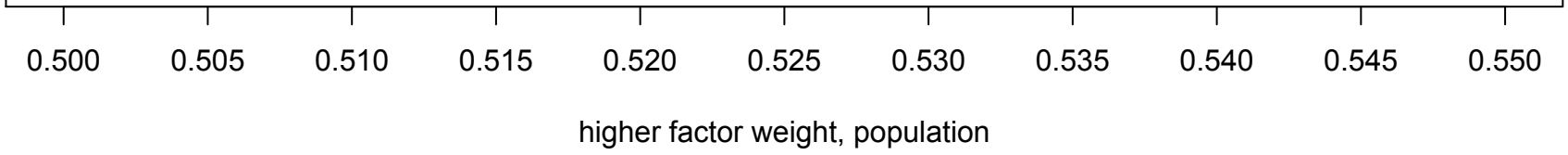
- R-squared values are biased* too high (compared to the population)
- test statistics do not have the correct distribution (F, chi-squared):
  - p-values are biased too small
  - standard errors (if reported) are biased too low
  - confidence intervals (if reported) are biased too wide
- multiple comparisons (not only a problem with stepwise; more widespread)
  - Bonferroni correction (conservative): α/n
  - Sidak correction (if predictors are independent): $1 - (1-\alpha)^{1/n}$
- regression coefficients are biased too high, even with a single predictor:
  - a predictor is more likely to be included if coefficient is overestimated
  - a predictor is less likely to be included if coefficient is underestimated
  - mainly relevant for marginally-significant predictors
- when there is multicollinearity, variable selection becomes arbitrary
- removing "insignificant" variables sets their coefficient(s) to zero, which may be implausible (or it may not)
- allows us not to think about the problem:
  - of multicollinearity
  - of forming and testing hypotheses more generally

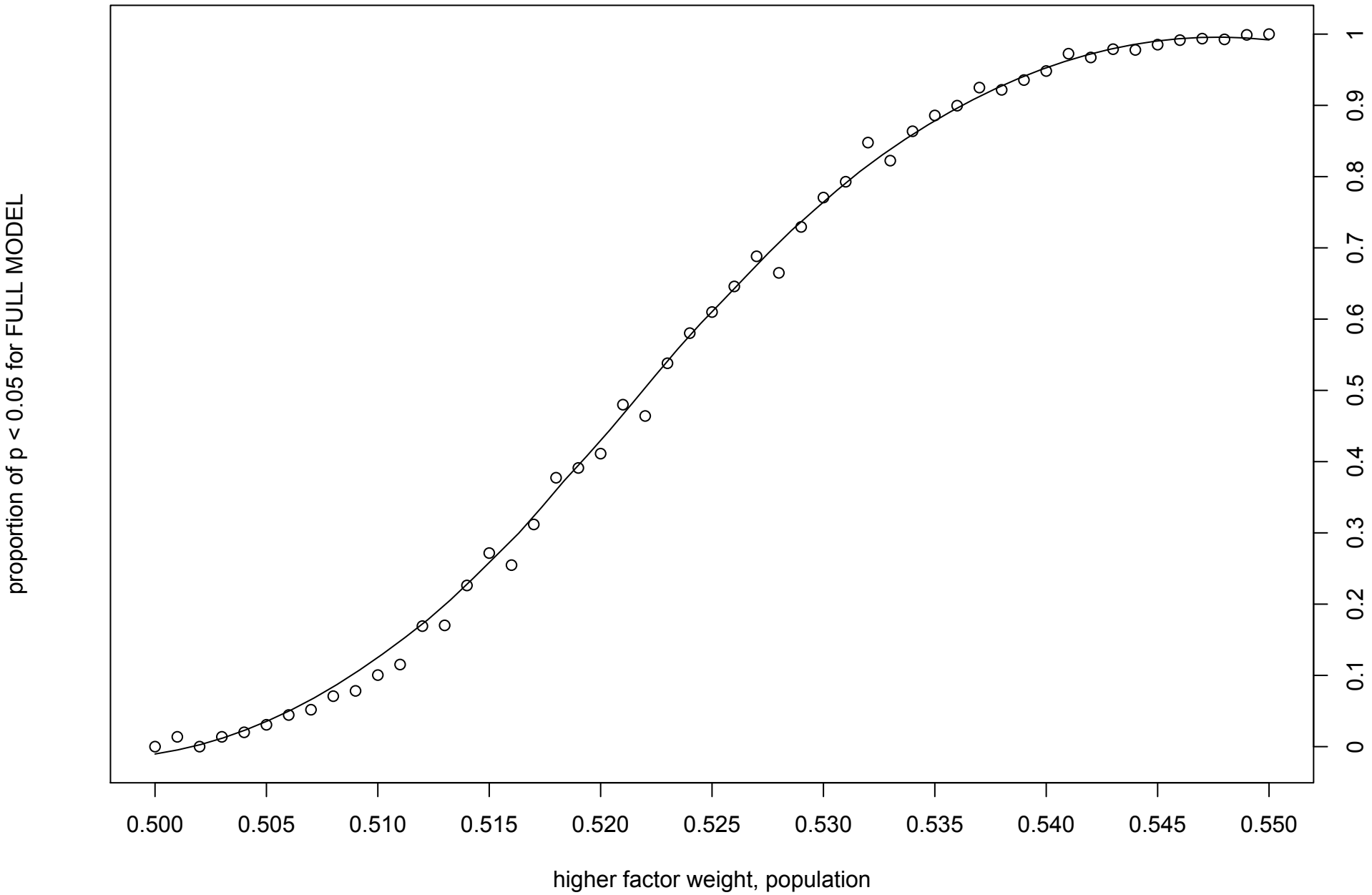
multiple comparisons

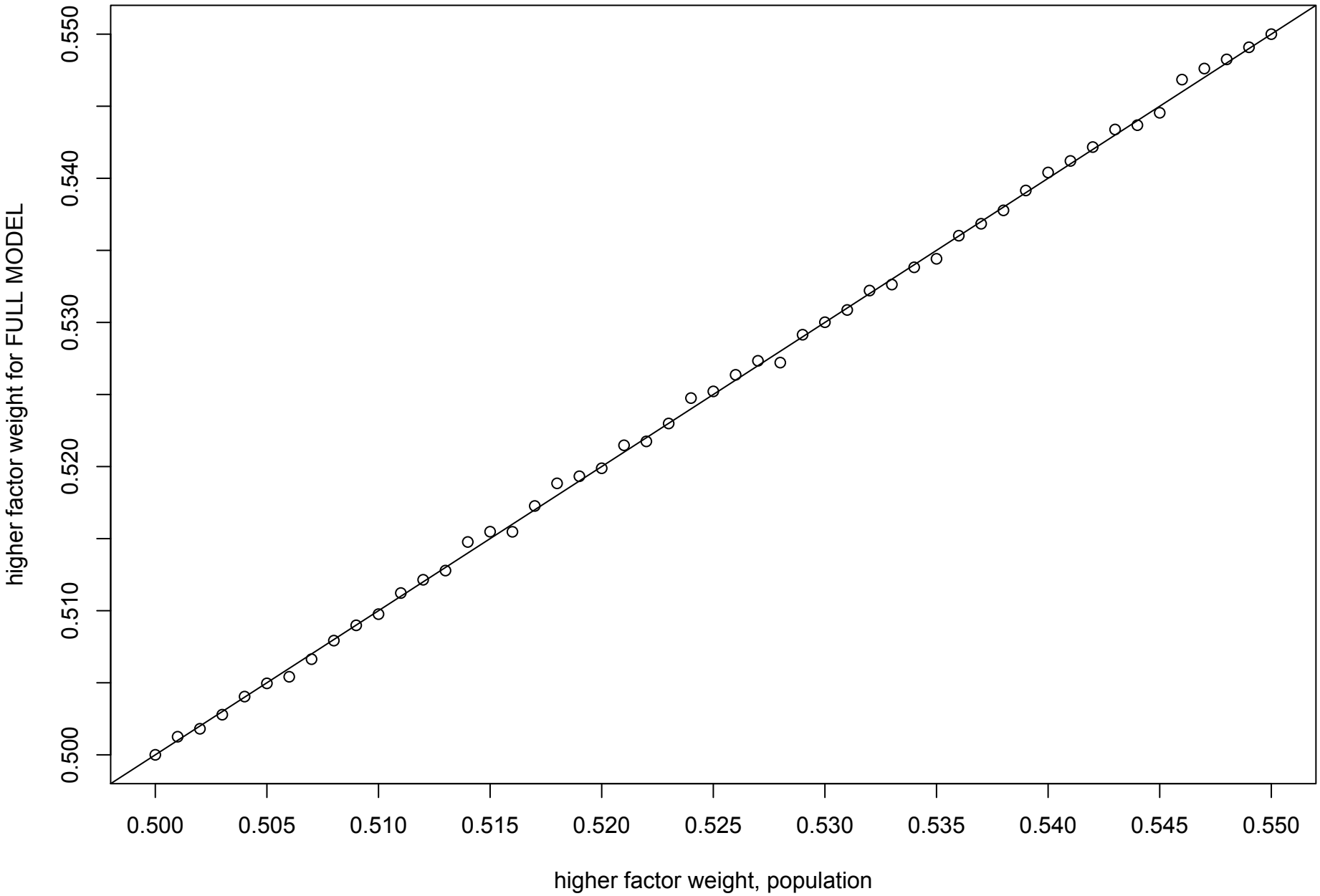
- say we test for age, gender, race, class
- using $p < .05$ for each
- assuming predictors are independent and have no real effect, chance of finding one or more "significant" predictors is .185
- using Sidak correction, $p < .013$
  the chance of finding one or more spuriously "significant" predictors is .05
- if fishing, don't hide it in reporting results
- accurate inference is based on all candidates

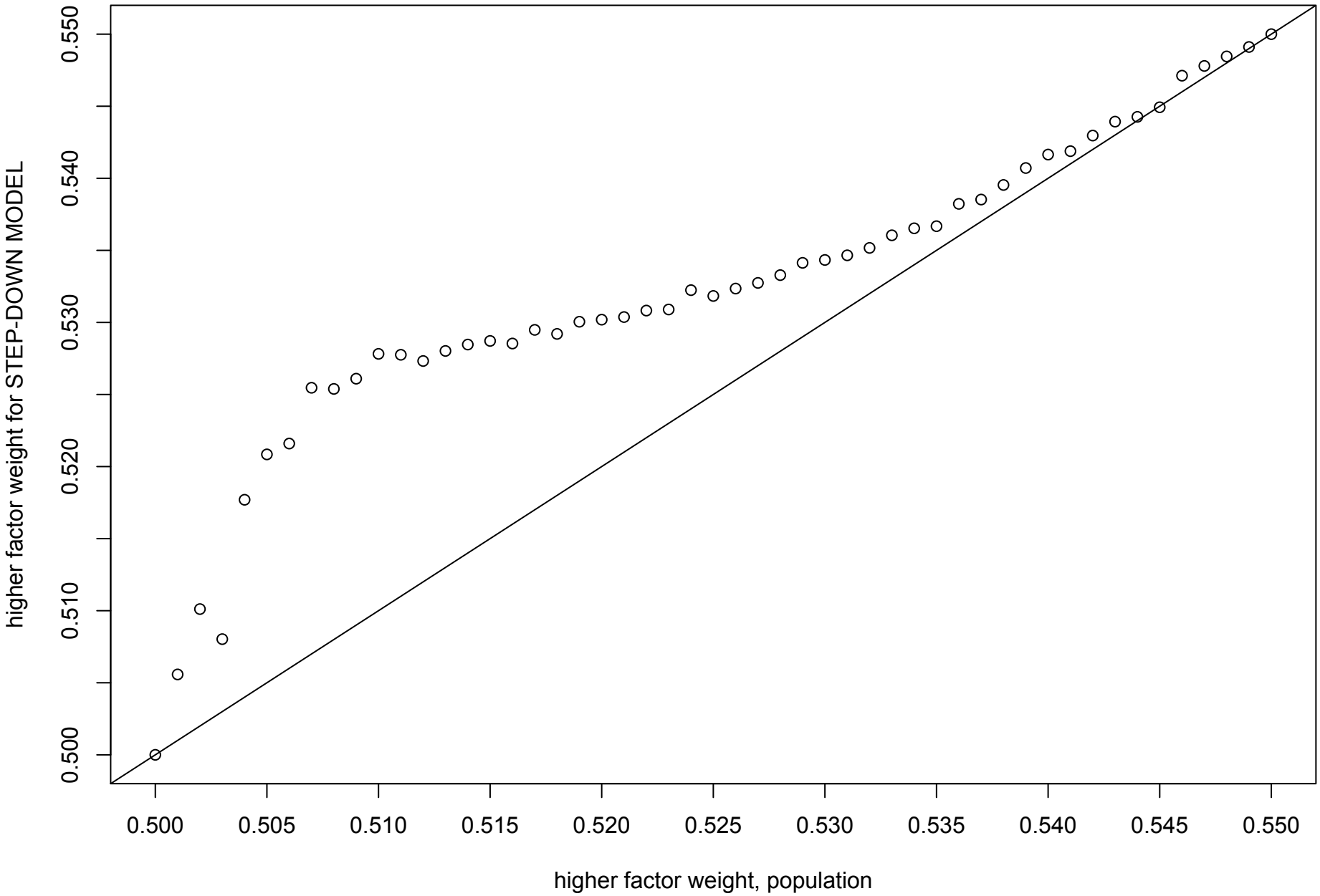
simulation* illustrating bias

imaginary binary response
one predictor (between-speaker)
20 speakers (10 ‘male’, 10 ‘female’)
no individual-speaker variation
100 tokens per speaker
range of true effect sizes (population)
1000 runs per effect size



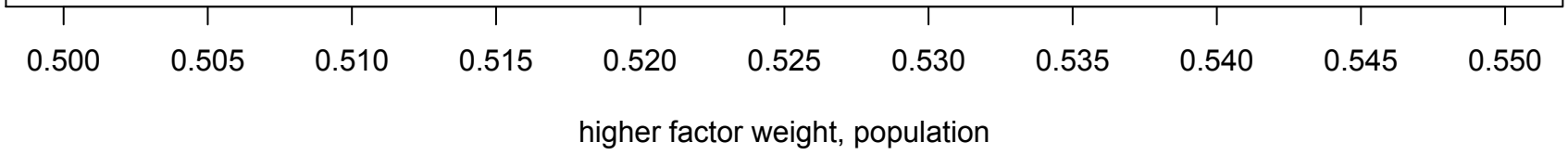

proportion of p < 0.05 for FULL MODEL
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.500
0.505
0.510
0.515
0.520
0.525
0.530
0.535
0.540
0.545
0.550
higher factor weight, population

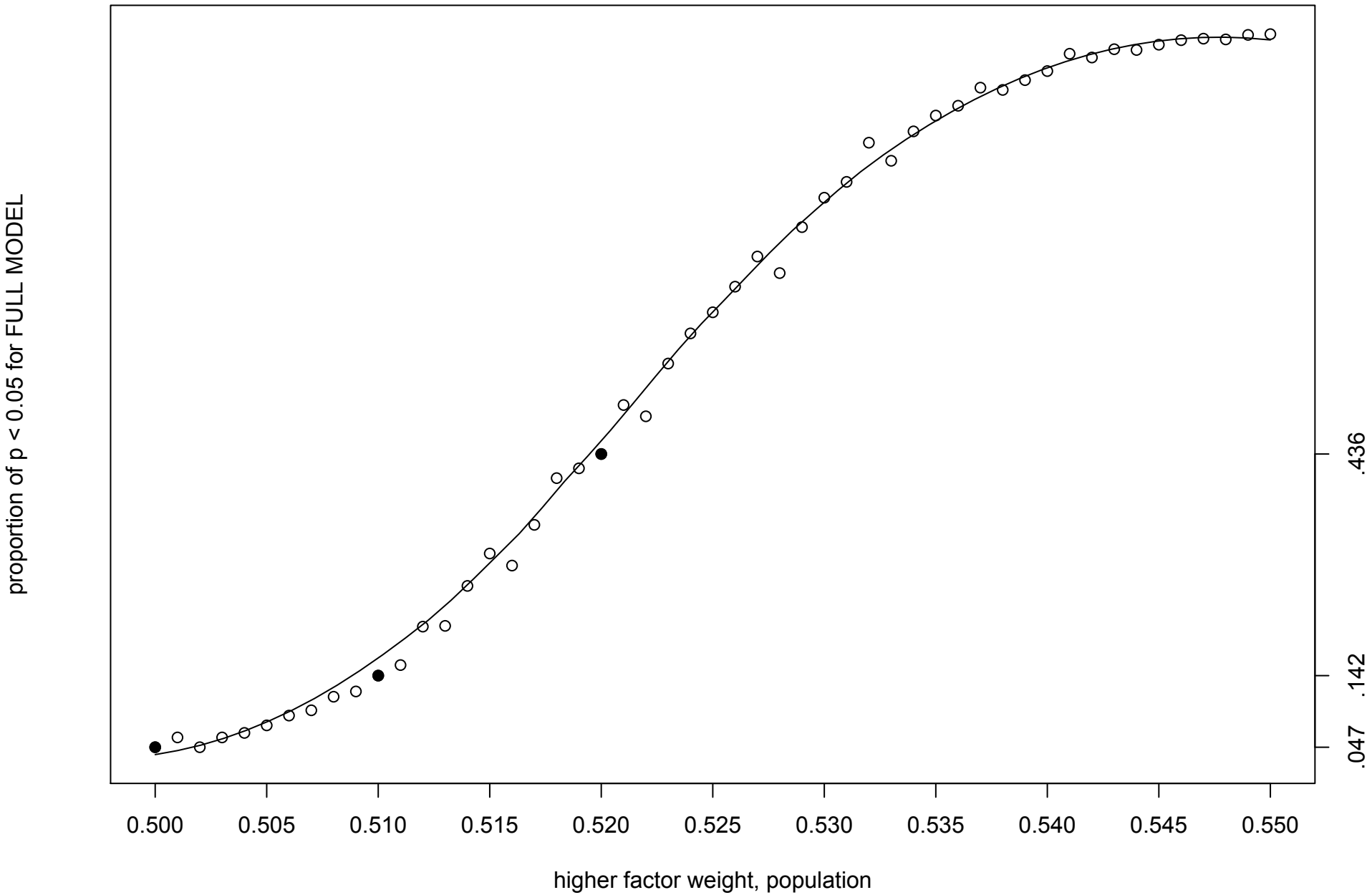

higher factor weight for FULL MODEL
0.500
0.510
0.520
0.530
0.540
0.550
0.500
0.505
0.510
0.515
0.520
0.525
0.530
0.535
0.540
0.545
0.550
higher factor weight, population

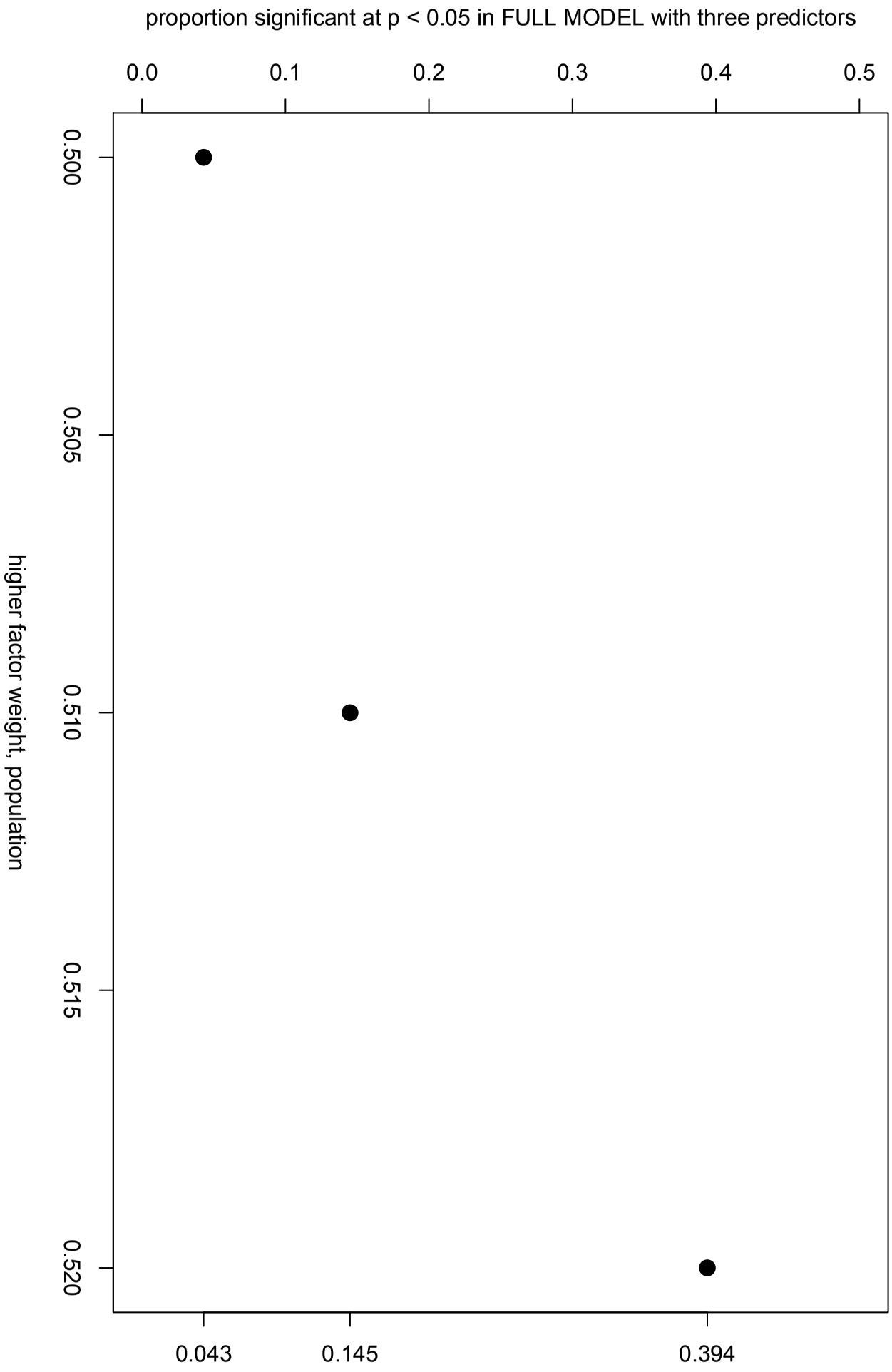

higher factor weight for STEP-DOWN MODEL
0.500
0.510
0.520
0.530
0.540
0.550
0.500
0.505
0.510
0.515
0.520
0.525
0.530
0.535
0.540
0.545
0.550
higher factor weight, population

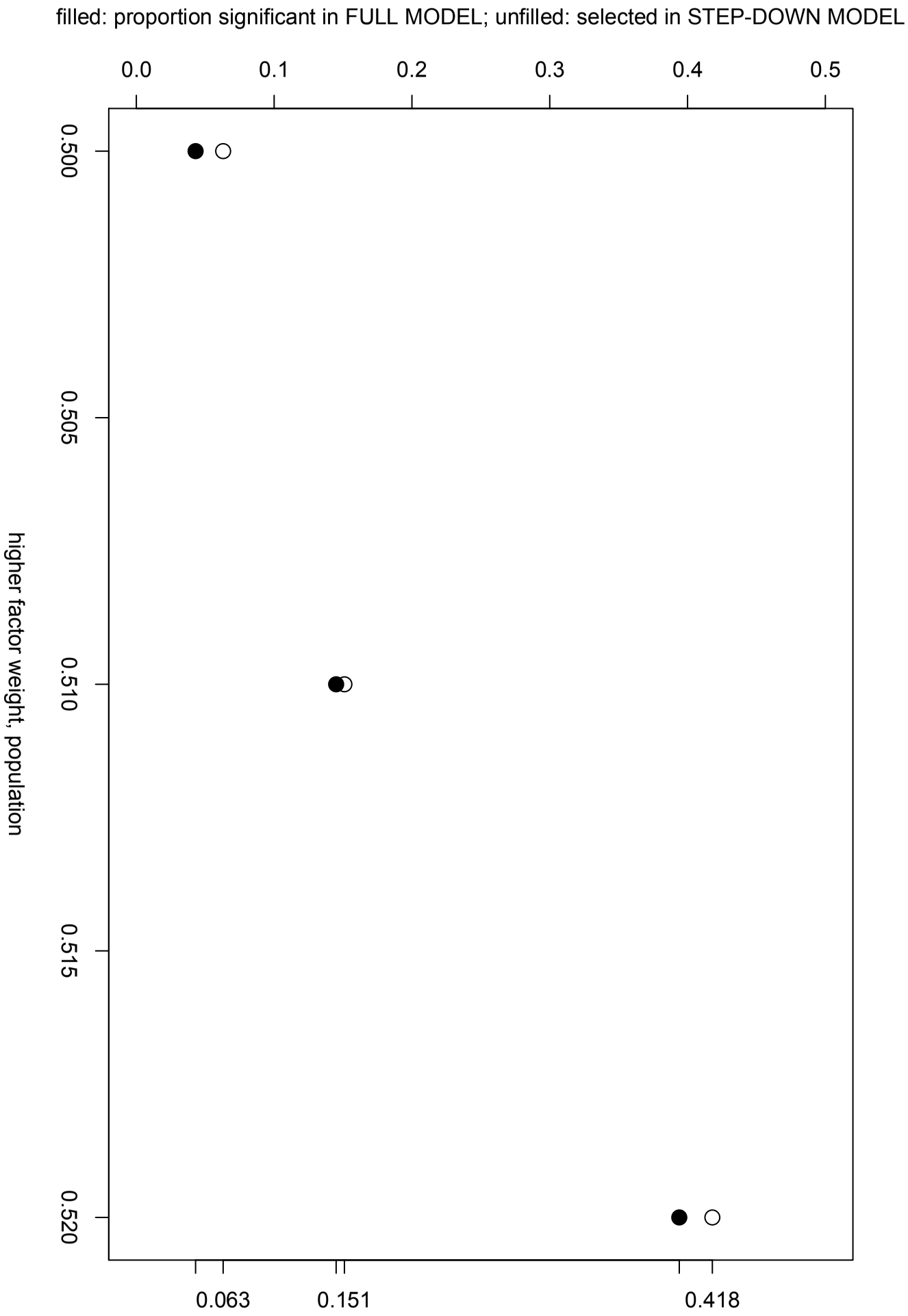
simulation illustrating selection problems

imaginary binary response
three predictors (between-speaker)
20 speakers (10 ‘male’, 10 ‘female’)
no individual-speaker variation
100 tokens per speaker
true effect sizes: .500, .510, .520
1000 runs per effect size




proportion of p < 0.05 for FULL MODEL
.047
.142
.436
0.500
0.505
0.510
0.515
0.520
0.525
0.530
0.535
0.540
0.545
0.550
higher factor weight, population


proportion significant at p < 0.05 in FULL MODEL with three predictors
0.0
0.1
0.2
0.3
0.4
0.5
higher factor weight, population
0.500
0.505
0.510
0.515
0.520
0.043
0.145
0.394


filled: proportion significant in FULL MODEL; unfilled: selected in STEP-DOWN MODEL
0.0
0.1
0.2
0.3
0.4
0.5
0.500
0.505
0.510
0.515
0.520
higher factor weight, population
0.063
0.151
0.418

general suggestions (FH)

- full model fit
  - use a pre-specified model without simplification
  - p-values (confidence intervals) more accurate
- "data reduction"
  - combine collinear variables into one variable
- only remove variables with $\alpha > 0.5$
- only remove if sign (+/-) is not sensible
- only remove if a coefficient of zero is plausible
- step-up: forget about it
- "step-down": Lawless & Singhal fastbw()

sociolinguistics suggestions (DEJ)

- often compare set of significant predictors, coefficients across groups, varieties
- don't compare coefficients from diff. models
- "significance" depends on many things
  - number of tokens
  - chance (distribution of speakers, words)
  - chance (plain chance)
- don't force binary distinction, signif. vs. n.s.
- best: compare models to test hypotheses
- publish complete data, allow re-analysis?

- thanks to:
- Kyle Gorman
- Sali Tagliamonte
- workshop participants

- please contact me:
- danielezrajohnson @gmail.com
- these slides are at:
- danielezrajohnson.com/ stepwise.pdf