# HERC CyberSeminar

# Log Odds and Ends: Marginal Effects in Logit Models

Edward C. Norton

University of Michigan and NBER

June 17, 2020

# Odds Ratios: Introduction

- Many ways to express strength of association between risk factors and binary outcome
  - Probability
  - Odds and Odds ratio
  - Risk and Risk ratio and Risk difference
  - Marginal effect
- Goal: provide insight into interpretation of ORs, their limitations, alternatives

# Odds

- **Odds** are ratio of probability that outcome occurs to probability that it does not occur
- Odds = $p/(1-p)$

- Examples
  - $\Pr(death) = \ 0.2$
  - $Odds(death) \ = \ 0.2/0.8 \ = \ 0.25$

  - $\Pr(survival) \ = 0.8$
  - $Odds(survival) = 0.8/0.2 \ = \ 4.0$

# Odds = $p/(1-p)$

- If *p* is small then odds $\approx$ *p*
- Odds and probabilities diverge as *p* grows
- “Even odds” (odds = 1) are when $p = 0.5$
- Odds are never negative
- No upper bound (as p $\rightarrow$ 1 then odds $\rightarrow \infty$)

# Gamblers

- If randomly pick one card from a deck of 52 cards, then probability of selecting spade ♠ is $p = 0.25 = 13/52$
- Odds of selecting spade are 13/39 = 1:3
- If bet $1, then need payoff of $3 (if spade) to break even

# Poll Question

- Which best describes your comfort with odds ratios and logistic regression?

1. I teach quantitative methods, very familiar
2. I write papers that use logistic regression
3. I read papers that use logistic regression
4. What is logistic regression?

# Logistic Regression

- Parameter $\beta$ is neither probability nor OR
- Parameter $\beta$ is log odds
- Odds = exp($\beta$)
- If $\beta = 0.4$ then OR = 1.5
- If $\beta = 0.0$ then OR = 1
- If $\beta = –0.2$ then OR = 0.82

# Odds Ratio (OR)

- OR for a risk factor has interpretation of whether someone with that risk factor is more or less likely than someone without that risk factor to have the outcome

# Tringale et al. (*JAMA* 2017)

- Studied industry payments to physicians
- Men:  50.8% received industry payment
- Women: 42.6% received industry payment
- Odds for men = 1.03 = .508/.492
- Odds for women = 0.74 = .426/.574
- OR (men to women) is 1.39 = 1.03/0.74
- Odds that men receive industry payment is about 40% higher than for women

# Tringale et al. (2017)

- Greater odds for men could be due to differences in specialty and other factors
- After controlling for other factors, OR reduced to 1.28 (CI: 1.26, 1.31)

# First Main Problem with OR

- Confusion of odds with probabilities
  - Odds are not probabilities
  - Odds ratios are not risk ratios
    - ~~Men are 40% more likely to get industry payments~~
- When probabilities are not close to zero, these differ a lot, can be confusing
- Well-known problem in literature

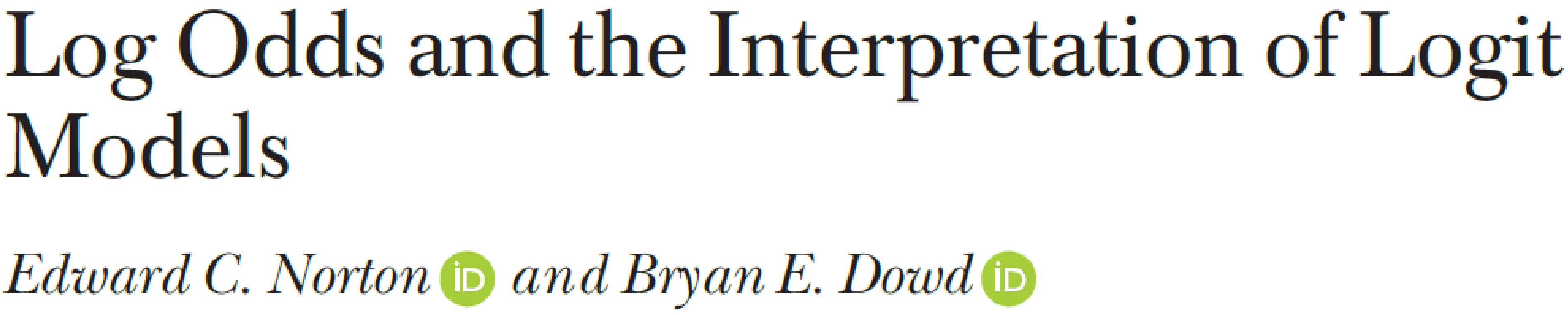
# Odd Thing About Odds

- Suppose OR = 2 for men compared to women
- Case 1: 1% for men, 0.5% for women
- Case 2: 50% for men, 33% for women
- Case 3: 80% for men, 67% for women

- *Important to know underlying probability*

# Second Main Problem with OR

- OR are scaled by an arbitrary factor
- ***Scaling factor equals square root of variance of the unexplained part of binary outcome***



Log Odds and the Interpretation of Logit Models

*Edward C. Norton and Bryan E. Dowd*

# Other Main Problem

- Scaling factor changes when variables added to logistic regression (they explain some of variance, so less is unexplained)
- Adding variables increases the odds ratio (if independent of variable of interest)

# Consequences of Arbitrary Scaling

1. There is no unique odds ratio (OR)
2. Cannot compare OR from same study if using different models (different variables)
3. Cannot compare OR from one study to OR from another study
4. Cannot use standard robustness checks to see if estimated coefficient is stable
5. Problem: arbitrary scaling factor $\sigma$

# Return to OLS Regression

- Consider an OLS regression
- Dependent variable $y^*$ is continuous
- $y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$
- Variance of $\varepsilon$ is called $\sigma^2$ (sigma squared)
- It is the variance of the unexplained part
- If add more variables, $\sigma^2$ decreases
- If add more variables, β **unchanged** (if those variables are independent of other **x**)

# From OLS to Logit

- The latent (unobserved, underlying) dependent variable $y^*$ is continuous
- Observe binary outcome $y$, value depends on whether $y$ exceeds threshold $T$

$$y_i^* = \mathbf{x}_i'\beta + \varepsilon_i$$

$$y_i = \begin{cases} 0 \text{ if } y_i^* \leq T \\ 1 \text{ if } y_i^* > T \end{cases}$$

# Latent Variable

- Write probability in terms of error term taking on range of values

$$\begin{aligned}\Pr(y_i^* > T) &= \Pr(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i > T) \\ &= \Pr(\varepsilon_i > T - \mathbf{x}_i'\boldsymbol{\beta}) \\ &= \Pr(\varepsilon_i < \mathbf{x}_i'\boldsymbol{\beta} - T)\end{aligned}$$

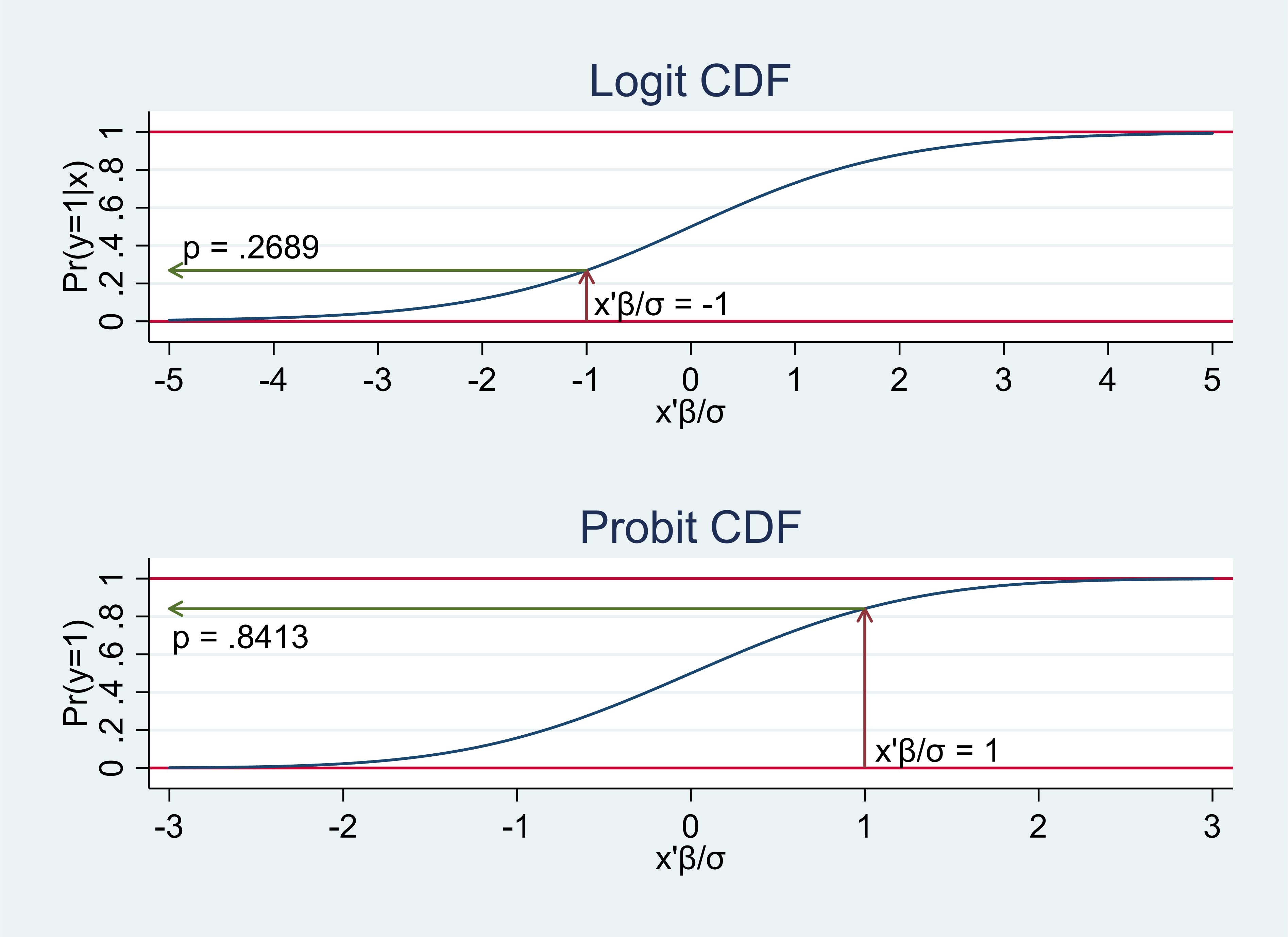
- Standardize the error by dividing by the standard deviation σ because can only make precise statements about **standardized distributions** (mean 0, variance 1)

- *Hmm, that sounds innocuous … what could go wrong?*

# Logit Normalization

- Drop *T* from equation (absorbed in constant)
- Typical assumption is that σ $= 1$
- Let’s not assume that, σ does not disappear

$$\begin{aligned}\Pr\left(y_i^* > T|\text{logistic}\right) &= \Pr\left(\frac{\varepsilon_i}{\sigma} < \frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) \\ &= \frac{1}{1+\exp\left(\frac{-(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)}\end{aligned}$$



Logit CDF

Probit CDF

# Odds

- Odds = $p/(1-p)$
- Write in terms of β and σ
- Notice that σ does not disappear
- Typo in paper (no minus sign)

$$\text{odds} = \frac{p_i}{1-p_i} = \frac{\dfrac{1}{1+\exp(-\boldsymbol{x}_i'\frac{\beta}{\sigma})}}{\dfrac{\exp(-\boldsymbol{x}_i'\dfrac{\beta}{\sigma})}{1+\exp\left(-\boldsymbol{x}_i'\dfrac{\beta}{\sigma}\right)}} = \exp\left(\quad \boldsymbol{x}_i'\frac{\beta}{\sigma}\right)$$

# Odds Ratio

- Useful interpretation for dummy variable

$$\text{odds for smoker} = \exp\left(\frac{\beta_0 + \beta_{\text{smoke}}\text{smoke}_{1i} + \beta_2 x_{2i} + \ldots \beta_K x_{Ki}}{\sigma}\right)$$

$$\text{odds for nonsmoker} = \exp\left(\frac{\beta_0 + \beta_2 x_{2i} + \ldots \beta_K x_{Ki}}{\sigma}\right)$$

$$\text{OR} = \text{odds ratio} = \frac{\text{odds for smoker}}{\text{odds for nonsmoker}} = \exp\left(\frac{\beta_{\text{smoke}}}{\sigma}\right)$$

$$\text{Log odds} = \left(\frac{\beta_{\text{smoke}}}{\sigma}\right)$$

# Sigma σ

- Logit models estimate β/σ
- σ = standard deviation of the error term
- **σ is unknown**

# What Changes Sigma?

- $\sigma$ is measure of unexplained variation
- If add variables to the model, then $\sigma \downarrow$
- But $\beta/\sigma$ changes by unknown amount

# Implication 1 of $\frac{\beta}{\sigma}$

- There is no single odds ratio
- An OR is not an absolute number (e.g., $\pi$)
- An OR is **conditional on sample and model**
- A study that aims or claims to estimate *the* OR is misguided
- An OR shows sign and magnitude, as does $\frac{\beta}{\sigma}$
- See Norton and Dowd (2018)

# Implication 2 of $\frac{\beta}{\sigma}$

- OR estimated from different data sets are not directly comparable
- OR estimated with different model specifications (same data set) are not directly comparable
- Different models have different $\sigma$
- (See Allison 1999; Mood 2010)

# Implication 3 of $\frac{\beta}{\sigma}$

- A statement like “*The* OR is 1.5.” is factually incorrect
- A correct, precise interpretation might be “The estimated OR is 1.5, conditional on demographics and health, but a different OR would be found if the model included a different set of explanatory variables. This estimated OR may not be used to compare OR from other data sets with the same explanatory variables, or even OR estimated from this same data set with a different model specification.”

# Implication 4 of $\frac{\beta}{\sigma}$

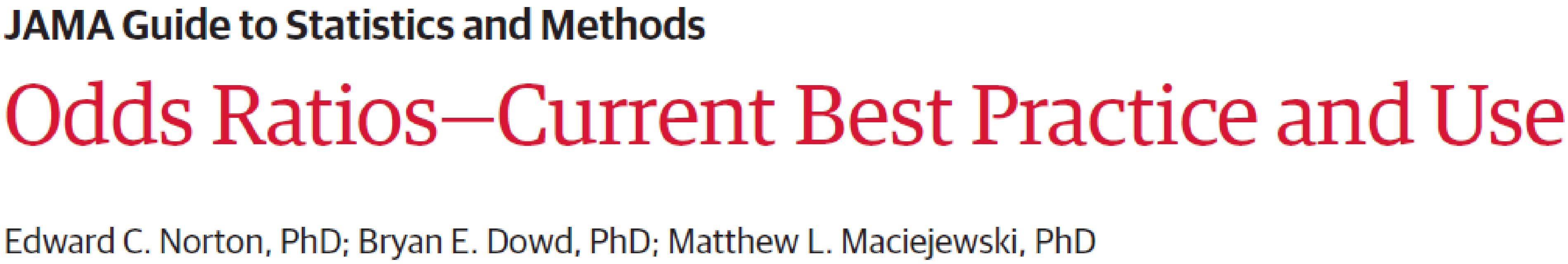
- Some authors progressively add more variables to see if results are robust
- Cannot compare the OR from these models directly
- Expect $\beta/\sigma$ to be different when you add variables

# Implication 5 of $\frac{\beta}{\sigma}$

- This understanding of importance of sigma enhances already strong criticisms of OR
- Most prior critical papers have focused on differences between OR and RR (risk ratios)

# Summary

- These five implications are not widely appreciated in the literature
- Papers frequently report findings of the odds ratio, as if it were an absolute number that could be estimated without explicit conditioning on the model and covariates



JAMA Guide to Statistics and Methods

Odds Ratios—Current Best Practice and Use

Edward C. Norton, PhD; Bryan E. Dowd, PhD; Matthew L. Maciejewski, PhD

# Precise Meaning

- False to say that ORs have no meaning
- In contrast, they have an extremely precise meaning
  - ***OR applies to that data set and that model specification only, but no other***
- The magnitude is not generalizable

# Intuition for Why OR changes

- What is Pr(tested COVID-19 positive)?

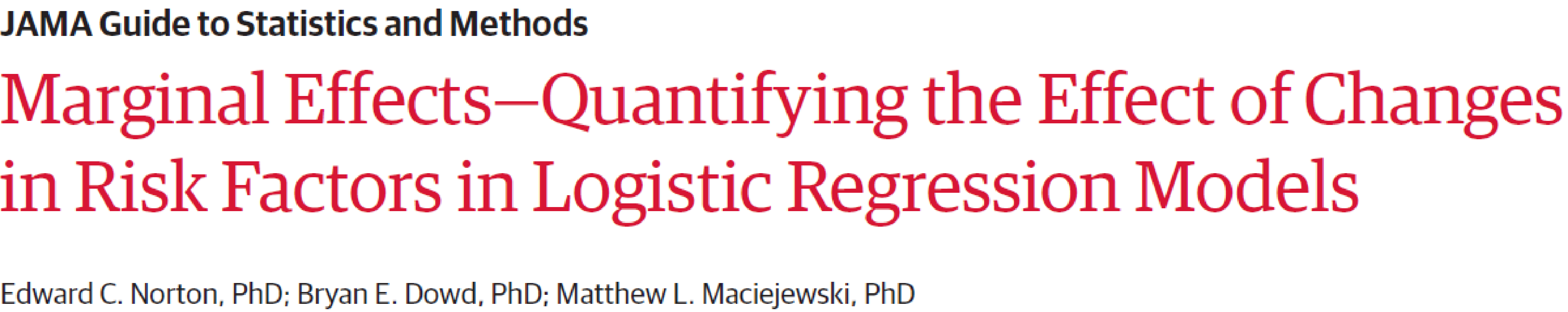
- Start with entire population of world
- What happens as add controls?
- Add controls removes explanations, makes study design more homogeneous
- In extreme, left with identical twins

# ORs are Sometimes Appropriate

- Appropriate for case-control studies
- Chamberlain FE logit model

# Marginal Effects (ME)

- Marginal effects are like risk differences
- Interpreted as **percentage point** changes
- Average ME not sensitive to changes in $\sigma$
- Therefore, use ME whenever possible



JAMA Guide to Statistics and Methods

Marginal Effects—Quantifying the Effect of Changes in Risk Factors in Logistic Regression Models

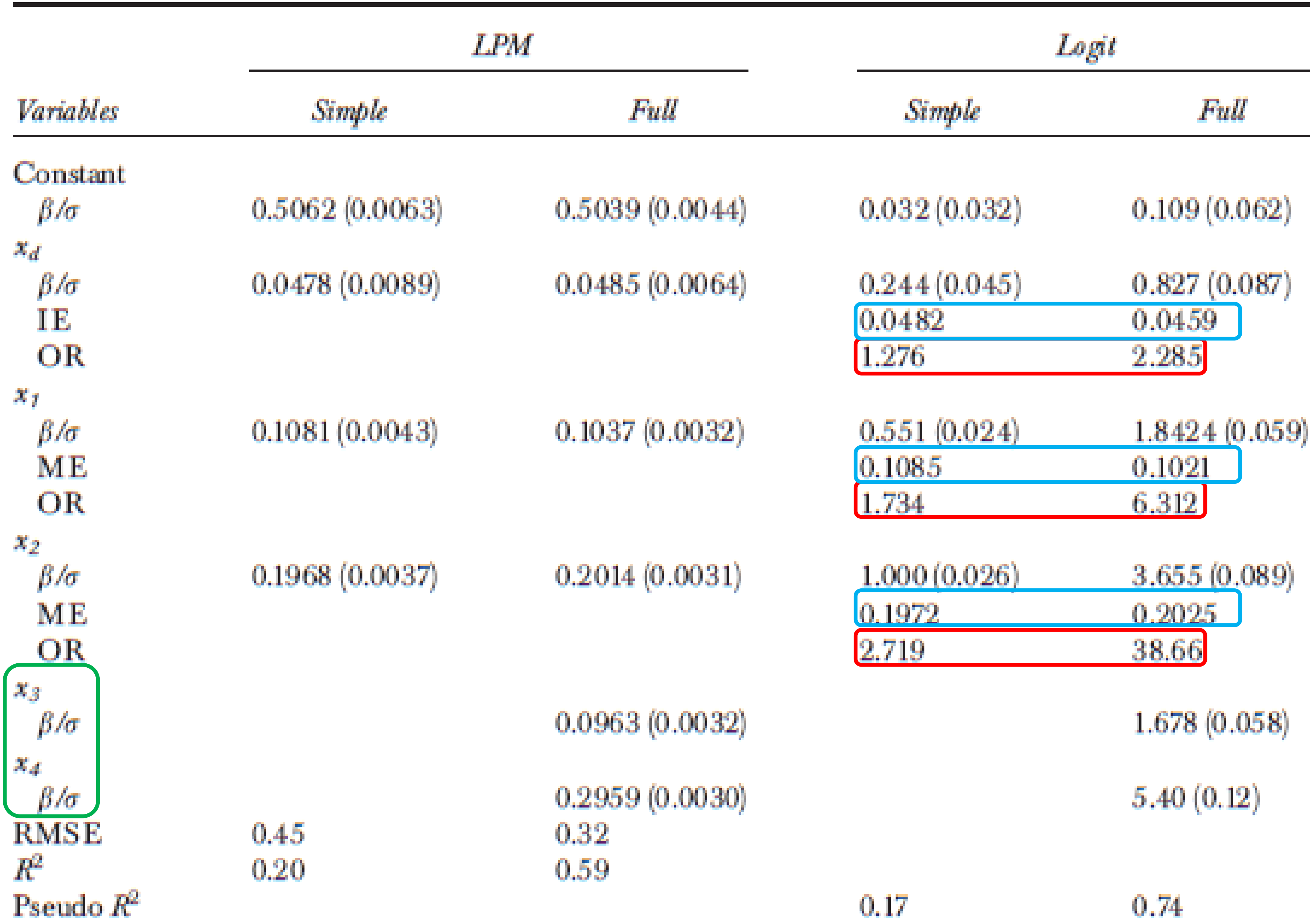
Edward C. Norton, PhD; Bryan E. Dowd, PhD; Matthew L. Maciejewski, PhD

# Marginal Effects (ME)

- Simulation shows what happens to OR and ME when add additional variables
- Marginal effects stay same, ORs increase

- Norton and Dowd (2018) Table 1

Table 1: Comparison of Coefficient Estimates, Marginal Effects, and Odds Ratio and Probit Models for Two Different Model Specifications



| | *LPM* | | *Logit* | |
|---|---|---|---|---|
| *Variables* | *Simple* | *Full* | *Simple* | *Full* |
| Constant | | | | |
| $\beta/\sigma$ | 0.5062 (0.0063) | 0.5039 (0.0044) | 0.032 (0.032) | 0.109 (0.062) |
| $x_d$ | | | | |
| $\beta/\sigma$ | 0.0478 (0.0089) | 0.0485 (0.0064) | 0.244 (0.045) | 0.827 (0.087) |
| IE | | | 0.0482 | 0.0459 |
| OR | | | 1.276 | 2.285 |
| $x_1$ | | | | |
| $\beta/\sigma$ | 0.1081 (0.0043) | 0.1037 (0.0032) | 0.551 (0.024) | 1.8424 (0.059) |
| ME | | | 0.1085 | 0.1021 |
| OR | | | 1.734 | 6.312 |
| $x_2$ | | | | |
| $\beta/\sigma$ | 0.1968 (0.0037) | 0.2014 (0.0031) | 1.000 (0.026) | 3.655 (0.089) |
| ME | | | 0.1972 | 0.2025 |
| OR | | | 2.719 | 38.66 |
| $x_3$ | | | | |
| $\beta/\sigma$ | | 0.0963 (0.0032) | | 1.678 (0.058) |
| $x_4$ | | | | |
| $\beta/\sigma$ | | 0.2959 (0.0030) | | 5.40 (0.12) |
| RMSE | 0.45 | 0.32 | | |
| $R^2$ | 0.20 | 0.59 | | |
| Pseudo $R^2$ | | | 0.17 | 0.74 |

# Marginal Effect Party Trick

- $\text{Logit ME} \; = \; \beta p(1-p)$

- Simple formula for overall marginal effect
- Example:  mean outcome is 0.1 (10%)
  - Then $p(1-p)$ is 0.09, or about 10%
  - Suppose $\beta = .2$, the ME is about 2 percentage pts.

# Conclusions

- Odds ratios often reported without proper discussion of conditioning, arbitrary scaling
- Odds ratios are conditional on data and model specification
- Cannot compare odds ratios
- Consider estimating marginal effects, which are usually not that sensitive to $\sigma$

# References

- Norton, EC, BE Dowd. 2018. “Log odds and the interpretation of logit models.” ***Health Services Research*** 53(2):859–878.
- Norton, EC, BE Dowd, ML Maciejewski. 2018. “Odds Ratios—Current Best Practice and Use.” ***JAMA*** 320(1):84–85.
- Norton, EC, BE Dowd, ML Maciejewski. 2019. “Marginal effects—Quantifying the effect of changes in risk factors in logistic regression models.” ***JAMA*** 321(13):1304–1305.

# Thank You!

- Contact information
- ecnorton@umich.edu
- Prof. Edward C. Norton
- University of Michigan

PUBLIC
HEALTH