Ottawa Health Research Institute
OHRI IRSO
Institut de recherche en santé d'Ottawa

Bias in logistic regression due to omitted covariates

Dean **Fergusson,** Medicine, University of Ottawa
Tim **Ramsay**, Epidemiology, University of Ottawa
George Alex **Whitmore**, Management, McGill University

This is not a new discovery

Gail *et al* (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika, 71(3):432—444

- "Important nonlinear regression models lead to biased estimates….if needed covariates are omitted"
  - linear or exponential regression unbiased
  - bias always towards the null
  - for proportional hazards, bias depends on amount of censoring
  - unrelated to imbalance or confounding

Lagakos & Schoenfeld (1984). *Properties of proportional-hazards score tests under misspecified regression models*. Biometrics; 40:1037—1048.

Begg & Lagakos (1990). *On the consequences of model misspecification in logistic regression*. Env Health Persp; 87:69—75.

Robinson & Jewell (1991). *Some surprising results about covariate adjustment in logistic regression models*. Int Stat Rev; 58(2):227—240.

Hauck *et al* (1991). *A consequence of omitted covariates when estimating odds ratios*. J Clin Epidemiol; 44(1):77—81.

Hauck *et al* (1998). *Should we adjust for covariates in nonlinear regression analyses of randomized trials?* Controlled Clin Trials; 19:249—256.

Johnston *et al* (2004). *Risk adjustment effect on stroke clinical trials*. Stroke;35:e43—e45.

Steyerberg & Eijkemans (2004). *Heterogeneity bias: the difference between adjusted and unadjusted effects*. Med Decis Making;24:102—104.

Martens *et al* (2008). *Systematic differences in treatment effect estimates between propensity score methods and logistic regression*. Int J Epidemiol;37:1142—1147.

Kent *et al* (2009). *Are unadjusted analyses of clinical trials inappropriately biased towards the null?* Stroke;40:672—673.

How bad can it be?

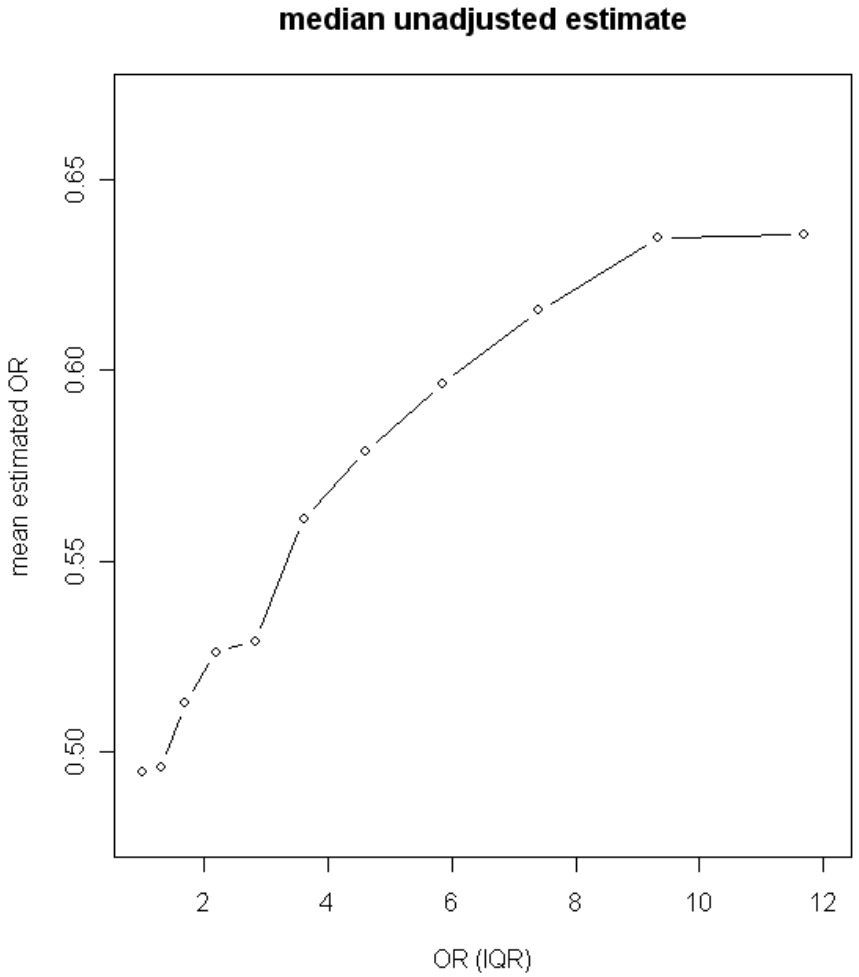
Let's try a little simulation:

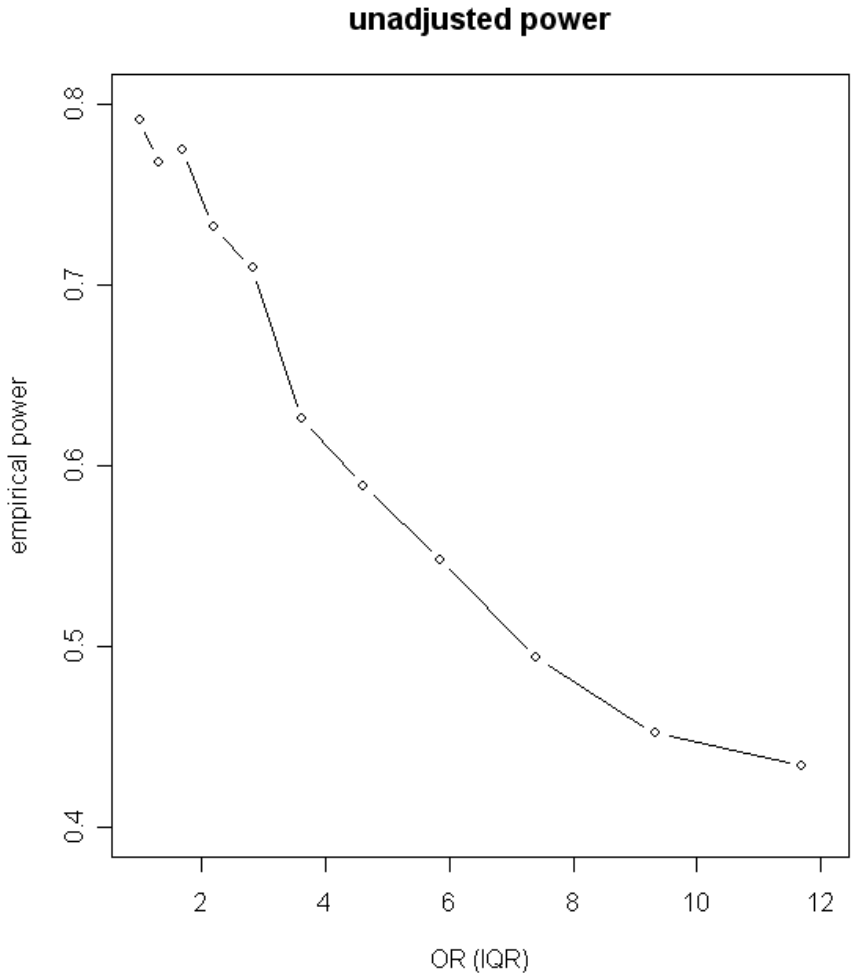
- dichotomous outcome
- dichotomous treatment: OR=0.5
- covariate (age) ~ N(40,10)
  - independent of treatment
  - balanced between groups

- n= 133 per group (80% power)
- age effect defined in terms of OR associated with IQR
  - range from OR=1 to OR=12

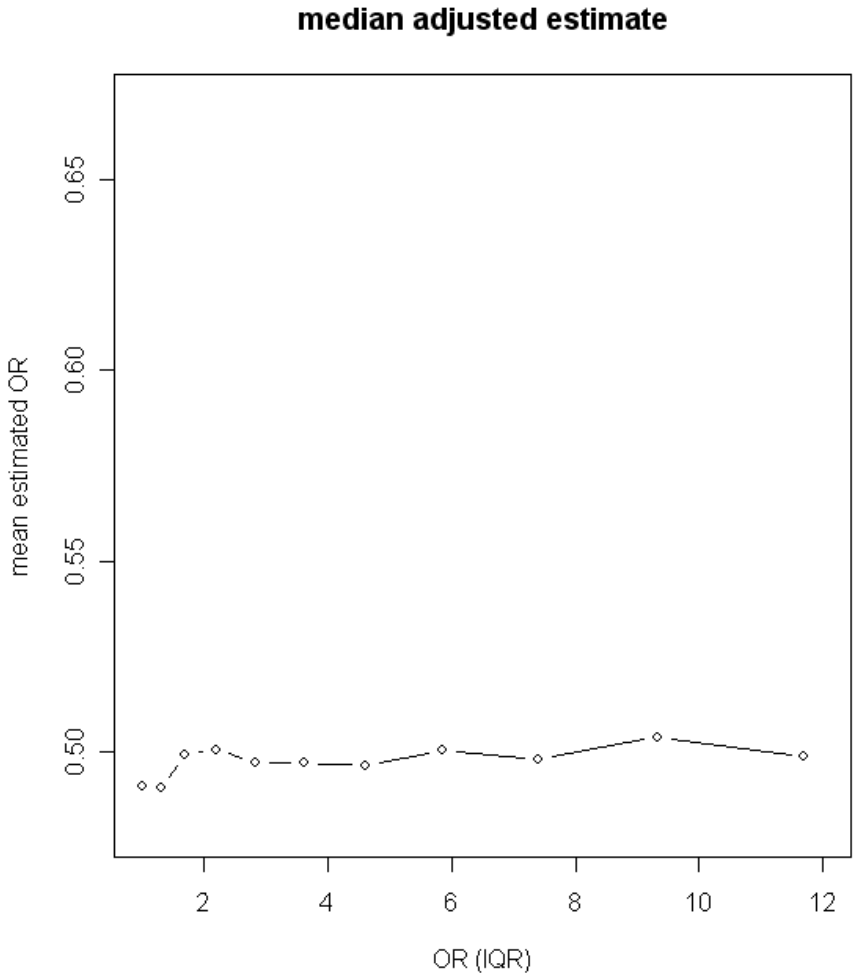
- Simulate 1000 trials per test age effect

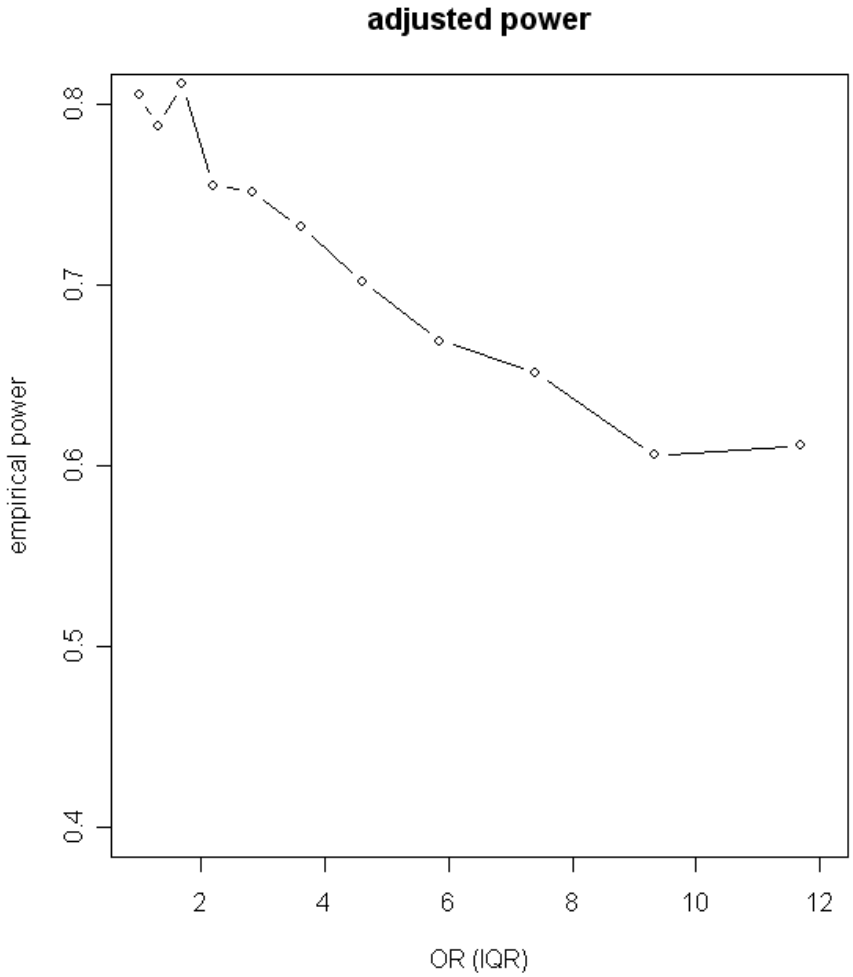
Unadjusted Analysis





Adjusted Analysis





WTF?

- Linear regression:
  - omitting balanced, independent covariates doesn't bias effect estimates
  - including important covariates increases precision of effect estimate

- Logistic regression:
  - omitting balanced, independent covariates does bias effect estimates (towards the null)
  - including these covariates decreases precision of effect estimate

Is this really bias?

marginal treatment effect

- (population-averaged effect)
- what effect will this treatment have on prevalence?

conditional treatment effect

- (individual effect)
- what effect will this treatment have on me?

Exact bias expression

RCT:

- 2 arms, *j=0,1*
- indicator variable $I_{ij}=0,1$ (i^{th} individual, j^{th} treatment)
- let $\boldsymbol{z}_{ij}$ be a vector of covariates
- Assume $\boldsymbol{Z}$ perfectly balanced between arms ($\boldsymbol{z}_{i0}=\boldsymbol{z}_{i1}$)
- Let c_i denote the total number of events in the i^{th} arm

$$p_{ij} = \frac{\exp\ \alpha_0 + \alpha_1 I_{ij} + z_{ij}\beta}{1+\exp\ \alpha_0 + \alpha_1 I_{ij} + z_{ij}\beta} = \frac{A_0 A_1^{\ I_{ij}} Z_{ij}}{1+ A_0 A_1^{\ I_{ij}} Z_{ij}}$$

Differentiating the log-likelihood with respect to A_0 and A_1, we can derive the maximum likelihood estimators for these quantities as the solutions to these two equations

$$c_0 - \sum_{i=1}^{n} \frac{Z_{i0}\hat{A}_0}{1+ Z_{i0}\hat{A}_0} = 0$$

$$c_1 - \sum_{i=1}^{n} \frac{Z_{i1}\hat{A}_0\hat{A}_1}{1+ Z_{i0}\hat{A}_0\hat{A}_1} = 0$$

Exact bias expression

$$\hat{A}_1^U = \left(\frac{n-c_0}{c_0}\right)\left(\frac{c_1}{n-c_1}\right)$$

$$\frac{c_0}{n-c_0} = \hat{A}_0^U \frac{\sum_{i=0}^{n} \hat{w}_{i0} Z_i}{\sum_{i=0}^{n} \hat{w}_{i0}} = \hat{A}_0 \bar{Z}_0$$

$$\frac{c_1}{n-c_1} = \hat{A}_0^U \hat{A}_1^U \frac{\sum_{i=0}^{n} \hat{w}_{i1} Z_i}{\sum_{i=0}^{n} \hat{w}_{i1}} = \hat{A}_0 \hat{A}_1 \bar{Z}_1$$

$$\hat{A}_1^U = \hat{A}_1 \left(\frac{\bar{Z}_1}{\bar{Z}_0}\right)$$

$$\hat{w}_{i0} = \frac{1}{1+Z_i \hat{A}_0}$$

$$\hat{w}_{i1} = \frac{1}{1+Z_i \hat{A}_0 \hat{A}_1}$$

- We refer to the weighted averages in the numerator and denominator as *logistic means*, and observe that the bias will always be towards the null.
- We also observe that the bias will be greater when the 'average' effect of the omitted covariates is larger.

So what should we do?

1. Design phase
    - Need to decide which variables to capture
    - Need to think carefully about power and sample size

So what should we do?

Analysis phase

- Need to decide which variables to include in the model
- May be an ideal application for propensity scores

  Martens, EP *et al* (2008). Int J Epid; 37:1142—1147

Above all, we don't want to open the door to p-value shopping

"Hmmm… Which of these covariates can I include to get the result I want???"

Alternative approach

Abandon logistic regression altogether!

Zou G (2004). *A modified Poisson regression approach to prospective studies with binary data*. Am J Epid; 159(7):702—706.

- Directly estimate relative risk (I hate odds ratios)
- Generalized estimating equations
- Robust variance estimator

Not clear that this doesn't suffer from the same problems as logistic regression.

Conclusion

- Heterogeneity bias is the elephant in the room that nobody talks about
- Probably because we don’t know what to do about it
- Logistic regression will **always** underestimate individual treatment effects