Comparing effects across nested logistic regression models

CADC Scholars Meeting

March 12, 2013

Steve Gregorich

Example from the literature of nested model comparisons

J Acquir Immune Defic Syndr • Volume 57, Number 4, August 1, 2011 *ZDV-Sparing HAART in Pregnancy*

TABLE 3. Crude and AORs for Maternal-to-Child Transmission Comparing ZDV-Sparing With ZDV-Containing HAART in Pregnancies

	Univariable Model (Numbers Vary Due to Missing Data)			Multivariable Model (n = 6111)*		
	n (%)	OR (95% CI)	*P*	n (%)	AOR (95% CI)	*P*
ART group			0.64			0.18
ZDV-containing HAART	5227 (0.9)	1		5214 (0.9)	1	
ZDV-sparing HAART	903 (0.8)	0.83 (0.37 to 1.83)		897 (0.8)	1.81 (0.77 to 4.26)	
Study			0.68			0.72
NSHPC	5261 (0.9)	1		5247 (0.9)	1	
ECS	869 (1.0)	1.16 (0.57 to 2.38)		864 (0.9)	1.15 (0.53 to 2.47)	
Mode of delivery						
Elective CS	3515 (0.8)	1	—	3515 (0.8)	1	—
Emergency CS	1095 (1.7)	2.28 (1.26 to 4.12)	<0.01	1095 (1.7)	2.07 (1.13 to 3.76)	0.02
Vaginal	1051 (0.9)	0.78 (0.37 to 1.66)	0.52	1051 (0.9)	0.80 (0.38 to 1.72)	0.58
Duration of HAART (wks)						
≥24	2087 (0.2)	1	—	2078 (0.2)	1	—
12–23	2416 (0.7)	2.95 (1.09 to 8.01)	0.03	2410 (0.7)	3.44 (1.20 to 9.86)	0.02
8–11	1020 (1.0)	4.12 (1.41 to 12.09)	0.01	1017 (1.0)	5.10 (1.65 to 15.77)	0.01
2–7	607 (4.0)	17.14 (6.51 to 45.12)	<0.001	606 (4.0)	20.09 (7.22 to 55.93)	<0.001

*Adjusted for study, mode of delivery, and duration of HAART.
ART, antiretroviral therapy; CS, cesarean section; OR, odds ratio.

Care is required when comparing ORs across nested models

Outline

. Comparing Parameter Estimates Across Nested Linear Models

. Example Data

. Example Application: Linear Modeling Framework

. Example Application: Logistic Modeling Framework

. Reconciliation

. Conclusions

. Resources

Comparing parameter estimates across nested *linear* models

Initial concepts regarding linear regression

$y_i = \text{intercept} + x_i b + e_i$

The total variance of *y* is decomposed into
- . the variance explained by *x*, plus
- . residual (unexplained) variance

$VAR(y_i) = VAR(x_i) \times b^2 + VAR(e_i)$

When you add additional *x* variables to the model,

- . the explained variation increases and
- . the residual variation decreases
- . ALWAYS: total variation = explained + residual variation

Comparing parameter estimates across nested *linear* models

Two models are *nested* if the parameters of one are a subset of the other

Unadjusted model: y_i = intercept + $x_i b_U$ + e_i

Adjusted model: y_i = intercept + $x_i b_A$ + $cov_i b$ + e_i

. The Unadjusted model is nested within the Adjusted model

What effect does adjustment for *cov* have on the modeled effect of *x*?

. Compare b_A to b_U, either formally or just 'eyeball' the difference

. *Unadjusted vs. Adjusted models are just one type of nested model comparison*

Example data (L Karliner)

N=8077 patient discharges from UCSF 14th Floor
January 2007 - January 2010

Main explanatory variable

Binary patient sex indicator (MalePt): 49% male

Additional *x* variable

scale	Mean	Median	Variance	Min	Max
LOS (days)	5.5	3.7	32.6	0.1	39.8
*ln*LOS	1.3	1.3	0.7	-2.8	3.7

. LOS = length of (hospital) stay

Outcome: Total Costs

scale	Mean	Median	Min	Max
Costs ($)	19,073	10,483	1,335	393,304
*ln*Costs	9.37	9.26	7.20	12.88
binaryCosts	0.50	--	0	1

Example Application: *Linear* Model

Modeling the effect of patient sex on *ln*Costs

	Unadjusted Model			Adjusted Model		
	b	e^b	*p*	b	e^b	*p*
MalePt	0.08	1.08	.0016	0.06	1.06	<.0001
*ln*LOS	--	--	--	0.92	1.09†	<.0001

† $e^{b*\ln(1.1)}$: a 10% increase in LOS → about a 9% increase in costs

. In the unadjusted model, compared to female patients
male patients had total costs expected to be about 8% higher

. After conditioning on *ln*LOS, compared to female patients
male patients had total costs expected to be about 6% higher

. *A fraction of the sex effect might be explained by LOS*

Example Application: *Logistic* Model

Modeling the effect of patient sex on binaryCosts

	Unadjusted Model			Adjusted Model		
	b	UOR	*p*	b	AOR	*p*
MalePt	0.14	1.15	.0016	0.30	1.35	<.0001
*ln*LOS	--	--	--	4.54	1.54†	<.0001

† $e^{b*\ln(1.1)}$: a 10% increase in LOS → about 54% increased odds of higher costs

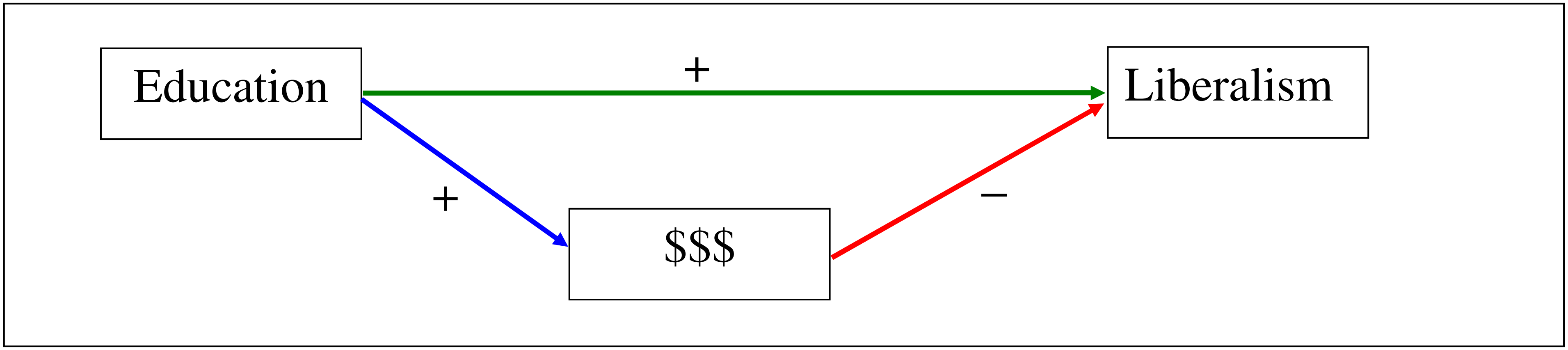
. In the unadjusted model, compared to female patients
male patients had about 15% higher odds of 'high costs'

. After conditioning on *ln*LOS, compared to female patients
male patients had about 35% higher odds of 'high costs'

Change in AOR away from 1.0 might suggest negative confounding

Reconciliation: Is negative confounding a possibility?

An example of negative confounding



. Negative confounding: Multiple effects of X with different signs

. *Direct Effect* is positive: Education → Liberalism

. *Indirect Effect* is negative: Education → $$$ → Liberalism

. *Total Effect = Direct Effect + Indirect Effect*

Under negative confounding,
Direct and Indirect Effects tend to cancel one another

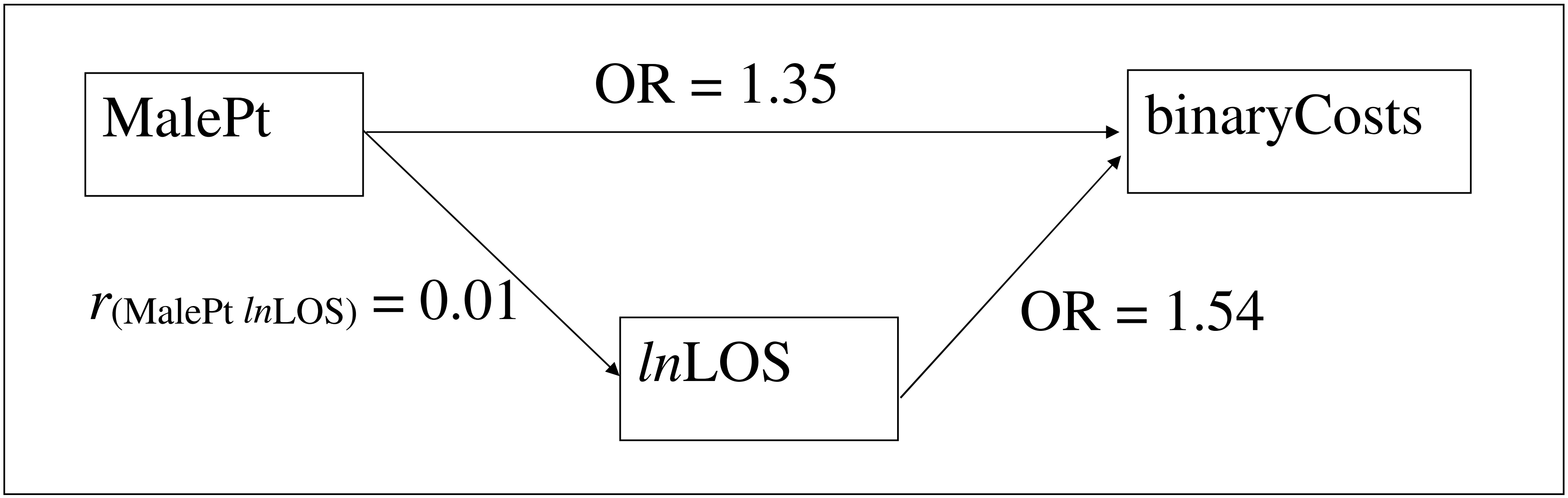
Reconciliation: Is negative confounding a possibility?

Correlations	MalePt	*ln*LOS
*ln*LOS	**0.01** (*p*=0.3464)	.
*ln*Costs	**0.04** (*p*=.0002)	**0.90** (*p*<.0001)
binaryCosts	**0.04** (*p*=.0016)	**0.72** (*p*<.0001)

. Here we see that $r_{(\text{MalePt } ln\text{LOS})} = 0.01$
potential for confounding of any type is low

Reconciliation: Is negative confounding a possibility?

Based upon information from the previous two slides,
we can create a path diagram



<u>Summary</u>

. The direct effect (OR = 1.35) is positive

. The indirect effect—a function of r=0.01 and OR=1.54—is also positive

. No negative confounding

Reconciliation: Underlying cause of discrepant results

Linear model: $y_i = \text{intercept}_{lin} + x_i b_{lin} + e_i$

. The residual variance is estimated and shrinks as the explanatory power of the model increases

Logistic model: $\text{logit}[\Pr(y_i=1 \mid x_i)] = \text{intercept}_{log} + x_i b_{log}$

. The residual variance of the logistic regression model is *fixed*

. equal to $\pi^2/3$ (variance of the standard logistic distribution)

. this is done for the purpose of model identification

Reconciliation: Underlying cause of discrepant results

Implications of fixing the residual variation in logistic regression

As additional *x* variables are added to a logistic regression model…

. Residual variance cannot be reduced (it is fixed by assumption)

. Something has to 'give'…

. Implied variation of the outcome increases: is *rescaled*

. Parameter estimates and ORs are also rescaled

All else being equal…

. Rescaled parameter estimates move away from zero

. Rescaled ORs (e.g., AORs) move away from 1.0

Reconciliation: Underlying cause of discrepant results

Implications of fixing the residual variation in logistic regression

Comparing parameters across nested logistic regression models
. e.g., AOR versus UOR

Operating Condition	Expectation: AOR v UOR
parameter rescaling*	AOR further away from 1.0 than UOR
negative confounding	AOR further away from 1.0 than UOR
confounding	AOR closer to 1.0 than UOR
combination	??? *possible counteracting effects*

* *with nested logistic models, some degree of parameter rescaling is always present*

Reconciliation: Underlying cause of discrepant results

…returning to the logistic regression example…

Modeling the effect of patient sex on binaryCosts

	Unadjusted Model			Adjusted Model		
	b	UOR	*p*	b	AOR	*p*
MalePt	0.14	1.15	.0016	0.30	1.35	<.0001
*ln*LOS	--	--	--	4.54	1.54†	<.0001

† $e^{b*\ln(1.1)}$: a 10% increase in LOS → about 54% increased odds of higher costs

. The UOR and AOR are on different scales; not directly comparable

. AOR > UOR: ?? neg. confounding, parameter rescaling, or both ??

. *For a direct comparison we need a UOR estimate for MalePt that is on the same scale as the corresponding AOR estimate*

Reconciliation: A way forward

Estimating a *rescaled* unadjusted effect of MalePt

. KHB method (Karlson, Holm, and Breen)

Step 1. Regress *ln*LOS onto MalePt and save residuals: $ln\text{LOS}_{resid}$

Step 2. Add $ln\text{LOS}_{resid}$ as an *x* variable:

i.e.,

$$\text{logit}[\Pr(\text{binaryCosts}_i = 1)] = \text{intercept} + \text{MalePt}_i b_1 + ln\text{LOS}_{resid\,i} b_2$$

The above model estimates an *unadjusted* effect of MalePt on an equivalent scale as the original AOR for MalePt

. *Method extends to accommodate multiple x variables & covariates*

Reconciliation: A way forward

The logic underlying the KHB method

$$\text{logit}[\Pr(\text{binaryCosts}_i = 1)] = \text{intercept} + \text{MalePt}_i b_1 + ln\text{LOS}_{\text{resid}i} b_2$$

. 1a: MalePt and lnLOS$_{resid}$ are uncorrelated

i.e., the effect of MalePt is not adjusted by lnLOS$_{resid}$

. 1b: any shared variation between MalePt and lnLOS is retained by MalePt, but removed from lnLOS$_{resid}$

i.e., the 'total' effect of MalePt is estimated

. 2: VAR(MalePt + lnLOS$_{resid}$) ≈ VAR(MalePt + lnLOS)

i.e., scaling of the parameter estimates is equivalent across the rescaled unadjusted and the original adjusted model

Reconciliation: A way forward

Returning to the logistic regression example…

Modeling the effect of patient sex on binaryCosts

	Rescaled Unadjusted Model (-2LL = 4811.75)			Adjusted Model (-2LL = 4811.75)		
	b	$UOR_{rescaled}$	*p*	b	AOR	*p*
MalePt	0.38	1.47	<.0001	0.30	1.35	<.0001
lnLOS_{resid}	4.54	1.54†	<.0001	--	--	--
*ln*LOS	--	--	--	4.54	1.54†	<.0001

† $e^{b*\ln(1.1)}$: a 10% increase in LOS → about 54% increased odds of higher costs

. The $UOR_{rescaled}$ and AOR are comparably scaled

. *Evidence of slight confounding: consistent with the observed correlation between MalePt & lnLOS. No negative confounding*

Conclusions

Given nested logistic regression models

. Parameter/OR rescaling always occurs as x variables are added

. When all added x variables have near zero effects,
then the degree of rescaling will be negligible

. When an added x variable has a substantial effect,
then some substantial rescaling will occur
The degree of rescaling will increase with variance of x

I.e., added x variables with large variance and large effects will induce largest levels of rescaling

Again, all types of nested logistic models,
not just unadjusted vs. adjusted

Conclusions

Be wary when comparing effects across nested logistic models…

…that appear to suggest negative confounding
(e.g., AOR is further away from 1.0 than UOR)

. *Probably, you are observing effects of rescaling*

…where the OR for a particular *x* variable does not appreciably change across nested models

. *Probably you are observing counteracting effects of adjustment by confounders and rescaling*

However, if the UOR is substantial, but the AOR is near 1.0, then you can attribute the change to adjustment for confounders.

Conclusions

. KHB method is simple to implement

. KHB method seems to do a good job of obtaining rescaled,
unadjusted point estimates

. Quality of KHB method standard errors/coverage
Coverage of rescaled unadjusted x effects was just OK
in a limited simulation that I conducted

. If you want to emphasize any tests of rescaled unadjusted effects,
the bootstrap should be considered

. Parameter/OR rescaling concerns go beyond logistic regression. Any model with residual variance fixed by assumption will have the same issues.

Resources

KHB papers (contact Kristian Karlson: kbk@sfi.dk)

1. Kristian Bernt Karlson, Anders Holm, and Richard Breen. (2012). Comparing Regression Coefficients Between Same-Sample Nested Models Using Logit and Probit: A New Method. *Sociological Methodology*, *42*, 286-313.

2. Kohler, U., Karlson, K.B., Holm, A. (2011). Comparing coefficients of nested nonlinear probability models. *The Stata Journal*, *11*, 420-438.

3. Breen, R., Karlson, K.B., Holm, A. (April 11, 2011). Total, Direct, and Indirect Effects in Logit Models. Abstract available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1730065

4. Karlson, K.B. and Holm, A. (2011). Decomposing primary and secondary effects: A new decomposition method. *Research in Social Stratification and Mobility*, *29*, 221-237. http://www.sciencedirect.com/science/article/pii/S0276562410000697

KHB Stata ado

http://ideas.repec.org/c/boc/bocode/s457215.html