Resource Labs 2&3: Generalized Linear Models (GLM)

To complete Labs 2 & 3, you will use a series of generalized linear models (GLM) to estimate risks, risk differences (RD), risk ratios (RR) and incidence odds ratios (IOR) with 95% confidence intervals (CI). This document provides an overview of STATA commands that fit these models and generate useful statistics for reporting results.

Summary of GLM

GLM are a family of modeling methods that can fit linear and non-linear models. They can be classified according to the distribution of the outcome (dependent, response) variable and the link function which specifies the relationship between the dependent variable (Y) and a linear combination of covariates (X_1, …, X_k), as summarized below.

Outcome	Regression model	Distribution	Link (g(Y))	Form
Continuous	linear	Normal	identity	$Y = \beta_0 + \beta_1X_1 \ldots + \beta_kX_k$
Binary	linear risk	binomial	identity	$R = \beta_0 + \beta_1X_1 \ldots + \beta_kX_k$
Binary	log-risk	binomial	log	$\ln(R) = \beta_0 + \beta_1X_1 \ldots + \beta_kX_k$
Binary	logit-risk	binomial	logit	$\text{logit}(R) = \beta_0 + \beta_1X_1 \ldots + \beta_kX_k$
Positive Integer (count)	Poisson	Poisson	log	$\ln(Y) = \beta_0 + \beta_1X_1 \ldots + \beta_kX_k$

Y = a continuous dependent variable (outcome) or count variable (Poisson models)

R = probability of a binomial outcome, e.g., risk (incidence proportion) or prevalence.

Distribution refers to the outcome variable.

Link: the functional relation between the dependent variable and the linear combination of covariates

(which is referred to as the linear predictor: $\beta_0 + \beta_1X_1 \ldots + \beta_kX_k$)

Model forms and estimation for linear, log and logit risk GLM models.

Model Type	Linear	Log Risk	Logit risk
Model Form	$\text{Risk} = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k$	$\ln(\text{Risk}) = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k$	$\text{logit}(\text{Risk}) = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k$
Outcome distribution	binomial	binomial	binomial
Link function	identity	ln()	logit()
Simple model	$\text{Risk}(\text{outcome} \mid X_1) = \beta_0 + \beta_1 X_1$	$\ln[\text{Risk}(\text{outcome} \mid X_1)] = \beta_0 + \beta_1 X_1$	$\text{logit}[\text{Risk}(\text{outcome} \mid X_1)] = \beta_0 + \beta_1 X_1$
		$\ln(R_0) = \beta_0 + 0 * \beta_1 = \beta_0$	$\ln(\text{Odds}_0) = \beta_0 + 0 * \beta_1 = \beta_0$
$\text{Risk}(\text{outcome} \mid X_1 = 0)$	$R_0 = \beta_0 + 0 * \beta_1 = \beta_0$		
		so $R_0 = \exp[\beta_0]$	so $\text{Odds}_0 = \exp[\beta_0]$
		$\ln(R_1) = \beta_0 + 1 * \beta_1 = \beta_0 + \beta_1$	$\ln(\text{Odds}_1) = \beta_0 + 1 * \beta_1 = \beta_0 + \beta_1$
$\text{Risk}(\text{outcome} \mid X_1 = 1)$	$R_1 = \beta_0 + 1 * \beta_1 = \beta_0 + \beta_1$		
		so $R_1 = \exp[\beta_0 + \beta_1]$	so $\text{Odds}_1 = \exp[\beta_0 + \beta_1]$
	Risk Difference	**Risk Ratio**	**Odds Ratio**
Risk comparison			
		$\ln(RR) = \ln(R_1 / R_0) = \ln(R_1) - \ln(R_0)$	$\ln(OR) = \ln(O_1 / O_0) = \ln(O_1) - \ln(O_0)$

	RD = R_1 - R_0 = [$\beta_0 + 1*\beta_1$] - [$\beta_0 + 0*\beta_1$] = [$\beta_0 + \beta_1$] - [β_0] = β_1	= [$\beta_0 + 1*\beta_1$] - [$\beta_0 + 0*\beta_1$] = [$\beta_0 + \beta_1$] - [β_0] = β_1	= [$\beta_0 + 1*\beta_1$] - [$\beta_0 + 0*\beta_1$] = [$\beta_0 + \beta_1$] - [β_0] = β_1
		so RR = exp(β_1)	so OR = exp(β_1)
	95% CI = β_1 +/- (1.96*SE(β_1))	and 95% CI = exp(β_1 +/- (1.96*SE(β_1)))	and 95% CI = exp(β_1 +/- (1.96*SE(β_1)))
STATA command	glm death bord5, fam(binomial) link(identity)	glm death bord5, fam(binomial) link(log) eform	glm death bord5, fam(binomial) link(logit) eform

Confidence Limit Difference (CLD) and Confidence Limit Ratio (CLR) are quick and useful measures of the precision of a parameter estimate and make it easier to compare estimates across studies. For example two studies may report similar point estimates, but one may have a much wider 95% CI, meaning it's value is less precise.

CLD = [upper 95% CI] - [lower 95% CI]

CLR = [upper 95% CI] / [lower 95% CI]

STATA commands to obtain useful information from models

1. Generating predicted values (risks) from log risk models

   - Use the ‘glm’ command to run a log risk model.

   - Use the ‘predict’ command (after running a glm model) to generate a new variable (predvar). This command calculates the predicted risk of the outcome for each participant/patient on the dataset, based on the covariable values for each individual, and stores the result in the variable predvar. For the model below, bord5 is a binary variable, so only 2 values of risk are possible – one each for the 2 levels of bord5.

   - Now use the ‘predict’ command to generate a new variable (*se*), which is the standard error for each value of bord5.

   - Generate variables for the upper and lower confidence limits for the estimated risk (pul for the upper and pll for the lower limit).

   - The command summ predvar generates a univariate description of the values of the predicted probability of the outcome.

   - The other sum commands give the estimated risk and 95% CI limits for the different values of the exposure variable bord5.

   - STATA commands

```
glm death bord5, fam(binomial) link(log) eform
predict p
predict se, stdp
gen pul = p + (1.96*se)
gen pll  = p  - (1.96*se)
summ p, de
summ p pll pul if bord5 == 0
summ p pll pul if bord5 == 1
```

2. Generating contrasts from CI from GLM models

   - The ‘lincom’ command can generate point estimates and CI for any combination of covariable values in the model; these are referred to as contrasts.

   - Use the ‘glm’ command to run a linear risk model.

   - Use the ‘lincom’ command to calculate the predicted risk of death for a child with birth order 1 – 4 (bord5 = 0 and R = β_0)

```
lincom _cons
```

   - Use the ‘lincom’ command to calculate the predicted risk of death for a child with birth order >= 5 (bord5 = 1 and R = $\beta_0 + 1*\beta_1$)

```
lincom  cons + bord5
```

- Use the 'lincom' command to calculate the difference of death for birth order >= 5 vs 1 – 4 (RD = $\beta_0+\beta_1 - \beta_0 = \beta_1$)

```
lincom  bord5
```

- Use the 'lincom' command option eform when using log or logit risk models

- These above are trivial examples with one covariable. The 'lincom' command is very handy for multivariable models.

Nominal Variables, Ordinal Variables and Disjoint Indicator Variables

1. Categorical variables can be represented as having values 1, 2, 3, …, but one must be careful with such representations. Nominal variables have no inherent ordering (e.g., male = 0, female = 1, Kleinfelter's = 2) and ordinal variables may be qualitatively ordered but may not have uniform linear spacing (e.g., low = 0, medium = 1, high = 2). Including such variables in models as linear terms means that the model is mis-specified and can lead to erroneous inference because the relationship between the outcome and categorical is assumed to be linear.

2. Disjoint indicator (aka, dummy) variables derived from nominal or ordinal categoricals removes the linear assumption and allows more flexibility in the shape of the outcome-predictor association.

3. Disjoint indicator variables are derived from categoricals by generating k new variables, one for each of the k levels of the categorical, as illustrated in the table below. You could leave out one of the indicator variables (the reference level), but I prefer to code all levels to allow flexibility in changing the reference level as needed.

			Indicator variables	
Original variable		ed_none	ed_primary	ed_post
education = 0	no primary school	1	0	0
education = 1	primary school only	0	1	0
education = 2	post-primary education	0	0	1

4. Example – linear-risk regression model for education, coded with indicators

- Risk(preterm|education) = $\beta_0 + \beta_1 X_1 + \beta_2 X_2$
  where X_1 = ed_primary and X_2 = ed_post and the referent category is no primary school

- In this model

  - No education $R_0 = [\beta_0 + 0*\beta_1 + 0*\beta_2] = \beta_0$

- Primary school $R_1 = [\beta_0 + 1*\beta_1 + 0*\beta_2] = \beta_0 + \beta_1$
- Post primary $R_2 = [\beta_0 + 0*\beta_1 + 1*\beta_2] = \beta_0 + \beta_2$

- To calculate RDs with "no primary school" as the reference group and "primary school only" race as the index group

  - RD [primary vs. no primary] $= R_1 – R_0$
    $= [\beta_0 + 1*\beta_1 + 0*\beta_2] - [\beta_0 + 0*\beta_1 + 0*\beta_2]$
    $= [\beta_0 + \beta_1] - [\beta_0] = \beta_1$

- To calculate RDs with "no primary school" as the reference group and "post-primary education" as the index group

  - RD post-primary vs. no primary] $= R_2 – R_0$
    $= [\beta_0 + 0*\beta_1 + 1*\beta_2] - [\beta_0 + \beta_1(0) + 0*\beta_2]$
    $= [\beta_0 + 1*\beta_2] - [\beta_0] = \beta_2$

- RR and OR are calculated similarly

STATA commands to generate indicator variables

There are at least 3 ways to make or model disjoint indicator variables in STATA.

- Use the 'tab' command to automatically generate indicator variables. For example

  ```
  tab education, gen(education_c)
  ```

  will create 3 indicator variables numbered consecutively from the lowest value of education to the highest, so that for education coded (0,1,2), the following indicator variables would be created:

  ```
  education_c1 = 1 if education==0, 0 otherwise
  education_c2 = 1 if education==1, 0 otherwise
  education_c3 = 1 if education==2, 0 otherwise
  ```

  To estimate RR for primary vs. no primary and post-primary vs. no primary you would then specify the model:

  ```
  glm death eduation_c2 education_c3, fam(bin) link(log)
  ```

  Or, to model the data with primary education as the referent:

  ```
  glm death education_c1 education_c3, fam(bin) link(log)
  ```

- You can also manually create k-1 indicator variables for an exposure with k categories.

  ```
  gen     education_c1  = 0 if education ~= .
  replace education_c1  = 1 if education == 1
  label def education_c1 0 "Primary school or post-primary" 1 "No education", modify
  label val education_c1 education_c1

  gen     education_c2 = 0 if education ~= .
  replace education_c2 = 1 if education == 2
  label def education_c2 0 "No primary school or Post primary school" 1 "Primary
  education", modify
  label val education_c2 education_c2

  gen     education_c3 = 0 if education ~= .
  replace education_c3 = 1 if education == 3

  label def education_c3 0 "No primary school or Primary school" 1 "Post-primary
  education", modify
  label val education_c3 education_c3
  ```

  To estimate RR for primary vs. no primary and post-primary vs. no primary you would then specify

  ```
  glm death education_c2 education_c3, fam(bin) link(log)
  ```

- **NOT RECOMMENDED FOR THIS LAB** [This is what we did in lab 1]. Indicator variables can be created automatically using the 'xi:' model command prefix, with the 'i.' prefix used to indicate each categorical exposure variable for which indicator variables are needed [*Note, in more recent versions of Stata, you don't need the 'xi' command, only the 'i.' command*]. This method can be used with any Stata model command.I include this here because many researchers no doubt use this method, but I strongly recommend against using this method. I prefer to precisely control the specification of indicator variables because in doing so there is no doubt about what these variables represent.

  For example

  ```
  xi:glm death i.education, fam(binomial) link(identity)
  ```

  will automatically generate and model k-1 indicator variables for education, with the lowest-value category of education used as the common referent category by default. The new variables will be listed in the variable window and shown in model output. For the model specified above, education = 0 would be the referent, and the two indicator variables would be _Ieducation_1 and _Ieducation_2, where:

  ```
  _Ieducation_1 = 1 if education==1, 0 otherwise
  _Ieducation_2 = 1 if education==2, 0 otherwise
  ```

  You can use the 'char' command to designate a different exposure category as the referent group. For example, to designate education=1 (Primary school only) as the common referent category:

  ```
  char education[omit] 1
  xi:glm death i.education, fam(binomial) link(identity)
  ```

  To revert back to the default referent (i.e., education = 0), enter:

  ```
  char education[omit] 0
  ```

  Note: It is not necessary to save the indicator variables that are created when you use the 'xi' model prefix since it is easier to re-generate them using the 'xi' prefix than it is to write out the new variable names the next time you run the same model. However, they will be maintained in your data set unless or until they are overwritten by new indicator variables.

Categorical vs. Continuous Coding

Some exposure variables may be either analyzed as continuous or categorized according to customary or clinical cut points (e.g., body mass index (BMI) categories based on values used to classify people as ‘underweight’, ‘normal’, ‘overweight’ or ‘obese’), empirical cut points (e.g. tertiles or quartiles), a *priori* cut points relevant to biologic mechanisms, potential public health interventions, other factors of interest or by cut points used commonly in the literature (this allows comparisons between your results and what has been published previously). The least restrictive way to model continuous variables would be to estimate separate risks for each value (e.g. each year of maternal age). However, this approach would yield highly imprecise estimates and is not recommended.

Graphing risks versus exposures: categorical variables

Refer back to section 2 of “STATA commands to obtain useful information from models” where we generated upper and lower confidence limits from a model. Now use the ‘scatter’ command to generate a scatter plot of the risks and upper and lower confidence limits (predvar, pll and pul as in the previous example) on the y axis and bord5 on the x axis.

```
scatter pll predvar pul bord5
```

There are several options you can use to add titles and other features to your graph, as shown in the example below. Note that there are options within the title options (e.g., margin(medium)) that will affect how things are displayed:

```
scatter pul predvar pll mage, ytitle("Estimated Risks (95% CI) for Death",
margin(medium) size(small)) xtitle(Maternal age in years, size(small))
title("Figure 1: Risk of Death in Association with Maternal Age (Categorical)", size(small))
subtitle("DHS Kenya, 2008", size(small)) msymbol(o o o)  mcolor(gs10 gs0 gs10)
lcolor(gs10 gs0 gs10) lstyle(p1 p1 p1) legend(row(2) notextfirst order(2 1)) sort(mage)
connect(J J J i)
```

The ‘sort’ and ‘connect’ options are used to connect the lines between each value of p, pul and pll. The (JJJ i) part of the ‘connect’ option tells Stata that you want to connect the first three variables (pll, p, and pul in the ‘scatter’ command statement above) in a stair-step fashion (indicated by ‘J’s), but do not want to connect the fourth variable in the command statement (mage, indicated by the ‘i’ in the fourth position). This code also includes a title and subtitle- these would not be appropriate for publication, but can be helpful when you generate graphs for your own use.

Drop the variables p, se, pll and pul after generating the graph so you can use the same variable names the next time you create predicted risk variables for a graph.

Assess effect measure modification: Main effects models, product interaction terms and the constancy assumption

- Generalized linear models can be used to estimate adjusted effect measures. Like pooling and the Mantel-Haenszel method, model-based estimates assume constancy. In this section of the lab, we will begin by testing the constancy assumption using likelihood ratio tests (LRT).

- We assess effect modification by comparing two models, such as:
    - Main effects only P(death) = $\beta0 + \beta\ 1(X1) + \beta\ 2(X2)$
    - Main effects with interaction terms P(death) = $\beta0 + \beta\ 1(X1) + \beta\ 2(X2) + \beta\ 3(X1 * X2)$

- We evaluate the difference between the models using a) likelihood ratio test or b) the test statistic for the interaction term estimate β 3 (more on this topic below).

- The main effects model assumes constancy (i.e., the association between the outcome and X1 is the same for all levels of X2 (and the converse).

  Consider the main effects model: Risk(death) = $\beta_0 + \beta_1(\text{bord5}) + \beta_2(\text{male})$:

The RD for high birth order vs. low birth order among male children is:

$\text{RD}_{\text{bord5=1 vs. bord5=0}} = [\beta_0 + \beta_1(1) + \beta_2(1)] - [\beta_0 + \beta_1(0) + \beta_2(1)] = \beta_1(1)$

The RD for high birth order vs. low birth order among female children is:

$\text{RD}_{\text{bord5=1 vs. bord5=0}} = [\beta_0 + \beta_1(1) + \beta_2(0)] - [\beta_0 + \beta_1(0) + \beta_2(0)] = \beta_1(1)$

The RD for male children vs. female children among those with high birth order is:

$\text{RD}_{\text{male=1 vs. male=0}} = [\beta_0 + \beta_1(1) + \beta_2(1)] - [\beta_0 + \beta_1(1) + \beta_2(0)] = \beta_2(1)$

The RD for male children vs. female children among those with low birth order is:

$\text{RD}_{\text{male=1 vs. male=0}} = [\beta_0 + \beta_1(0) + \beta_2(1)] - [\beta_0 + \beta_1(0) + \beta_2(0)] = \beta_2(1)$

In contrast, a product interaction model relaxes the constancy assumption (i.e., the association between the outcome and X_1 is the different among levels of X_2 and the converse)

Consider the interaction term model: Risk(death) = $\beta_0 + \beta_1(\text{bord5}) + \beta_2(\text{male}) + \beta_3(\text{bord5*male})$

The RD for high birth order vs. low birth order among male children is:
RDbord5=1 vs. bord5=0 = [b0 + b1(1) + b2(1) + b3(1*1)] – [b0 + b1(0) + b2(1) + b3(0*1)] = b1(1) + b3(1*1)
The RD for high birth order vs. low birth order among female children is:
RDbord5=1 vs. bord5=0 = [b0 + b1(1) + b2(0) + b3(1*0)] - [b0 + b1(0) + b2(0) + b3(0*0)] = b1(1)
The RD for male children vs. female children among those with high birth order is:
RDmale=1 vs. male=0 = [b0 + b1(1) + b2(1) + b3(1*1)] – [b0 + b1(1) + b2(0) +b3(1*0)] = b2(1) + b3(1*1)
The RD for male children vs. female children among those with low birth order is:
RDmale=1 vs. male=0 = [b0 + b1(0) + b2(1) + b3(0*1)] – [b0 + b1(0) + b2(0) + b3(0*0)] = b2(1)

Note : The model including the interaction term bord5*male is a **saturated model** because we obtain all 4 possible estimates from the included variables (2x2 = 4 for all levels of bord5 by male) by including 4 terms in the model (β_0 counts as a term). The previous model without the interaction term yields only 2 estimates because of the constancy constraint.

If the constancy assumption is valid, adding the product interaction term contributes very little additional information to the model. Consequently, the coefficient for the product interaction term (β_3) will be very small, so that $\beta_1(1) + \beta_3(1*1)$ is approximately equal to $\beta_1(1)$ and the p-value of the interaction term will be non-significant.

In addition, the model that includes the interaction term will not “fit” the data substantially better than the model with the lower order terms only, so that there will be very little difference in the likelihood of the observed data between the two models. The null hypothesis for a likelihood ratio test is that the likelihood of the observed data is the same for the two models being compared; therefore, a small p-value by the LRT indicates that the data are not consistent with the homogeneity assumption.

In general the coefficient p-value will be similar to the LRT p-value

The Likelihood Ratio Test (LRT)

In general, the LRT can be used to determine whether a larger model with extra variable(s) fits the data better than a smaller model without the extra variable(s) (e.g., 2 models with and without an interaction term or 2 models with and without age as a term).

Specifically, the LRT statistic tests the null hypothesis that a larger model maximizes the likelihood of the observed data
no better than a reduced model that includes fewer covariates.

Critical point – the models must be strictly comparable, so

- the variables in the reduced model must be a **strict subset** of those in the larger model
- the observations (e.g., people) in both models must be **identical**; the models are not strictly comparable if you have missing data in a variable included in the larger model because you are making estimates on different datasets
- if either of these conditions is not met, the LRT is invalid because you are comparing apples to oranges.

In this section of the lab, we will use the LRT to determine whether a model that allows effect estimates to vary across covariate strata (heterogeneity) fits the data better than a model that assumes a constant RD, RR or OR (homogeneity or constancy).

The likelihood ratio test statistic is 2X the difference between the log likelihoods of [the reduced model – the full model]

LRT statistic = (-2*log likelihood (reduced model)) – (-2*log likelihood (full model)) = (deviance (reduced model) - (deviance (full model)

The LRT statistic is distributed as a Chi-square with degrees of freedom (df) = (# variables in full model) – (# variables in the reduced model)

Likelihood ratio test calculation in STATA - you can use the ‘lrtest’ command to perform a likelihood ratio test comparing the two models.

- Run your ‘reference’ model and store its estimates as ‘A’ using the ‘est store’ command

```
gen pnc5_smoker = pnc5 * smoker
glm preterm pnc5 smoker pnc5_smoker, link(ident) fam(bin)
est store A
```

Note – ‘A’ will appear as a new variable in your variables list

Run the model you want to compare with model ‘A’ and use the ‘lrtest’ command to run the LRT. Note, the commands in [ ] are optional because the ‘lrtest’ command assumes the most recent model as the one to use for comparison.

```
glm preterm pnc5 smoker pnc5_smoker, link(ident) fam(bin)
[est store B]
lrtest A
[lrtest A B]
```

If you enter ‘lrtest A’ after a fitting a third model, Stata will perform a new likelihood ratio test comparing the likelihood of the third (most recent) model to that of the first model, since the first model's results are still stored as 'A'. You can re-use the same variable name (A) to store results of a new reference model by repeating 'est store A' after fitting the new model – STATA will simply overwrite the previous data. To store estimates from multiple models, assign new variable names to each set of estimates (e.g., est store B, est store C, etc.) Drop these temporary variables before you save your data so you can re-use these names again, since they will no longer be recognized as temporary variables once the dataset is saved.

Evaluating the validity-precision trade-off assuming constancy

Validity is how closely an estimate of an association comes to the 'truth' (i.e., the opposite of bias) and we cannot generally know the validity of a measure in observational epidemiology because

- we cannot make measurements in the entire population of the planet (the 'truth') because the observed measurements depend on the sample we are studying and the luck of the sample we draw
- measures may change over time (asthma, autism, etc.)

Precision is a measure of the accurately we measure an estimate of association (e.g., a 95% confidence interval), regardless of validity.

Consider a dartboard, where the center ('cork' in darts-speak) is the 'truth'. You might throw 3 darts into the board that are widely spread, but average to the center. This situation is analogous to a valid, but not very precise, measure. On the other hand, you might throw 3 darts into the triple (inner ring) 20 segment. This situation is analogous to a precise, but significantly biased, measure.

There is a trade-off between validity and precision. We control for confounding to reduce bias in our estimate of an association, but the trade-off here is that the precision of an estimate generally decreases as we add variables to a model. The question to address is how to balance validity and precision as we model our system.

So, if the difference between a crude and adjusted main exposure effect estimate is "large", then the amount of bias due to confounding is "large" and we might be inclined to sacrifice some precision to obtain a less biased estimate.

However, if the difference between a crude and adjusted main exposure effect estimate is "small" but the loss of precision is significant, then we must evaluate whether or not to adjust for the confounder.

On the other hand, if the difference between a crude and adjusted main exposure effect estimate is "small" and the loss of precision is minor, the penalty for adjustment may be negligible.

We will use a "mean squared error" approach to contrast the "penalty" from reduced precision with the "gain" from increased validity that occurs when we model the minimally sufficient adjustment set of confounders identified using a DAG.

The mean squared error (MSE) of an effect estimate is approximated as (MSE = bias^2 + variance), where the "bias" is the difference between an adjusted estimate and a crude (or reduced) estimate.

Here, the component of bias we are considering is just the confounding by one or more adjustment variables in the minimally sufficient conditioning set we identified by analyzing our DAG. We are assuming that our DAG is correct, that we have measured all our variables accurately, and that we have specified them properly in our model.

For the RD, the reduction in bias (i.e., the reduction in confounding) is the change in the RD, $B = RD_{adjusted} - RD_{reduced}$.
We square this value and add it to the change in the variance, $\Delta V = var(RD_{adjusted}) - var(RD_{reduced})$ and call this value M.
For two models, unadjusted model 1 (with $MSE=M1=B1^2 + \Delta V1$) and more-adjusted model 2 (with $MSE=M2=B2^2 + \Delta V2$).
If M1>M2, the validity-precision tradeoff favors adjustment. If M1<M2, the tradeoff favors not adjusting.

For the RR and IOR, we assess the validity-precision tradeoff on the natural log scale.

For the RR: $B = \ln RR_{adjusted} – \ln RR_{reduced}$ and $\Delta V = var(\ln RR)_{adjusted} – var(\ln RR)_{reduced}$.

For the IOR, $B = \ln IOR_{adjusted} – \ln IOR_{reduced}$ and $\Delta V = var(\ln IOR)_{adjusted} – var(\ln IOR)_{reduced}$

The validity-precision tradeoff can be assessed starting with a fully-adjusted estimate from a model that includes all of the covariates in the minimally sufficient adjustment set, and then considering the validity-precision tradeoff for each adjustment variable by deleting them one-by-one from the model. If this approach is taken and more than one adjustment variable is dropped, the tradeoff should be assessed again, this time comparing "full adjustment" with adjustment for the reduced set of covariates. You will be doing this in section B of this lab.

- Note that all change in estimate methods, including the one described above, assume that "adjusted" estimates are less biased than unadjusted estimates; however, "adjustment" can increase bias if the covariate is poorly measured or incorrectly modeled. In addition, change in estimate methods cannot identify covariates that are affected by the outcome or the exposure. Always use prior information to determine whether to adjust and how to adjust for covariates, before using change in estimate methods to assess confounding.
- If confounders are identified (i.e., using a DAG), measured, specified and modeled appropriately, a "fully adjusted" estimate will be the most unbiased; therefore, it should be the standard to which less adjusted estimates are compared.

This may be referred to as "backward deletion" when used to select a subset of confounders for adjustment, since change is assessed as covariates are removed from a fully-adjusted model.

Alternatively, a "forward selection" strategy may be used when it is not possible to model all potential confounders simultaneously (e.g., when data are sparse or covariates are highly correlated), with changes in precision and validity evaluated relative to a crude effect estimate as confounders are added to a model.

These methods apply to the situation where you have a main exposure of interest. If your interest is in prediction or if numerous covariables are of equal interest, there are other considerations involved to evaluate model building.