PUBH 8343 Notes

Causal inference

We want to estimate Pr[Y|X=1] – Pr[Y|X=0], which is the predicted outcome in a counterfactual setting

We can only observe [Y| set X=1] – [Y| set X=0] in our observational studies

Causal inference gives us tangible assumptions that we need to work through to use our observational studies as tools to assess potential associations using a causal inference framework

Causal inference Assumptions

Types of people in our counterfactual world

Type	Name	Y|X=1	Y|X=0	Y|X=1 - Y|X=0
Type 1	Doomed	1	1	0
Type 2	Causative	1	0	1
Type 3	Protective	0	1	-1
Type 4	Immune	0	0	0

Exchangeability: Can I use the control group as a believable counterfactual substitute for the exposed group?

Type	Name	Y|X=1	Y|X=0	Proportion of the Population
Type 1	Doomed	1	1	p1
Type 2	Causative	1	0	p2
Type 3	Protective	0	1	p3
Type 4	Immune	0	0	p4

In our population, if we force everyone to X = 1, then p1 + p2 will get the outcome Y=1

In our population, if we force everyone to X = 0, then p1 + p3 will get the outcome Y=1

RD = Pr[Y|X=1] – Pr[Y|X=0] Becomes [p1+p2] – [p1+p3] = p2-p3 in a counterfactual framework

However, this is not possible in the real world, so we need to have two populations where we forced everyone to X=1 in population A and everyone to X=0 in population B

Type	Name	Y|X=1	Y|X=0	Population A	Population B
Type 1	Doomed	1	1	p1	q1
Type 2	Causative	1	0	p2	q2
Type 3	Protective	0	1	p3	q3
Type 4	Immune	0	0	p4	q4

Effect of picking different target populations

Target population A+B > RD = (p2 + q2) – (p3 + q3)

Target population A > RD = p2 – p3

Target Population B > RD = q2 – q3

In reality, Population A would be our exposed group and population B would be our control group, so

RD = Pr[Y|X=1] – Pr[Y|X=0] becomes [p1+p2] – [q1+q3]

Assuming exchangeability (i.e. B is a good counterfactual substitute for A) we can estimate this observed RD

Strong exchangeability: p1 = q1, p2=q2, p3=q3, p4=q4. This means that we can substitute p1 + p3 in our counterfactual formula for q1 + q3. Technically, you only need the weak exchangeability (p1 + p3) = (q1 + q3), but the best way to achieve weak exchangeability is to achieve strong exchangeability through randomization.

Lack of exchangeability makes our estimates confounded. Furthermore, confounding depends on the target population.

Type	Name	Y|X=1	Y|X=0	Population A	Observed Proportion	Population B	Observed Proportion
Type 1	Doomed	1	1	p1	0.2	p1	0
Type 2	Causative	1	0	p2	0.5	p2	0.4
Type 3	Protective	0	1	p3	0.1	p3	0.3
Type 4	Immune	0	0	p4	0.2	p4	0.3

If A is the target population

- Counterfactual RD = (P1 + P2) – (P1 + P3) = 0.4
- Observed RD = (P1 + P2) – (Q1 + Q3) = 0.4

However, if B is the Target population

- Counterfactual RD = (Q1 + Q2) – (Q1 + Q3) = 0.1
- Observed RD = (P1 + P2) – (Q1 + Q3) = 0.4 > This is because P1 + P2 =/= Q1 + Q2

Positivity: The probability of being exposed is non zero for each group. Otherwise, you would not have a comparison group and this would be a positivity violation. More realistically, this happens when we over stratify our data so you end up with strata without enough participants, or you can have logical positivity violation, whereby there can be no observations (stratifying HRT on gender wouldn't make sense as men would not be exposed). This is also called structural confounding.

Consistency: What would happen in the counterfactual world should be confirmed when we set the participant to the corresponding exposure in the real world. In practical terms, this is important because the way we evaluate exposures in a group should be consistent, using well defined interventions and and consistent dosage.

Confounding

In observational epidemiology, we do not randomize, but we assume that people in our sample were randomly exposed. It is not a reliable assumption so we need to own our assumptions, use weighting methods (standardizations, IPW), and doing sensitivity analyses.

Confounding is a state of non exchangeability: A confounder is a factor that is a risk factor for the disease, associated with the exposure, and not in the causal pathway. Accounting for confounders improves exchangeability within the strata of confounder.

Comparing crude Risk differences to standardized differences when you have a confounder

With Z we can't calculate P[Y=1|X=1] – P[Y=1|X=0)

Instead, we stratify on Z and standardize, assuming that Z has the same distribution in X and not X:

P[Y|X,Z]*P[Z] + P[Y|X,not Z]*P[not Z] - P[Y|not X,Z]*P[Z] + P[Y|not X, not Z]*P[not Z]

If we assumed that Z had a different distribution in X and not X, we would have:

P[Y|X,Z]*P[Z|X] + P[Y|X, not Z]*P[not Z|X] - P[Y|not X,Z]*P[Z|not X] + P[Y|not X, notZ]*P[not Z|notX]

However, since P[A|B]*P[B] = P[A,B] the formula above becomes

P[Y,Z|X] + P[Y, not Z|X] - P[Y,Z|not X] + P[Y, not Z|not X]

And since P[A, B] + P[A, not B] = P(A), we have P[Y|X] - P[Y|not X] which means that standardizing when the Z distribution is not held constant between X and not X would have no effect.

Interpreting the rules of a confounder using the standardization formula

P[Y|X,Z]*P[Z] + P[Y|X, not Z]*P[not Z] - P[Y|not X,Z]*P[Z] + P[Y|not X, not Z]*P[not Z]

Z must be a risk factor of Z within the reference level of exposure: If Z is not related to X, then P[Y|not X, Z] would be the same as P[Y|not X, not Z] so the formula would not work. In practice, this is why you look at the association between Z and X in your unexposed.

Z must be associated with the exposure: The crude RD is P(Y|X,Z)*P(Z|X) + P(Y|X, not Z)*P(not Z|X). If Z and X are not associated, then P(Z|X) would be the same at P(Z).

You do not want to condition on X and Z on Y, because Y it is a collider and that would always throw off your results.

Z cannot be affected by the exposure or disease: otherwise, it would become an intermediate in the former, and a collider in the later

Hypothesis Testing in a counterfactual framework

Average null Hypothesis: Pr(Type 2) = Pr (Type 3) so Pr(Type 2) - Pr (Type 3) = 0

Sharp null hypothesis Pr(Type 2) = Pr (Type 3) = 0; this means that there is nobody causal or doomed in our population. This means that everyone we see with Y = 1 would be doomed, and that a random

proportion of them were assigned to being exposed. Let M be the proportion of doomed people in our population, and A be the proportion of Doomed people in our population assigned to X = 1.

	Y=1	Y=0	Marginal of X
X=1	A	N1 - A	N1 (Exposed)
X=0	M - A	(N-M) - (N1 - A)	N0 (Unexposed)
Marginal of Y	M (Doomed)	N-M (Immune)	N (Total)

In order to estimate P(A), we use the marginal N1 and M and the potential values of P(A) take a hypergeometric distribution.

Combination formula for the permutation test (read M choose a, a being a subset of values for M)

$$\binom{M}{A} = \frac{M!}{a! * (M-a)!}$$

Where M! represents all of the ways you can combined things in M, and a! gets rid of the redundant combinations in M. Let us say M = 3, and a =2. We would want all of the possible combinations of 2 numbers between 1, 2 and 3 for M! = 3! = 6, i.e. (1,2) (1,3) (2,3) (2,1) (3,1) (3,2) and a! = 2! = 2 would remove the duplicate combinations, (2,1) (3,1) (3,2).

These distributions are used to replace the Chi-squared test (not reliable with low cell counts) with Fisher's exact test using a hypergeometric distribution and the combination formula.

Example: Let us assume that N = 6, M = 4 and N1 = 3. The marginals will be fixed, and we are interested in all the possible values that A can take

	Y=1	Y=0	Marginal of X
X=1	2	1	3
X=0	2	1	3
Marginal of Y	4	2	6

	Y=1	Y=0	Marginal of X
X=1	1	2	3
X=0	3	0	3
Marginal of Y	4	2	6

	Y=1	Y=0	Marginal of X
X=1	3	0	3
X=0	1	2	3
Marginal of Y	4	2	6

Let us say we wanted a p-value for P(A = 2). We would need to determine the probability distribution function for all values of A (A=1. A=2 and A= 3), in order to find the corresponding p-values.

You could use stata to find the probability of (A = a) given all of the values of A

- di comb (X, a) ***for each individual value of A
- di hypergeomcomb (N, M, N1, A=a) *to have stata determine the hypergeometric distribution at once and calculate all the relevant probabilities for A = a.

Then you can just use laws of probability to find each corresponding p-value

- Right sided p = P(A>=2) = P(A=2 or A=3)
- Left sided p = P(A=<2) = P(A=1 or A=2)
- Two sided p = P(A=<2 or A>=2) = P(A) + P(B) – P(A and B)

Probability review

P(A,B) = Joint probability

P(A|B) = Conditional probability

P(A|B) = P(A,B) / P(B)

P(A) = P(A,B) + P(A,not B) = Law of total probability

Relationship between different probability rules

				P(B) = P(A,B) + P(not A, B)				
				^				
P(B|A)	<	P(A,B) / P(A)	<	P(A,B)	>	P(A,B)/P(B)	>	P(A|B)
				v				
				P(A) = P(A,B) + P(A, not B)				

Going from P(B|A) to P (A|B) is known as Bayes theorem

	Y=1	Y=0	Marginal of X
X=1	P(A,B)	P(notA, B)	P(B)
X=0	P(A,notB)	P(Not A, Not B)	P(notB)
Marginal of Y	P(A)	P(Not A)	1

Let Z be a confounder that we stratify on. Going from P(A|Z) to P(A) is known as standardization

G methods introduction

Non Parametric G-formula (i.e, standardization)

- Point exposure to continuous exposures
- Time varying exposures

Parametric G formula (i.e. marginal structural models)

- Point exposure to continuous exposures
- Time varying exposures

Controlling for confounding is a substantive knowledge question, not a statistical one. However, we do have analytical techniques which allow us to control for confounding. During our analysis, we first look for effect measure modification, and if we do not have any, we keep the Z variables which are potential confounder and report a pooled estimate.

Stratification methods allow us to look at the effect of X or B1 on Y within levels of Z. For example, in regression, we are looking at the stratum specific effects of X on Y in Z because it is held constant in our model formula. However, using an interaction term between X and Z allows Z to vary, which makes it more similar to a pooled estimate effect.

Standardization methods allow us to average an effect over all the levels of Z, by making the distribution of Z matchup across all levels of X. This implies uniformity, which may be problematic if interaction is present, because interaction indicates that there is a difference in the effect of X on Y between levels of Z.

- Crude Risk Difference: P(Y|X) – P(Y|notX)
- Standardized Difference: [P(Y|X,Z)*P(Z) + P(Y|X, not Z)*P(not Z)] - [P(Y | not X,Z)*P(Z) + P(Y| not X, not Z)*P(not Z)]

Assuming P(Z) = P(B) and P(A,B) = P(Y|X,Z), then we know that P(A,B)*P(B) + P(A, not B) * P(not B) = P(A) which would be P(Y|X) in this example.

G Formula Factorization: We are going from P(B|A) for Z to P(A,B) for X to P(A) for Y. We will look at this in only a portion of the formula, but the full G formula is

[P(Y=1|X=x, Z=z)*P(X=x|Z=z)*P(Z=z) + P(Y=1|X=x, Z=not z)*P(X=x|Z=not z)*P(Z= not z)] –

[P(Y=1|X=not x, Z=z)*P(X=not x|Z=z)*P(Z=z) + P(Y=1|X= not x, Z=not z)*P(X=not x|Z=not z)*P(Z=not z)]

- G Formula for the exposed: P(Y=1|X=x, Z=z)*P(X=x|Z=z)*P(Z=z)
- P(X=x|Z=z)*P(Z=z) = P(A|B) * P(B) = P(A,B) = P(X=x,Z=z)
- So the G formula becomes: P(Y=1|X=x, Z=z)*P(X=x,Z=z)
- Again, P(Y=1|X=x, Z=z)*P(X=x,Z=z) = P(A|B) * P(B) = P(A,B) = P(Y=1, X=x, Z=z)
- When we force X =1 or X = 0, P(X=x|Z=z) becomes 1, and the G-formula becomes the Standardization formula: P(Y=1|X=x, Z=z)*P(X=x|Z=z)*P(Z=z) > P(Y=1|X=1,Z=z)*P(Z=z) for the exposed, and the same for the unexposed.

G Formula Analysis

** Use the cs command for 2x2 tables between your exposure and your outcome (Gives you an OR and RD)

Overall CS / CS for Z = 1 / CS for Z = 0

**Expand the dataset to generate observations for your pseudo population

Flag = 1 for the original observations

Flag = 2 for all observations where we force X = 1

Flag = 3 for all observations where we force X = 0

** Run a regression (glm for continuous, logistic for binary outcomes) and use its results to generate predicted values for the outcome for the pseudopopulation (Flag = 2 and Flag = 3)

Replace the predicted outcomes for X = 0 and X = 1 separately

**Calculate the RD among all values in the pseudo population: Pr(Y=1|X=1) – Pr(Y=1|X=0)

Monte Carlo Method

Follows the chain of events in your DAG is a chronological manner, from Z to X to Y.

Create a large population (N = 10000): The larger the N, the better the Y estimate will be

Determine the distribution of Z in your original dataset (sum Z: gets the min, max, std and mean) by saving these values as a temporary variable to generate values for Z in the simulated dataset

Store the coefficient from the regression of Y on X and Z in the original dataset to later use them for predicting y in the simulated dataset (logistic Y X Z > matrix list e(b)

Generate IDs for everyone in the new dataset

Generate values for Z in a simulated dataset

Force X to take predetermined values > get mean values for X = 0 and X =1

Simulate values of Y given Z and X. The predicted values of Y given X and Y will come from a initial regression of Y on X and Z in the original dataset

Marginal structural models

In G formula, you model the outcome Y after removing the association between X and Z

In Marginal structural models, you model the exposure by running a regression with the exposure as the outcome, and the confounder Z as the exposure. Then, these values Pr(X|Z) are used to calculate a propensity score and weight the data

How to calculate the weights in STATA

Calculate the weights: Logistic x i.z followed by predict p will give you P(X|Z).

Then you generate them using (tabulate weight by x. remember that p(X=0|Z=z) is 1 – p(X=1|z=z)

Finally, you use the weights in regression by using the [pw=wt] command logistic y x [pw=wt]

Stabilizations

You would always stabilize the weights to minimize outliers, where a few people tend to count disproportionately for others in the weight. You do so by using P(X=xi) in the numerator instead of 1

Marginal Structural Models vs G formula

G formula: Model the outcome (Y) , predict p, and remove the association between Z and X by generating simulated datasets

Marginal Structural Models: Model the exposure X, calculate propensity scores, weight the data to remove the association between Z and X

Marginal structural models

Create the weight in the original population via propensity scores: if X is dicothomous, you can use logistic regression / GLM with identity link for risk difference ? GLM with log link for relative risk

Propensity Score: Pr(X=1|Z1=z1, Z2=z2)

MSM Weights = 1 / Pr(X=1|Z1=z1, Z2=z2) for X=1 and 1/ Pr(X=0|Z1=z1, Z2=z2) for X = 0

Example: Let Y, X and Z be 3 dicothomous variables

Z=0

	Y=1	Y=0
X=1	600	400
X=0	200	100

P(X|notZ) = 1000/1300

P(notX|notZ) = 300/1300

Z=1

	Y=1	Y=0
X=1	20	100
X=0	90	490

P(X|Z) = 120/610

P(notX|Z) = 490/610

The 4 weights are given by X=1 and Z=1, X=0 and Z=1, X=1 and Z=0 , X=0 and Z=0. These probabilities can be obtained from the table above

Once you have the 4 probabilities, the weights are calculated by 1/P(Xi|Zi), which you use to multiply the values in each cells in the table above

G formula and Marginal structural models terminology

Parametric methods: non saturated models

Non parametric methods: saturated models (ie all possible interactions are included)

G formula methods are well suited for simple analyses and time varying confounding

MSMs are well suited for longitudinal data and loss to follow up analyses

Marginal structural model example

Cross tabulate the data (cs command) to make 2x2 tables for Z=0, Z=1 and the overall dataset(cs y x, if z==1, if z==0)

Calculate the propensity scores for the weights my modeling the exposure based on the confounder (logistic x i.z) then get the propensity score value for each observations (predict px) and generate your weight (remember that P(X=0|Z=z) = 1-p(X=1|Z=z)

Finaly you can recalculate the association using the new weights (glm y x [pw=wt] > family (binomial) link (id) for risk difference / link (log) for relative risk / link (logit) for odds ratio

Using G formula, you can do the same results via standardization by logistic y i.x##i.z then marginsr.x

Stabilizations

You would always stabilize the weights to minimize outliers, where a few people tend to count disproportionately for others in the weight. You do so by using P(X=xi) in the numerator instead of 1 (aka P(X=1)/P(X=1|Z=z) and P(X=0)/P(X=0|Z=z).

Interactions in G-formula

Example using censoring weights (C0 for censoring at time 0, and C1 for censoring at time 1)

Assign the weights in a temporal order (C0 > A0 > C1 > A1). You need to restrict the regression using if == (A=1). Remember to stabilize your weights!

Logit (PrC0=1) = B0 + B1*L0

Logit (PrA0=1) = B0 + B1*L0|C0=0 (we cannot use B2C0 because B0=1 means that there is censoring so no info on those participant)

Logit (PrC1=1) = B0 + B1*A0 + B2*L1 + B3*L0 | C0 = 0

Logit (PrA1=1) = B0 + B1*A0 + B2*L1 + B3*L0 | C0 = 0 and C1 = 0

For X=0

- Denominator: logistic x0 z0, predict p_den, pr
- Numerator: logistic x0, predict p_num, pr
- G wt = num/den if x==1
- Replace (1-num) / den if x==0

For C=0

- Denominator: logistic c0 i.x0##i.z0, predict p_den, pr
- Numerator: logistic c, predict p_num, pr
- G wt = num/den if c==1
- Replace (1-num) / (1-den) if c==0

For X=1

- Denominator: logistic x1 z0 z1 if c0==0, predict p_den, pr
- Numerator: logistic x if c0==0, predict p_num, pr
- G wt = num/den if x==1
- Replace (1-num) / den if x==0

For C=1

- Denominator: logistic c1 x1 x0 if c0 ==0, predict p_den, pr
- Numerator: logistic c1 if co==0, predict p_num, pr

- G wt = num/den if x==1
- Replace (1-num) / den if x==0

Make the final weight

- G weight = wt(xo)*wt(co)*wt(x1)*wt(c1)

The average weight should be around 1, check for outliers

Regression for binary outcomes

Model based inference: Y ~ Bernoulli(p), logit (p) = B0 + B1x + B2z (with E=p and V=p(1-p)

Modeling assumptions: X and Z vary linearily on the log odds scale, and there is no interaction between X and Z

Statistical Association: Calculate P(Y|X=1,Z=0) versus P(Y|X=0,Z=0) (which we observe)

Epidemiological measure of effect: Pr(Y|X set to 1, Z=0) versus Pr(Y|X set to 1, Z=0) in a randomized controlled trial fashion

We can relate the observed statistical association to the desired epidemiological measure of effect if exchangeability, consistency and positivity holds

Stata example

Case control function, not cs for cross sectional tabulations)

Cc y x if z ==0

Cc y x if z==1

Cc x z (if Z is a confounder, there would be an association between X and Z. if the effect measures are 1 for the RR/OR, or 0 for the RD, there is no confounding)

How to get various measures of effect for binary outcomes

- RR: use log binomial regression (log(p) instead of log(p/1-p)). This is done using the log link / Poisson regression / cox regression with a constant time variable. However, using this model can yield probabilities over 1
- RD: Linear regression. This is done using the identity link. However, using this model can yield probabilities over 1

G formula, Standardization and IPW summary

- Parametric g formula: standardization (3ple population)
- Parametric g formula: Monte Carlo Simulation (Large simulation dataset)
- Non parametric g formula: IPW (model X based on Z the use the weights to find the effect of X on Y)

The Margins command in stata uses the 3ple population and delta method for estimation

Should we standardize over non homogeneous RDs?

- Option 1: Present stratified results (based on you're a priori specifications)
- Option 2: Standardize and present your weighted average effect estimate (using the margins command)

Hypothetical population Case control studies

	Y=1	Y=0	N	Person Time (2 Year f/u, outcomes happen @ year 1)
X=1	20	80	100	180
X=0	10	60	70	130

Every case control study is nested within a larger cohort

- OR = 20x60/10x80
- RR = (20/100)/(10/70)
- Rate = (20/180) / (10/130)

For the OR, you sample from the Y=0 group but the ratio of P(Y=0|X=1) to P(Y=0|X=1) must be the same as in the total cohort to get the same OR as in the total population

	Y=1	Y=0
X=1	20	8
X=0	10	6

For the RR, you sample from the Y=0 group but the ratio of P(Y=0|X=1) to P(Y=0|X=1) must be the same as in the total cohort to get the same OR as in the total population

	Y=1	Y=0
X=1	20	10
X=0	10	7

For the Rate, you sample from the Y=0 group but the ratio of P(Y=0|X=1) to P(Y=0|X=1) must be the same as in the total cohort to get the same OR as in the total population

	Y=1	Y=0
X=1	20	18
X=0	10	13

Instrumental Variables

Uncontrolled confounding is bad in observational research, because we cannon estimate a causal effect of X on Y due to the pathway through U. We can get an unbiased effect of Z(our IV) on Y, if it meets the

following three criteria: 1) associated with the exposure 2) only associated with the outcome through the exposure 3) independent of any unmeasured confounders. In the following example, we use the randomization scheme as our IV for actually taking the drug

Type	Z=0	Z=1	% in the population
Always	X=1	X=1	Pa
Compliers	X=0	X=1	Pc
Defiers	X=1	X=0	Pd
Never	X=0	X=0	Pn

Randomization does nothing for the always user and the never users. However, based on the table we can estimate the following.

P(Y=1|X=1) = Sum of (P(Y|Z and Type) * P(Type)

= P(Y|Z,A)*P(Always) + P(Y|Z,C)*P(Compliers) + P(Y|Z,D)*P(Defiers) + P(Y|Z,N)*P(Never)

For Z=1, We have

- P(Y|X=1,A)*P(Always)
- P(Y|X=1,C)*P(Compliers)
- P(Y|X=0,D)*P(Defiers)
- P(Y|X=0,N)*P(Never)

For Z=0, We have

- P(Y|X=1,A)*P(Always)
- P(Y|X=0,C)*P(Compliers)
- P(Y|X=1,D)*P(Defiers)
- P(Y|X=0,N)*P(Never)

Since randomization does nothing for the always and never portions of the population. Let us calculate P(Y|Z) – P(Y|Not Z). The first half is part of Z=1 and the second comes from Z = 0

- [P(Y|X=1,C)*P(Compliers) + P(Y|X=0,D)*P(Defiers)] - [P(Y|X=0,C)*P(Compliers) + P(Y|X=1,D)*P(Defiers)]

The key assumption of instrumental variables is that there are no defiers in our population (P(Defiers) = 0). Our equation thus becomes

- [P(Y|X=1,C)*P(Compliers)] - [P(Y|X=0,C)*P(Compliers)] > Unknown
- P(Y|Z=1) – P(Y|Z=0) > Known

Since we assume that the always and never cancel each other out from Z=1 and Z=0, and that there are no defiers, we can get some probabilities from the following table

	Z=1	Z=0
X=1	Always or Compliers	Always or Defiers
X=0	Defiers or Never	Compliers or Nerver

- P(X=1|Z=1) = P(Always) + P(Compliers)
- P(X=1|Z=0) = P(Compliers)
- Therefore, P(Compliers) = P(X=1|Z=1) – P(X=1|Z=0)

Going back to the Instrumental Variable Equation, we now have

- RD = [P(Y|X=1,C)*P(Compliers)] - [P(Y|X=0,C)*P(Compliers)]
- RD = [P(Y|X=1,C) – P(Y|X=0,C)]*P(Compliers)
- RD = [P(Y|X=1,C) – P(Y|X=0,C)]* P(X=1|Z=1) – P(X=1|Z=0)

In Summary, we have

- RD (effect of x on y) = [P(Y|X=1,C) – P(Y|X=0,C)] = RD(effect of z on y) / P(Compliers)
- RD(effect of z on y) / P(Compliers) = P(Y|Z=1) – P(Y|Z=0) / P(X=1|Z=1) – P(X=1|Z=0)
- Remember that the effect of Z on Y can be directly be estimated in the dataset

Quantitative Bias Analyses

Sources of (systematic) Bias that can be addressed using quantitative bias analysis

- Selection Bias
- Measurement error misclassification
- Uncontrolled confounding

The direction of Bias follows a multiplicative rule with uncontrolled confounding U

U->X	U->Y	Total Bias
+	+	+
+	-	-
-	+	-
-	-	+

Measurement Error vs Misclassification

- Misclassification is a type of measurement error
- Measurement error = continuous or categorical variables
- Misclassification error = categorical variables

- Epidemiologists really only talk about misclassification
- Lots of measurement error corrections exist in the literature, though

Measures of Misclassification

- Sensitivity: Pr(X*=1|X=1)
- False Negative Probability: Pr(X*=0|X=1)
  - 1-sensitivity
- Specificity: Pr(X*=0|X=0)
- False Positive Probability: Pr(X*=1|X=0)
  - 1-specificity
- Positive Predictive Value (PPV): Pr(X=1|X*=1)
- Negative Predictive Value (NPV): Pr(X=0|X*=0)
- All of these measures can be estimated for any covariate: exposure, disease or confounder

Non/differential & Sens/Spec

- Non Differential Exposure Misclassification:
  1. Sensitivity is same for disease/non-diseased
     - Pr(X*=1|X=1,Y=1) = Pr(X*=1|X=1,Y=0)
  2. Specificity is same for disease/non-diseased
     - Pr(X*=0|X=0,Y=1) = Pr(X*=0|X=0,Y=0)
- Differential Exposure Misclassification
  1. At least one of these equalities does not hold
- Similar Equations hold for Outcome misclassification

Dependent & Independent Misclassification

- Differential misclassification: Sensitivity/specificity of exposure (disease) don't depend on true disease (exposure)
- Dependent misclassification: probability of misreporting exposure (disease) is higher if also misreport disease (exposure)
- Imagine a survey given to a large number of school students – Minnesota Student Survey
- Is it more plausible to believe that mistakes on the survey are independent of one another
- Or perhaps students who make one mistake are more likely to exaggerate everywhere

- Or perhaps they're running out of time and making 'innocent' mistakes everywhere
- The general idea behind dependent misclassification: some overarching characteristic that controls misreporting

What Do We Do About these Biases?

- Typical Approach:
  - Ignore them
- Less Common Approach:
  - Mention they are possible
- Really Uncommon Approach:
  - Qualitatively mention the possible impact of the biases
- Vanishingly Uncommon Approach:
  - Quantify the extent of bias you might have and adjust for the source of bias

Quantitative Bias Analysis

- A set of techniques to adjust the observed effect estimate for sources of error (misclassification, selection, uncontrolled confounding)
- Simple Bias Analysis: Adjust for 1 source of error
- Multiple Bias Analysis: Adjust for multiple sources of error
- Probabilistic Bias Analysis: Adjust for sources of error and account for the uncertainty in your correction

Example 1: Misclassification Bias Parameters

- Sensitivity among the Y=1 is (X*=1 | X=1, Y=1)
- Sensitivity among the Y=0 is (X*=1 | X=1, Y=0)
- Specificity among the Y=1 is (X*=0 | X=0, Y=1)
- Specificity among the Y=0 is (X*=0 | X=0, Y=0)
  - Positive Predictive Value among Y=1/Y=0
  - Negative Predictive Value among the Y=1/Y=0
  - We choose sensitivity and specificity over PPV and NPV because sensitivity and specificity is less dependent on prevalence and can be used from population to population

Simple Correction for Misclassification

1. Specify Bias Parameters
2. Use bias parameters to Impute the 2x2 table you would have observed if there had been no misclassification
3. Estimate effects of interest in the imputed table

Exposure Misclassification Process

	D=1	D=0	Total
X=1	A	B	A+B
X=0	C	D	C+D
Total	N1	N0	

se_0, se_1, sp_0, sp_1

	D=1	D=0	Total
X*=1	a	b	a+b
X*=0	c	d	c+d
Total	N1	N0	

Exposure Misclassification Correction

	D=1	D=0	Total
X=1	A	B	A+B
X=0	C	D	C+D
Total	N1	N0	

se_0, se_1, sp_0, sp_1

	D=1	D=0	Total
X*=1	a	b	a+b
X*=0	c	d	c+d
Total	N1	N0	

Exposure Misclassification Process

$$A = \frac{a - (1 - sp_1)N_1}{se_1 + sp_1 - 1}$$
$$B = \frac{b - (1 - sp_0)N_0}{se_0 + sp_0 - 1}$$
$$C = N_1 - A$$
$$D = N_0 - B$$

	D=1	D=0	Total
X*=1	a	b	a+b
X*=0	c	d	c+d
Total	N1	N0	

Exposure Misclassification Process

$$A = \frac{a - (1 - sp_1)N_1}{se_1 + sp_1 - 1}$$
$$B = \frac{b - (1 - sp_0)N_0}{se_0 + sp_0 - 1}$$
$$C = N_1 - A$$
$$D = N_0 - B$$

	D=1	D=0	Total
X*=1	a	b	a+b
X*=0	c	d	c+d
Total	N1	N0	

Disease Misclassification

- This is a direct analog of exposure misclassification
- You just turn the 2x2 table on its side and everything else is the same
- Specify sensitivity/specificity conditional on exposure status

Disease Misclassification Correction

	D=1	D=0	Total
X=1	A	B	A+B
X=0	C	D	C+D
Total	N1	N0	

se_0, se_1, sp_0, sp_1

	D*=1	D*=0	Total
X=1	a	b	M1
X=0	c	d	M0
Total	a+c	b+d	

Disease Misclassification Process

$$A = \frac{a - (1 - sp_1)M_1}{se_1 + sp_1 - 1}$$

$$B = M_1 - A$$

$$C = \frac{c - (1 - sp_0)M_0}{se_0 + sp_0 - 1}$$

$$D = M_0 - C$$

	D*=1	D*=0	Total
X=1	a	b	M1
X=0	c	d	M0
Total	a+c	b+d	

Selection Bias

- It was the best of times, it was the worst of times
- Selection bias adjustment is computationally very simple
- The information you need is almost never actually available to you!
- Remember we've actually done this already using inverse probability weights

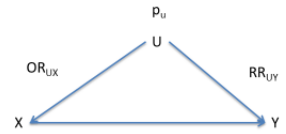
Bias Parameters: Selection Bias

- If S=1 is selection into the study, bias parameters are:
  - Pr(S=1|E=1,D=1)
  - Pr(S=1|E=0,D=1)
  - Pr(S=1|E=1,D=0)
  - Pr(S=1|E=0,D=0)
- You just have to multiply each respective cell by the probability to go from the truth (unobserved) to the study population (observed). Then, to get back to your unbiased sample, you would divide each cell from the observed population by the corresponding probability above.

Confounding

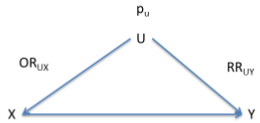
- Enormous energy has been spent developing methods to control for confounding
  - Measured – regression, standardization, g-methods, MH, etc
  - Unmeasured – matched twin studies
- But what about the common case we worry about of unmeasured confounding?

Limits of Confounding



- RR_{crude} is the crude RR
- RR_{adj} is the adjusted RR
- $RR_{conf}=RR_{crude}/RR_{adj}$
- OR_{UX} is confounder – exposure association
- RR_{UY} is confounder – outcome association
- P_u is prevalence of confounder in the unexposed

Limits of Confounding



- RR_{conf} is bounded by (the minimum of)
  - OR_{UX}
  - RR_{UX}
  - $1/p_u$
- There are other conditions as well

Confounding Bias Parameters

- If we want to adjust for confounding bias, we need bias parameters
- These will be specified differently depending on the effect measure of interest but always deal with
  1. Confounder-disease association
  2. Confounder-exposure association
  3. Prevalence of the confounder

General approach: Relative Risk Confounding Adjustment

$$RR_{adj} = RR_{obs}\frac{RR_{UD}p_0 + (1-p_0)}{RR_{UD}p_1 + (1-p_1)}$$

- Compute observed RR
- Specify confounder-disease association
- Specify p_0 and p_1:prevalence of confounder in unexposed and exposed, respectively
- Plug in estimates and calculated RR_{adj}
- What happened to the confounder-exposure association? The ratio of P0 / P1 will give us the estimate of the confounder exposure association (it is a relative risk)

Relative Risk Adjustment

- It is often more instructive to impute the table we would have observed, just like with selection bias and misclassification examples
- In this case, we need to impute tables within stratum of the confounder
  - 2x2x2 table
- Homogeneity assumption is often made

Imputing Data: Uncontrolled Confounding, RR

	Total		C_1		C_0	
	E_1	E_0	E_1	E_0	E_1	E_0
D=1	a	b	A_1	B_1	A_0	B_0
D=0	c	d	C_1	D_1	C_0	D_0
	m	n	M_1	N_1	M_0	N_0

Lower Case = observed
Upper Case = imputed
$RR_{UD} = (B_1/N_1)/(B_0/N_0)$
$P_0=N_1/n$
$P_1=M_1/m$

Imputing Data: Uncontrolled Confounding, RR

	Total		C_1		C_0	
	E_1	E_0	E_1	E_0	E_1	E_0
D=1	a	b	A_1	B_1	A_0	B_0
D=0	c	d	C_1	D_1	C_0	D_0
	m	n	M_1	N_1	M_0	N_0

$$B_1 = \frac{RR_{UD} N_1 b}{RR_{UD} N_1 + n - N_1}$$
$$B_0 = b - B_1$$
$$A_1 = \frac{RR_{UD} M_1 a}{RR_{UD} M_1 + m - M_1}$$
$$A_0 = a - A_1$$
$$M_1 = mp_1$$
$$M_0 = m - M_1$$
$$N_1 = np_0$$
$$N_0 = n - N_1$$

Imputing Data: Uncontrolled Confounding, OR

	Total		C_1		C_0	
	E_1	E_0	E_1	E_0	E_1	E_0
D=1	a	b	A_1	B_1	A_0	B_0
D=0	c	d	C_1	D_1	C_0	D_0
	m	n	M_1	N_1	M_0	N_0

Lower Case = observed
Upper Case = imputed
$OR_{UD} = (B_1/D_1)/(B_0/D_0)$
$P_0=D_1/d$
$P_1=C_1/d$

Imputing Data: Uncontrolled Confounding, RD

	Total		C_1		C_0	
	E_1	E_0	E_1	E_0	E_1	E_0
D=1	a	b	A_1	B_1	A_0	B_0
D=0	c	d	C_1	D_1	C_0	D_0
	m	n	M_1	N_1	M_0	N_0

$$B_1 = \frac{RD_{UD} n(n - N_1) + bN_1}{n}$$
$$B_0 = b - B_1$$
$$A_1 = \frac{RD_{UD} m(m - M_1) + aM_1}{m}$$
$$A_0 = a - A_1$$
$$M_1 = mp_1$$
$$M_0 = m - M_1$$
$$N_1 = np_0$$
$$N_0 = n - N_1$$

Multiple Bias Modeling: Example

- Case Control Study of Alcohol and OFC
- Cases: infants born with OFC
- Controls: infants sampled from the general population
- ETOH ascertained by maternal self-report following delivery
- Errors:
    1. Misclassification of exposure
    2. Confounding by smoking (not measured)

Multiple Bias Modeling

- Often we will have multiple biases
- We want to adjust for each of them
- Adjusting for each separately doesn't answer the question
    1. One answer adjusted for uncontrolled confounding
    2. Another adjusted for misclassification
- Instead, we need to adjust in sequence
- Impute the table without misclassification
- Then use that table to impute the table stratified by the confounder

Order is Important

- Adjust for 1) confounding then 2) misclassification
- Adjust for 1) misclassification then 2) confounding
- These may give different answers!
- There's no single answer for which order to do the corrections
- In general, you want to <u>do the correction in the reverse order they occurred</u>

Error Hierarchy

1. Confounding – population level
2. Selection – who gets into the study
3. Misclassification – measurement within the study

Corrections will often go:

1. Misclassification

2. Selection
3. Confounding

Meta Analysis

Meta Analysis

- The quantitative part
  - A good meta analysis will largely mirror a good "regular" analysis
  - Systematic review = data collection
    - Define outcome
    - Define exposure
    - etc
- Meta analysis
  - Combining effects across studies (strata)
  - Methods to determine whether we should aggregate results
  - Methods to aggregate those results

Systematic Review

- What is your research question?
- Define your outcome – precisely
- Define your exposure – precisely
- What effect?
  - OR? RR? HR? RD?

Data Collection

- Abstract relevant data from papers
- Good to have multiple reviewers, if possible
- What to abstract?
  - Effect
  - 95% CIs or Std Errs
  - Confounders that were accounted for (matching, regression, etc)
  - Important characteristics that could influence the effect of interest
    - Year, Country of Origin, Demographics

Analysis: Big Picture

- Each study is a separate stratum of data
  - Akin to estimating the effect of TV viewing on mortality among underweight, normal, over, obese
    - Should we aggregate those estimates or not?
- Meta Analysis
  - Should we aggregate these different study estimates or not?

Lumpers and Splitters

- Everyone falls somewhere on the lumper-splitter spectrum
- Lumpers like to aggregate all effects
- Splitters like to present stratum specific effects
- Epidemiologists tend to be lumpers
- Psychologists seem to be splitters
- Our tendency to lump isn't always the best approach in meta analysis
- (courtesy Charlie Poole)

Meta Statistics

- Say θ_i is the effect estimate from the ith study
  - RD_i perhaps (though that would be less common)
  - RR_i, HR_i, OR_i, will all be more common
    - Transform relative measures to the log scale
    - $\theta_i=\log(RR_i)$
- We also need to know the variance of the effect estimate
  - $s_i=\text{var}(\theta_i)$
- How do I get s_i if I only have 95%CI's?

Fixed Effects Meta Analysis

- Assume there is 1 true effect: Ψ
- Each of the θ_i are an estimate of Ψ

  E(θ)=Ψ

  !!!!Key Assumption!!!!
- If we believe all the θ_i are estimating Ψ, we just need to figure out how to combine them
- We have lots of ways of doing this already!
  - Regression (Likelihood methods)
  - Mantel Haenzel
  - Weighting

Combining Study Specific Estimates

- Likelihood: requires more assumptions (Normality) and is quite similar to weighting anyhow
- M-H: no reason not to, but no one does
- Weighting: Simple, intuitive (also matches up to likelihood in a Normal setting)

- Combine study specific effects so our summary estimate is a weighted average
  - What weights?

Fixed Effects – inverse variance weights

- Each study: θ_i , s_i, and $w_i=1/s_i$
- Meta Estimates:

$$\theta = \frac{\sum_{i=1}^{n} w_i \theta_i}{\sum_{i=1}^{n} w_i} \qquad V(\theta) = \frac{1}{\sum_{i=1}^{n} w_i}$$

Example – TV use and Mortality

- Grontved and Hu. JAMA 2011

Study	RR	Lower 95%CI	Upper 95%CI
Dunstan	1.17	1.00	1.37
Stamatakis	1.14	1.06	1.23
Wijndaele	1.10	1.02	1.19

- What is the fixed effects summary estimate and 95%CI?
- Is the summary estimate statistically different than zero? How do we test that?

Homogeneity

- I told you this relied on an assumption of homogeneity
- How can we test that?

$$Q = \sum_{i=1}^{n} w_i \left(\theta_i - \theta\right)^2$$

- Chi-square, with df=n-1
- Very low powered test!

I

$$I^2 = \frac{Q - df}{Q}$$

- I^2=0 if Q<df
- 0: No important heterogeneity
- 100: considerable heterogeneity

Meta Analysis in STATA

```
rename upper up

rename low low


*Use the code below to log transform the abstracted estimates and 95% bounds

*from your data

g theta =log(rr)

g up_theta=log(up)

g low_theta=log(low)


*Manually determine the std from the 95% CI values

g var=((up_theta-low_theta)/(2*1.96))^2


*create your weight values to estimate your composite estimate

g wt=1/var

g wt_theta=log(rr)*wt


sum wt

local sumwt=r(sum)

local v_psi=1/`sumwt'

sum wt_theta

local sumwt_theta=r(sum)


di `sumwt_theta'

di `sumwt'

di exp(`sumwt_theta'/`sumwt')

di exp(`sumwt_theta'/`sumwt'-1.96*sqrt(`v_psi'))

di exp(`sumwt_theta'/`sumwt'+1.96*sqrt(`v_psi'))
```

```
*This is how you can automate a meta analysis in STATA

metan theta low_theta up_theta, fixedi  eform


* meta analysis day 2

use "/Users/presentation/Dropbox/Classes/8343/PuBH 8343 fall 2016/week 11/slides/metadata2.dta", clear


g logrr=log(rr)

g logup=log(upper)

g loglow=log(lower)

metan logrr loglow logup, fixedi nograph eform

metan logrr loglow logup, randomi nograph eform


*BCG EXample

use "/Users/presentation/Dropbox/Classes/8343/PuBH 8343 fall 2016/week 11/slides/bcgtrial.dta", clear


g control1=tot1-cases1

g control0=tot0-cases0


metan cases1 control1 cases0 control0, rr fixedi nograph

metan cases1 control1 cases0 control0, rr randomi nograph

metan cases1 control1 cases0 control0, rd randomi nograph


*publication bias

metan cases1 control1 cases0 control0, rr fixedi nograph

g logrr=log(_ES)

g se=_selogES

g wt=_WT
```

```
metafunnel logrr se
metabias logrr se, egger

*double check with alternate model: E(theta)=b0+b1*se
regress logrr se [pw=wt]
metabias logrr se, begg
metan cases1 control1 cases0 control0 if se<0.4, rr fixedi nograph

****meta regression
*fixed effects
vwls logrr lat, sd(se)
*RE regression
metareg logrr lat, wsse(se)

metareg logrr lat start, wsse(se)
```

Bootstrap

The goal is to estimate the distribution of an estimator by iteratively sapling from a distribution with replacement > calculate the estimate for each iteration > calculate the variance for the distribution of all the estimates you iterated.

Variance = Sum (Xi – Xbar)^2 / N or Sum (Xi – Xbar)^2 / N-1 for the unbiased version

Standard Deviation = sqrt (Variance / N)

Standard Error (For a single statistic, since the population variance is not actually known)

Let us use an example of boot strapping, based on Monte Carlo Integration

***montecarlo Integration**

Clear

*200 obs will be our sample, and 1000 will be the number of boostrap iteration that we want to do

set obs 1200

*The following means that the last 200 observation will have a bstrap value = to 0, based on Boolean logic which implies that for every n under 1000, bstrap ==1 else bstrap == 0.

g bstrap=_n<=1000

*This will generate the blank columns to hold our observations

g theta=.

g bmi=.

*The first thing we do is create our original dataset for our sample observations (which we would have collected from real data in other cases. Here, we randomly pick them from a random normal distribution. Then we sum them up (the 200 observation where bstrap = 0, and compute a mean, which is stored in n=1. This is one iteration of the procedure. If we had real data, we would use the values calculated from it in the mean and SD, and we would have to make sure that the n(real data) = n(bootstrap data) in order not to bias the variance.

replace bmi=rnormal(27,6) if bstrap==0

qui sum bmi

replace theta=r(mean) in `i'

*now we have to create this for all 1000 bootstrap iterations, which we do by using a loop function

foreach i of numlist 1/1000{

replace bmi=rnormal(27,6) if bstrap==0

qui sum bmi

replace theta=r(mean) in `i'

*** parametric bootstrap**

* when we are using the parametric bootstrap we are directly specifying which distribution we are using. Here, let us say our data from our 2x2 looked as follows

	Cases	Controls
Exposed	9	33
Unexposed	5	38

*First, we would use the csi command to enter them into stata by csi A B C D

csi 9 5 33 38

*Then we would pick the number of bootstrap iterations we wanted

clear

set obs 4000

*The next step would be to generate the parameters of our sampling distribution, using the data we collected.

g y1=rbinomial(42,9/42)

g y0=rbinomial(43,5/43)

*Then, for all 4000 observations, we could iteratively sample from the Y1 distribution of cases, and Y0 distribution of controls to calculate and store an estimate.

g p1hat=y1/42

g p0hat=y0/43

g rr=p1hat/p0hat

*Once we have our bootstrapped sample, we can calculate the standard deviation from our bootstrapped iterations, then use the original RR (from our 2x2 table, and the estimated bootstrapped SD to create the new CI.

sum rr

local rrsd=r(sd)

di `rrsd'

di 1.84-1.96*`rrsd'

di 1.84+1.96*`rrsd'

```
g lrr=log(rr)
sum lrr
local lrrsd=r(sd)
di `lrrsd'
di exp(log(1.84)-1.96*`lrrsd')
di exp(log(1.84)+1.96*`lrrsd')
```

Bayesian Analysis (and Penalized regression)

Probability

$$\Pr(Heads) = 0.5$$

- What is this probability, as a frequentist?
  - Likelihood of having the outcome as heads, if we could repeat the same experiment a large number of times
- What is this probability as a Bayesian?
  - Likelihood of having the outcome as heads, given prior knowledge or content area expertise

Frequentist Statistics

$$y_i \sim N(\theta, \tau^2)$$

- Frequentist statistics is everything you've learned so far: How does it work?
  - Theta is fixed (truth) but we do not get to estimate it directly. We find thetahat (estimator, usually our mean), from collected observation of a random sample of the population.
  - We specify a distribution a priori that our data follows (normal, binomial, exponential) and find the parameter that maximizes the likelihood of our outcome given the data we observe
  - Each F(yi) is given a density, summed up overall to a distribution ln[F(yi)] from i=1 to n, which is the log likelihood, and theta hat is the parameter that maximizes the density distribution based on the data we observe.

95% Confidence Intervals

- What are they?
  - If we were to conduct the experiment an infinite number of time, 95% of our 95% CI would contain the true value of theta
    - In publication, we can either interpret them as a hypothesis test (wheter the 95% CI contain the null as an indication of significance) -> Very limited
    - The 95% CI is the range of estimates that are roughly compatible with the data -> More realistic
    - The ratio of the upper bound / lower bound is an indication of the precision of our measure (smaller is better)
- Are they frequentist?
  - No, because we only get to compute them once whereas the frequentist approach requires an infinite number of samples.

Frequentist Statistics is Indispensable

- What types of questions does frequentist statistics answer?
  - It is very useful to tell us how estimators will act like in the long run

Bayesian Statistics

- What is it?
  - We specify a prior (based on metaanalysis of the existing data, or expert knowledge)
  - We collect our data in our experiment (Using MLE just like in frequentist statistics)
  - We get a posterior estimate, which is an inverse variance weighted average of our prior with the data collected
  - It goes from $Y|\theta > \theta > \theta|Y$ (which is Bayes' Theorem)
- What types of questions does it answer?
  - Determines if our study is more informative, while being more efficient by incorporating prior knowledge

Schools of Thought

- Subjective Bayesians: Bayesians who explicitly incorporate prior knowledge into a model
- Objective Bayesians: Try to get a posterior distribution without having to incorporate prior knowledge
- Pragmatic Bayesians: Screw philosophy, MCMC is super awesome

Bayesian Critiques

- Science should be objective!!
  - That'd be great, but its not the world we live in
  - Frequentists make objective decisions all the time
    - What variables go in the model
    - What model is chosen
    - What results to report
  - Objective inference is a good ideal, but its not a possible one
  - Bayesian inference is (when done well) transparent about many of these modeling choices
    - The subjectivity is transparent, rather than hidden

Bayesian Critiques

- Your prior is arbitrary!
  - My prior represents my belief in the magnitude and uncertainty around the parameter of interest, discarding my data
  - A mathematical representation of what I know prior to collecting data
  - Based on previous literature (like a meta-analysis)
  - Or based on expert opinion

Betting

- Greenland (2006) describes priors in terms of bets placed on the effect of interest
- The median of the prior is the value for which you would give even odds to the truth being on either side
  - $Pr(RR<RR_{median})=Pr(RR>RR_{median})$
- 95% limits give 95:5=19:1 odds on $Pr(RR_{lower}<RR<RR_{upper})$
- Your prior should incorporate all available information to maximize the chance that you win your bet

Bayesian Components, Briefly

- Prior $f(\theta)$
- Likelihood $f(Data|\theta)$
- Posterior $f(\theta|Data)$

Likelihood

$$f(Data|\theta)$$

- Common to frequentists and Bayesians
- Consider binary outcome

$$y \sim Binomial(N, p)$$

$$f(Y|p)=\binom{N}{y}p^{y}(1-p)^{N-y}$$

Posterior Distribution

$$f(\theta|Data)$$

- A mathematical expression of our subjective belief about the parameter, having seen the data
- Computing the EXACT posterior distribution is very hard, typically
  - So is computing the distribution of $\hat{\theta}$ in frequentist statistics
- Markov Chain Montecarlo Techniques allow us to draw random samples from the posterior distribution, even if we cannot directly specify it

Posterior Distribution

- Another approach is to assume the posterior distribution is asymptotically normal
  - The central limit theorem holds for the posterior distribution as well, because as the sample data collected gets larger, the variance of the data gets smaller, and the inverse variance weight that will be used for the final posterior estimates is more important.
- This makes life so much easier, as we'll see

A Simple Bayesian Example

$$Y \sim N(\theta, \sigma^2)$$
$$\theta \sim N(\mu, \tau^2)$$
$$\theta | Y \sim N\left(\frac{1/\sigma^2 Y + 1/\tau^2 \mu}{1/\sigma^2 + 1/\tau^2}, \frac{1}{1/\sigma^2 + 1/\tau^2} \right)$$

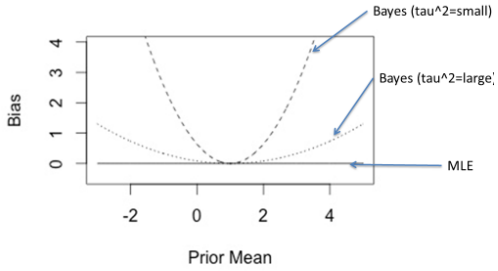
- What's the prior?
  - It is the normal distribution specified for Theta
- What's the Likelihood?
  - It is the normal distribution specified for Y
- What's the posterior?
  - It is the distribution specified for θ|Y
- Explain the posterior parameters in words
  - The Mean parameter is the weighted average of the likelihood and prior means, weighted according to the inverse variance

A Simple Bayesian Example

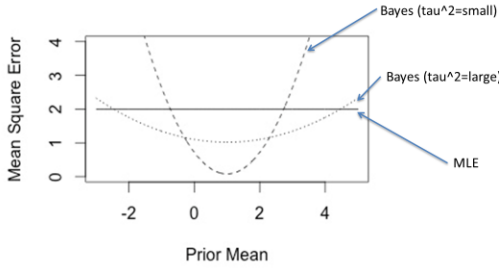
$$Y \sim N(\theta, \sigma^2)$$
$$\theta \sim N(\mu, \tau^2)$$
$$\theta | Y \sim N\left(\frac{1/\sigma^2 Y + 1/\tau^2 \mu}{1/\sigma^2 + 1/\tau^2}, \frac{1}{1/\sigma^2 + 1/\tau^2} \right)$$

- What does this remind you of?
  - This is exactly like the variance weighted approach for meta analyses!

Bayesian Bias



Bayesian MSE



Shrinkage

- Bayesian estimators in common models have a natural shrinkage property
- The frequentist estimate is "shrunk" toward the prior mean
- This induces bias
- It can also reduce mean squared error
- Bayesian estimators CAN be closer to the truth on average
  - CAN is the operative word
  - This property depends on choosing a decent prior

Shrinkage in Practice

- Why might we use shrinkage?
- Sparse data is a really great reason!
  - Greenland and Poole (1994) used shrinkage techniques to estimate effects of multiple environmental hazards
- Shrink to the null to limit false positives
- Highly correlated data, shrinkage can stabilize estimation considerable
- Regression models with lots of covariates
- There's a wealth of theoretical data and simulation studies to show that shrinkage performs very well in these situations
- When might shrinkage not be important?

Shrinkage in Regression

- Shrinkage via Bayesian techniques has a long history in statistics
  - Even frequentists love it (they just use a different name for these techniques)
- Ridge Regression is an old regression technique used to stabilize regression estimates.
- These techniques place a 'penalty' on the likelihood function
- The 'penalty' is literally added to the end of the log-likelihood

Penalized (Ridge) Model

$$\propto \sum (y_i - \beta_0)^2$$
$$\propto \sum (y_i - \beta_0)^2 + \lambda\beta_0^2$$

$$y_i \sim N(\beta_0, \sigma^2)$$
$$\beta_0 \sim N(0, 1/\lambda)$$

Summary

- Bayesian models induce shrinkage
  - The MLE estimate is pulled toward the prior mean
- When the prior mean is zero, these align with a class of penalized regression techniques in the frequentist literature
- Shrinkage can dramatically improve MSE when you have:
  - Small data set
  - Highly correlated predictors
  - Lots of predictors
  - Sparse data, generally

Bayesian Analysis

- Usually, Bayesian analysis proceeds via Markov chain Monte Carlo techniques
  - A way to draw random samples from a posterior distribution
  - Even if you can't write down a 'closed form solution' for the distribution
  - Even if the distribution doesn't have a 'name'
  - Even if it is uuuuuugly
- The drawback:
  - MCMC takes a long time
  - Requires special skills to make sure it worked OK

Approximate Bayesian Analysis

- Pretend the posterior distribution is normally distributed
  - Akin to frequentist maximum likelihood methods
  - In fact, we'll use maximum likelihood algorithms to estimate the Bayesian models a little later
- There are methods to improve the normal approximation
- These approximations are very accurate
  - When they become inaccurate, you have small sample sizes and the inaccuracy of the distribution will be swamped by the uncertainty in the data anyway

Data (likelihood)

	D=1	D=0
E=1	A1	B1
E=0	C1	D1

Prior

	D=1	D=0
E=1	A2	B2
E=0	C2	D2

$$RR_{data} = \frac{A1/(A1+B1)}{C1/(C1+D1)}$$

$$V[\log(RR_{data})] = \frac{1}{A1} - \frac{1}{A1+B1} + \frac{1}{C1} - \frac{1}{C1+D1}$$

$$RR_{prior} = \frac{A2/(A2+B2)}{C2/(C2+D2)}$$

$$V[\log(RR_{prior})] = \frac{1}{A2} - \frac{1}{A2+B2} + \frac{1}{C2} - \frac{1}{C2+D2}$$

$$\log(RR_{post}) = \frac{\log(RR_{data})/V[\log(RR_{data})] + \log(RR_{prior})/V[\log(RR_{prior})]}{1/V[\log(RR_{data})] + 1/V[\log(RR_{prior})]}$$

$$V[\log(RR_{post})] = \frac{1}{1/V[\log(RR_{data})] + 1/V[\log(RR_{prior})]}$$

Take Home Message

- This type of Bayesian analysis is akin to a (fixed effect) meta analysis (with inverse variance weights)
- The prior is study 1
- The data is study 2
- The meta estimate is the posterior
- Treat it like a meta analysis!
- Check to see if the prior and data are compatible
  - What do we do if they aren't?

- How do we come up with RR_{prior} and $V[\log(RR_{prior})]$?
- Choose a RR_{prior}
- Choose CI for RR_{prior}
  - (L - U)

Prior

	D=1	D=0
E=1	A2	B2
E=0	C2	D2

$$RR_{prior} = \frac{A2/(A2+B2)}{C2/(C2+D2)}$$

$$V[\log(RR_{prior})] = \frac{1}{A2} - \frac{1}{A2+B2} + \frac{1}{C2} - \frac{1}{C2+D2}$$

$$se[\log(RR_{prior})] = \frac{\log(U) - \log(L)}{2 \times 1.96}$$

Prior Data

- We've specified a prior based on prior knowledge
  - Log(RR) is the best guess before we analyse the data
  - V(log(RR)) comes from specifying prior CIs
- What if we translate this prior information into data?
- Fabricate a dataset that, when analyzed, will return the RR and CIs we specify

$$RR_{prior} = \frac{A/(N1)}{C/(N0)}$$

$$V[\log(RR_{prior})] = \frac{1}{A} - \frac{1}{N1} + \frac{1}{C} - \frac{1}{N0}$$

Prior

	D=1	D=0	Total
E=1	A	B	N1
E=0	C	D	N0

- How do we translate log(RR)=0, v(log(RR))=0.25 into data?
- Many tables will yield this answer, which do we choose?
- Step 1: Assume A = C
  - This helps with asymptotic normality
- Step 2: Assume N1 and N0 are large
  - Simplify Variance formula

$$RR = (A/N1)/(C/N0)$$

$$RR = N0/N1$$

$$RR = N0/10^5$$

$$V(\log(RR)) = \frac{1}{A} + \frac{1}{C}$$

$$V(\log(RR)) = 2/A$$

$$Log(RR) = 0 \rightarrow RR = 1$$

$$V(\log(RR)) = 0.25$$

Prior

	D=1	D=0	Total
E=1	A	B	N1
E=0	C	D	N0

$$RR = N0/N1$$

$$1 = N0/10^5$$

$$N0 = 10^5$$

$$0.25 = \frac{2}{A}$$

$$A = 2/0.25 = 8$$

$$RR = (A/N1)/(C/N0)$$
$$RR = N0/N1$$
$$RR = N0/10^5$$
$$V(\log(RR)) = \frac{1}{A} + \frac{1}{C}$$
$$V(\log(RR)) = 2/A$$

Prior

	D=1	D=0	Total
E=1	8	99,992	10^5
E=0	8	99,992	10^5

Check this in Stata

Combining Data & Prior

	D=1	D=0	Total
E=1	7	91	98
E=0	4	130	134

	D=1	D=0	Total
E=1	8	99,992	10^5
E=0	8	99,992	10^5

Treat these as separate strata in a dataset

Bayes Analysis Part II

	Disease	No Disease	
Exposed	A	B	N1
Unexposed	C	D	N0

Assumptions for making a prior dataset in STATA

- A=C
- Set N0 = 10^5 (need a large sample size)
- For the prior, set RR = 1, Standard Error = 0.25

Since RR = (A/N1) / (C/N0) = 1, and A=C, then RR=N0/N1

- RR=N0/N1=1, so N0 = N1 = 10^5
- Therefore, B = N1-A and D=N0-C

Finally, the standard deviation of the relative risk is 1/A – 1/N1 + 1/C – 1/N0

- Given that N1 = N0 = 10^5, 1/N1 = 1/N0 ~ 0
- Thus, V(Log(RR)) = 1/A + 1/C = 2/A

Data Entry for the Data Editor in STATA

	Prior	Disease	Exposure	2x2 table cell	Count
Strata for Prior	1	1	1	A	8
Strata for Prior	1	1	0	C	8
Strata for Prior	1	0	1	B	99992
Strata for Prior	1	0	0	D	99992
Strata by your Data	0	1	1	A	7
Strata by your Data	0	1	0	C	4
Strata by your Data	0	0	1	B	91
Strata by your Data	0	0	0	D	130

How to make the full dataset from the collapsed 2x2 tables above:

Expand count will make the number of rows matched with the value in the cell, while duplicating the values for all other columns.

Once you have created the data, you should double check that the prior and full data estimates match your specifications

- CS disease exposure if prior ==1 (for your artificial data)
- CS disease exposure if prior==0 (for your real data)
- This is the same as CSI ABCD for each strata (gives you a RD, RR and OR if specified)

Bayesian Approach 1: Mantel Haenzel stratification

- Treat the prior as a strata for confounding: CS disease exposure, by (prior)
- The adjusted MH is the combined/adjusted estimate, aka our posterior distribution. The MH test of homogeneity tells us if our prior and real data are too far apart to be combined.

Bayesian Approach 2: Regression Approach

- First, look at your stratum specific estimates to check your data: Need log link for RR, and eform to exponentiate the value
    - Glm disease exposure if prior ==1, link(log) eform
    - Glm disease exposure if prior ==0, link(log) eform

- Then, include prior as a covariate. Adjusting for the prior gives you the same estimate as the MH approach, while giving the ability to adjust for covariates.
    - Glm disease exposure prior, link(log) eform

The constant/prior from the Bayesian model are not interpretable, because they are arbitrary and set up by the investigator

The variance of the prior does not depend on the size of n chosen, because the standard deviation of the relative risk is 1/A – 1/N1 + 1/C – 1/N0, which becomes 1/A + 1/C when N1 and N0 is large

How to integrate confounders into the bayesian analysis

- First, you have to make a variable for your confounder
- Generate confounder = rbinom(1,0.5) where 1 is the value of a success, and 0.5 is the probability of success
- Replace confounder=0 if prior =1. You need to a constant value for your confounder (doesn't matter which) for your confounder in the prior dataset so that there is no association between the confounder and the disease in the dataset.
- If you want to build a prior for your confounder, the value for your exposure in the prior strata where you have a prior for your confounder must be constant, so that there is no association between exposure and disease in the prior for your confounder.
- Remember to create a new count variable to expand on the prior for the confounder, and setting all the previous rows count to 1 for other rows (replace count2=1 if count2=.) so that they do not get expanded twice.

Method 3: Bayes regression using the metropolis hastings algorithm

The metropolis hastings algorithm is an application of the metropolis algorithm, which falls under the monte carlo markov chain umbrella. What it does is that you have an unknown distribution centered around θ, and you randomly pick a value for θ1. You generate random noise around θ1, then pick a value for θ2. You will then calculate the likelihood for both θ1 and θ2 and pick whichever fits the data for θ best. You keep repeating this loop until you get enough samples (thousands) to estimate θ. In practice, you use a frequentist approach to get an MLE, which is a good starting point to estimate θ

bayesmh d e if prior==0,

the Stata command is bayesmh (bayes metropolis hastings). You set if prior ==0 because you do not want to include the simulated data for your prior.

likelihood(logit)

We are using the logit model to get Odds Ratio

prior({d:_cons}, normal(0, 10000))

We have to specify a prior for every RV, so for the constant, we have a normal distribution of mean 0 and SD 10000: this is super non specific, which we call an uninformed prior.

prior({d:e}, normal(0,.25)) mcmcsize(100000)

we specify a prior the exposure and disease association: here we use a normal distribution, even if we want odds ratio, because the odds ratio are normal on the log odds scale (thus Log(1) = 0 which is the null)

burnin(1000)

we set a burnin of 1000 so our first 1000 iteratations do not count towards the final estimation.

saving(mcmc.dta, replace)

The code below allows us to directly exponentiate the estimates to get the odds ratio. It is better to use the median instead of the means (OR and RR are often right skewed) and you can use HPD confidence intervals for your credible interval bounds.

bayesstats summary (OR: exp({d:e}))

bayesgraph diagnostics {d:e}

bayesgraph diagnostics {d:_cons}

bayesgraph matrix _all

STATA Code Part 1

```
G_Methods including Standardization G formula Monte Carlo Simulation and
Inverse Probability Weighting or Marginal Structural Modeling

***** Standardization

*Standardization = non parametric g-formula

*Example of standardization using margins in STATA
*The example is the effect of tolbutamide on all cause mortality
stratified by age<55 vs age>=55

use "/Users/maclehose/Dropbox/Classes/8343/pubh 8300 - 2015/Week 2 -
Binary Data/Programs&Data/tolbutamide.dta", clear

*Always examine your data first by using 2x2 tables if possible. You can
use the cs command  to do so

* Crude/collapsed table
cs dead tolb

* Stratified table on Z
cs dead tolb if age==0
cs dead tolb if age==1

* Mantel-Haensel adjusted RR
cs dead tolb, by(age)

*Standardization

*create weights for standardizing to the total population
*these weights are the count for the total number of people in age=1 and
age=0
g stand=183 if age==1
replace stand=226 if age==0

* use CS in stata to compute the standardized (to the total population)
RR
cs dead tolb, by(age) standard(stand)
cs dead tolb, by(age) standard(stand) rd

*G formula method 1
*Replicate your dataset 2 times
*The first replicate is the actual data; 2nd has x=1; 3rd has x=0
*Regress Y as a function of X, Z in first dataset
*Predict Y in 2nd dataset and 3rd dataset
*Compute contrast in mean Y in 2nd and 3rd datasets

**********************************
*
* Example of G-formula
* Method 1: Replicating the dataset
*
```

```
******************

*make 3 copies of this data in the same dataset
g f=3
expand f
*tag the data so we know have flag=1 for the original data, flag=2,
flag=3 for the created observations
bysort id: g flag=_n
tab flag
*remove the outcome from datasets 2&3 so we don't accidentally analyze
them
replace dead=. if flag>1
*data with flag=2 is our hypothetical X=1 intervention
replace tolb=1 if flag==2
*data with flag=3 is our hypothetical X=0 intervention
replace tolb=0 if flag==3

*Predict Y in 2nd dataset and 3rd dataset: the local values will not be
displayed in STATA
logistic dead i.age##i.tolb
predict p, pr
qui sum p if flag==2
local y1=r(mean)
qui sum p if flag==3
local y0 =r(mean)
*Compute contrast in mean Y in 2nd and 3rd datasets
di "Standardized RD= "`y1'-`y0'

*I prefer to do it in a way were I can see the values saved in my
dataset. Remember that you need to include interactions to have a
saturated model
logistic dead i.age##i.tolb
predict p, pr
sum p if flag==2
g y1=r(mean)
sum p if flag==3
g y0 =r(mean)
g srd=y1-y0

***************************
*
*Method 2: Monte Carlo
*
*****************************
use "/Users/maclehose/Dropbox/Classes/8343/pubh 8300 - 2015/Week 2 -
Binary Data/Programs&Data/tolbutamide.dta", clear

*get the distribution of age in the total population
sum age
local pr_age=r(mean)

*fit a logistic model to get the Pr(Y|age, tolb)
logistic dead i.age i.tolb
*save coefficient matrix
```

```
matrix b=e(b)
*extract coefficients
local b0=b[1,5]
local bage=b[1,2]
local btolb=b[1,4]
di `b0'

*create a dataset for tolb=1
clear
set obs 1000000
g id=_n
*create an age
g age=rbinomial(1,`pr_age')
*intervene to expose them
g tolb=1
*predict the outcome
g pr_dead=invlogit(`b0'+`bage'*age+`btolb'*tolb)
g dead=rbinomial(1,pr_dead)
sum dead
*save the prevalence of death in this group
local dead1=r(mean)

*repeat for tolb=0
clear
set obs 1000000
g id=_n
g age=rbinomial(1,`pr_age')
g tolb=0
g pr_dead=invlogit(`b0'+`bage'*age+`btolb'*tolb)
g dead=rbinomial(1,pr_dead)
sum dead
local dead0=r(mean)

di `dead1'-`dead0'

** again, you can use the "g" command instead of the "local" command
above to get permanent values saved to your dataset

*************
*
*     Example 3: Using MSM
*
*           Simple toy example Y=outcome X=exposure Z=confounder
*
***************

*Step 1: Examine the simple crosstabs
cs y x
cs y x if z==0
cs y x if z==1

*MSM. Marginal Structural Models = IPW (inverse probability weighting)
*Specify a logistic exposure model
logistic x i.z
```

```
predict px
*Calcualte the MSM weights
g wt=1/px if x==1
replace wt=1/(1-px) if x==0

*Now generate the risk difference (this is the glm equivalent to logistic
regression)
glm y x [pw=wt], link(id) fam(binomial) robust

*Compare this RD and notice that its the same
tab z
g total=610 if z==1
replace total=1300 if z==0
cs y x, by(z) stand(total) rd

* or parametric g-formula
logistic y i.x##i.z
margins r.x

*We can multiply the weights by a constant to standardize to different
populations
g wtexp=px*wt
glm y x [pw=wtexp], link(id) fam(binomial) robust

*We can stabilize the weights by multiplying by pr(X=x), which we get by
running a regression on X alone for Bo
logistic x
predict stab

g wtstab=wt*stab if x==1
replace wtstab=wt*(1-stab) if x==0
tab x z,row
glm y x [pw=wtstab],link(id) fam(binomial) robust

* Stata v 13 makes MSMs even easier. You don't need to calculate the
weights yourself
* standardized to the total
teffects ipw (y) (x z)
* standardized to the exposed
teffects ipw (y) (x z),atet



*Dealing with a normal exposure
* the effect of sysbp on death
use "/Users/maclehose/Dropbox/Classes/8343/PuBH 8343 fall 2016/Week
3/frmgham2.dta", clear
d
regress sysbp bmi cursmoke diabetes
predict xb,xb
g density_den=normalden(sysbp,xb,e(rmse))
regress sysbp
predict xb2,xb
g density_num=normalden(sysbp,xb2,e(rmse))
```

```
g wt=density_num/density_den
sum wt
glm death sysbp [pw=wt], family(binomial) link(id)

teffects ipw (death) (sysbp bmi cursmoke diabetes)

*A more complicated exposure model (more covariates)
*the effect of diabetes on death
use "/Users/maclehose/Dropbox/Classes/8343/PuBH 8343 fall 2016/Week
3/frmgham2.dta", clear
logistic diabetes c.age##i.cursmoke##c.totchol##c.sysbp i.sex##c.educ
predict p_diab,pr
logistic diabetes
predict p_num,pr
g wt=1/p_diab if diabetes==1
replace wt=1/(1-p_diab) if diabetes==0
g swt=wt*p_num if diabetes==1
replace swt=wt*(1-p_num) if diabetes==0
sum wt swt
hist swt
sum swt,d

glm death diabetes [pw=swt], family(binomial) link(id)
glm death diabetes [pw=swt] if swt>=0.29 & swt<=1.29, family(binomial)
link(id)


**Censoring examples

use "/Users/presentation/Dropbox/Classes/8343/PuBH 8343 fall 2016/week
5/msmtv.dta", clear



* generate x0 weights

logistic x0 z0
predict x0_den, pr


logistic x0
predict x0_num, pr

g wt_x0=x0_num/x0_den if x0==1
replace wt_x0 = (1-x0_num)/(1-x0_den) if x0==0

* generate C0 weights


logistic c0 i.x0 c.z0
predict c0_den, pr

logistic c0
predict c0_num, pr
```

```
g wt_c0=(1-c0_num)/(1-c0_den)


* generate X1 weights

logistic x1 x0 z0 z1  if c0==0
predict x1_den, pr

logistic x1 if c0==0
predict x1_num, pr

g wt_x1=x1_num/x1_den if x1==1
replace wt_x1=(1-x1_num)/(1-x1_den) if x1==0

g fwt=wt_x0*wt_x1*wt_c0

sum wt* fwt


* final model

logistic y i.x0 i.x1 [pw=fwt]
glm y i.x0##i.x1 [pw=fwt], fam(binomial) link(id)
```

STATA Code Part 2.

Meta Analysis and Bayes by extension

```
rename upper up

rename low low


*Step 1 of a metanalysis project: Use the code below to log transform the abstracted estimates and 95% bounds from your data

g theta =log(rr)

g up_theta=log(up)

g low_theta=log(low)


*Manually determine the std from the 95% CI values

g var=((up_theta-low_theta)/(2*1.96))^2


*create your weight values to estimate your composite estimate

g wt=1/var

g wt_theta=log(rr)*wt


sum wt

local sumwt=r(sum)

local v_psi=1/`sumwt'

sum wt_theta

local sumwt_theta=r(sum)


di `sumwt_theta'

di `sumwt'

di exp(`sumwt_theta'/`sumwt')

di exp(`sumwt_theta'/`sumwt'-1.96*sqrt(`v_psi'))

di exp(`sumwt_theta'/`sumwt'+1.96*sqrt(`v_psi'))
```

*This is how you can automate a meta analysis in STATA

metan theta low_theta up_theta, fixedi eform

* meta analysis day 2

use "/Users/presentation/Dropbox/Classes/8343/PuBH 8343 fall 2016/week 11/slides/metadata2.dta", clear

* In this first step, we are using the automated stata function to run a data analysis, after placing all the effect estimates on the log scale

g logrr=log(rr)

g logup=log(upper)

g loglow=log(lower)

* we are running a meta analysis using both a fixed effect approach, and a random effect approach

* in presence of heterogeneity (ie the strata or different studies are too different to aggregate together)

* a common practice is to do a random effect analysis (the wider 95% CI help incorporate the heterogeneity

* stata uses the DerSimonian and Laird RE model to implement the RE in case of heterogeneity in meta analysis.

metan logrr loglow logup, fixedi nograph eform

metan logrr loglow logup, randomi nograph eform

*****

*BCG EXample

**

```
use "/Users/presentation/Dropbox/Classes/8343/PuBH 8343 fall 2016/week 11/slides/bcgtrial.dta", clear


g control1=tot1-cases1

g control0=tot0-cases0


* Here, we are using a different syntax (we actually have all the data for a 2x2 so we did not need to log transform the estimate, LL and UL from different studies

* We also have the option to pick between rr, rd, or for our aggregated estimates. However, the measure of interest should match the one used in the studies which form the strata of our metaanalysis


metan cases1 control1 cases0 control0, rr fixedi nograph

metan cases1 control1 cases0 control0, rr randomi nograph

metan cases1 control1 cases0 control0, rd randomi nograph


*publication bias

*to test for publication bias, we can use the egger regression model with y=log(rr) and x=var(log(rr)).

*the first step would be to run a  regular meta analysis

*the second step would be to transform the aggregate meta analysis estimate into log(rr) for our y variable, and var(log(rr)) as our x variable

*once we have these variables, we can use the metafunnel command to make a graph, and get a formal test for publication bias using the metabias command

*asymetric funnel plots can be a consequence if true heterogeneity (by study size), poor quality studies, random error and sometimes non collapsibility of OR (baseline risk can vary by study size)


metan cases1 control1 cases0 control0, rr fixedi nograph

g logrr=log(_ES)

g se=_selogES

g wt=_WT


metafunnel logrr se
```

metabias logrr se, egger

*double check with alternate model: E(theta)=b0+b1*se

regress logrr se [pw=wt]

metabias logrr se, begg

metan cases1 control1 cases0 control0 if se<0.4, rr fixedi nograph

****meta regression

*we sometimes try to explain substantial heterogeneity with regression:

*fe meta analysis is just a mean model, so by adding another variable that could explain that variability across out studies, we can include it into the model as an additional variable

*this can be done using the vwls for fixed effect meta analysis, ir metareg with the wsse option for the metareg command

*fixed effects

vwls logrr lat, sd(se)

*RE regression

metareg logrr lat, wsse(se)

metareg logrr lat start, wsse(se)