

My contact information

- Eric Vittinghoff
- E-mail: eric@biostat.ucsf.edu
- Office 5723

1

Multiple predictor linear regression

- Models dependence of the mean of continuous outcome on multiple predictors simultaneously
- By including multiple predictors we can try to
 - *estimate causal mechanisms from observational data*: control confounding, examine mediation, assess interaction
 - *risk stratify patients*: make well-calibrated predictions with good discrimination
 - *also*: account for stratification in RCT, make treatment estimate more precise

3

Biostat 208 Session 4 Outline

- Quick review of the multi-predictor linear model
- How models deal with confounding and mediation
- Log-transformed outcomes and predictors

2

Review: components of the linear model

- Systematic:
 - how does the mean value of outcome y depend on values of the predictors?
- Random:
 - at each observed value of the predictors, values of y are distributed about the predicted mean
 - assumed distribution of deviations underlies hypothesis tests, p-values, and confidence intervals

4

Review: systematic part of the model

- In abstract terms, model written as

$$E[y|\mathbf{x}] = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p$$

- $E[y|\mathbf{x}]$: Expected value or mean of y for a given set of values for predictors $\mathbf{x} = (x_1, x_2, \dots, x_p)$
- β_j : change in $E[y|\mathbf{x}]$ per unit increase in predictor x_j , holding all other predictors constant
- β_0 (the intercept): $E[y|\mathbf{x}]$ when all predictors = 0

5

Review: assumptions about the predictors

- No Normality assumption
 - predictors can be continuous, discrete (e.g. counts), categorical (dichotomous, nominal, ordinal)
 - however, non-Normality can result in violations of linearity, influential points (Session 6)

7

Review: random part of the model

$$y_i = E[y|\mathbf{x}_i] + \epsilon_i$$

- Error term ϵ assumed to be
 1. Normally distributed
 2. mean zero at every value of \mathbf{x}
 3. constant variance
 4. statistically independent across observations

6

Review: assumptions about the predictors

- No assumption of independence within observations
 - correlated predictors are the main reason for using multi-predictor models
 - but very high correlation (co-linearity) can cause problems (Sessions 6, 7)
- Measured without error
 - measurement error in predictors can induce regression dilution bias and residual confounding

8

Standard errors of linear regression coefficients

$$SE(\hat{\beta}_j) = \frac{\sigma_e}{\sigma_{x_j} \sqrt{(n-1)(1-\rho_j^2)}}$$

- Precision increases with larger sample size (n), greater variability in x_j (σ_{x_j})
- But decreases with higher multiple correlation of x_j with covariates (ρ_j), greater outcome variability left unexplained (σ_e)
- In linear models, adjustment can *increase* power if reduction in σ_e outweighs the *variance inflation factor* $1/(1-\rho_j^2)$ – but does not hold for logistic, Cox models

9

Confounding

- *Confounder*: a common cause of both primary predictor and outcome, or a surrogate for a common cause
- Can account for the some or all of the unadjusted association between a predictor and an outcome
- Controlling confounding the main reason for doing multi-predictor regression
- Only an association adjusted for confounders can be viewed as possibly causal

11

Summary of multipredictor model

- A tool for estimating how mean of outcome depends on multiple predictors simultaneously
- Inferential machinery evaluates precision of estimates, and whether sampling error can account for findings
- Coefficients interpretable as change in mean of outcome per unit increase in predictor, holding other predictors constant

10

Rubin's causal model and counterfactuals

- Suppose we could observe *counterfactual* outcomes for each individual at every possible level of a causal variable – for example, treatment and placebo
- Differences between these counterfactual outcomes are what we mean by causal effects
- Other causal influences would intrinsically be *held constant* in this counterfactual experiment
- Thus the populations contributing outcomes at each predictor level are *exchangeable* and *completely overlap*

12

Average causal effects

- Average causal effect (ACE): difference in individual counterfactual outcomes at different levels of a primary predictor, *averaged across the population*
- The causal effect may vary across individuals
- It may also vary according to measured effect modifiers
- The averaging gets us to the public health perspective
- *Both* ACE and within-subgroup causal effects are of interest

13

Causal effects and modeling

- RCTs mimic counterfactual experiment by balancing other causal influences
 - treatment groups are exchangeable, at least on average in large trials
- Exchangeability in observational data is only achieved by accurately modeling all confounding effects
 - within modeled strata, sub-populations are exchangeable
 - *unconfoundedness* means that primary predictor is independent of counterfactual outcomes, given covariates

14

Adjusted regression coefficients and ACE

- An adjusted regression coefficient is unbiased for ACE *if*:
 - there are no unmeasured or poorly modeled confounders
 - the sample is representative of the target population
 - primary predictor does not interact with covariates, or interactions are properly handled (Session 5); otherwise coefficient estimates a subgroup causal effect
 - the model is linear; more about logistic models later

15

Directed Acyclic Graphs (DAGs)

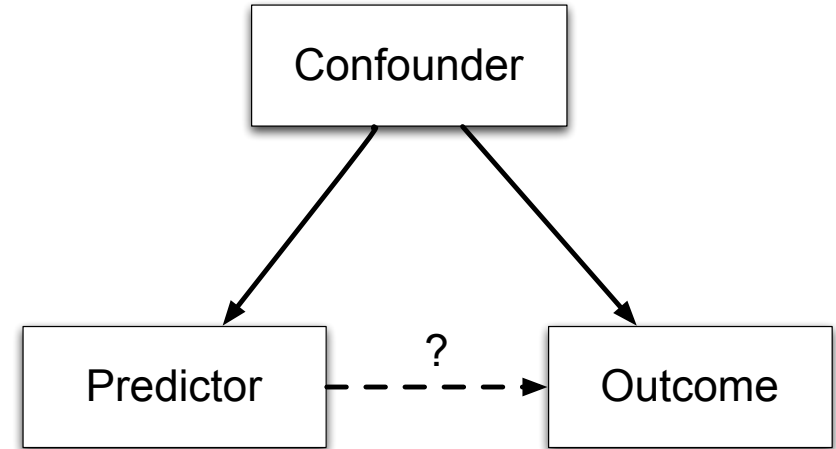
- DAGs help us to identify what we do and do not need to control for to achieve unconfoundedness
- Primary predictor, outcome, potential confounders and mediators represented as *nodes*
- Causal relationships depicted as *directed edges*
- No factor can cause itself, hence graph is *acyclic*

16

Directed Acyclic Graphs (DAGs)

- A good DAG requires *a lot* of prior substantive information about what causes what
- We analyze DAG for *backdoor paths* between primary predictor and outcome
- Unblocked backdoor paths can induce a statistical association between predictor and outcome in cases where no causal link exists
- Controlling for a confounder on a backdoor path *blocks* it, preventing the artefactual association

17

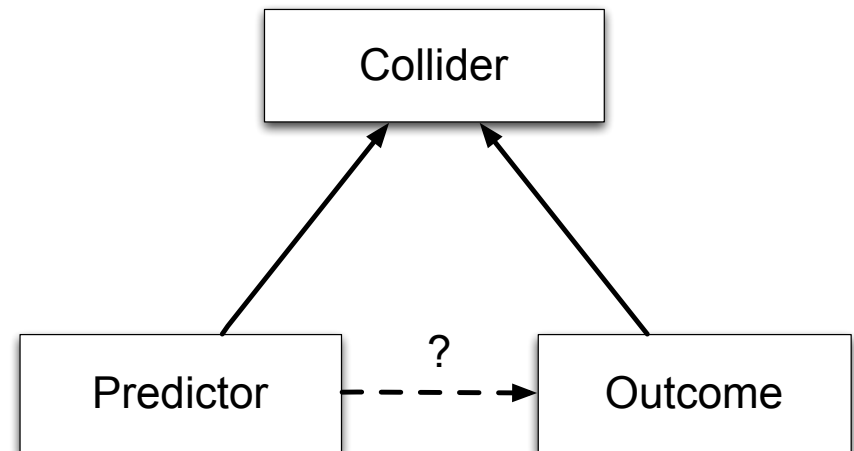


18

Colliders

- A *collider* on a backdoor path from predictor to outcome is a node with incoming arrows from both directions
- A collider blocks the backdoor path if it is *not* adjusted for
- Adjusting for a collider *opens the path* and potentially *induces* an artefactual association
- Also controlling for a non-collider on same the backdoor path re-blocks it

19



20

Insights based on colliders

- Common effects of predictor and outcome are colliders – *don't control for them*
- Do control confounding of mediator→outcome relationship
- Selection biases can also be represented as colliders:
 - restriction of sample according to effects of both predictor and outcome opens a backdoor path
- See Hernán *et al.*, A structural approach to selection bias, *Epidemiology*, 2004;15(5):615–625

21

A simple model with confounding

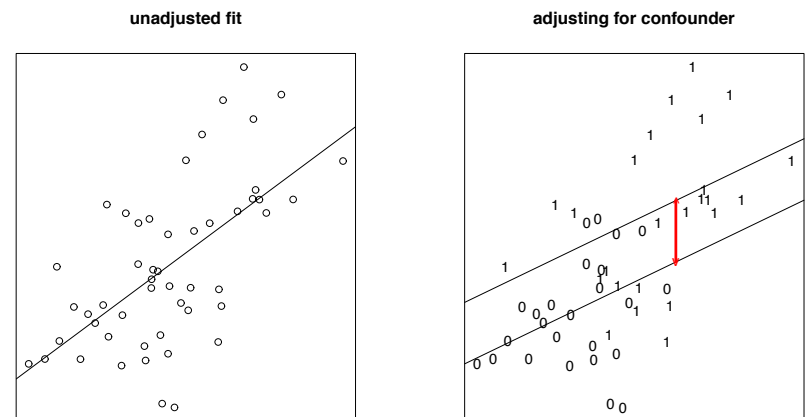
- Continuous primary predictor x_1
- Correlated binary confounder x_2 , either a cause of x_1 or a surrogate for one
- Continuous outcome y
- True model is $E[y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

22

A simple model with confounding

- Unadjusted effect of x_1 partly reflects higher prevalence of $x_2 = 1$ at higher values of x_1
- When x_1 increases, so does $E(x_2)$
- Adjustment fixes the problem by modeling effect of x_2
- Model allows us to estimate effect of x_1 holding x_2 constant

23



24

Unadjusted model

```
. reg outcome predictor
```

Source	SS	df	MS			
Model	49222.1184	1	49222.1184	Number of obs =	50	
Residual	53059.449	48	1105.40519	F(1, 48) =	44.53	
Total	102281.567	49	2087.37893	Prob > F =	0.0000	
				R-squared =	0.4812	
				Adj R-squared =	0.4704	
				Root MSE =	33.248	

outcome	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
predictor	6.165333	.9239258	6.67	0.000	4.307655	8.02301
_cons	-246.862	60.56897	-4.08	0.000	-368.6441	-125.0799

- True value of β_1 (predictor) is 2.5

25

Model adjusting for the confounder

```
. reg outcome predictor confounder
```

Source	SS	df	MS			
Model	62764.6569	2	31382.3285	Number of obs =	50	
Residual	39516.9104	47	840.785328	F(2, 47) =	37.33	
Total	102281.567	49	2087.37893	Prob > F =	0.0000	
				R-squared =	0.6136	
				Adj R-squared =	0.5972	
				Root MSE =	28.996	

outcome	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
predictor	3.31254	1.074505	3.08	0.003	1.150915	5.474165
confounder	43.89191	10.93647	4.01	0.000	21.89058	65.89325
_cons	-82.3544	66.86232	-1.23	0.224	-216.864	52.15524

- True value of β_1 (predictor) is 2.5
- True value of β_2 (confounder) is 25

26

Confounding: interpretation of results

- Unadjusted estimate for primary predictor
 - estimates an observable trend in whole population
 - but a causal interpretation misleading in most contexts
- Adjusted estimate has a causal interpretation in this case
- More generally, a causal interpretation if unconfoundedness is convincing

27

Confounding: interpretation of results

- Regression lines for subgroups with $x_2 = 0$ and $x_2 = 1$:
 - slopes estimate predictor/outcome association within each subgroup (i.e., holding x_2 constant)
 - assumed parallel (no interaction – same effect in both subgroups)

28

Behavior of regression coefficients for this case

- When the primary predictor and confounder are
 - positively correlated
 - both predict higher (or lower) values of the outcome
 adjusted coefficient for primary predictor is attenuated.
- Same result if predictor, confounder negatively correlated, have opposite effects
- Typical pattern for confounding

29

Negative confounding

- Confounding can also mask an independent association
- Overall, AZT prophylaxis does not predict HIV transmission after needlestick
 - AZT use associated with severity of injury
 - injury severity increases transmission risk
- AZT protection unmasked after controlling for severity

30

Negative confounding: two scenarios

Negative confounding may arise between predictors that are

1. *Positively correlated, with opposite effects on outcome:*
 Example: injury severity, AZT, and HIV transmission
2. *Inversely correlated, with similar effects on outcome:*
 Example: mean BMI decreases with age in HERS cohort, but both predict increased SBP

31

Unadjusted association of BMI and SBP

```
. reg sbp BMI
```

Source	SS	df	MS			
Model	3573.21875	1	3573.21875	Number of obs =	2758	
Residual	993677.692	2756	360.550687	F(1, 2756) =	9.91	
Total	997250.911	2757	361.715963	Prob > F =	0.0017	
				R-squared =	0.0036	
				Adj R-squared =	0.0032	
				Root MSE =	18.988	

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
BMI	.2063226	.065539	3.148	0.002	.077812	.3348332
_cons	129.1723	1.907635	67.713	0.000	125.4318	132.9129

32

BMI and SBP adjusted for age

. reg sbp BMI age

Source	SS	df	MS			
Model	33945.7704	2	16972.8852	Number of obs =	2758	
Residual	963305.14	2755	349.657038	F(2, 2755) =	48.54	
Total	997250.911	2757	361.715963	Prob > F =	0.0000	
				R-squared =	0.0340	
				Adj R-squared =	0.0333	
				Root MSE =	18.699	

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
BMI	.3034786	.0653778	4.642	0.000	.1752842	.431673
age	.5056204	.0542507	9.320	0.000	.3992443	.6119965
_cons	92.69495	4.341354	21.352	0.000	84.18231	101.2076

33

Confounding is difficult to rule out

- Were all important confounders adjusted for?
- Were they measured accurately?
- Were their effects modeled adequately – no omitted interactions (Session 5) or unmodeled non-linearities (Session 6)?
- Was there enough “overlap” between the groups being compared to avoid extrapolation (Session 6)?

35

Summary: negative confounding

- Negative confounding can mask an independent association in unadjusted analysis
- Not all that uncommon
- Implications for predictor selection: univariate screening, forward selection procedures may miss some negatively confounded predictors (Session 7)
 - hence our recommendation to use backwards deletion in most circumstances

34

Getting the model right

- Plausible DAGS may omit both nodes and edges even in well-studied areas, leading to omitted confounders
- Non-linearities and interactions common but hard to detect
 - default linear, no-interaction model may well be too simple
- Avoiding extrapolation may require *restriction* of scope of inference (e.g., ACE in treated, Complier ACE)
- Difficulty of getting the model right a primary motivation for methods based on *propensity scores*

36

Confounding: summary

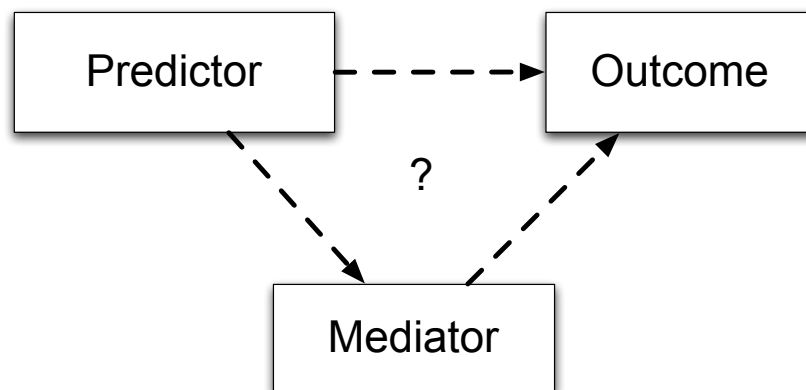
- Ideally, regression controls for confounding by jointly modeling effects of exposure and confounders
- Unbiased estimation of causal effects depends on getting the model right
- Bigger samples don't reduce unmodeled confounding, but do make modeling well-measured confounders easier
- *Residual confounding is difficult to rule out*

37

Mediation

- *Confounder*: a common cause of both primary predictor and outcome, or a surrogate for a common cause
- *Mediator*: on causal pathway from predictor to outcome
- In models, mediation and confounding behave alike
 - must be distinguished on substantive grounds
- Goal: estimate *overall* and *direct* effects of primary predictor, evaluate *indirect pathway via mediator*

38



39

Examining mediation

- Use a series of models to figure out whether:
 - primary predictor independently affects mediator
 - mediator affects outcome independently of primary predictor
 - the estimated effect of the primary predictor changes after adjustment for the mediator

40

Examining mediation

- The models:
 1. regress mediator on predictor and confounders
 2. regress outcome on predictor and confounders
 3. regress outcome on predictor, confounders, and mediator
- Models 2 and 3 must both
 - use the same observations
 - include confounders of mediator→outcome relationship

41

Model 1: BMI predicts higher glucose

```
. reg glucose BMI age exercise drinkany
```

Source	SS	df	MS			
Model	321116.455	4	80279.1137	Number of obs =	2756	
Residual	3426880.46	2751	1245.68537	F(4, 2751) =	64.45	
Total	3747996.91	2755	1360.43445	Prob > F =	0.0000	
				R-squared =	0.0857	
				Adj R-squared =	0.0843	
				Root MSE =	35.294	

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
BMI	1.6902	.1256441	13.45	0.000	1.443833	1.936566
age	-.2144159	.102671	-2.09	0.037	-.4157359	-.0130959
exercise	-2.774221	1.400946	-1.98	0.048	-5.521234	-.0272081
drinkany	-7.15434	1.385168	-5.16	0.000	-9.870414	-4.438266
_cons	82.01119	8.38539	9.78	0.000	65.5689	98.45349

43

Mediation

- Interpretation of coefficient estimates for primary predictor:
 - before adjustment for mediator: *overall effect*
 - after adjustment: *direct effect* via other pathways
- Assess mediation by $\beta_{overall} - \beta_{direct}$
- Alternatively use PTE: $(\beta_{overall} - \beta_{direct})/\beta_{overall}$
- Example: do glucose levels mediate effects of BMI on SBP?

42

Model 2: overall effect of BMI on SBP

```
. reg SBP BMI age exercise drinkany
```

Source	SS	df	MS			
Model	42202.7194	4	10550.6799	Number of obs =	2756	
Residual	954711.522	2751	347.041629	F(4, 2751) =	30.40	
Total	996914.241	2755	361.856349	Prob > F =	0.0000	
				R-squared =	0.0423	
				Adj R-squared =	0.0409	
				Root MSE =	18.629	

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
BMI	.245644	.0663176	3.70	0.000	.1156067	.3756813
age	.4933303	.0541919	9.10	0.000	.3870694	.5995913
exercise	-2.911396	.7394488	-3.94	0.000	-4.361327	-1.461465
drinkany	-2.022715	.7311205	-2.77	0.006	-3.456316	-.5891144
_cons	97.07514	4.425984	21.93	0.000	88.39656	105.7537

```
. estimates store m2
. scalar b_overall = _b[BMI]
```

44

Model 3: direct BMI effect via other pathways

```
. reg SBP BMI age exercise drinkany glucose
```

Source	SS	df	MS
Model	50555.9327	5	10111.1865
Residual	946358.308	2750	344.130294
Total	996914.241	2755	361.856349

```

Number of obs = 2756
F( 5, 2750) = 29.38
Prob > F = 0.0000
R-squared = 0.0507
Adj R-squared = 0.0490
Root MSE = 18.551

```

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
BMI	.1621961	.0681763	2.38	0.017	.0285141 .2958781
age	.5039164	.0540069	9.33	0.000	.3980183 .6098145
exercise	-2.774428	.7368653	-3.77	0.000	-4.219294 -1.329563
drinkany	-1.669494	.7315689	-2.28	0.023	-3.103974 -.2350138
glucose10	.4937161	.1002102	4.93	0.000	.2972212 .690211
_cons	93.02612	4.483349	20.75	0.000	84.23505 101.8172

```

. estimates store m3
. scalar b_direct = _b[BMI]
. display (b_overall - b_direct) / b_overall * 100
33.971067

```

Mediation of BMI by glucose levels

- BMI independently predicts glucose (Model 1)
- Glucose independently predicts SBP (Model 3)
- Overall BMI effect: 0.25 mmHg per kg/m² (Model 2)
- Direct BMI effect: 0.16 mmHg per kg/m² (Model 3)
- PTE: glucose explains $(.25-.16)/.25*100 = 34\%$ of the effect of BMI on SBP

Testing for mediation

- We don't test for confounding – why test for mediation?
- Statistically significant difference between overall, direct effects helps validate pathway
- Overall and direct effect estimates based on *separate models* using the *same data*, so they are correlated
- Test of equality accounting for correlation done using `suest` command in Stata

Comparing overall and direct BMI effects

```
. suest m2 m3
Simultaneous results for m2, m3
```

	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
m2_mean					
BMI	.245644	.0665409	3.69	0.000	.1152263 .3760617
m3_mean					
BMI	.1621961	.0682976	2.37	0.018	.0283353 .2960569

```

. test [m2_mean]BMI = [m3_mean]BMI
( 1) [m2_mean]BMI - [m3_mean]BMI = 0
      chi2( 1) = 20.61
      Prob > chi2 = 0.0000

```


Negative mediation

- TZDs cause bone loss in mouse models for diabetes
- Among diabetics in Health ABC, TZD use not associated with bone loss, after controlling for confounders by indication
- TZDs also cause weight gain (protective against bone loss)
- TZD use predicts bone loss, after controlling for weight gain
- Focus on overall or direct effect?

53

Summary: mediation

- Regression coefficients change when either a confounder or a mediator is added to the model; which is which depends on how you draw the causal arrows
- Models 2 and 3 must use the same observations, and must both control for confounders of mediator
- Estimated independent effect of primary predictor
 - before adjustment for mediator: overall effect
 - after adjustment: direct effect via other pathways

55

Confounding or mediation?

Primary Predictor	Adjustment Variable	Outcome
visceral fat	adipocytokines	insulin resistance
visceral fat	total body fat	insulin resistance
diabetes	nonfatal MI	heart failure
waist circumference	exercise	glucose level
statin use	LDL level	MI
HAART	CD4 ⁺ cell count	AIDS

54

Interpreting results for log-transformed variables

- Positive continuous variables commonly log-transformed
 - outcomes: normalize and equalize variance
 - predictors: get rid of non-linearity, interaction
 - more about this is session 6
- Both \log_{10} (HIV viral load) and natural log transformations used
- How does this affect interpretation of regression coefficients?

56

Log-transformed predictors

Log-transformed predictors

- For natural-log or \log_{10} transformed predictor x_j
 - $\hat{\beta}_j$ estimates the increase in $E(y)$ for a 1-unit increase in $\log(x_j)$
 - equivalently a 2.7-fold or 10-fold increase in x_j .

57

SBP and log-transformed creatinine

```
. reg sbp age10 lncreat diabetes
```

Source	SS	df	MS	Number of obs = 2761		
Model	62724.9657	3	20908.3219	F(3, 2757)	=	61.70
Residual	934341.037	2757	338.897728	Prob > F	=	0.0000
Total	997066.003	2760	361.255798	R-squared	=	0.0629
				Adj R-squared	=	0.0619
				Root MSE	=	18.409

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.444083	.0536661	8.27	0.000	.3388531	.5493129
lncreat	9.211717	1.682269	5.48	0.000	5.913081	12.51035
diabetes	6.426345	.8007529	8.03	0.000	4.856209	7.996481
_cons	103.2765	3.600996	28.68	0.000	96.21558	110.3374

```
. * increase in SBP associated with a 50% increase in creatinine
. nlcom _b[lncreat]*log(1.5)
```

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_nl_1	3.73503	.6821016	5.48	0.000	2.397548	5.072511

59

- For natural-log or \log_{10} transformed predictor x_j
 - $\hat{\beta}_j \ln(1 + k/100)$ estimates the change in $E(y)$ for a $k\%$ increase in x_j .
 - e.g., $\hat{\beta}_j \ln(1.5)$ estimates the change in $E(y)$ for a 50% increase in x_j .
- Use $\hat{\beta}_j \log_{10}(1 + k/100)$ if x_j is \log_{10} -transformed
- Use `nlcom` to get interpretable estimates with confidence interval (`lincom` does not allow `log()` as argument)

58

Natural log-transformed outcomes

- Outcome is $\ln(y)$
- $e^{\hat{\beta}_j}$ estimates the *relative increase* in $E(y)$ – the mean of the *untransformed outcome* – for a 1-unit increase in x_j
 - if $e^{\hat{\beta}_j} = 1.5$, a 1.5-fold increase in $E(y)$, from 1 to 1.5, 10 to 15, 100 to 150, ... for a 1-unit increase in x_j
- $100(e^{\hat{\beta}_j} - 1)$ estimates the *percent increase* in $E(y)$ for a 1-unit increase in x_j
 - a 50% increase in $E(y)$ for a 1-unit increase in x_j

60

Natural log-transformed outcomes

- For untransformed predictors:
 - use `eform("exp(beta)")` to have `regress` display $e^{\hat{\beta}_j}$
 - use `lincom` with `eform` option to get relative increase in outcome per k -unit increase in predictor
 - use `nlcom` to get percent increase in outcome per k -unit increase in predictor

61

Model for log-creatinine

```
. reg lncreat age lntgl diabetes, eform("exp(beta)")
```

Source	SS	df	MS			
Model	6.08561714	3	2.02853905	Number of obs =	2757	
Residual	118.993186	2753	.043223097	F(3, 2753) =	46.93	
Total	125.078803	2756	.045384181	Prob > F =	0.0000	
				R-squared =	0.0487	
				Adj R-squared =	0.0476	
				Root MSE =	.2079	

lncreat	exp(beta)	Std. Err.	t	P> t	[95% Conf. Interval]	
age	1.005867	.0005998	9.81	0.000	1.004692	1.007044
lntgl	1.04169	.0103602	4.11	0.000	1.021573	1.062204
diabetes	1.049131	.0095069	5.29	0.000	1.030654	1.067939

63

Outcome and predictor both log-transformed

- $e^{\hat{\beta}_j \ln(1+k/100)}$ estimates the relative increase in $E(y)$ for a $k\%$ increase in x_j
- $100(e^{\hat{\beta}_j \ln(1+k/100)} - 1)$ estimates the percent increase in $E(y)$ for a $k\%$ increase in x_j
- Use `eform` option with `regress`, and `nlcom` to get interpretable estimates with confidence intervals
- See VGSM, Sect. 4.7.5

62

Model for log-creatinine (continued)

```
. * relative increase in creatinine associated with 10-year increase in age
. lincom age*10, eform
```

lncreat	exp(b)	Std. Err.	t	P> t	[95% Conf. Interval]	
(1)	1.060246	.0063224	9.81	0.000	1.047921	1.072715


```
. * percent increase in creatinine associated with 10-year increase in age
. nlcom 100*(exp(_b[age]*10)-1)
```

lncreat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_nl_1	6.024558	.6322391	9.53	0.000	4.784847	7.264269


```
. * percent increase in creatinine for a 25% increase in tryglyceride levels
. nlcom 100*(exp(_b[lntgl]*log(1.25))-1)
```

lncreat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_nl_1	.9155889	.2239611	4.09	0.000	.4764403	1.354738

64