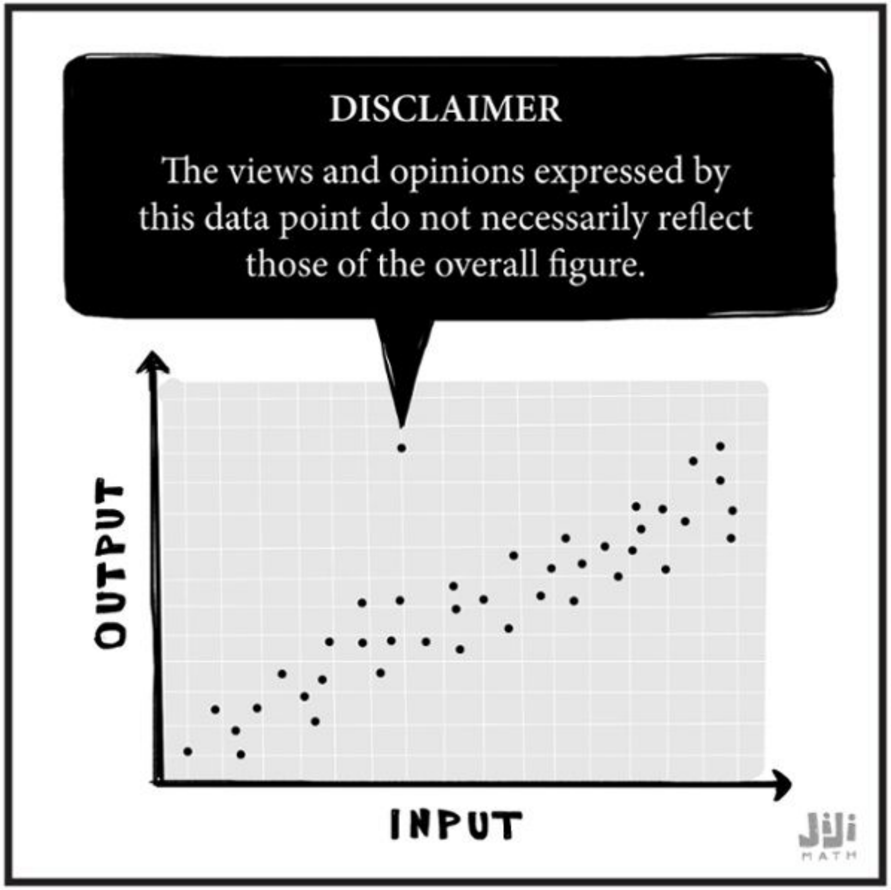
BST 210
Applied Regression Analysis


DISCLAIMER
The views and opinions expressed by this data point do not necessarily reflect those of the overall figure.
OUTPUT
INPUT
Jiji MATH

© MIND Research Institute 11/19/2015

Lecture 3

Plan for Today

- Questions from last class
- Recap: Framework for Analyses
- Multiple Linear Regression
- Confounding
- Effect Modification (interaction)

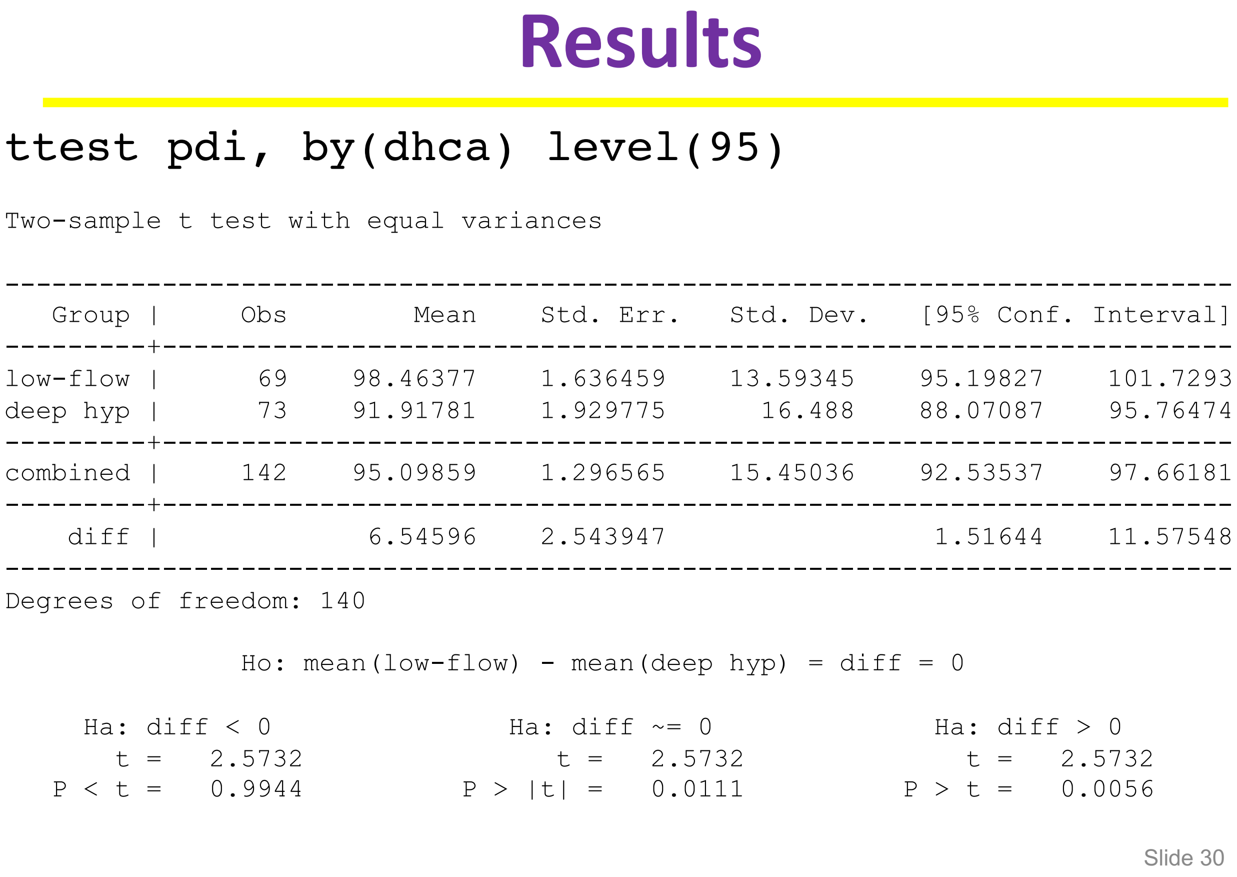
Questions from/since last class

- When comparing multiple means, students are sometimes advised to compare confidence intervals to see whether the intervals overlap. When 95% confidence intervals for the means of two independent populations don’t overlap, there will indeed be a statistically significant difference between the means (at the 0.05 level of significance).

- * However the opposite (or converse) is not always true. The CI’s may overlap, yet there may be a statistically significant difference between the means. That is, if two test statistics have non-overlapping confidence intervals, they are necessarily significantly different *but* if they have overlapping confidence intervals, it is not necessarily true that they are *not* significantly different. *

- The discrepancy arises since distance from the mean is calculated in a different way for the t-statistic than it is for mean confidence intervals.

Questions from/since last class



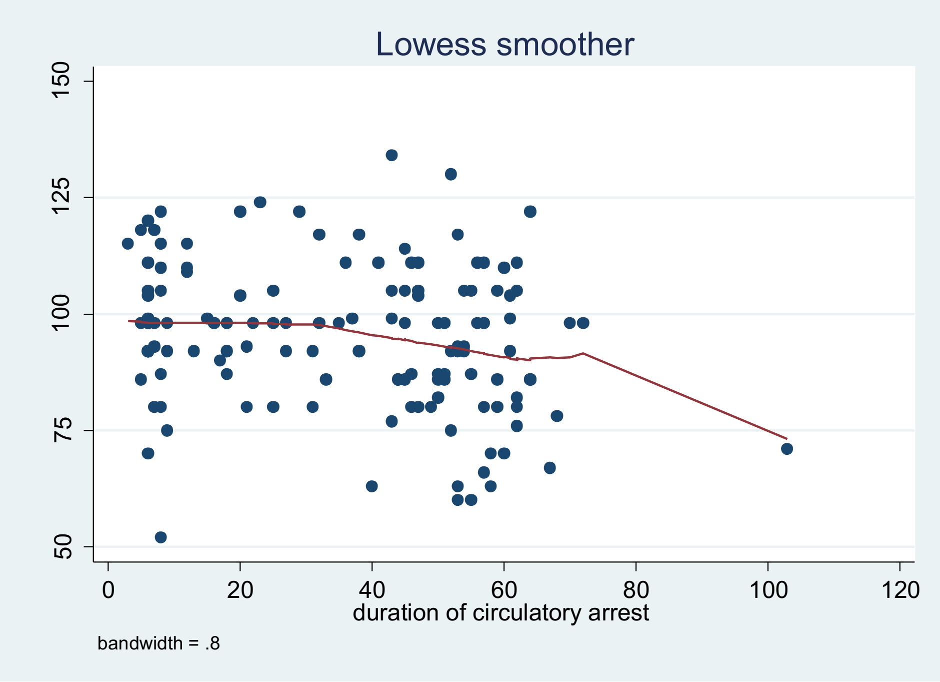
Results

```
ttest pdi, by(dhca) level(95)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
low-flow |      69    98.46377    1.636459    13.59345    95.19827    101.7293
deep hyp |      73    91.91781    1.929775      16.488    88.07087    95.76474
---------+--------------------------------------------------------------------
combined |     142    95.09859    1.296565    15.45036    92.53537    97.66181
---------+--------------------------------------------------------------------
    diff |             6.54596    2.543947                1.51644    11.57548
------------------------------------------------------------------------------
Degrees of freedom: 140

               Ho: mean(low-flow) - mean(deep hyp) = diff = 0

     Ha: diff < 0               Ha: diff ~= 0              Ha: diff > 0
       t =   2.5732               t =   2.5732               t =   2.5732
   P < t =   0.9944          P > |t| =   0.0111          P > t =   0.0056
```

Slide 30

Questions from/since last class

- Lowess smoothing: bandwidth = .8 is default
- What does larger bandwidth do? Smaller?



Review: True of false?

- R^2 = amount of variability in Y that our fitted model is unable to explain.

- Correlation => Causation

- One assumption of linear regression is that the outcomes Y are dependent.

- A mnemonic acronym for remembering the assumptions of linear regression is LANE.

Review: Is It a Linear Model?

- $\mathrm{E}(Y_\mathrm{i}) = \beta_0$
- $\mathrm{E}(Y_\mathrm{i}) = \beta_0 + \beta_1 \cdot age_\mathrm{i} + \beta_2 \cdot age_\mathrm{i}^2$
- $\mathrm{E}(Y_\mathrm{i}) = \beta_0 + \beta_1 \cdot age_\mathrm{i} + \beta_2 \cdot female_\mathrm{i} + \beta_3 \cdot age_\mathrm{i} \cdot female_\mathrm{i}$
- $\mathrm{E}(Y_\mathrm{i}) = \beta_0 + \beta_1 \cdot \exp(age_{\mathrm{i}1})$
- $\mathrm{E}(Y_\mathrm{i}) = \beta_0 + \exp(\beta_1 \cdot age_{\mathrm{i}1})$
- $\mathrm{E}(Y_\mathrm{i}) = \exp(\beta_0 + \beta_1 \cdot x_{\mathrm{i}1})$
- $\mathrm{E}(Y_\mathrm{i}) = (\beta_0 + \beta_1 \cdot age_\mathrm{i} + \beta_2 \cdot age_\mathrm{i}^2\,)^2$

Continue to develop general framework for approaching analyses

First -

- Learn the topic/study well, really well
- Collaborate to define motivating questions of interest, check PubMed, other sources
- What techniques might help to achieve answers? Which do the data warrant? (develop intuition, read literature)
- Possible Confounding or Effect Modification to account for?
- Keep an open mind, and the larger picture – there is no recipe

Continue to develop general framework for approaching analyses

Next -

- Diagnostics/Checking Assumptions:
  - Scatterplot, summary statistics
  - Boxplots, histograms
  - Correlations
  - Smoothing (example: Lowess)
  - Residual Analysis
- Hypothesis testing/modeling:
  - t-test?
  - Correlation (r)?
  - ANOVA useful?
  - Nonparametric approach better?
  - Linear regression or extensions (multiple reg.)?
  - Generalizations

Recall Motivating Questions in last example – how do we accommodate?

- Is PDI related to the (continuous) duration of CA?
- Is PDI related to the (categorical) treatment group (CA vs. LFB)?
- Is PDI related to treatment group, after adjusting for diagnosis group (IVS and VSD)?
- Other predictor variables?

Recall Motivating Questions in last example – how do we accommodate?

- Considering all of the possible covariates, which factors are most predictive of PDI?
- What are the final conclusions regarding treatment group comparisons, adjusting for other factors?
- Need to build multiple linear regression models to predict PDI

Multiple Linear Regression

- **Simple linear regression**: a single independent variable (**Y**) is used to predict the value of a dependent variable (**X**).
- **Multiple linear regression**: two or more independent variables (**X**) are used to predict the value of a dependent variable (**Y**).
- The difference between the two is simply the number of independent variables.

- Data: $(x_{i1}, x_{i2}, \ldots, x_{ip}, Y_i)$, $i = 1, \ldots, N$

  x_{ij} = jth predictor variable for the ith subject, measured without error

  Y_i = outcome for the ith subject, random, continuous, may have error

  N = number of subjects

Multiple Linear Regression

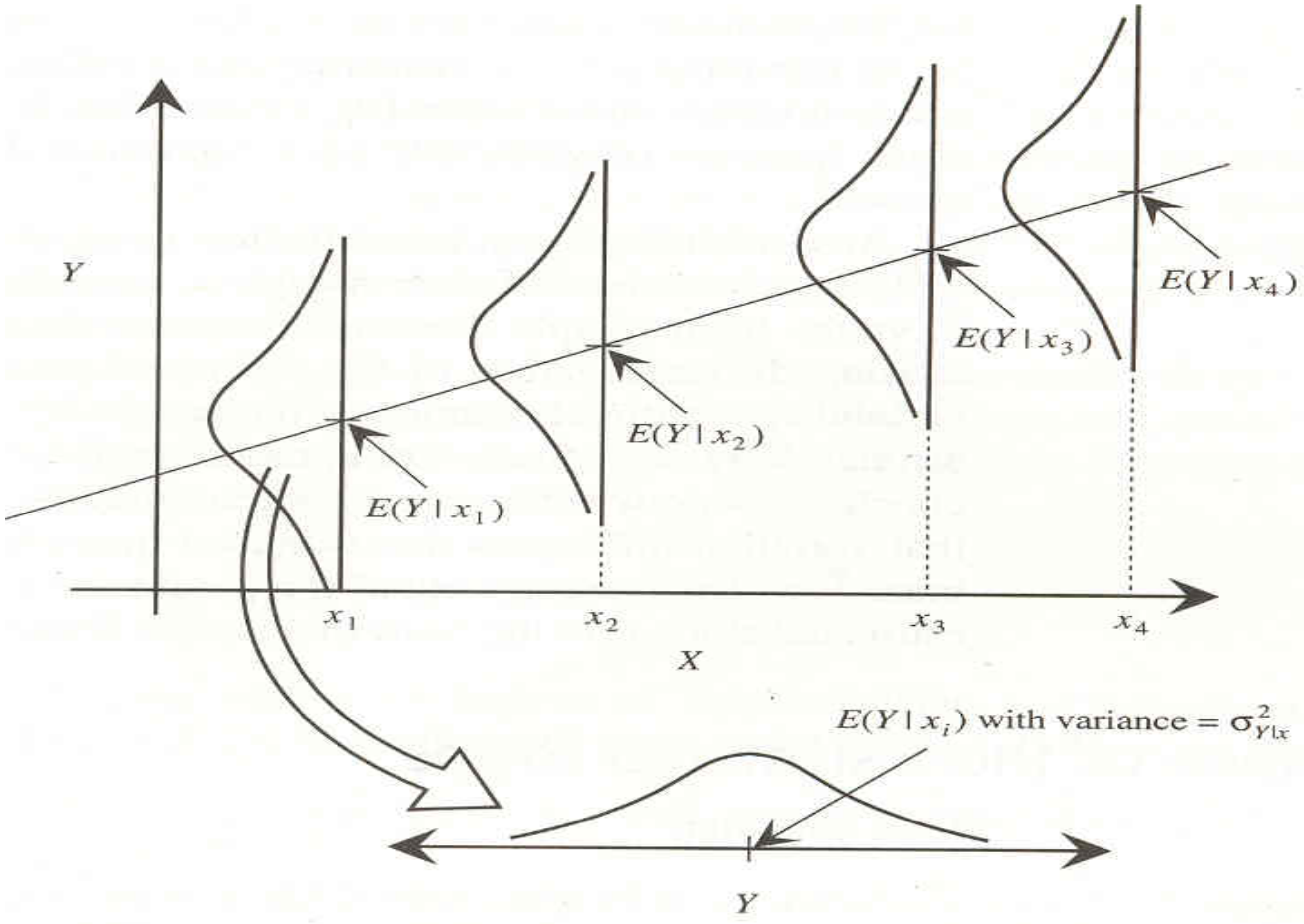
- Model: $\mathbf{E(\boldsymbol{Y}_i) = \beta_0 + \beta_1 \boldsymbol{x}_{i1} + \beta_2 \boldsymbol{x}_{i2} + \ldots + \beta_p \boldsymbol{x}_{ip}}$

  - $E(Y_i)$ = expected value of Y for a given set of covariates, x_{i1} , x_{i2} ,…, x_{ip}
  - β_0 = intercept, or constant term, corresponding to the mean value of Y when all covariates = 0
  - β_j = slope, or the change in Y corresponding to a 1 unit increase in the jth covariate, x_j, holding all the other covariates constant

Multiple Linear Regression

- Assumptions:
  - **(L)** the mean of Y_i is an unknown, but **linear**, function of x_{i1} , x_{i2} ,…, and x_{ip}
  - **(I)** all responses are ***independent***
  - **(N)** the distribution of *Y* about its mean value is ***normally** distributed*
  - **(E)** the variability of *Y* about its mean value is **equal** for all *x* values (*homoscedasticity*)
  - **Existence**: For any fixed value of the variable *X*, *Y* is a random variable with a certain probability distribution having finite mean and variance

For any fixed value of the variable *X*, *Y* is a random variable with a certain probability distribution having finite mean and variance

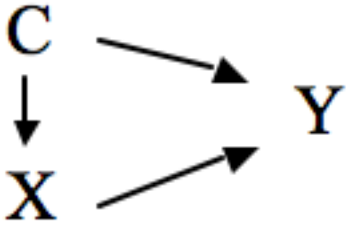


Let’s back up: Confounding

- Suppose we are interested in the association between an exposure and outcome
- But there may be other factors that distort the relationship between exposure and outcome
- What to do?

Confounding Review

- A variable is a **confounding variable** if it satisfies two conditions (classical definition of confounding):
  - It is a risk factor for the outcome
  - It is associated with exposure, but not a consequence of exposure
- Failure to control for confounding can lead to **bias**



Control of Confounders: Study Design

- **Randomization** (clinical trials)
  - – Should balance confounders in groups being compared
- **Restriction**
  - – Select a restricted subgroup to study
- **Matching** (case-control studies)
  - – Cases and controls have same confounding characteristics (hence balanced)

Control of Confounders: Data Analysis

- **Stratification**
  - Split the data into strata, make within-strata comparisons, then recombine to get overall estimates
  - Compare crude (unadjusted) and stratified (adjusted) estimates to assess confounding
- **Multivariable analysis**
  - Include covariate in multiple linear regression

$$\mathbf{E}(Y_i) = \beta_0 + \beta_1 C_i + \beta_2 X_i$$

Multiple Linear Regression

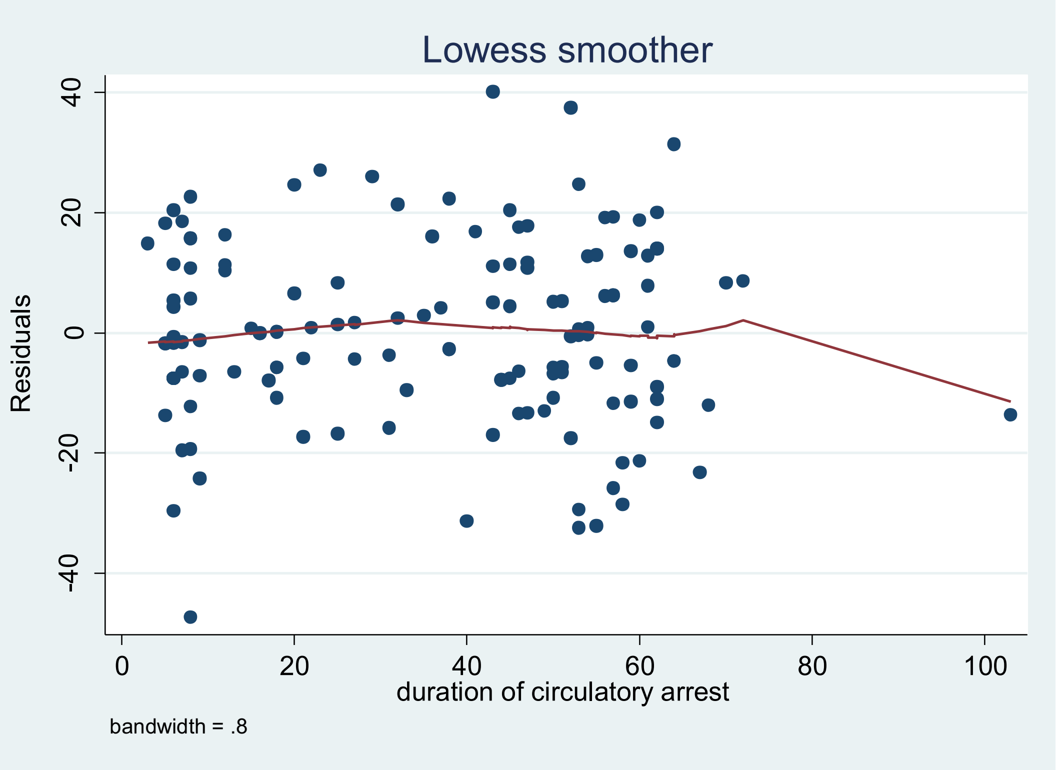
- Used to investigate the relationship between a response variable and *several* explanatory variables
- Model: $E(Y) = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_p \cdot x_p$
- The intercept β_0 is the predicted value of Y when all covariates = 0

Multiple Linear Regression

- The slope β_j is the change in Y corresponding to a 1 unit change in x_j, assuming all other covariates are held constant
- We say that we are adjusting for, or controlling for, the other covariates

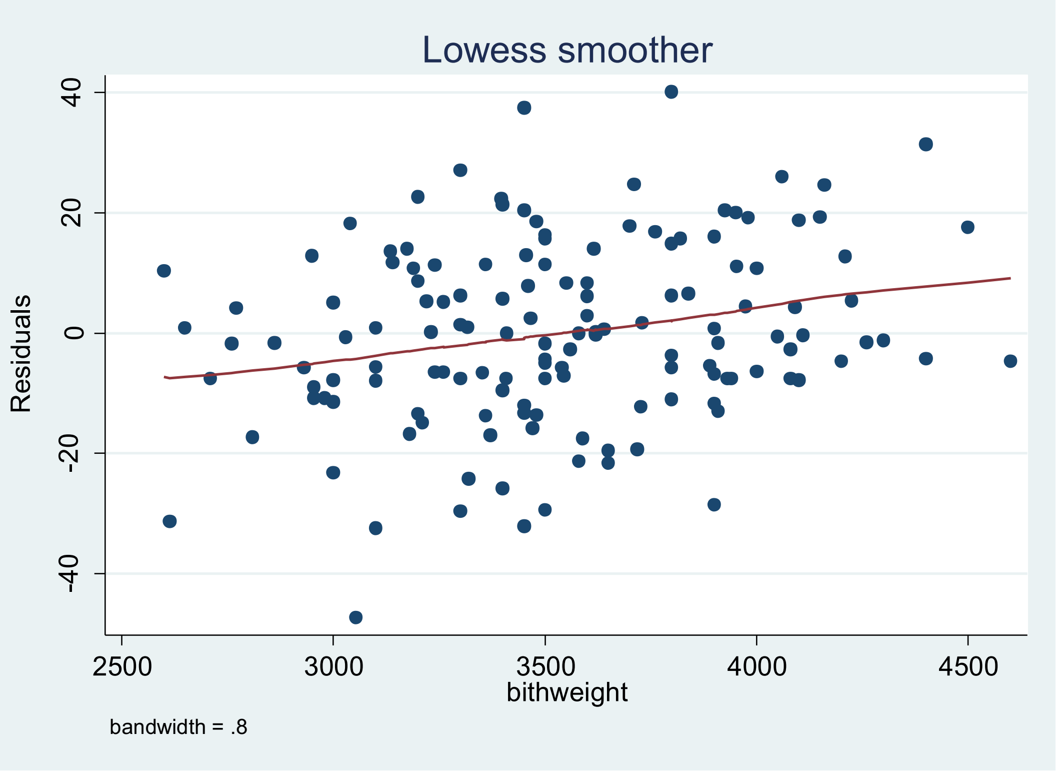
Recall: significant relationship

```
predict resids, residuals
lowess resids minutes
```



What about birth weight?

```
lowess resids birthwt
```



Results

- We found a significant relationship between minutes of CA and PDI
- Residual plots suggested a possible association with birth weight
- After accounting for minutes of CA, does birth weight improve our ability to predict PDI?

Results

```
regress pdi minutes birthwt
```

```
      Source |       SS       df       MS              Number of obs =     142
-------------+------------------------------           F(  2,   139) =    7.75
       Model |  3377.28384     2  1688.64192           Prob > F      =  0.0006
    Residual |  30281.3359   139  217.851337           R-squared     =  0.1003
-------------+------------------------------           Adj R-squared =  0.0874
       Total |  33658.6197   141  238.713615           Root MSE      =   14.76

------------------------------------------------------------------------------
         pdi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     minutes |   -.164923   .0566858    -2.91   0.004     -.277001   -.052845
     birthwt |   .0084129   .0029668     2.84   0.005     .0025471   .0142788
       _cons |   71.20498   10.62018     6.70   0.000       50.207   92.20297
------------------------------------------------------------------------------
```

Results

PDI = 71.2 – 0.165 minutes + 0.0084 birthwt

- For two infants with identical minutes of CA, a birth weight difference of 1000 grams would yield an 8.4 point change in predicted PDI score (*P* = 0.005)
- Not sensible to interpret intercept (71.2) here, as no birth weights are zero

Results

PDI = 71.2 – 0.165 minutes + 0.0084 birthwt

- The coefficient of minutes adjusting for birthwt (– 0.165) is fairly close to the unadjusted, or crude, coefficient (– 0.155); thus birthwt does not appreciably confound the association between minutes of CA and PDI

Confounding

- If an adjusted analysis gives an appreciably different result than a crude (unadjusted) analysis, we say the added variable is a confounder of the exposure-outcome association; use the adjusted analysis!
- Confounding bias can be large or small (and can even reverse direction of effect)
- Generally, define a confounder based on prior knowledge or biological reasoning

Confounding

- Leads to bias in your estimate of the exposure-outcome association if you fail to control for the effects of the confounder
- Can sometimes be controlled for in the analysis or avoided by design
- Confounder versus Covariate
- Is a bias, and worth avoiding!

Indicator Variables

- When there are categorical or binary predictor variables, we create indicator variables (or dummy variables or design variables)
- Examples: Diagnosis (IVS vs. VSD), Sex (F vs. M), Age group (< 1 mo, 1-2 mo, 3-9 mo)
- We create variables with numeric 0/1 coding

Results

```
regress pdi minutes vsd
predict yhat
gen yhativs=yhat if vsd==0
gen yhatvsd=yhat if vsd==1
gen pdiivs=pdi if vsd==0
gen pdivsd=pdi if vsd==1
sort minutes

scatter pdiivs pdivsd yhativs yhatvsd
minutes,xlabel(0(20)120) ylabel(50(2)150)
symbol (O T i i i) c(. . l l)
```

Results

```
regress pdi minutes vsd
```

```
      Source |       SS       df       MS              Number of obs =     142
-------------+------------------------------           F(  2,   139) =    5.02
       Model |  2266.77325     2  1133.38662           Prob > F      =  0.0079
    Residual |  31391.8465   139  225.840622           R-squared     =  0.0673
-------------+------------------------------           Adj R-squared =  0.0539
       Total |  33658.6197   141  238.713615           Root MSE      =  15.028

------------------------------------------------------------------------------
         pdi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     minutes |  -.1351676   .0587281    -2.30   0.023    -.2512836   -.0190516
         vsd |  -5.245585   3.112904    -1.69   0.094    -11.40035      .90918
       _cons |   101.0308   2.413607    41.86   0.000     96.25865    105.8029
------------------------------------------------------------------------------
```

Results

- Here vsd = 1 for VSD diagnosis, vsd = 0 for IVS diagnosis
- Fitted regression model is:

$$\text{PDI} = 101.0 - 0.135\ \text{minutes} - 5.25\ \text{vsd}$$

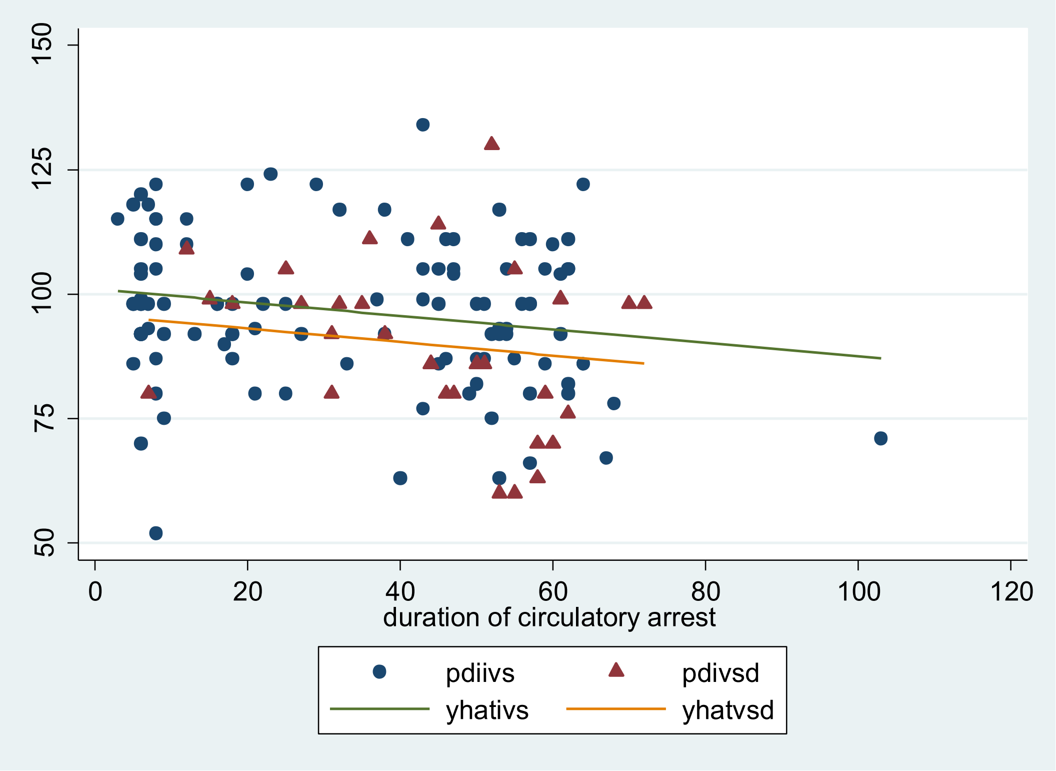
- For IVS, PDI = 101.0 – 0.135 minutes
- For VSD, PDI = 95.8 – 0.135 minutes
- Parallel lines, different intercepts

Results

$$\text{PDI} = 101.0 - 0.135 \text{ minutes} - 5.25 \text{ vsd}$$

- *P*-value for vsd effect is marginally significant (*P* = 0.09), and minutes is still significant (*P* = 0.023)
- Study surgeons and cardiologists thought that diagnosis was an important factor to consider

Results


150
125
100
75
50
0
20
40
60
80
100
120
duration of circulatory arrest
● pdiivs
▲ pdivsd
yhativs
yhatvsd

Results

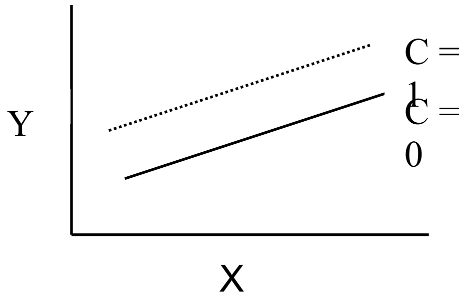
- In particular, infants with VSD (relative to IVS) were:
  - older at time of surgery
  - looked better preoperatively
  - had more complex surgeries with longer duration of CA
- Is diagnosis a confounder of the effect of minutes of CA?

Effect Modification

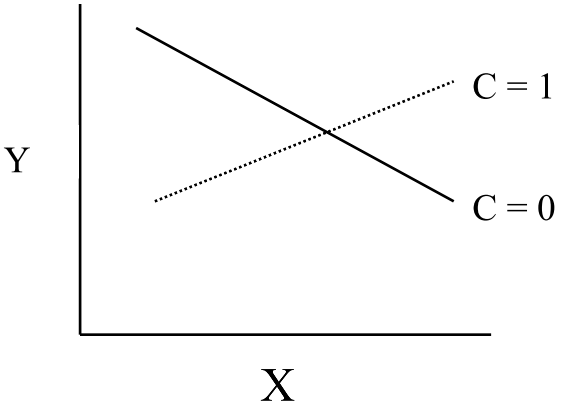
- It is not necessarily true that the effect of minutes of CA should be the same for both diagnosis groups
- Models including effect modification (or interaction) allow the effects of one variable to vary depending on the levels of another
- Modelled using product terms

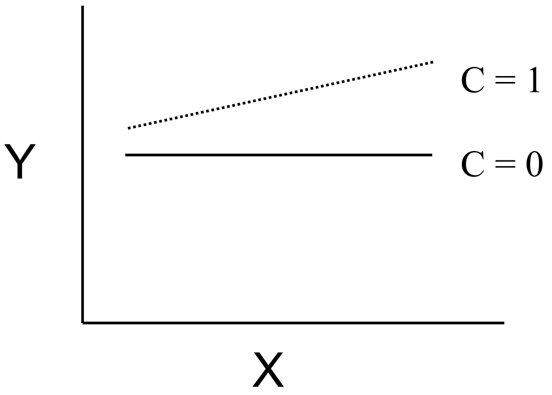
Effect Modification Review

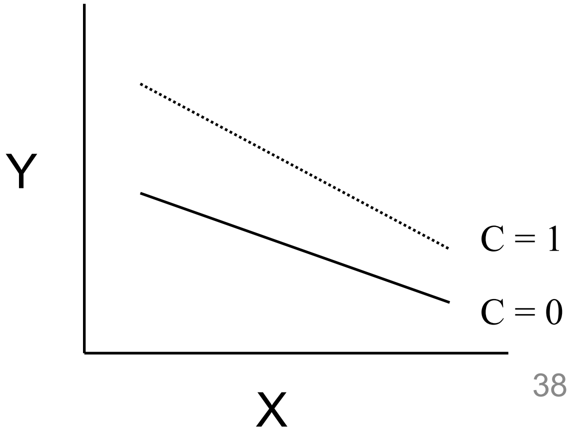
- Relationship between variable (X) and outcome (Y) differs by level of third variable (C)
- Example: No effect modification (parallel slopes)



- Example: effect modification (NOT parallel slopes)







Results

```
generate interact = minutes * vsd
regress pdi minutes vsd interact
```

```
      Source |       SS       df       MS              Number of obs =     142
-------------+------------------------------           F(  3,   138) =    3.62
       Model |  2456.25578     3  818.751927           Prob > F      =  0.0148
    Residual |  31202.3639   138  226.104087           R-squared     =  0.0730
-------------+------------------------------           Adj R-squared =  0.0528
       Total |  33658.6197   141  238.713615           Root MSE      =  15.037

------------------------------------------------------------------------------
         pdi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     minutes |  -.1141839   .0630748    -1.81   0.072     -.238902    .0105343
         vsd |   1.451135   7.950786     0.18   0.855    -14.26998    17.17225
    interact |  -.1588876    .173564    -0.92   0.362    -.5020763    .1843011
       _cons |   100.3351   2.531758    39.63   0.000     95.32905    105.3412
------------------------------------------------------------------------------
```

Results

$$\text{PDI} = 100.3 - 0.114\ \text{minutes} + 1.45\ \text{vsd} - 0.159\ \text{minutes} \cdot \text{vsd}$$

- For IVS, PDI = 100.3 – 0.114 minutes
- For VSD, PDI = 101.8 – 0.273 minutes
- Lines not parallel, though minutes effect is negative for both diagnosis groups

Results

$$\text{PDI} = 100.3 - 0.114\ \text{minutes} + 1.45\ \text{vsd} - 0.159\ \text{minutes} \cdot \text{vsd}$$

- *P*-value for the interaction is only 0.36, so no statistical evidence to support the interaction
- Reasonable to drop nonsignificant interactions, and to only test for those thought interesting in advance

Effect Modification Review

- Models including effect modification (or interaction) allow the effects of one variable to vary depending on the levels of another
  - No Interaction: $\mathbf{E(\mathit{Y}_i) = \beta_0 + \beta_1 \mathit{C}_i + \beta_2 \mathit{X}_i}$

    For C=0: $E(Y_i) = \beta_0 + \beta_2 X_i$ *For C=1:* $E(Y_i) = (\beta_0 + \beta_1) + \beta_2 X_i$

    -Different Intercepts and Same Slopes

  - Interaction: $\mathbf{E(\mathit{Y}_i) = \beta_0 + \beta_1 \mathit{C}_i + \beta_2 \mathit{X}_i + \beta_3 \mathit{C}_i \mathit{X}_i}$

    For C=0: $E(Y_i) = \beta_0 + \beta_2 X_i$

    For C=1: $E(Y_i) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_i$

    -Different Intercepts and Different Slopes

Coming Up

- Development of LS regression results
- Model fit assessment
- Residual analysis
- More multiple regression