Covariate Adjustment in Observational Studies

UC Irvine - ISI BUDS 2023

Presented July 24, 2023

Daniel L. Gillen
Chancellor's Professor and Chair
Department of Statistics
University of California, Irvine

Review of Linear Regression

Goal and assumptions

- Construct a model for the dependence of a response Y on predictors $X_1, X_2, ..., X_p$
  - Two components to the model:
    1. The *systematic* component (mean model)

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}$$

    2. The *random* component (error term)

$$Y_i = \mu_i + \epsilon_i, \quad \text{where } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Review of Linear Regression

Goal and assumptions

- Under this scenario, μ_i denotes the expected value of Y_i conditional on covariates $X_{1i}, X_{2i}, ..., X_{pi}$

$$E(Y_i | X_{1i}, X_{2i}, ..., X_{pi}) = E(\mu_i) + E(\epsilon_i) = \mu_i$$

- Question: Why is the assumption that $E(\epsilon_i) = 0$ a reasonable one in the above model formulation?

UCIrvine
University of California, Irvine

Parameter interpretation

- Consider the model

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

$$E[Y|X_1 = x_1 + 1, X_2, ..., X_p] - E[Y|X_1 = x_1, X_2, ..., X_p]$$

- In general, β_i is the expected (average) difference in Y for two populations with the same value for $x_k, k \neq i$ and whose value of x_i differs by 1 unit

Review of Linear Regression

Estimation of Model Parameters

- We consider parameter estimates that minimize the sum of squared errors

$$\sum_{i=1}^{n}(Y_i - \mu_i)^2 = \sum_{i=1}^{n}(Y_i - X_i\vec{\beta})^2$$

where X_i is the i^{th} row of the design matrix (the row vector of covariate values corresponding to the i^{th} observation) and $\vec{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$

- Why focus on the sum of squared errors?
  - It results in the MLE if $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
  - It is reasonable and mathematically convenient!

Review of Linear Regression

Estimation of Model Parameters

- To minimize the sum of squares with respect to $\vec{\beta}$,

$$\frac{\partial}{\partial \vec{\beta}} \sum_{i=1}^{n} (Y_i - X_i \vec{\beta})^2 \equiv 0$$

$$\Rightarrow \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \vec{\beta}} (Y_i - X_i \vec{\beta}) = 0$$

$$\Rightarrow \sum_{i=1}^{n} X_i (Y_i - X_i \vec{\beta}) = 0$$

$$\Rightarrow X^T Y - X^T X \vec{\beta} = 0$$

$$\Rightarrow \hat{\vec{\beta}} = (X^T X)^{-1} X^T Y$$

UCIrvine University of California, Irvine

Mean and Variance of the OLS Estimator

Mean of the OLS Estimator

- Proposition: $\hat{\vec{\beta}}$ is unbiased for $\vec{\beta}$ (ie. $\mathrm{E}[\hat{\vec{\beta}}] = \vec{\beta}$ or the average value of $\hat{\vec{\beta}}$'s computed across repeated experiments is the true $\vec{\beta}$)

Proof:

$$\begin{aligned} \mathrm{E}[\hat{\vec{\beta}}] &= \mathrm{E}[(X^T X)^{-1} X^T Y] \\ &= (X^T X)^{-1} X^T \mathrm{E}[Y] \\ &= (X^T X)^{-1} X^T X \vec{\beta} \\ &= \vec{\beta} \end{aligned}$$

Mean and Variance of the OLS Estimator

Variance of the OLS Estimator

- Proposition: If we assume constant variance in the errors then the variance of $\hat{\vec{\beta}}$ is given by

$$\begin{aligned}
\text{Var}[\hat{\vec{\beta}}] &= \text{Var}[(X^T X)^{-1} X^T Y] \\
&= (X^T X)^{-1} X^T \text{Var}[Y] X (X^T X)^{-1} \\
&= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

where $\widehat{\text{Var}}[\hat{\vec{\beta}}]$ is given by replacing σ^2 with

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$$

Types of Adjustment Variables

Effect modifiers (interaction terms)

- Suppose that we are interested in modeling the association between an outcome variable Y and a predictor X
- I tend to classify adjustment covariates into four broad categories (this terminology is not universal)
- Effect modifiers (interaction variables)
  - An effect modifier (W) is a covariate for which the association between the predictor of interest (X) *and* the outcome of interest (Y) differs with each level of W

UCIrvine
University of California, Irvine

Types of Adjustment Variables

Example: Effect modification

- Example: The association between gender and the risk of chd differs by systolic blood pressure

```
##
#####        Odds ratio describing the association between gender and CHD
#####        at 4 different SBP levels
##
------------------------------------
            OR (lower  95% upper)
[80,126]  0.39    0.32       0.48
(126,146] 0.43    0.34       0.54
(146,166] 0.60    0.43       0.82
(166,270] 0.74    0.48       1.16
------------------------------------
Mantel-Haenszel OR =0.46 95% CI ( 0.4,0.52 )
Test for heterogeneity: X^2( 3 ) = 9.62 ( p-value 0.0221 )
```

How do we deal with effect modifiers?

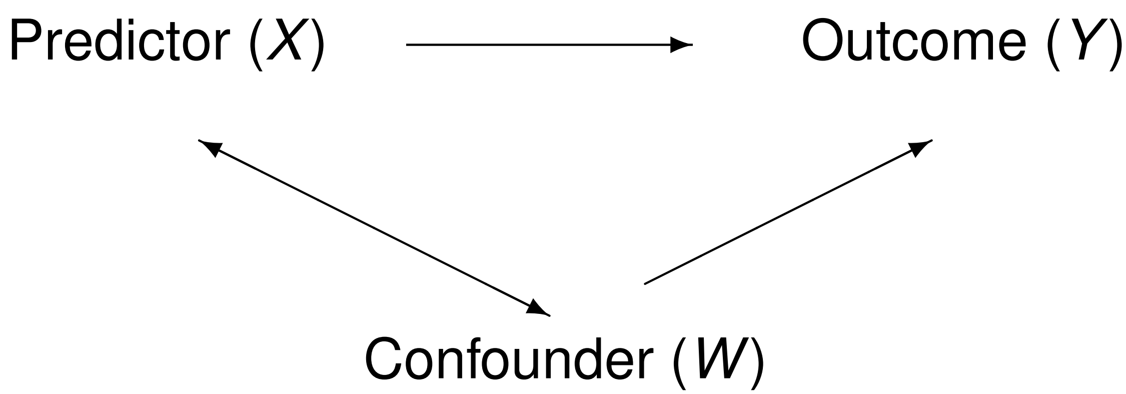
- Present stratified point estimates

UCIrvine
University of California, Irvine

Types of Adjustment Variables

Confounders

- One definition: A **confounder** is a variable that is associated with the predictor of interest (X) *and* causally related to the outcome of interest (Y).



Types of Adjustment Variables

Example: Confounding

- Example: Weight may be a confounder in the relationship between diabetes and blood pressure:
    - Diabetics tend to be heavier than non-diabetics
    - Increased weight is associated with higher blood pressure

How do we deal with confounding?

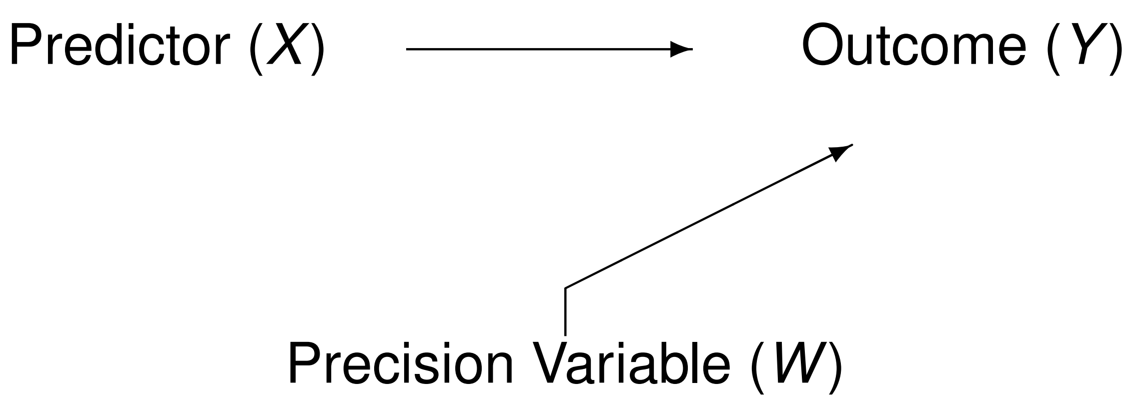
- Adjust for the confounder

UCIrvine
University of California, Irvine

Types of Adjustment Variables

Precision variables

- I define a **precision variable** as a covariate that is related to the outcome Y, but independent of the predictor of interest X.



Types of Adjustment Variables

Example: Precision variable

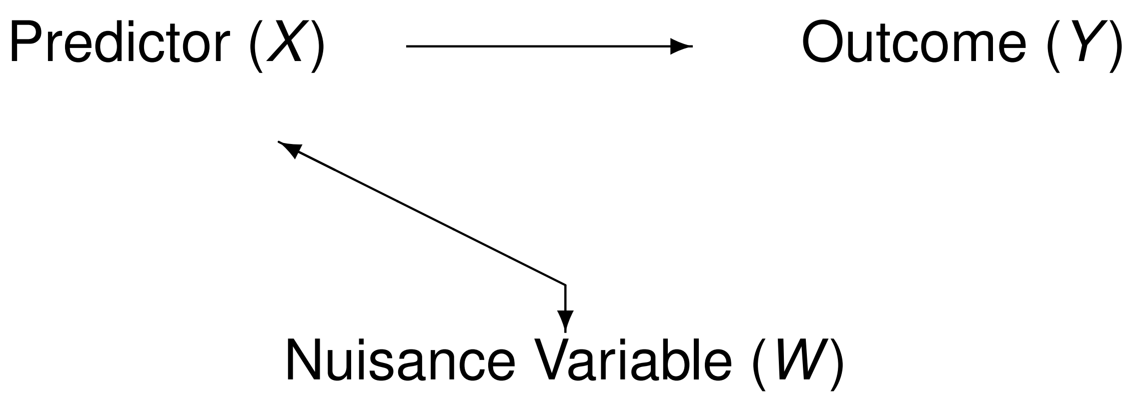
- Example: In a controlled experiment, we randomize patients to an experimental cancer treatment or placebo and look at the proportion of patients who relapse on each arm:
  - Age may be associated with the probability of relapse
  - Because of randomization, age is independent of whether treatment was received

Why precision?

- Why do I refer to this as a precision variable? Coming soon...
- In many cases, adjustment for a precision variable is a good idea!

Nuisance variables

- I define a **nuisance variable** as a covariate that is independent of the outcome Y, but may or may not be related to the predictor of interest X.



Types of Adjustment Variables

Example: Nuisance variable

- Example In a controlled experiment, we randomize patients to an experimental cancer treatment or placebo and look at the proportion of patients who relapse on each arm:
  - Shoe color on the day of randomization is not likely to be associated with the probability of relapse

Adjustment for nuisance parameters is not a good thing

- We are trying to model the outcome Y
- We do not want to intentionally include covariates that (we believe) are not associated with Y

Result of Confounding

Adjusted vs. unadjusted covariate effects

- Consider the following linear regression models:
  1. Unadjusted model: $\text{E}[Y_i] = \beta_0 + \beta_1 X_i$
     - β_1 is the difference in the mean of Y for groups differing by 1-unit in X
  2. Adjusted model: $\text{E}[Y_i] = \gamma_0 + \gamma_1 X_i + \gamma_2 W_i$
     - γ_1 is the difference in the mean of Y for groups differing by 1-unit in X, but agreeing in their value of W

UCIrvine
University of California, Irvine

Result of Confounding

Adjusted vs. unadjusted covariate effects

- Proposition 1: Let $\hat{\beta}_1$ denote the OLS estimate of β_1. Then under the adjusted model,

$$\mathrm{E}[\hat{\beta}_1] = \gamma_1 + \frac{\mathrm{cov}(X, W)}{\mathrm{var}(X)}\gamma_2$$

$$= \gamma_1 + r_{XW}\sqrt{\frac{\mathrm{var}(W)}{\mathrm{var}(X)}}\gamma_2$$

where r_{XW}, $\mathrm{var}(X)$, and $\mathrm{var}(W)$ are the sample correlation between X and W, sample variance of X, and sample variance of W, respectively.

UCIrvine
University of California, Irvine

Result of Confounding

The implication...

- $\hat{\beta}_1$ is biased (and inconsistent) for γ_1 unless at least one of the following hold

  1. $r_{XW} = 0$: X and W are uncorrelated (in the sample), OR
  2. $\gamma_2 = 0$: W is not related to Y

- In either case, $\hat{\beta}_1$ is unbiased (and consistent) for β_1

- Implication for confounders?

  - By definition, a confounder is related to the predictor of interest and the response

  - This implies that if W is a confounder, then both conditions above fail

  - Hence the parameter from the reduced model is biased for the adjusted estimate

UCIrvine
University of California, Irvine

Result of Confounding

The real question...

- We never know the 'true' model!
- Big Question: What do we *want* to hold constant when estimating the association between Y and X?
- The answer to this defines the interpretation of our result...

Precision of Estimators

Relationship between the precision of unadjusted and adjusted estimates

- Consider the following linear regression models:

1. Unadjusted model: $E[Y_i] = \beta_0 + \beta_1 X_i$
2. Adjusted model: $E[Y_i] = \gamma_0 + \gamma_1 X_i + \gamma_2 W_i$

Precision of Estimators

Relationship between the precision of unadjusted and adjusted estimates

- Proposition 2:
  1. For the unadjusted model,

$$\text{Var}[\hat{\beta}_1] = \frac{\sigma^2_{Y|X}}{n\text{var}(X)}$$

  2. For the adjusted model,

$$\text{Var}[\hat{\gamma}_1] = \frac{\sigma^2_{Y|X,W}}{n\text{var}(X)(1 - r^2_{XW})}$$

  where $\sigma^2_{Y|X,W} = \sigma^2_{Y|X} - \gamma_2^2 \text{var}(W|X)$

- Hence, if $\gamma_2 \neq 0$ then $\sigma^2_{Y|X,W} < \sigma^2_{Y|X}$

UCIrvine
University of California, Irvine

To adjust or not to adjust...

Implications of Propositions 1 & 2 (generalizeable to p coviarate case)

- Case 1: $r_{XW} = 0$ (X and W uncorrelated) and $\gamma_2 = 0$ (W and Y unrelated)
  - From Proposition 1, $\hat{\beta}_1$ unbiased for γ_1
  - From Proposition 2, $\text{Var}[\hat{\beta}_1] = \text{Var}[\hat{\gamma}_1]$
  - Conclusion: Lose 1 degree of freedom for hypothesis tests and CIs if adjusting for W

UCIrvine
University of California, Irvine

To adjust or not to adjust...

Implications of Propositions 1 & 2 (generalizeable to *p* coviarate case)

- Case 2: $r_{XW} \neq 0$ (X and W correlated) and $\gamma_2 = 0$ (W and Y unrelated)
  - From Proposition 1, $\hat{\beta}_1$ unbiased for γ_1
  - From Proposition 2, $\text{Var}[\hat{\beta}_1] < \text{Var}[\hat{\gamma}_1]$
  - Conclusion: Mathematically estimating the same quantity but *lose* precision when adjusting for W (nuisance variable)

To adjust or not to adjust...

Implications of Propositions 1 & 2 (generalizeable to p coviarate case)

- Case 3: $r_{XW} = 0$ (X and W uncorrelated) and $\gamma_2 \neq 0$ (W and Y related)
  - From Proposition 1, $\hat{\beta}_1$ unbiased for γ_1
  - From Proposition 2, $\text{Var}[\hat{\beta}_1] > \text{Var}[\hat{\gamma}_1]$
  - Conclusion: Mathematically estimating the same quantity but *gain* precision when adjusting for W (precision variable)

UCIrvine
University of California, Irvine

To adjust or not to adjust...

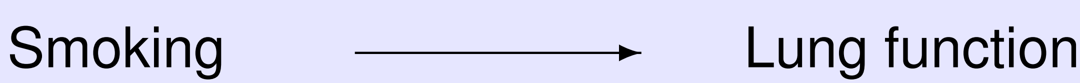
Implications of Propositions 1 & 2 (generalizeable to p coviarate case)

- Case 4: $r_{XW} \neq 0$ (X and W correlated) and $\gamma_2 \neq 0$ (W and Y related)
  - From Proposition 1, $\hat{\beta}_1$ biased for γ_1
  - From Proposition 2, no definitive statement about the variances
  - Conclusion: W is a confounder and decision to adjust should be based on what you are trying to estiamte.

Example - FEV Data

Is there an association between smoking and lung function in children?

- Scientific justification
  - Longterm smoking is associated with lower lung function
  - Are similar effects observed in short term smoking in children?
- Causal pathway of interest
  - Interested in whether smoking will cause a decrease in lung function



UCIrvine
University of California, Irvine

Example - FEV Data

Is there an association between smoking and lung function in children?

- Statistical analyses, however, can only detect associations between smoking and lung function
  - In a randomized trial, we could infer from the design that any association must be causal (not likely to happen)
  - In an observational study, we must try to isolate causal pathways of interest by adjusting for covariates

Example - FEV Data

Study design

- Observation study
  - Measurements obtained on a sample of 654 healthy children
  - Children were sampled while being seen for a regular checkup
  - Predictor of interest: Self-reported smoking
  - Response: FEV (Forced Expository Volume)
  - Additional covariates
    - Effect modifiers, potential confounders, precision variables

Example - FEV Data

Effect modifiers

- There are no covariates currently of scientific interest for their potential for effect modification
  - Might consider an age by smoking interaction (duration of exposure effect)
- Not generally advisable to go looking for different effects of smoking in subgroups before we have established that an effect exists overall
  - We may sometimes delay discovery of important facts, but most times this seems the logical strategy

Example - FEV Data

Potential confounders

- Necessary requirements for confounders
  - Associated causally with response
  - Associated with predictor of interest in sample
- Prior to looking at data, we cannot be sure of the second criterion
- Clearly, any strong predictor of the response has the potential to be a confounder
- Strategy: First consider known predictors of response
- Remember: In an observational study, known associations in the population will likely also be in the sample

UCIrvine
University of California, Irvine

Example - FEV Data

'Known' associations with smoking in the population

1. Height: Smoking may stunt growth
2. Age: Older children smoke
3. Gender: Girls smoke more than boys??? (used to be true)

- Bottom line
  - Comparing non-smokers to smokers of the same age will reduce a large amount of confounding
  - Comparing non-smokers to smokers of the same age and sex will reduce the majority of confounding

Example - FEV Data

Precision variables

- What about height?
  - In an observatinal study, all predictors of response should be considered potential confounders
  - Plus, we know that even if strong predictors of response are not confounding (i.e., not associated with the predictor of interest in the sample), we might want to consider adjusting the analysis to gain precision

Example - FEV Data

Precision variables

- Height is probably the strongest predictor of response that we have
    - The amount of air exhaled in 1 second (FEV) involves
        - Lung size (may not be of as much interest)
        - Lung function (probably more affected by smoking)
- Height is a reasonable surrogate for lung size
- Adjusting for height may allow comparisons that are more directly related to lung function

UCIrvine
University of California, Irvine

Example - FEV Data

Precision variables

- After adjusting for age, height is primarily a precision variable
  - After adjusting for age, there may be some residual confounding through any tendency for one sex to smoke more
- Note: If we adjust for height, we do lose one of the ways that smoking might have affected FEV
  - Smoking may stunt growth, which could lead to lower FEV

Example - FEV Data

Analysis plan

- Based on these issues, a priori we might plan an analysis adjusting for age and height (and sex?)
  - If that had not been specified a priori, I would perform the unadjusted analysis and then report the observed confounding from exploratory analyses

Example - FEV Data

Data analysis in R

- Let's implement our analysis plan a step at a time
- Start with recoding the data to make it more descriptive

```
> ##
> #####        FEV example
> ##
> #####        Preliminary data description and management
> ##
> summary( fev )
       id             age              fev             height        sex          smoke
 Min.    :  201   Min.    : 3.00   Min.    :0.791   Min.    :46.0   F:318   nosmoker:589
 1st Qu.:15811   1st Qu.: 8.00   1st Qu.:1.981   1st Qu.:57.0   M:336   smoker  : 65
 Median :36071   Median :10.00   Median :2.547   Median :61.5
 Mean   :37170   Mean   : 9.93   Mean   :2.637   Mean   :61.1
 3rd Qu.:53638   3rd Qu.:12.00   3rd Qu.:3.119   3rd Qu.:65.5
 Max.   :90001   Max.   :19.00   Max.   :5.793   Max.   :74.0
```

UCIrvine
University of California, Irvine

Example - FEV Data

```
## Recode gender so that it is intuitive
> fev$male <- as.numeric(fev$sex) - 1
> table( fev$sex, fev$male )

      0   1
  F 318   0
  M   0 336


## Recode smoking status so that it is intuitive
> fev$smoker <- as.numeric(fev$smoke) - 1
> table( fev$smoke, fev$smoker )

             0   1
  nosmoker 589   0
  smoker     0  65

## Drop 'sex' and 'smoke' from the dataset
> fev <- fev[ , !is.element(names(fev), c("sex", "smoke")) ]
```

Example - FEV Data

Data analysis in R

- Simple descriptive statistics and error checking

```
> summary( fev )
       id              age              fev            height           male           smoker      
 Min.   :  201   Min.   : 3.00   Min.   :0.791   Min.   :46.0   Min.   :0.000   Min.   :0.0000  
 1st Qu.:15811   1st Qu.: 8.00   1st Qu.:1.981   1st Qu.:57.0   1st Qu.:0.000   1st Qu.:0.0000  
 Median :36071   Median :10.00   Median :2.547   Median :61.5   Median :1.000   Median :0.0000  
 Mean   :37170   Mean   : 9.93   Mean   :2.637   Mean   :61.1   Mean   :0.514   Mean   :0.0994  
 3rd Qu.:53638   3rd Qu.:12.00   3rd Qu.:3.119   3rd Qu.:65.5   3rd Qu.:1.000   3rd Qu.:0.0000  
 Max.   :90001   Max.   :19.00   Max.   :5.793   Max.   :74.0   Max.   :1.000   Max.   :1.0000  
```

UCIrvine
University of California, Irvine

Example - FEV Data

Transformations for FEV and height

- Based upon the previously reported scientific relationship between FEV and its strongest predictor (height), we will log-transform both covariates
- Effects will be multiplicative (on median)

```
## Create log-transformed versions of FEV and height
> fev$logfev <- log( fev$fev )
> fev$loght <- log( fev$height )
```

Example - FEV Data

Restrict age of sample

- We will restrict our analyses to children 9 and older
  - The dataset included children as young as 3!
  - The youngest smoker was 9
- Dilemma
  - Younger children may help predict "normal" FEV, if our modeling of age and height is correct
  - If we are wrong, then we may not remove all confounding
- Reasoning behind decision
  - We only have 65 smokers, so that is the limiting factor in precision of our analysis
  - Having young nonsmokers does not add much

Example - FEV Data

Simple unadjusted analysis

- Use `lm()` to compute OLS estimates
- Use `subset` option to restrict dataset
- Use `lmCI()` on course webpage as one way to obtain CI's for parameter estimates

```
> ##
> #####        Unadjusted comparison of log-fev by smoking status
> ##
> fit.unadj <- lm( logfev ~ smoker, subset=age>=9, data=fev )
> summary( fit.unadj )$coef
             Estimate Std. Error t value     Pr(>|t|)
(Intercept)  1.05817   0.012806 82.6323 7.2253e-269
smoker       0.10231   0.033280  3.0741  2.2437e-03

> hist( fev$logfev[ fev$age >= 9 ] )

> ## Use the lmCI() function (in course code) as one way to obtain CI's
> ##
>
> lmCI( fit.unadj )
               Est ci95.lo ci95.hi t value Pr(>|t|)
(Intercept) 1.0582  1.0330  1.0833 82.6323   0.0000
smoker      0.1023  0.0369  0.1677  3.0741   0.0022
```

Example - FEV Data

Interpretation of smoking effect

- Note that our model is:

$$E[\log(\text{FEV})] = \beta_0 + \beta_1 I_{\text{smoker}}$$

- Common error is to assume that

$$E[\log(\text{FEV})] = \log(E[\text{FEV}])$$

  - In this case, we would (INCORRECTLY!) have that

$$\begin{aligned} & E[\log(\text{FEV} \mid \text{smoker=1})] - E[\log(\text{FEV} \mid \text{smoker=0})] \\ & = \log(E[\text{FEV} \mid \text{smoker=1}]) - \log(E[\text{FEV} \mid \text{smoker=0}]) \\ & = \log(E[\text{FEV} \mid \text{smoker=1}] \,/\, E[\text{FEV} \mid \text{smoker=0}]) \\ & = \beta_1 \end{aligned}$$

UCIrvine
University of California, Irvine

Example - FEV Data

Interpretation of smoking effect

- Thus, e^{β_1} *would* denote the ratio of mean FEV comparing a smoker to a non-smoker
- Problem: Jensen's inequality says that $E[g(X)] \geq g(E[X])$ for a convex function g. If g is concave (eg. $g(x) = \log(x)$), then $E[g(X)] \leq g(E[X])$.
- One way to get around this is to interpret the medians for each group
  - Note that if the distribution of $\log(FEV)$ is (roughly) symmetric then we have

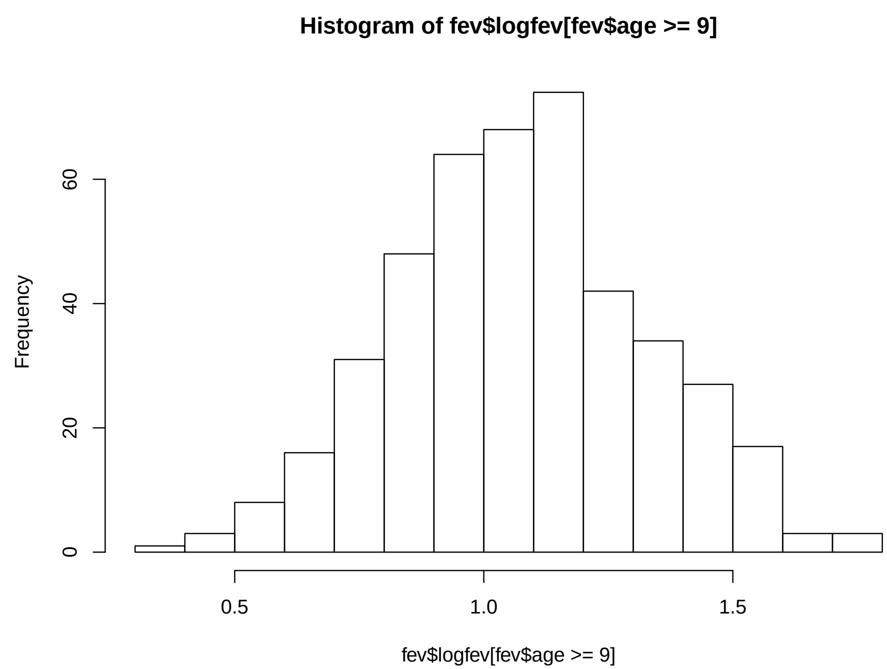
$$E[\log(\text{FEV})] \approx \text{median}[\log(\text{FEV})] = \log(\text{median}[\text{FEV}])$$

Example - FEV Data

Interpretation of smoking effect

- Let's look at the distribution of log(FEV):

```
. hist logfev if age>=9
```



- This is pretty symmetric, which allows us to interpret the effect of smoking on the ratio of medians

Example - FEV Data

Interpretation of smoking effect

- Use the `lmCI()` function on the course webpage to exponentiate the coefficient for smoking and CI

```
> ## Again, use lmCI() but with the expcoef=TRUE option
> lmCI( fit.unadj, expcoef=TRUE )
             exp( Est ) ci95.lo ci95.hi t value Pr(>|t|)
(Intercept)      2.8811  2.8095  2.9545 82.6323   0.0000
smoker           1.1077  1.0376  1.1826  3.0741   0.0022
```

- Interpretation: The median FEV of a smoker is estimated to be 10.8% higher than that of a non-smoker (95% CI: 1.04, 1.18). This difference is statistically significant $p = 0.002$.

UCIrvine
University of California, Irvine

Example - FEV Data

Adjustment for age

- The finding that smokers have better lung function is quite unintuitive and is likely due to confounding by age.
- Let's adjust for age in our analysis and look at the effect of smoking

```
> ##
> #####        Comparison of log-fev by smoking status with adjustment for age
> ##
> fit.age <- lm( logfev ~ smoker + age, subset=age>=9, data=fev )

> summary( fit.age )$coef
              Estimate Std. Error t value   Pr(>|t|)
(Intercept)  0.351817  0.0545238  6.4525 2.9309e-10
smoker      -0.051349  0.0304568 -1.6860 9.2516e-02
age          0.063596  0.0048111 13.2185 8.8010e-34


> lmCI( fit.age, expcoef=TRUE )
            exp( Est ) ci95.lo ci95.hi t value Pr(>|t|)
(Intercept)     1.4216  1.2772  1.5825  6.4525   0.0000
smoker          0.9499  0.8948  1.0085 -1.6860   0.0925
age             1.0657  1.0556  1.0758 13.2185   0.0000
```

Example - FEV Data

Adjustment for age

- Interpretation of smoking (age adjusted): The median FEV of a smokers is estimated to be 5.0% lower than that of non-smokers similar in age (95% CI: 0.90, 1.01). This difference is not statistically significant at the .05 level ($p = 0.093$).

- Interpretation of age (smoking adjusted): Median FEV is estimated to be 6.6% higher for each year difference in age between two groups with similar smoking status (95% CI: 1.06 to 1.08) This difference is statistically significant ($p < 0.001$).

Example - FEV Data

Comparison of unadjusted and age adjusted analyses

- Marked difference in effect of smoking suggests that there was indeed confounding
- Age is a relatively strong predictor of FEV
- Age is associated with smoking in the sample
  - Mean (SD) of age in analyzed nonsmokers: 11.1 (2.04)
  - Mean (SD) of age in analyzed smokers: 13.5 (2.34)
- Effect of age adjustment on precision
  - Lower Root MSE (.209 vs .248) tends to increase precision of estimate of smoking effect
  - Association between smoking and age tends to lower precision
  - Net effect: Slightly increased precision (SE 0.031vs 0.033)

Example - FEV Data

Adjustment for age and height

- After adjustment for age, height should have little association with smoking status but is still likely to have an association with FEV.
- Plan is to adjust for log(height) as a precision variable.

```
> ##
> #####        Additional adjustment for loght as a precision variable
> ##
> fit.adj <- lm( logfev ~ smoker + age + loght, subset=age>=9, data=fev )
> summary( fit.adj )$coef
              Estimate  Std. Error   t value    Pr(>|t|)
(Intercept) -11.094618  0.5201258 -21.3306 1.2784e-69
smoker       -0.053590  0.0209462  -2.5584 1.0852e-02
age           0.021529  0.0038187   5.6379 3.1014e-08
loght         2.869659  0.1300580  22.0645 6.0112e-73

> lmCI( fit.adj, expcoef=TRUE )
            exp( Est ) ci95.lo ci95.hi  t value Pr(>|t|)
(Intercept)     0.0000  0.0000  0.0000 -21.3306   0.0000
smoker          0.9478  0.9096  0.9877  -2.5584   0.0109
age             1.0218  1.0141  1.0295   5.6379   0.0000
loght          17.6310 13.6541 22.7663  22.0645   0.0000
```

Example - FEV Data

Adjustment for age and height

- Interpretation of smoking (age and height adjusted): The median FEV of smokers is estimated to be 5.2% lower than that of non-smokers similar in age and height (95% CI: 0.91, 0.99). This difference is statistically significant at the .05 level ($p = 0.011$).

- Interpretation of age (smoking and height adjusted): Median FEV is estimated to be 2.2% higher for each year difference in age between two groups with similar smoking status and similar in height (95% CI: 1.01 to 1.03) This difference is statistically significant ($p < 0.001$).

Example - FEV Data

Comparison of age and age-height adjusted analyses

- No difference in effect of smoking suggests there was no more confounding after age adjustment
- Marked difference in the effect of age on FEV, suggesting confounding by height, but there is still an independent effect of age.
- Effect of height adjustment on precision
  - Lower Root MSE (.144 vs .209) would tend to increase precision of estimate of smoking effect
  - Little association between smoking and height after adjustment for age will not tend to lower precision
  - Net effect: Much greater precision (SE 0.021 vs 0.031)

Example - FEV Data

Adjustment for age, height, and gender

- Is there residual confounding by gender?

```
> ##
> #####        Additional adjustment for loght as a (potential?) precision variable
> ##
> fit.gender <- lm( logfev ~ smoker + age + loght + male, subset=age>=9, data=fev )
> summary( fit.gender )$coef
              Estimate Std. Error  t value   Pr(>|t|)
(Intercept) -10.895107  0.5567732 -19.5683 1.3767e-61
smoker       -0.050883  0.0211187  -2.4094 1.6395e-02
age           0.022117  0.0038633   5.7250 1.9328e-08
loght         2.818043  0.1398450  20.1512 3.1515e-64
male          0.014977  0.0149137   1.0042 3.1583e-01

> lmCI( fit.gender, expcoef=TRUE )
            exp( Est ) ci95.lo ci95.hi  t value Pr(>|t|)
(Intercept)     0.0000  0.0000  0.0001 -19.5683   0.0000
smoker          0.9504  0.9117  0.9907  -2.4094   0.0164
age             1.0224  1.0146  1.0302   5.7250   0.0000
loght          16.7441 12.7201 22.0409  20.1512   0.0000
male            1.0151  0.9858  1.0453   1.0042   0.3158
```

Adjustment for age, height, and gender

- Interpretation of smoking: The median FEV of smokers is estimated to be 5.0% lower than that of non-smokers similar in age, height, and gender (95% CI: 0.91, 0.99). This difference is statistically significant at the .05 level ($p = 0.016$).

- Interpretation of age: Median FEV is estimated to be 2.2% higher for each year difference in age between two groups with similar smoking status and similar in height and gender (95% CI: 1.02 to 1.03) This difference is statistically significant ($p < 0.001$).

- Interpretation of gender: The median FEV of males is estimated to be 1.5% higher than that of females similar in smoking status, height, and age (95% CI: 0.99, 1.05). This difference is not statistically significant ($p = 0.316$).

Example - FEV Data

Comparison of age/height and age/height/gender adjusted analyses

- No suggestion of further confounding by sex
- Effect of sex adjustment on precision
  - Root MSE (.144 vs .144) suggests that sex adds virtually no precision to the model

Example - FEV Data

Final comments

- Choosing the model for analysis
  - Confirmatory vs Exploratory analyses
  - Every statistical model answers a different question
  - Data driven choice of analyses requires later confirmatory analyses

Example - FEV Data

Final comments

- Best strategy
  - Choose appropriate primary analysis based on scientific question identified a priori
  - Provide most robust statistical inference regarding this question (still to come)
  - Further explore your data to generate new hypotheses and speculate on mechanism
    - Regard these statistics as descriptive