Lecture 6

Case Study - Role of Adjustment Variables
Statistics 211 - Statistical Methods II

Presented January 31, 2018

Dan Gillen
Department of Statistics
University of California, Irvine
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

![Diagram showing Smoking leading to Lung function]
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
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 observational 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
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
### 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

```r
> ##
> ###### 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

## 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
```
Example - FEV Data

Data analysis in R

- Simple descriptive statistics and error checking

```r
> summary(fev)

   id      age     fev   height     smoke       male
Min. : 201 Min. : 3.00 Min. :0.791 Min. :46.0 nosmoker:589 Min. :0.000
1st Qu.:15811 1st Qu.: 8.00 1st Qu.:1.981 1st Qu.:57.0 smoker : 65 1st Qu.:0.000
Median :36071 Median :10.00 Median :2.547 Median :61.5 Median :1.000
Mean :37170 Mean : 9.93 Mean :2.637 Mean :61.1 Mean :0.514
3rd Qu.:53638 3rd Qu.:12.00 3rd Qu.:3.119 3rd Qu.:65.5 3rd Qu.:1.000
Max. :90001 Max. :19.00 Max. :5.793 Max. :74.0 Max. :1.000
```
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

```r
# Unadjusted comparison of mean fev by smoking status
> fit.unadj <- lm( fev ~ smoke, subset=age>=9, data=fev )
> summary( fit.unadj )$coef

              Estimate Std. Error    t value  Pr(>|t|)
(Intercept) 2.973100   0.039415  75.43150 1.072e-252
smokesmoker 0.303760   0.102431   2.96553 3.187e-03

> lmCI( fit.unadj )

         Est ci95.lo ci95.hi   t value Pr(>|t|)
(Intercept) 2.9731  2.8956  3.0506  75.4315  1.072e-252
smokesmoker 0.3038  0.1024  0.5051   2.9655   3.187e-03
```
Unadjusted estimate

- Interpretation of smoking (unadjusted model): The mean FEV of a smoker is estimated to be 0.3038 liters/sec higher than that of a non-smoker (95% CI: 0.1024, 0.5051). This difference is statistically significant \( p = 0.0032 \).
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.

```r
> ##
> ###### Comparison of mean fev by smoking status with adjustment for age
> ##
> fit.age <- lm( fev ~ smoke + age, subset=age>=9, data=fev )
> summary( fit.age )$coef

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 0.7669   | 0.166806   | 4.5963  | 5.6411e-06 |
| smokesmoker    | -0.1762  | 0.093178   | -1.8911 | 5.9279e-02 |
| age            | 0.19865  | 0.014719   | 13.4964 | 6.2228e-35 |

> lmCI( fit.age )

|                | Est     | ci95.lo | ci95.hi | t value | Pr(>|t|) |
|----------------|---------|---------|---------|---------|----------|
| (Intercept)    | 0.7667  | 0.4388  | 1.0945  | 4.5963  | 0.0000   |
| smokesmoker    | -0.1762 | -0.3593 | 0.0069  | -1.8911 | 0.0593   |
| age            | 0.1987  | 0.1697  | 0.2276  | 13.4964 | 0.0000   |

> cbind( summary(fit.age)$sigma, summary(fit.unadj)$sigma )

<table>
<thead>
<tr>
<th></th>
<th>[,1]</th>
<th>[,2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>0.64089</td>
<td>0.76224</td>
</tr>
</tbody>
</table>
```
### Adjustment for age

- **Interpretation of smoking (age adjusted):** The mean FEV of a smokers is estimated to be 0.1762 liters/sec less than that of non-smokers similar in age (95% CI: -0.3593, 0.0069). This difference is not statistically significant at the .05 level ($p = 0.0593$).

- **Interpretation of age (smoking adjusted):** Mean FEV is estimated to be 0.1987 liters/sec higher for each year difference in age between two groups with similar smoking status (95% CI: 0.1697, 0.2276) 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 (0.64089 vs 0.76224) tends to increase precision of estimate of smoking effect
  - Association between smoking and age tends to lower precision
  - Net effect: Slightly increased precision (SE 0.093178 vs 0.102431)
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 height as a precision variable.

```r
> ##
> ###### Additional adjustment for height as a precision variable
> ##
> fit.adj <- lm( fev ~ smoke + age + height, subset=age>=9, data=fev )
> summary( fit.adj )$coef

  Estimate Std. Error   t value  Pr(>|t|)
(Intercept)  -6.431311  0.3559590  -18.0676 7.6048e-55
smokesmoker   -0.178429  0.0651068   -2.7406 6.3862e-03
   age         0.071201  0.0118844    5.9911 4.3738e-09
   height        0.135303  0.0063222   21.4012 6.1162e-70

> lmCI( fit.adj )

  Est ci95.lo   ci95.hi   t value  Pr(>|t|)
(Intercept) -6.4313 -7.1309 -5.7317  -18.0676 7.6048e-55
smokesmoker   -0.1784 -0.3064 -0.0505  -2.7406 6.3862e-03
  age          0.0712  0.0478  0.0946    5.9911 4.3738e-09
  height      0.1353  0.1229  0.1477   21.4012 6.1162e-70

> cbind( summary(fit.adj)$sigma, summary(fit.age)$sigma )
    [,1]    [,2]
[1,]  0.44782  0.64089
```
**Example - FEV Data**

**Adjustment for age and height**

- Interpretation of smoking (age and height adjusted): The mean FEV of smokers is estimated to be 0.1784 liters/sec less than that of non-smokers similar in age and height (95% CI: -0.3064, -0.0505). This difference is statistically significant at the .05 level \( (p = 0.0064) \).
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 (0.44782 vs 0.64089) 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 0.0651068 vs 0.093178).
Example - FEV Data

Adjustment for age, height, and gender

► Is there residual confounding by gender?

```r
> ##
> ####### Additional adjustment for gender as a (potential?) precision variable
> ##
> fit.gender <- lm( fev ~ smoke + age + height + male, subset=age>=9, data=fev )
> summary( fit.gender )$coef

Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.142998 0.3729172 -16.4728 1.0465e-47
smokesmoker -0.158084 0.0652631 -2.4223 1.5833e-02
age 0.075881 0.0119694 6.3396 5.7833e-10
height 0.128974 0.0067947 18.9816 6.1909e-59
male 0.113691 0.0463341 2.4537 1.4531e-02

> lmCI( fit.gender )

Est ci95.lo ci95.hi t value Pr(>|t|)
(Intercept) -6.1430 -6.8759 -5.4100 -16.4728 0.0000
smokesmoker -0.1581 -0.2864 -0.0298 -2.4223 0.0158
age 0.0759 0.0524 0.0994 6.3396 0.0000
height 0.1290 0.1156 0.1423 18.9816 0.0000
male 0.1137 0.0226 0.2048 2.4537 0.0145

> cbind( summary(fit.gender)$sigma, summary(fit.adj)$sigma )

[,1]     [,2]
[1,] 0.44525 0.44782
```
Example - FEV Data

Adjustment for age, height, and gender

- Interpretation of smoking: The mean FEV of smokers is estimated to be 0.1581 liters/sec less than that of non-smokers similar in age, height, and gender (95% CI: -0.2864, -0.0298). This difference is statistically significant at the .05 level ($p = 0.0158$).
Example - FEV Data

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

- No real suggestion of further confounding by sex

- Effect of sex adjustment on precision
  - Root MSE (0.44525 vs 0.44782) 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
Remedies For Nonconstant Variance

What can we do in the linear regression scenario?

- There are three ways that we can address the nonconstant variance problem:

  1. Transform the outcome variable
     - Basically a trial and error approach
     - Typical choices for transforms include log or square-root
     - See HW 1

  2. Robust variance estimators
     - Try to ‘fix up’ our estimate of the variance

  3. Weighted least squares
     - If $\text{Var}(\epsilon_i) \propto 1/\sigma_i^2$ then $\text{Var}(\sqrt{w_i} \epsilon_i) \propto 1$ if $w_i = 1/\sigma_i^2$
     - In this case we can consider the model
       $$\sqrt{w_i} Y_i = \sqrt{w_i} \tilde{X}_i^T \beta + \sqrt{w_i} \epsilon_i$$
Remedies For Nonconstant Variance

Robust Variance Estimators

► Suppose that $\text{Var}[\vec{Y}] = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2) = \Sigma$. Then the usual variance estimate of the OLS estimator

$$
\hat{\text{Var}}[\hat{\beta}] = \hat{\sigma}^2 (X^T X)^{-1}
$$

is not correct.

► The true variance of $\hat{\beta}$ is given by

$$
\text{Var}[\hat{\beta}] = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}
$$

where

$$
\Sigma = 
\begin{pmatrix}
\sigma_1^2 & 0 & \ldots & 0 \\
0 & \sigma_2^2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \sigma_n^2
\end{pmatrix}
$$
Remedies For Nonconstant Variance

Robust Variance Estimators

- Plan: Estimate \( \Sigma \) with \( \hat{\Sigma} \) and plug-in to obtain:

\[
\hat{\text{Var}}[\hat{\beta}] = (X^T X)^{-1} X^T \hat{\Sigma} X (X^T X)^{-1}
\]

- A natural estimator of \( \sigma_i^2 \) is the squared residual \( e_i^2 = (Y_i - \hat{Y}_i)^2 \), so that we could take

\[
\hat{\Sigma} = \begin{pmatrix}
  e_1^2 & 0 & \cdots & 0 \\
  0 & e_2^2 & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & e_n^2
\end{pmatrix}
\]
Remedies For Nonconstant Variance

Robust Variance Estimators

- This is called the
  - *Huber-White* or
  - *sandwich* or
  - *robust* variance estimator

- The function `robust.se.lm` on the course webpage will compute the robust variance estimator for linear regression models

- Because we are estimating the entire covariance matrix of $Y$ (as opposed to a single quantity when we assume constant variance) the robust variance estimator requires a slightly larger sample use to produce reliable estimates (see HW 3!)

- Also, by Gauss-Markov we know that OLS will not be BLUE under heteroscedasticity...
Ex: Robust Variance Estimator Applied to FEV Analysis

▶ Again, consider the FEV analysis where our final model adjusted model included the indicator of smoking, age, and height

▶ Let’s examine the model for heteroscedasticity...

```
##
#### Remedies For Nonconstant Variance
####

Ex: Robust Variance Estimator Applied to FEV Analysis

▶ Again, consider the FEV analysis where our final model adjusted model included the indicator of smoking, age, and height

▶ Let’s examine the model for heteroscedasticity...

```
```"
Remedies For Nonconstant Variance

Ex: Robust Variance Estimator Applied to FEV Analysis
Ex: Robust Variance Estimator Applied to FEV Analysis

- Signs of heteroscedasticity implying incorrect variance estimates
- Let’s apply the robust variance estimator and compare with the model based variance estimates

```r
> summary(fit.adj)$coef

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -6.431311| 0.3559590  | -18.0676| 7.6048e-55 |
| smokersmoker   | -0.178429| 0.0651068  | -2.7406 | 6.3862e-03 |
| age            | 0.071201  | 0.0118844  | 5.9911  | 4.3738e-09 |
| height         | 0.135303  | 0.0063222  | 21.4012 | 6.1162e-70 |

> robust.se.lm(fit.adj)

|                | Estimate | Robust SE | ci95.lo  | ci95.hi  | t value | Pr(>|t|) |
|----------------|----------|-----------|----------|----------|---------|----------|
| (Intercept)    | -6.431311| 0.3778403 | -7.17393 | -5.688692| -17.0212| 3.6652e-50|
| smokersmoker   | -0.178429| 0.0785244 | -0.33276 | -0.024094| -2.2723 | 2.3558e-02|
| age            | 0.071201  | 0.0123028 | 0.04702  | 0.095381 | 5.7874  | 1.3697e-08|
| height         | 0.135303  | 0.0065733 | 0.12238  | 0.148222 | 20.5837 | 3.1415e-66|
```
Remedies For Nonconstant Variance

Ex: Robust Variance Estimator Applied to FEV Analysis

- We can see that the robust variances are larger for all covariates
  - This is not always true
  - The direction of the correction depends on the relative sample size and variance within each subpopulation being compared
- The resulting confidence intervals incorporate the robust variance estimate as well
Remedies For Nonconstant Variance

Weighted least squares

> Again, suppose that $\text{Var}[\vec{Y}] = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2) = \Sigma$

> Consider minimizing the weighted sum of squared errors in order to obtain parameter estimates:

$$\sum_{i=1}^{n} w_i (Y_i - \mu_i)^2 = \sum_{i=1}^{n} w_i (Y_i - \hat{X}_i^T \hat{\beta})^2$$

where $w_i = 1/\sigma_i^2$.

> This leads to the weighted least squares estimator given by

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

where $W = \text{diag}(w_1, \ldots, w_n)$. 
Remedies For Nonconstant Variance

Weighted least squares

- The variance of $\hat{\beta}_W$ is given by

$$\text{Var}[\hat{\beta}_W] = \text{Var}[(X^T W X)^{-1} X^T W \tilde{Y}]$$

$$= (X^T W X)^{-1} X^T W \text{Var}[\tilde{Y}] W X (X^T W X)^{-1}$$

$$= (X^T W X)^{-1} X^T W \Sigma W X (X^T W X)^{-1}$$

- In general, if we choose $w_i \propto 1/\sigma_i^2$, then

$$\text{Var}[\hat{\beta}_W] \propto (X^T W X)^{-1}$$

- In this case, the standard errors computed via usual (weighted) least squares will be ‘correct’. (why?)
Remedies For Nonconstant Variance

How well does WLS ‘work’?

- When we choose $w_i = 1/\sigma_i^2$, we also get back all of the conclusion we had with the OLS estimator under the assumptions of independent, mean zero, constant variance errors
  - That is, $\hat{\beta}_W$ is a best linear unbiased estimator

- Problem: How often do we know $\sigma_i^2$?
Remedies For Nonconstant Variance

Could try to estimate $\sigma_i^2$ ...

- $\hat{\beta}_W$ is unbiased for any fixed choice of weights, even if $w_i \neq 1/\sigma_i^2$

- If $w_i \neq 1/\sigma_i^2$ then standard errors are not correct and the estimator is less efficient (higher variance than other unbiased estimators)

  - General rule: Getting the weights roughly correct will give roughly correct standard errors and a nearly efficient estimator

- How to estimate it $\sigma_i^2$? If $\sigma_i^2 = V(\mu_i)$ (ie. only dependent upon the mean of $Y_i$) for some smooth function then we can

  1. Estimate $V(\mu)$ by grouping $\mu_i$ into quantiles and taking $\hat{V}(\mu)$ to be the variance of the residuals in that quantile
  2. Smooth $\epsilon_i^2$ against $\hat{\mu}_i$ to get $\hat{V}(\mu)$