Modeling Goals and Purposes

There are two goals of regression models, both equally important:

1. Prediction
   - want the model to fit data well
   - want replicability
   - mechanism is not important
2. Explanation
   - need accurate estimates of coefficients
   - the "correct" form of the model is one goal in itself
   - fitted model may be used for important policy decisions

Thus far, we have focused on models with explanation in mind, proposing and testing models that made sense and had plausibility we could justify on the basis of biological, ecological, or socio-economic theories.

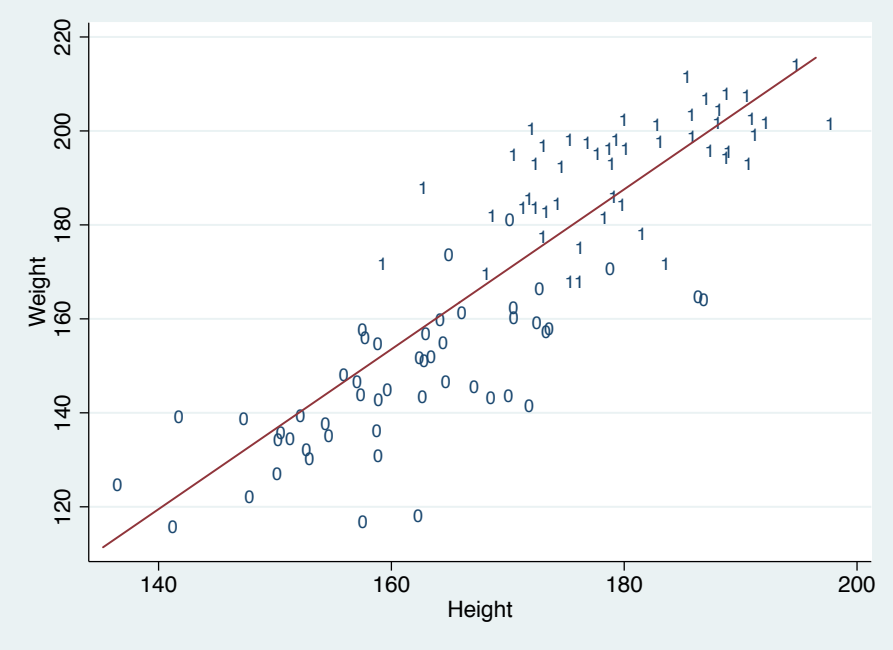
Control of confounding is key in explanatory modeling.

Confounding

- Again, a confounder is a variable that is related to the predictor as well as the response (even in the absence of the predictor). It is not 'caused' by the predictor

  For example, examine the weight-height relationship in a dataset: the marginal relationship is given with the following regression line:

```
. regress w h
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  264.06
       Model |  51652.0816     1  51652.0816           Prob > F      =  0.0000
    Residual |   19169.393    98  195.606051           R-squared     =  0.7293
-------------+------------------------------           Adj R-squared =  0.7266
       Total |  70821.4746    99   715.36843           Root MSE      =  13.986
------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           h |   1.701886   .1047316    16.25   0.000      1.49405    1.909723
       _cons |  -118.7635   17.78881    -6.68   0.000    -154.0649   -83.46223
------------------------------------------------------------------------------
.  twoway (scatter w h, mlabel(sex) msymbol(none))  (lfit w h), xtitle("Height")
 ytitle("Weight") legend(off)
```



- males (symbol=1) and females (symbol=0) form somewhat distinct groups with respect to height. What do we know about the relationship between sex and weight? Males tend to be heavier (for any given height) than women. Males also tend to be taller.

Confounding

- The slopes for these subgroups might also be different. So, is sex is a potential confounder?

- Add sex to the model:

```
. regress w h sex
```

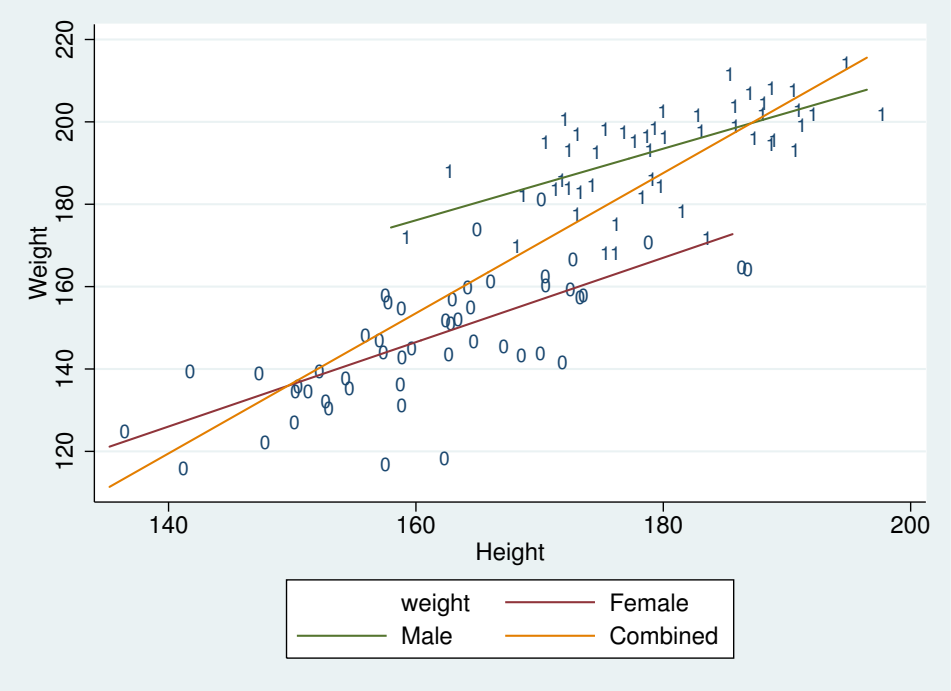
```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =  311.78
       Model |  61287.6579     2   30643.829           Prob > F      =  0.0000
    Residual |   9533.8167    97  98.2867701           R-squared     =  0.8654
-------------+------------------------------           Adj R-squared =  0.8626
       Total |  70821.4746    99   715.36843           Root MSE      =   9.914
------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           h |   .9639984   .1051921     9.16   0.000     .7552211    1.172776
         sex |   27.81749   2.809484     9.90   0.000     22.24144    33.39354
       _cons |  -7.728952   16.87486    -0.46   0.648    -41.22088    25.76298
------------------------------------------------------------------------------
```

- How much does slope change? Less effect at 0.963 now vs. 1.702

Correction of Confounding

- The similar but less steep slopes in men and women separately:

```
. twoway (scatter w h, mlabel(sex) msymbol(none)) (lfit w h if (sex==0))
 (lfit w h if (sex==1)) (lfit w h), xtitle("Height")
ytitle("Weight")  legend(order(1 "Weight" 2 "Female" 3 "Male" 4 "Combined"))
```



Correction of Confounding

- We can also examine the relationship of height to weight separately by sex, performing a **stratified analysis**

```
. sort sex
. by sex: reg w h

-------------------------------------------------------------------------------
-> sex = 0

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   50.56
       Model |  5692.98528     1  5692.98528           Prob > F      =  0.0000
    Residual |  5404.60962    48  112.596034           R-squared     =  0.5130
-------------+------------------------------           Adj R-squared =  0.5028
       Total |  11097.5949    49  226.481529           Root MSE      =  10.611
------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           h |   1.024336   .1440569     7.11   0.000     .7346905    1.313982
       _cons |  -17.37486   23.07847    -0.75   0.455    -63.77722    29.02751
------------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------
-> sex = 1

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   30.74
       Model |  2612.09064     1  2612.09064           Prob > F      =  0.0000
    Residual |  4078.44897    48  84.9676868           R-squared     =  0.3904
-------------+------------------------------           Adj R-squared =  0.3777
       Total |   6690.5396    49  136.541625           Root MSE      =  9.2178

------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           h |   .8692904   .1567825     5.54   0.000      .554058    1.184523
       _cons |   37.02104   28.06089     1.32   0.193    -19.39915    93.44124
------------------------------------------------------------------------------
```

- Note that the slope overall (adjusting for sex) of 0.963/cm is approximately an average of the sex-specific slopes of 1.024 (females) and 0.869 (males)

Correction of Confounding

- We can formally test whether the slopes are different (how?).

- Turns out that the slopes are not different (by statistical or material criteria) as the plot seems to indicate. Adjusting for sex lets us examine the true relationship between weight and height more accurately.

- Note that age and sex are the confounding usual suspects in medical and epidemiologic studies, and so we often adjust for them in analyses.)

• Question: Why is sex not considered an effect modifier?

Confounding vs. Effect Modification: Again

- Here, while sex is an important predictor of weight ...
  - There is clearly no differential effect of height on weight according to sex. The slopes for height within males and within females are about the same. There is no interaction effect
  - Both of these slopes are different from the marginal slope or unadjusted effect for height (e.g., ignoring sex)
- Thus, the effect of height on weight is said to be confounded by sex

Confounding in Observational Studies

- Framework for many observational studies: three types of variables:
  a) Response (outcome, dependent variable) Y
  b) Predictor variable X - exposure of interest (and sometimes the interaction term)
  c) Covariates: may be confounders, sometimes called control variable(s), Z, and may also include other nuisance variables (e.g., a suspect confounder, or the main effect of the effect modifier)

- Distinction between X and Z is because we CARE about predictors while the covariates are considered nuisance variables that we must control to avoid bias, wrong conclusions

- To address confounders in studies.
  a Must carefully consider context, conceptual model, and **collect suspect factors**
  b. Practically, might analyze with and without adjustment for a suspect confounder

Confounding in Observational Studies

- Models to contrast (often presented in epidemiologic studies)
  unadjusted model: $Y = \beta_0 + \beta_1 X + \epsilon$
  adjusted model: $Y = \beta_0' + \beta_1' X + \beta_2' Z + \epsilon$
- Check to see whether β_1 and β_1' are different from each other (not strictly a statistical question!)
- If yes: Z could be a confounder. What if β_2' is not statistically significant? Does not mean that Z is not a confounder. We may retain Z in the model to maximally control bias
- So what variable should we consider the "usual suspects"?
  a) Factors known or generally thought to influence Y ("the risk factors" for the response)
  b) Factors thought to be important for interpretability, credibility of findings

Consequences of Ignoring Confounders

- When confounding is present, the contributing effect of X is the same for each value of Z (i.e., in a linear main effects model), but not taking Z into account distorts the true effect.
  Wrong estimation and misleading conclusions.

Effect Modification (Interaction) and Confounding - Summary

- Confounding is a bias that we hope to prevent or control - makes X seem related to Y but it is not
- Effect modification is a real effect - differential effect on Y of X_1 in presence/absence/at value of X_2
- Confounding is something to avoid and so confounders need to be included in the analysis
- Effect modification, if not accounted for, provides an 'average effect' ignoring the third variable, may not be wrong but is much less informative. With qualitative interactions, conclusion may be wrong

How to adjust confounding effect?

- If you DON'T KNOW what the potential confounders are
  1. Before the experiment is conducted (data are collected), randomization is the best protection. Randomization eliminates the potential links between the exposure of interest and potential confounders. But randomization is not always possible.
  2. After the experiment (when data are already collected), there's not much you can do.

- If you can't randomize but KNOW what the potential confounders are, potential confounders must be measured as part of study. There are statistical methods to help control or adjust for confounders.
  - Include confounder in the analysis as covariate
  - Stratified analysis
  - Matched study

Summary

- Before you collect or see the data, think about the causal diagram. A confounder is causally related to the response (Y) and is associated with the predictor of interest (X). An effect modifier will delineate effects of X on Y for different groups or under different conditions (slope changes for different groups).
- You must consider (measure or using randomization to eliminate the effect of) confounders in experimental design to reduce bias.
- After you see the data, the coefficients for confounders does not need to be significant, nor the coefficients for X on Y. But adjusting confounders or not will give you different estimated effects of X on Y. And this difference in estimates does not need to be significant either.

Examples for you to read

1. A classic economics example:
   Gender discrimination in salary. Without adjusting for education and experience, there appears to be salary discrimination against female. With the adjustment, often the "discrimination" disappears.
   So, education and experience are confounders.

2. Another example: Nutritional intake (calcium) is associated with osteoporosis among women. Among men this association is not so strong because men's bone mineral content is not affected as much by nutritional intake.

   That is, the overall association correctly estimates the average effect of the exposure, but that effect is different in different subgroups.
   Here gender is an effect modifier of the association between nutritional intake and osteoporosis.
   If the separate associations are of interest, then a stratified analysis is called for. If the main scientific interest is in the average effect across the population, then a stratified analysis is unnecessary.