PM 592
Regression Analysis for Public Health Data Science

Week 6

Confounding & Interaction

1

1

Confounding & Interaction

Estimating Association

Confounding

Effect Modification

2

Lecture Objectives

- Describe the two possible goals of regression analysis and how the modeling approach would differ for each.
- Discern whether a variable could sensibly be a confounder.
- Assess the extent to which a variable acts as a confounder.
- Specify an interaction term in a linear regression model.
- Interpret continuous-by-categorical and continuous-by-continuous interactions.

3

1. Review 4

- ✓ Shared covariance of independent variables
- ✓ The multiple regression model
- ✓ Interpretation of multiple regression coefficients
- ✓ Interpretation of R^2 from a multiple regression model
- ✓ Diagnosing collinearity
- ✓ Methods for diagnosing outliers and influential points

4

2. Regression: Estimating Association 5

What is the primary purpose of multiple regression models?

This depends on your research question!

1. **Estimating Associations**.
   - Multiple regression can model the relationship between a dependent variable and a set of independent variables of **scientific interest**.
   - The variables are chosen due to **hypothesis-driven** associations.
   - Covariates are chosen to best capture the **true effect** of the IV on the outcome.
   - We include **confounding variables** (variables that confound the relationship between X and Y) and **effect modifiers** (variables that modify the relationship between X and Y).

5

2. Regression: Estimating Association 6

When estimating associations:

- The model coefficients reflect the slope **adjusting for all other variables in the model**.
- The validity of this adjustment is contingent on:
  - The correct covariates being included
  - The covariates being modeled correctly (e.g., correct assumption of linearity, interactions modeled correctly)

With this approach your goal is to model the true association between the independent variable of interest and the outcome.

6

What is the primary purpose of multiple regression models?

2. Predicting Outcome

- The main goal isn't to examine a hypothesized association.
- Rather, the main goal is to find the best way to predict outcome.
- Good prediction models use a parsimonious set of independent variables.

With this approach your goal is to find a set of variables that best predicts outcome.

7

How do we know if an additional variable improves model fit?

1. Extra Sums of Squares Test

Idea: compare the SS of the model with an additional variable vs. the SS of the model without it.

We've already used the this test to determine:

- If a single variable is statistically significant (compared to the null hypothesis)
- If a dummy variable set is statistically significant (i.e., joint significance of all variables)

8

The F-test shown in the output reflects the significant improvement in sum of squares **compared to a model with nothing in it**.

For example, the extra SS (F) test shows that the model below is significantly better than a model with no predictors.

```
> m1 <- lm(enjoy_ex1 ~ intervention, data = places)
> summary(m1)

Call:
lm(formula = enjoy_ex1 ~ intervention, data = places)

Residuals:
   Min     1Q Median     3Q    Max 
-3.273 -1.273 -0.079  1.727  3.115 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.88507    0.13748  28.259   <2e-16 ***
intervention  0.19396    0.09612   2.018    0.044 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.854 on 586 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.006901,	Adjusted R-squared:  0.005207 
F-statistic: 4.072 on 1 and 586 DF,  p-value: 0.04405
```

Overall, this model has significant predictive ability.

9

2. Regression: Estimating Association

10

What does a null (unconditional) model look like?

```
> m0 <- lm(enjoy_ex1 ~ 1, data = places)
> summary(m0)

Call:
lm(formula = enjoy_ex1 ~ 1, data = places)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1157 -1.1157 -0.1157  1.8843  2.8843 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.11565    0.07666   53.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.859 on 587 degrees of freedom
  (27 observations deleted due to missingness)
```

This is the null model (everybody is assigned the mean value of Y as their predicted value) so there is no F test.

10

2. Regression: Estimating Association

11

Recall the Extra Sums of Squares F-test statistic is given by:

$$F_{(DFE_0-DFE_1,DF\ \ _1)} = \frac{\frac{SSE_0 - SSE_1}{DFE_0 - DFE_1}}{\frac{SSE_1}{DFE_1}}$$

```
> anova(m1)
Analysis of Variance Table

Response: enjoy_ex1
             Df Sum Sq Mean Sq F value  Pr(>F)  
intervention   1   14.0 13.9970  4.0723 0.04405 *
Residuals    586 2014.1  3.4371                  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(m0)
Analysis of Variance Table

Response: enjoy_ex1
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 587 2028.1  3.4551               

F-statistic: 4.072 on 1 and 586 DF,  p-value: 0.04405
```

$$F_{(587-586,586)} = \frac{\frac{2028-2014}{587-586}}{\frac{2014}{586}} = \frac{14}{\frac{2014}{586}} = 4.07$$

11

2. Regression: Estimating Association

12

In this past example we used the Extra SS test to determine how well our model fit, compared to a null model.

We can also use the Extra SS test to determine if the addition of an independent variable – or multiple independent variables – improves our model.

This is especially useful when examining the effect of adding multiple independent variables.

Note: when performing this test, the **models must be nested** (e.g., the model with intervention is nested within the model with intervention + factor(race))

12

2. Regression: Estimating Association

Suppose we want to test the effect of the addition of categorical race (4 dummy variables).

```
> m2 <- lm(enjoy_ex1 ~ intervention + racecat, data = places)
> summary(m2)

Call:
lm(formula = enjoy_ex1 ~ intervention + racecat, data = places)

Residuals:
   Min     1Q Median     3Q    Max 
-3.483 -1.341 -0.126  1.659  3.416 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     3.91121    0.17024  22.975   <2e-16 ***
intervention                    0.21479    0.09991   2.150   0.0320 *  
racecatAsian                   -0.25413    0.30411  -0.836   0.4037    
racecatBlack/African American  0.10563    0.24776   0.426   0.6700    
racecatNon-Hispanic White     -0.32693    0.19561  -1.671   0.0952 .  
racecatOther                    0.35754    0.30810   1.160   0.2463    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.851 on 582 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.01726,	Adjusted R-squared:  0.008817 
F-statistic: 2.044 on 5 and 582 DF,  p-value: 0.07085

> anova(m1, m2)
Analysis of Variance Table

Model 1: enjoy_ex1 ~ intervention
Model 2: enjoy_ex1 ~ intervention + racecat
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    586 2014.1                           
2    582 1993.1  4    21.009 1.5336 0.1909
```

M1 is the nested model with just intervention.

M2 is the full model with intervention and categorical race.

$$F_{(586-582,58\ \ )} = \frac{\frac{2014-1993}{586-582}}{\frac{1993}{582}} = \frac{\frac{21}{4}}{\frac{1993}{582}} = 1.53$$

13

2. Regression: Estimating Association

The Extra SS test is quite useful when determining if a set of variables improves the model.

When only one variable is added to the model, this test will mirror the Wald (t) test.

14

2. Regression: Estimating Association

Recap

- Models can be created either for 1) estimating associations or 2) prediction.
- The Extra Sums of Squares test can be used to determine if additional variable(s) improve the model.
- This test must be used on nested models (i.e., the parameters in the nested model must also appear in the full model).

15

Recap

- Explain the difference between an association model and a prediction model
- Compute the Extra Sums of Squares test, including the F-statistic, degrees of freedom, and p-value
- Explain what the Extra Sums of Squares test is testing

16

Confounding

- When the association of interest between the outcome variable (Y) and a specific independent variable (X) is distorted by the influence of a third (or more) variable.

Criteria

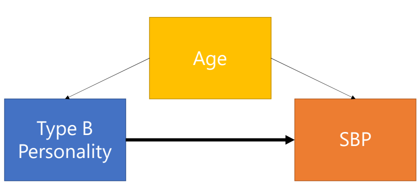
1. Change in effect estimate criterion: the adjusted estimate is at least 10-20% different than the unadjusted estimate AND
2. The variable must sensibly be able to be a cause of both X and Y

17

Example

Is the relationship between personality type and SBP confounded by age?

Step 1: Can age reasonably be a confounder of this relationship?



18

3. Confounding

19

Step 2: Assess confounding through change in parameter estimates.

```
> m3 <- lm(sbp ~ factor(dibpat), data = wcgs)
> m3a <- lm(sbp ~ factor(dibpat) + age, data = wcgs)
> summary(m3)

Call:
lm(formula = sbp ~ factor(dibpat), data = wcgs)

Residuals:
    Min      1Q  Median      3Q     Max
-29.782  -9.782  -1.782   8.455 102.534

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          129.7823     0.3782 343.163  < 2e-16 ***
factor(dibpat)Type B  -2.3164     0.5369  -4.315 1.65e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.08 on 3152 degrees of freedom
Multiple R-squared:  0.005871,	Adjusted R-squared:  0.005556
F-statistic: 18.62 on 1 and 3152 DF,  p-value: 1.649e-05

> summary(m3a)

Call:
lm(formula = sbp ~ factor(dibpat) + age, data = wcgs)

Residuals:
    Min      1Q  Median      3Q     Max
-33.833 -10.244  -2.503   8.022 102.002

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          109.27669    2.28336  47.858  < 2e-16 ***
factor(dibpat)Type B  -1.88867    0.53213  -3.549 0.000392 ***
age                    0.43850    0.04817   9.103  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.88 on 3151 degrees of freedom
Multiple R-squared:  0.03134,	Adjusted R-squared:  0.03073
F-statistic: 50.98 on 2 and 3151 DF,  p-value: < 2.2e-16
```

$$\%\text{change} = \frac{\beta_{unadjusted} - \beta_{adjusted}}{\beta_{unadjusted}}$$

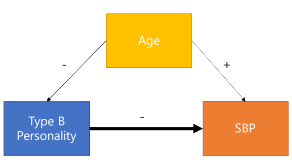
$$\frac{-2.32 - (-1.89)}{-2.32} = 18.5\%$$

19

3. Confounding

20

The parameter estimate for personality type changed by 18.5% after including age into the model. It appears that age confounds the relationship between personality type and SBP.



A correlation matrix can help you with this.

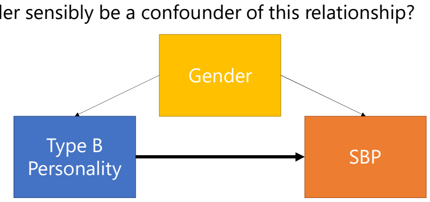
20

3. Confounding

21

Example

Could gender sensibly be a confounder of this relationship?



We don’t have gender in the data set, but it seems like this could be a valid model.

21

How do we pick confounders to examine? (Not a simple question!)

- **Philosophically -** Sometimes variables are *a priori* included automatically because they have traditionally been examined as confounders (e.g., age, gender, race).
- Demographic variables are most likely to be confounders.

22

How do we pick confounders to examine? (Not a simple question!)

- **Statistically –** We sometimes don't have adequate information to choose all potential confounders.
- In the CHS study, variables such as pets, carpet type, air conditioning, multivitamin use, etc. can be potential confounders of relationships with lung function.
- Including a large number of "adjustment" variables can pose several problems.

23

Problems with Adding Several Adjustment Variables

- When individuals are **missing** on certain adjustment variables, those individuals may be eliminated from the analysis, reducing power and potentially biasing effect estimates.

24

3. Confounding 25

Problems with Adding Several Adjustment Variables

- Adding several independent variables changes our power to detect effects of interest.
  - **Reduced power.** As the number of predictors in a model increases, the Error DF decreases. This leads to heavier tails in the t-distribution and the observed t-value must be larger to reject H_0 ($p<.05$).
  - **Increased power**. The standard error of the slope coefficients are a function of the estimated standard deviation of the residuals. If adjustment variables are added and they are able to reduce this standard deviation, then the SE will be smaller and the t-value will be larger.

$$SE(\hat{\beta}) = \frac{s_{Y|X}}{s_X \cdot \sqrt{n-1}}$$

25

3. Confounding 26

Problems with Adding Several Adjustment Variables

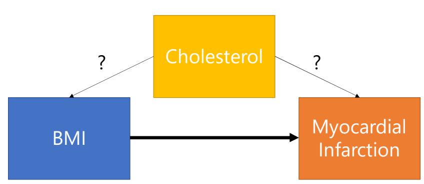
- A large model with many nonsignificant X-variables violates our desire to have a minimal (i.e. "parsimonious") model that does a good job predicting our outcome.
- In model-building, we ultimately want to find the most parsimonious model

26

3. Confounding 27

Confounder?

Could this variable sensibly be a confounder?



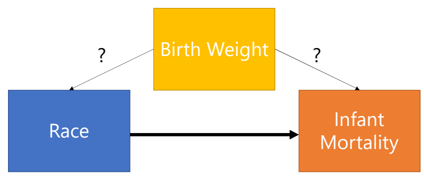
It's not entirely clear whether cholesterol affects BMI or vice versa…

27

3. Confounding 28

Confounder?

Could this variable sensibly be a confounder?



28

3. Confounding 29

Study Design

- The potential for confounding is reduced in experimental studies. This is because individuals are randomized into treatment groups, so the treatment/exposure groups are similar in terms of potential confounding variables.

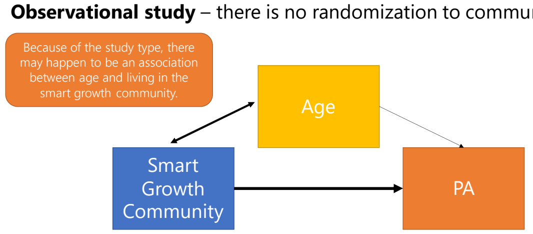
29

3. Confounding 30

Example

Is moving to a Smart Growth community related to increases in physical activity (PA)?

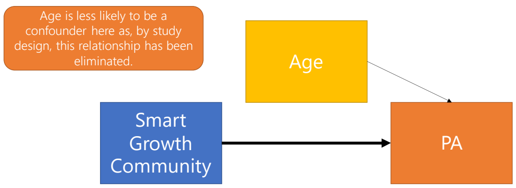
Observational study – there is no randomization to community type.



30

3. Confounding 31

Experimental study – all individuals who want to move to the smart growth community are included into a lottery. Half move and half remain where they had been living.



31

3. Confounding 32

In an observational study there may be several pre-existing associations that we want to control for.

For example, in the CHS:

- Individuals who live in a more polluted community may happen to be of a particular racial/ethnic group (race confounds the association between pollution and lung function).
- Individuals who live in a more polluted community may happen to be of a particular gender (sex confounds the association between pollution and lung function).

32

3. Confounding 33

Mediators

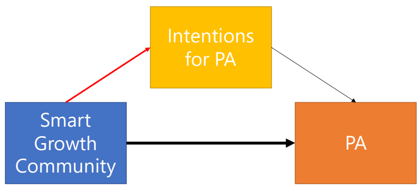
- <u>Confounders</u> are variables that cause a spurious association between an X and Y variable of interest.
- <u>Mediators</u> are variables that are intermediary steps in a causal pathway between X and Y.

33

3. Confounding 34

The Theory of Planned Behavior posits that the relationship between an X variable and a behavioral outcome is mediated by intentions.

In this example, intentions for PA is a mediator in a causal pathway between moving to a smart growth community and physical activity.



34

3. Confounding 35

Dealing with Mediators

- As mediators are within the causal pathway, they do not lead to spurious relationships between X and Y.
- The analysis of mediators is beyond the scope of this course, but their effects can be assessed with path analysis.

35

3. Confounding 36

Recap

- In order to be a confounder, the variable under consideration must:
  - Change the slope parameter of interest appreciably (>10-20%)
  - Sensibly be a confounder (i.e., not be a causal intermediary)
- By including a confounder in a regression of Y on X, you will be able to calculate the effect of X on Y that is not due to the confounder.

36

3. Confounding 37

Recap

- Determine whether a given variable is a confounder of a relationship between X and Y
- Determine whether a given variable is a causal intermediary of a relationship between X and Y

37

4. Effect Modification 38

Effect Modification

- When the association of interest between the outcome variable (Y) and a specific independent variable (X) is depends on a third variable (Z).

Criteria

- An interaction term will be significant within the model.

38

4. Effect Modification 39

Example

30 families were asked about their ice cream consumption in the past month. The following variables were obtained:

CONS – ice cream consumption in pints per capita

INCOME – family income (in $1,000)

PRICE – price per ounce of ice cream ($)

TEMP – average temperature in the past month

39

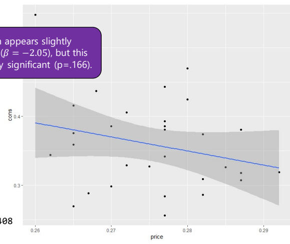
4. Effect Modification

40

Let's examine whether the price of ice cream is related to consumption of ice cream.

```
> ic.m1 <-
+   lm(cons ~ price, data = ic)
> summary(ic.m1)

Call:
lm(formula = cons ~ price, data = ic)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.11152 -0.03707 -0.00991  0.03666  0.15724 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   0.9230     0.3964   2.329   0.0273 *
price        -2.0472     1.4393  -1.422   0.1660  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06466 on 28 degrees of freedom
Multiple R-squared:  0.06739,	Adjusted R-squared:  0.03408 
F-statistic: 2.023 on 1 and 28 DF,  p-value: 0.166
```

Ice cream consumption appears slightly negatively related to price ($\beta = -2.05$), but this association is not statistically significant (p=.166).



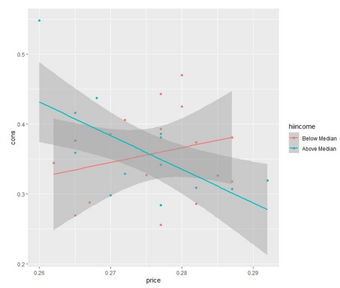
40

4. Effect Modification

41

What happens when we look at this relationship by income status (high vs. low, dichotomized at the median)? It appears that there are two different effects between price and consumption:

1) For high-income families, higher price is associated with lower consumption.
2) For low-income families, higher price is slightly associated with higher consumption.



41

4. Effect Modification

42

If we run the regression separately for those with high income and low income, we come to two separate effects, by income category:

```
> ic.m2a <-
+   lm(cons ~ price, data = ic %>%
+   filter(hiincome == "Below Median"))
> summary(ic.m2a)

Call:
lm(formula = cons ~ price, data = ic %>%
filter(hiincome == "Below Median"))

Residuals:
      Min        1Q    Median        3Q       Max 
-0.103842 -0.056842  0.003614  0.048810  0.103832 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2243     0.6462  -0.347    0.734
price         2.1088     2.3404   0.901    0.384


Residual standard error: 0.0662 on 13 degrees of freedom
Multiple R-squared:  0.05878,	Adjusted R-squared:  -0.01362 
F-statistic: 0.8119 on 1 and 13 DF,  p-value: 0.384
```

For families with low income, price does not affect consumption.

```
> ic.m2b <-
+   lm(cons ~ price, data = ic %>%
+   filter(hiincome == "Above Median"))
> summary(ic.m2b)

Call:
lm(formula = cons ~ price, data = ic %>%
filter(hiincome == "Above Median"))

Residuals:
      Min        1Q    Median        3Q       Max 
-0.085236 -0.037943  0.005448  0.033898  0.116715 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   1.6806     0.4232   3.971  0.00160 **
price        -4.8049     1.5404  -3.119  0.00814 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05344 on 13 degrees of freedom
Multiple R-squared:  0.4281,	Adjusted R-squared:  0.3841 
F-statistic:  9.73 on 1 and 13 DF,  p-value: 0.00814
```

For families with high income, higher price decreases consumption.

42

4. Effect Modification 43

Instead of analyzing these two models separately, we write this as one model in the form:

$$\mu_{Y|X1,X2} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

The variables X1 and X2 are said to interact if $\beta_3 \neq 0$.

43

4. Effect Modification 44

Here we want to test:

$$\mu_{Y|X1,X2} = \beta_0 + \beta_1 X_{HIINCOME} + \beta_2 X_{PRICE} + \beta_3 X_{HIINCOME} X_{PRICE}$$

44

4. Effect Modification 45

We will arrive at to distinct effects depending on income category:

For <u>low income</u> ($X_{HIINCOME}$=0):

$$\mu_{Y|X1,X2} = \beta_0 + \beta_1(0) + \beta_2 X_{PRICE} + \beta_3(0) X_{PRICE}$$
$$= \beta_0 + \beta_2 X_{PRICE}$$

For <u>high income</u> ($X_{HIINCOME}$=1):

$$\mu_{Y|X1,X2} = \beta_0 + \beta_1(1) + \beta_2 X_{PRICE} + \beta_3(1) X_{PRICE}$$
$$= \beta_0 + \beta_1 + \beta_2 X_{PRICE} + \beta_3 X_{PRICE}$$
$$= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_{PRICE}$$

45

4. Effect Modification

46

Our fit equation becomes:

$$\hat{Y} = \mu_{Y|X1,X2} = -0.22 + 1.90X_{HIINCOME} + 2.11X_{PRICE} - 6.91X_{HIINCOME}X_{PRICE}$$

```
# Created price_hiincome:
# mutate(price_hiincome = price * (income > median(income)))

> summary(ic.m2)

Call:
lm(formula = cons ~ price + hiincome + price_hiincome, data = ic)

Residuals:
      Min        1Q    Median        3Q       Max
-0.103842 -0.047352  0.004531  0.039789  0.116715

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.2243     0.5872  -0.382   0.7056
price                 2.1088     2.1269   0.991   0.3306
hiincomeAbove Median  1.9048     0.7562   2.519   0.0183 *
price_hiincome       -6.9137     2.7441  -2.519   0.0182 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06016 on 26 degrees of freedom
Multiple R-squared:  0.2504,    Adjusted R-squared:  0.1639
F-statistic: 2.895 on 3 and 26 DF,  p-value: 0.05426
```

Because the p-value for the interaction term is <.05, we can conclude there is a statistically significant interaction between ice cream price and income category on ice cream consumption.

46

4. Effect Modification

47

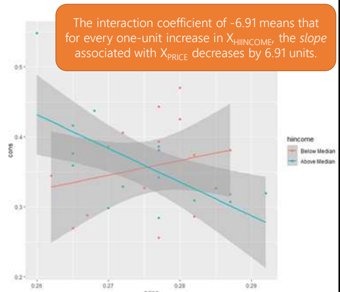
Another way to think about this is that the interaction term describes how the slope changes.

For those with income below median ($X_{HIINCOME}$ = 0):

$$\hat{Y} = -0.22 + 1.90(0) + 2.11X_{PRICE} - 6.91(0)X_{PRICE}$$
$$= -0.22 + 2.11X_{PRICE}$$

For those with income above median ($X_{HIINCOME}$ = 1):

$$\hat{Y} = -0.22 + 1.90(1) + 2.11X_{PRICE} - 6.91(1)X_{PRICE}$$
$$= 1.68 - 4.80X_{PRICE}$$



The interaction coefficient of -6.91 means that for every one-unit increase in $X_{HIINCOME}$, the *slope* associated with X_{PRICE} decreases by 6.91 units.

47

4. Effect Modification

48

In R, interaction terms can be created directly in the equation formula:

```
> summary(lm(cons ~ price*hiincome, data = ic))

Call:
lm(formula = cons ~ price * hiincome, data = ic)

Residuals:
      Min        1Q    Median        3Q       Max
-0.103842 -0.047352  0.004531  0.039789  0.116715

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -0.2243     0.5872  -0.382   0.7056
price                        2.1088     2.1269   0.991   0.3306
hiincomeAbove Median         1.9048     0.7562   2.519   0.0183 *
price:hiincomeAbove Median  -6.9137     2.7441  -2.519   0.0182 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06016 on 26 degrees of freedom
Multiple R-squared:  0.2504,    Adjusted R-squared:  0.1639
F-statistic: 2.895 on 3 and 26 DF,  p-value: 0.05426
```

48

4. Effect Modification 49

Notes about Interaction Terms

- **Main Effects** refer to the effects (variables) in the model without modeling any interaction (e.g., the effect of price on consumption regardless of income category).
- If the **interaction term is significant** then you should retain the effects of all variables included in the interaction term – regardless of significance.
- Detecting an interaction requires **larger sample size** than for main effects; typically 4n, where n is the sample size required to detect a main effect at 80% power. You may want to be more liberal with the α level for interactions (e.g., $\alpha = 0.15$).

49

4. Effect Modification 50

Continuous by Continuous Interactions

- We previously examined the interaction of a continuous variable with a dichotomous variable.
- Continuous-by-dichotomous interactions are easier to interpret because we can stratify by the dichotomous variable for interpretation.

50

4. Effect Modification 51

Suppose we want to examine the same interaction as before, but we will leave income in as a continuous variable.

Let's fit the following model:

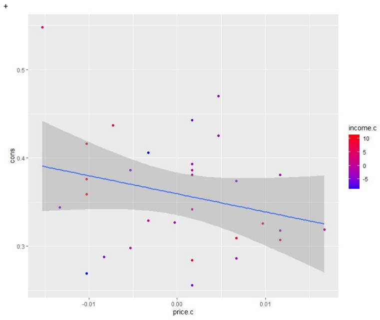
$$\hat{Y} = \beta_0 + \beta_1 X_{INCOME.C} + \beta_2 X_{PRICE.C} + \beta_3 X_{INCOME.C} X_{PRICE.C}$$

Note: it's a good idea to center variables in general, but especially when performing a continuous-by-continuous interaction. Here the ".C" in the variable name is my notation for saying that variable is mean-centered.

51

4. Effect Modification


```
> ic %>%
+   ggplot(aes(x = price.c, y = cons, color = income.c)) +
+   geom_point() +
+   geom_smooth(method = "lm") +
+   scale_color_gradient(low = "blue", high = "red")
```



52

4. Effect Modification

We see that the interaction between price and income is statistically significant (p=.006):

```
> ic.m3 <-
+   lm(cons ~ price.c*income.c, data = ic)
> summary(ic.m3)

Call:
lm(formula = cons ~ price.c * income.c, data = ic)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.116865 -0.038381 -0.007629  0.033801  0.127481 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       0.355666   0.010633  33.448  < 2e-16 ***
price.c          -1.285239   1.318446  -0.975  0.33864    
income.c         -0.002278   0.001919  -1.187  0.24585    
price.c:income.c -0.695851   0.232148  -2.997  0.00592 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05783 on 26 degrees of freedom
Multiple R-squared:  0.3072,	Adjusted R-squared:  0.2273 
F-statistic: 3.843 on 3 and 26 DF,  p-value: 0.02113
```

53

4. Effect Modification

And our model-fit equation becomes:

$$\hat{Y} = 0.356 - 0.002X_{INCOME.C} - 1.285X_{PRICE.C} - 0.696X_{INCOME.C}X_{PRICE.C}$$

What does this tell us?

Two ways to interpret main effects:

1. At the mean value of income ($X_{INCOME.C}$=0), a one-unit increase in $X_{PRICE.C}$ is associated with a predicted decrease ice cream consumption of 1.285 units (p=0.34).
2. At the mean value of price ($X_{PRICE.C}$=0), a one-unit increase in $X_{INCOME.C}$ is associated with a predicted decrease in ice cream consumption of 0.002 units (p=0.25).

54

4. Effect Modification

55

And our model-fit equation becomes:

$$\hat{Y} = 0.356 - 0.002X_{INCOME.C} - 1.285X_{PRICE.C} - 0.696X_{INCOME.C}X_{PRICE.C}$$

What does this tell us?

Two ways to interpret the interaction term:

1. For every one-unit increase in $X_{INCOME.C}$, the slope associated with $X_{PRICE.C}$ decreases by 0.696.
2. For every one-unit increase in $X_{PRICE.C}$, the slope associated with $X_{INCOME.C}$ decreases by 0.696.

55

4. Effect Modification

56

One way to examine the effect of a continuous-by-continuous interaction term is **simple slopes**.

In this analysis, we look at the effect of one variable by levels of another variable – typically at the mean, -1SD, and +1SD.

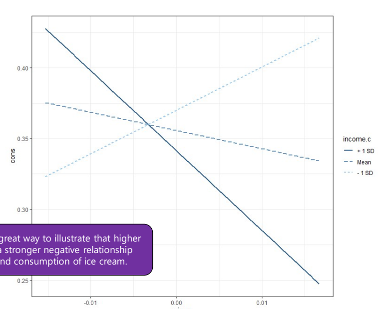
56

4. Effect Modification

57

One way to examine the effect of a continuous-by-continuous interaction term is with **simple slopes**.

In this analysis, we look at the effect of one variable by levels of another variable – typically at the mean, -1SD, and +1SD.



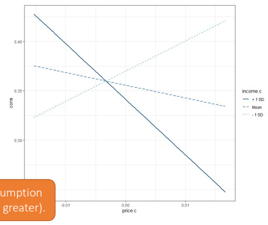
This approach is a great way to illustrate that higher income leads to a stronger negative relationship between price and consumption of ice cream.

57

4. Effect Modification 58

The calculation of $\beta_{PRICE.C}$ at the mean, -1SD, and +1SD of income is straightforward, but the significance of the slope at these values is a bit tricky. I recommend using a package that will calculate these for you.

```
> sim_slopes(ic.m3, pred = price.c, modx = income.c, johnson_neyman = FALSE)
SIMPLE SLOPES ANALYSIS

Slope of price.c when income.c = -6.25 (- 1 SD):

  Est.   S.E.   t val.      p
------ ------ -------- ------
  3.06   2.14     1.43   0.16

Slope of price.c when income.c =  0.00 (Mean):

   Est.   S.E.   t val.      p
------- ------ -------- ------
  -1.29   1.32    -0.97   0.34

Slope of price.c when income.c =  6.25 (+ 1 SD):

   Est.   S.E.   t val.      p
------- ------ -------- ------
  -5.63   1.77    -3.19   0.00
```



This tells us that price is related to consumption only for high levels of income (+1SD and greater).

58

4. Effect Modification 59

Recap

- To determine if the effect of X on Y changes based on the value of another variable Z, you can include an interaction term in the model.
- The significance of the beta coefficient for the interaction term (X*Z) determines whether there is a significant interaction.
- Compared to main effects, interactions require larger sample size to detect significant effects.

59

4. Effect Modification 60

Recap

- Determine the presence of an interaction through adding an interaction term in a regression model
- Explain the effect of X on Y based on:
  - Levels of Z (if Z is categorical)
  - The value of Z (if Z is continuous)

60

5. Recap

61

Goals of Data Analysis and Recommendations

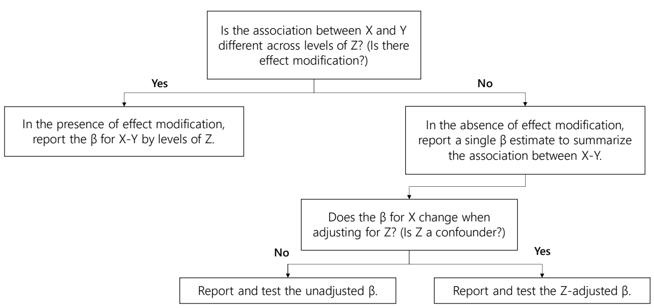
	Model of Association	Prediction Model
Goal	Explain the true relationship between an X (or set of X) variable and an outcome.	Use a set of X variables to find the model that can best predict the outcome.
Variables	Choose based on theoretically meaningfulness of associations (e.g., which variables may be distorting the relationship between X and Y).	Choose based on what might be associated with the outcome.
Method	Control for confounding. Examine effect modification. Keep variables that explain the "true" relationship between X and Y.	Confounding is not important. Examine effect modification. Keep variables that improve model fit.
Metrics	R^2, parsimony, ensuring you have considered all possible confounders.	R^2, parsimony, validation with an external data set.

61

5. Recap

62

Flowchart for assessing effect modification and confounding



62

5. Recap

63

- **Keep variables continuous when possible.** In this lecture we categorized income to examine its interaction with price and ice cream consumption – categorizing a variable results in a loss of information, but with added interpretability. Because income is continuous, it may be more desirable to keep a continuous-by-continuous interaction with price.
- **Keep all lower-order terms for significant interactions.** When an interaction is included in the model, keep all variables included in the interaction term (regardless of significance).

63

- **Check interactions first.** Because interactions reflect strata-specific effects, check for interaction variables before assessing confounding.
- **Don't get overwhelmed.** With several independent variables it may be tempting to examine many possible interactions. Stick to *a priori* hypothesized interactions, or interactions that make the most sense. This will reduce the amount of possible variables you have to look at.
- **Stick to parsimony.** A good model attempts to simplify reality while retaining accuracy. A parsimonious model explains as much of the relationship of interest as possible with as few variables as possible.

64

Additional Reading

- Explanatory vs. Prediction Modeling
  https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf

65

Packages and Functions

- `interactions::interact_plot()`

66