ENV 710: Lecture 14

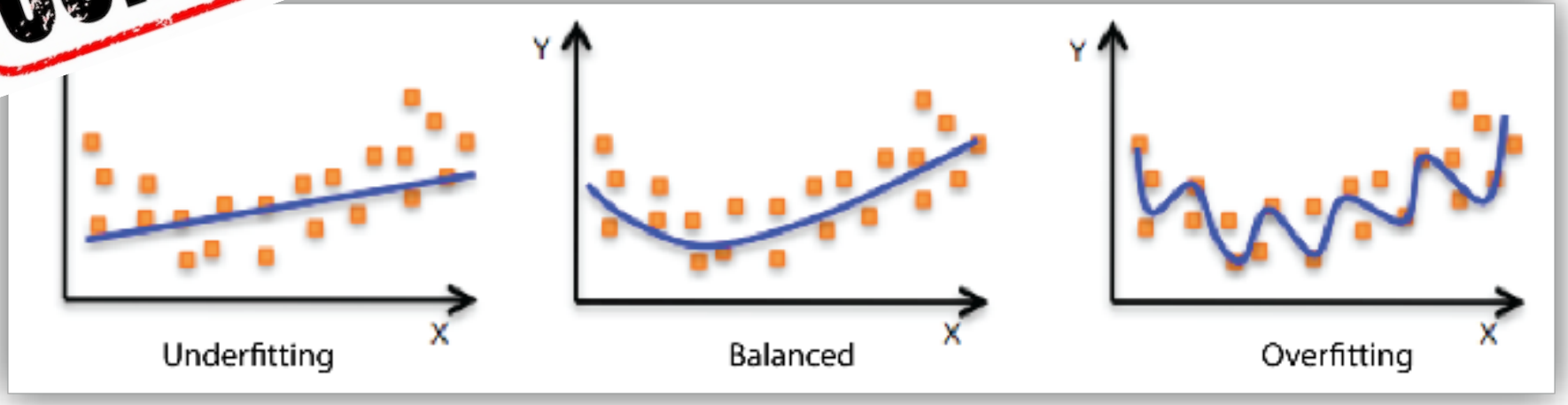
interactions

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

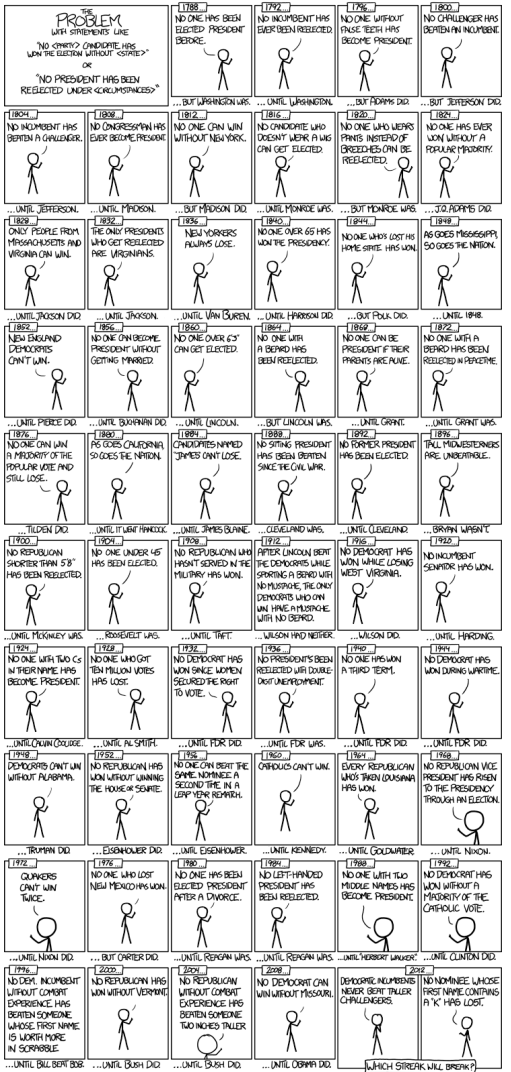
linear models

interactions, etc.

COMMON SENSE

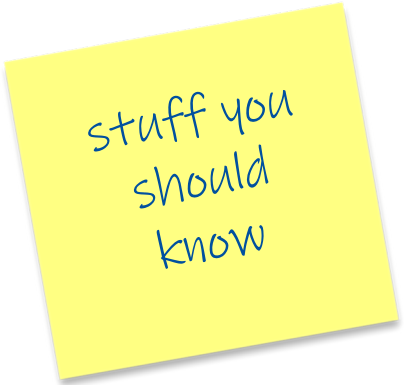


COMMON SENSE



learning goals

- what do interactions mean?
  - continuous IV x categorical IV
  - continuous IV x continuous IV
- when and why do we scale independent variables?
  - centering, standardizing



interactions

- an interaction effect exists when the effect of an independent variable on a dependent variable changes, depending on the value(s) of one or more other independent variables
- a third variable influences the relationship between an independent and dependent variable

Satisfaction = Food Type x Condiment

types of interactions

categorical IV x categorical IV
categorical IV x continuous IV
continuous IV x continuous IV

interaction: continuous and categorical IV's

example cognitive test scores

Can the cognitive test scores of 3 & 4-year olds (`kid.score`) be predicted from characteristics of their mothers?

- mother's IQ (`mom.iq`),
- completion of high school (`mom.hs`): 1 = completed, 0 = failed
- mother's age (`mom.age`)

```
lm(kid.score ~ mom.hs*mom.iq)
```

$$Y = \beta_0 + \beta_1 HS + \beta_2 IQ + \beta_3 \cdot HS \cdot IQ$$

what do the coefficients mean?

```
lm(kid.score ~ factor(mom.hs) * mom.iq)
```

```
lm(formula = kid.score ~ mom.hs + mom.iq + mom.hs:mom.iq)

Coefficients:
              Estimate   Std. Error t value Pr(>|t|)
(Intercept)   -11.4820    13.7580  -0.835 0.404422
mom.hs         51.2682    15.3376   3.343 0.000902 ***
mom.iq          0.9689     0.1483   6.531 1.84e-10 ***
mom.hs:mom.iq  -0.4843     0.1622  -2.985 0.002994 **
---
```

- intercept is the predicted `kid.score` when `mom.hs` and `mom.iq` are both 0

- coefficients for main effects reflect conditional relationships
- β_1 is the effect of X_1 on Y when $X_2 = 0$
- β_2 is the effect of X_2 on Y when $X_1 = 0$

- interaction coefficient, β_3, is included when neither X_1 or $X_2 = 0$; *then the effect of one variable depends on the other*

what do the coefficients mean?

```
lm(kid.score ~ factor(mom.hs) * mom.iq)

lm(formula = kid.score ~ mom.hs + mom.iq + mom.hs:mom.iq)

Coefficients:
              Estimate   Std. Error t value Pr(>|t|)
(Intercept)   -11.4820    13.7580  -0.835 0.404422
mom.hs         51.2682    15.3376   3.343 0.000902 ***
mom.iq          0.9689     0.1483   6.531 1.84e-10 ***
mom.hs:mom.iq  -0.4843     0.1622  -2.985 0.002994 **
---
```

- β_1 is the effect of X_1 on Y when $X_2 = 0$

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 0 + \beta_3 X_1 \cdot 0$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 + 0 + 0$$

- *when both* X_1 and X_2 are not 0, β_3 becomes important, and the effect of X_1 *varies with* X_2

$$\hat{Y} = \beta_0 + \beta_1 \cdot 1 + \beta_2 X_2 + \beta_3 \cdot 1 \cdot X_2$$

$$\hat{Y} = \beta_0 + \beta_1 + \beta_2 X_2 + \beta_3 X_2$$

$$\hat{Y} = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) \cdot X_2$$

- β_2 is the effect of X_2 on Y when $X_1 = 0$

$$\hat{Y} = \beta_0 + \beta_1 \cdot 0 + \beta_2 X_2 + \beta_3 \cdot 0 \cdot X_2$$

$$\hat{Y} = \beta_0 + 0 + \beta_2 X_2 + 0$$

example cognitive test scores

Calculate the cognitive score for a child with a mother who completed high school and has an IQ of 100…

$$Y = -11.48 + 51.27 \cdot HS + 0.97 \cdot IQ - 0.48(HS \cdot IQ)$$

```
lm(formula = kid.score ~ mom.hs + mom.iq +
             mom.hs:mom.iq)

Estimate Std. Error t value Pr(>|t|)
(Intercept)   -11.4820    13.7580  -0.835 0.404422
mom.hs         51.2682    15.3376   3.343 0.000902 ***
mom.iq          0.9689     0.1483   6.531 1.84e-10 ***
mom.hs:mom.iq  -0.4843     0.1622  -2.985 0.002994 **
```

use `predict()`

```
x.new <- data.frame(mom.hs = 1, mom.iq = 100)

predict (fit.4, x.new, interval = "confidence",
level=0.95)
```

```
    fit       lwr       upr
 88.24766 86.31365 90.18167
```

fit equation in R

```
fit.4$coef[1]+fit.4$coef[2]*1 + fit.4$coef[3]*100 +
            fit.4$coef[4]*1*100
```

```
(Intercept)
  88.24766
```

example

cognitive test scores

Calculate the cognitive score for a child with a mother who graduated and has an IQ of 100…

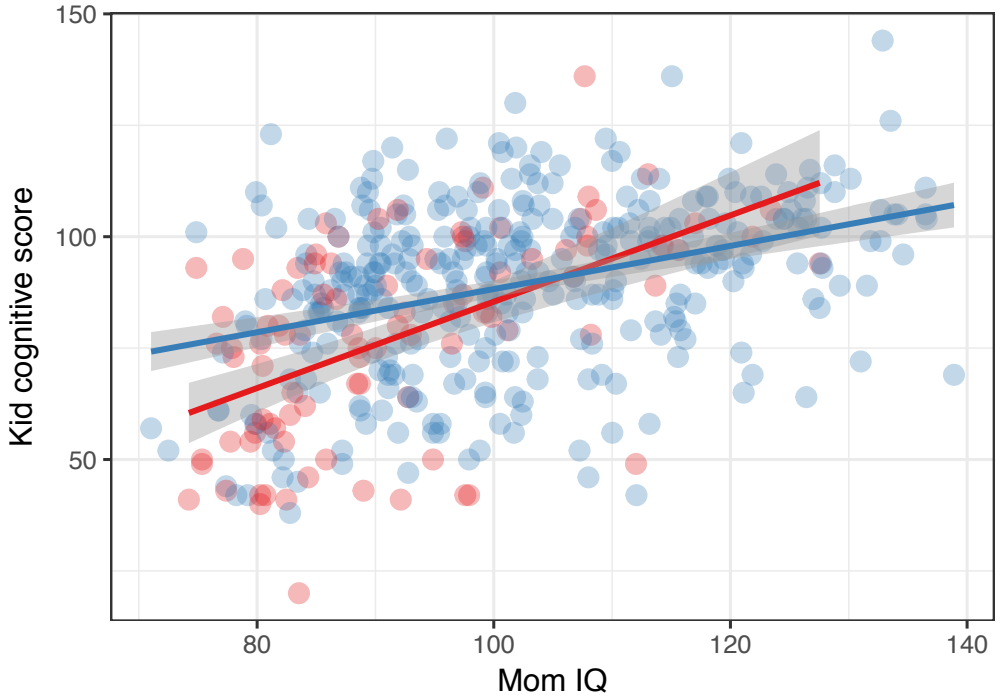
$$Y = -11.48 + 51.27 \cdot HS + 0.97 \cdot IQ - 0.48(HS \cdot IQ)$$

$$\text{no high school: } y = -11.48 + 0.97 \cdot IQ$$

$$\text{high school: } y = 39.79 + 0.97 \cdot IQ - 0.48 \cdot IQ =$$

$$39.79 + 0.49 \cdot IQ$$

interaction: categorical & continuous IV's

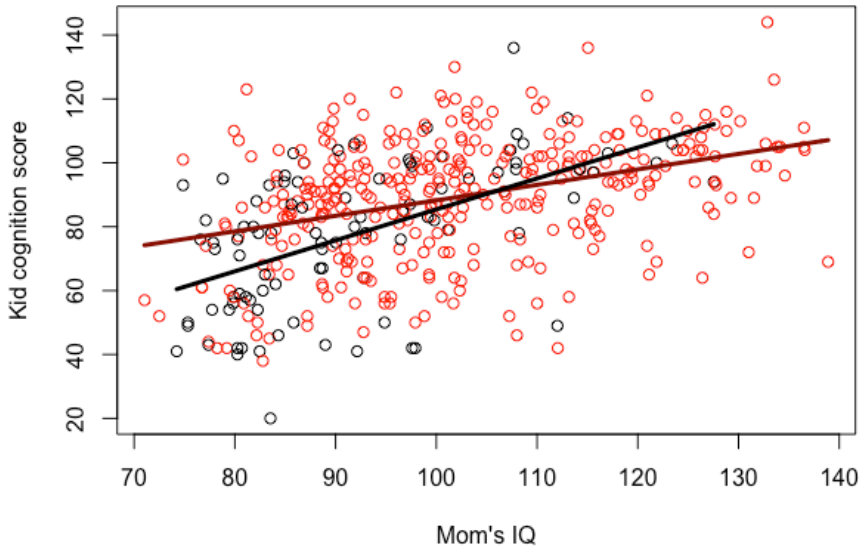


no high school: $y = -11.48 + 0.97 \cdot IQ$

high school: $y = 39.79 + 0.97 \cdot IQ - 0.48 \cdot IQ = 39.79 + 0.49 \cdot IQ$

no high school vs. high school

interaction: categorical & continuous IV's




```
cf <- coef(fit.4)
  r1 <- range(mom.iq[mom.hs == 1])
   r0 <- range(mom.iq[mom.hs == 0])

plot(x = mom.iq, y = kid.score, col = factor(mom.hs),
      ylab = "Kid cognition score", xlab = "Mom's IQ")

curve(cf[1] + cf[2] + (cf[3] + cf[4])*x, from = r1[1],
to = r1[2], add = T, col = "darkred", lwd = 3)

curve(cf[1] + cf[2]*0 + cf[3]*x, from = r0[1], to =
r0[2], add = T, col = "black", lwd = 3)
```

notes on interactions

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \beta_3(X_1)(X_2)$$

X_1 is continuous, X_2 is categorical

- to test if lines are parallel, test if interaction term is significant:

$$H_0 : \beta_3 = 0$$

- interaction term → effect of covariate X_1 on the response is different for different values of X_2

- if interaction term is significant, do not consider the tests for "main effects"

- no interaction term → test if there are two separate parallel lines for two groups, or if one line can describe the data

$$H_0 : \beta_2 = 0$$

example cognitive test scores

interaction: continuous IV's

```
lm(formula = kid.score ~ mom.iq * mom.age)

Coefficients:   Estimate Std. Error t value Pr(>|t|)
(Intercept)    -32.69123   51.67605  -0.633    0.527
mom.iq           1.09947    0.50508   2.177    0.030 *
mom.age          2.55288    2.21404   1.153    0.250
mom.iq:mom.age  -0.02131    0.02155  -0.989    0.323

Residual standard error: 18.26 on 430 degrees of freedom
Multiple R-squared:  0.2054,        Adjusted R-squared:  0.1998
F-statistic: 37.04 on 3 and 430 DF,  p-value: < 2.2e-16
```

unique effect of `mom.iq`

$\beta_1 + \beta_3 \cdot mom.age$

unique effect of `mom.age`

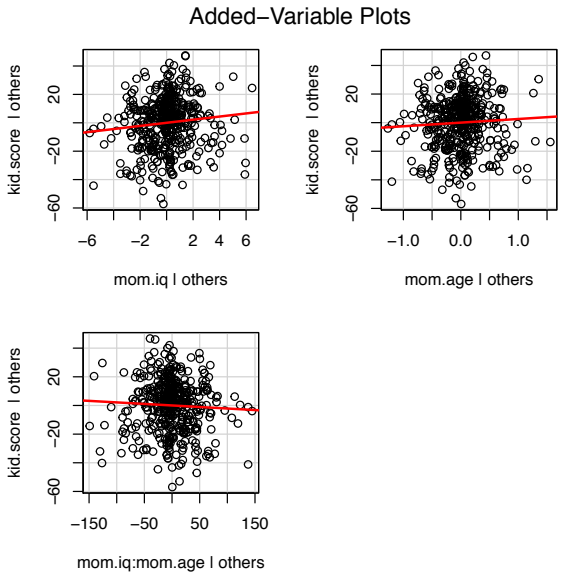
$\beta_2 + \beta_3 \cdot mom.iq$

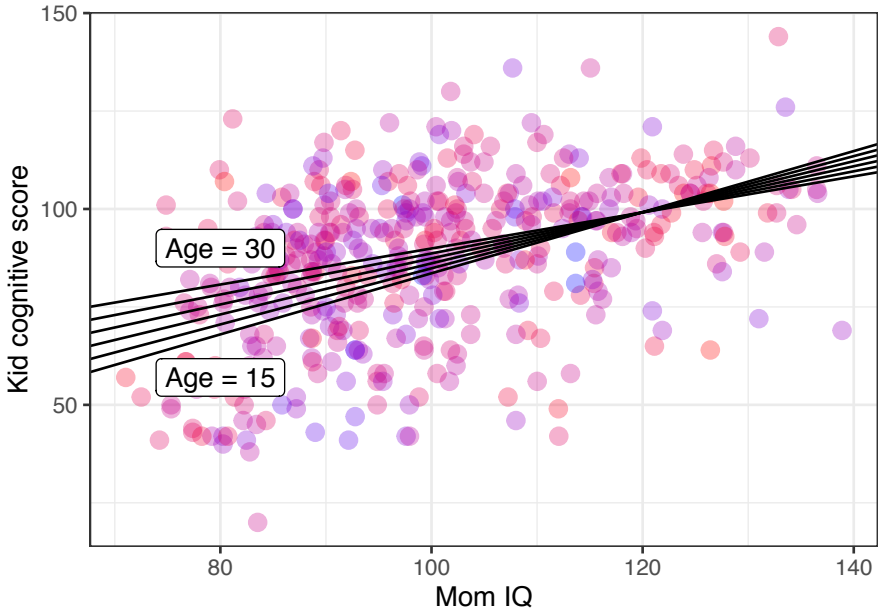
what are your conclusions from this model?

interpret each model coefficient

continuous IV's

- effect of mother IQ on kid cognitive score varies over values of mother age (but not significantly)





scaling/centering IV's

- regression coefficient β_j represents the average difference in *Y* due to a 1-unit change on the j^{th} predictor

- in some cases, a difference of 1 unit on the *X*-scale is not the most relevant comparison

- scale or center IV's to make them more sensible -- does not fundamentally change the model, but may make it more computationally tractable

regression of height vs. earnings

earnings = -61000 + 51 • height(mm) + ε

scaling/centering IV's

- regression coefficient β_j represents the average difference in *Y* due to a 1-unit change on the j^{th} predictor

- in some cases, a difference of 1 unit on the *X*-scale is not the most relevant comparison

- scale or center IV's to make them more sensible -- does not fundamentally change the model, but may make it more computationally tractable

regression of height vs. earnings

earnings = -61000 + 51 • height(mm) + ε
earnings = -61000 + 82,000,000 • height(miles) + ε

$51 doesn't seem to matter, $82,000,000 is a lot!
better in inches?

scaling/centering IV's

- regression coefficient β_j represents the average difference in *Y* due to a 1-unit change on the j^{th} predictor
- in some cases, a difference of 1 unit on the *X*-scale is not the most relevant comparison
- scale or center IV's to make them more sensible -- does not fundamentally change the model, but may make it more computationally tractable

regression of height vs. earnings

earnings = -61000 + 51 • height(mm) + ε
earnings = -61000 + 82,000,000 • height(miles) + ε
earnings = -61000 + 1,295 • height(inches) + ε

$1295 per inch is reasonable for interpretation

scaling/centering IV's

- standardize by centering: each main effect corresponds to a predictive difference with the other input at its average value

```
x.cent <- (x - mean(x))
```

- standardize to z-score: coefficients are units of standard deviations

```
x.z <- (x - mean(x))/sd(x)
```

- standardize using reasonable scales (inches, dollars, years)

centering IV's

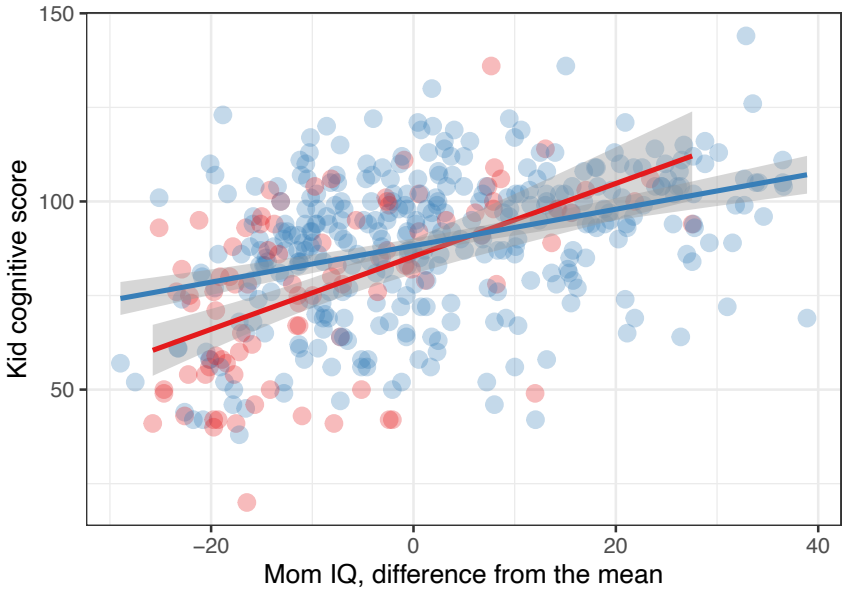
```
mom.iqc <- mom.iq - mean(mom.iq)
summary(lm(kid.score ~ mom.hs + mom.iqc + mom.hs:mom.iqc))
```

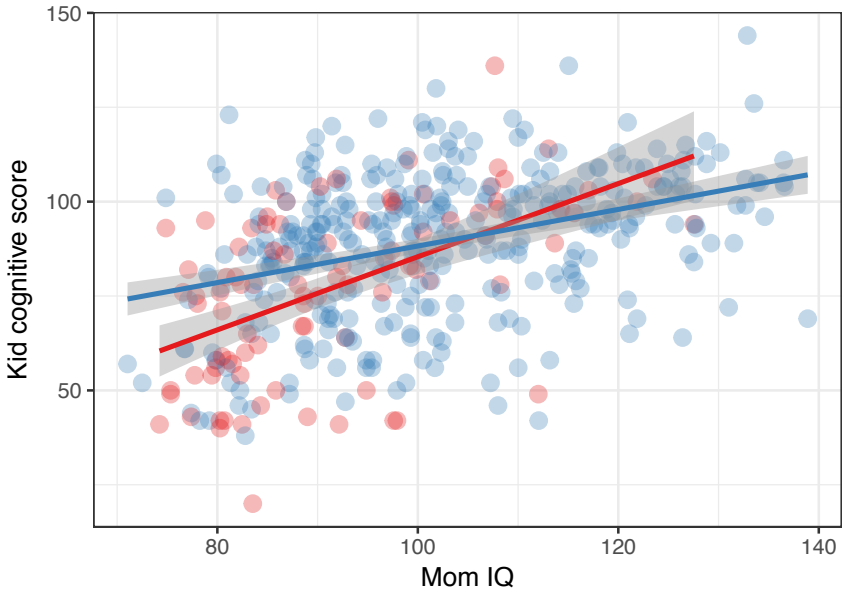
```
lm(formula = kid.score ~ mom.hs + mom.iqc + mom.hs:mom.iqc)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    85.4069     2.2182  38.502  < 2e-16 ***
mom.hs          2.8408     2.4267   1.171  0.24239
mom.iqc         0.9689     0.1483   6.531 1.84e-10 ***
mom.hs:mom.iqc -0.4843     0.1622  -2.985  0.00299 **
---
Residual standard error: 17.97 on 430 degrees of freedom
Multiple R-squared:  0.2301, Adjusted R-squared:  0.2247
F-statistic: 42.84 on 3 and 430 DF,  p-value: < 2.2e-16
```

```
(Intercept)    -11.4820
mom.hs          51.2682
mom.iq           0.9689
mom.hs:mom.iq   -0.4843
```

notes on standardizing IV's





no high school vs. high school

Questions?