Lecture 4 Handout

Elizabeth Colantuoni

2/4/2021

I. Introduction

In real estate, there are three principles: "location, location, location".

In data analysis (empirical science, generally), the corresponding principles are: "question, question, question".

In this lecture, we will look at two questions about Nepali children's growth using the Nepal Children's Anthropomety data kindly provided by Joanne Katz, Professor of International Health and her colleagues.

The questions are:

1. How does the population mean (i.e. average) arm circumference (AC) vary as a function of child's age? Is the AC-age relationship the same for boys and girls?
2. Among children of the same height, how does the population mean AC vary as a function of age and is the relationship the same for boys and girls?

We will address Question 1 in Lecture 3 and Question 2 in Lecture 4.

II. The Data

In this section, we will create the analysis dataset using similar steps as in Lecture 3. We will focus our attention on arm circumference (AC), age, gender and height!

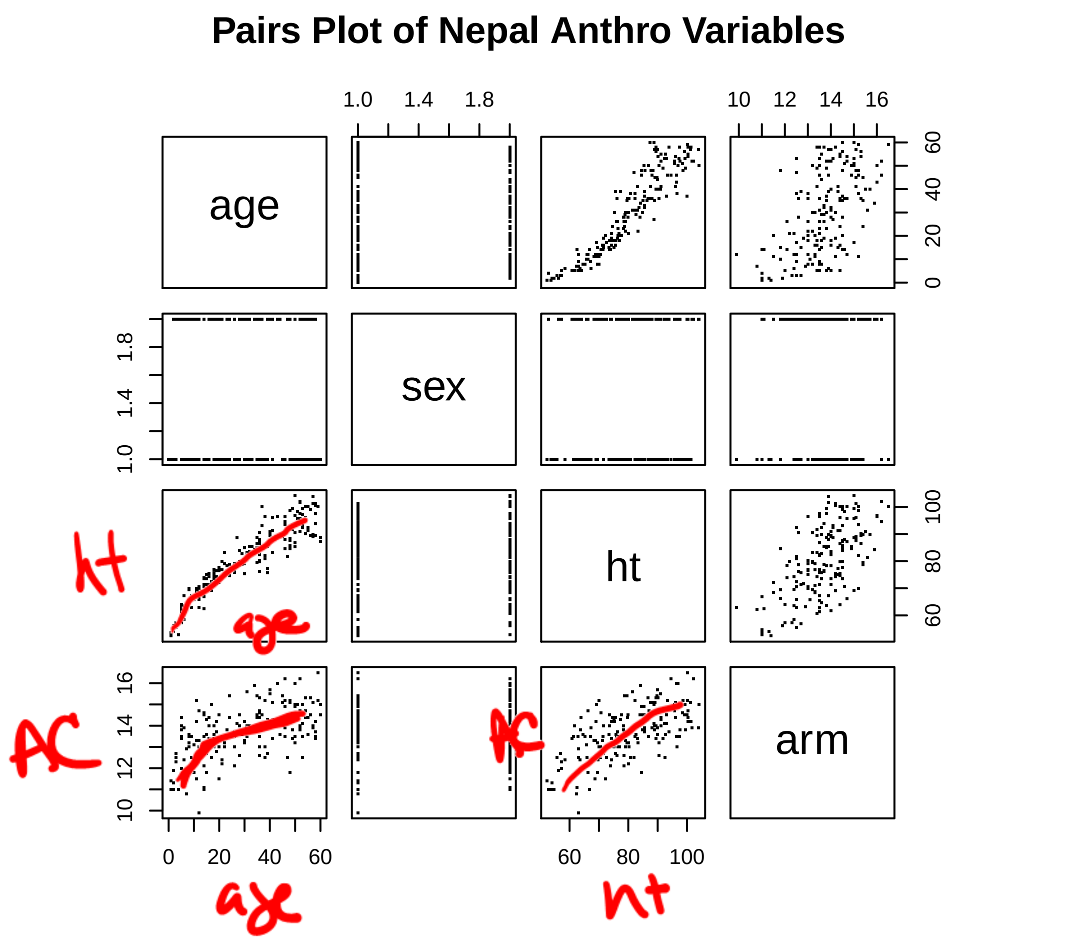
```
load(".\\NepalAnthro.rdata")
d= nepal.anthro %>% select(., arm,age,sex,ht,num) %>% filter(.,num==1)
```

Display key variables

The pairs plot (you find the ggplot version; see ggpairs) is a convenient way to see the pairwise scatterplots in the dataset.

It is a good idea to include the Y and X variables, putting the Y variable last so the bottom row is the plot of Y against each individual X.

```
pairs(select(d,age,sex,ht,arm),pch=".",main="Pairs Plot of Nepal Anthro Variables")
```



Q1: Describe the relationship between a) AC and age, b) AC and height, and c) age and height

Positive and non-linear

- strongest relationship ht vs age

III. Using regression for adjustment

The second question for our analysis is: Among children of the same height, how does the population mean arm circumference vary as a function of age and is the relationship the same for boys and girls?

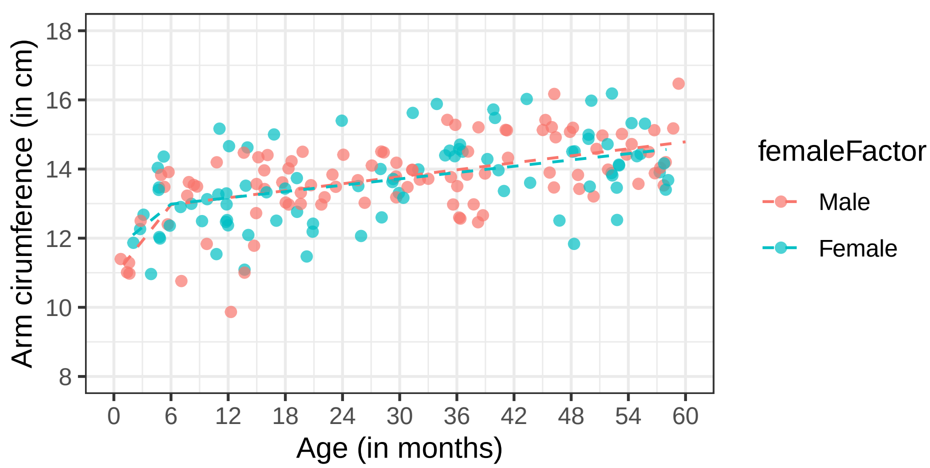
How do we start to explore or visualize "among children of the same height"?

A. A simpler "adjustment" example

Let's take a step back and think about a simpler adjustment to start: Suppose the question was "among children of the same gender, how does the population average arm circumference vary as a function of age?".

We have visualized this in Lecture 3:

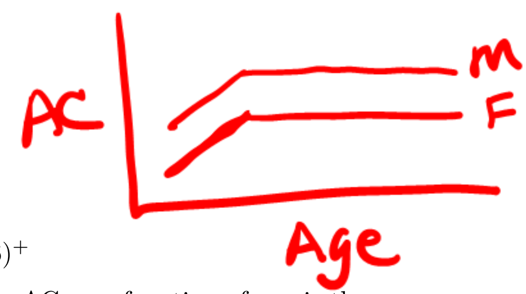
```r
cc=complete.cases(select(d,age,arm))
d.cc=filter(d,cc) %>%
  mutate(female=sex-1,
    agesp6=ifelse(age-6>0, age-6,0),
    int.female.age=female*age,
    int.female.agesp6=female*agesp6,
    femaleFactor = factor(female,levels=c(0,1),labels=c("Male","Female")))
reg4=lm(data=d.cc,arm~female + age + agesp6 + int.female.age + int.female.agesp6)
ggplot(d.cc,aes(x=age, y=arm, color=femaleFactor)) +
    geom_jitter(alpha = 0.7) + theme_bw() +
    geom_line(aes(x=age, y = reg4$fitted.values,
color=femaleFactor),linetype="dashed") +
  scale_y_continuous(breaks=seq(8,18,2),limits=c(8,18)) +
  scale_x_continuous(breaks=seq(0,60,6),limits=c(0,60)) +
  labs(y = "Arm cirumference (in cm)", x = "Age (in months)")
```



Each of the lines in the figure above represents the relationship between arm circumference and age "among children with the same gender".

When we make "adjustment" we assume that the relationship between arm circumference and age is the same after "adjustment for gender"; that is, the slope or rate of change in arm circumference as a function of age is the same regardless of gender.

How can we express this in the form of a model?



$E(AC|age, female) = \beta_0 + \beta_1 female + \beta_2 age + \beta_3 (age - 6)^+$

which we can decompose into:

- Boys: $E(AC|age, female = 0) = \beta_0 + \beta_2 age + \beta_3 (age - 6)^+$
- Girls: $E(AC|age, female = 1) = (\beta_0 + \beta_1) + \beta_2 age + \beta_3 (age - 6)^+$

Notice that in both models, the rate of change for the population mean AC as a function of age is the same for both gender.

Notice that we are allowing the boys and girls to be different from each other in their mean AC; but the rate in which arm circumference changes with age is the same for both genders.

B. Back to our question

Among children of the same height, how does the population mean AC vary as a function of age and is the relationship the same for boys and girls?

1. Coarse adjustment

How should we make the "adjustment" for height?

- We could break height into quintiles? or deciles?

We do this below and compare the coefficients for *age* and *agesp6* unadjusted for height and adjusted for height using quintiles and deciles.

```
d.cc$break5 = cut(d.cc$ht,breaks=quantile(d.cc$ht,seq(0,1,0.2)),labels=seq(1,5))
d.cc$break10 = cut(d.cc$ht,breaks=quantile(d.cc$ht,seq(0,1,0.1)),labels=seq(1,10))
reg.noadj = lm(arm~age+agesp6,data=d.cc)
reg.adj5 = lm(arm~age+agesp6+as.factor(break5),data=d.cc)
reg.adj10 = lm(arm~age+agesp6+as.factor(break10),data=d.cc)
summary(reg.noadj)$coeff;summary(reg.adj5)$coeff;summary(reg.adj10)$coeff
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 11.1208946 0.50958653 21.823369 1.079068e-52
## age          0.3114066 0.09264467  3.361300 9.453805e-04
## agesp6      -0.2795773 0.09441211 -2.961244 3.472606e-03

##                      Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)        11.2873991  0.5676853 19.883198 5.620102e-47
## age                 0.2553234  0.1093653  2.334591 2.068687e-02
## agesp6             -0.2567177  0.1108456 -2.315993 2.170490e-02
## as.factor(break5)2  0.5813649  0.2650206  2.193659 2.956300e-02
## as.factor(break5)3  0.9312077  0.3348778  2.780739 6.010683e-03
## as.factor(break5)4  1.5851476  0.4424987  3.582265 4.400372e-04
## as.factor(break5)5  1.7490374  0.5054182  3.460575 6.755512e-04

##                          Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)           11.71375677  0.5937359 19.7289019 4.912193e-46
## age                    0.08132895  0.1323258  0.6146113 5.396236e-01
## agesp6                -0.08951192  0.1337004 -0.6694964 5.040766e-01
## as.factor(break10)2    0.91551489  0.3835910  2.3866951 1.808602e-02
## as.factor(break10)3    1.18780077  0.3991223  2.9760319 3.340820e-03
## as.factor(break10)4    1.38010626  0.4272217  3.2304220 1.481057e-03
## as.factor(break10)5    1.58765708  0.4444844  3.5719076 4.598088e-04
## as.factor(break10)6    1.86691291  0.5078685  3.6759771 3.163964e-04
## as.factor(break10)7    2.44113837  0.5502725  4.4362355 1.629367e-05
## as.factor(break10)8    2.48070921  0.5975168  4.1516982 5.184795e-05
## as.factor(break10)9    2.50943013  0.6159293  4.0742179 7.037637e-05
## as.factor(break10)10   2.82842001  0.6564933  4.3083762 2.759832e-05
```

Q5 Q4 Q3 Q2 Q1

ht

age

11.1 cm at birth for all children

.31 cm/month

.04 cm/month

11.3 cm at birth among lowest quintile of height

.26 cm/month

0 cm/month

11.7 cm at birth among children in lowest decile of height

.08 cm/month

0 cm/month

Q2: How do the coefficients for age and agesp6 compare without and with adjustment for height?

with adjustment for height, the coefficients are approaching 0

=> after accounting for height, age explains very little about AC

Q3: Is there a way to make a more smooth adjustment for height?

Quintile model

$E(AC \mid age, ht) = \beta_0 + \beta_1 age + \beta_2 agesp6 + \beta_3 I(Q=2) + \beta_4 I(Q=3) + \beta_5 I(Q=4) + \beta_6 I(Q=5)$

3. Smooth adjustment

$\beta_0 + \beta_1 age + \beta_2 agesp6 + ns(height, 3)$

Instead of using the coarse adjustment for height, we could include a smooth function of height. We will try a natural spline (aka natural cubic spline) with 3 degrees of freedom.

```
reg.adjsmooth = lm(arm~age+agesp6+ns(ht,3),data=d.cc)
summary(reg.adjsmooth)
```

```
##
## Call:
## lm(formula = arm ~ age + agesp6 + ns(ht, 3), data = d.cc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6147 -0.6073 -0.0581  0.6893  1.9864
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.602723   0.520775  22.280  < 2e-16 ***
## age         -0.011574   0.157284  -0.074 0.941421
## agesp6      -0.002885   0.157563  -0.018 0.985413
## ns(ht, 3)1   2.729914   0.634244   4.304 2.75e-05 ***
## ns(ht, 3)2   5.433659   1.464733   3.710 0.000277 ***
## ns(ht, 3)3   2.911156   0.644534   4.517 1.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9232 on 179 degrees of freedom
## Multiple R-squared:  0.4127, Adjusted R-squared:  0.3963
## F-statistic: 25.16 on 5 and 179 DF,  p-value: < 2.2e-16
```

Among children of the same height,
$-.01 \approx 0$ Δ in mean AC per month < 6 months
≈ 0 Δ in mean AC per month ≥ 6 months

Q4: What do you conclude regarding the height-adjusted relationship between arm circumference and age?

After accounting for childrens height, AC is not associated with the child's age.

C. Visualization of "adjustment"

How do we visualize the height-adjusted relationship between AC and age?

We can construct the adjusted variable plot; i.e. we want to remove information about height from both arm circumference and age and then examine the relationship with what is left over!

→ Lab 3

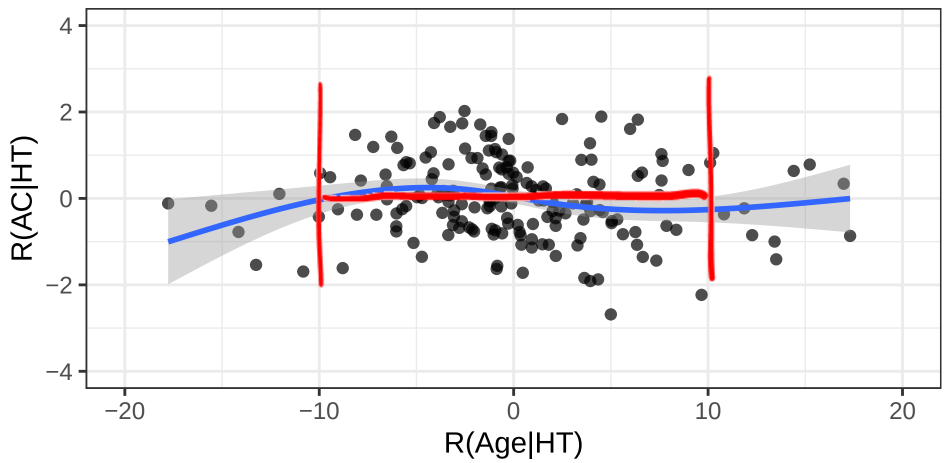
Steps for creating an Adjusted Variable Plot of Y on X1, "controlling for X2,...Xp"

1. Regress Y on X2,...Xp, save residuals as R(Y|X2,...Xp)
2. Regress X1 on X2,...Xp, save residuals as R(X1|X2,...Xp)
3. Plot R(Y|X2,...Xp) vs R(X1|X2,...Xp)

The plot you create in 3. represents the "adjusted" information between Y and X1.

The figure below displays the height adjusted relationship between AC and age.

```
d.cc$resid.arm = lm(arm~ns(ht,3),data=d.cc)$residuals
d.cc$resid.age = lm(age~ns(ht,3),data=d.cc)$residuals
ggplot(d.cc, aes(x = resid.age, y = resid.arm)) +
    geom_jitter(alpha = 0.7) + theme_bw() +
    geom_smooth(method="gam",formula=y ~ splines::ns(x, 3)) +
    scale_y_continuous(breaks=seq(-4,4,2),limits=c(-4,4)) +
    scale_x_continuous(breaks=seq(-20,20,10),limits=c(-20,20)) +
    labs(y = "R(AC|HT)", x = "R(Age|HT)")
```



Q5. What patterns do you see from the adjusted variable plot?

No strong relationship between the height adjust AC and age variables

Q6. Can you identify any challenges in interpreting the adjusted variable plot?

D. Height-adjusted interaction model

The second part of our original question is: is the height-adjusted relationship between arm circumference and age different for boys and girls?

Q7: Can you write out the regression model you want to fit? Call this model the "Model Extended"

$$E(AC \mid age, F, height) = \beta_0 + \beta_1 age + \beta_2 agesp6 + \beta_3 F + \beta_4 F\,age + \beta_5 F\,agesp6 + ns(height, 3)$$

Q8: What model do you want to compare "Model Extended" to to answer the question?

Null

$$E(AC \mid age, height) = \beta_0 + \beta_1 age + \beta_2 agesp6 + ns(height, 3)$$

1. Fit adjusted interaction model

Now fit the models you specified above:

```
reg.adjsmoothint = lm(arm~female + age + agesp6 +
                        int.female.age +
                        int.female.agesp6 +
                        ns(ht,3),data=d.cc)
summary(reg.adjsmooth)$coeff;summary(reg.adjsmoothint)$coeff
```

null model

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 11.602722772  0.5207747 22.27973387 1.683908e-53
## age         -0.011574130  0.1572838 -0.07358756 9.414207e-01
## agesp6      -0.002884691  0.1575633 -0.01830813 9.854134e-01
## ns(ht, 3)1   2.729913677  0.6342437  4.30420288 2.754576e-05
## ns(ht, 3)2   5.433659247  1.4647330  3.70965842 2.767626e-04
## ns(ht, 3)3   2.911156437  0.6445338  4.51668525 1.137171e-05
```

extended model

```
##                      Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)       11.322029922  0.6030362 18.77504060 6.955038e-44
## female             1.010698863  1.0654978  0.94856964 3.441402e-01
## age               -0.002852731  0.1663440 -0.01714959 9.863367e-01
## agesp6            -0.008886066  0.1671911 -0.05314916 9.576733e-01
## int.female.age    -0.139329123  0.1929743 -0.72200872 4.712473e-01
## int.female.agesp6  0.133021858  0.1963867  0.67734661 4.990754e-01
## ns(ht, 3)1         2.834123148  0.6467146  4.38233994 2.013673e-05
## ns(ht, 3)2         5.731490323  1.4895033  3.84792064 1.664379e-04
## ns(ht, 3)3         2.974559707  0.6512404  4.56752940 9.250705e-06
```

Q9. Do you think the data supports the hypothesis that the height adjusted relationship between the population mean AC and age differs for male and female children?

data supports the null model compared to the model allowing for a sex-specific height-adjusted AC vs age

E.Summarize your findings!

→ There is an unadjusted relationship between mean AC and age

→ After adjusting for height, age provides no additional information about AC

→ no evidence of gender-specific unadjusted or height adjusted relationship