Lecture 3

Elizabeth Colantuoni and Jiawei Bai

1/29/2021

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 Lectures 3 and 4, we will look at two questions about Nepali children's growth using the Nepal Children's Anthropometry 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 read in the data and perform some basic processing and data visualizations to prepare for the analysis.

A. Read in and look at the data

```
load("NepalAnthro.rdata")
dim(nepal.anthro)
```

```
## [1] 1000   30
```

```
head(nepal.anthro)
```

```
##       id sex   wt    ht  arm bf day month year mage lit died alive t2 age num
## 1 120011   1 12.8  91.2 14.3  0  15     9   46   35   0    2     5  9  41   1
## 2 120011   1 12.8  93.9 13.5  0  22     1   47   35   0    2     5 13  45   2
## 3 120011   1 13.1  95.2 14.5  0   3     5   47   35   0    2     5 17  49   3
## 4 120011   1 13.8  96.9 14.1  0   6     9   47   35   0    2     5 21  53   4
## 5 120011   1   NA    NA   NA  0  20     1   48   35   0    2     5 25  57   5
## 6 120012   2 14.9 103.9 13.9  0  15     9   46   35   0    2     5  9  57   1
##   xmin agemin agemax age_sp1 age_sp2 age_sp3 age1 age2 age3 age4   armpr1
## 1   41     41     57      29      17       0    0    0   41    0 14.15903
## 2   41     41     57      33      21       0    0    0   45    0 14.30622
## 3   41     41     57      37      25       1    0    0    0   49 14.42511
## 4   41     41     57      41      29       5    0    0    0   53 14.45913
## 5   41     41     57      45      33       9    0    0    0   57 14.49315
## 6   57     57     73      45      33       9    0    0    0   57 14.57972
##   futime fuvisit fuvisit1
## 1      1       0        0
## 2      5       1        1
## 3      9       2        2
## 4     13       3        3
## 5     17       4        4
## 6      1       0        0
```

B. Analysis sample

Extract the key variables we need for our analysis and only the first row of data for each child. NOTE: The data provides multiple observations over time for each child. Initially, we will evaluate the first assessment (baseline) for each child.

```
d= nepal.anthro %>% select(.,age,sex,ht,wt,arm,num) %>% filter(.,num==1)
d <- d[complete.cases(d),] # drop cases without one or more of these variables
d <- d[order(d$age),-6] # Sort the data by age and drop "num"
dim(d)
```

```
## [1] 185   5
```

```
head(d)
```

```
##     age sex   ht  wt  arm
## 446   1   1 52.4 4.1 11.4
## 671   1   1 53.6 3.8 11.0
## 241   2   2 56.2 4.6 11.9
## 261   2   1 54.7 4.1 11.0
## 301   2   1 54.1 4.1 11.3
## 196   3   2 57.0 5.2 12.7
```

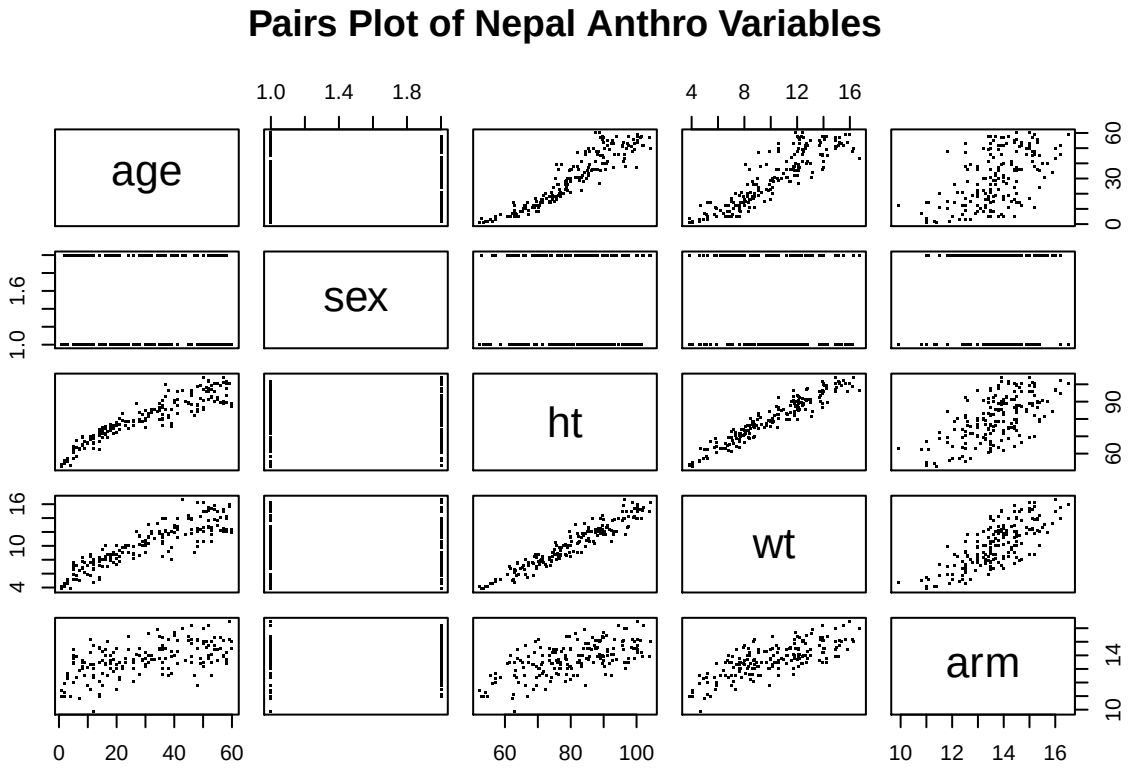
C. Display key variables

You should ALWAYS visualize your data. One quick approach is to make pairwise scatterplots where you visualize the association between each pair of variables.

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

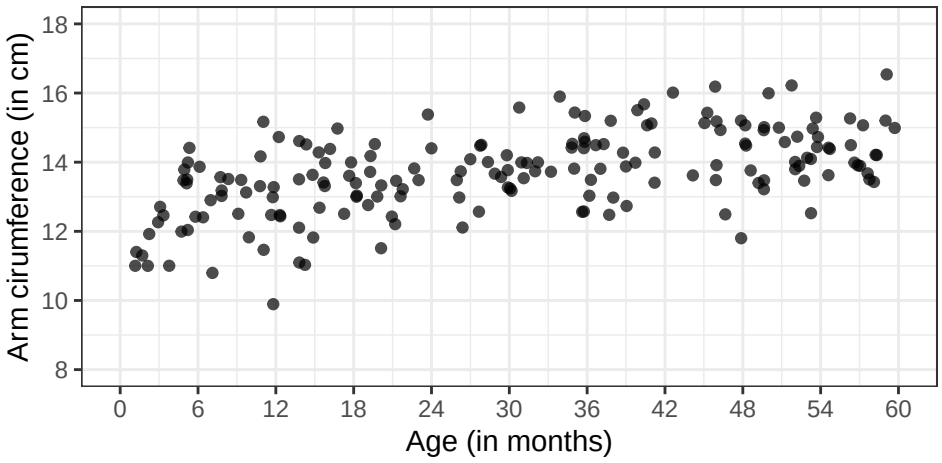
It is a good idea to include the Y and X variables, putting the Y variable last so the bottom row is the plot of Y against each individual X.

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



For Question 1, we will focus on AC and age. Here we make a plot to take a closer look the relationship between these two variables.

```
ggplot(d, aes(x = age, y = arm)) +
    geom_jitter(alpha = 0.7) +
    theme_bw() +
    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)")
```



Q1: Describe the relationship between AC and age.

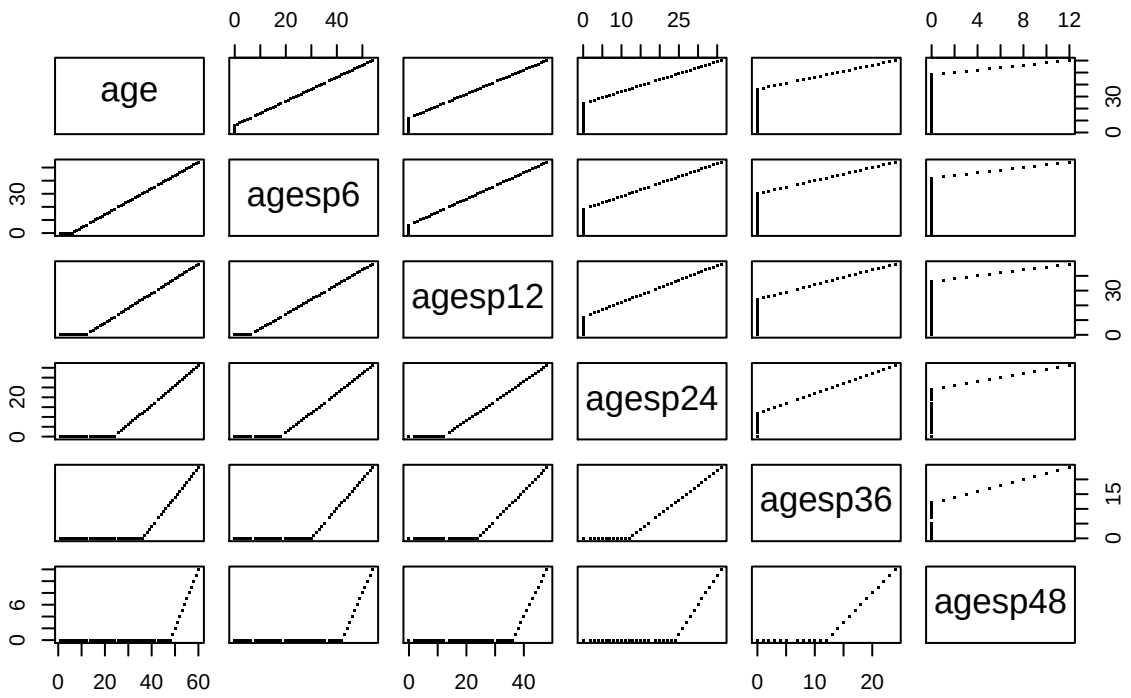
III. Define, fit and interpret a linear spline model

In this section, we will define the necessary variables for, fit and interpret a linear spline model to the relationship between population mean AC and age.

A. Define the spline terms for inclusion in model

We will start with knots at 6, 12, 24, 36, 48 months and then drop unnecessary knot points (i.e. where allowing for a chane in slope is not necessary).

```
d=mutate(d,
agesp6=ifelse(age-6>0, age-6,0),
agesp12=ifelse(age-12>0, age-12,0),
agesp24=ifelse(age-24>0, age-24,0),
agesp36=ifelse(age-36>0,age-36,0),
agesp48=ifelse(age-48>0, age-48,0)
)
# check what predictors for linear splines look like
pairs(select(d,age,agesp6,agesp12,agesp24,agesp36, agesp48),pch=".",
main="Pairs Plot of Age Linear Spline Vars")
```



Fit, interpret and visualize the linear spline regression model

Include all the spline terms at once allowing for 5 slope changes over the 5 years of age, call this model Model 1.

Q2: Write down the mathematical representation for the model allowing for 5 slope changes over the 5 years of age.

```r
cc=complete.cases(select(d,age,arm))
d.cc=filter(d,cc)
d.cc = arrange(d.cc,age)
reg1<-lm(data=d.cc, arm~age+agesp6+agesp12+agesp24+agesp36+agesp48)
summary.lm(reg1)
```

```
##
## Call:
## lm(formula = arm ~ age + agesp6 + agesp12 + agesp24 + agesp36 +
##     agesp48, data = d.cc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.00372 -0.58480  0.06702  0.60947  2.24463
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.942480   0.546698  20.016  < 2e-16 ***
## age          0.378517   0.119322   3.172  0.00178 **
## agesp6      -0.430159   0.166500  -2.584  0.01058 *
## agesp12      0.109313   0.085560   1.278  0.20305
## agesp24     -0.009497   0.054665  -0.174  0.86227
## agesp36     -0.030832   0.048821  -0.632  0.52851
## agesp48     -0.013044   0.055423  -0.235  0.81421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9689 on 178 degrees of freedom
## Multiple R-squared:  0.3567, Adjusted R-squared:  0.3351
## F-statistic: 16.45 on 6 and 178 DF,  p-value: 4.792e-15
```

Q3: Interpret the value of the intercept

Q4: Interpret the coefficient for "age"

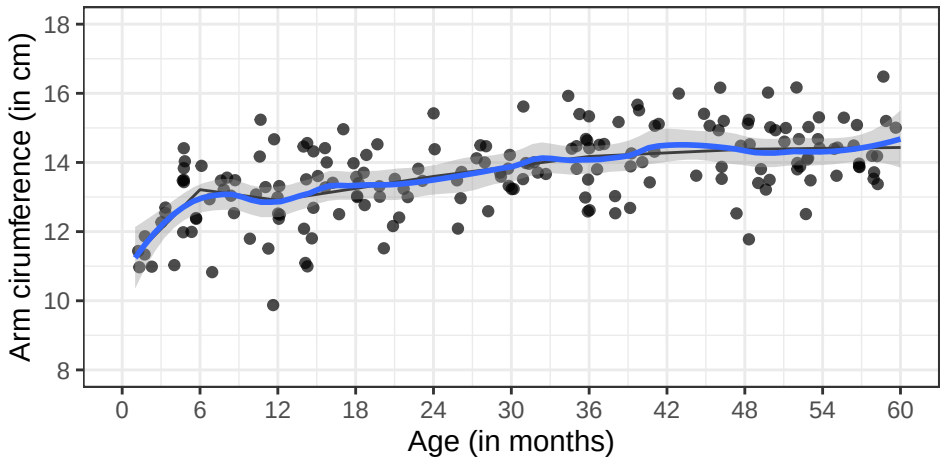
Q5: Interpret the coefficient for "agesp6"

Q6: What is the estimate of the population average, standard deviation and variance of AC among 12-month old children?

The figure below overlays the observed data with the estimated population mean AC at each age.

```
ggplot(d.cc, aes(x = age, y = arm)) +
  theme_bw() +
  geom_jitter(alpha = 0.7) +
  geom_line(aes(x = age, y = reg1$fitted.values)) +
  geom_smooth(span=0.3) +
  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)")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Q7: Do you think your Model 1 is a useful model? Are there any trends in how the estimated population mean AC changes with age that you find unusual or unlikely?

IV. Fine-tuning the model

Because there is little reason to think that mean AC goes down between ages 6 and 12 months and because there is little evidence in the data supporting changes in slope after 12 months, let's fit a second model with only two slope changes at 6 and 12 months, then a third with only a break at 6 months.

```
reg2<-lm(data=d.cc, arm~age+agesp6+agesp12)
reg3<-lm(data=d.cc, arm~age+agesp6)
summary.lm(reg2); summary.lm(reg3)
```

```
##
## Call:
## lm(formula = arm ~ age + agesp6 + agesp12, data = d.cc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2229 -0.6229  0.0499  0.6271  2.0703
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

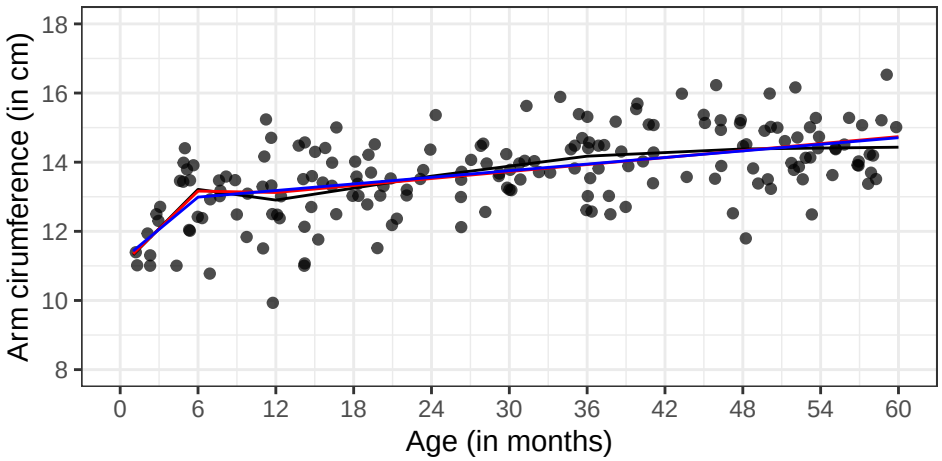
```
## (Intercept) 10.98212    0.54772  20.051  < 2e-16 ***
## age          0.36360    0.11920   3.050  0.00263 **
## agesp6      -0.37041    0.16093  -2.302  0.02249 *
## agesp12      0.04048    0.05804   0.697  0.48640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9721 on 181 degrees of freedom
## Multiple R-squared:  0.3415, Adjusted R-squared:  0.3306
## F-statistic: 31.29 on 3 and 181 DF,  p-value: 2.41e-16
```

```
##
## Call:
## lm(formula = arm ~ age + agesp6, data = d.cc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2803 -0.5988 -0.0076  0.6375  2.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.12089    0.50959  21.823  < 2e-16 ***
## age          0.31141    0.09264   3.361 0.000945 ***
## agesp6      -0.27958    0.09441  -2.961 0.003473 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9708 on 182 degrees of freedom
## Multiple R-squared:  0.3397, Adjusted R-squared:  0.3325
## F-statistic: 46.83 on 2 and 182 DF,  p-value: < 2.2e-16
```

Q8: Without doing a formal statistical test, consider the size and standard error of the estimate for "agesp12" and make a recommendation about the need to allow the AC vs. age relationship to vary comparing children 6 to 12 months of age to children over 12 months of age.

Make a figure of the observed data and the 3 linear spline models.

```
ggplot(d.cc, aes(x = age, y = arm)) + theme_bw() +
  geom_jitter(alpha = 0.7) +
  geom_line(aes(x = age, y = reg1$fitted.values),color="black") +
  geom_line(aes(x = age, y = reg2$fitted.values),color="red") +
  geom_line(aes(x = age, y = reg3$fitted.values),color="blue") +
  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)")
```



Q9: How does the population mean AC vary as a function of child's age? Write an answer using the results of your favorite model among 1-3. Write in scientific terms, use units, be numerate. This is an exercise to put into your own words the results of a simple regression analysis.

V. Does the AC vs. age relationship vary by gender?

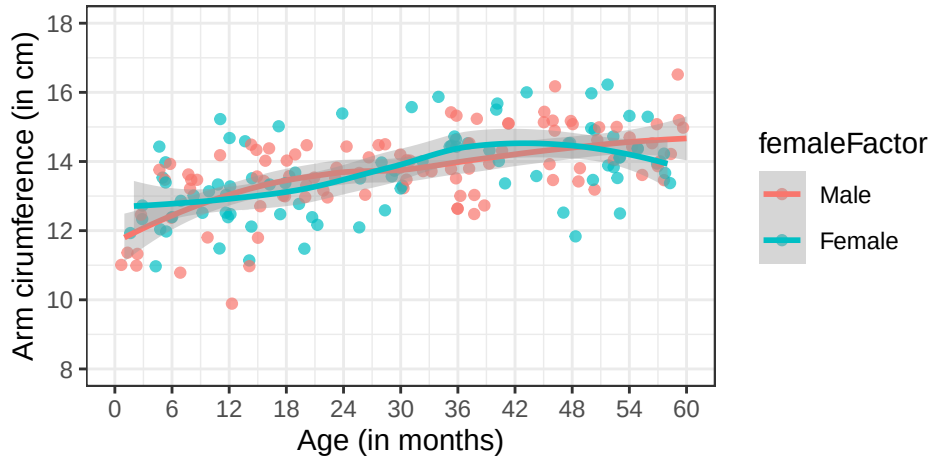
Now we want to explore whether the relationship between the population mean AC vs. age differs for males vs. females.

A. Visualize the question

Start with a visualization: plot the association between mean AC and age, separately for each gender.

```
d.cc$female=d.cc$sex-1
d.cc$femaleFactor = factor(d.cc$female,levels=c(0,1),labels=c("Male","Female"))
ggplot(d.cc,aes(x=age, y=arm, color=femaleFactor)) +
  theme_bw() + geom_jitter(alpha = 0.7) +
  geom_smooth() +
  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)")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Q10: From your visualization, do you think the population mean AC vs. age relationship differs by gender?

B. Fit Model 3, separately for each gender

To numerically explore the question, we can fit Model 3 separately to the data for females and males.

```
reg3.boy=lm(arm~age+agesp6,data=d.cc,subset=female==0)
reg3.girl=lm(arm~age+agesp6,data=d.cc,subset=female==1)
summary.lm(reg3.boy); summary.lm(reg3.girl)
```

```
##
## Call:
## lm(formula = arm ~ age + agesp6, data = d.cc, subset = female ==
##     0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2701 -0.5186  0.0951  0.6614  1.8845
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.8864     0.5720  19.034  < 2e-16 ***
## age           0.3469     0.1044   3.324  0.00126 **
## agesp6       -0.3132     0.1067  -2.936  0.00416 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9083 on 96 degrees of freedom
## Multiple R-squared:  0.414,  Adjusted R-squared:  0.4018
## F-statistic: 33.91 on 2 and 96 DF,  p-value: 7.235e-12
```

```
##
## Call:
## lm(formula = arm ~ age + agesp6, data = d.cc, subset = female ==
##     1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.46193 -0.66229 -0.03179  0.61947  2.06398
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.6457     0.9999  11.647   <2e-16 ***
## age           0.2231     0.1811   1.232    0.222
## agesp6       -0.1927     0.1841  -1.047    0.298
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.05 on 83 degrees of freedom
## Multiple R-squared:  0.264,  Adjusted R-squared:  0.2462
## F-statistic: 14.88 on 2 and 83 DF,  p-value: 2.994e-06
```

Q11: Based on the fit of the two models, do you think the hypothesis that the population mean AC vs. age relationship differs by gender is supported by the data?

C. Interaction model

Instead of fitting a gender-specific model, we can fit a single model to address the question of interest.

Q12: Write out the mathematical model representing the hypothesis that the population mean AC vs. age relationship differs by gender, where the mean AC vs. age is given by a linear spline model with knot at 6-months of age.

To fit the model above, we will create interaction terms in the model to allow for a separate AC and age association for each gender, will allow us to define a hypothesis test(s) to determine if there are any differences by gender.

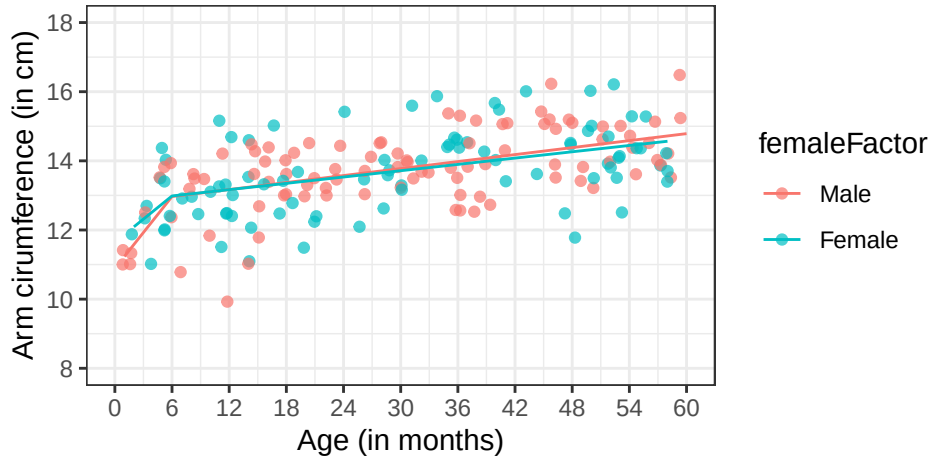
```
d.cc$int.female.age=d.cc$female*d.cc$age
d.cc$int.female.agesp6=d.cc$female*d.cc$agesp6
reg4=lm(data=d.cc,arm~female + age + agesp6 + int.female.age + int.female.agesp6)
summary.lm(reg4)
```

```
##
## Call:
## lm(formula = arm ~ female + age + agesp6 + int.female.age + int.female.agesp6,
##     data = d.cc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2701 -0.5803  0.0116  0.6336  2.0640
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        10.8864     0.6149  17.705  < 2e-16 ***
## female              0.7593     1.1149   0.681  0.49675
## age                 0.3469     0.1122   3.092  0.00231 **
## agesp6             -0.3132     0.1147  -2.731  0.00694 **
## int.female.age     -0.1239     0.2024  -0.612  0.54140
## int.female.agesp6   0.1206     0.2061   0.585  0.55914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9765 on 179 degrees of freedom
## Multiple R-squared:  0.3429, Adjusted R-squared:  0.3245
## F-statistic: 18.68 on 5 and 179 DF,  p-value: 6.42e-15
```

D. Visualize the model fit

The figure below displays the estimated gender-specific relationship between the population mean AC and age.

```
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)) +
  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)")
```



VI. Summarize your findings

Q13 Based on your interaction model, express the following hypotheses with respect to coefficients from your regression model.

- The rate of change in AC as a function of age differs by gender

- The mean AC as a function of age differs for male to female children.

Q14: Summarize your analyses to address the question: how does the population mean AC vary as a function of child's age? Is the relationship between AC and age the same for boys and girls? Write to a public health audience; no unnecessary statistical jargon; be numerate.