Visualization with scatterplots

ANALYZING SURVEY DATA IN R

Kelly McConville

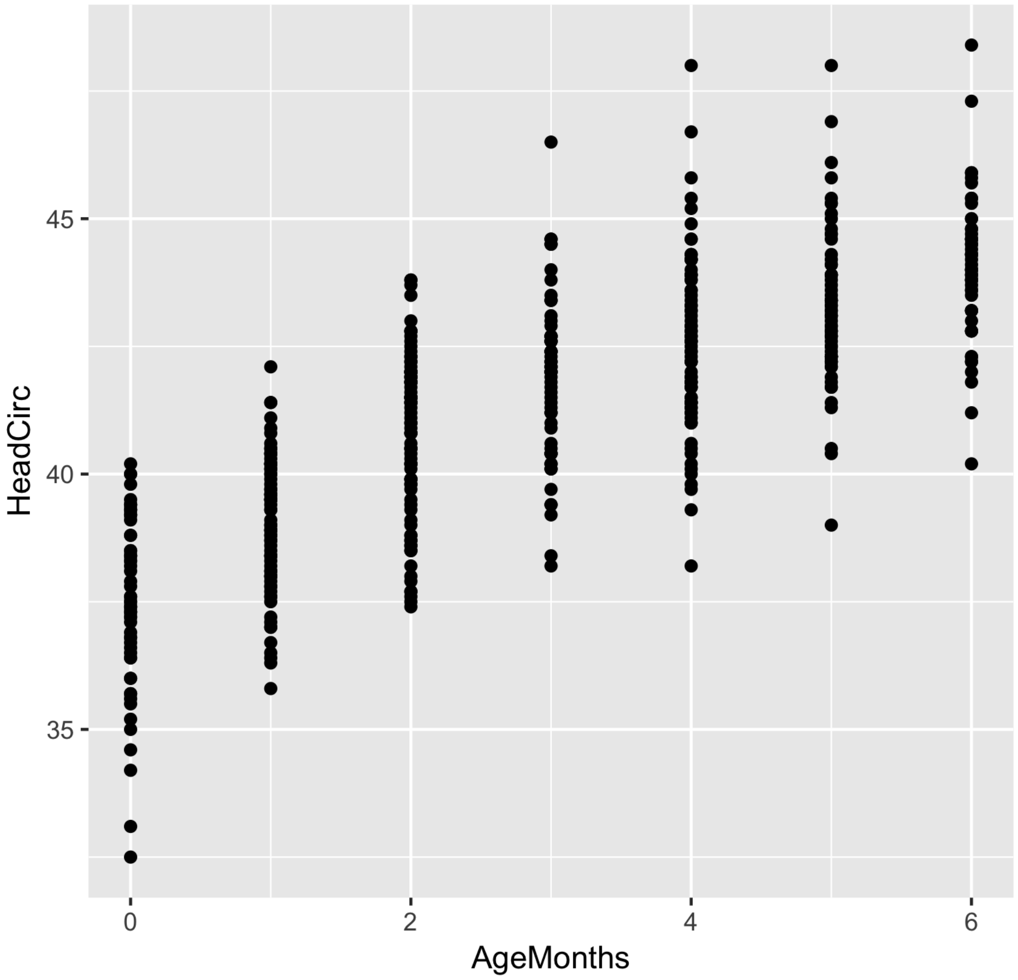
Assistant Professor of Statistics

```r
babies <- filter(NHANESraw, AgeMonths <= 6) %>%
  select(AgeMonths, HeadCirc)
babies
```

```
# A tibble: 484 x 2
   AgeMonths HeadCirc
       <int>    <dbl>
 1         3     42.7
 2         4     42.8
 3         2     38.8
 4         0     36.0
 5         5     42.7
 6         2     41.9
 7         6     44.3
 8         3     42.0
 9         2     41.3
10         1     38.9
# ... with 474 more rows
```

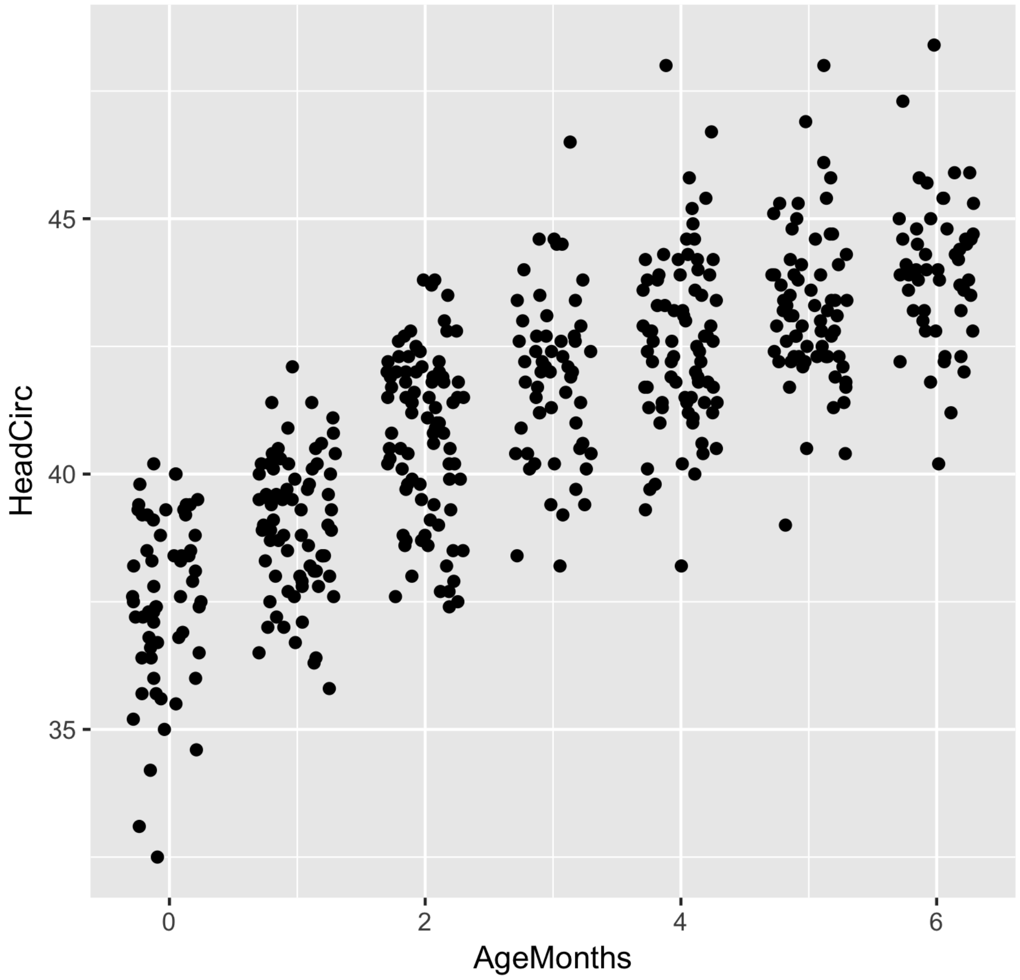
Scatterplots

```
ggplot(data = babies, mapping = aes(x = AgeMonths, y = HeadCirc)) +
  geom_point()
```



Jittering

```
ggplot(data = babies, mapping = aes(x = AgeMonths, y = HeadCirc)) +
  geom_jitter(width = 0.3, height = 0)
```

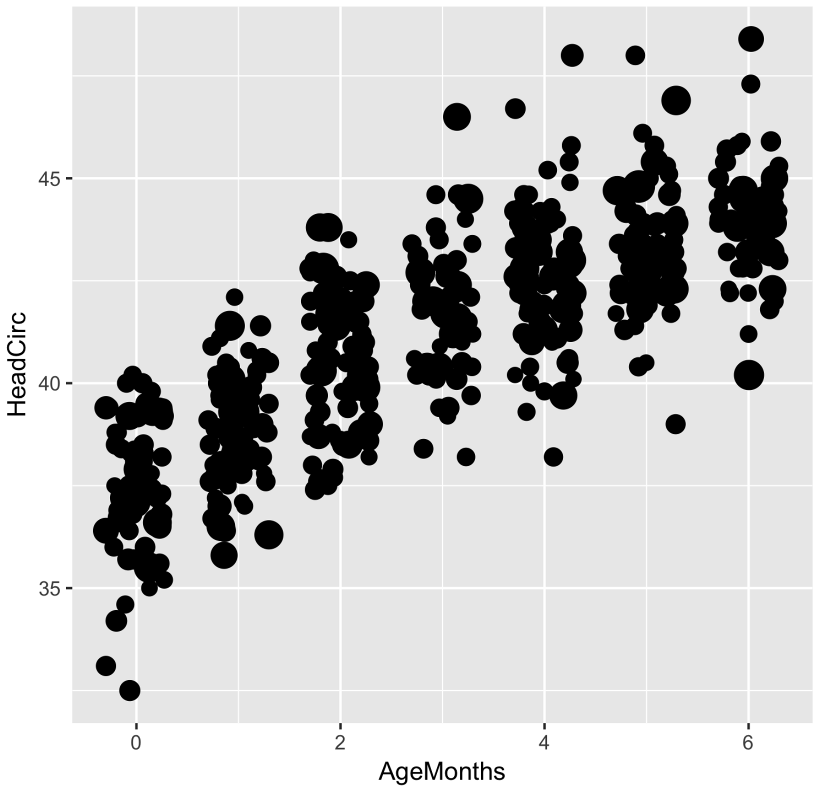


```r
babies <- filter(NHANESraw, AgeMonths <= 6) %>%
  select(AgeMonths, HeadCirc, WTMEC4YR)
babies
```

```
# A tibble: 484 x 3
   AgeMonths HeadCirc WTMEC4YR
       <int>    <dbl>    <dbl>
 1         3     42.7    12915
 2         4     42.8    12791
 3         2     38.8     2359
 4         0     36.0     4306
 5         5     42.7     2922
 6         2     41.9     5561
 7         6     44.3    10416
 8         3     42.0     9957
 9         2     41.3     4503
10         1     38.9     3718
# ... with 474 more rows
```

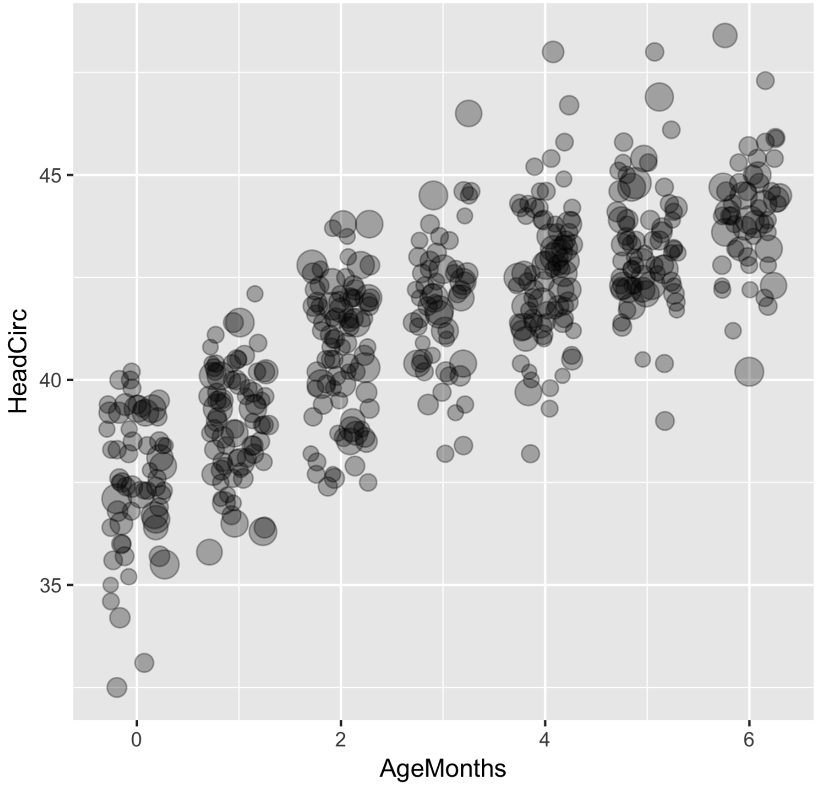
Bubble plots

```
ggplot(data = babies, mapping = aes(x = AgeMonths, y = HeadCirc,
                                    size = WTMEC4YR)) +
  geom_jitter(width = 0.3, height = 0) +
  guides(size = FALSE)
```



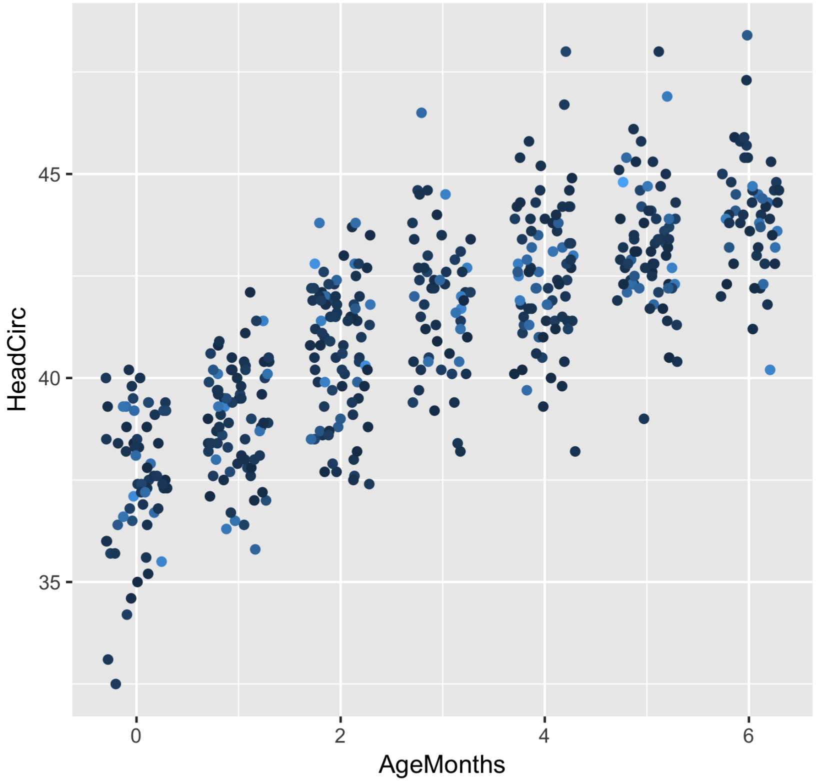
Bubble plots

```
ggplot(data = babies, mapping = aes(x = AgeMonths, y = HeadCirc,
                                    size = WTMEC4YR)) +
  geom_jitter(width = 0.3, height = 0, alpha = 0.3) +
  guides(size = FALSE)
```



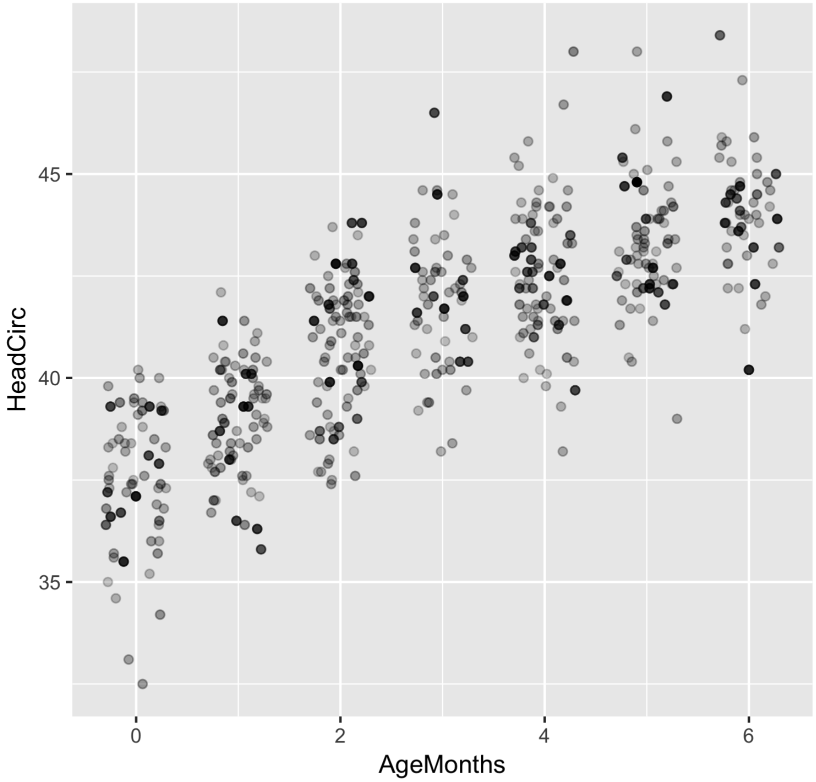
Survey-weighted scatterplots

```
ggplot(data = babies, mapping = aes(x = AgeMonths, y = HeadCirc,
                                    color = WTMEC4YR)) +
  geom_jitter(width = 0.3, height = 0) +
  guides(color = FALSE)
```



Survey-weighted scatterplots

```
ggplot(data = babies, mapping = aes(x = AgeMonths, y = HeadCirc,
                                    alpha = WTMEC4YR)) +
  geom_jitter(width = 0.3, height = 0) +
  guides(alpha = FALSE)
```



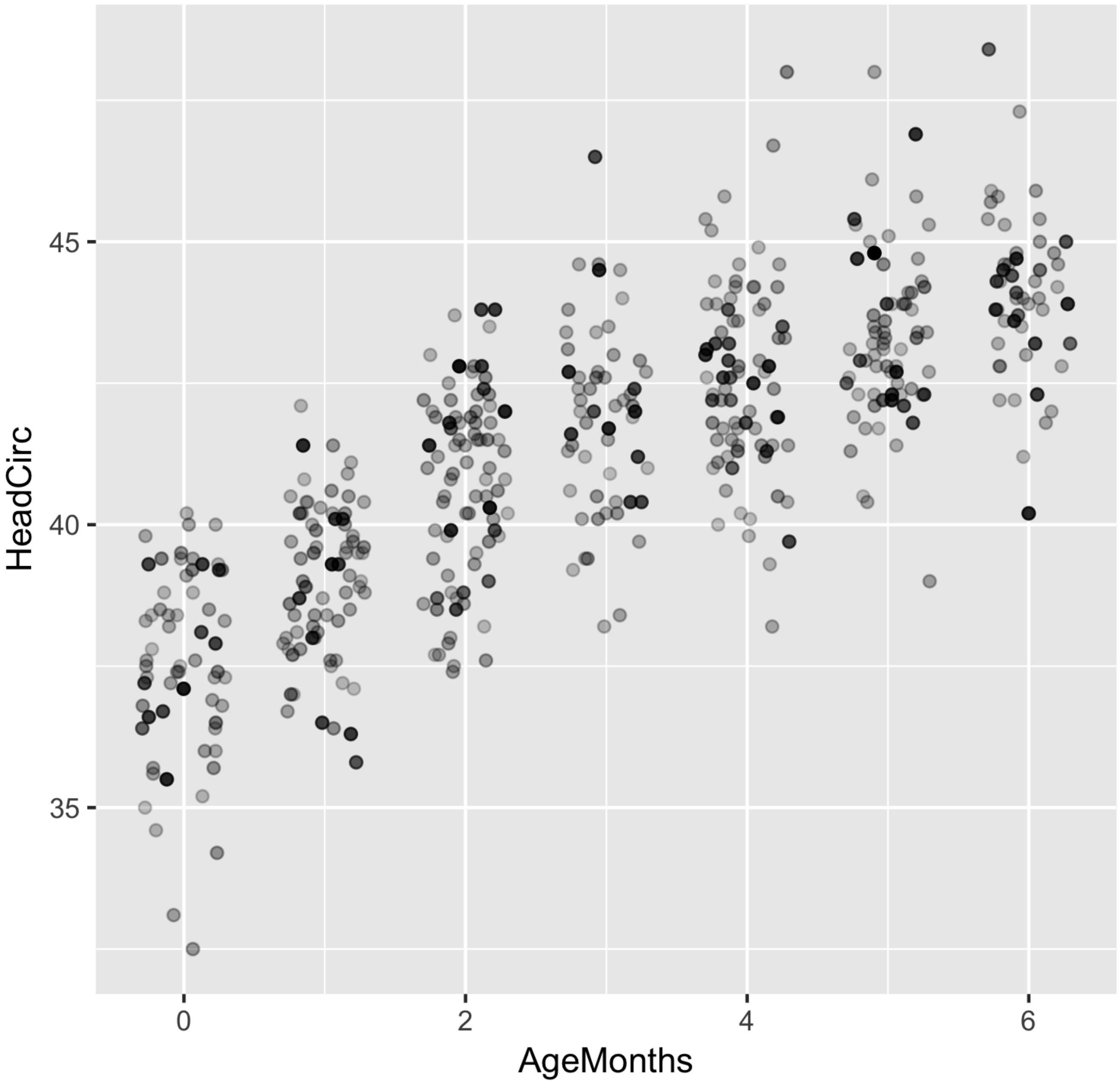
Let's practice!

Visualizing trends

ANALYZING SURVEY DATA IN R

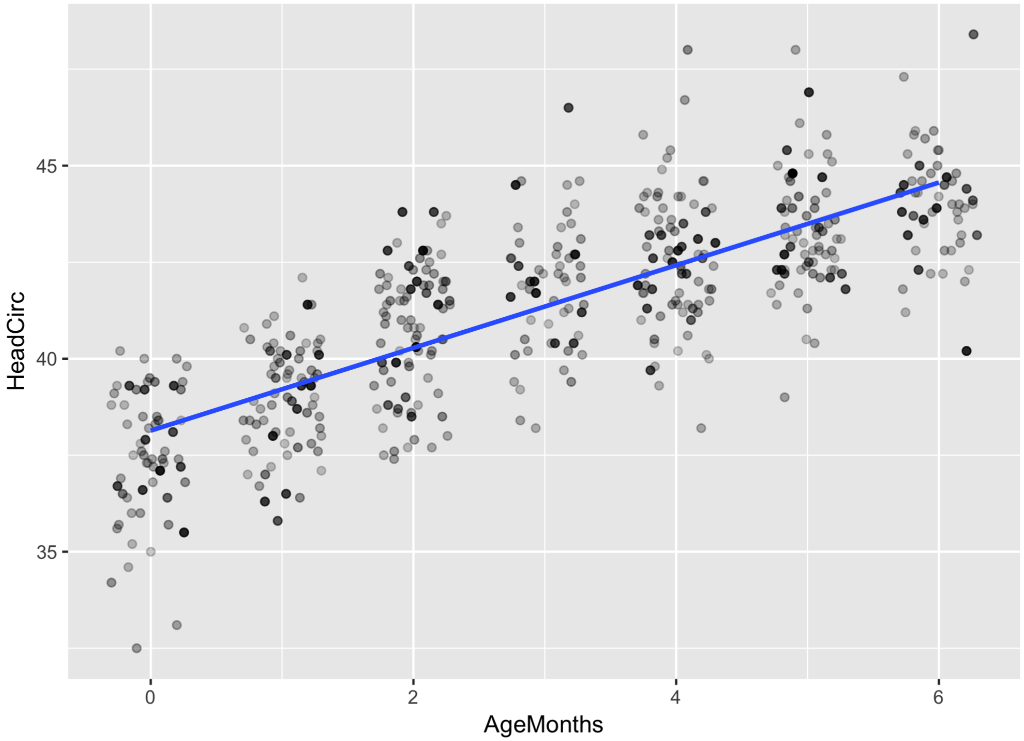
Kelly McConville
Assistant Professor of Statistics

Scatter plots



Survey-Weighted Line of Best Fit

```
ggplot(data = babies, mapping = aes(x = AgeMonths, y = HeadCirc,
                                    alpha = WTMEC4YR)) +
  geom_jitter(width = 0.3, height = 0) + guides(alpha = FALSE) +
  geom_smooth(method = "lm", se = FALSE, mapping = aes(weight = WTMEC4YR))
```

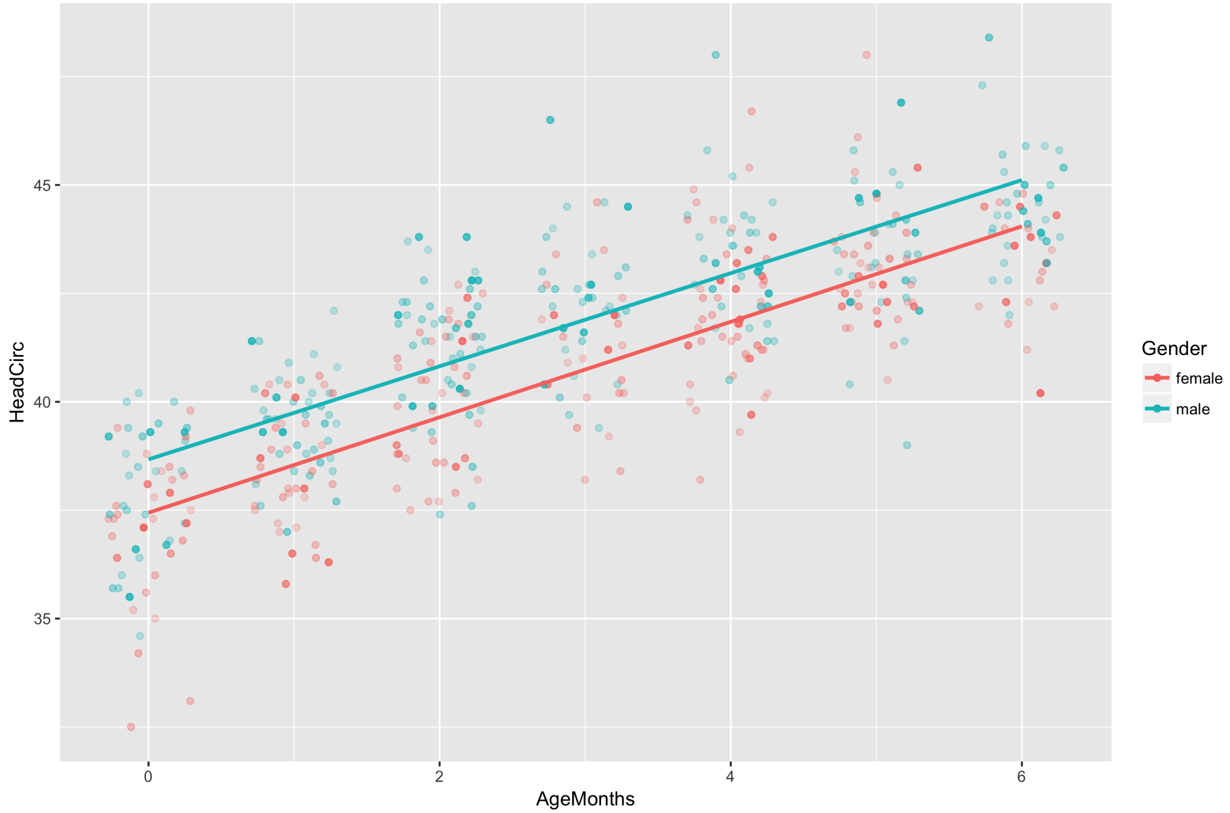


```r
babies <- filter(NHANESraw, AgeMonths <= 6) %>%
  select(AgeMonths, HeadCirc, WTMEC4YR, Gender)
babies
```

```
# A tibble: 484 x 4
   AgeMonths HeadCirc WTMEC4YR Gender
       <int>    <dbl>    <dbl> <fct>
 1         3     42.7   12915. male
 2         4     42.8   12791. female
 3         2     38.8    2359. female
 4         0     36.0    4306. female
 5         5     42.7    2922. female
 6         2     41.9    5561. male
 7         6     44.3   10416. female
 8         3     42.0    9957. female
 9         2     41.3    4503. male
10         1     38.9    3718. female
# ... with 474 more rows
```

Trend Lines

```
ggplot(data = babies, mapping = aes(x = AgeMonths, y = HeadCirc,
                                    alpha = WTMEC4YR, color = Gender)) +
  geom_jitter(width = 0.3, height = 0) + guides(alpha = FALSE) +
  geom_smooth(method = "lm", se = FALSE, mapping = aes(weight = WTMEC4YR))
```



Let's practice!

ANALYZING SURVEY DATA IN R

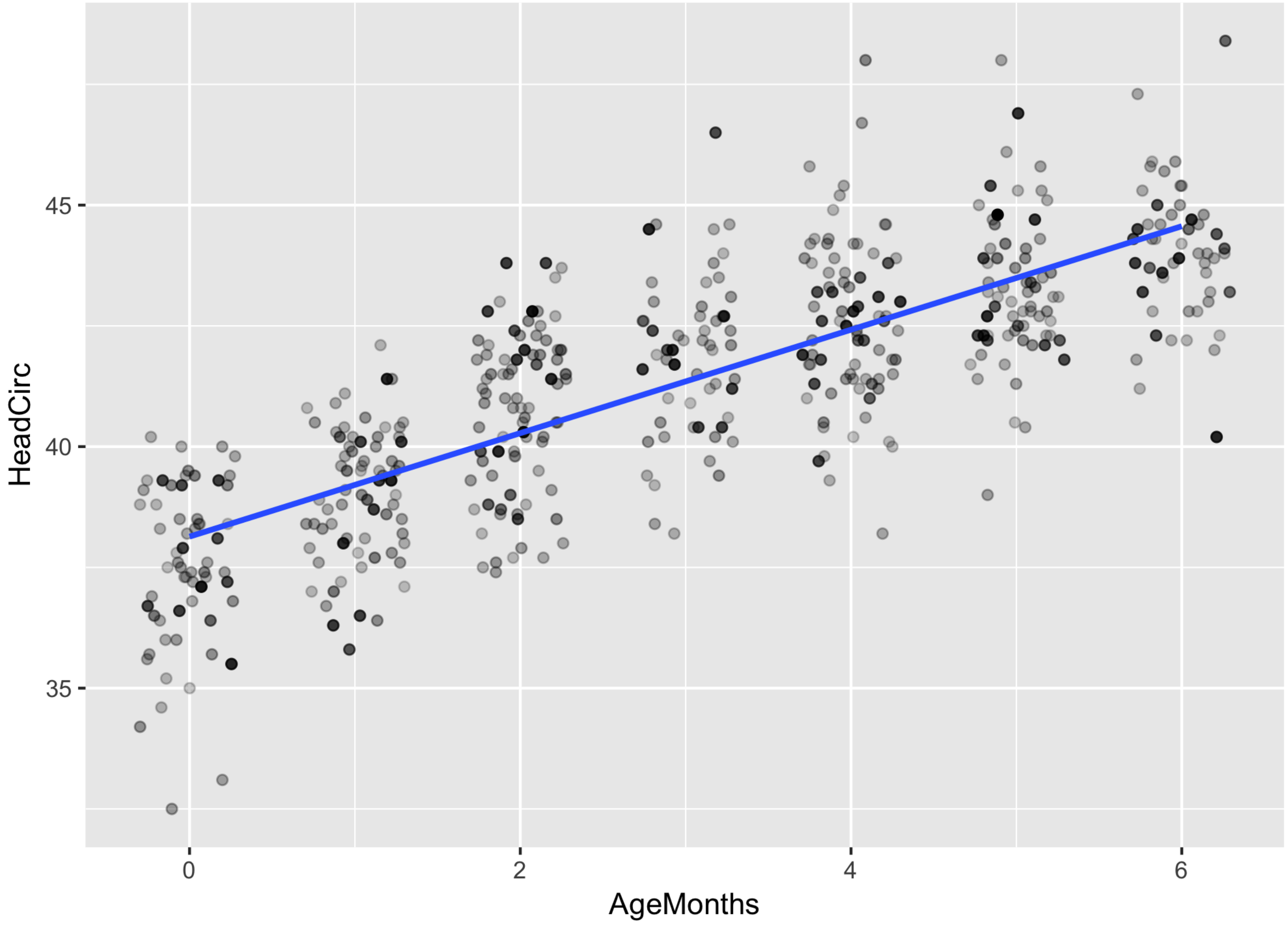
Modeling with linear regression

ANALYZING SURVEY DATA IN R

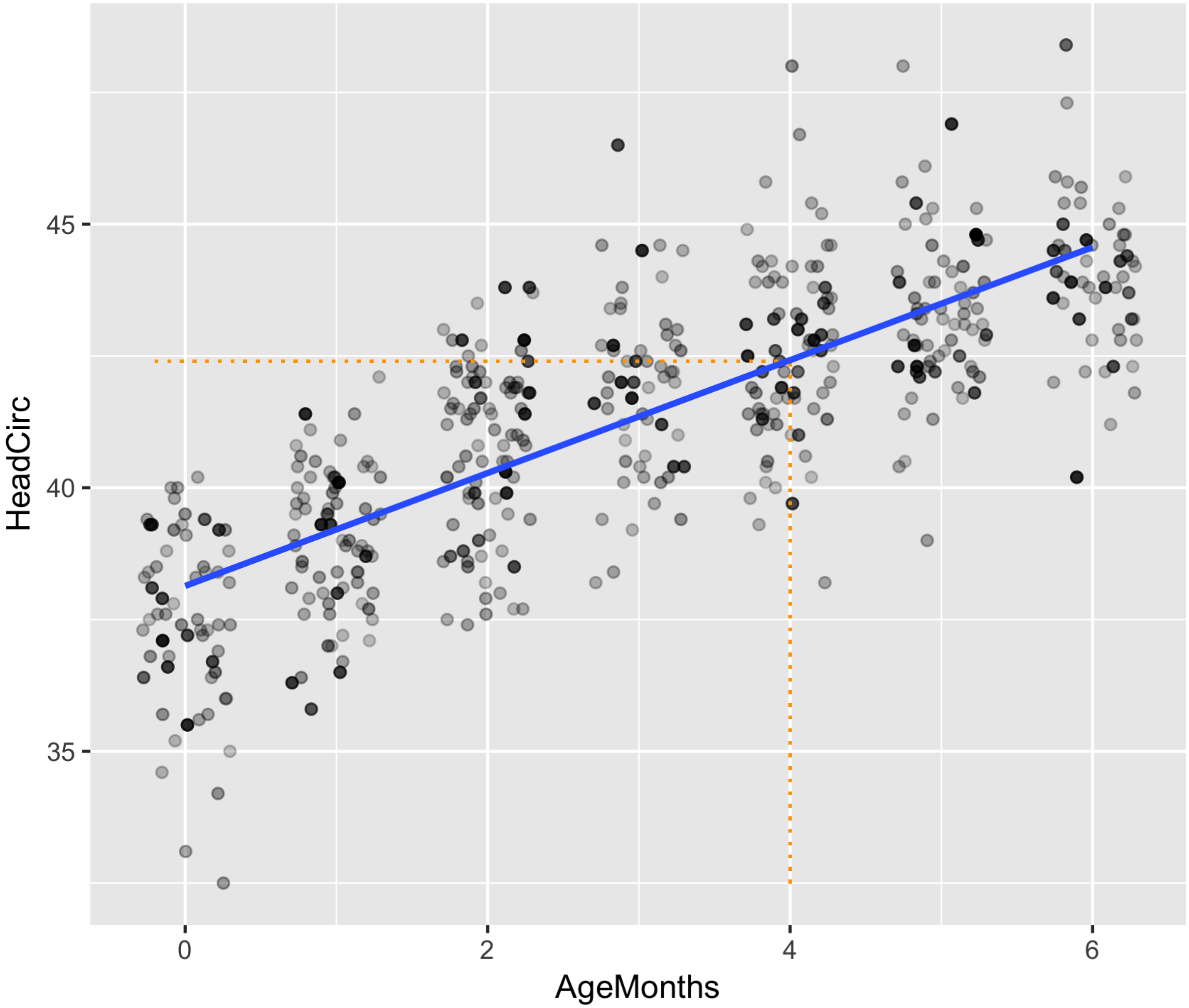
Kelly McConville

Assistant Professor of Statistics

Regression line



Regression line



Regression equation

- Regression equation is given by:

$$\hat{y} = a + bx$$

- Find a and b by minimizing

$$\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

Fitting regression model

```r
mod <- svyglm(HeadCirc ~ AgeMonths, design = NHANES_design)
summary(mod)
```

```
svyglm(formula = HeadCirc ~ AgeMonths, design = NHANES_design)

Survey design:
svydesign(data = NHANESraw, strata = ~SDMVSTRA, id = ~SDMVPSU,
    nest = TRUE, weights = ~WTMEC4YR)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.1376     0.2004   190.3   <2e-16 ***
AgeMonths     1.0708     0.0593    18.1   <2e-16 ***
(Some output omitted)
```

Linear regression inference

- **Estimated** regression equation is given by:

$$\hat{y} = a + bx$$

- **True** regression equation is given by:

$$E(y) = A + Bx$$

- $E(y)$ is the average value of y and the variance is sd $(y) = \sigma$.

Linear regression inference

Null Hypothesis: Head size and age are not linearly related (i.e., $B = 0$).

Alternative Hypothesis: Head size and age are linearly related (i.e. $B \neq 0$).

```
mod <- svyglm(HeadCirc ~ AgeMonths, design = NHANES_design)
summary(mod)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.1376     0.2004   190.3   <2e-16 ***
AgeMonths     1.0708     0.0593    18.1   <2e-16 ***
(Some Output Omitted)
```

Test statistic: $t = \frac{b}{SE}$

Let's practice!

ANALYZING SURVEY DATA IN R

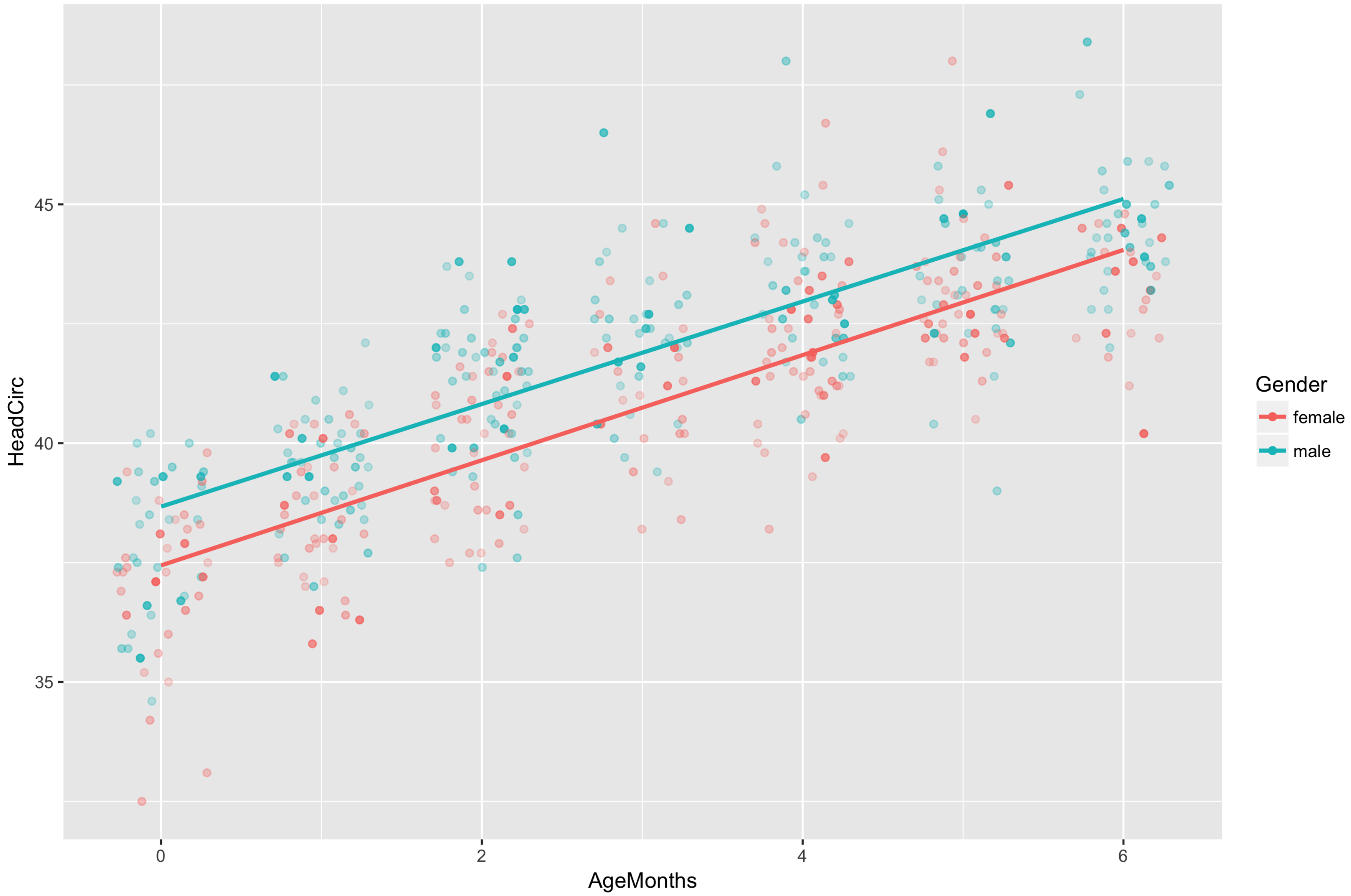
More complex modeling

ANALYZING SURVEY DATA IN R

Kelly McConville

Assistant Professor of Statistics

Multiple linear regression



Multiple linear regression

- Multiple linear regression equation is given by:

$$E(y) = B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_p x_p$$

```
babies
```

```
# A tibble: 484 x 4
   AgeMonths HeadCirc WTMEC4YR Gender
       <int>    <dbl>    <dbl> <fct>
 1         3     42.7   12915. male
 2         4     42.8   12791. female
 3         2     38.8    2359. female
 4         0     36.0    4306. female
 5         5     42.7    2922. female
 6         2     41.9    5561. male
 7         6     44.3   10416. female
# ... with 477 more rows
```

Multiple linear regression

- Multiple linear regression equation is given by:

$$E(y) = B_0 + B_1 x_1 + B_2 x_2$$

```
babies
```

```
# A tibble: 484 x 4
   AgeMonths HeadCirc WTMEC4YR Gender
       <int>    <dbl>    <dbl> <fct>
 1         3     42.7   12915. male
 2         4     42.8   12791. female
 3         2     38.8    2359. female
 4         0     36.0    4306. female
 5         5     42.7    2922. female
 6         2     41.9    5561. male
 7         6     44.3   10416. female
# ... with 477 more rows
```

Multiple linear regression

```
babies <- mutate(babies, Gender2 = case_when(
  Gender == "male" ~ 1,
  Gender == "female" ~ 0))
babies
```

```
# A tibble: 484 x 5
   AgeMonths HeadCirc WTMEC4YR Gender Gender2
       <int>    <dbl>    <dbl> <fct>    <dbl>
 1         3     42.7   12915. male        1.
 2         4     42.8   12791. female      0.
 3         2     38.8    2359. female      0.
 4         0     36.0    4306. female      0.
 5         5     42.7    2922. female      0.
 6         2     41.9    5561. male        1.
 7         6     44.3   10416. female      0.
# ... with 477 more rows
```

Multiple linear regression

- Multiple linear regression equation is given by:

$$E(y) = B_0 + B_1 x_1 + B_2 x_2$$

- Line for males:

$$E(y) = (B_0 + B_2) + B_1 x_1$$

- Line for females:

$$E(y) = B_0 + B_1 x_1$$

Multiple linear regression

```
mod <- svyglm(HeadCirc ~ AgeMonths + Gender, design = NHANES_design)
summary(mod)
```

```
svyglm(formula = HeadCirc ~ AgeMonths + Gender, design = NHANES_design)

Survey design:
svydesign(data = NHANESraw, strata = ~SDMVSTRA, id = ~SDMVPSU,
    nest = TRUE, weights = ~WTMEC4YR)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.48508    0.18320 204.613  < 2e-16 ***
AgeMonths    1.08658    0.05379  20.200  < 2e-16 ***
Gendermale   1.15034    0.16298   7.058  6.3e-08 ***
(Some output omitted)
```

Multiple linear regression

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.48508    0.18320 204.613  < 2e-16 ***
AgeMonths    1.08658    0.05379  20.200  < 2e-16 ***
Gendermale   1.15034    0.16298   7.058  6.3e-08 ***
(Some output omitted)
```

Null hypothesis: Given age is in the model, gender should not be included $(B_2 = 0)$.

Alternative hypothesis: Given age is in the model, gender should be included $(B_2 \neq 0)$.

Test statistic: $t = \frac{b_2}{SE}$

Multiple linear regression

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.48508    0.18320 204.613  < 2e-16 ***
AgeMonths    1.08658    0.05379  20.200  < 2e-16 ***
Gendermale   1.15034    0.16298   7.058  6.3e-08 ***
(Some output omitted)
```

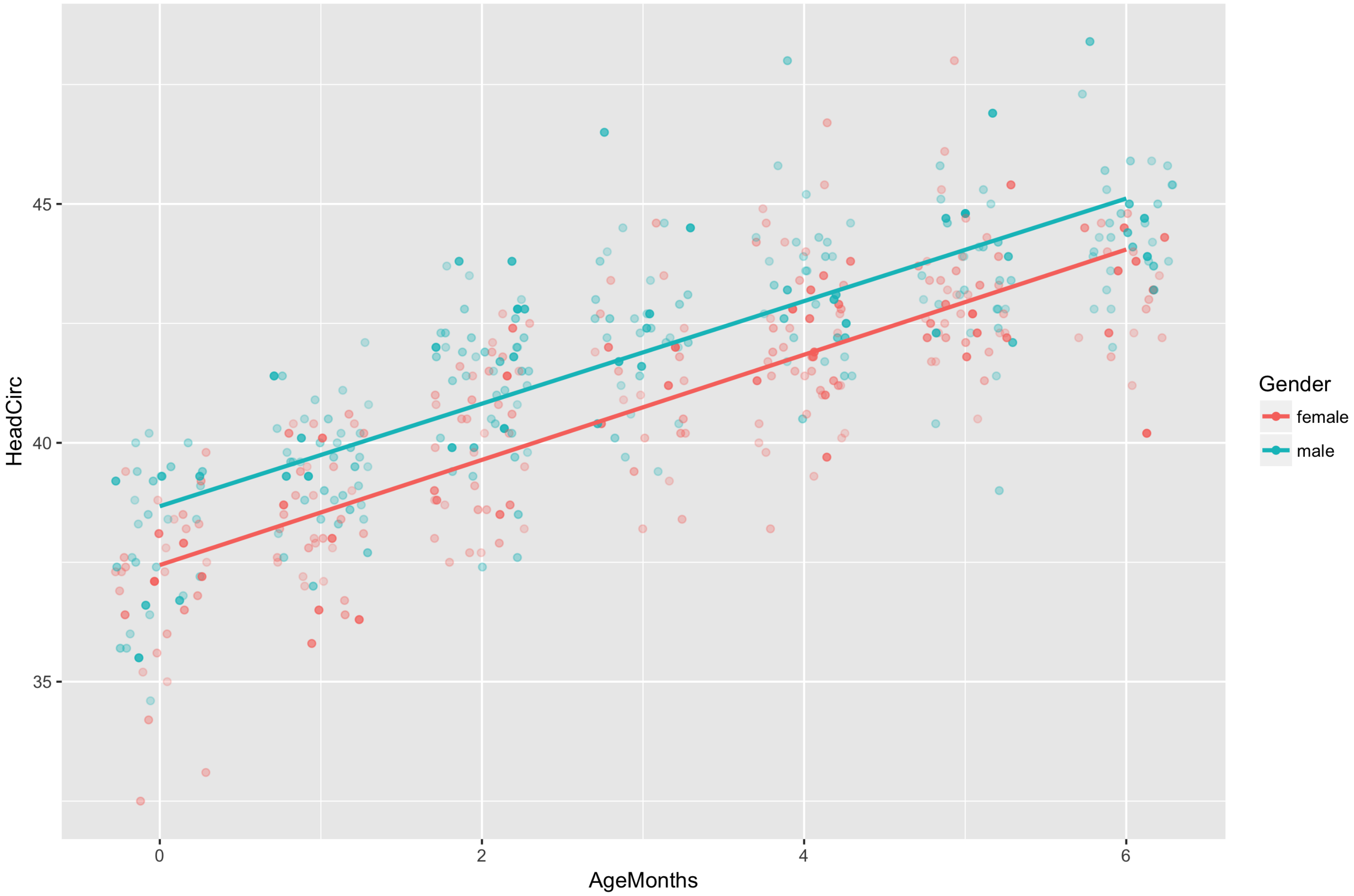
Null hypothesis: Given gender is in the model, age should not be included $(B_1 = 0)$.

Alternative hypothesis: Given gender is in the model, age should be included $(B_1 \neq 0)$.

Test statistic: $t = \frac{b_1}{SE}$

Multiple linear regression

$$E(y) = B_0 + B_1 x_1 + B_2 x_2$$



Let's practice!

ANALYZING SURVEY DATA IN R

Wrap-up

ANALYZING SURVEY DATA IN R

Kelly McConville

Assistant Professor of Statistics

R packages

- `survey` : To analyze survey data
- `dplyr` : To wrangle data
- `ggplot2` : To graph the data

Course summary

- Ch 1: Survey fundamentals
  - Common design features: clustering, stratification
  - Survey weights
  - Telling R about your `svydesign()`
- Ch 2: Categorical data
  - Frequency and contingency tables with `svytable()`
  - Bar graphs with `geom_col()`
  - Inference with `svychisq()`

Course summary

- Ch 3: Quantitative and categorical data
  - Summary stats with `svymean()` , `svytotal()` , `svyquantile()`
  - Domain estimates with `svyby()`
  - Describing shape with `geom_histogram()` , `geom_density()`
  - Inference with `svyttest()`
- Ch 4: Modeling trends
  - Mapping survey weights in `geom_point()`
  - Linear trends with `geom_smooth(method = "lm")`
  - Linear regression with `svyglm()`

Extensions

- Estimating more complex population quantities.
  - EX: `svyratio()`
- Building more complex models
  - EX :

    ```
    svyglm(Diabetes ~ Age, design = NHANES_design,
    family = quasibinomial)
    ```

Congratulations!

ANALYZING SURVEY DATA IN R