JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section D

Simple Linear Regression Model: Estimating the Regression Equation—Accounting for Uncertainty in the Estimates

Example: Hemoglobin and Packed Cell Volume

- So in the last section, we showed the results from several simple linear regression models

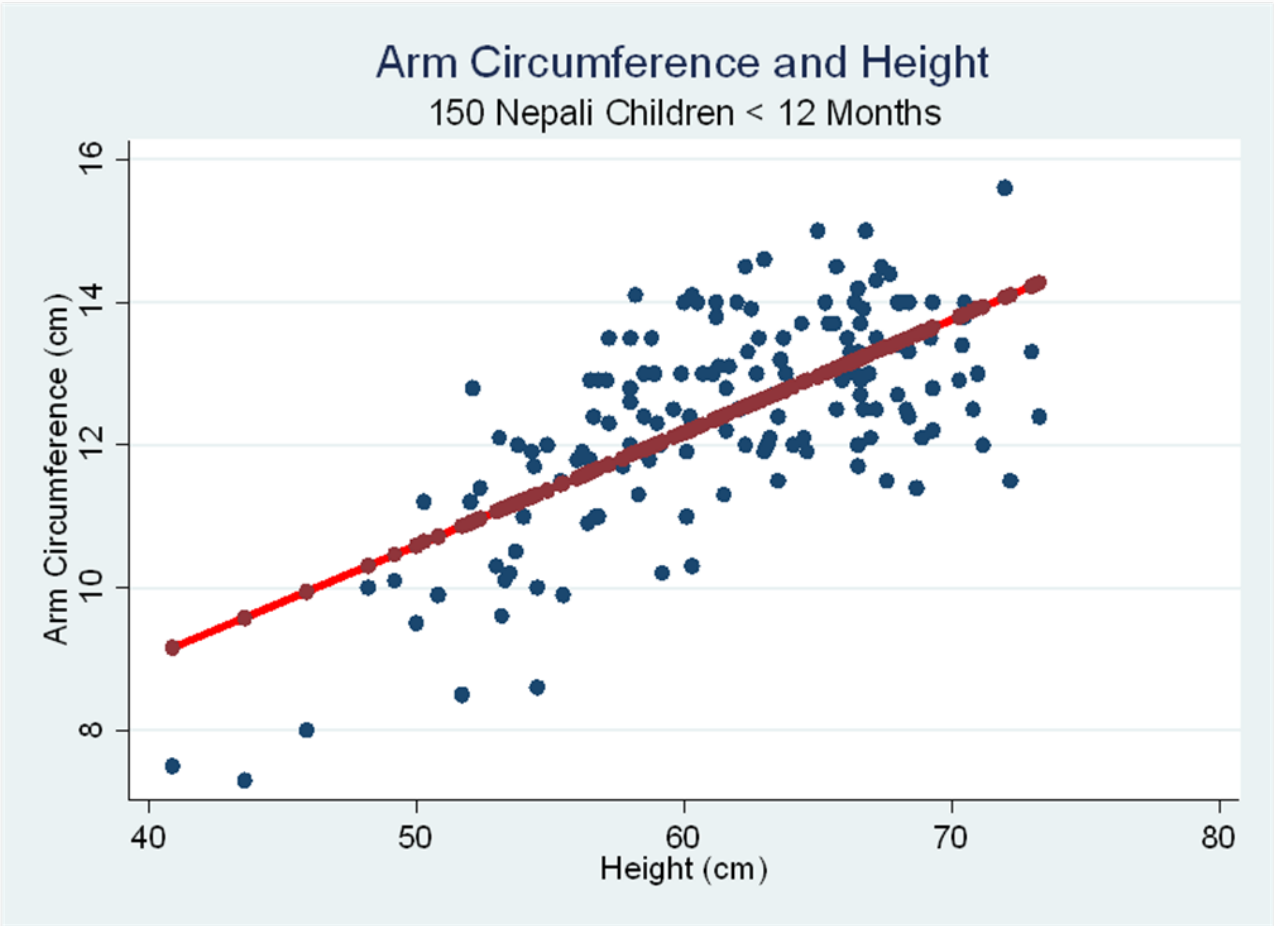
- For example, when relating arm circumference to height using a random sample of 150 Nepali children < 12 months old, I told you that the resulting regression equation was . . .

$$\hat{y} = 2.7 + 0.16x$$

- I told you this came from Stata—and will show you how to do regression with Stata shortly—but how does Stata estimate this equation?

Example: Arm Circumference and Height

- There must be some algorithm that will always yield the same results for the same data set



Example: Arm Circumference and Height

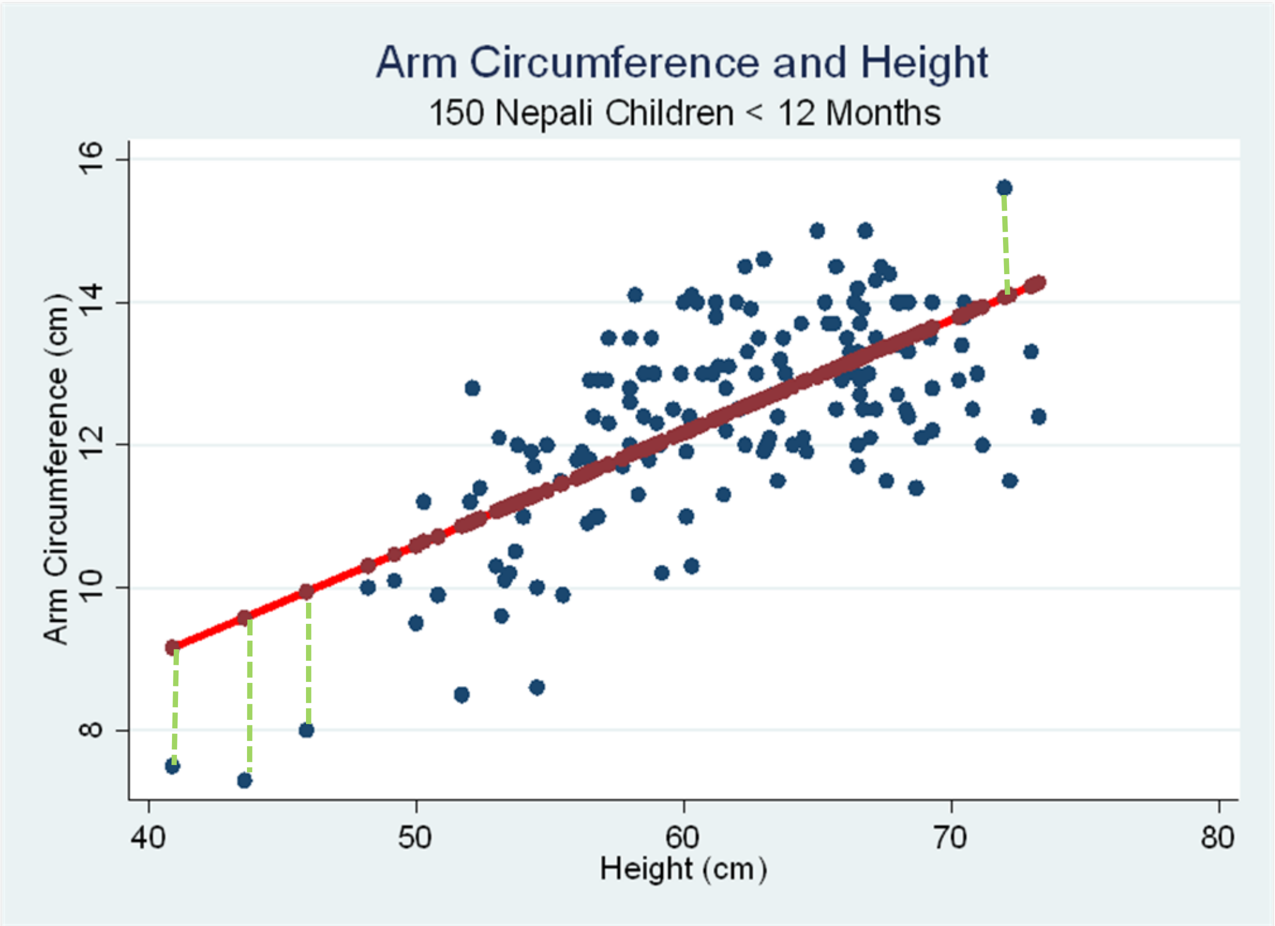
- The algorithm to estimate the equation of the line is called the "least squares" estimation

- The idea is to find the line that gets "closest" to all of the points in the sample

- How to define closeness to multiple points?
  - In regression, closeness is defined as the cumulative squared distance between each point's y-value and the corresponding value of $\hat{y}$ for that point's x
  - In other words the squared distance between an observed y-value and the estimated y-value for all points with the same value of x

Example: Arm Circumference and Height

- Each distance is $y - \hat{y} = y - (\hat{\beta}_o + \hat{B}_1 x)$: this is computed for each data point in the sample



Example: Arm Circumference and Height

- The algorithm to estimate the equation of the line is called the "least squares" estimation

- The values chosen for $\hat{\beta}_o$ *and* $\hat{\beta}_1$ are the values that minimize the cumulative distances squared:

$$\min\left[\sum_{i=1}^{n}\left(y_i - (\hat{\beta}_o + \hat{\beta}_1 x_i)\right)^2\right]$$

Example: Arm Circumference and Height

- The values chosen for $\hat{\beta}_o$ *and* $\hat{\beta}_1$ are just estimates based on a single sample
  - If you were to have a different random sample of 150 Nepal children from the same population of <12 month olds, the resulting estimate would likely be different (i.e., the values that minimized the cumulative squared distance from this second sample of points would likely be different)

- As such, all regression coefficients have an associated standard error that can be used to make statements about the true relationship between mean y and x (for example, the true slope β_1) based on a single sample

Example: Arm Circumference and Height

- The estimated regression equation relating arm circumference to height using a random samples of 150 Nepali children < 12 months old, I told you that the resulting regression equation was . . .

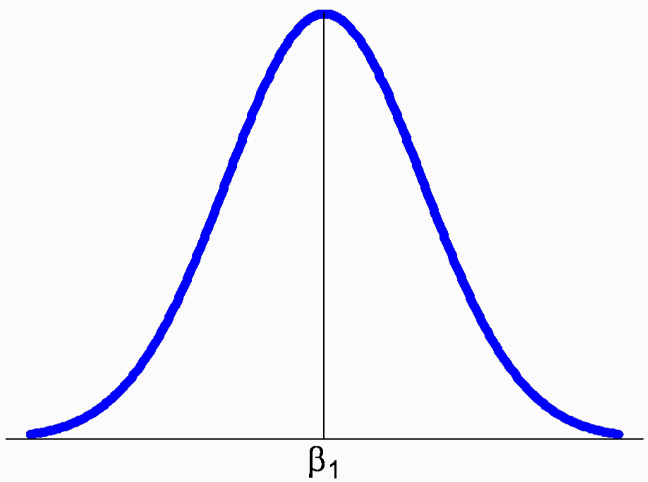
$$\hat{y} = 2.7 + 0.16x$$

$$\hat{\beta}_1 = 0.16 \; and \; S\hat{E}(\hat{\beta}_1) = 0.014$$

$$\hat{\beta}_o = 2.70 \; and \; S\hat{E}(\hat{\beta}_o) = 0.88$$

Example: Arm Circumference and Height

- Random sampling behavior of estimated regression coefficients is normal for large samples (n > 60), and centered at true values



- As such, we can use the same ideas to create 95% CIs and get p-values

Example: Arm Circumference and Height

- The estimated regression equation relating arm circumference to height using a random samples of 150 Nepali children < 12 months old, I told you that the resulting regression equation was:

$$\hat{y} = 2.7 + 0.16x$$

$$\hat{\beta}_1 = 0.16 \text{ and } S\hat{E}(\hat{\beta}_1) = 0.014$$

- 95% CI for β_1

$$\hat{\beta}_1 \pm 2 \times S\hat{E}(\hat{\beta}_1) \rightarrow 0.16 \pm 2 \times 0.014 \approx (0.13, 0.19)$$

Example: Arm Circumference and Height

- p-value for testing:
  - $H_o: \beta_1 = 0$
  - $H_o: \beta_1 = 0$

- Assume the null is true and calculate standardized "distance" of $\hat{\beta}_1$ from 0

$$t = \frac{\hat{\beta}_1 - 0}{S\hat{E}(\beta_1)} = \frac{\hat{\beta}_1}{S\hat{E}(\beta_1)} = \frac{0.16}{.014} \approx 11.4$$

- P-value is the probability of being 11.4 or more standard errors away from mean of 0 on a normal curve: very low in this example, $p < .001$

Summarizing Findings: Arm Circumference and Height

- This research used simple linear regression to estimate the magnitude of the association between arm circumference and height in Nepali children less than 12 months old, using data on a random sample of 150

- A statistically significant positive association was found ($p<.001$)

- The results estimate that two groups of such children who differ by 1 cm in height will differ on average by 0.16 cm in arm circumference (95% CI 0.13 cm to 0.19 cm)

Summarizing Findings: Arm Circumference and Height

- Finally: Stata!

- If you have your “y” and “x” values entered in Stata, then to do linear regression use the regress command:
  - regress y x

- Data snippet from Stata

```
+-------------------+
     | armcirc   height |
     |-------------------|
  1. |      12     71.2 |
  2. |     9.9     55.5 |
  3. |    12.5     70.8 |
  4. |    11.2       52 |
  5. |    14.1     58.2 |
     +-------------------+
```

Using Stat: Arm Circumference and Height

- regress armcirc height

```
. regress armcirc height

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  1,   148) =  124.30
       Model |  148.874597     1  148.874597           Prob > F      =  0.0000
    Residual |  177.263335   148  1.19772523           R-squared     =  0.4565
-------------+------------------------------           Adj R-squared =  0.4528
       Total |  326.137932   149  2.18884518           Root MSE      =  1.0944

------------------------------------------------------------------------------
     armcirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .1579469   .0141671    11.15   0.000     .1299511    .1859428
       _cons |   2.695906   .8774225     3.07   0.003     .9620116      4.4298
------------------------------------------------------------------------------
```

$$\hat{y} = 2.7 + 0.16x$$

Using Stat: Arm Circumference and Height

- regress armcirc height

```
. regress armcirc height

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  1,   148) =  124.30
       Model |  148.874597     1  148.874597           Prob > F      =  0.0000
    Residual |  177.263335   148  1.19772523           R-squared     =  0.4565
-------------+------------------------------           Adj R-squared =  0.4528
       Total |  326.137932   149  2.18884518           Root MSE      =  1.0944

------------------------------------------------------------------------------
     armcirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .1579469   .0141671    11.15   0.000      .1299511    .1859428
       _cons |   2.695906   .8774225     3.07   0.003      .9620116      4.4298
------------------------------------------------------------------------------
```

$$\hat{y} = 2.7 + 0.16x$$

Using Stat: Arm Circumference and Height

- regress armcirc height

```
. regress armcirc height

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  1,   148) =  124.30
       Model |  148.874597     1  148.874597           Prob > F      =  0.0000
    Residual |  177.263335   148  1.19772523           R-squared     =  0.4565
-------------+------------------------------           Adj R-squared =  0.4528
       Total |  326.137932   149  2.18884518           Root MSE      =  1.0944

------------------------------------------------------------------------------
     armcirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .1579469   .0141671    11.15   0.000     .1299511    .1859428
       _cons |   2.695906   .8774225     3.07   0.003     .9620116      4.4298
------------------------------------------------------------------------------
```

$\hat{\beta}_o$ (points to _cons 2.695906)

$$\hat{y} = 2.7 + 0.16x$$

Using Stat: Arm Circumference and Height

- regress armcirc height

```
. regress armcirc height

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  1,   148) =  124.30
       Model |  148.874597     1  148.874597           Prob > F      =  0.0000
    Residual |  177.263335   148  1.19772523           R-squared     =  0.4565
-------------+------------------------------           Adj R-squared =  0.4528
       Total |  326.137932   149  2.18884518           Root MSE      =  1.0944

------------------------------------------------------------------------------
     armcirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .1579469   .0141671    11.15   0.000     .1299511    .1859428
       _cons |   2.695906   .8774225     3.07   0.003     .9620116      4.4298
------------------------------------------------------------------------------
```

$\hat{\beta}_1$ (height coefficient)

$$\hat{y} = 2.7 + 0.16x$$

Using Stat: Arm Circumference and Height

- regress armcirc height

```
. regress armcirc height

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  1,   148) =  124.30
       Model |  148.874597     1  148.874597           Prob > F      =  0.0000
    Residual |  177.263335   148  1.19772523           R-squared     =  0.4565
-------------+------------------------------           Adj R-squared =  0.4528
       Total |  326.137932   149  2.18884518           Root MSE      =  1.0944

------------------------------------------------------------------------------
     armcirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .1579469   .0141671    11.15   0.000     .1299511    .1859428
       _cons |   2.695906   .8774225     3.07   0.003     .9620116      4.4298
------------------------------------------------------------------------------
```

$$\hat{y} = 2.7 + 0.16x$$

Example 2: Arm Circumference and Height

- Give an estimate and 95% CI for the mean difference in arm circumference for children 60 cm tall compared to children 50 cm tall
  - From previous set we know this estimated mean difference is

$$(60-50)\times\hat{\beta}_1 = 10\hat{\beta}_1 = 10\times 0.16 = 1.6\,cm$$

  - How to get standard error? Well as it turns out:

$$S\hat{E}(10\hat{\beta}_1) = 10\times S\hat{E}(\hat{\beta}_1)$$

$$S\hat{E}(10\hat{\beta}_1) = 10\times 0.014 = 0.14$$

  - 95% CI for the mean difference

$$10\hat{\beta}_1 \pm 2S\hat{E}(10\hat{\beta}_1)$$

$$1.6 \pm 2\times 0.14$$

Example 2: Hemoglobin and "Packed Cell Volume"

- Equation of regression line relating estimated mean Hemoglobin (g/dL) to packed cell volume: from Stata

$$\hat{y} = 5.77 + 0.20x$$

- Snippet of data in Stata

```
     +------------+
     |   Hb   PCV |
     |------------|
  1. | 13.5    35 |
  2. | 10.5    30 |
  3. |  9.6    25 |
  4. | 13.5    35 |
  5. |   12    35 |
     +------------+
```

Example 2: Hemoglobin and “Packed Cell Volume”

- regress Hb PCV

```
. regress Hb PCV

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =   19.81
       Model |  53.7803079     1  53.7803079           Prob > F      =  0.0003
    Residual |  51.5711174    19  2.71426934           R-squared     =  0.5105
-------------+------------------------------           Adj R-squared =  0.4847
       Total |  105.351425    20  5.26757126           Root MSE      =  1.6475

------------------------------------------------------------------------------
          Hb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PCV |   .2033502   .0456835     4.45   0.000     .1077335    .2989668
       _cons |    5.77645   1.913624     3.02   0.007     1.771188    9.781712
------------------------------------------------------------------------------
```

Example 2: Hemoglobin and "Packed Cell Volume"

- Same idea with computation of 95% CI and p-value as we saw before

- However, with small ($n < 60$) samples, a slight change analogous to what we did with means and differences in means before

- Sampling distribution of regression coefficients not quite normal, but follow a t-distribution with n-2 degrees of freedom

- 95% for β_1

$$\hat{\beta}_1 \pm t_{.95,n-2} \times S\hat{E}(\hat{\beta}_1)$$

  - In this example

$$\hat{\beta}_1 \pm t_{.95,19} \times S\hat{E}(\hat{\beta}_1) \rightarrow 0.20 \pm 2.09 \times .046 \approx (0.10, 0.30)$$

Example: Hemoglobin and "Packed Cell Volume"

- p-value for testing:
  - H_o: $\beta_1 = 0$
  - H_o: $\beta_1 = 0$

- Assume the null is true and calculate standardized "distance" of $\hat{\beta}_1$ from 0

$$t = \frac{\hat{\beta}_1 - 0}{S\hat{E}(\beta_1)} = \frac{\hat{\beta}_1}{S\hat{E}(\beta_1)} = \frac{0.20}{.046} \approx 4.35$$

- P-value is the probability of being 4.35 or more standard errors away from mean of 0 on a t curve with 19 degrees of freedom: very low in this example, $p < .001$

Interpreting Result of 95% CI

- So, the estimated slope is 0.2 with 95% CI 0.10 to 0.30

- How to interpret results?
  - Based on a sample of 21 subjects, we estimated that PCV(%) is positively associated with hemoglobin levels
  - We estimated that a one-percent increase in PCV is associated with a 0.2 g/dL increase in hemoglobin on average
  - Accounting for sampling variability, this mean increase could be as small as 0.10 g/dL, or as large as 0.3 g/dL in the population of all such subjects

Interpreting Result of 95% CI

- In other words:
  - We estimated that the average difference in hemoglobin levels for two groups of subjects who differ by one-percent in PCV to be 0.2 g/dL on average (higher PCV group relative to lower)
  - Accounting for sampling variability, the mean difference could be as small as 0.10 g/dL, or as large as 0.3 g/dL, in the population of all subjects

What about Intercepts?

- In this section, all examples have confidence intervals for the slope, and multiples of the slope
- We can also create confidence intervals/p-values for the intercept in the same manner (and Stata presents it in the output)
- However as we noted in the previous section, many times the intercept is just a placeholder and does not describe a useful quantity: as such, 95% CIs and p-values are not always relevant