JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section C

Simple Linear Regression: More Examples

Example: Hb and PCV

- Linear regressions performed with a single predictor (one x) are called simple linear regressions

- Linear regressions performed with more than one predictor (more than one x) are called multiple linear regressions

- In this set of lectures, we are dealing with simple linear regression
  - In this section we will give three more examples

Example: Hb and PCV

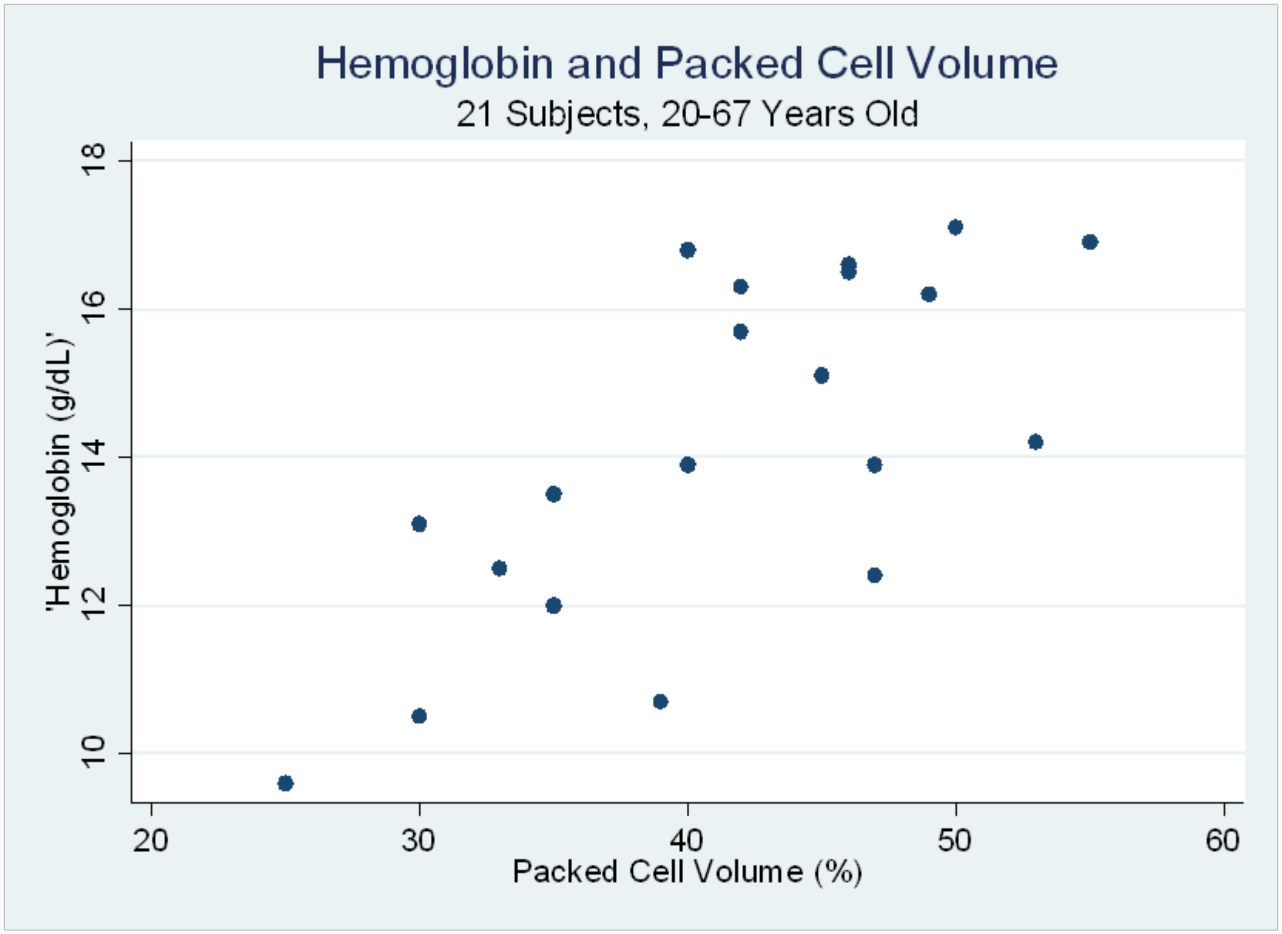
- Data on laboratory measurements on a random sample of 21 clinical patients, 20-67 years old

- Question—what is the relationship between hemoglobin levels (g/dL) and packed cell volume (percent of packed cells)

- Data
  - Hemoglobin (Hb): mean 14.1 g/dl, SD 2.3 g/dL, range 9.6 g/dL – 17.1 g/dL
  - Packed Cell Volume (PCV): mean 41.1%, SD 8.1%, range 25% to 55%

Visualizing Hb and PCV Relationship

- Scatterplot display



Example: Hb and PCV

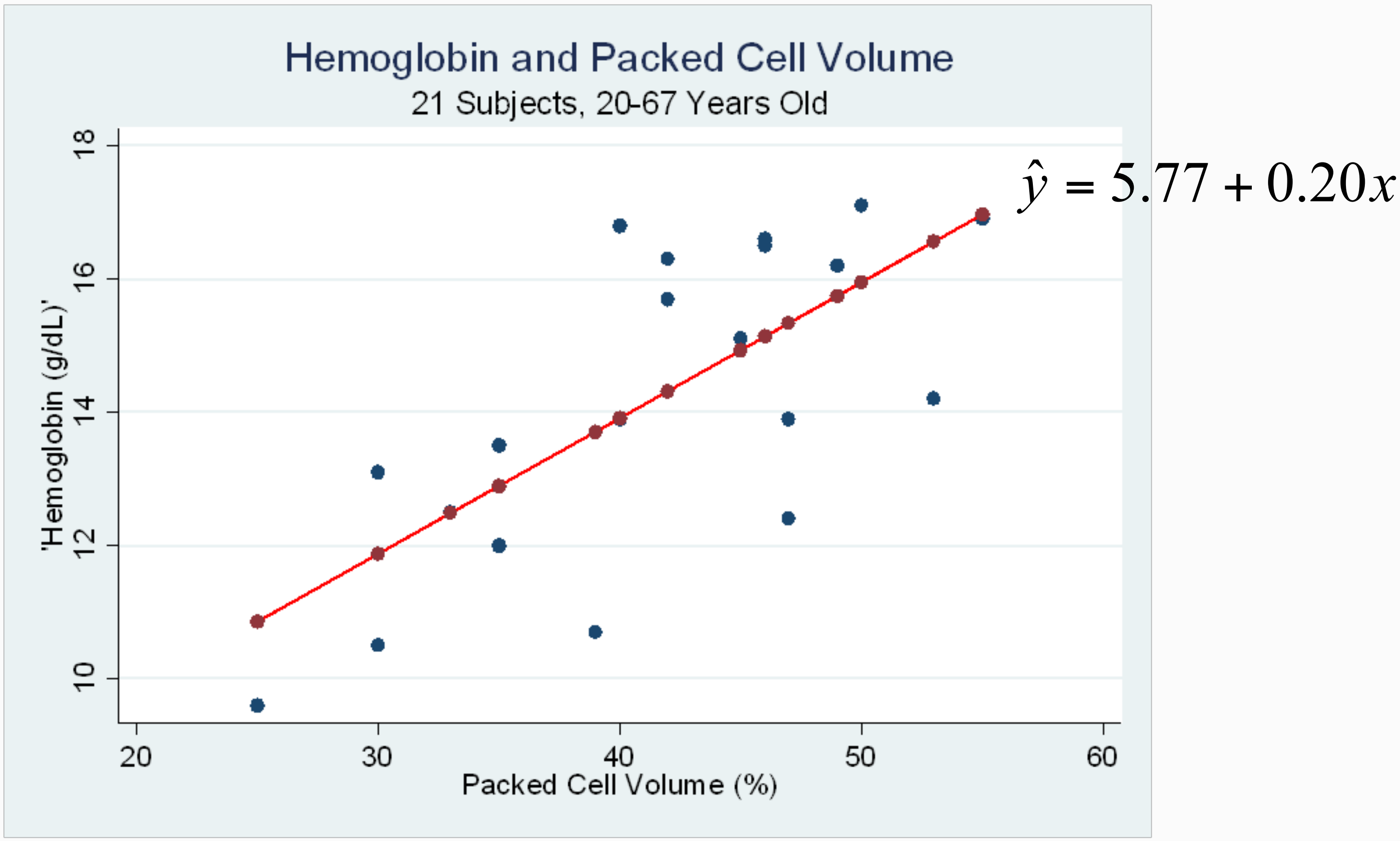
- Equation of regression line relating estimated mean hemoglobin (g/dL) to packed cell volume: from Stata
  - $\hat{y} = 5.77 + 0.20x$
  - Here, $\hat{y}$ = estimated average hemoglobin (like what we previously would call $\bar{y}$), x = height, $\hat{\beta}_o = 5.77$ and $\hat{\beta}_1 = 0.20$
  - This is the estimated line from the sample of 21 subjects

Example: Hb and PCV

- Equation of regression line relating estimated mean hemoglobin (g/dL) to packed cell volume: from Stata
  - $\hat{y} = 5.77 + 0.20x$
  - $\hat{\beta}_1 = 0.20$: what are the units?
  - Well, $\hat{y}$ is in g/dL, x in percent; so $\hat{\beta}_1$ is in units if g/dL per percent
    - This result estimates that the mean difference in hemoglobin levels for two groups of subjects who differ by 1% in PCV is 0.20 g/dL: subjects with greater PCV have greater Hb levels in average

Visualizing Hb and PCV Relationship

- Scatterplot display with regression line



Example: Hb and PCV

- What is the average difference in Hb levels for subjects with PCV of 40% compared to subjects with 32%?

- $\hat{\beta}_1 = 0.20$: compares groups of subjects who differ in PCV by 1% (it is positive, so those with the greater PCV have hemoglobin levels of . 20 g/dL greater on average)

- To compare subjects with PCV of 40% versus subjects with 32%, which is an eight unit difference in x, take

$$8 \times \hat{\beta}_1 = 8 \times 0.20 = 1.6\ g / dL$$

Example: Hb and PCV

- What is estimated Hb level for subjects with PCV of 41%?

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

- Plugging 41% into the equation:

$$\hat{y} = 5.77 + 0.20 \times 41 = 13.97\ g/dL$$

Example: Wages and Education Level

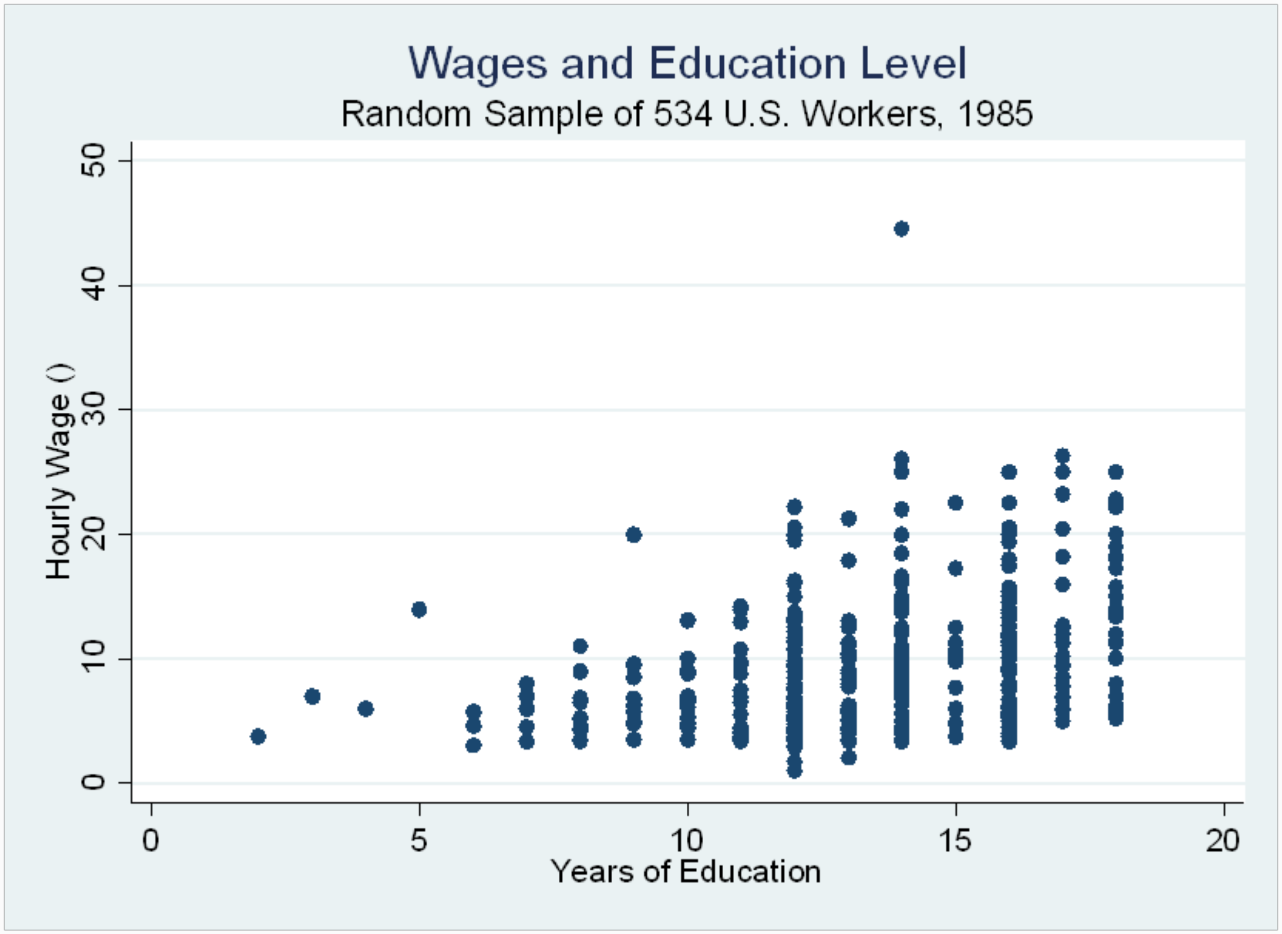
- Data on hourly wages from a random sample of 534 U.S. workers in 1985

- Question: what is the relationship between hourly wage (U.S. $) and years of formal education

- Data:
  - Hourly wages: mean $9.04/hour, SD $5.13/hour, range $1.00/hour–$44.50/hr
  - Year of formal education: mean 13.0 years, SD 2.6 years, range 2 years–18 years

Visualizing Wages and Education Level Relationship

- Scatterplot display

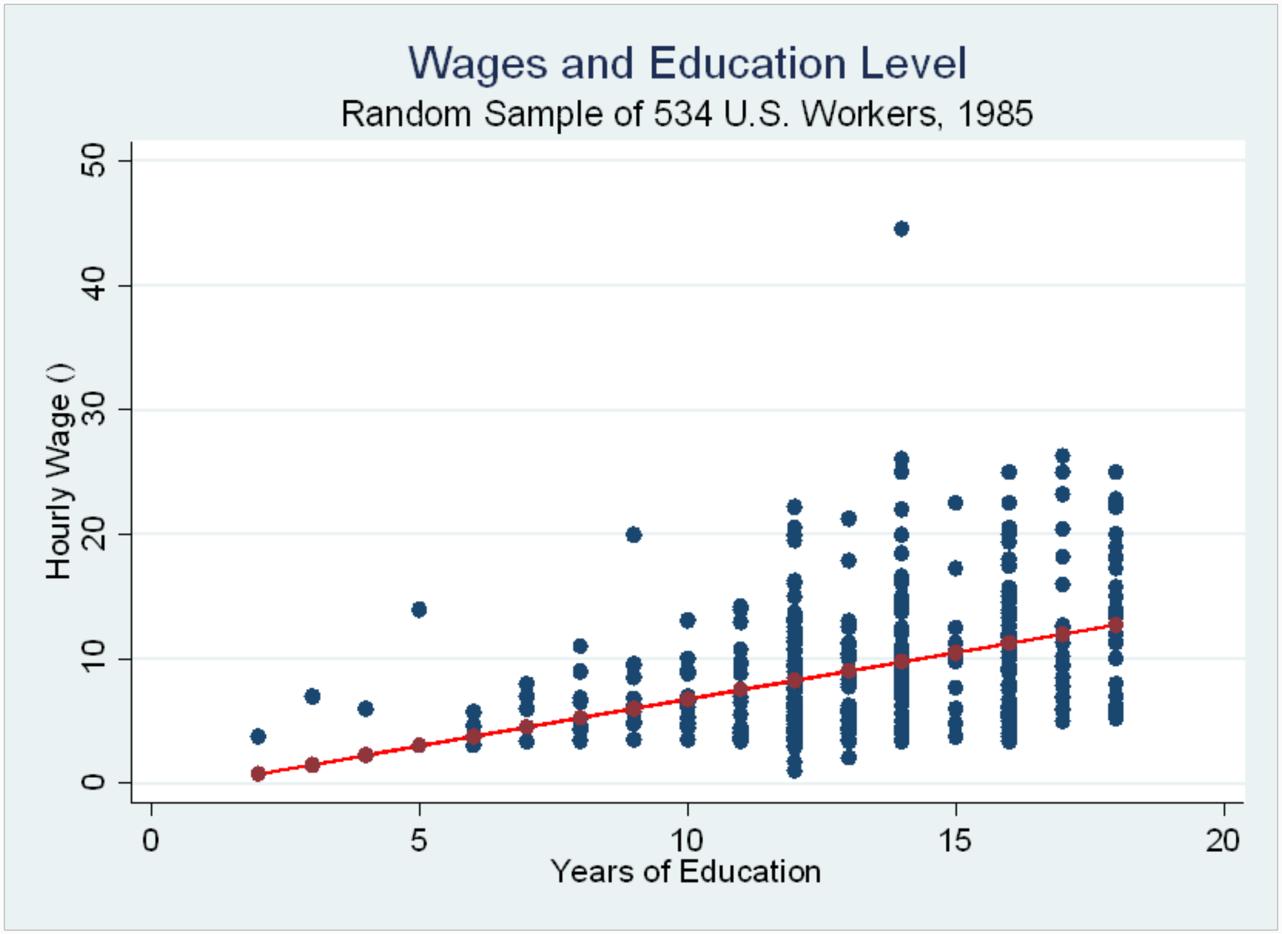


Example: Wages and Education Level

- Equation of regression line relating estimated mean hourly wages (U.S. $) to years of education: from Stata
  - $\hat{y} = -0.75 + 0.75x$
  - Here, $\hat{y}$ = estimated average hourly wage (like what we previously would call $\overline{y}$), x = years of formal education, $\hat{\beta}_o = -0.75$ and $\hat{\beta}_1 = 0.75$
  - This is the estimated line from the sample of 534 subjects

Visualizing Wages and Education Level Relationship

- Scatterplot display with regression line



Example: Arm Circumference and Sex

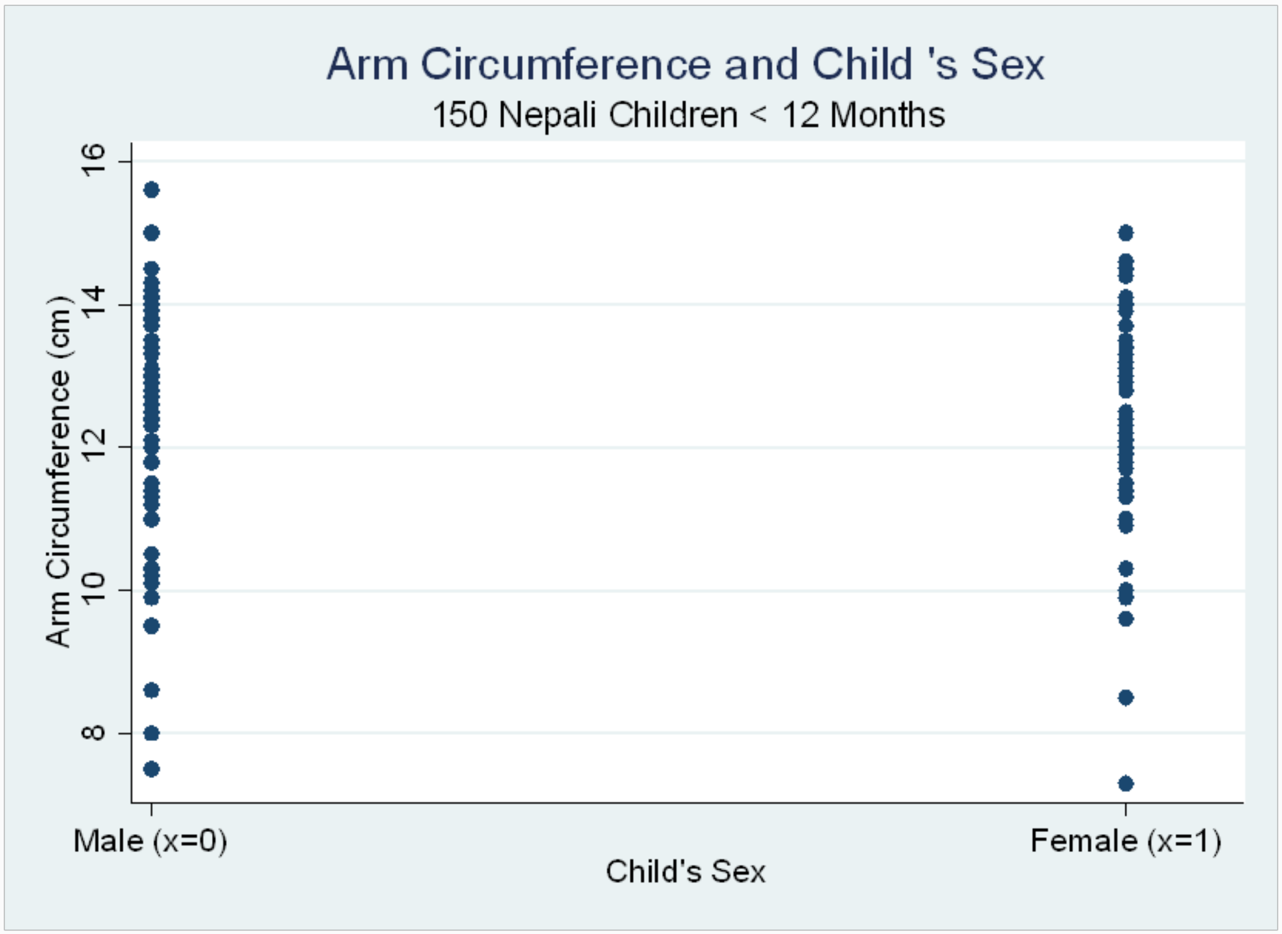
- Data on anthropomorphic measures from a random sample of 150 Nepali children (0, 12) months old

- Question: what is the relationship between average arm circumference and sex of a child

- Data:
  - Arm circumference: mean 12.4 cm, SD 1.5 cm, range 7.3 cm – 15.6 cm
  - Sex: 51% female

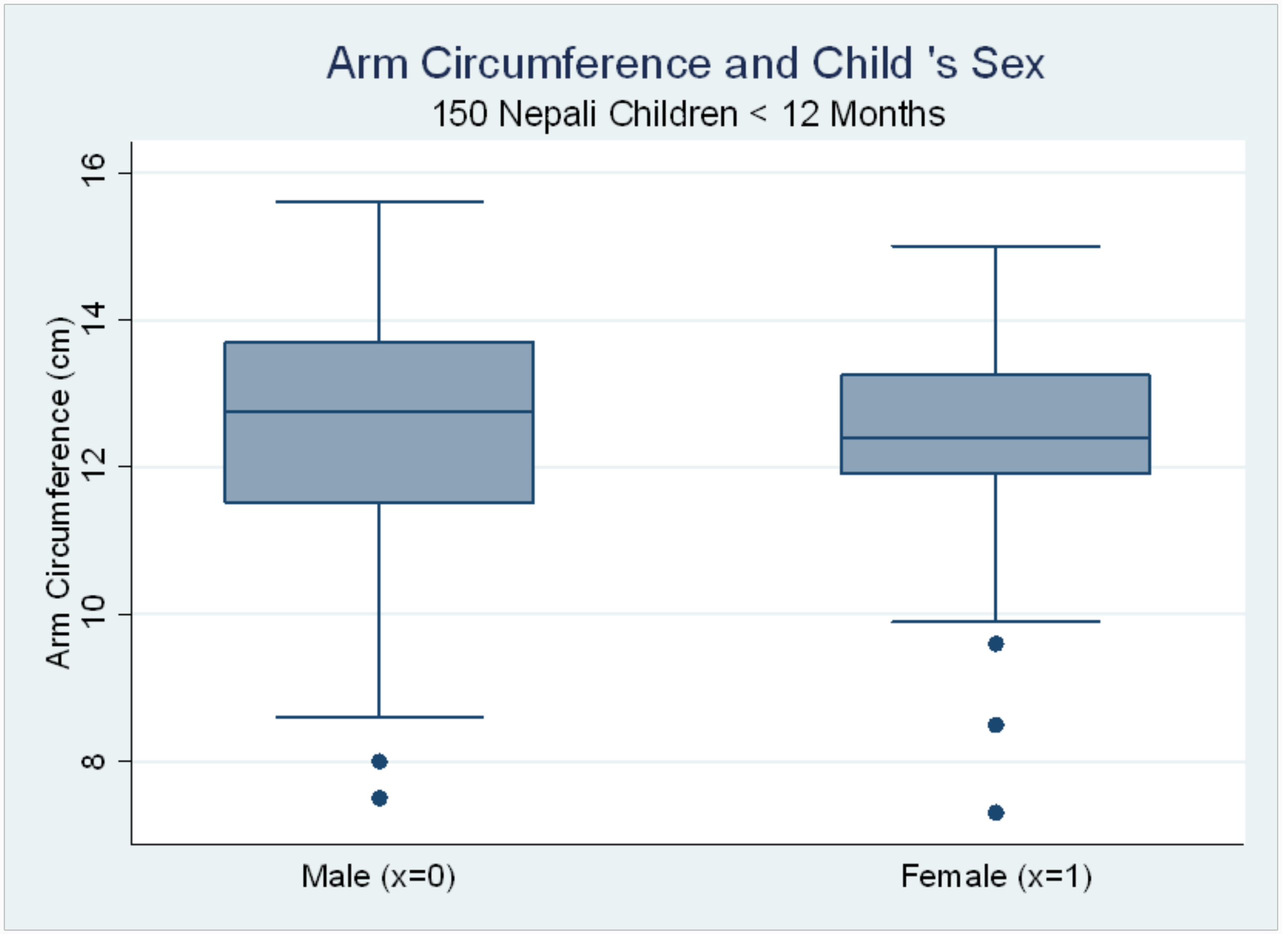
Visualizing Arm Circumference and Sex Relationship

- Scatterplot display



Visualizing Arm Circumference and Sex Relationship

- Boxplot display



Example: Arm Circumference and Sex

- Here, *y* is arm circumference, a continuous measure; x is not continuous, but binary (male or female)

- How to handle sex as an "x" in regression?
  - One possibility is x = 0 for male children and x = 1 for female children

- The equation we will estimate

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- How to interpret regression coefficients?

Example: Arm Circumference and Sex

- Notice, this equation is only estimating two values: mean arm circumference for male children, and the mean for female children

- For female children: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times 1 = \hat{\beta}_0 + \hat{\beta}_1$

- For male children: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times 0 = \hat{\beta}_0$

- So $\hat{\beta}_1$ is still a slope estimating mean difference in y for one-unit difference in x
  - But only possible one-unit difference is 1 (females) to 0 (males)

- $\hat{\beta}_o$ actually has substantive meaning in this example; it is the average arm circumference for male children

Example: Arm Circumference and Sex

- The resulting equation $\hat{y} = 12.5 + -0.13x$

- $\hat{\beta}_1 = -0.13$: the estimated mean difference in arm circumference for female children compared to male children is -0.13 cm; female children have lower arm circumference by 0.13 cm on average

- $\hat{\beta}_o = 12.5$: the mean arm circumference for male children is 12.5 cm

Visualizing Arm Circumference and Sex Relationship

- Scatterplot display with regression line