CIVL 7012/8012

Simple Linear Regression

Lecture 3

OLS assumptions - 1

- Model of population $y = \beta_0 + \beta_1 x + \varepsilon$
- Sample estimation (best-fit line) $\hat{y} = b_0 + b_1 x$
- We want $\boldsymbol{E}(\boldsymbol{b_1}) = \boldsymbol{\beta_1}$---> (1)
- Meaning we want $b_1$to be “unbiased”

- 5 assumptions of OLS to ensure “unbiasedness” of our slope parameter.

OLS assumptions - 2

1- Linear in parameters: in OLS,

- we can not have $y = \beta_0 + \beta_1^2 x + \varepsilon$ *(not linear)*
- we could have $y = \beta_0 + \beta_1 x^2 + \varepsilon$ *(linear-in-parameters)*

2- Random sampling

3- Zero conditional mean

$$E(\varepsilon/x) = 0$$ ---> (2)

- Error is random with an expected average value of 0 given the IV (x).

OLS assumptions - 3

3- Zero conditional mean (cont.)

- Before we state how ε and x are related, we can make one assumption about ε – As long as intercept β_0 is included in the equation, nothing is lost by assuming that the average value of ε in the population is zero.
- i.e. $E(\varepsilon) = 0$ ---> (3)
- (remember from SLR lec.1 errors above and below the best-fit line averaged 0).
- Eq. (3) suggests that the distribution of unobserved factors in the population is zero.
- Combining equations 2, and 3, we get that:

$E(\varepsilon/x) = E(\varepsilon) = 0$ ---> (4)

OLS assumptions - 4

3- Zero conditional mean (cont.)

- $E(\varepsilon/x) = E(\varepsilon) = 0$ ---> (4)
- Equation (4) suggests that average value of ε does not depend on the value of x.
- If equation (4) holds true, then we can say that ε is “mean independent” of x
- When equations (3) and (4) are met, we can state the zero conditional mean assumption is met.

OLS assumptions - 5

- 3- Zero conditional mean (cont.)

$$E(\varepsilon/x) = E(\varepsilon) \;=\; 0 \text{ ---> (4)}$$

- The error is random with an expected average value of 0 given the IV
- **Example (1):** predict wage based on individual's height.
- So, for a certain height $(h_i = 70\ inches)$, we will have different predicted values of wages (individuals with different wages but same height)..
- Individuals 1,2,3 ---> wages = 40k, 42k, 38k.
- Why do we have these differences in wages? Because of other factors that are not known to the analyst, but combined in the ε term (factors such as: more/less talented, productive, etc.). .

OLS assumptions - 6

3- Zero conditional mean (cont.)

$$E(\varepsilon/x) = E(\varepsilon) \ = \ 0 \text{ ---> (4)}$$

- **Example (2): I**n an effort to determine income as a function of education, we can state that $Income \ = \ \beta_0 \ + \beta_1(education) \ + \varepsilon$
- Let us say ε is same as innate ability
- Let $E(ability/8)$ represents average ability for the group of the population with 8 years of education.
- Similarly, let $E(ability/16)$ represents average ability for the group of the population with 16 years of education.
- As per equation (4): $E(ability/8) \ = \ E(ability/16) \ = 0$

OLS assumptions - 7

3- Zero conditional mean (cont.)

$$E(\varepsilon/x) = E(\varepsilon) \;=\; 0 \text{ ---> (4)}$$

- As per equation (4): $E(ability/8) \;=\; E(ability/16) \;= 0$
- As we can not observe innate ability, we have no way of knowing whether or not average ability is same for all education levels. • So for all unobserved factors we consider that $E(\varepsilon/x) \;=\; 0$

OLS assumptions - 8

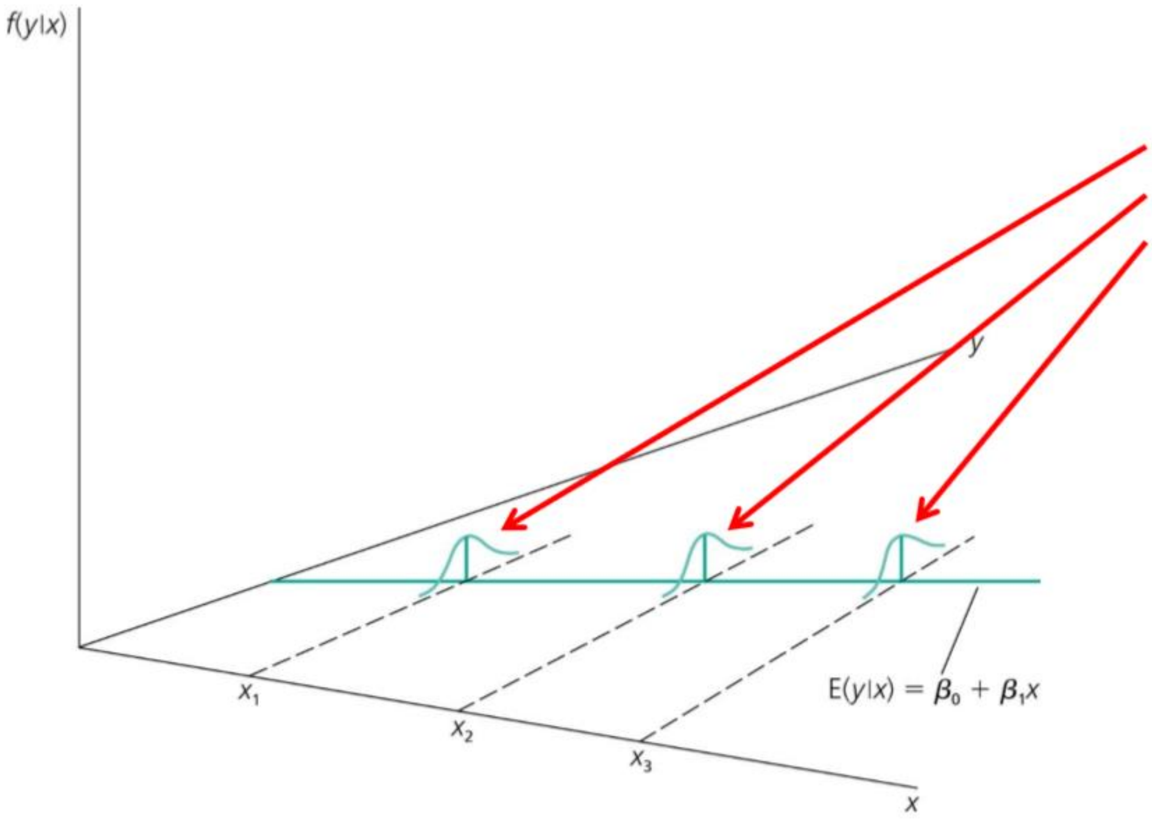
4- Sample variation

- We need to have different y's and different x's.
- We can not just have 1 observation and say multiply it by 100 and claim 100 observation – observations need to be unique and have variations in both y and x.

5- Homoscedasticity: homo=same, scedasticity=variance

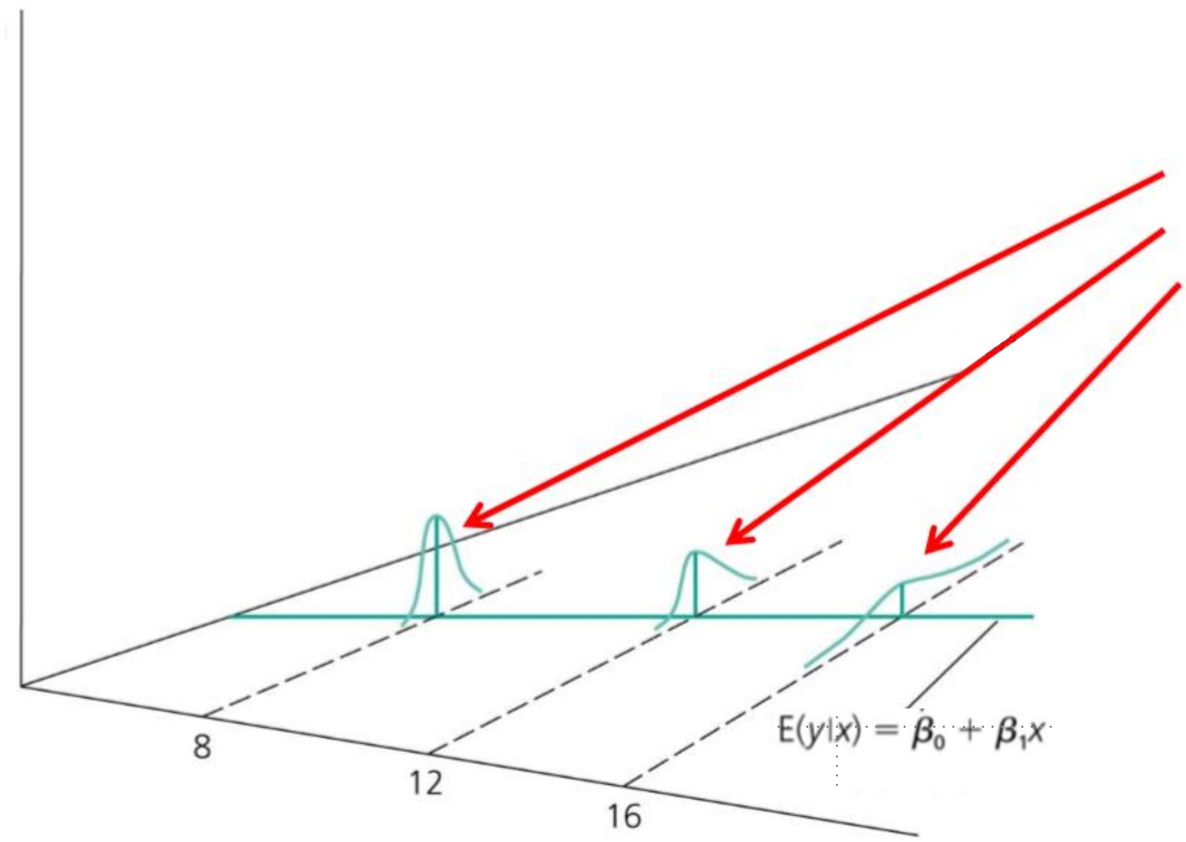
- $var\left(\frac{\varepsilon}{x}\right) = \sigma^2$ ---> constant variance.
- So, this means that the variance in the y variable is constant across the range of the x variable.

OLS assumptions – 9 Homoscedasticity



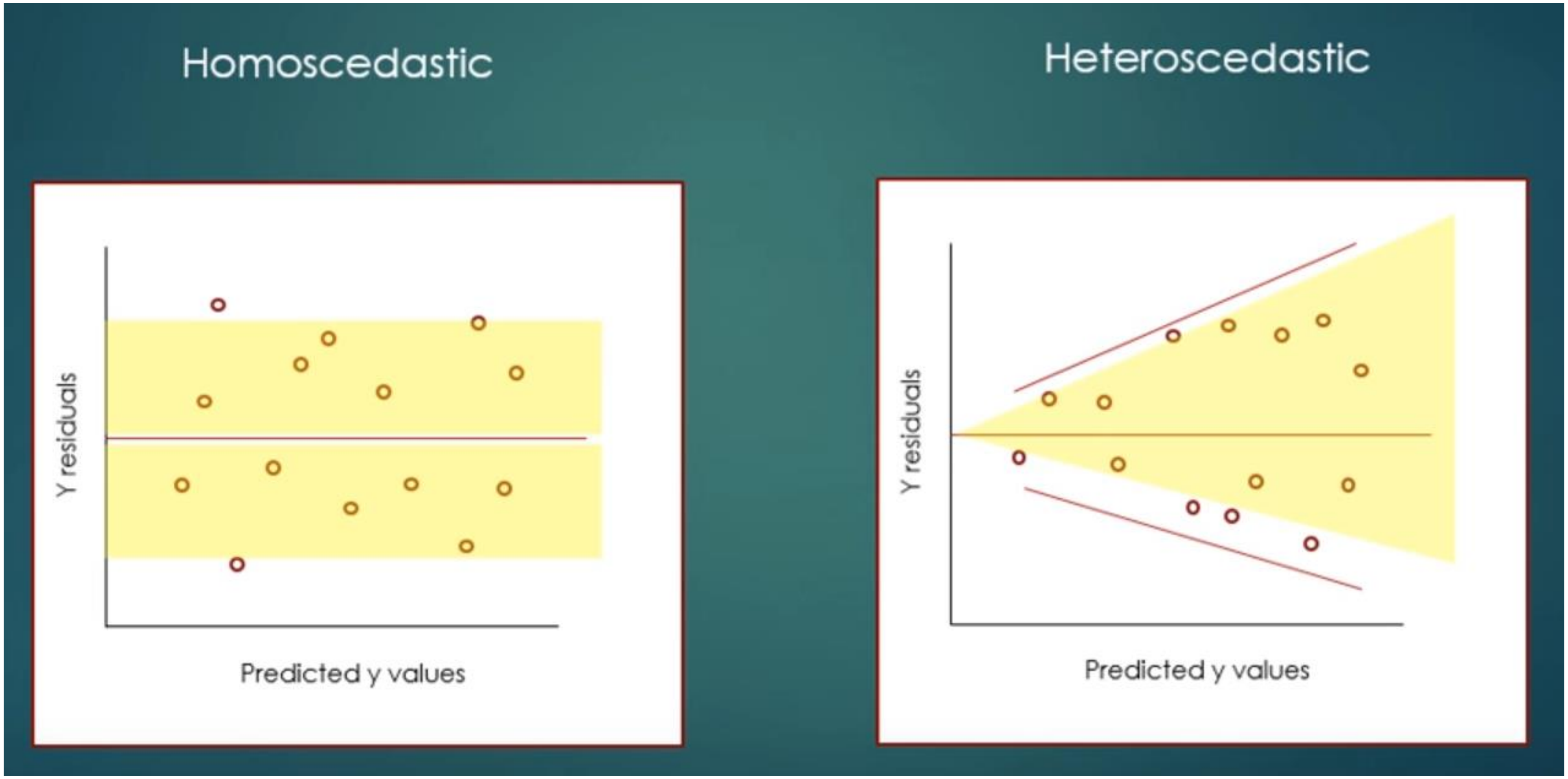
The variability of the unobserved influences (error) does not dependent on the value of the explanatory variable.

OLS assumptions – 10 Heteroscedasticity



The variance of the unobserved determinants of y increases with the change in x.

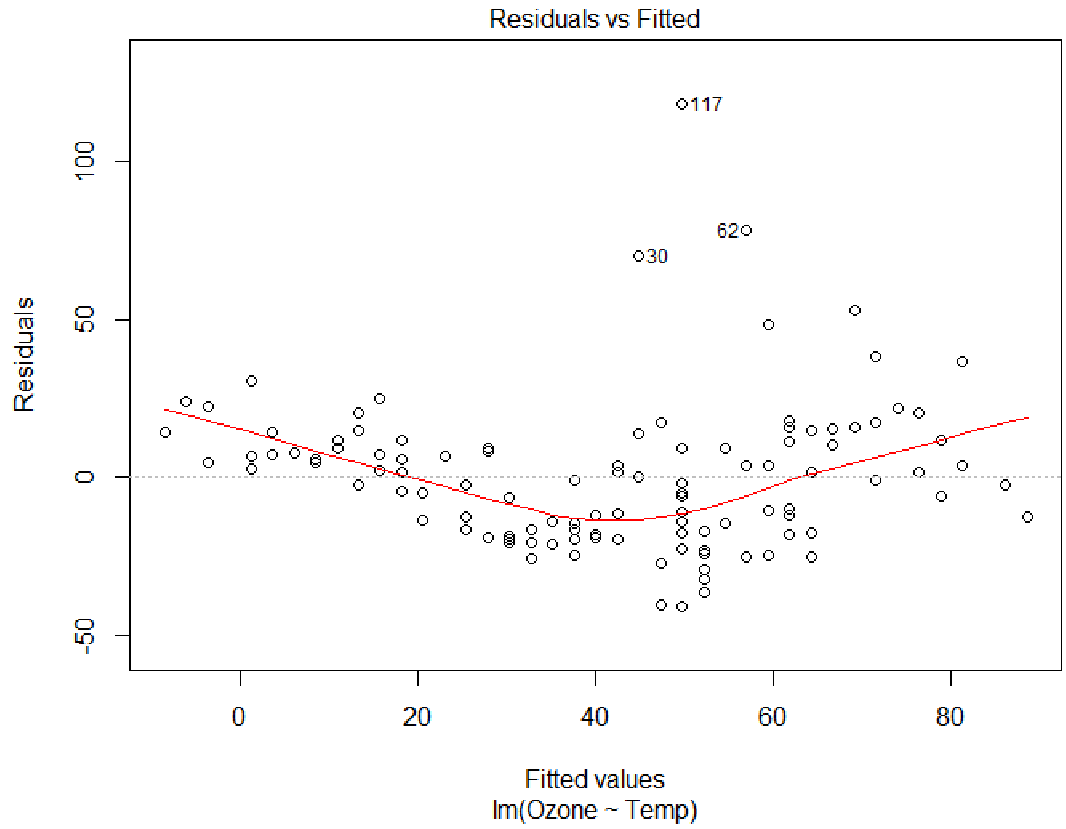
OLS assumptions - 11



OLS assumptions - 12

This is R generated plot for the ozone/temp example from SLR lecture 2. Simply type in the R console:

`plot(m1)` for a model named m1 in R



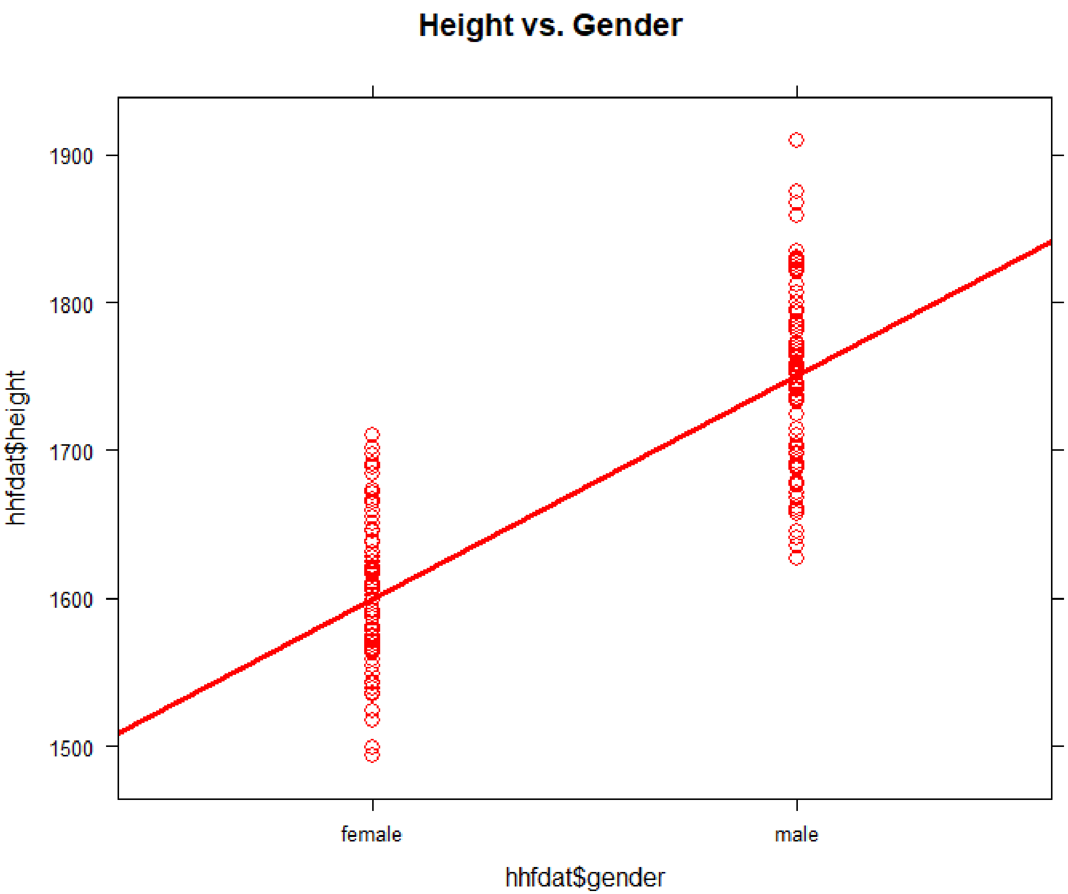
Variable types in regression

- We first discussed SLR and did an example with 2 continuous variables x and y.
- What happens if your IV is not continuous? If your IV is categorical?
- **Binary** variable (takes only 2 values).
  - 1 / 0.
  - Yes / No.
  - Male / Female.
- **Nominal** categorical (2 or more categories). Typically with no numeric values.
  - The blood type of a person: A, B, AB or O.
  - The state that a person lives in.
- So there is flexibility in the choices of the types of variables.

Example – binary variable

- Example: Is height associated with gender?
- Two variable:
  - $x =$ IV with two levels (male, female)
  - $y =$ DV- continuous.(height).
- This IV has no values.
- Does our model change? $\hat{y} = b_0 + b_1 x$ *NO*
- If so, then all of the 5 OLS assumptions still hold and we can use SLR.
- Let's plot our variables
-

Example – binary variable



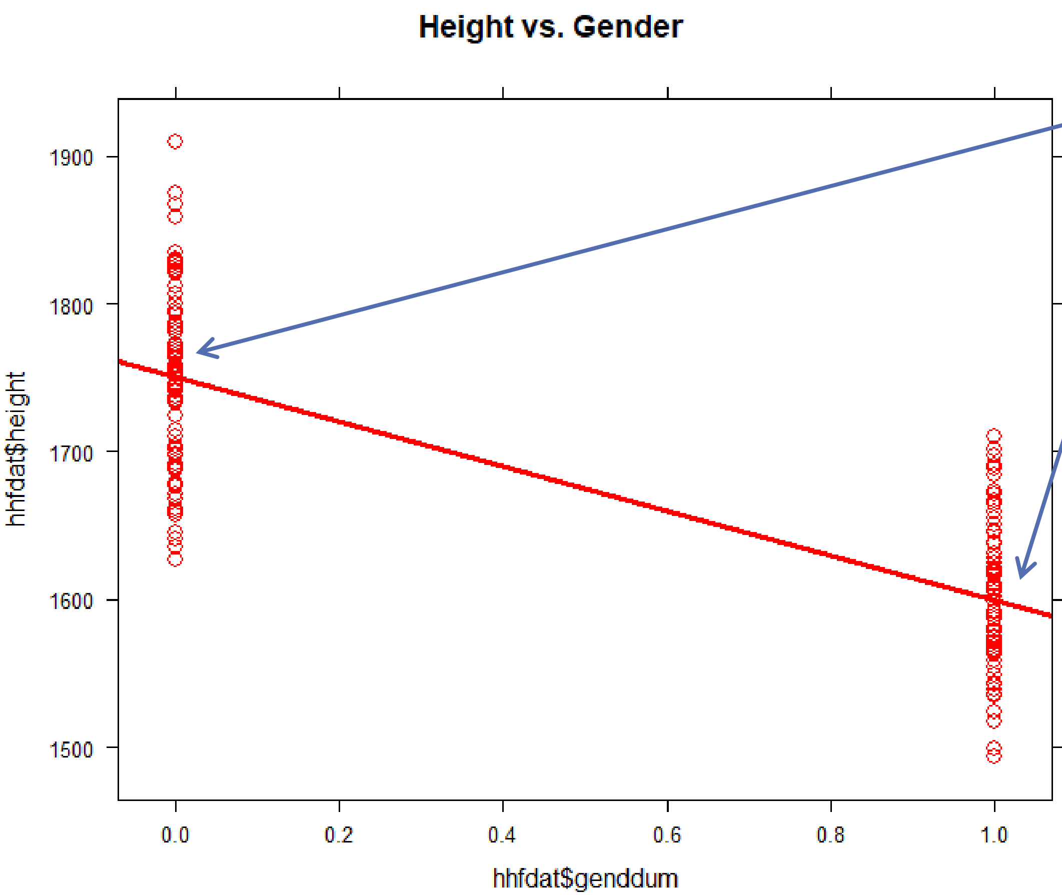
At what value does the line cross for males? And for females? The average

What does the slope now represents? The difference in average y moving from males to females (between genders).

Note: before the slope represented a change of 1 unit in x, now it represents the average from males to females.

Slope interpretation of binary SLR

- How does R handle categorical variables?



When $x = 0$ (males), the model becomes $\hat{y} = \beta_0$, so the intercept is the average of the males.

When $x = 1$ (females), the model becomes $\hat{y} = \beta_0 + \beta_1$, so the sum of the parameters is the average of the females.

So, my β_1 now is the difference between ave. of females – avg. of males $= \beta_0 + \beta_1 - \beta_0 = \boldsymbol{\beta_1}$

If betas were similar, there is no difference in height between male and female.

R - SLR - binary variable

- How does R handle categorical variables?

```
Call:
lm(formula = height ~ relevel(gender, ref = "male"), data = hhfdat,
    na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-123.361  -35.785    1.839   33.439  158.939

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1750.561      6.237  280.68   <2e-16 ***
relevel(gender, ref = "male")female -150.952      8.966  -16.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.78 on 153 degrees of freedom
Multiple R-squared:  0.6494,    Adjusted R-squared:  0.6472
F-statistic: 283.5 on 1 and 153 DF,  p-value: < 2.2e-16
```

Male=0 is the reference case

We can generally conclude that female are less in height compared to males. But more preciously:

$\beta_0 = 1750.561$ is the y-intercept at $male = 0$

$\beta_1 = -150.952$ is the coefficient for $female = 1$

β_1 is representing the difference of females relative to males (the difference).

As we are moving from males to females, the average is decreasing by -150.952

R - SLR - binary variable

- How does R handle categorical variables?

```
Call:
lm(formula = height ~ relevel(gender, ref = "male"), data = hhfdat,
    na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max 
-123.361  -35.785    1.839   33.439  158.939 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         1750.561      6.237  280.68   <2e-16 ***
relevel(gender, ref = "male")female -150.952      8.966  -16.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.78 on 153 degrees of freedom
Multiple R-squared:  0.6494,	Adjusted R-squared:  0.6472 
F-statistic: 283.5 on 1 and 153 DF,  p-value: < 2.2e-16
```

Order matters!

β_1= always non-reference level minus reference level.

There is a significate difference between females and males.

There is a significate difference in height between those who ate females and those who are males ($p - value = 2e^{-16}$)

Hypothesis test with β_1

to conduct a t-test on slope β_1:

Specify hypothesis:

- H_0: $\beta_1 = 0$ $vs.$ H_1: $\beta_1 \neq 0$ at $\alpha = 0.05$ (similar to previous class).
- If we can reject our null, this mean our slope is different than 0 and that there is a difference between both categories of gender.
- Luckily we have software that can do all that for us.

- Conclusion: We can conclude that there is a significate linear association between gender and height
- Interpretation (proper): There is a significant difference in the average height for females compared to males. Females were 150.952 mm shorter in height than males.

What happens when we have more than two categories to a variable?

- R will split the variable into smaller binary variables.
- For example, if a categorical variable X has 3 levels, some software will create 3 binary variables (x_1, x_2, x_3) representing the categories of X.
- This technique of mimicking Xrepresents level 1 vs level 3 and level 2 vs level 3. Variables created with this procedure are called “Dummy Variables”.
- Note that both levels 1, 2 are being compared to level 3, thus level 3 is our reference level.
- Other types of software are smart enough to recognize the categorical levels of a variable.
- And other times, you have to code the variable so that the software can understand the 3 levels.