

# Lesson 18: Simple Linear Regression (SLR)

Nicky Wakim

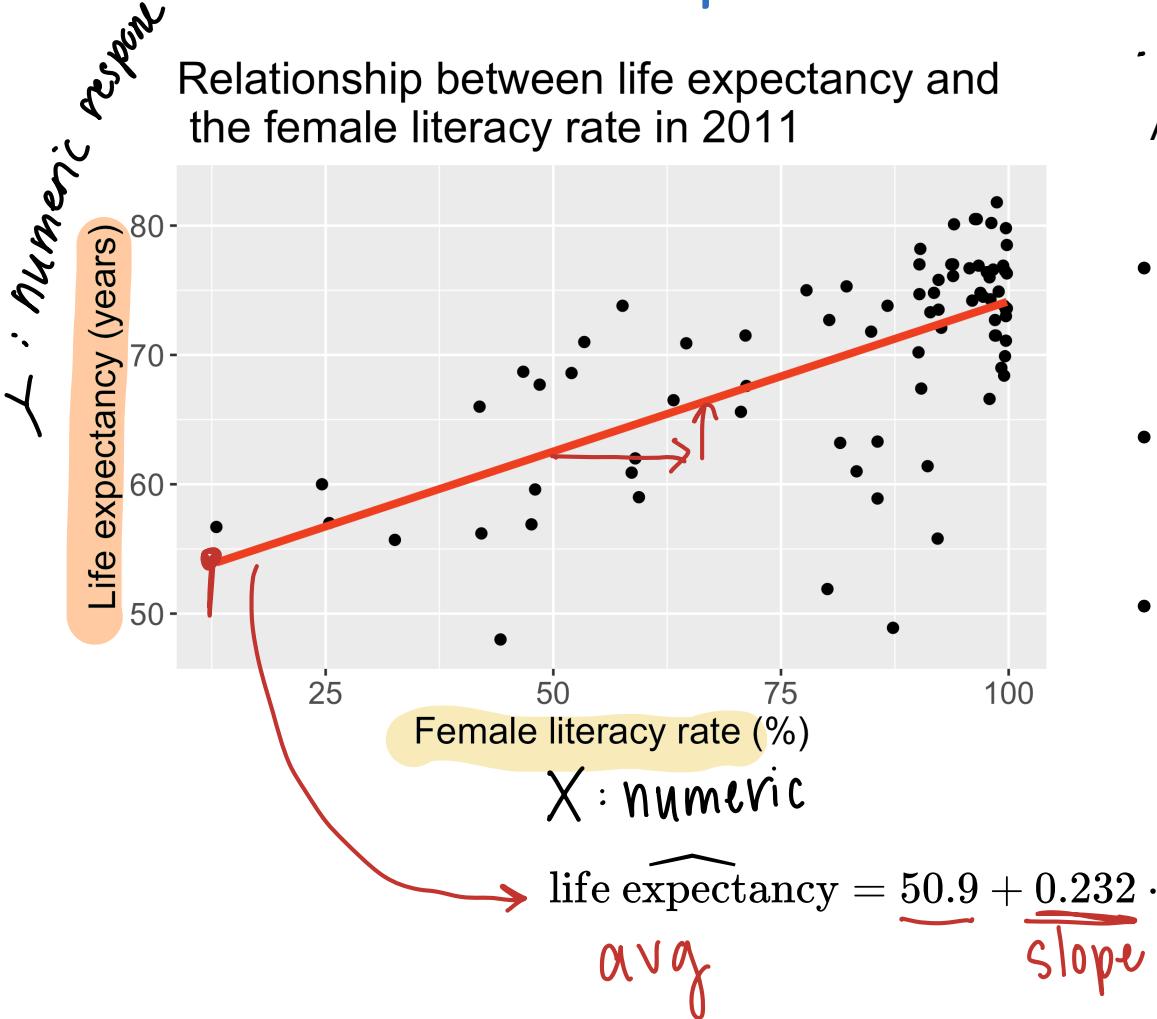
2024-12-09

# Learning Objectives

1. Identify the simple linear regression model and define statistics language for key notation
2. Illustrate how ordinary least squares (OLS) finds the best model parameter estimates
3. Apply OLS in R for simple linear regression of real data
4. Using a hypothesis test, determine if there is enough evidence that population slope  $\beta_1$  is not 0 (applies to  $\beta_0$  as well)
5. Calculate and report the estimate and confidence interval for the population slope  $\beta_1$  (applies to  $\beta_0$  as well)

# Let's start with an example

Relationship between life expectancy and the female literacy rate in 2011



two sample:

· categorial: 2 options

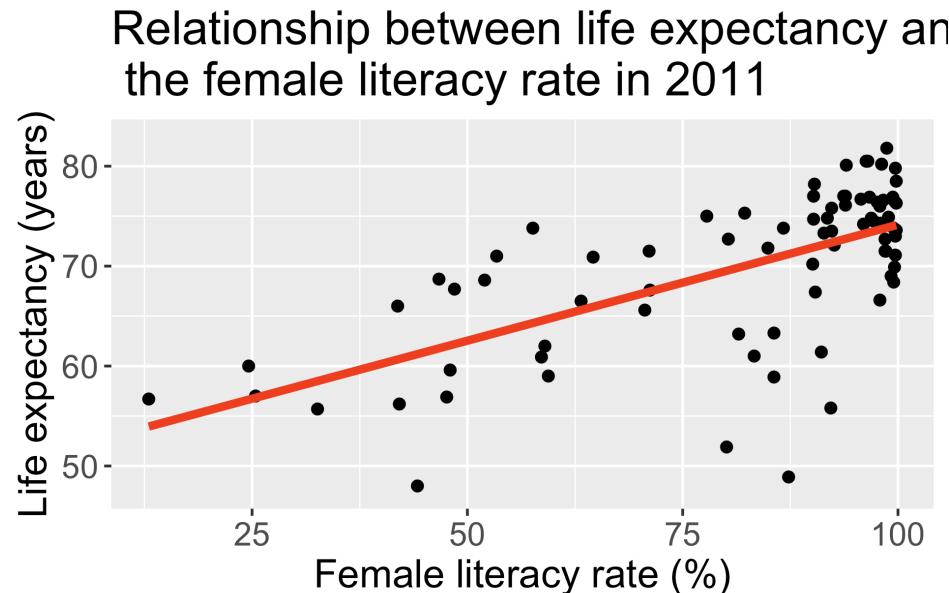
· numeric : response

Average life expectancy vs. female literacy rate

- Each point on the plot is for a different country
- $X$  = country's adult female literacy rate
- $Y$  = country's average life expectancy (years)

# Reference: How did I code that?

```
1 ggplot(gapm, aes(x = female_literacy_rate_2011,
2                   y = life_expectancy_years_2011)) +
3   geom_point(size = 4) +
4   geom_smooth(method = "lm", se = FALSE, size = 3, colour="#F14124") +
5   labs(x = "Female literacy rate (%)",
6        y = "Life expectancy (years)",
7        title = "Relationship between life expectancy and \n the female literacy rate in 2011") +
8   theme(axis.title = element_text(size = 30),
9         axis.text = element_text(size = 25),
10        title = element_text(size = 30))
```



# Dataset description

- Data files
  - Cleaned: `lifeexp_femlit_2011.csv`
  - Needs cleaning: `lifeexp_femlit_water_2011.csv`
- Data were downloaded from [Gapminder](#)
- 2011 is the most recent year with the most complete data
- **Life expectancy** = the average number of years a newborn child would live if current mortality patterns were to stay the same.
- **Adult literacy rate** is the percentage of people ages 15 and above who can, with understanding, read and write a short, simple statement on their everyday life.

# Get to know the data (1/2)

- Load data

```
1 gapm_original <- read_csv(here::here("data", "lifeexp_femlit_2011.csv"))
```

- Glimpse of the data

```
1 glimpse(gapm_original)
```

```
Rows: 188
Columns: 3
$ country           <chr> "Afghanistan", "Albania", "Algeria", "Andor...
- $ life_expectancy_years_2011 <dbl> 56.7, 76.7, 76.7, 82.6, 60.9, 76.9, 76.0, 7...
- $ female_literacy_rate_2011 <dbl> 13.0, 95.7, NA, NA, 58.6, 99.4, 97.9, 99.5,...
```

- Note the missing values for our variables of interest

## Get to know the data (2/2)

- Get a sense of the summary statistics

```
1 gapm_original %>%
2   select(life_expectancy_years_2011, female_literacy_rate_2011) %>%
3   summary()
```

	life_expectancy_years_2011	female_literacy_rate_2011
Min.	:47.50	Min. :13.00
1st Qu.	:64.30	1st Qu.:70.97
Median	:72.70	Median :91.60
Mean	:70.66	Mean :81.65
3rd Qu.	:76.90	3rd Qu.:98.03
Max.	:82.90	Max. :99.80
NA's	:1	NA's :108

# Poll Everywhere Question 1

13:15 Mon Dec 9

Join by Web [PollEv.com/nickywakim275](https://PollEv.com/nickywakim275)

What are other ways you would get to know your data? (Hint: What else have we learned to visualize or summarize the data?)

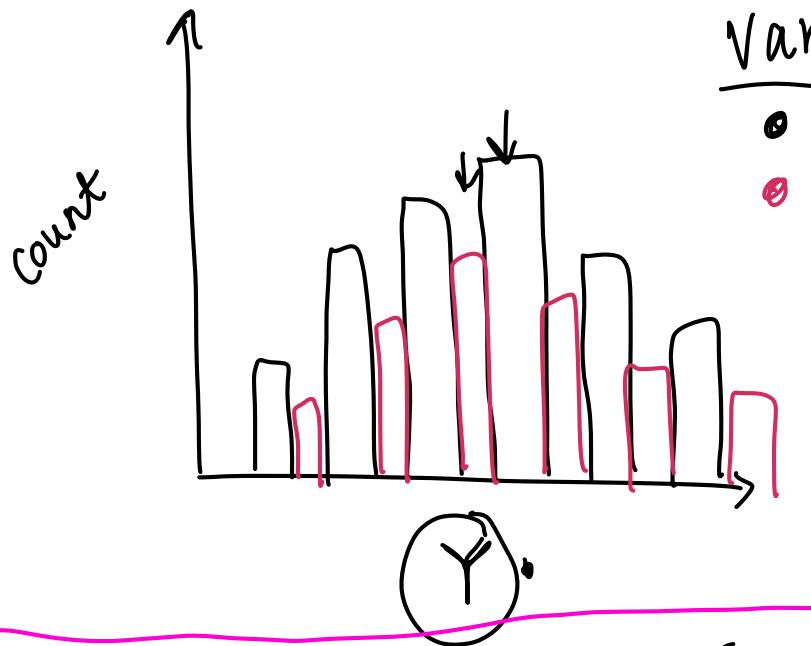
Plot scatter plot

0 likes 0 dislikes

Getsummarystats()

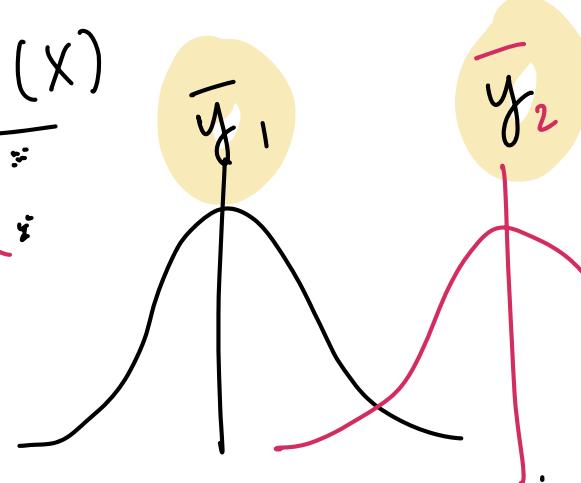
0 likes 0 dislikes

Powered by  Poll Everywhere

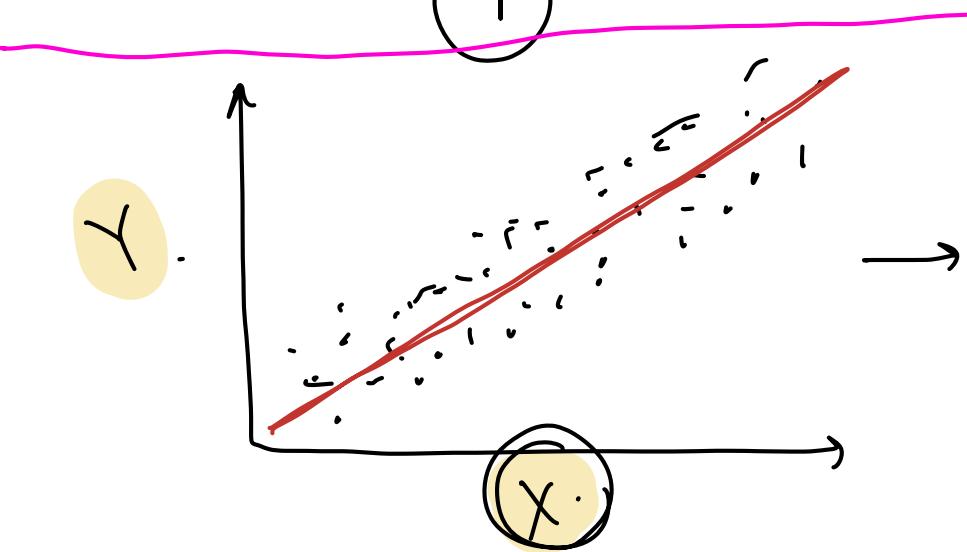


Variable 2 (X)

- cat 1
- cat 2



two, t-test  
sample



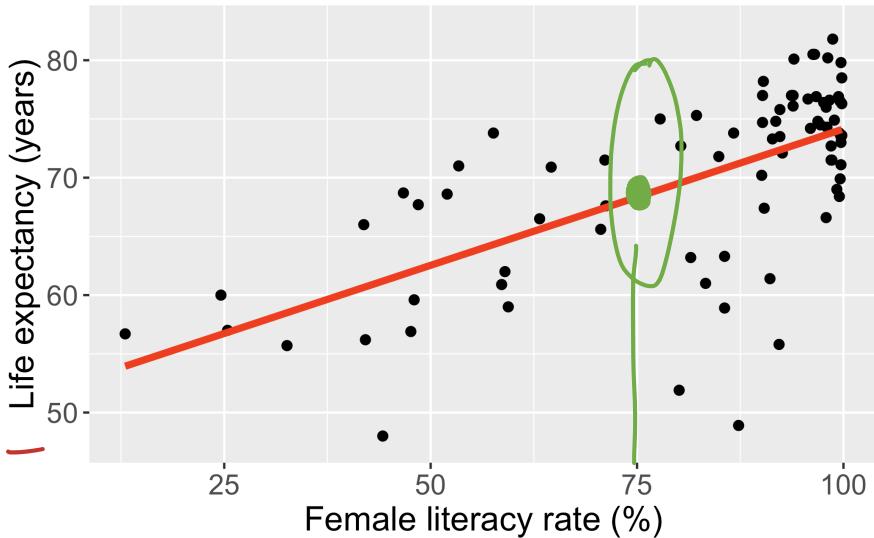
estimate  
slope

(Y)  
outcome =  
cont/num  
explanatory:  
(X): binary/  
cat.

outcome (Y):  
cont/num  
explanatory (X):  
cont/num

# Questions we can ask with a simple linear regression model

Relationship between life expectancy and the female literacy rate in 2011



$$\widehat{\text{life expectancy}} = 50.9 + 0.232 \cdot \text{female literacy rate}$$

- How do we...
  - calculate slope & intercept? ✓
  - interpret slope & intercept? ✓
  - do inference for slope & intercept?
    - CI, p-value
  - do prediction with regression line?
    - CI for prediction?
- Does the model fit the data well?
  - Should we be using a line to model the data? ✓
- Should we add additional variables to the model?
  - multiple/multivariable regression

BSTA  
512

# Learning Objectives

1. Identify the simple linear regression model and define statistics language for key notation
2. Illustrate how ordinary least squares (OLS) finds the best model parameter estimates
3. Apply OLS in R for simple linear regression of real data
4. Using a hypothesis test, determine if there is enough evidence that population slope  $\beta_1$  is not 0 (applies to  $\beta_0$  as well)
5. Calculate and report the estimate and confidence interval for the population slope  $\beta_1$  (applies to  $\beta_0$  as well)

# Simple Linear Regression Model

The (population) regression model is denoted by:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Diagram illustrating the components of the population regression model:

- Y**: Dependent variable (pink circle)
- β₀**: Intercept (red circle)
- β₁X**: Slope (red circle)
- ϵ**: Error term (red circle)

Annotations in red:

- intercept**: Points to the  $\beta_0$  term.
- slope**: Points to the  $\beta_1 X$  term.

## Unobservable population parameters

- $\beta_0$  and  $\beta_1$  are **unknown** population parameters
- $\epsilon$  (epsilon) is the error about the line
  - It is assumed to be a random variable with a...
    - Normal distribution with mean 0 and constant variance  $\sigma^2$
    - i.e.  $\epsilon \sim N(0, \sigma^2)$

*pop param:  
variance of error*

## Observable sample data

- $Y$  is our dependent variable
  - Aka outcome or response variable
- $X$  is our independent variable
  - Aka predictor, regressor, exposure variable

# Simple Linear Regression Model (another way to view components)

The (population) regression model is denoted by:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

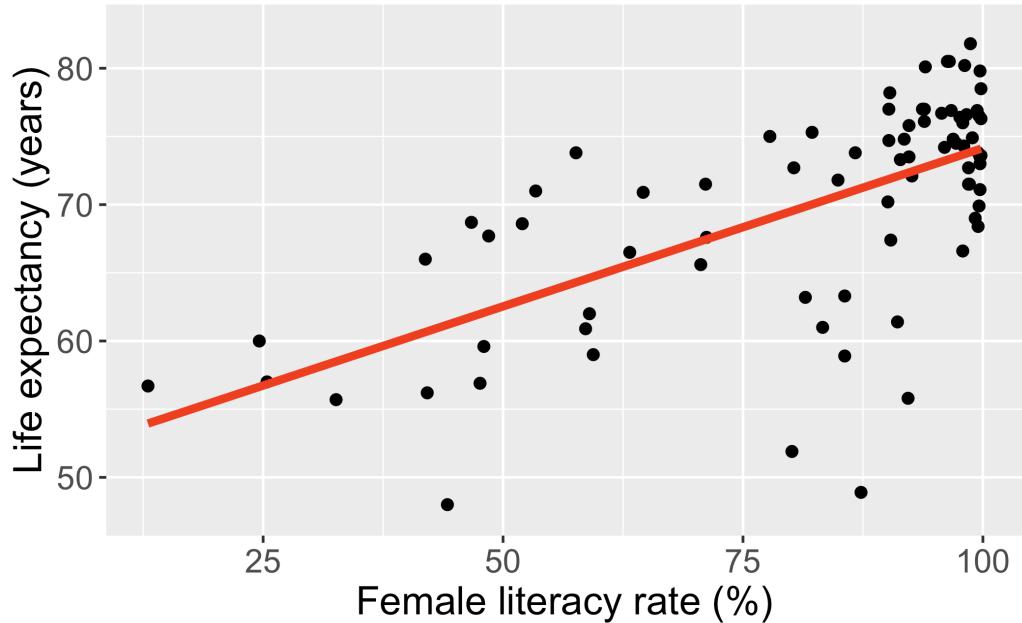
Component	Type	Name
$Y$	Observed	response, outcome, dependent variable
$\beta_0$	Pop. parameter	intercept
$\beta_1$	Pop. parameter	slope
$X$	Observed	predictor, covariate, independent variable
$\epsilon$	Pop. parameter	residuals, error term

# If the population parameters are unobservable, how did we get the line for life expectancy?

Note: the **population model** is the true, underlying model that we are trying to estimate using our sample data

- Our goal in simple linear regression is to estimate  $\beta_0$  and  $\beta_1$

Relationship between life expectancy and the female literacy rate in 2011



# Poll Everywhere Question 2

13:29 Mon Dec 9

Join by Web [PollEv.com/nickywakim275](https://PollEv.com/nickywakim275)

estimated  
What do we label as the slope of the red line?

Relationship between life expectancy and the female literacy rate in 2011

Life expectancy (years)

Female literacy rate (%)

$\beta_1$

9%

$\hat{\beta}_1$

73%

$\hat{\beta}_0$

9%

Powered by Poll Everywhere

estimated  
intercept :  $\hat{\beta}_0$

estimated =  $\wedge$

# Regression line = best-fit line

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

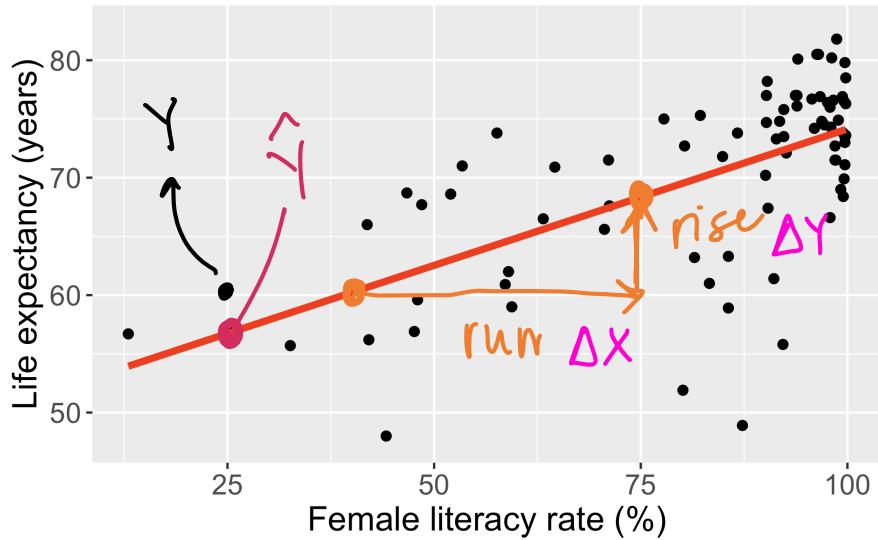
25

- $\hat{Y}$  is the predicted outcome for a specific value of  $X$
- $\hat{\beta}_0$  is the intercept of the best-fit line
- $\hat{\beta}_1$  is the slope of the best-fit line, i.e., the increase in  $\hat{Y}$  for every increase of one (unit increase) in  $X$ 
  - slope = rise over run

$$\text{slope} = \frac{\Delta Y}{\Delta X}$$

"change in"

Relationship between life expectancy and the female literacy rate in 2011



# Simple Linear Regression Model

Population regression *model*    theoretical

$$Y = \beta_0 + \beta_1 X + \epsilon$$

## Components

$Y$  response, outcome, dependent variable

$\beta_0$  intercept

$\beta_1$  slope

$X$  predictor, covariate, independent variable

$\epsilon$  residuals, error term

to be fitted

Estimated regression *line*

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

## Components

$\hat{Y}$  estimated expected response given predictor  $X$

$\hat{\beta}_0$  estimated intercept

$\hat{\beta}_1$  estimated slope

$X$  predictor, covariate, independent variable

fitted: used the data to find best fit line

# Learning Objectives

1. Identify the simple linear regression model and define statistics language for key notation
2. Illustrate how ordinary least squares (OLS) finds the best model parameter estimates
3. Apply OLS in R for simple linear regression of real data
4. Using a hypothesis test, determine if there is enough evidence that population slope  $\beta_1$  is not 0 (applies to  $\beta_0$  as well)
5. Calculate and report the estimate and confidence interval for the population slope  $\beta_1$  (applies to  $\beta_0$  as well)

# It all starts with a residual...

- Recall, one characteristic of our population model was that the residuals,  $\epsilon$ , were Normally distributed:

$$\epsilon \sim N(0, \sigma^2)$$

- In our population regression model, we had:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- We can also take the average (expected) value of the population model  $\beta_0$  &  $\beta_1$  are constants!
- We take the expected value of both sides and get:

$$\rightarrow E[Y] = E[\beta_0 + \beta_1 X + \epsilon]$$

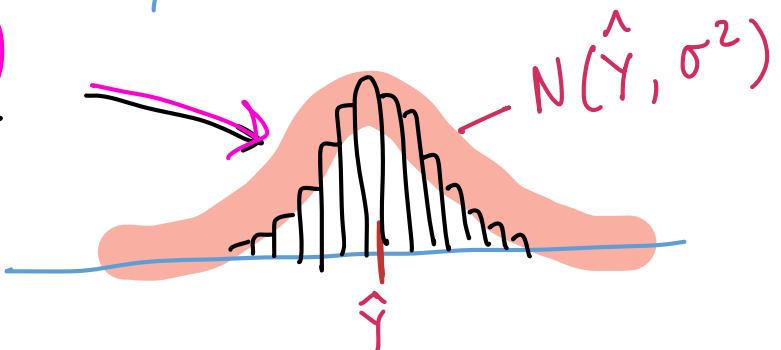
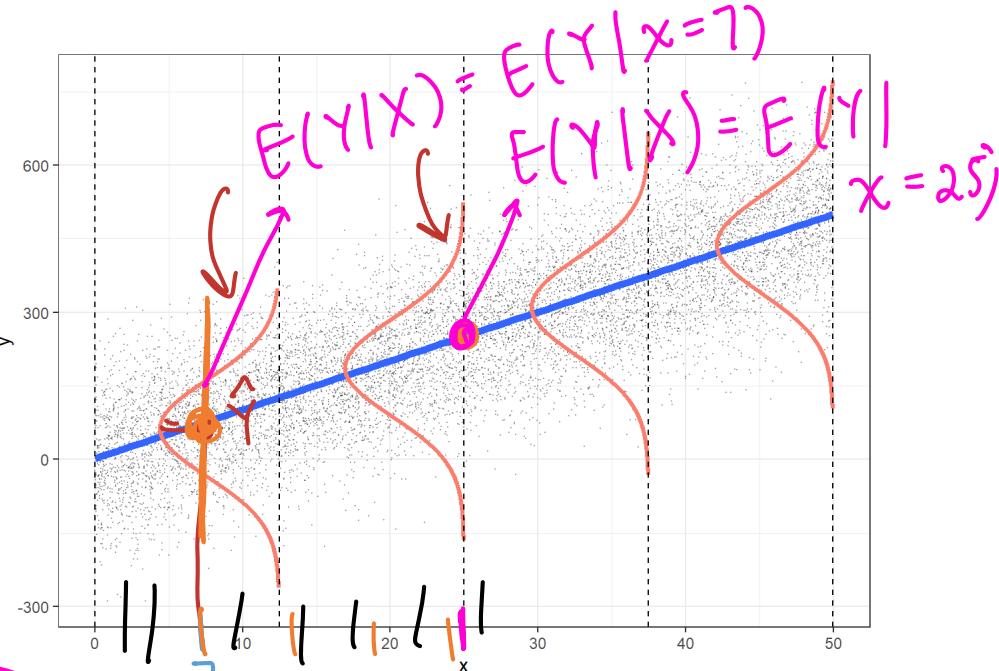
$$E[Y] = E[\beta_0] + E[\beta_1 X] + E[\epsilon]$$

$$E[Y] = \beta_0 + \beta_1 \underbrace{E[X]}_{=0}$$

$$E[Y|X] = \underline{\beta_0 + \beta_1 X}$$

@  $X=7$

- We call  $E[Y|X]$  the expected value of  $Y$  given  $X$



# So now we have two representations of our population model

With observed  $Y$  values and residuals: *pop model*

$$Y = \beta_0 + \beta_1 X + \epsilon$$

*residual error*

*pop model*  
With the population expected value of  $Y$  given  $X$ :

$$\rightarrow E[Y|X] = \beta_0 + \beta_1 X$$

Using the two forms of the model, we can figure out a formula for our residuals:

$$\begin{aligned} Y &= (\beta_0 + \beta_1 X) + \epsilon \\ Y &= E[Y|X] + \epsilon \\ Y - E[Y|X] &= \epsilon \\ \epsilon &= \underbrace{Y}_{\text{obs}} - \underbrace{E[Y|X]}_{\text{exp value (pop)}} \end{aligned}$$

And so we have our true, population model, residuals!

This is an important fact! For the **population model**, the residuals:  $\epsilon = Y - E[Y|X]$

## Back to our **estimated model** fitted model

We have the same two representations of our estimated/fitted model:

With observed values:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\epsilon}$$

With the estimated expected value of  $Y$  given  $X$ :

$$\begin{aligned}\hat{E}[Y|X] &= \hat{\beta}_0 + \hat{\beta}_1 X \\ \widehat{E[Y|X]} &= \hat{\beta}_0 + \hat{\beta}_1 X \\ \hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 X\end{aligned}$$

Using the two forms of the model, we can figure out a formula for our estimated residuals:

$$Y = (\hat{\beta}_0 + \hat{\beta}_1 X) + \hat{\epsilon}$$

$$Y = \hat{Y} + \hat{\epsilon}$$

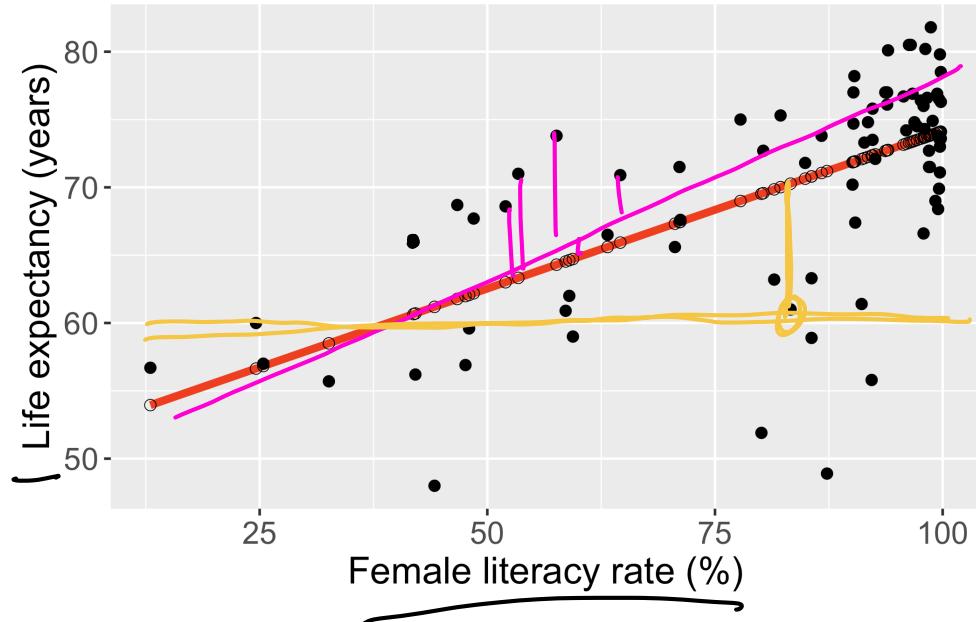
$$\hat{\epsilon} = Y - \hat{Y}$$

This is an important fact! For the **estimated/fitted model**, the residuals:  $\hat{\epsilon} = Y - \hat{Y}$

# Individual $i$ residuals in the estimated/fitted model

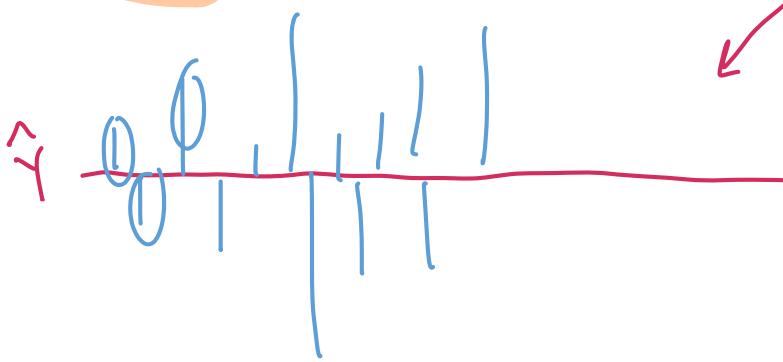
- Observed values for each individual  $i$ :  $Y_i$ 
  - Value in the dataset for individual  $i$
- Fitted value for each individual  $i$ :  $\hat{Y}_i$ 
  - Value that falls on the best-fit line for a specific  $X_i$
  - If two individuals have the same  $X_i$ , then they have the same  $\hat{Y}_i$

Relationship between life expectancy and the female literacy rate in 2011



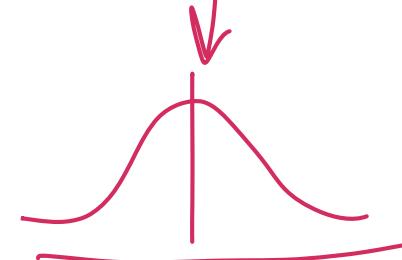
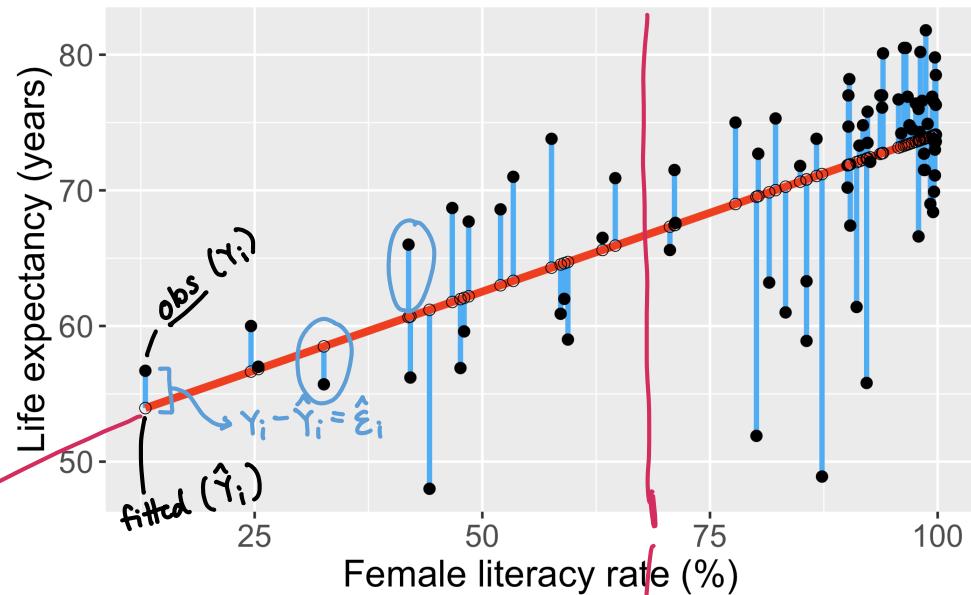
# *Individual $i$ residuals in the estimated/fitted model*

- Observed values for each individual  $i$ :  $Y_i$ 
  - Value in the dataset for individual  $i$
- Fitted value for each individual  $i$ :  $\hat{Y}_i$ 
  - Value that falls on the best-fit line for a specific  $X_i$
  - If two individuals have the same  $X_i$ , then they have the same  $\hat{Y}_i$
- Residual for each individual:  $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ 
  - Difference between the observed and fitted value



$$\varepsilon \sim N(0, \sigma^2)$$

Relationship between life expectancy and the female literacy rate in 2011



# Poll Everywhere Question 3

13:51 Mon Dec 9

Join by Web [PollEv.com/nickywakim275](https://PollEv.com/nickywakim275)

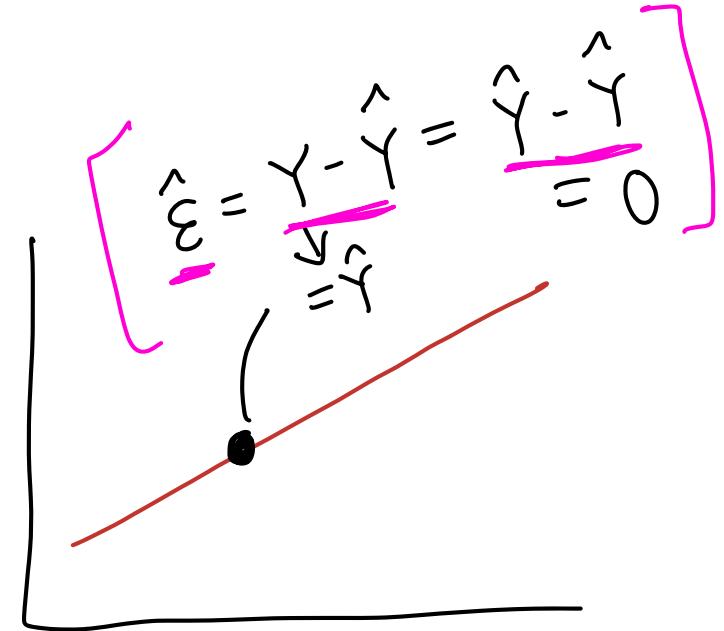
QR code

If our observed  $\hat{Y}$  value fell exactly on the best-fit line, what would the residual be?

0

0

Powered by  Poll Everywhere



# So what do we do with the residuals?

- We want to minimize the sum of residuals
  - Aka minimize the difference between the observed  $Y$  value and the estimated expected response given the predictor ( $\hat{E}[Y|X]$ )
- We can use ordinary least squares (OLS) to do this in linear regression!
- Idea behind this: reduce the total error between the fitted line and the observed point (error between is called residuals)
  - Vague use of total error: more precisely, we want to reduce the sum of squared errors ✓
  - ~~Think back to my R Shiny app!~~
  - We need to mathematically define this!
- Note: there are other ways to estimate the best-fit line!!
  - Example: Maximum likelihood estimation

# Break

# Learning Objectives

1. Identify the simple linear regression model and define statistics language for key notation
2. Illustrate how ordinary least squares (OLS) finds the best model parameter estimates
3. Apply OLS in R for simple linear regression of real data
4. Using a hypothesis test, determine if there is enough evidence that population slope  $\beta_1$  is not 0 (applies to  $\beta_0$  as well)
5. Calculate and report the estimate and confidence interval for the population slope  $\beta_1$  (applies to  $\beta_0$  as well)

# Setting up for ordinary least squares

- Sum of Squared Errors (SSE)

$$SSE = \sum_{i=1}^n \hat{\epsilon}_i^2$$

$$SSE = \sum_{i=1}^n (Y_i - \underline{\widehat{Y}_i})^2$$

$$SSE = \sum_{i=1}^n (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i))^2$$

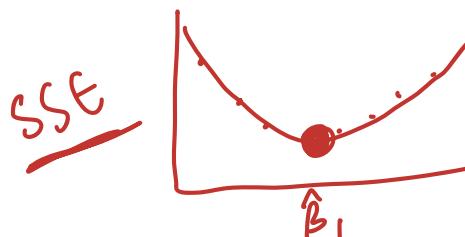
$$\textcircled{SSE} = \sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

error/  
residual

- $\hat{\epsilon}_i = Y_i - \widehat{Y}_i$
- $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$

find  $\widehat{\beta}_0$  &  $\widehat{\beta}_1$  that minimize  
this sum!

Then we want to find the estimated coefficient values that minimize the SSE!



# Poll Everywhere Question 4

14:09 Mon Dec 9

X Join by Web PollEv.com/nickywakim275

What do  $\hat{\beta}_0$  and  $\hat{\beta}_1$  mean for our model?

undescribed values  
(and does not have  
a set value in terms of real #s)

They are the coefficient estimates that minimize every residual value 0%

They are the coefficient estimates that are closest to the population parameters 38%

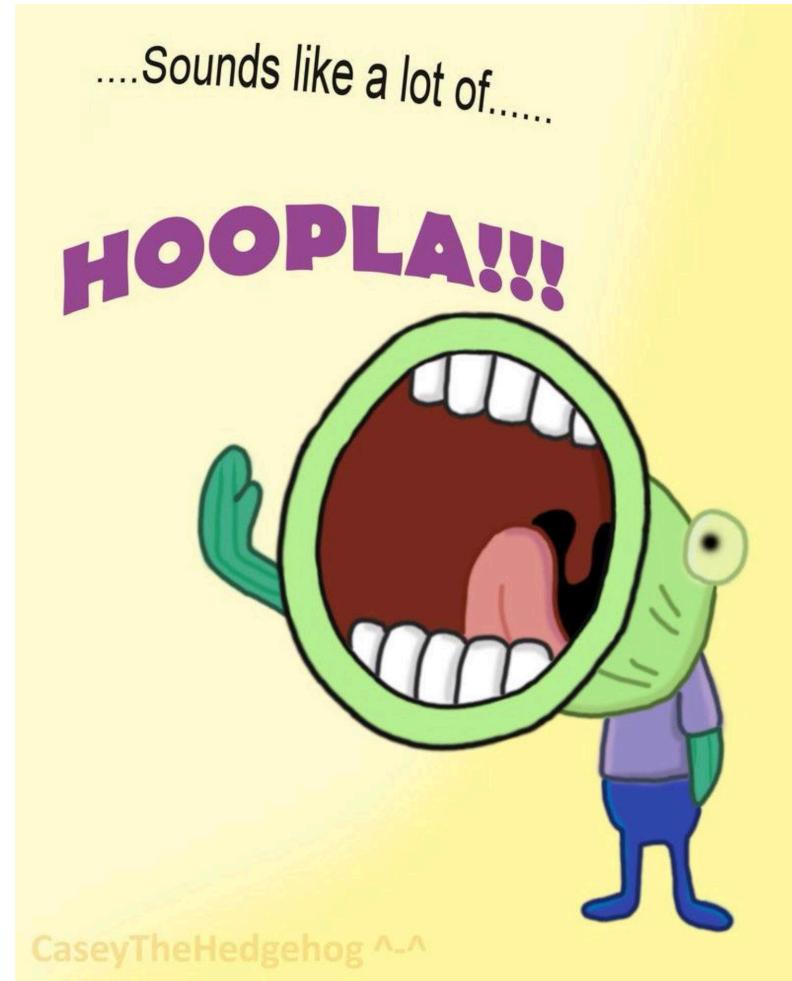
They are the coefficient estimates that perfectly fit our data 0%

They are the coefficient estimates that minimize the sum of the squared residual error 63%

Powered by Poll Everywhere

$H_0: \underline{\mu} = \text{prescribed value}$

So how do I find the coefficient estimates that minimize the SSE?



# Regression in R: lm() - linear model

- Let's discuss the syntax of this function

```
1 model1 <- lm(life_expectancy_years_2011 ~ female_literacy_rate_2011,  
2 data = gapm)
```

distributed  
as

math

$$X \sim N(0, 1)$$

$Y \sim X$   
"tilde"  
**RESPONSE**

t.test ( $Y \sim X$ )

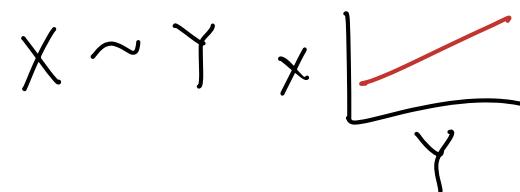
explanatory /  
independent /  
exposure /  
predictor

model w/  
diff in means

R

$$Y \sim X$$
  
**model**

using linear regression



# Regression in R: `lm()` + `summary()`

```
1 modell <- lm(life_expectancy_years_2011 ~ female_literacy_rate_2011,  
2                           data = gapm)  
3 summary(modell)
```

Call:

```
lm(formula = life_expectancy_years_2011 ~ female_literacy_rate_2011,  
  data = gapm)
```

Residuals:

E  
Min 1Q Median 3Q Max  
-22.299 -2.670 1.145 4.114 9.498

Mean : 0

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	50.92790	2.66041	19.143	< 2e-16 ***
female_literacy_rate_2011	0.23220	0.03148	7.377	1.5e-10 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.142 on 78 degrees of freedom  
(108 observations deleted due to missingness)

Multiple R-squared: 0.4109, Adjusted R-squared: 0.4034

F-statistic: 54.41 on 1 and 78 DF, p-value: 1.501e-10



# Regression in R: `lm()` + `tidy()`

```
1 tidy(modell) %>%
2   gt() %>%
3   tab_options(table.font.size = 45)
```

*Coeff:  
part of  
Summary()*

term	estimate	std.error	statistic	p.value
(Intercept)	50.9278981	2.66040695	19.142898	3.325312e-31
female_literacy_rate_2011	0.2321951	0.03147744	7.376557	1.501286e-10

- Regression equation for our model (which we saw a looong time ago):

$$\text{life expectancy} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{female literacy rate}$$

$\hat{\beta}_0$        $\hat{\beta}_1$

# How do we interpret the coefficients?

$$\widehat{\text{life expectancy}} = 50.9 + 0.232 \cdot \text{female literacy rate}$$

- Intercept
  - The expected outcome for the  $Y$ -variable when the  $X$ -variable is 0
  - Example: The expected/average life expectancy is 50.9 years for a country with 0% female literacy.

- Slope
  - For every increase of 1 unit in the  $X$ -variable, there is an expected increase of, ~~on average~~,  $\hat{\beta}_1$  units in the  $Y$ -variable.
  - We only say that there is an expected increase and not necessarily a causal increase.
  - Example: For every 1 percent increase in the female literacy rate, life expectancy increases, on average, 0.232 years.

- You can say either expected OR average

# Learning Objectives

1. Identify the simple linear regression model and define statistics language for key notation
2. Illustrate how ordinary least squares (OLS) finds the best model parameter estimates
3. Apply OLS in R for simple linear regression of real data
4. Using a hypothesis test, determine if there is enough evidence that population slope  $\beta_1$  is not 0 (applies to  $\beta_0$  as well)
5. Calculate and report the estimate and confidence interval for the population slope  $\beta_1$  (applies to  $\beta_0$  as well)

# Steps in hypothesis testing

1. Check the assumptions regarding the properties of the underlying variable(s) being measured that are needed to justify use of the testing procedure under consideration.
2. State the null hypothesis  $H_0$  and the alternative hypothesis  $H_A$ .
3. Specify the significance level  $\alpha$ .
4. Specify the test statistic to be used and its distribution under  $H_0$ .

↓ Critical region method

5. Form the decision rule for rejecting or not rejecting  $H_0$  (i.e., specify the rejection and nonrejection regions for the test, based on both  $H_A$  and  $\alpha$ ).
6. Compute the value of the test statistic from the observed data.

↓

7. Draw conclusions regarding rejection or nonrejection of  $H_0$ .

↓ p-value method

5. Compute the value of the test statistic from the observed data.
6. Calculate the p-value

↓

parallel to one sample mean process!!

## General steps for hypothesis test for population slope $\beta_1$ (t-test)

1. For today's class, we are assuming that we have met the underlying assumptions (checked in our Model Evaluation step)

2. State the null hypothesis.

Often, we are curious if the coefficient is 0 or not:

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ \text{vs. } H_A : \beta_1 &\neq 0 \end{aligned}$$

3. Specify the significance level.

Often we use  $\alpha = 0.05$

4. Specify the test statistic and its distribution under the null

The test statistic is  $t$ , and follows a Student's t-distribution.

5. Compute the value of the test statistic

The calculated **test statistic** for  $\hat{\beta}_1$  is

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{SE}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\text{SE}_{\hat{\beta}_1}}$$

when we assume  $H_0 : \beta_1 = 0$  is true.

6. Calculate the p-value

We are generally calculating:  $2 \cdot P(T > t)$

7. Write conclusion for hypothesis test

We (reject/fail to reject) the null hypothesis that the slope is 0 at the  $100\alpha\%$  significance level. There is (sufficient/insufficient) evidence that there is significant association between ( $Y$ ) and ( $X$ ) (p-value =  $P(T > t)$ ).

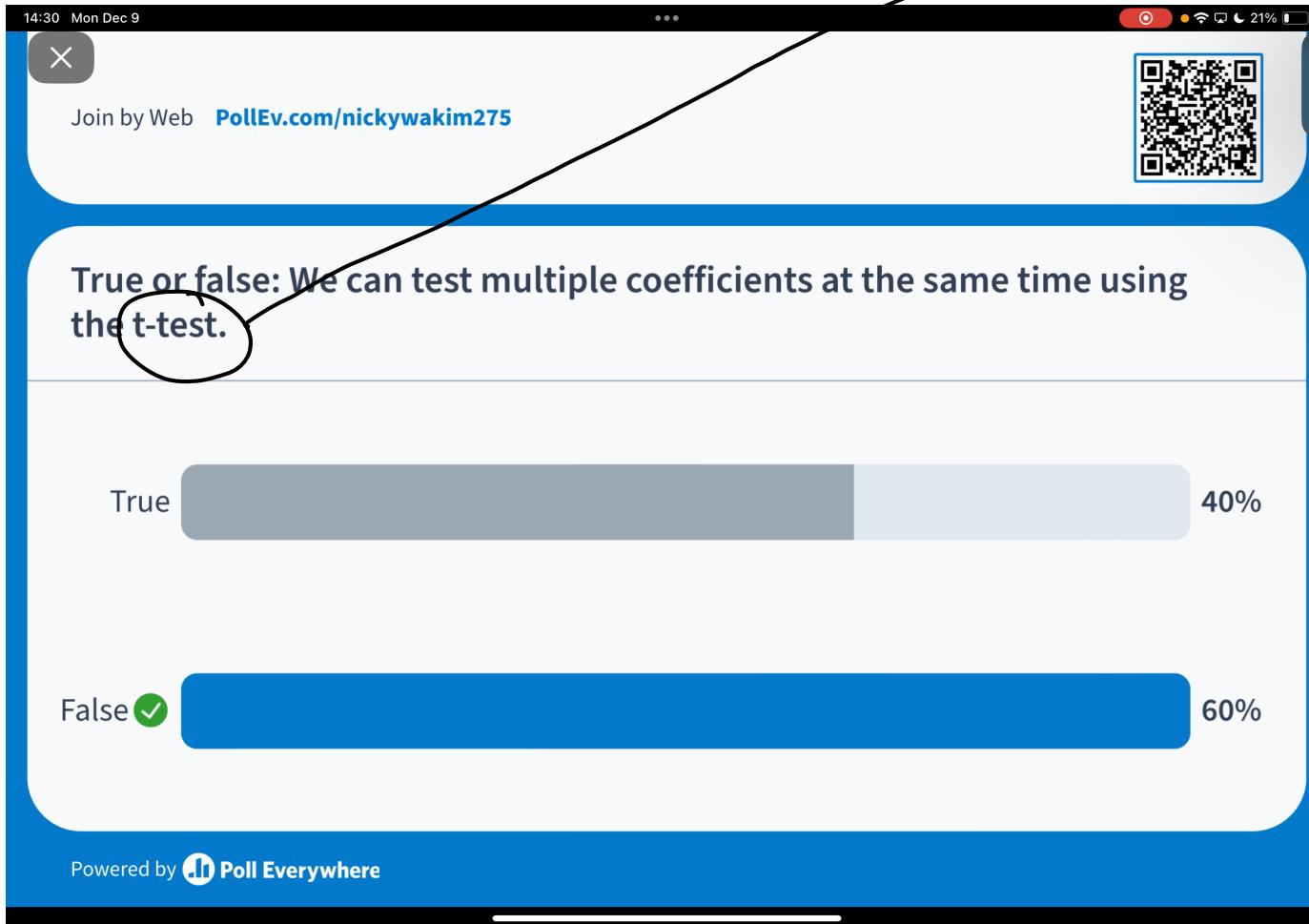
# Some important notes

- Today we are discussing the hypothesis test for a **single** coefficient
- The test statistic for a single coefficient follows a Student's t-distribution
  - It can also follow an F-distribution, but we will discuss this more with multiple linear regression and multi-level categorical covariates
- Single coefficient testing can be done on any coefficient, but it is most useful for continuous covariates or binary covariates
  - This is because testing the single coefficient will still tell us something about the overall relationship between the covariate and the outcome
  - We will talk more about this with multiple linear regression and multi-level categorical covariates

$$\begin{array}{ll} \left[ H_0 : \beta_1 = 0 \right] & \beta_1 = \beta_2 = \beta_3 = 0 \\ \left[ H_a : \beta_1 \neq 0 \right] & \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \dots \end{array}$$

## Poll Everywhere Question 5

only used for a  
single  
coefficient



# Life expectancy example: hypothesis test for population slope $\beta_1$ (1/4)

- Steps 1-4 are setting up our hypothesis test: not much change from the general steps

1. For today's class, we are assuming that we have met the underlying assumptions (checked in our Model Evaluation step)

2. State the null hypothesis.

We are testing if the slope is 0 or not:

$$H_0 : \beta_1 = 0$$

vs.  $H_A : \beta_1 \neq 0$

3. Specify the significance level.

Often we use  $\alpha = 0.05$

4. Specify the test statistic and its distribution under the null

The test statistic is  $t$ , and follows a Student's t-distribution.

# Life expectancy example: hypothesis test for population slope $\beta_1$ (2/4)

5. Compute the value of the test statistic

- **Option 1:** Calculate the test statistic using the values in the regression table

```
1 # recall model1_b1 is regression table restricted to b1 row
2 model1_b1 <- tidy(model1) %>% filter(term == "female_literacy_rate_2011")
3 model1_b1 %>% gt() %>%
4   tab_options(table.font.size = 40) %>% fmt_number(decimals = 2)
```

term	estimate	std.error	statistic	p.value
female_literacy_rate_2011	0.23	0.03	7.38	0.00

```
1 (TestStat_b1 <- model1_b1$estimate / model1_b1$std.error)
[1] 7.376557
```

$$t = \frac{\hat{\beta}_1}{SE \hat{\beta}_1}$$

- **Option 2:** Get the test statistic value ( $t^*$ ) from R

```
1 model1_b1 %>% gt() %>%
2   tab_options(table.font.size = 40) %>% fmt_number(decimals = 2)
```

term	estimate	std.error	statistic	p.value
female_literacy_rate_2011	0.23	0.03	7.38	0.00

# Life expectancy example: hypothesis test for population slope $\beta_1$ (3/4)

## 6. Calculate the p-value

- The *p*-value is the probability of obtaining a test statistic just as extreme or more extreme than the observed test statistic assuming the null hypothesis  $H_0$  is true
- We know the probability distribution of the test statistic (the *null distribution*) assuming  $H_0$  is true
- Statistical theory tells us that the test statistic  $t$  can be modeled by a *t*-distribution with  $df = n - 2$ .
  - We had 80 countries' data, so  $n = 80$
- Option 1:** Use `pt()` and our calculated test statistic

```
1 (pv = 2*pt(TestStat_b1, df=80-2, lower.tail=F))  
[1] 1.501286e-10
```

- Option 2:** Use the regression table output

```
1 modell_b1 %>% gt() %>%  
2 tab_options(table.font.size = 40)
```

term	estimate	std.error	statistic	p.value
female_literacy_rate_2011	0.2321951	0.03147744	7.376557	1.501286e-10

## Life expectancy example: hypothesis test for population slope $\beta_1$ (4/4)

7. Write conclusion for the hypothesis test

We reject the null hypothesis that the slope is 0 at the 5% significance level. There is sufficient evidence that there is significant association between female life expectancy and female literacy rates ( $p\text{-value} < 0.0001$ ).

There is suff evidence that for 1% increase in FLR there is an expected inc of 0.232 yrs in life expectancy ( $p\text{-val} < 0.009$ )

# Life expectancy ex: hypothesis test for population intercept $\beta_0$ (1/4)

- Steps 1-4 are setting up our hypothesis test: not much change from the general steps
1. For today's class, we are assuming that we have met the underlying assumptions (checked in our Model Evaluation step)
  2. State the null hypothesis.

We are testing if the intercept is 0 or not:

$$H_0 : \beta_0 = 0$$

vs.  $H_A : \beta_0 \neq 0$

3. Specify the significance level

Often we use  $\alpha = 0.05$

4. Specify the test statistic and its distribution under the null

This is the same as the slope. The test statistic is  $t$ , and follows a Student's t-distribution.

# Life expectancy ex: hypothesis test for population intercept $\beta_0$ (2/4)

5. Compute the value of the test statistic

- **Option 1:** Calculate the test statistic using the values in the regression table

```
1 # recall model1_b1 is regression table restricted to b1 row
2 model1_b0 <- tidy(model1) %>% filter(term == "(Intercept)")
3 model1_b0 %>% gt() %>%
4   tab_options(table.font.size = 40) %>% fmt_number(decimals = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	50.93	2.66	19.14	0.00

```
1 (TestStat_b0 <- model1_b0$estimate / model1_b0$std.error)
[1] 19.1429
```

- **Option 2:** Get the test statistic value ( $t^*$ ) from R

```
1 model1_b0 %>% gt() %>%
2   tab_options(table.font.size = 40) %>% fmt_number(decimals = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	50.93	2.66	19.14	0.00

# Life expectancy ex: hypothesis test for population intercept $\beta_0$ (3/4)

## 6. Calculate the p-value

- **Option 1:** Use `pt()` and our calculated test statistic

```
1 (pv = 2*pt(TestStat_b0, df=80-2, lower.tail=F))  
[1] 3.325312e-31
```

- **Option 2:** Use the regression table output

```
1 model1_b0 %>% gt() %>%  
2 tab_options(table.font.size = 40)
```

term	estimate	std.error	statistic	p.value
(Intercept)	50.9279	2.660407	19.1429	3.325312e-31

## Life expectancy ex: hypothesis test for population intercept $\beta_0$ (4/4)

7. Write conclusion for the hypothesis test

We reject the null hypothesis that the intercept is 0 at the 5% significance level. There is sufficient evidence that the intercept for the association between ~~average female~~ life expectancy and female literacy rates is different from 0 (p-value < 0.0001).

- Note: if we fail to reject  $H_0$ , then we could decide to remove the intercept from the model to force the regression line to go through the origin (0,0) if it makes sense to do so for the application.

50.9

# Learning Objectives

1. Identify the simple linear regression model and define statistics language for key notation
2. Illustrate how ordinary least squares (OLS) finds the best model parameter estimates
3. Apply OLS in R for simple linear regression of real data
4. Using a hypothesis test, determine if there is enough evidence that population slope  $\beta_1$  is not 0 (applies to  $\beta_0$  as well)
5. Calculate and report the estimate and confidence interval for the population slope  $\beta_1$  (applies to  $\beta_0$  as well)

# Inference for the population slope: hypothesis test and CI

Population model

line + random "noise"

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

with  $\varepsilon \sim N(0, \sigma^2)$

$\sigma^2$  is the variance of the residuals

Sample best-fit (least-squares) line

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

Note: Some sources use  $b$  instead of  $\hat{\beta}$

We have two options for inference:

1. Conduct the **hypothesis test**

$$H_0 : \beta_1 = 0$$

$$\text{vs. } H_A : \beta_1 \neq 0$$

Note: R reports p-values for 2-sided tests

2. Construct a **95% confidence interval** for the population slope  $\beta_1$

# Confidence interval for population slope $\beta_1$

Recall the general CI formula:

$$\mu, + t^* \cdot SE$$
$$\widehat{\beta}_1 \pm t_{\alpha, n-2}^* \cdot SE_{\widehat{\beta}_1}$$

To construct the confidence interval, we need to:

- Set our  $\alpha$ -level
- Find  $\widehat{\beta}_1$
- Calculate the  $t_{n-2}^*$
- Calculate  $SE_{\widehat{\beta}_1}$

# Calculate CI for population slope $\beta_1$ (1/2)

$$\hat{\beta}_1 \pm t^* \cdot SE_{\beta_1}$$

where  $t^*$  is the  $t$ -distribution critical value with  $df = n - 2$ .

- **Option 1:** Calculate using each value

Save values needed for CI:

```
1 b1 <- modell_b1$estimate
2 SE_b1 <- modell_b1$std.error
1 nobs(modell) # sample size n
[1] 80
1 (tstar <- qt(.975, df = 80-2))
[1] 1.990847
```

Use formula to calculate each bound

```
1 (CI_LB <- b1 - tstar*SE_b1)
[1] 0.1695284
1 (CI_UB <- b1 + tstar*SE_b1)
[1] 0.2948619
```

## Calculate CI for population slope $\beta_1$ (2/2)

$$\hat{\beta}_1 \pm t^* \cdot SE_{\beta_1}$$

where  $t^*$  is the  $t$ -distribution critical value with  $df = n - 2$ .

- Option 2: Use the regression table

```
1 tidy(modell, conf.int = T) %>% gt() %>%
2   tab_options(table.font.size = 40) %>% fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	50.928	2.660	19.143	0.000	45.631	56.224
female_literacy_rate_2011	0.232	0.031	7.377	0.000	0.170	0.295

# Reporting the coefficient estimate of the population slope

- When we report our results to someone else, we don't usually show them our full hypothesis test
  - In an informal setting, someone may want to see it
- Typically, we report the estimate with the confidence interval
  - From the confidence interval, your audience can also deduce the results of a hypothesis test
- Once we found our CI, we often just write the interpretation of the coefficient estimate:

## General statement for population slope inference

For every increase of 1 unit in the  $X$ -variable, there is an expected average increase of  $\hat{\beta}_1$  units in the  $Y$ -variable (95%: LB, UB).

- In our example: For every 1% increase in female literacy rate, life expectancy increases, on average, 0.232 years (95% CI: 0.170, 0.295).

# Many options for how to word our results (Reference)

1. In our example: For every 1% increase in female literacy rate, life expectancy increases, on average, 0.232 years (95% CI: 0.170, 0.295).

2. In our example: For every 1% increase in female literacy rate, life expectancy is expected to increase 0.232 years (95% CI: 0.170, 0.295).

2. In our example: For every 1% increase in female literacy rate, the average life expectancy increases 0.232 years (95% CI: 0.170, 0.295).

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

expected or avg  $\hat{Y}$

The handwritten notes illustrate different ways to interpret the coefficient  $\hat{\beta}_1$ . The first interpretation, 'on average', focuses on the overall mean shift. The second interpretation, 'is expected to', focuses on the individual prediction shift. The third interpretation, 'average', focuses on the population mean shift.

## Poll Everywhere Question 6



## For reference: quick CI for $\beta_0$

- Calculate CI for population intercept  $\beta_0$ :  $\hat{\beta}_0 \pm t^* \cdot SE_{\beta_0}$   
where  $t^*$  is the  $t$ -distribution critical value with  $df = n - 2$

- Use the regression table

```
1 tidy(modell, conf.int = T) %>% gt() %>%
2   tab_options(table.font.size = 40) %>% fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	50.928	2.660	19.143	0.000	45.631	56.224
female_literacy_rate_2011	0.232	0.031	7.377	0.000	0.170	0.295

General statement for population intercept inference

The expected outcome for the  $Y$ -variable is  $(\hat{\beta}_0)$  when the  $X$ -variable is 0 (95% CI: LB, UB).

- For example: The average life expectancy is 50.9 years when the female literacy rate is 0 (95% CI: 45.63, 56.22).  
*yrs.*

