

Simple Linear Regression (SLR)

Nicky Wakim

2023-01-17

Wednesday 1/17

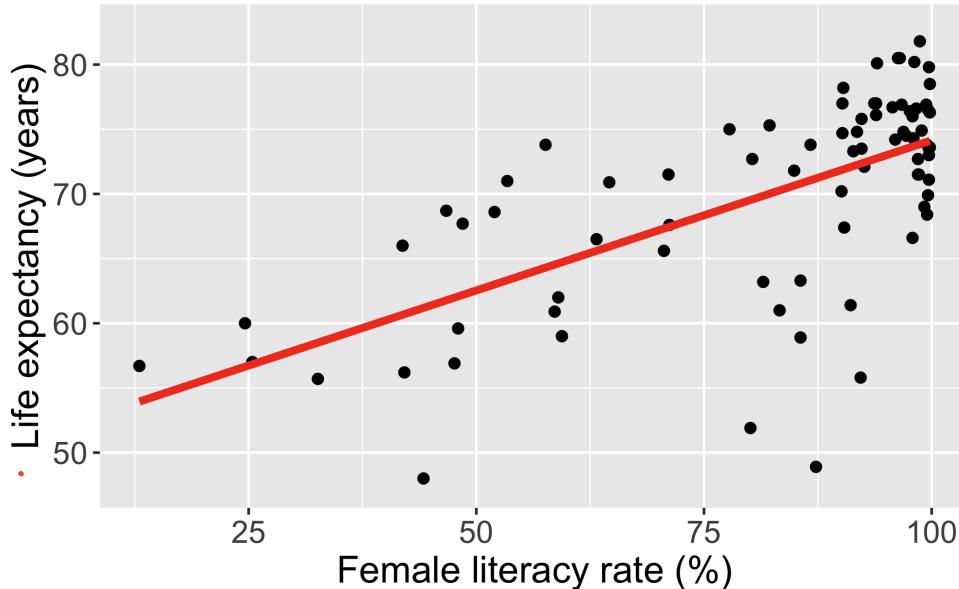
- Our physical classroom space will be changing...
 - It's a little confusing - our time will be split between three classrooms in the RLSB
 - 2 are right next to each other
 - To start, our classes for next week are in:
 - on Monday, 1/22: RLSB 3A003B
 - on Wednesday, 1/22: RLSB 3A003A
 - HW 1 IS NOT DUE THIS WEEK!!! This is my mistake!!
 - Homework 1 is due 1/25!!
 - The finalized HW1 is finally up! Thank you for your patience!
 - Muddiest points for Week 1 are added ✓
 - Office hours starting this week!
 - First one is today at 4:30 with Antara
 - If you are in 612, the reading assignments are posted
 - Wanted to clear something up about attendance
 - If you miss the exit tickets for less than or equal to 5 classes, your grade will not be impacted
 - If you miss more than 5 exit tickets, then your attendance grade will be affected
 - Any questions on the lab? (10 minutes)
- Working on updating
Lab 1 tomorrow

Learning Objectives

1. Identify the aims of your research and see how they align with the intended purpose of simple linear regression
2. Identify the simple linear regression model and define statistics language for key notation
3. Illustrate how ordinary least squares (OLS) finds the best model parameter estimates
4. Solve the optimal coefficient estimates for simple linear regression using OLS
5. Apply OLS in R for simple linear regression of real data

Let's start with an example

Relationship between life expectancy and the female literacy rate in 2011



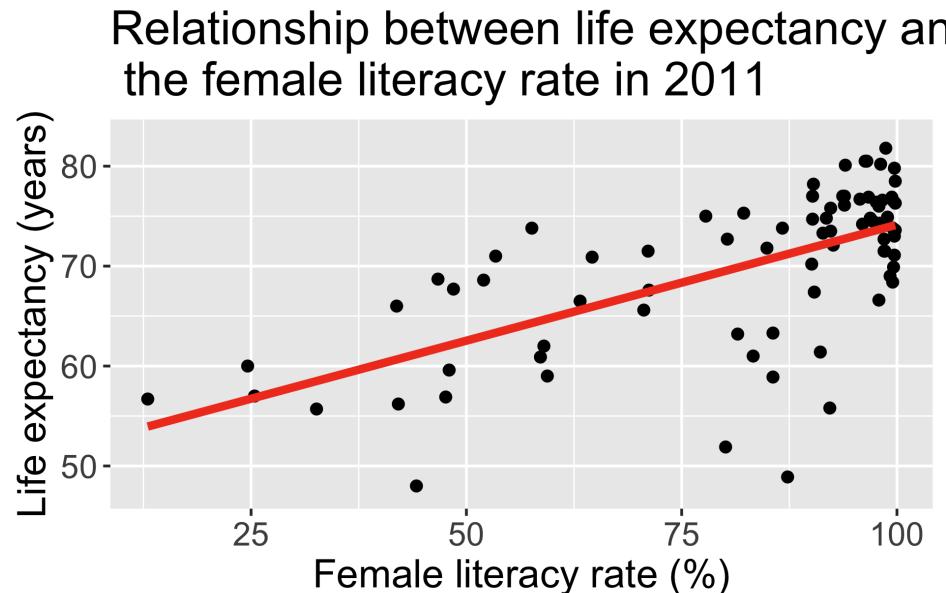
Average life expectancy vs. female literacy rate

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

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

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

read.csv()

```
1 gapm_original <- read_csv(here::here("data", "lifeexp_femlit_water_2011.csv"))
```

- Glimpse of the data

```
1 glimpse(gapm_original)
```

```
Rows: 194
Columns: 5
$ 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, ...
$ water_basic_source_2011    <dbl> 52.6, 88.1, 92.6, 100.0, 40.3, 97.0, 99.5, ...
$ water_2011_quart        <chr> "Q1", "Q2", "Q2", "Q4", "Q1", "Q3", "Q4", "...
```

- 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()      can also use gtsummary pkg
```

```
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   :7                      NA's   :114
```

Remove missing values (1/2)

- Remove rows with missing data for life expectancy and female literacy rate → [Complete case analysis]

```
1 gapm <- gapm_original %>%
2   drop_na(life_expectancy_years_2011, female_literacy_rate_2011)
3
4 glimpse(gapm)
```

drop rows w/ NAs in either variable

Rows: 80

Columns: 5

```
$ country                  <chr> "Afghanistan", "Albania", "Angola", "Antigu...
$ life_expectancy_years_2011 <dbl> 56.7, 76.7, 60.9, 76.9, 76.0, 73.8, 71.0, 7...
$ female_literacy_rate_2011 <dbl> 13.0, 95.7, 58.6, 99.4, 97.9, 99.5, 53.4, 9...
$ water_basic_source_2011    <dbl> 52.6, 88.1, 40.3, 97.0, 99.5, 97.8, 96.7, 9...
$ water_2011_quart          <chr> "Q1", "Q2", "Q1", "Q3", "Q4", "Q3", "Q3", "...
```

- No missing values now for our variables of interest

Remove missing values (2/2)

- And no more missing values when we look only at our two variables of interest

```
1 gapm %>% select(life_expectancy_years_2011, female_literacy_rate_2011) %>%
2   get_summary_stats()

# A tibble: 2 × 13
  variable      n    min    max median     q1     q3    iqr   mad mean     sd     se
  <fct>     <dbl> <dbl>
1 life_expec...    80    48   81.8  72.4  65.9  75.8  9.95  6.30  69.9  7.95  0.889
2 female_lit...    80    13   99.8  91.6  71.0  98.0  27.0  11.4  81.7  22.0  2.45
# i 1 more variable: ci <dbl>
```

Note

- Removing the rows with missing data was not needed to run the regression model. → R will do it for us
- I did this step since later we will be calculating the standard deviations of the explanatory and response variables for *just the values included in the regression model*. It'll be easier to do this if we remove the missing values now.

Poll Everywhere Question 1

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

summary()

Like 4 Dislike 0



glimpse()

Like 2 Dislike 0



Head

Like 1 Dislike 0



Group_by

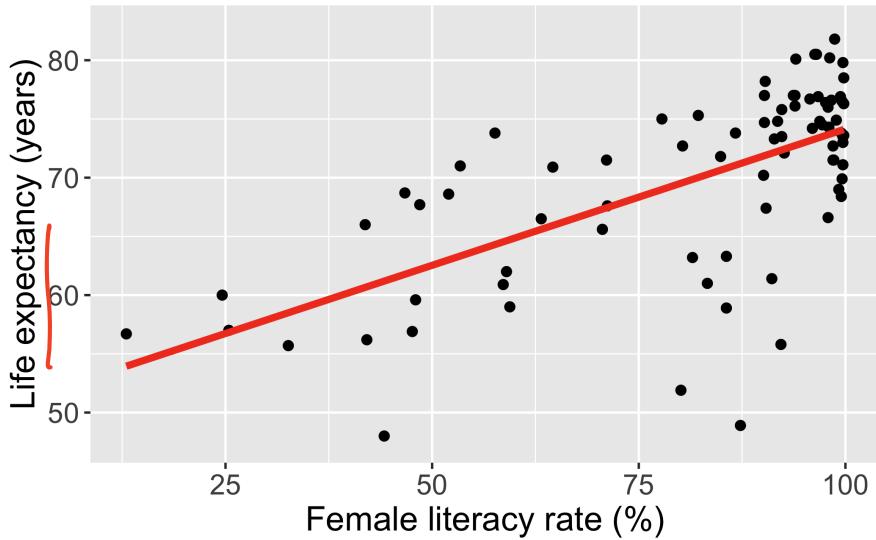


Learning Objectives

1. Identify the aims of your research and see how they align with the intended purpose of simple linear regression
2. Identify the simple linear regression model and define statistics language for key notation
3. Illustrate how ordinary least squares (OLS) finds the best model parameter estimates
4. Solve the optimal coefficient estimates for simple linear regression using OLS
5. Apply OLS in R for simple linear regression of real data

Questions we can ask with a simple linear regression model

Relationship between life expectancy and the female literacy rate in 2011



- 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

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

Association vs. prediction

Association

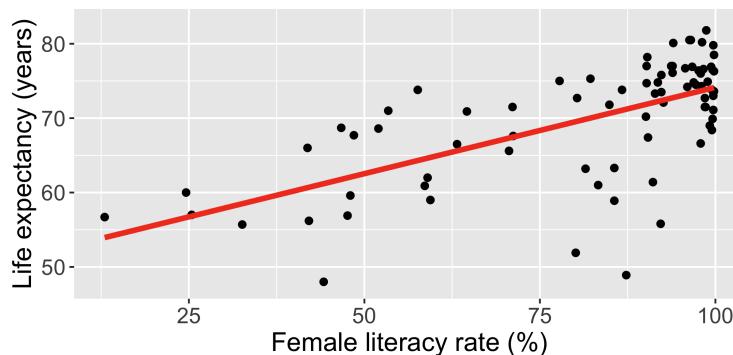
- What is the association between countries' life expectancy and female literacy rate?
- Use the slope of the line or correlation coefficient

Prediction

- What is the expected average life expectancy for a country with a specified female literacy rate?

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

Relationship between life expectancy and the female literacy rate in 2011



Three types of study design (there are more)

Experiment

for ex:
individual

- Observational units are randomly assigned to important predictor levels
 - Random assignment controls for confounding variables (age, gender, race, etc.)
 - “gold standard” for determining causality
 - Observational unit is often at the participant-level

Quasi-experiment

- Participants are assigned to intervention levels without randomization
- Not common study design

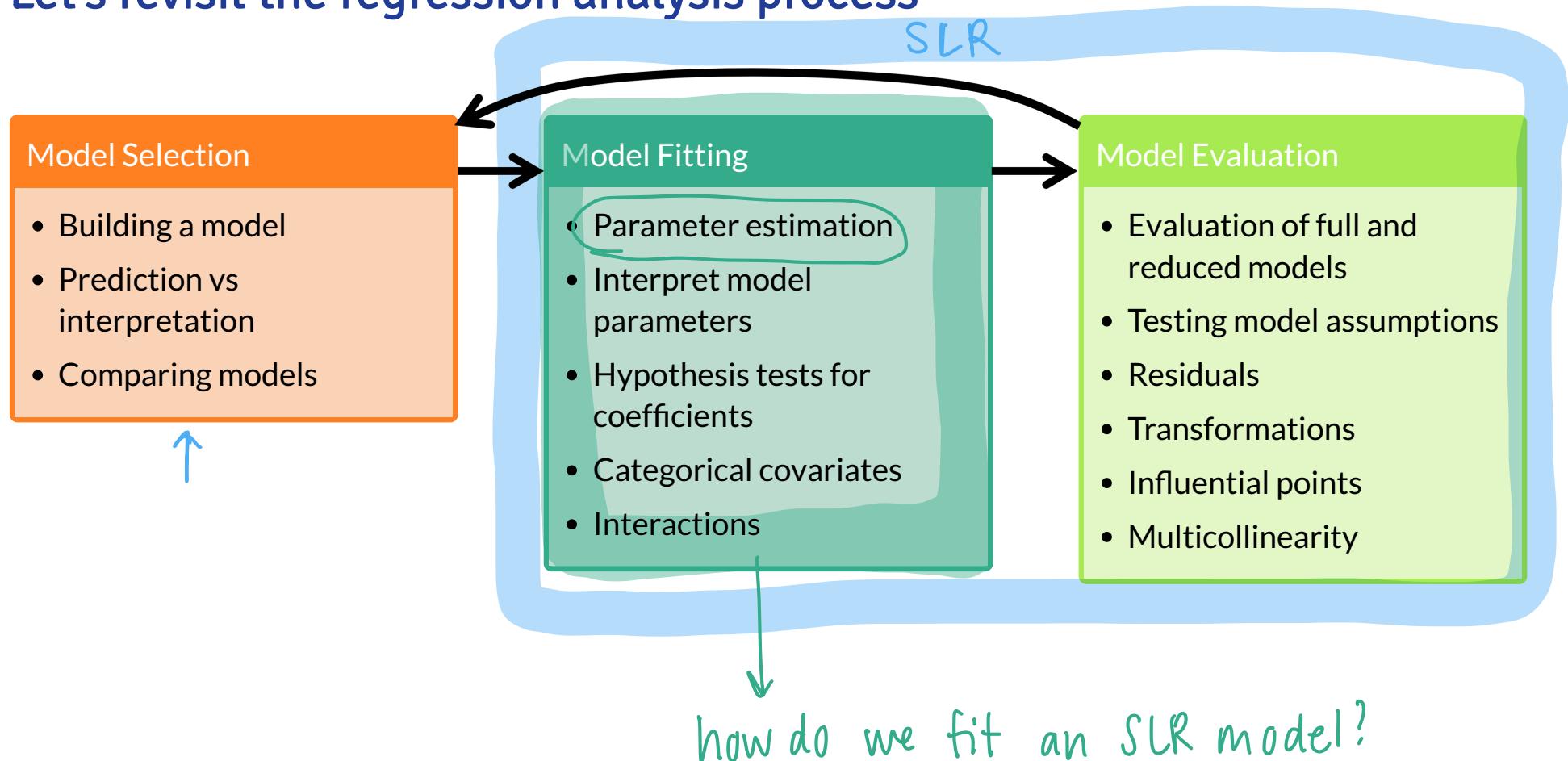


retrospective

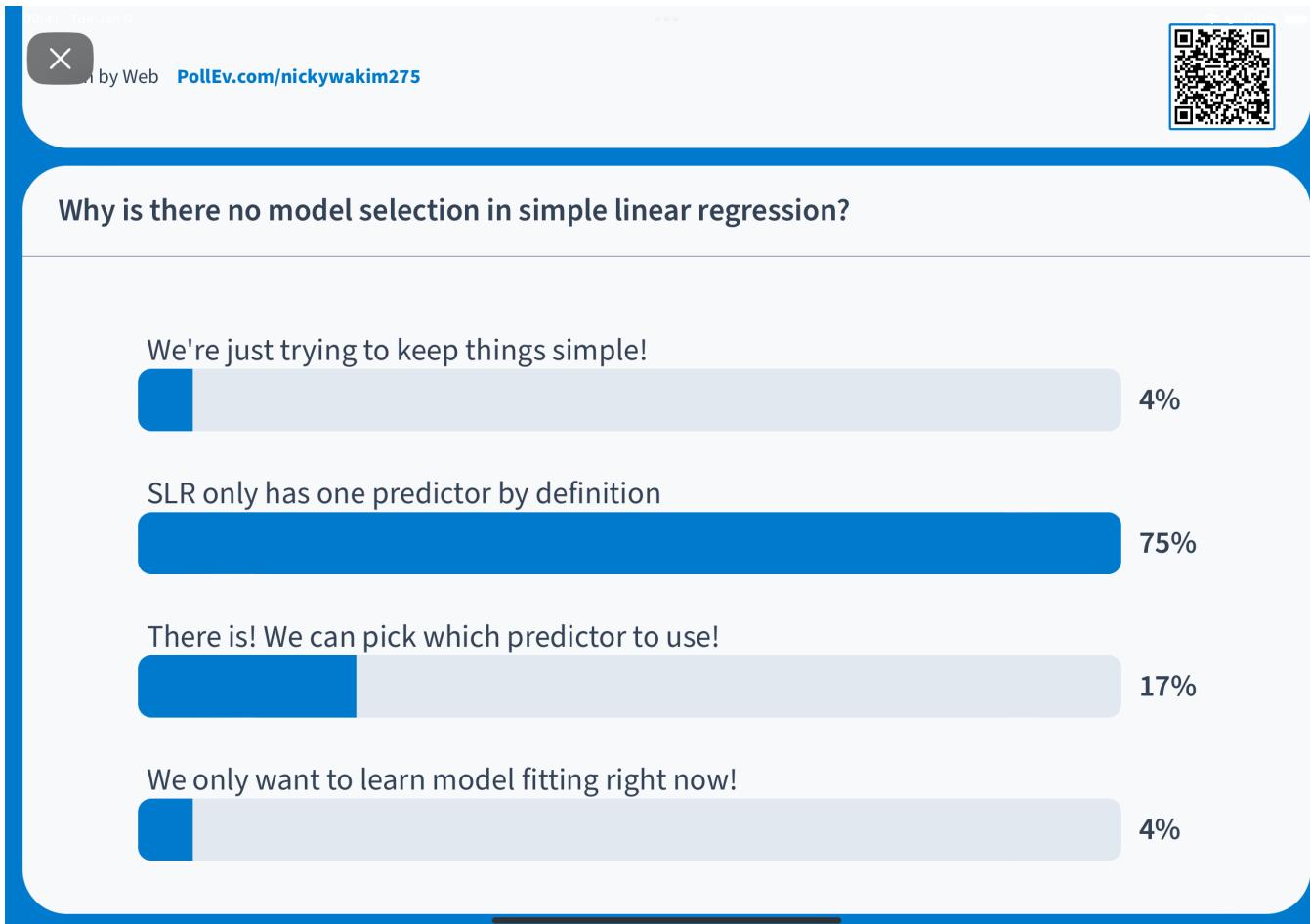
Observational

- No randomization or assignment of intervention conditions
- In general cannot infer causality
 - However, there are causal inference methods...

Let's revisit the regression analysis process



Poll Everywhere Question 2

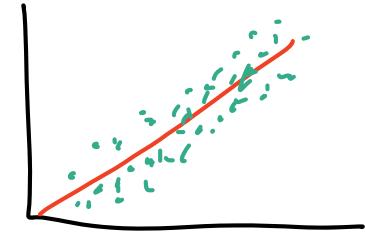


Learning Objectives

1. Identify the aims of your research and see how they align with the intended purpose of simple linear regression
2. Identify the simple linear regression model and define statistics language for key notation
3. Illustrate how ordinary least squares (OLS) finds the best model parameter estimates
4. Solve the optimal coefficient estimates for simple linear regression using OLS
5. Apply OLS in R for simple linear regression of real data

Simple Linear Regression Model

The (population) regression model is denoted by: *true, underlying model*



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

Observable sample data

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

Unobservable population parameters

- β_0 and β_1 are **unknown** population parameters
- ϵ (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 σ^2
 - i.e. $\epsilon \sim N(0, \sigma^2)$

Simple Linear Regression Model (another way to view components)

The (population) regression model is denoted by:

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

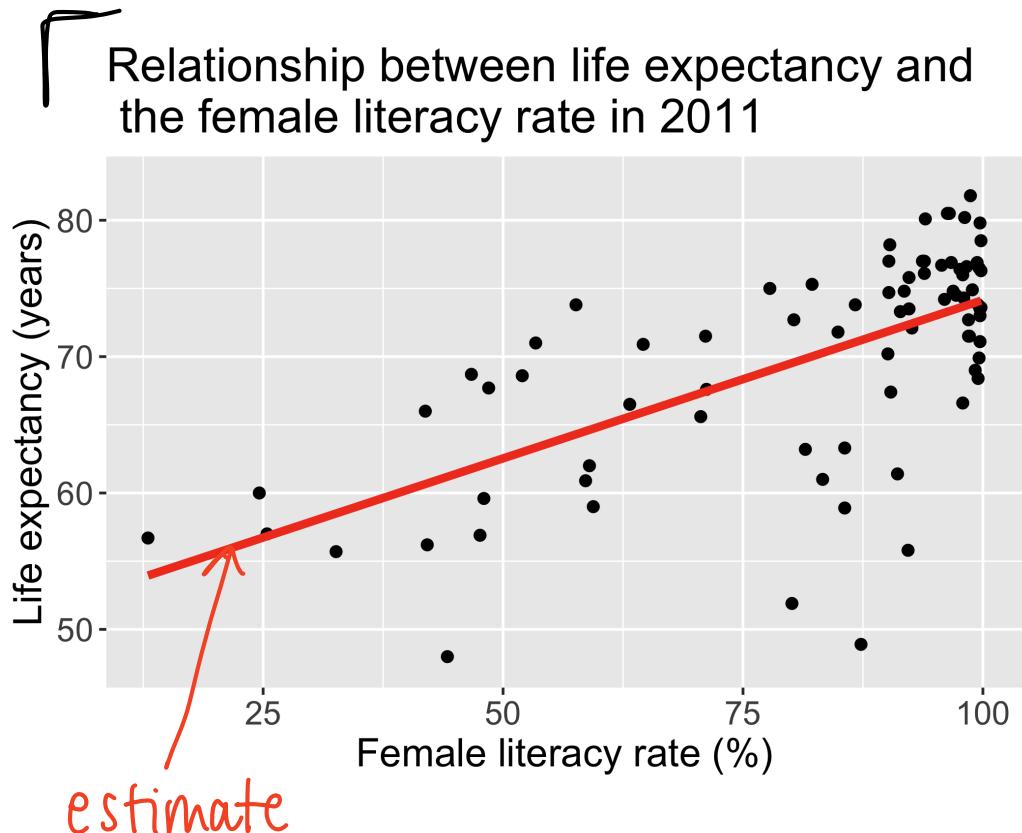
Components

Y	response, outcome, dependent variable
β_0	intercept
β_1	slope
X	predictor, covariate, independent variable
ϵ	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 β_0 and β_1



Poll Everywhere Question 3

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 (%)

β_0 → population intercept

$\hat{\beta}_0$ → sample estimated intercept

β_1 → population slope

$\hat{\beta}_1$: sample estimated slope

SEE MORE

9% 17% 61% 13%

red line is estimated from sample

A scatter plot titled "Relationship between life expectancy and the female literacy rate in 2011". The x-axis is labeled "Female literacy rate (%)" and ranges from 25 to 100. The y-axis is labeled "Life expectancy (years)" and ranges from 50 to 80. Numerous black dots represent individual data points. A single red line, representing the regression line, passes through the general trend of the data points. Handwritten text in orange points to the red line with the statement "red line is estimated from sample". To the right of the plot, there is a poll with four options represented by colored bars:

- Blue bar: β_0 → population intercept (9%)
- Blue bar: $\hat{\beta}_0$ → sample estimated intercept (17%)
- Dark blue bar: β_1 → population slope (61%)
- Light blue bar: $\hat{\beta}_1$: sample estimated slope (13%)

At the bottom left, there is a "SEE MORE" button with a downward arrow.

Okay, so how do we estimate the regression line?

At this point, we are going to move over to an R shiny app that I made.

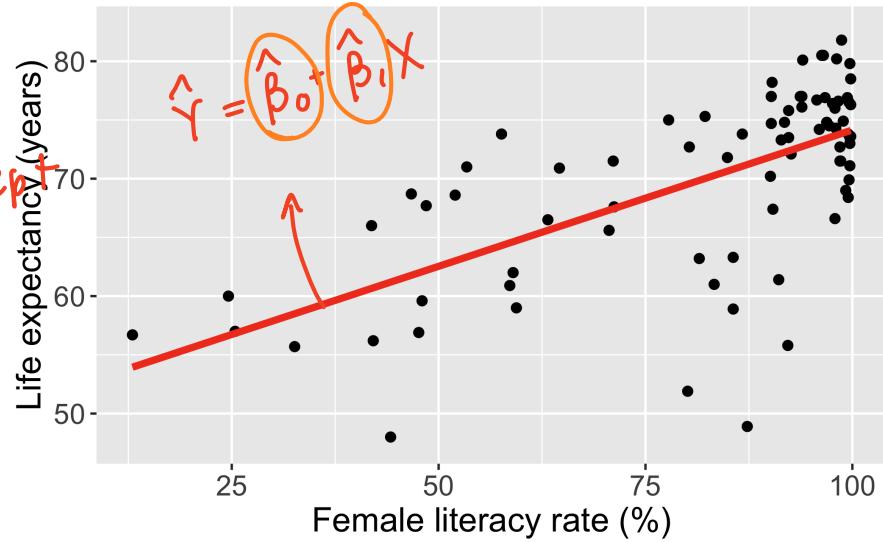
Let's see if we can eyeball the best-fit line!

Regression line = best-fit line

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

- \hat{Y} is the predicted outcome for a specific value of X
- $\hat{\beta}_0$ is the intercept of the best-fit line → estimated intercept
- $\hat{\beta}_1$ 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*

Relationship between life expectancy and the female literacy rate in 2011



Simple Linear Regression Model

Population regression *model*

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

Estimated regression line

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

Components

Y response, outcome, dependent variable



β_0 intercept



β_1 slope



X predictor, covariate, independent variable



ϵ residuals, error term

Components

\widehat{Y}

estimated expected response given predictor X

$\widehat{\beta}_0$

estimated intercept

$\widehat{\beta}_1$

estimated slope

X

predictor, covariate, independent variable

We get it, Nicky! How do we estimate the regression line?

First let's take a break!!

Learning Objectives

1. Identify the aims of your research and see how they align with the intended purpose of simple linear regression
2. Identify the simple linear regression model and define statistics language for key notation
3. Illustrate how ordinary least squares (OLS) finds the best model parameter estimates
4. Solve the optimal coefficient estimates for simple linear regression using OLS
5. Apply OLS in R for simple linear regression of real data

It all starts with a residual...

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

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

$$E(\epsilon) = 0 \text{ (mean is 0)}$$

- 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
- We take the expected value of both sides and get:

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

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

$$\underline{E[Y]} = (\beta_0 + \beta_1 X) + E[\epsilon]$$

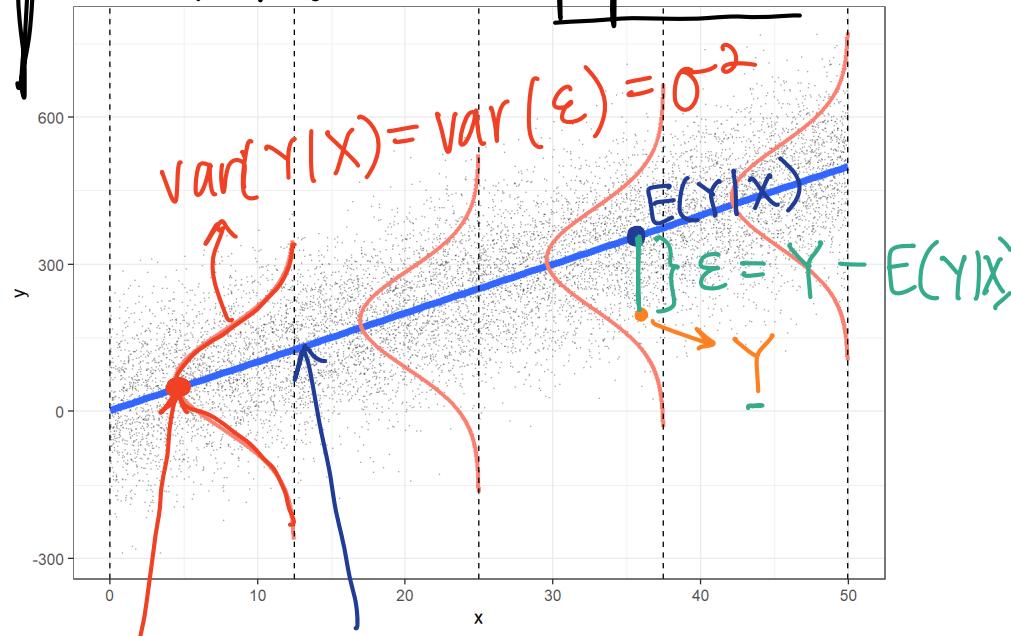
$$E[Y|X] = \beta_0 + \beta_1 X \quad \rightarrow = 0$$

mean/avg/expected Y given X

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

$$\text{var}(\epsilon) = \sigma^2 \text{ (variance is } \sigma^2)$$

If this is our population:



mean is
 $\beta_0 + \beta_1 X$

Variance is σ^2

So now we have two representations of our population model

With observed Y values and residuals:

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

With the population expected value of Y given X :

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

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

$$\begin{aligned} Y &= (\beta_0 + \beta_1 X) + \epsilon \\ Y &= E[Y|X] + \epsilon \end{aligned}$$

$$Y - E[Y|X] = \epsilon$$

$$\epsilon = Y - E[Y|X]$$

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



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 :

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

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 : $Y_i \rightarrow$ outcome for i

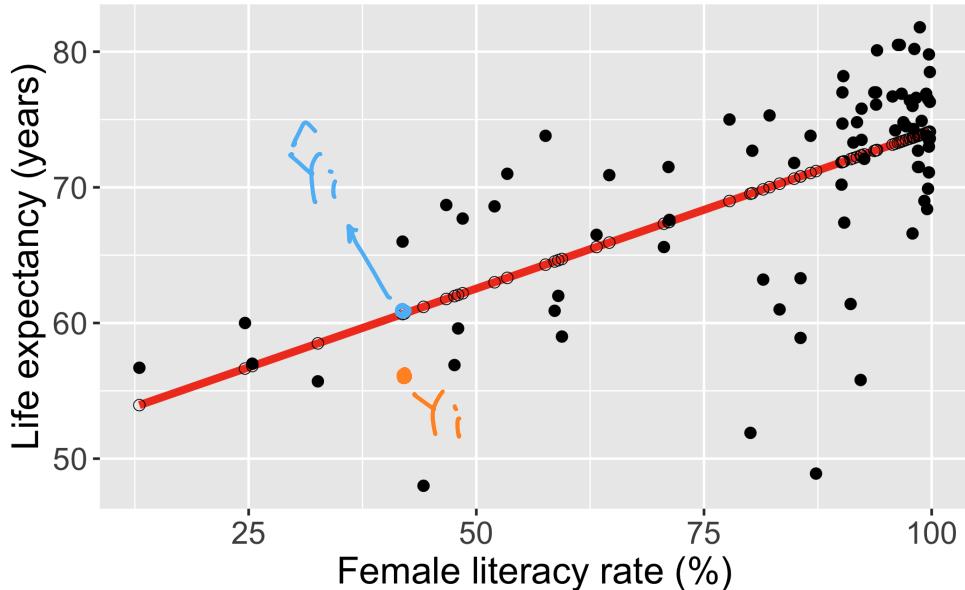
- Fitted value for each individual i : \hat{Y}_i

- 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

"fitted y i value" = \hat{Y}_i

Y & X
↳ whole sample (vector)

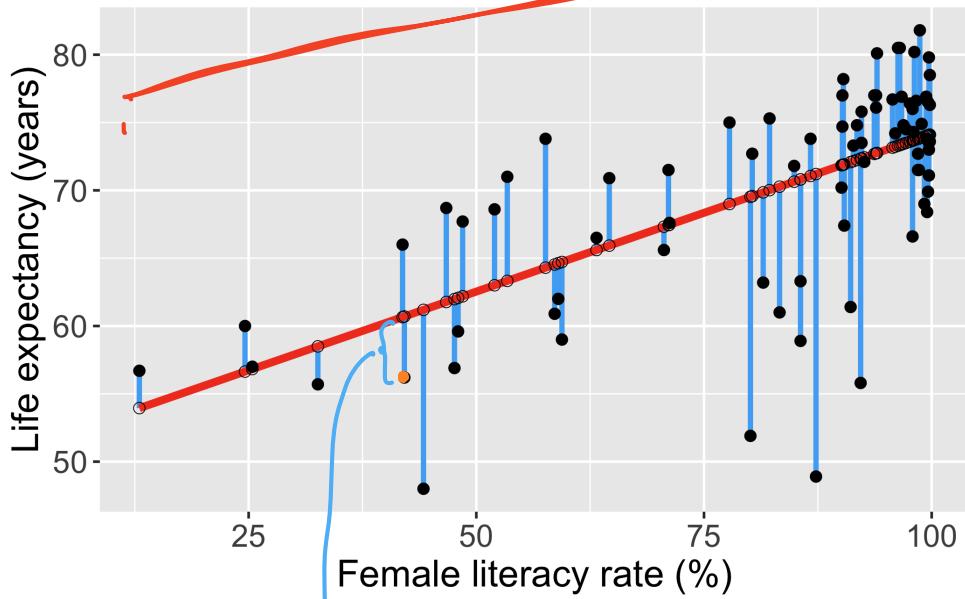
Relationship between life expectancy and the female literacy rate in 2011



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

Relationship between life expectancy and the female literacy rate in 2011



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

$$\hat{\epsilon}_i = Y_i - \hat{E}(Y_i | X_i)$$

$$\hat{\epsilon}_i = Y_i - E(Y_i | X_i)$$

Poll Everywhere Question 4

If our observed Y value fell exactly on the best-fit line, what would the residual be?

0

 1  0



0

 0  0



0

 0  0



0



So what do we do with the residuals?

- We want to minimize the 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

Learning Objectives

1. Identify the aims of your research and see how they align with the intended purpose of simple linear regression
2. Identify the simple linear regression model and define statistics language for key notation
3. Illustrate how ordinary least squares (OLS) finds the best model parameter estimates
4. Solve the optimal coefficient estimates for simple linear regression using OLS
5. Apply OLS in R for simple linear regression of real data

Setting up for ordinary least squares

- Sum of Squared Errors (SSE)

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

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

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

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

Things to use

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

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

Steps to estimate coefficients using OLS

1. Set up SSE (previous slide) ✓

2. Minimize SSE with respect to coefficient estimates

- Need to solve a system of equations

3. Compute derivative of SSE wrt $\hat{\beta}_0$

4. Set derivative of SSE wrt $\hat{\beta}_0 = 0$

5. Compute derivative of SSE wrt $\hat{\beta}_1$

6. Set derivative of SSE wrt $\hat{\beta}_1 = 0$

7. Substitute $\hat{\beta}_1$ back into $\hat{\beta}_0$

2. Minimize SSE with respect to coefficients

- Want to minimize with respect to (wrt) the potential coefficient estimates ($\hat{\beta}_0$ and $\hat{\beta}_1$)
- Take derivative of SSE wrt $\hat{\beta}_0$ and $\hat{\beta}_1$ and set equal to zero to find minimum SSE

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = 0 \text{ and } \frac{\partial SSE}{\partial \hat{\beta}_1} = 0$$

sys of eq

- Solve the above system of equations in steps 3-6

3. Compute derivative of SSE wrt $\hat{\beta}_0$

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

$$\begin{aligned}\frac{\partial SSE}{\partial \hat{\beta}_0} &= \frac{\partial \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\partial \hat{\beta}_0} = \sum_{i=1}^n \frac{\partial (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\partial \hat{\beta}_0} \\ &= \sum_{i=1}^n 2 (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) (-1) = \sum_{i=1}^n -2 (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)\end{aligned}$$

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

Things to use

- Derivative rule: derivative of sum is sum of derivative
- Derivative rule: chain rule

4. Set derivative of SSE wrt $\hat{\beta}_0 = 0$

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = 0$$

$$-2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\sum_{i=1}^n Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n X_i = 0$$

$$\frac{1}{n} \sum_{i=1}^n Y_i - \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n X_i = 0$$

$$\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Things to use

- $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$
- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

want to
separate out
 $\hat{\beta}_0$

5. Compute derivative of SSE wrt $\hat{\beta}_1$

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

$$\begin{aligned}\frac{\partial SSE}{\partial \hat{\beta}_1} &= \frac{\partial \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\partial \hat{\beta}_1} = \sum_{i=1}^n \frac{\partial (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\partial \hat{\beta}_1} \\ &= \sum_{i=1}^n 2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-X_i) = \sum_{i=1}^n -2X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\ &= -2 \underbrace{\sum_{i=1}^n X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)}_{\text{Set eq to 0}}\end{aligned}$$

Things to use

- Derivative rule: derivative of sum is sum of derivative
- Derivative rule: chain rule

6. Set derivative of SSE wrt $\hat{\beta}_1 = 0$

$$\frac{\partial SSE}{\partial \hat{\beta}_1} = 0$$

$$\sum_{i=1}^n \left(X_i Y_i - \hat{\beta}_0 X_i - \hat{\beta}_1 X_i^2 \right) = 0$$

separate out $\hat{\beta}_1$

$$\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \hat{\beta}_0 - \sum_{i=1}^n X_i^2 \hat{\beta}_1 = 0$$

$$\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \left(\bar{Y} - \hat{\beta}_1 \bar{X} \right) - \sum_{i=1}^n X_i^2 \hat{\beta}_1 = 0$$

$$\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \bar{Y} + \sum_{i=1}^n \hat{\beta}_1 X_i \bar{X} - \sum_{i=1}^n X_i^2 \hat{\beta}_1 = 0$$

$$\sum_{i=1}^n X_i (Y_i - \bar{Y}) + \sum_{i=1}^n (\hat{\beta}_1 X_i \bar{X} - X_i^2 \hat{\beta}_1) = 0$$

$$\sum_{i=1}^n X_i (Y_i - \bar{Y}) + \hat{\beta}_1 \sum_{i=1}^n X_i (\bar{X} - X_i) = 0$$

Things to use

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$
- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i (Y_i - \bar{Y})}{\sum_{i=1}^n X_i (X_i - \bar{X})}$$

7. Substitute $\hat{\beta}_1$ back into $\hat{\beta}_0$

Final coefficient estimates for SLR

Coefficient estimate for $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i(Y_i - \bar{Y})}{\sum_{i=1}^n X_i(X_i - \bar{X})}$$

Coefficient estimate for $\hat{\beta}_0$

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_0 &= \bar{Y} - \frac{\sum_{i=1}^n X_i(Y_i - \bar{Y})}{\sum_{i=1}^n X_i(X_i - \bar{X})} \bar{X}\end{aligned}$$

Poll Everywhere Question 5

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

0

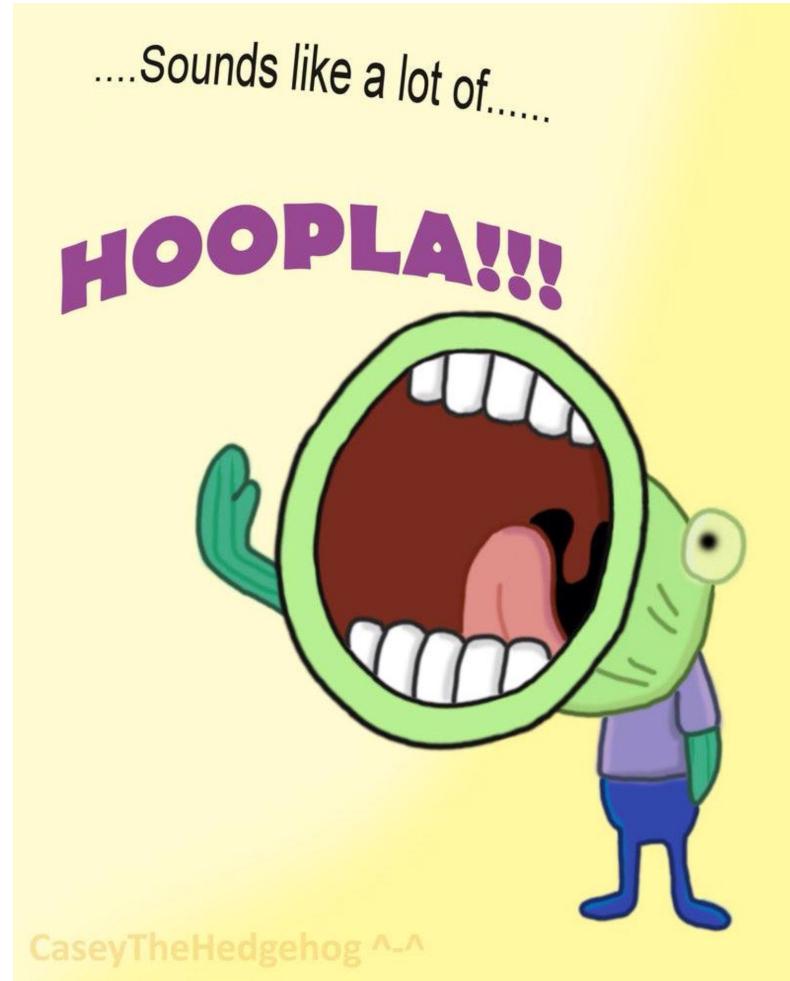
They are the coefficient estimates that minimize every residual value

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

They are the coefficient estimates that perfectly fit our data

They are the coefficient estimates that minimize the sum of the squared residuals

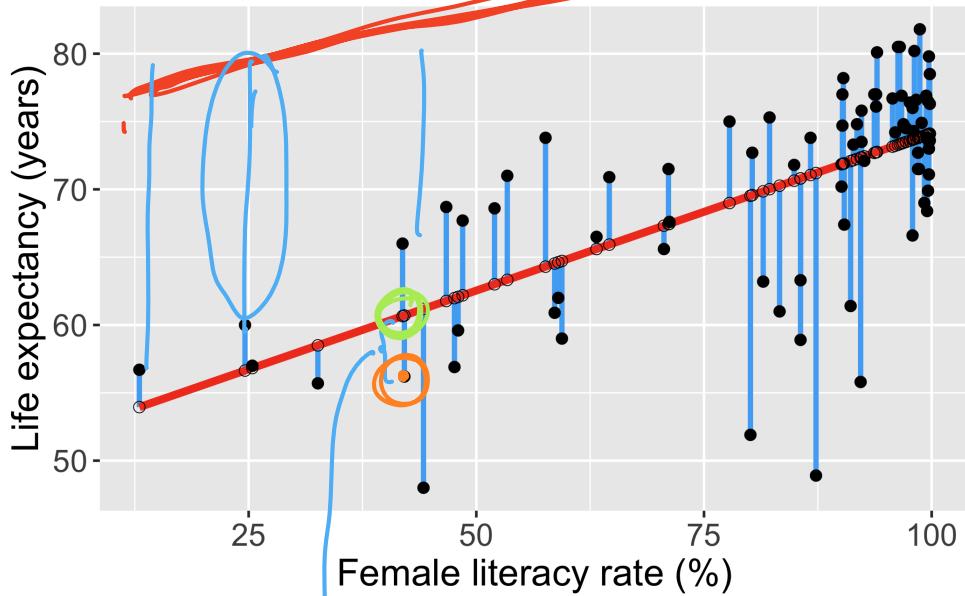
Do I need to do all that work every time??



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

Relationship between life expectancy and the female literacy rate in 2011



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

$$\hat{\epsilon}_i = Y_i - \hat{E}(Y_i | X_i)$$

$$\hat{\epsilon}_i = Y_i - E(Y_i | X_i)$$

Setting up for ordinary least squares

- Sum of Squared Errors (SSE)

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

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

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

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

Things to use

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

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

Regression in R: `lm()`

- Let's discuss the syntax of this function

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

$\text{lm}(\text{Y} \sim \text{X}, \text{data} = \text{dataset name})$

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

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

dictates that our outcome is continuous & normal (ish)

$\hat{\sigma}^2$: estimate of the variance of the ~~target~~ residuals

$\text{glm}()$
generalized

lm assume have $\varepsilon \sim N(0, \sigma^2)$
 glm will ask what distribution you are working w/

pop

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

pop parameters

$$\text{model } Y = \beta_0 + \beta_1 X + \varepsilon$$

line

$$E(Y|X) = \beta_0 + \beta_1 X$$

model = line + error

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

sample / fitted

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

↑ estimates of pop param

model

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

line

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

$$\hat{Y}_i = Y_i - \hat{\varepsilon}_i$$

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

$$\hat{E}(Y|X)$$

$$E(Y|X)$$

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

```
1 model1 <- lm(life_expectancy_years_2011 ~ female_literacy_rate_2011,  
2                           data = gapm)  
→ 3 summary(model1)
```

Call:

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

Residuals:

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

→ median of $\hat{\epsilon}$ ($\hat{\epsilon}$ is a vector of all $\hat{\epsilon}_i$)

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

$$\hat{\sigma}^2 = 6.142$$

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

```
1 tidy(model1) %>%
2   gt() %>%
3   tab_options(table.font.size = 45)
```

term	estimate	std.error	statistic	p.value
→ (Intercept)	50.9178981	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):

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

\widehat{Y} say as "expected life expectancy"

How do we interpret the coefficients?

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

- 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, the expected/average life expectancy increases, on average, 0.232 years.

Next time

- Inference of our estimated coefficients
- Inference of estimated expected Y given X
- Prediction
- Hypothesis testing!

