

# Lesson 5: SLR-ish: Categorical Covariates

Meike Niederhausen and Nicky Wakim

2025-01-22

# Learning Objectives

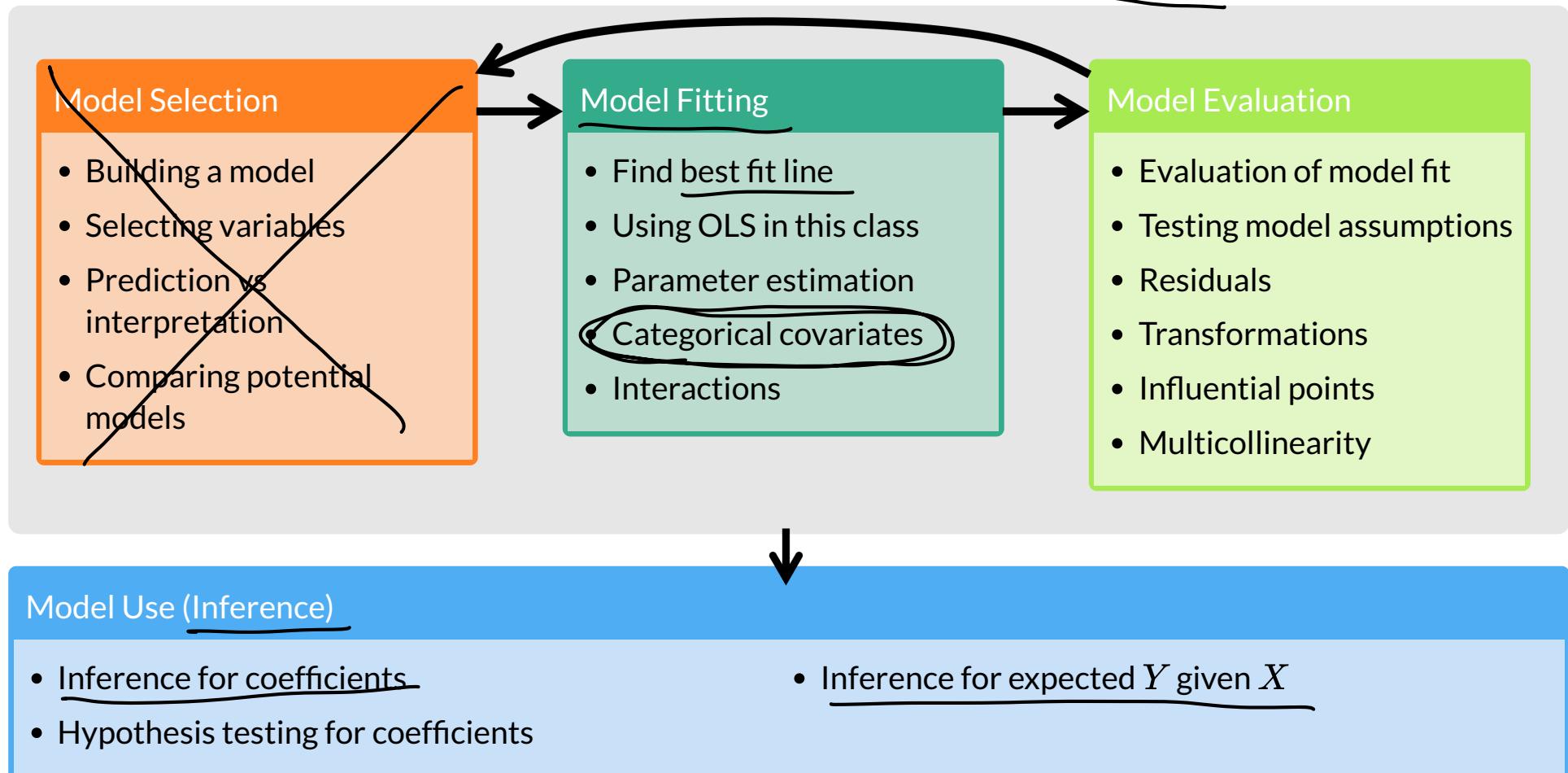
1. Understand why we need a new way to code categorical variables compared to continuous variables
2. Write the regression equation for a categorical variable using reference cell coding
3. Calculate and interpret coefficients for reference cell coding
4. Change the reference level in a categorical variable for reference cell coding
5. Create new variables and interpret coefficient for ordinal / scoring coding

# Why “SLR-*ish*”?

- The **strict** definition of simple linear regression: only two variables that are BOTH continuous
- Common (but kinda wrong) use of simple linear regression: only two variables with outcome continuous and predictor not specified
- I'm including multi-level categorical covariates in SLR mostly because it's easier to learn now!

# Let's map that to our regression analysis process

SLR



# Learning Objectives

1. Understand why we need a new way to code categorical variables compared to continuous variables
2. Write the regression equation for a categorical variable using reference cell coding
3. Calculate and interpret coefficients for reference cell coding
4. Change the reference level in a categorical variable for reference cell coding
5. Create new variables and interpret coefficient for ordinal / scoring coding

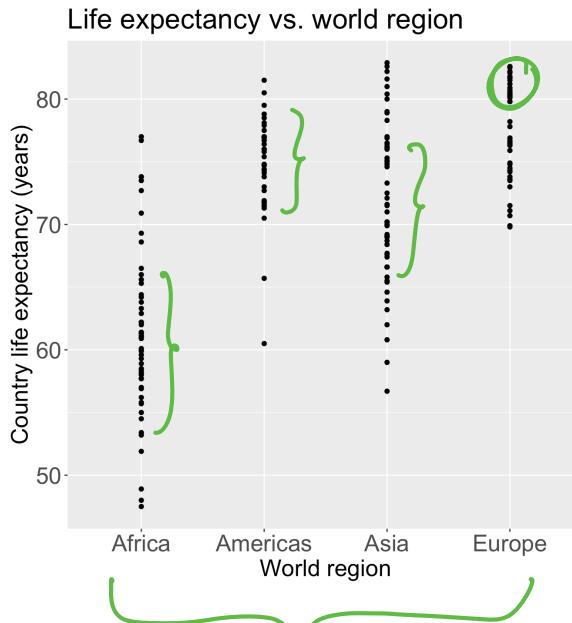
# Still looking at Gapminder Life Expectancy data

- We will look at life expectancy vs. these world regions
- Gapminder uses four world regions
  - Africa
  - The Americas
  - Asia
  - Europe
- World region is a multi-level categorical covariate: it has four regions  
*more than 2*
- Note: I am calling the expected life expectancy  $\widehat{LE}$ 
  - Previously, I have written life expectancy

# Linear regression with a categorical covariate (1/2)

Bad option for visualization:

- Code



Good option for visualization:

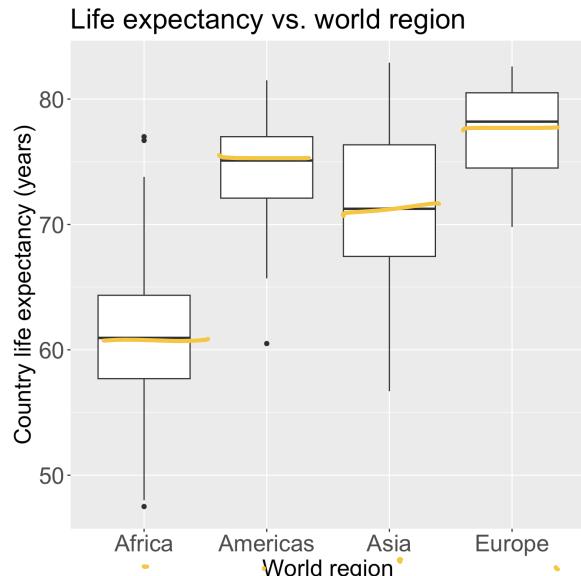
- Code



- Used `geom_jitter()`

Good option for visualization:

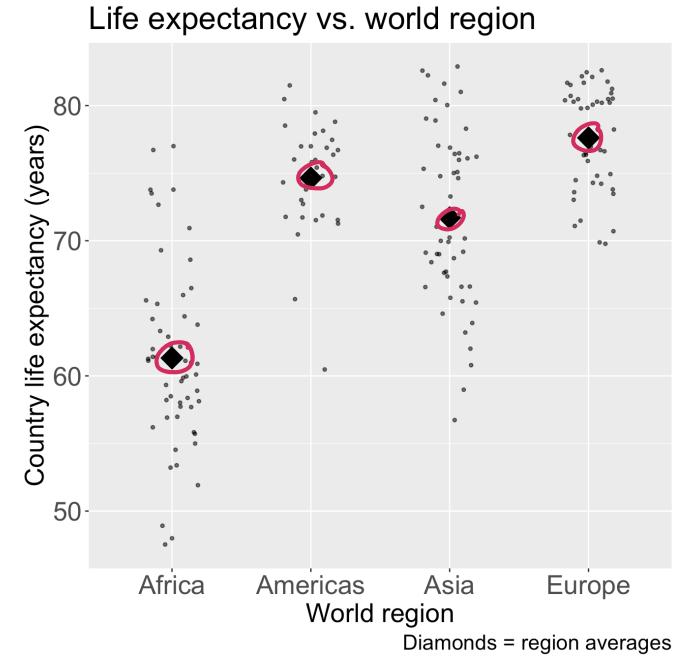
- Code



# Linear regression with a categorical covariate (2/2)

- When using a categorical covariate/predictor (that is not ordered),
  - We do NOT, technically, find a best-fit line
- Instead we model the means of the outcome
  - For the different levels of the categorical variable

- {
- In 511, we used Kruskal-Wallis test and our ANOVA table to test if groups means were statistically different from one another
  - We can do this **using linear models** AND we can include other variables in the model



# There are different ways to code categorical variables

- ★ • Reference cell coding (sometimes called dummy coding)
  - Compares each level of a variable to the omitted (reference) level
- Effect coding (sometimes called sum coding or deviation coding)
  - Compares deviations from the grand mean
  - Not covered in our class
- Ordinal encoding (sometimes called scoring)
  - Categories have a natural, even spaced ordering

If you want to learn more about these and other coding schemes:

- Coding Systems for Categorical Variables in Regression Analysis
- Categorical Data Encoding Techniques
- Coding Schemes for Categorical Variables

# Building the regression equation: problem with a single coefficient

Previously: simple linear regression

- Outcome  $Y$  = numerical variable
- Predictor  $X$  = numerical variable

The regression (best-fit) line is:

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

New: what if the explanatory variable is categorical?

Naively, we could write:  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$

Or, with our variables:

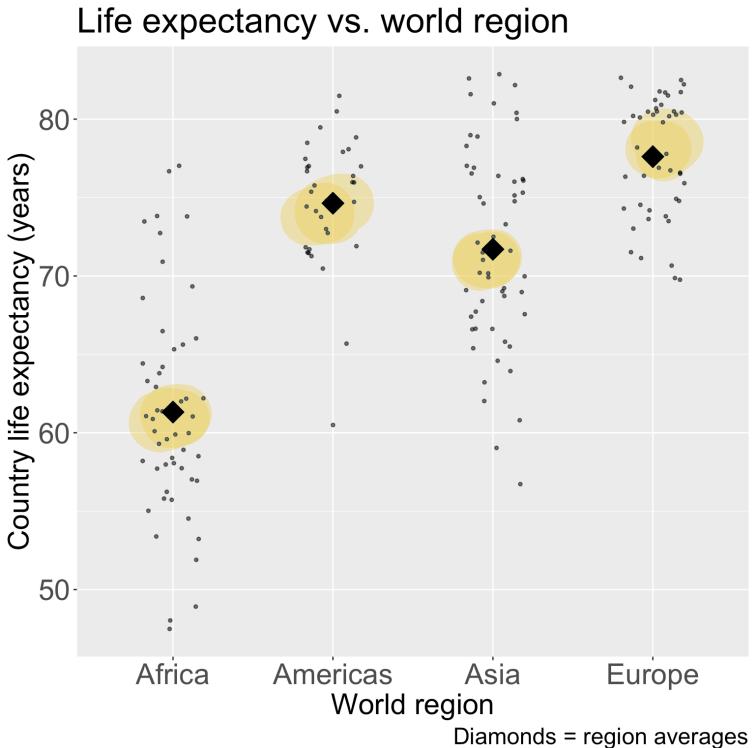
$$X \rightarrow [\widehat{LE} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{WR}]$$

- But what does **WR** (world regions) mean in this equation?
  - What values can it take? How do we represent each region?

- 
- Note: the above is **WRONG**

# Building the regression equation: how do we map categories to means?

- If we only have world region in our model and want to map it to an expected life expectancy...
- We want to create a function that can map each region to life expectancy
  - If in Africa:  $\widehat{LE} = 61.32$  years
  - If in the Americas:  $\widehat{LE} = 74.64$  years
  - If in Asia:  $\widehat{LE} = 71.70$  years
  - If in Europe:  $\widehat{LE} = 77.61$  years
- Can we make one equation for  $\widehat{LE}$  by putting the “if” statements within the equation?



# Building the regression equation: Indicator functions

- In order to represent each region in the equation, we need to introduce a new function:
  - Indicator function:

if  $X = x$

$$I(X = x) \text{ or } I(x) = \begin{cases} 1, & \text{if } X = x \\ 0, & \text{else } X \neq x \end{cases}$$

- This basically a binary yes/no if  $X$  is a specific value  $x$
- For example, if we want to identify a country ~~in~~ in the Americas region, we can make:

if  $I(WR = \text{Americas})$  or  $I(\text{Americas}) = \begin{cases} 1, & \text{if } WR = \text{Americas} \\ 0, & \text{else} \end{cases}$

Country	City	Region
1		Americas
2		Asia
3		Africa

	$I(WR = \text{Americas})$	$I(\text{Asia})$
	1	0
	0	1
	0	0

# Poll Everywhere Question 1

13:45 Wed Jan 22    ...    89%

X

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

QR code

To include world regions in our model, how many indicators do we need to make?

1 0%  
2 24%  
3 16%  
4 60%

Powered by  Poll Everywhere

one becomes  
our reference

# Learning Objectives

1. Understand why we need a new way to code categorical variables compared to continuous variables
2. Write the regression equation for a categorical variable using reference cell coding
3. Calculate and interpret coefficients for reference cell coding
4. Change the reference level in a categorical variable for reference cell coding
5. Create new variables and interpret coefficient for ordinal / scoring coding

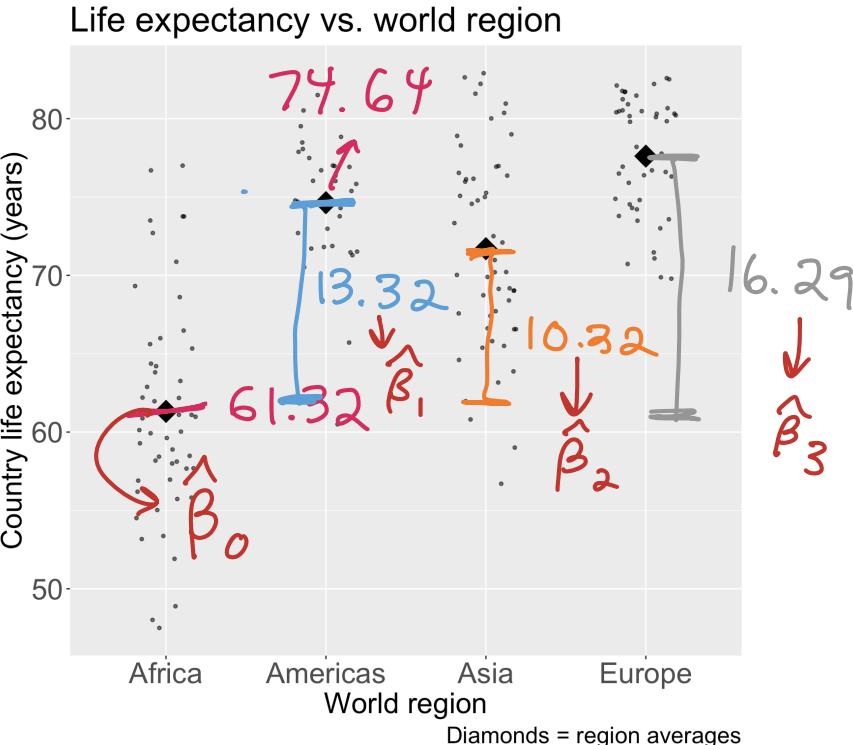
# Building the regression equation: Indicators in our equation

$$\widehat{LE} = 61.32 \cdot I(\text{Africa}) + 74.64 \cdot I(\text{Americas}) + 71.7 \cdot I(\text{Asia}) + 77.61 \cdot I(\text{Europe})$$

- However, a linear regression equation still requires an intercept!
- So one of our regions need to become our “reference” group
- We'll use Africa as our reference
- That means we need to adjust all the numbers

$$\widehat{LE} = 61.32 + 13.32 \cdot I(\text{Americas}) + 10.38 \cdot I(\text{Asia}) + 16.29 \cdot I(\text{Europe})$$

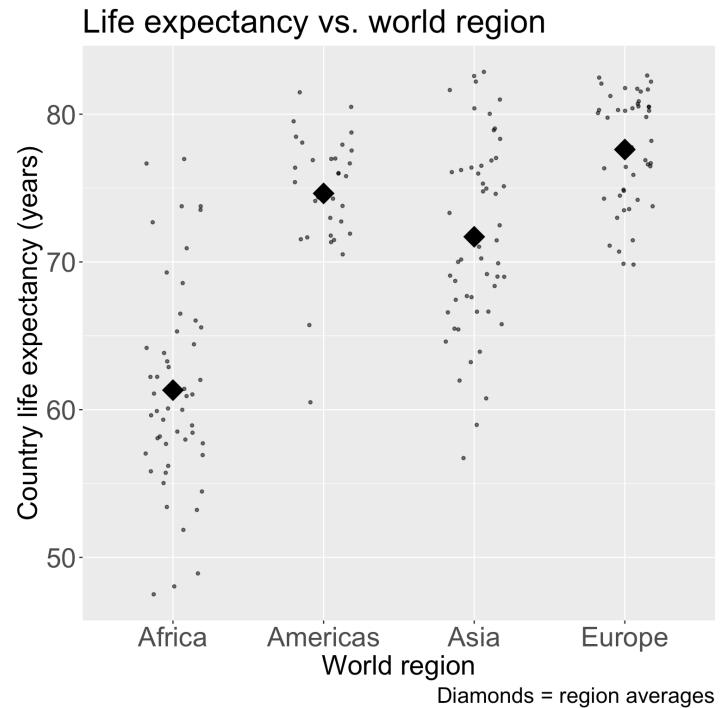
$$\widehat{LE} = \hat{\beta}_0 + \hat{\beta}_1 \cdot I(\text{Americas}) + \hat{\beta}_2 \cdot I(\text{Asia}) + \hat{\beta}_3 \cdot I(\text{Europe})$$



# Viewing the regression equation another way

$$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot I(\text{Americas}) + \widehat{\beta}_2 \cdot I(\text{Asia}) + \widehat{\beta}_3 \cdot I(\text{Europe})$$

<u>World region</u>	<u>Regression equation for WR</u>	<u>Average Life Expectancy for WR</u>
Africa	$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot 0 + \widehat{\beta}_2 \cdot 0 + \widehat{\beta}_3 \cdot 0$	$\widehat{LE} = \widehat{\beta}_0$
Americas	$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot 1 + \widehat{\beta}_2 \cdot 0 + \widehat{\beta}_3 \cdot 0$	$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1$
Asia	$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot 0 + \widehat{\beta}_2 \cdot 1 + \widehat{\beta}_3 \cdot 0$	$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_2$
Europe	$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot 0 + \widehat{\beta}_2 \cdot 0 + \widehat{\beta}_3 \cdot 1$	$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_3$



# Learning Objectives

1. Understand why we need a new way to code categorical variables compared to continuous variables
2. Write the regression equation for a categorical variable using reference cell coding
3. Calculate and interpret coefficients for reference cell coding
4. Change the reference level in a categorical variable for reference cell coding
5. Create new variables and interpret coefficient for ordinal / scoring coding

# Interpretation of regression equation coefficients

- Remember: expected, mean, and average are interchangeable

Coefficient	Interpretation
$\hat{\beta}_0$	Expected/mean/average life expectancy of Africa <i>OR OR</i>
$\hat{\beta}_1$	Difference in mean life expectancy of the Americas and Africa -OR- Mean difference in life expectancy of the Americas and Africa
$\hat{\beta}_2$	Difference in mean life expectancy between Asia and Africa -OR- Mean difference in life expectancy between Asia and Africa
$\hat{\beta}_3$	Difference in mean life expectancy between Europe and Africa -OR- Mean difference in life expectancy between Europe and Africa

## Poll Everywhere Question 2

13:54 Wed Jan 22

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

Would the coefficients for the mean difference in life expectancy between the Americas and Africa change if we included female literacy rate in the model?

Yes! 43%

No! 13%

We would need to fit the model to find out 43%

Powered by Poll Everywhere

$$\hat{LE} = \hat{\beta}_0 + \hat{\beta}_1 FLR$$

$$\hookrightarrow FLR = 0$$

$$\hat{LE} = \hat{\beta}_0 + \hat{\beta}_1 I(Am)$$

$$+ \hat{\beta}_2 I(As) +$$

$$+ \hat{\beta}_3 I(E)$$

$\hat{LE}$  country in Africa

# Regression table with `lm()` function

```
1 model1 <- lm(LifeExpectancyYrs ~ four_regions, data = gapm2)
2 tidy(model1, conf.int=T) %>% gt() %>% tab_options(table.font.size = 38) %>%
3   fmt_number(decimals = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	61.32	0.76	80.26	0.00	59.81	62.83
four_regionsAmericas	13.32	1.23	10.83	0.00	10.89	15.74
four_regionsAsia	10.38	1.08	9.61	0.00	8.25	12.51
four_regionsEurope	16.29	1.13	14.37	0.00	14.05	18.52

$$\widehat{LE} = 61.32 + 13.32 \cdot I(\text{Americas}) + 10.38 \cdot I(\text{Asia}) + 16.29 \cdot I(\text{Europe})$$

- Which world region did R choose as the reference level?
- How you would calculate the mean life expectancies of world regions using only the results from the regression table?

# Bringing in the numbers/units/95% CI

Coefficient	Interpretation
$\hat{\beta}_0$	Average life expectancy of countries in Africa is 61.32 years (95% CI: 59.81, 62.83).
$\hat{\beta}_1$	The difference in mean life expectancy between countries in the Americas and Africa is 13.32 (95% CI: 10.89, 15.74).
$\hat{\beta}_2$	The difference in mean life expectancy between countries in the Americas and Africa is 10.38 (95% CI: 8.25, 12.51).
$\hat{\beta}_3$	The difference in mean life expectancy between countries in Europe and Africa is 18.52 (95% CI: 14.05, 18.52).

- Don't forget that we can use the confidence intervals to assess whether the mean difference with Africa is significant or not

# We can also use R to report each region's average life expectancy

Find the 95% CI's for the mean life expectancy for the Americas, Asia, and Europe

- Use the base R `predict()` function (see Lesson 4 for more info)
- Requires specification of a `newdata` "value"

```
1 newdata <- data.frame(four_regions = c("Africa", "Americas", "Asia", "Europe"))  
1 (pred = predict(model1,  
2                  newdata=newdata,  
3                  interval="confidence"))
```

	fit	lwr	upr
1	61.32037	59.81287	62.82787
2	74.63824	72.73841	76.53806
3	71.70185	70.19435	73.20935
4	77.60889	75.95751	79.26027

Africa  
Americas  
Asia  
Europe

1            2            3            4

## Interpretations

- The average life expectancy for countries in the Americas is 74.64 years (95% CI: 72.74, 76.54).
- The average life expectancy for countries in Asia is 71.7 years (95% CI: 70.19, 73.21).
- The average life expectancy for countries in Europe is 77.61 years (95% CI: 75.96, 79.26).

## Another way to look at coefficient values

$$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot I(\text{Americas}) + \widehat{\beta}_2 \cdot I(\text{Asia}) + \widehat{\beta}_3 \cdot I(\text{Europe})$$

► Code

World regions	Average life expectancy	Difference with Africa
Africa	61.3	0.0
Americas	74.6	13.3
Asia	71.7	10.4
Europe	77.6	16.3

$$\widehat{LE} = 61.32 + 13.32 \cdot I(\text{Americas}) + 10.38 \cdot I(\text{Asia}) + 16.29 \cdot I(\text{Europe})$$

# 10 minute break here?

# Learning Objectives

1. Understand why we need a new way to code categorical variables compared to continuous variables
2. Write the regression equation for a categorical variable using reference cell coding
3. Calculate and interpret coefficients for reference cell coding
4. Change the reference level in a categorical variable for reference cell coding
5. Create new variables and interpret coefficient for ordinal / scoring coding

# Reference levels

Why is Africa not one of the variables in the regression equation?

$$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot I(\text{Americas}) + \widehat{\beta}_2 \cdot I(\text{Asia}) + \widehat{\beta}_3 \cdot I(\text{Europe})$$

- Categorical variables have to have at least 2 levels. If they have 2 levels, we call them *binary*
- We choose one level as our reference level to which all other levels of the categorical variable are compared
  - The levels Americas, Asia, Europe are compared to the level Africa
- The **intercept** of the regression equation is the *mean of the outcome restricted to the reference level*
  - Recall that the intercept is the mean life expectancy of Africa, which was our reference level
- If the categorical variable has  $r$  levels, then we need  $r - 1$  variables/coefficients to model it!

## We can change the reference level to Europe (1/2)

- Suppose we want to compare the mean life expectancies of world regions to the Europe level instead of Africa
- Below is the estimated regression equation for when Africa is the reference level

$$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot I(\text{Americas}) + \widehat{\beta}_2 \cdot I(\text{Asia}) + \widehat{\beta}_3 \cdot I(\text{Europe})$$

- Update the variables to make Europe the reference level:

$$\widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot I(\text{Africa}) + \widehat{\beta}_2 \cdot I(\text{Americas}) + \widehat{\beta}_3 \cdot I(\text{Asia})$$

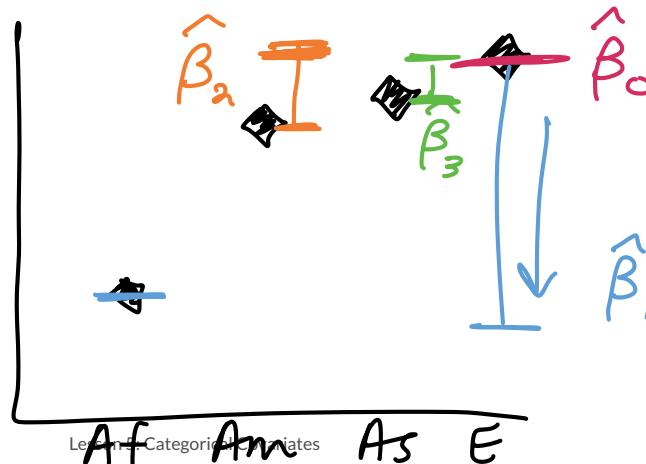
Europe

## We can change the reference level to Europe (2/2)

- Now update the coefficients of the regression equation using the output below.

World regions	Average life expectancy	Difference with Europe
Africa	61.32	-16.29
Americas	74.64	-2.97
Asia	71.70	-5.91
Europe	77.61	0.00

$$\widehat{LE} = 77.61 - 16.29 \cdot I(\text{Africa}) - 2.97 \cdot I(\text{Americas}) - 5.91 \cdot I(\text{Asia})$$



## R: Change reference level to Europe (1/2)

- `four_regions` data type was originally a `character` - check this with `str()`

```
1 str(gapm$four_regions)
```

→ chr [1:195] "asia" "europe" "africa" "europe" "africa" "americas" ...

- In order to change the reference level, we need to convert it to data type `factor`

- I also did this at the beginning to capitalize each region

```
1 gapm_ex = gapm %>%
2   mutate(
3     four_regions = factor(four_regions,
4       levels = c("africa", "americas", "asia", "europe"),
5       labels = c("Africa", "Americas", "Asia", "Europe"))
6   )
7 str(gapm_ex$four_regions)
```

Factor w/ 4 levels "Africa", "Americas", ... : 3 4 1 4 1 2 2 4 3 4 ...

```
1 levels(gapm_ex$four_regions) # order of factor levels
[1] "Africa"    "Americas"   "Asia"      "Europe"
```

## R: Change reference level to Europe (2/2)

- Now change the order of the factor levels
- Code below uses `fct_relevel()` from the `forcats` package that gets loaded as a part of the `tidyverse`
- Any levels not mentioned will be left in their existing order, after the explicitly mentioned levels.*

```
1 gapm2 <- gapm2 %>%
2   mutate(
3     four_regions = fct_relevel(four_regions, "Europe")
4   )
```

E, As, Am, Af

- Check the order:

```
1 levels(gapm2$four_regions)
[1] "Europe"    "Africa"     "Americas"   "Asia"
```

## R: Run model with Europe as the reference level

```
1 levels(gapm2$four_regions)
[1] "Europe"    "Africa"     "Americas"   "Asia"
2
3 model2 <- lm(LifeExpectancyYrs ~ four_regions, data = gapm2)
4 tidy(model2) %>% gt() %>% tab_options(table.font.size = 35) %>%
5   fmt_number(decimals = 2)
```

Europe —

term	estimate	std.error	statistic	p.value
(Intercept)	77.61	0.84	92.72	0.00
four_regionsAfrica	-16.29	1.13	-14.37	0.00
four_regionsAmericas	-2.97	1.28	-2.33	0.02
four_regionsAsia	-5.91	1.13	-5.21	0.00

$$\widehat{LE} = \hat{\beta}_0 + \hat{\beta}_1 \cdot I(\text{Africa}) + \hat{\beta}_2 \cdot I(\text{Americas}) + \hat{\beta}_3 \cdot I(\text{Asia})$$

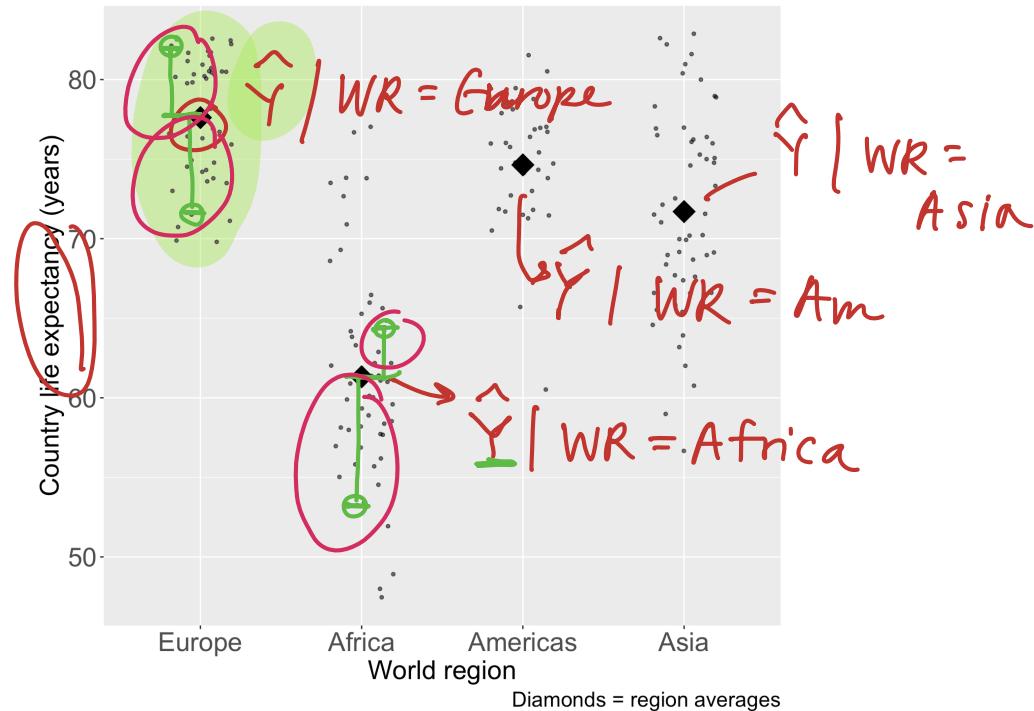
$$\widehat{LE} = 77.61 - 16.29 \cdot I(\text{Africa}) - 2.97 \cdot I(\text{Americas}) - 5.91 \cdot I(\text{Asia})$$

# Fitted values & residuals

Similar to as before:

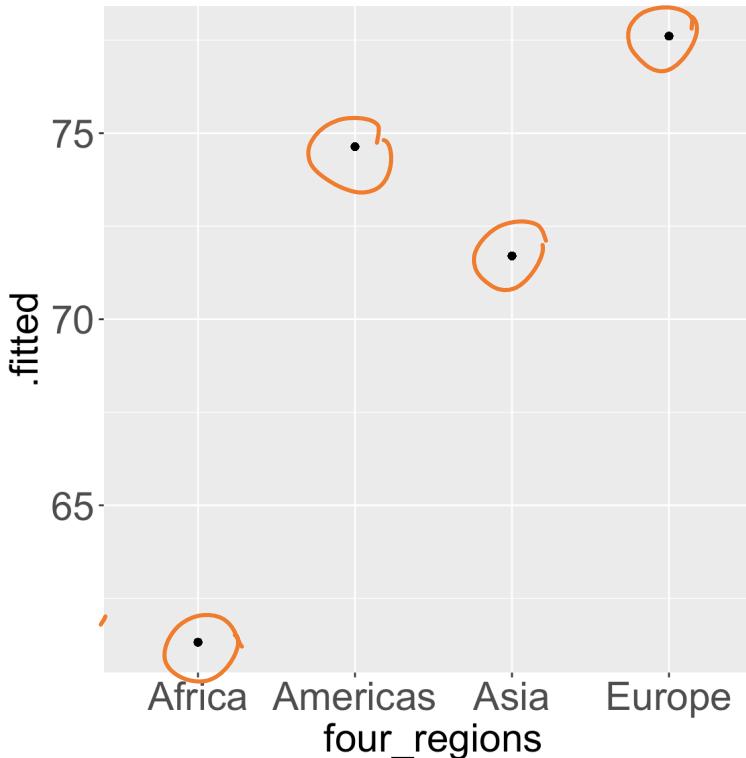
- Observed values  $\hat{Y}$  are the values in the dataset
- Fitted values  $\hat{Y}$  are the values that fall on the best fit line for a specific value of  $x$  are the means of the outcome stratified by the categorical predictor's levels
- Residuals ( $\hat{\epsilon} = Y - \hat{Y}$ ) are the differences between the two

Life expectancy vs. world region



# Fitted values are the same as the means

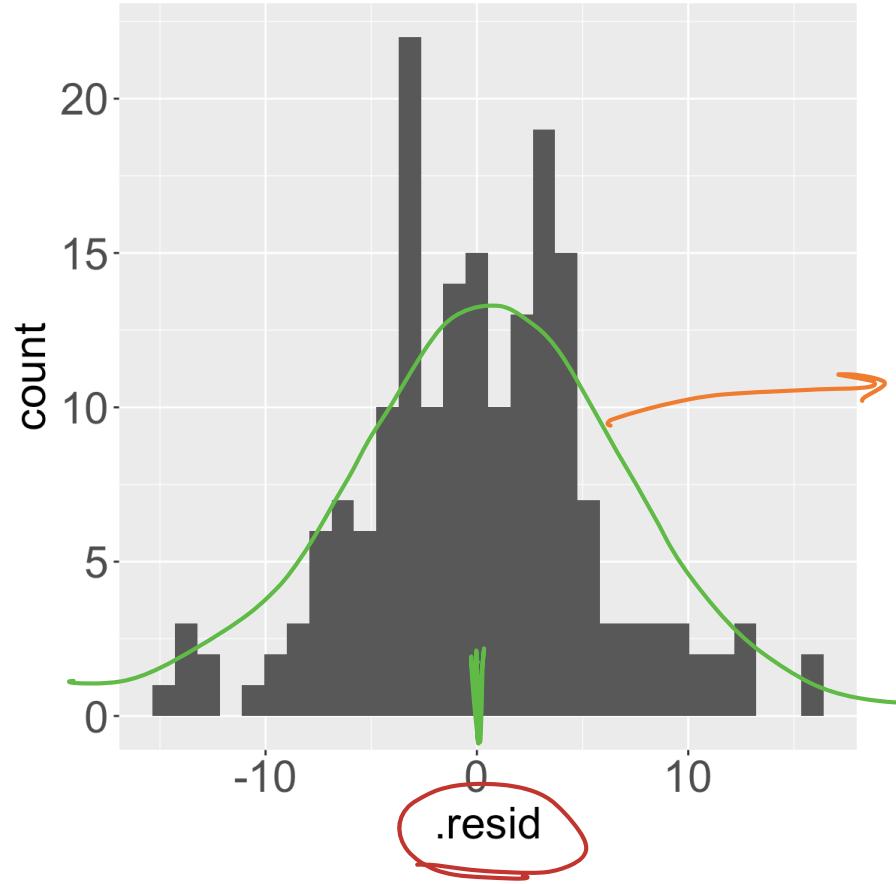
```
1 m1_aug <- augment(model1)
2
3 ggplot(m1_aug, aes(x = four_regions, y = .fitted)) + geom_point() +
4   theme(axis.text = element_text(size = 22), axis.title = element_text(size = 22))
```



## Residual plots (now the spread within each region)

```
1 ggplot(ml_aug, aes(x=.resid)) + geom_histogram() +  
2   theme(axis.text = element_text(size = 22), title = element_text(size = 22))
```

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



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

# Poll Everywhere Question 3

14:37 Wed Jan 22

X Join by Web [Pollev.com/nickywakim27](https://pollev.com/nickywakim27) "polleveverywhere.com" is in full screen. Swipe down to exit.

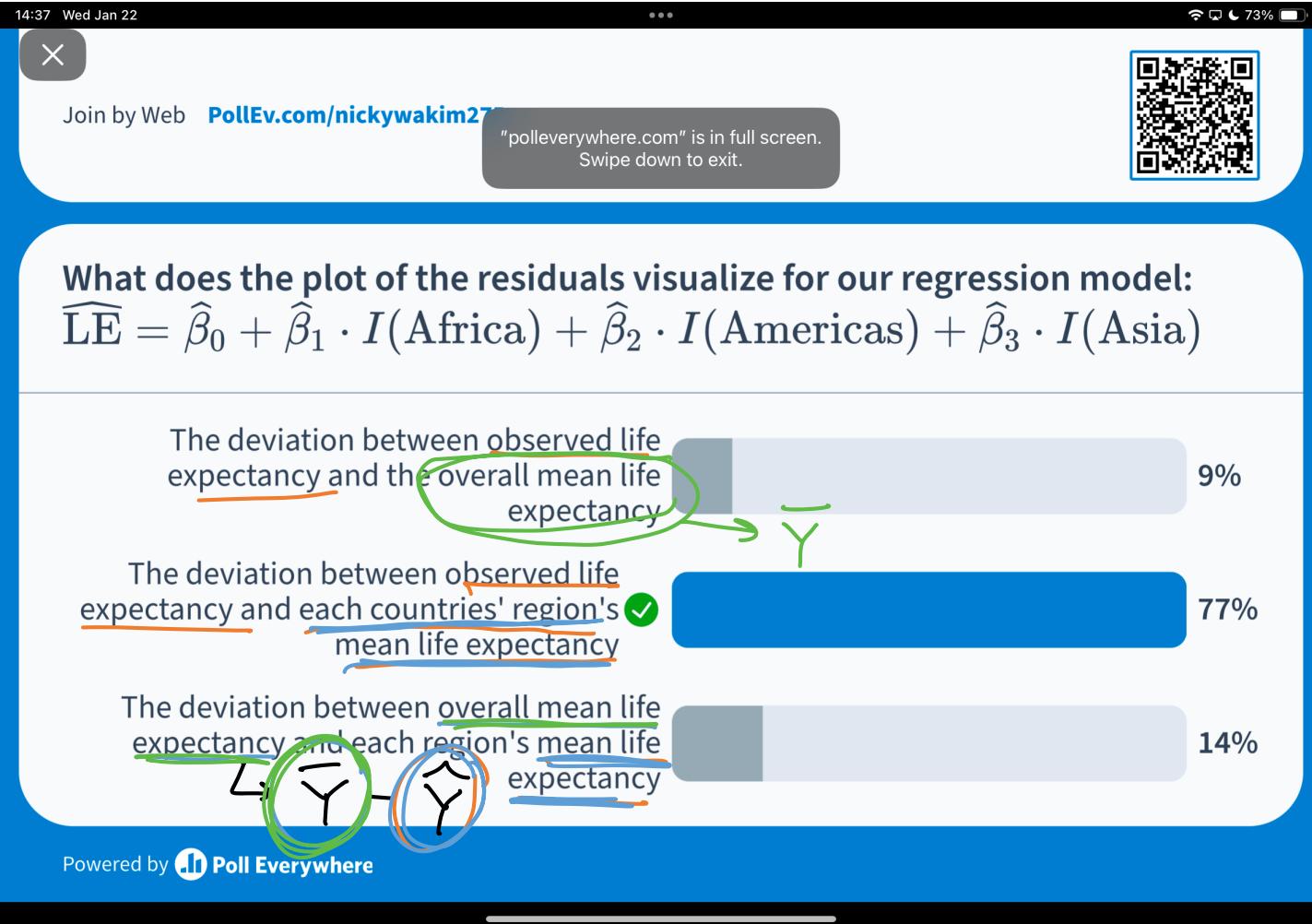
What does the plot of the residuals visualize for our regression model:  
 $\widehat{LE} = \hat{\beta}_0 + \hat{\beta}_1 \cdot I(\text{Africa}) + \hat{\beta}_2 \cdot I(\text{Americas}) + \hat{\beta}_3 \cdot I(\text{Asia})$

The deviation between observed life expectancy and the overall mean life expectancy 9%

The deviation between observed life expectancy and each country's region's mean life expectancy ✓ 77%

The deviation between overall mean life expectancy and each region's mean life expectancy 14%

Powered by  Poll Everywhere



$$\varepsilon_i = Y_i - (\hat{Y} | X_i)$$

↑  
WR<sub>i</sub>

Observed  
country's  
LE

mean  
LE given  
WR  
(of a specific  
country)

# Learning Objectives

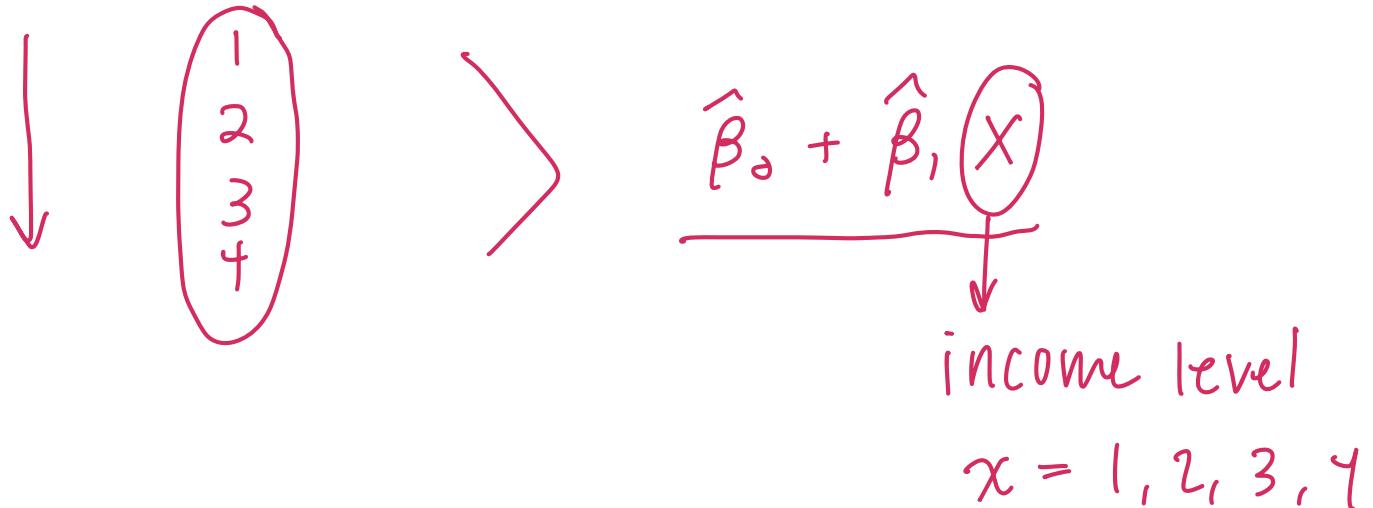
1. Understand why we need a new way to code categorical variables compared to continuous variables
2. Write the regression equation for a categorical variable using reference cell coding
3. Calculate and interpret coefficients for reference cell coding
4. Change the reference level in a categorical variable for reference cell coding
5. Create new variables and interpret coefficient for ordinal / scoring coding

# Let's look at life expectancy vs. four income levels

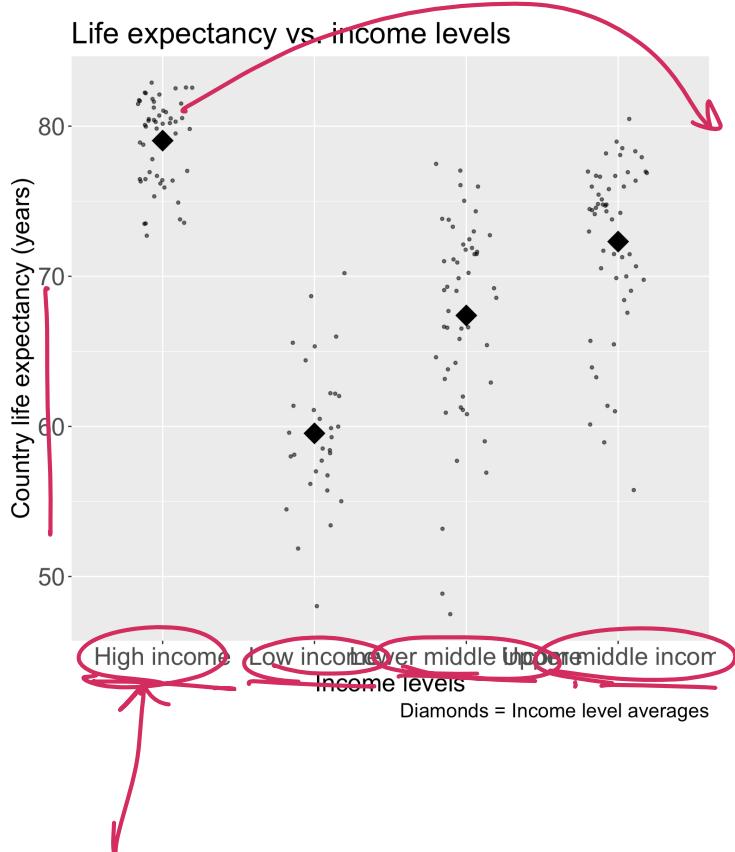
- Gapminder discusses individual income levels

- Income levels for a country is based on average GDP per capita, and grouped into:

- Low income
- Lower middle income
- Upper middle income
- High income



# Visualizing the ordinal variable, income levels



A few changes needed:

- Put the income levels in order

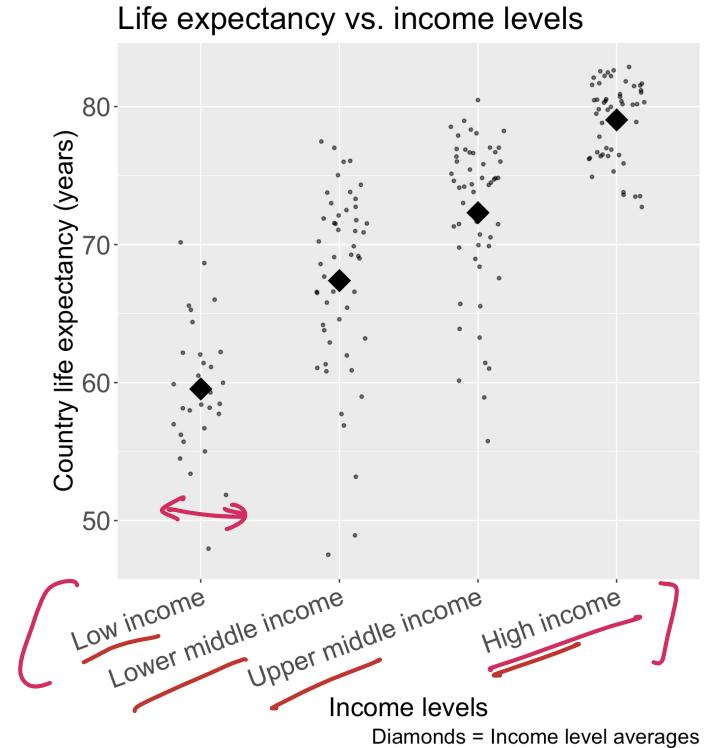
```
1 gapm2 = gapm2 %>%
2   mutate(income_levels = factor(income_levels,
3     ordered = T,
4     levels = c("Low income",
5       "Lower middle income",
6       "Upper middle income",
7       "High income")))
```

- Make the income levels readable

■ How to Rotate Axis Labels in ggplot2?

# Much better: Visualizing the ordinal variable, income levels

```
1 ggplot(gapm2, aes(x = income_levels, y = LifeExpectancyYrs)) +  
2   geom_jitter(size = 1, alpha = .6, width = 0.2) +  
3   stat_summary(fun = mean, geom = "point", size = 8, shape = 18) +  
4   labs(x = "Income levels",  
5        y = "Country life expectancy (years)",  
6        title = "Life expectancy vs. income levels",  
7        caption = "Diamonds = Income level averages") +  
8   theme(axis.title = element_text(size = 20),  
9         axis.text = element_text(size = 20),  
10        title = element_text(size = 20),  
11        axis.text.x=element_text(angle = 20, vjust = 1, hjust=1))
```



# How can we code this variable?

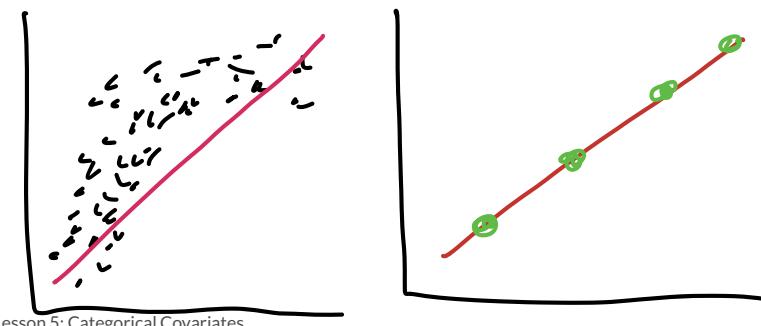
We have two options:

Treat the levels as nominal and use reference cell coding

- Like we did with world regions
- This option will not break the linearity assumption
- For  $g$  categories of the variable, we will have  $g - 1$  coefficients to estimate

Use the ordinal values to score the levels and treat as a numerical variable

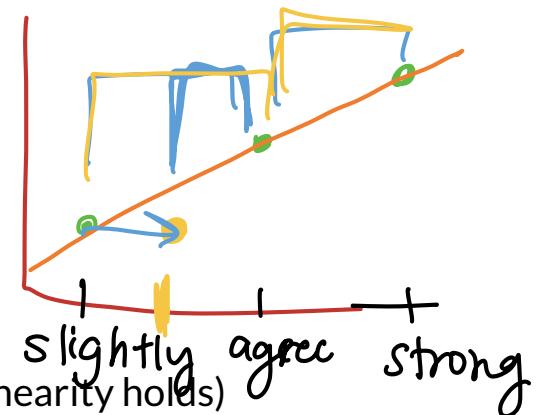
- Even if a variable is inherently ordered, we need to check that linearity holds if categories are represented as numbers
- This way of coding preserves more power in the model (less coefficients to estimate means more power)
- Only one coefficient to estimate



Lesson 5: Categorical Covariates

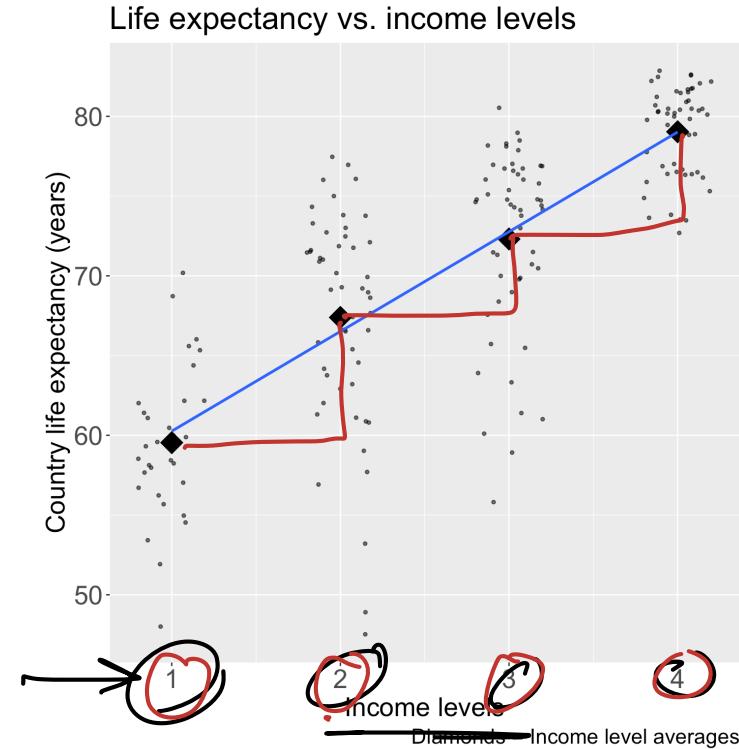
# Some important considerations when scoring ordinal variables

- Even if a variable is inherently ordered, we need to check that linearity holds if categories are represented as numbers (more in next lessons)
  - Linearity is an assumption of linear regression: that the relationship between X and Y is linear
- Assumes differences between adjacent groups are equal
  - Income levels are pre-set groups by Gapminder
  - Might be hard to interpret “every 1-level increase in income level”
- Is the variable part of the main relationship that you are investigating? (even if linearity holds)
  - If yes, consider leaving as reference cell coding unless the interpretation makes sense *nominal*
  - If no, and just needed as an adjustment in your model, then power benefit of scoring might be worth it!



# Check that linearity holds for income levels

- Using visual assessment, linearity holds for our income levels (more in next lessons)
- We can use the ordinal encoding for income levels



# Poll Everywhere Question 4

14:54 Wed Jan 22

X

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

QR code

In your IAT score project, for which variables might you consider using an ordinal coding/scoring?

something  
scale age category  
bmi  
groups bins weight best  
rating age-bins self thinness-fatness  
reported likert age-groups  
worst self-perception

Powered by  Poll Everywhere

# Ordinal coding / Scoring

- Map each income level to a number
- Usually start at 1

Income Level	Score
Low income	1
Lower middle income	2
Upper middle income	3
High income	4

```
1 gapm2 = gapm2 %>%
2   mutate(income_num = as.numeric(income_levels))
3 str(gapm2$income_num)
```

num [1:187] 1 3 3 4 2 4 3 2 4 4 ...

factor variable  
w/ set levels

# Run the model with the scored income

```
1 mod_inc2 = lm(LifeExpectancyYrs ~ income_num, data = gapm2)
2 tidy(mod_inc2) %>% gt() %>% tab_options(table.font.size = 37) %>%
3   fmt_number(decimals = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	54.01	1.06	51.03	0.00
income_num	6.25	0.37	16.91	0.00

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

$$\widehat{LE} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Income level}$$
$$\widehat{LE} = 54.01 + 6.25 \cdot \text{Income level}$$

1.5

- Keep in mind: We cannot calculate the expected outcome outside of the scoring values
  - For example, we cannot find the mean life expectancy for an income level of 1.5

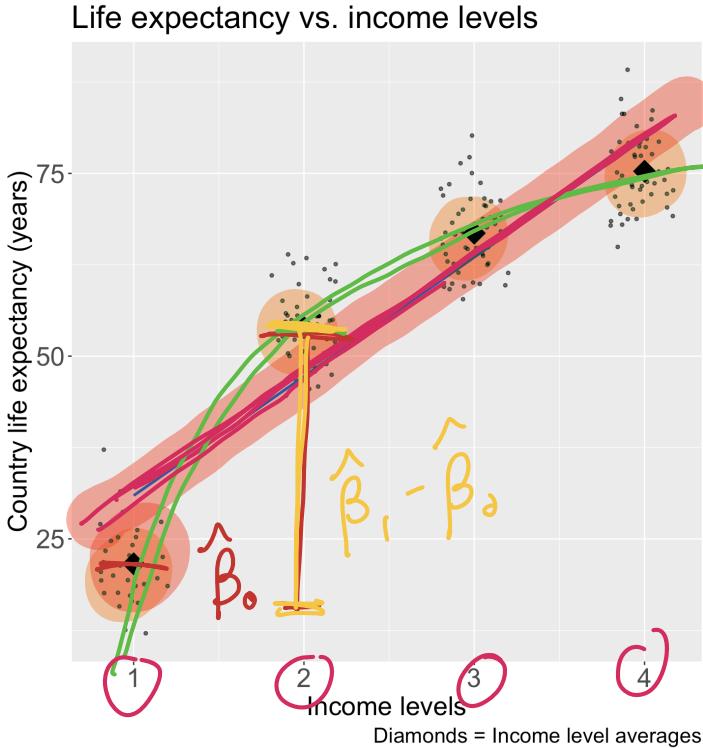
# Interpreting the model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	54.01	1.06	51.03	0.00	51.92	56.10
income_num	6.25	0.37	16.91	0.00	5.52	6.98

$$\widehat{LE} = 54.01 + 6.25 \cdot \text{Income level}$$

- **Interpreting the intercept:** At an income level of 0, mean life expectancy is 54.01 (95% CI: 51.92, 56.10).
  - Note: this does not make sense because there is no income level of 0!
- **Interpreting the coefficient for income:** For every 1-level increase in income level, mean life expectancy increases 6.25 years (95% CI: 5.52, 6.98).

# What if life expectancy vs. income level looked like this?



- No longer maintaining the linearity assumption
- Need to use reference cell coding
- We would fit the following model:

$$LE = \beta_0 + \beta_1 \cdot I(\text{Lower middle income}) + \beta_2 \cdot I(\text{Upper middle income}) + \beta_3 \cdot I(\text{High income}) + \epsilon$$

## If time...

Let's walk through categorical variables that have multiple selections

- So each group is not mutually exclusive
- We could make an indicator for each category, but individuals could be a part of multiple categories
- Also, thinking about income levels - can we combine two groups to make one??

