MULTIPLE REGRESSION

MPA 630: Data Science for Public Management

October 18, 2018

Fill out your reading report on Learning Suite

PLAN FOR TODAY

Slopes and intercepts again

Regression with categorical variables

Multiple regression

Interpretation practice

SLOPES AND INTERCEPTS AGAIN

GOAL OF REGRESSION

Explain (or predict) variation in an outcome using one or more explanatory variables

DRAWING LINES WITH STATS

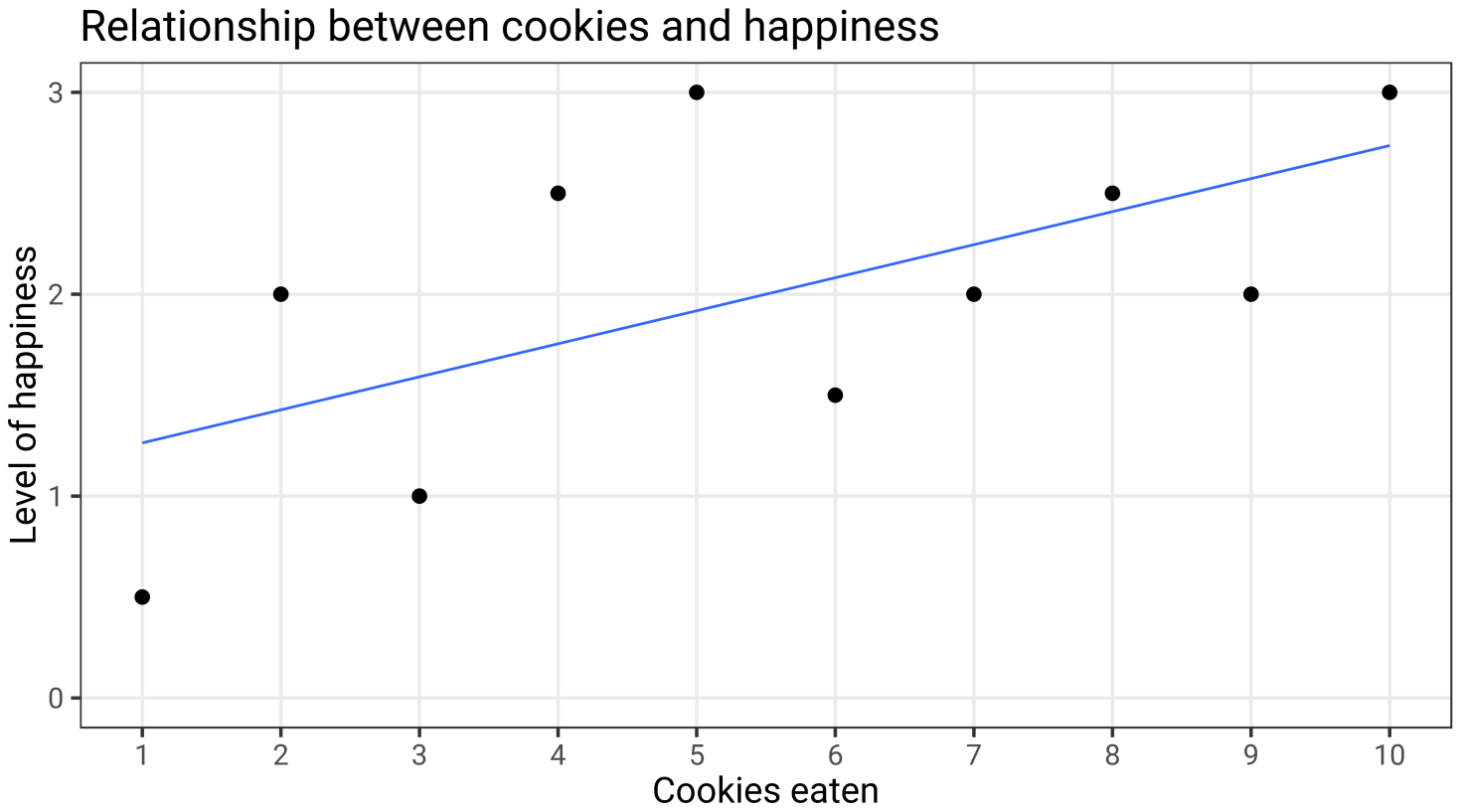
$$y = mx + b$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \varepsilon$$

y	$\hat{y}$	Outcome variable
x	x_1	Explanatory variable
m	β_1	Slope
b	β_0	y intercept
	ε	Error (residuals)

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{cookies} + \epsilon$$

$$\hat{\text{happiness}} = 1.1 + (0.164 \times \text{cookies}) + \epsilon$$



term	estimate	std_er	stati	p_va	lower	uppe
intercept	1.1	0	9	0	0.	4
cookies	0.164	76	159	63	11	338

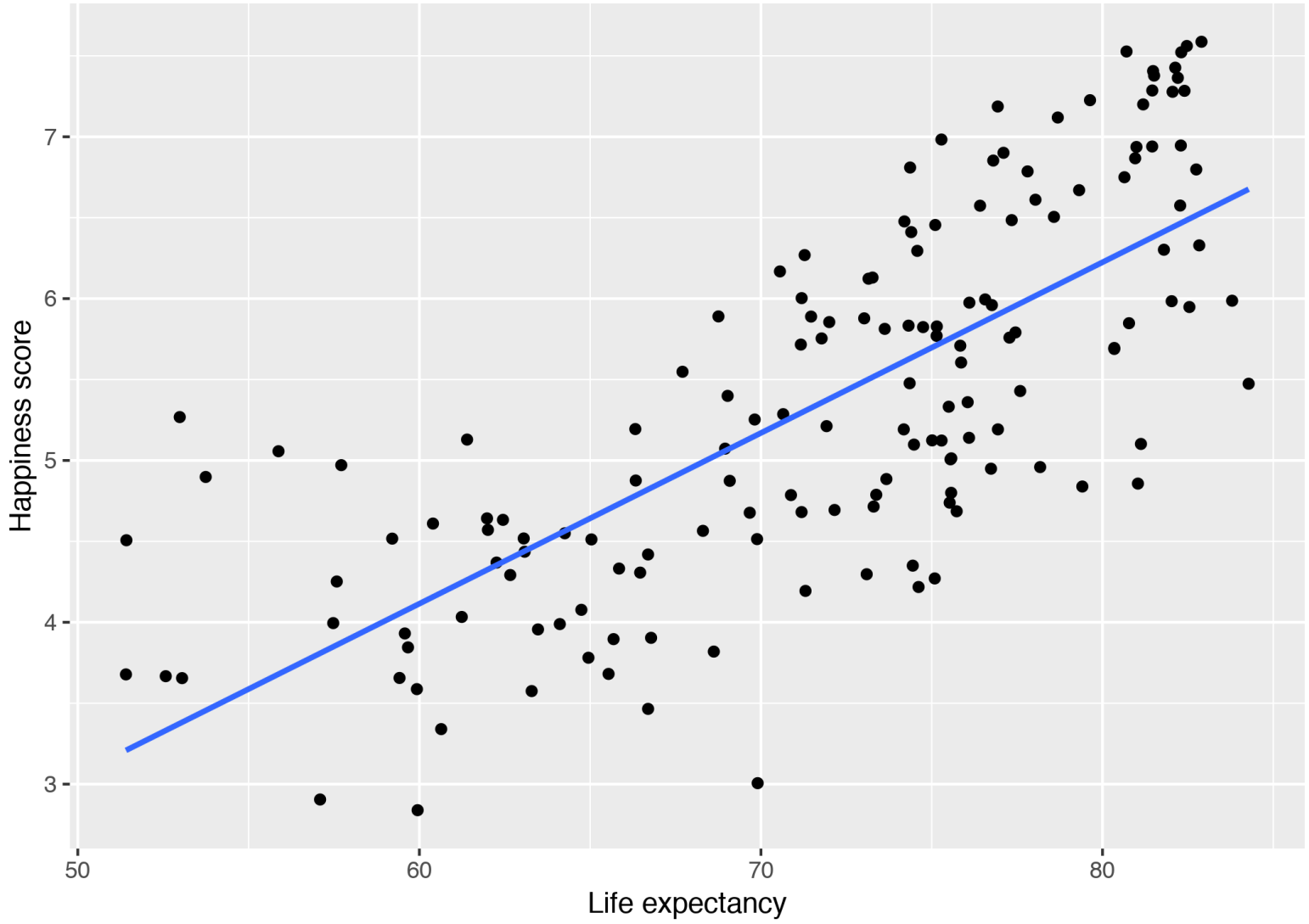
Chapter 10
Chapter 11
Chapter 11
Chapter 9
Chapter 9

TEMPLATE

A one unit increase in x is *associated* with a β_1 increase (or decrease) in y, on average

$$\hat{\text{happiness}} = 1.1 + (0.164 \times \text{cookies}) + \epsilon$$

WORLD HAPPINESS



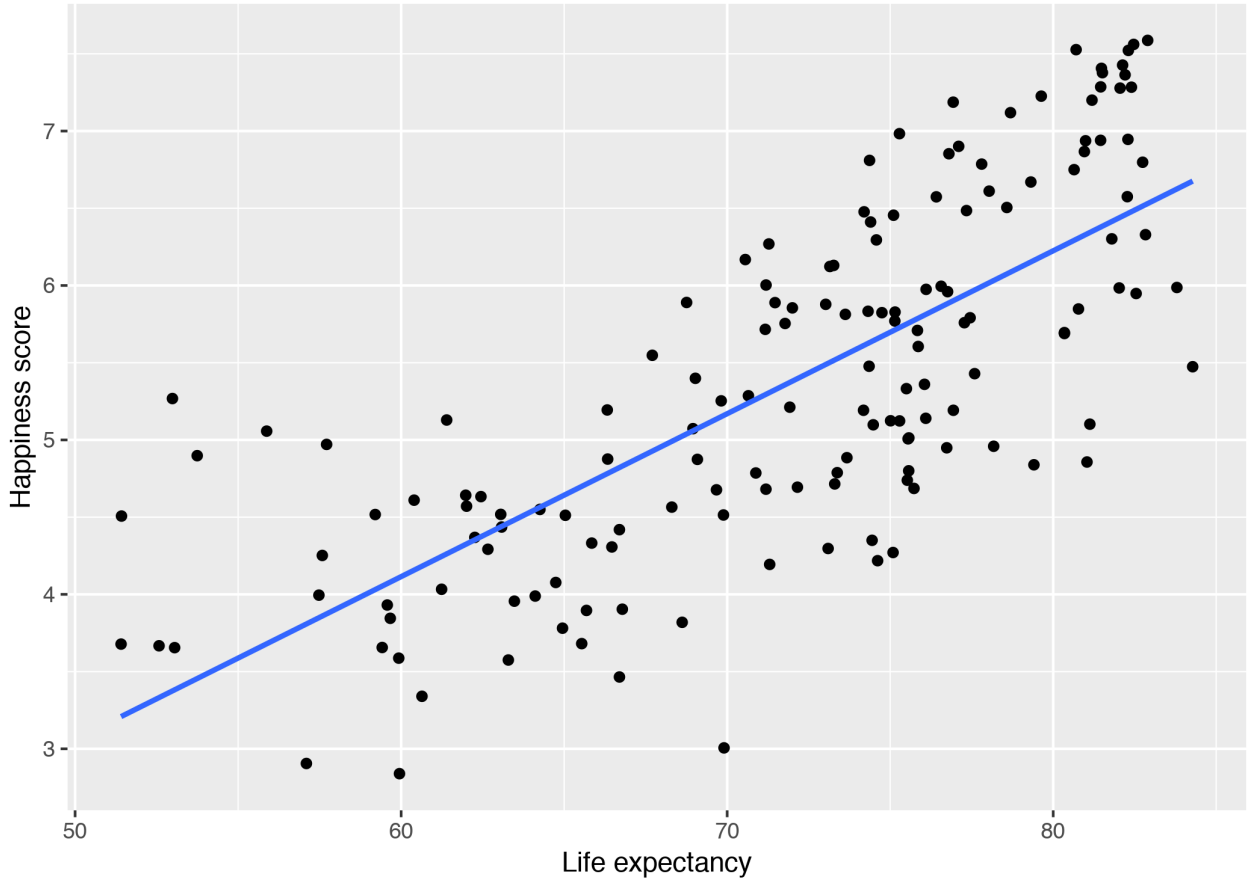
```
model1 <- lm(happiness_score ~ life_expectancy,
             data = world_happiness)
model1 %>%
  get_regression_table()
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-2.215	0.556	-3.983	0	-3.313	-1.116
life_expectancy	0.105	0.008	13.73	0	0.09	0.121

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \epsilon$$

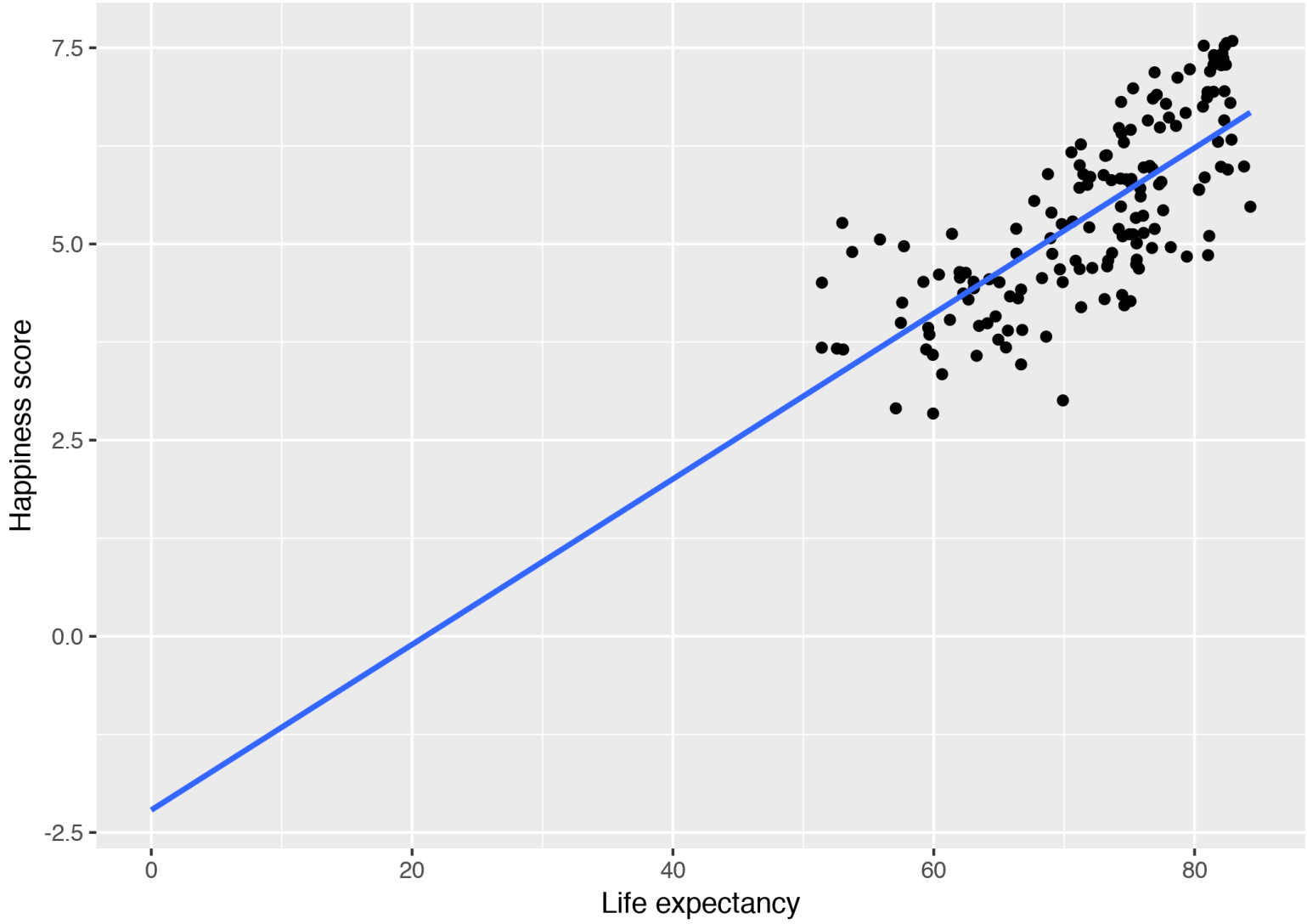
$$\hat{\text{happiness}} = -2.215 + (0.105 \times \text{life expectancy}) + \epsilon$$

WORLD HAPPINESS



$$\hat{\text{happiness}} = -2.215 + (0.105 \times \text{life expectancy}) + \epsilon$$

WORLD HAPPINESS



REGRESSION WITH CATEGORICAL VARIABLES

VARIABLE TYPES

Numeric variables	Categorical variables
(Continuous)	(Factors)
Numbers	**Not numbers**

NUMERIC OR CATEGORICAL?

Income

True/false

- 18–25
- 26–34
- 35–44
- 45–54

State

Weight

Tax rates

Political party

Gender

- Strongly agree
- Agree
- Disagree
- Strongly disagree

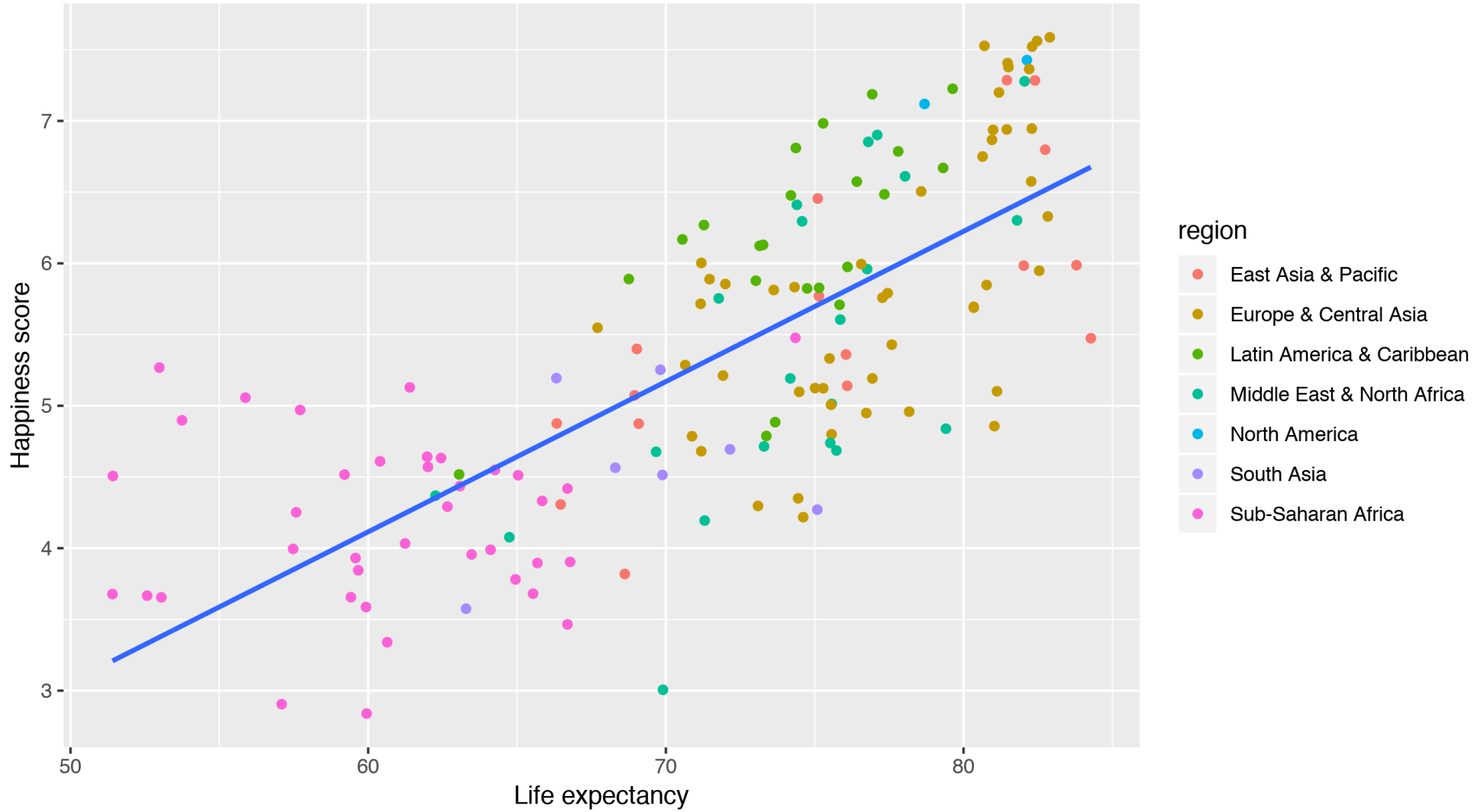
Year

Happiness

Age

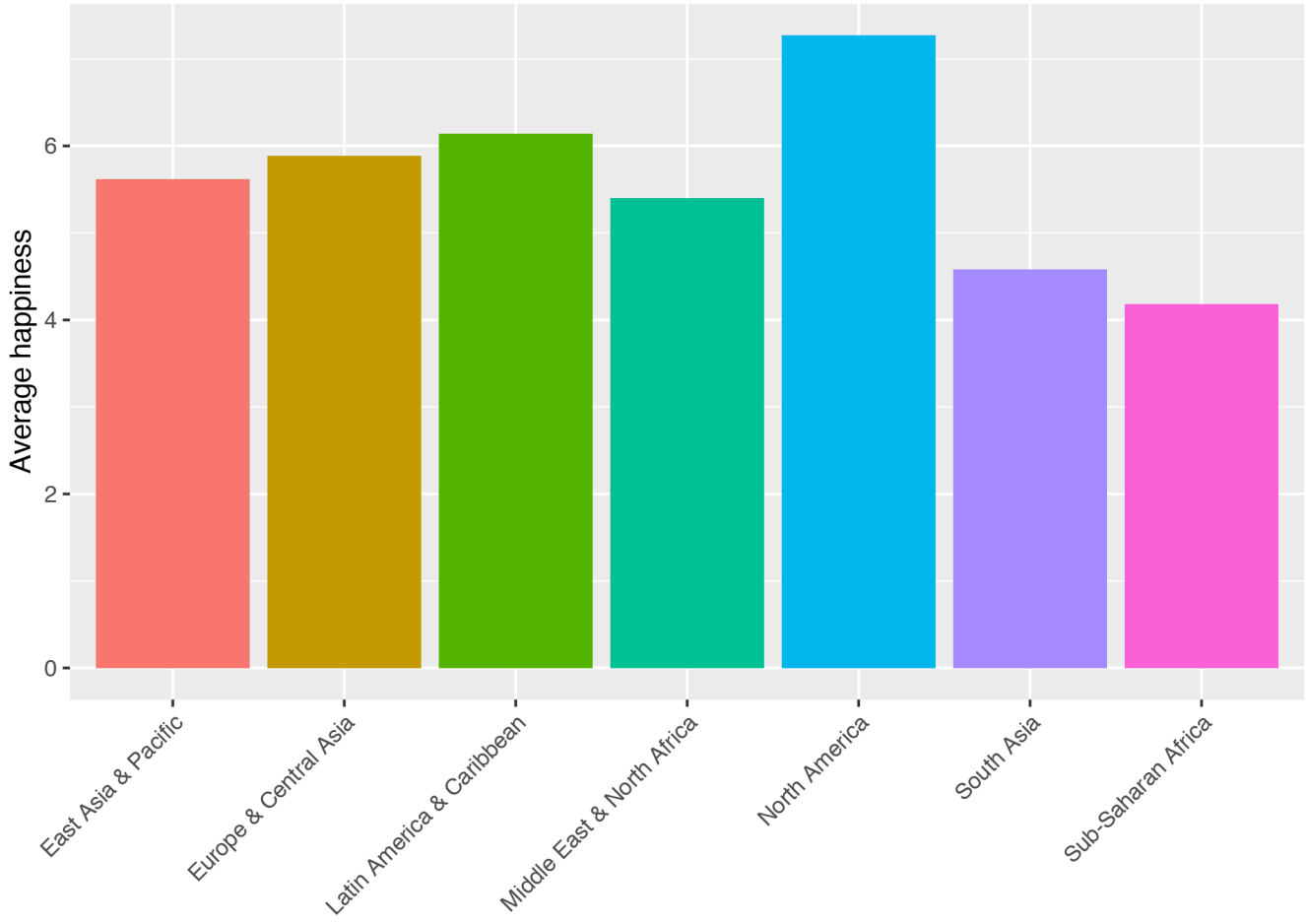
Day of the week

LIFE EXPECTANCY IS NOT THE FULL STORY



REGIONAL DIFFERENCES

region	avg
East Asia & Pacific	5.618
Europe & Central Asia	5.889
Latin America & Caribbean	6.145
Middle East & North Africa	5.404
North America	7.273
South Asia	4.581
Sub-Saharan Africa	4.181



```
model2 <- lm(happiness_score ~ region, data = world_happiness)
```

term	estimate	std_error	statistic	p_value
intercept	5.618	0.217	25.84	0
regionEurope & Central Asia	0.271	0.25	1.084	0.28
regionLatin America & Caribbean	0.527	0.286	1.844	0.067
regionMiddle East & North Africa	-0.214	0.289	-0.742	0.459
regionNorth America	1.655	0.652	2.538	0.012
regionSouth Asia	-1.037	0.394	-2.631	0.009
regionSub-Saharan Africa	-1.437	0.259	-5.544	0

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{Europe} + \beta_2 \text{Latin America} +$$
$$\beta_3 \text{MENA} + \beta_4 \text{North America} +$$
$$\beta_5 \text{South Asia} + \beta_6 \text{Sub-Saharan Africa} + \epsilon$$

```
model2 <- lm(happiness_score ~ region, data = world_happiness)
```

term	estimate	std_error	statistic	p_value
intercept	5.618	0.217	25.84	0
regionEurope & Central Asia	0.271	0.25	1.084	0.28
regionLatin America & Caribbean	0.527	0.286	1.844	0.067
regionMiddle East & North Africa	-0.214	0.289	-0.742	0.459
regionNorth America	1.655	0.652	2.538	0.012
regionSouth Asia	-1.037	0.394	-2.631	0.009
regionSub-Saharan Africa	-1.437	0.259	-5.544	0

$$\hat{\text{happiness}} = 5.618 + (0.271 \times \text{Europe}) + (0.527 \times \text{Latin America}) + (-0.214 \times \text{MENA}) + (1.655 \times \text{North America}) + (-1.037 \times \text{South Asia}) + (-1.437 \times \text{Sub-Saharan Africa}) + \epsilon$$

HAPPINESS IN EAST ASIA

$$\hat{\text{happiness}} = 5.618 + (0.271 \times \text{Europe}) + (0.527 \times \text{Latin America}) + (-0.214 \times \text{MENA}) + (1.655 \times \text{North America}) + (-1.037 \times \text{South Asia}) + (-1.437 \times \text{Sub-Saharan Africa}) + \epsilon$$

$$\hat{\text{happiness}} = 5.618 + (0.271 \times 0) + (0.527 \times 0) + (-0.214 \times 0) + (1.655 \times 0) + (-1.037 \times 0) + (-1.437 \times 0) + \epsilon$$

$$\hat{\text{happiness}} = 5.618$$

HAPPINESS IN EUROPE

$$\hat{\text{happiness}} = 5.618 + (0.271 \times \text{Europe}) + (0.527 \times \text{Latin America}) +$$
$$(-0.214 \times \text{MENA}) + (1.655 \times \text{North America}) +$$
$$(-1.037 \times \text{South Asia}) + (-1.437 \times \text{Sub-Saharan Africa}) + \epsilon$$

$$\hat{\text{happiness}} = 5.618 + (0.271 \times 1) + (0.527 \times 0) +$$
$$(-0.214 \times 0) + (1.655 \times 0) +$$
$$(-1.037 \times 0) + (-1.437 \times 0) + \epsilon$$

$$\hat{\text{happiness}} = 5.618 + (0.271 \times 1)$$
$$= 5.889$$

Regression coefficients

term	estimate
intercept	5.618
regionEurope & Central Asia	0.271
regionLatin America & Caribbean	0.527
regionMiddle East & North Africa	-0.214
regionNorth America	1.655
regionSouth Asia	-1.037
regionSub-Saharan Africa	-1.437

Averages

region	avg
East Asia & Pacific	5.618
Europe & Central Asia	5.889
Latin America & Caribbean	6.145
Middle East & North Africa	5.404
North America	7.273
South Asia	4.581
Sub-Saharan Africa	4.181

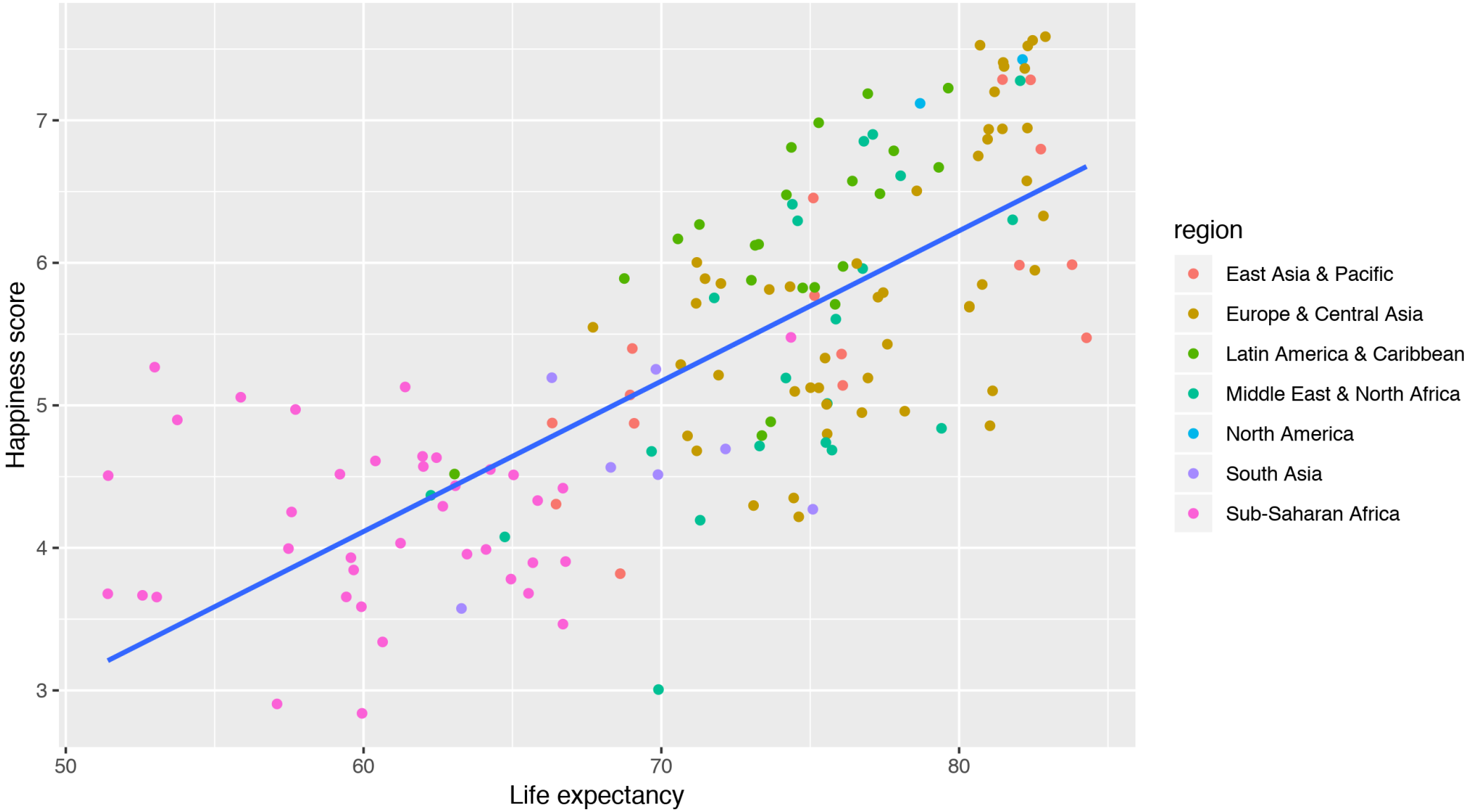
TEMPLATE

On average, y is β_n units larger (or smaller) in x_n, compared to x_0

On average, national happiness is 1.65 points higher in North America than in East Asia

On average, compared to East Asia, national happiness is 1.44 points lower in Sub Saharan Africa

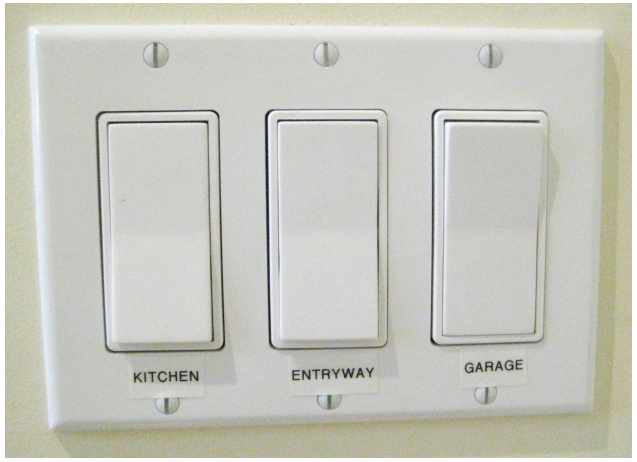
GETTING CLOSER



MULTIPLE REGRESSION

SLIDERS AND SWITCHES

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \epsilon$$

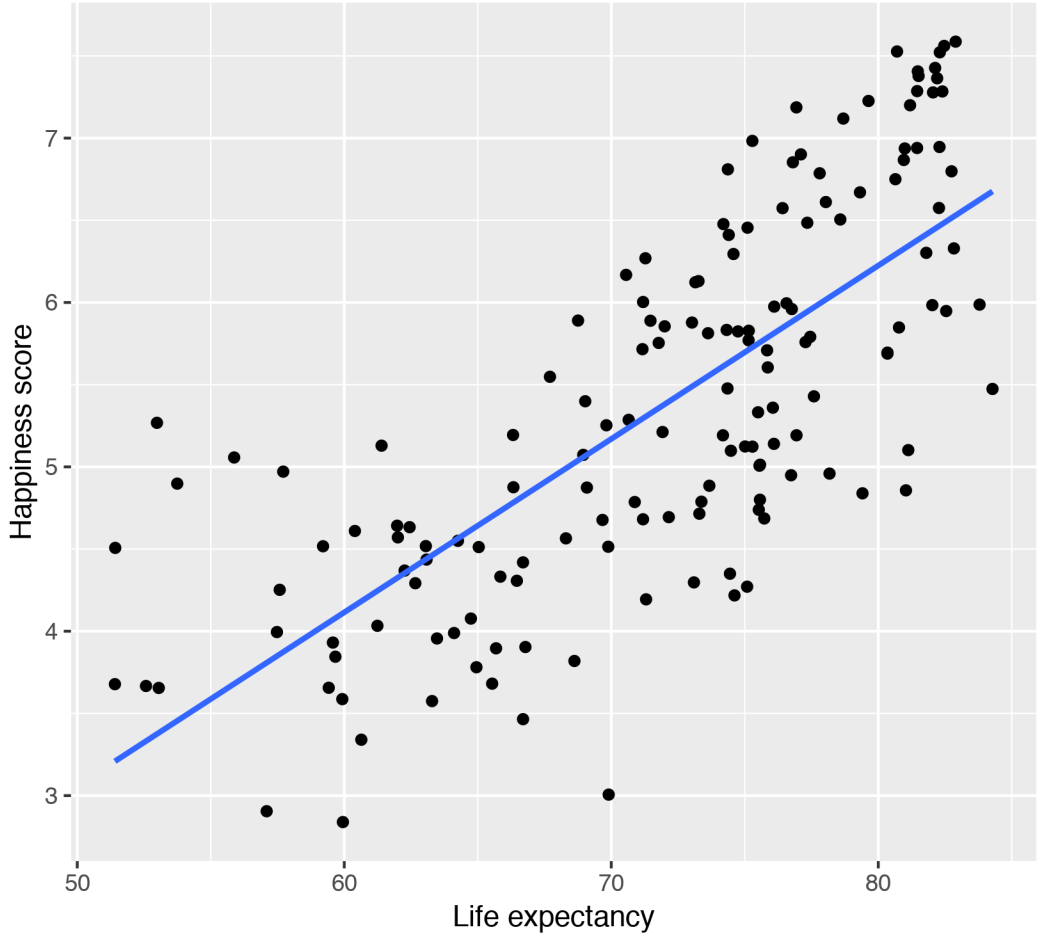


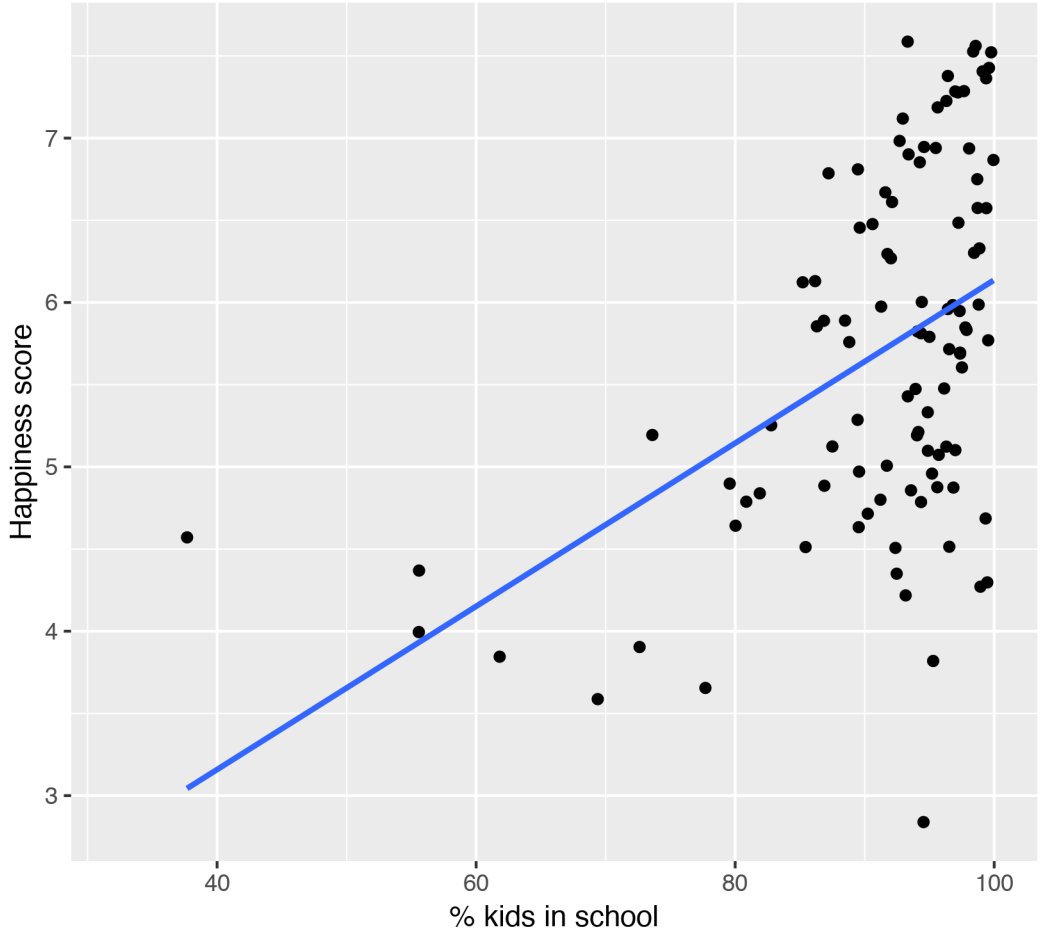
$$\begin{aligned}\hat{\text{happiness}} = \beta_0 &+ \beta_1 \text{Europe} + \beta_2 \text{Latin America} + \\ &\beta_3 \text{MENA} + \beta_4 \text{North America} + \\ &\beta_5 \text{South Asia} + \beta_6 \text{Sub-Saharan Africa} + \epsilon\end{aligned}$$

ALL AT ONCE!

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \beta_2 \text{school enrollment} +$$
$$\beta_3 \text{Europe} + \beta_4 \text{Latin America} + \beta_5 \text{MENA} +$$
$$\beta_6 \text{North America} + \beta_7 \text{South Asia} + \beta_8 \text{SSA} + \epsilon$$

HAPPINESS ~ LIFE + SCHOOL





```
model_life <- lm(happiness_score ~ life_expectancy,
                 data = world_happiness)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-2.215	0.556	-3.983	0	-3.313	-1.116
life_expectancy	0.105	0.008	13.73	0	0.09	0.121

```
model_school <- lm(happiness_score ~ school_enrollment,
                   data = world_happiness)
```

term	estimate	std_error	statistic	p_value	lower_ci
intercept	1.173	0.879	1.334	0.185	-0.571
school_enrollment	0.05	0.01	5.19	0	0.031

BOTH AT THE SAME TIME

Life expectancy and school enrollment both explain some variation in happiness

On its own, a 1 year increase in school enrollment is associated with a 0.105 point increase in happiness, on average

On its own, a 1% increase in school enrollment is associated with a 0.05 point increase in happiness, on average

Some of that explanation is shared!

```
model_life_school <- lm(happiness_score ~ life_expectancy +
                          school_enrollment,
                        data = world_happiness)
```

term	estimate	std_error	statistic	p_value	lower_ci
intercept	-2.111	0.835	-2.529	0.013	-3.767
life_expectancy	0.101	0.014	7.447	0	0.074
school_enrollment	0.003	0.01	0.331	0.741	-0.016

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \beta_2 \text{school enrollment} + \epsilon$$

$$\hat{\text{happiness}} = -2.11 + (0.101 \times \text{life expectancy}) + (0.003 \times \text{school enrollment}) + \epsilon$$

FILTERING OUT VARIATION

Each x in the model explains some portion of the variation in y

This will often change the simple regression coefficients

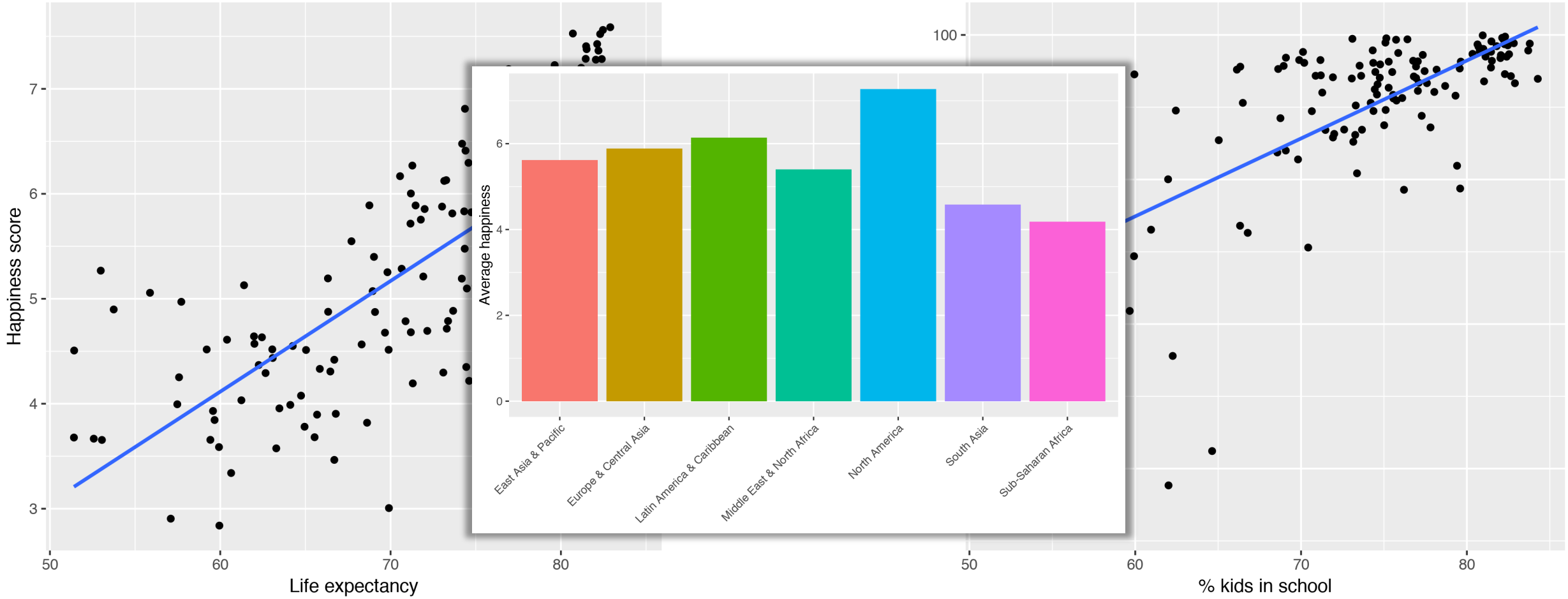
Interpretation is a little trickier, since you can only ever move **one** switch or slider (or variable)

TEMPLATE

Taking all other variables in the model into account, a one unit increase in x_n is associated with a β_n increase (or decrease) in y, on average

Controlling for school enrollment, a 1 year increase in life expectancy is associated with a 0.1 point increase in national happiness, on average

HAPPINESS ~ LIFE + SCHOOL + REGION



```
model_life_school_region <-
  lm(happiness_score ~ life_expectancy + school_enrollment + region,
     data = world_happiness)
```

term	estimate	std_error	statistic	p_value
intercept	-2.821	1.355	-2.083	0.04
life_expectancy	0.102	0.017	5.894	0
school_enrollment	0.008	0.01	0.785	0.435
regionEurope & Central Asia	0.031	0.255	0.123	0.902
regionLatin America & Caribbean	0.732	0.294	2.489	0.015
regionMiddle East & North Africa	0.189	0.317	0.597	0.552
regionNorth America	1.114	0.581	1.917	0.058
regionSouth Asia	-0.249	0.45	-0.553	0.582
regionSub-Saharan Africa	0.326	0.407	0.802	0.425

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \beta_2 \text{school enrollment} +$$
$$\beta_3 \text{Europe} + \beta_4 \text{Latin America} + \beta_5 \text{MENA} +$$
$$\beta_6 \text{North America} + \beta_7 \text{South Asia} + \beta_8 \text{SSA} + \epsilon$$

Controlling for school enrollment and region, a 1 year increase in life expectancy is associated with a 0.102 point increase in national happiness, on average

Controlling for life expectancy and region, a 1% increase in school enrollment is associated with a 0.008 point increase in national happiness, on average

On average, controlling for life expectancy and school enrollment, North America is 1.114 points happier than East Asia

On average, controlling for life expectancy and school enrollment, South Asia is 0.249 points less happier than East Asia

INTERPRETATION PRACTICE

Y = test scores

Not small vs. small

Class doesn't have aide vs. class has aide

Student not white/Asian vs. yes

Student is boy vs. girl

Student does not receive FRL vs. yes

Teacher not white/Asian vs. yes

Years (actual number)

Teacher does not have MA vs. yes

Table 2: OLS models for four standardized tests

VARIABLES	(1) Reading	(2) Math	(3) Listening	(4) Words
Small class	6.47***	8.84***	3.24**	6.99***
	(1.45)	(2.32)	(1.42)	(1.60)
Regular + aide class	1.00	0.42	-0.58	1.27
	(1.26)	(2.14)	(1.32)	(1.42)
White or Asian	7.85***	16.91***	17.98***	7.08***
	(1.61)	(2.40)	(1.70)	(1.91)
Girl	5.39***	6.46***	2.67***	5.03***
	(0.78)	(1.12)	(0.74)	(0.94)
Free/reduced lunch	-14.69***	-20.08***	-15.23***	-15.97***
	(0.91)	(1.33)	(0.90)	(1.07)
Teacher white or Asian	-0.56	-1.01	-3.68	0.46
	(2.66)	(3.80)	(2.59)	(3.07)
Years of teacher experience	0.30**	0.42**	0.25*	0.30**
	(0.12)	(0.20)	(0.15)	(0.14)
Teacher has MA	-0.75	-2.20	0.50	0.24
	(1.25)	(2.08)	(1.24)	(1.46)
School fixed effects	X	X	X	X
Constant	431.69***	475.52***	531.28***	428.97***
	(3.12)	(4.49)	(2.84)	(3.59)

FOLLOW ALONG IN R