SOC6302 Statistics for Sociologists

Monica Alexander

Week 8: Simple Linear Regression

Announcements

- Assignment 3 out (1 question)
- Details for research project proposal and EDA also out

Where are we at

- We are interested in making inferences about a population
- **Toolkit 1**: Probability
  - when answering questions using the data we have available, there is chance/randomness involved
  - e.g. data from a sample
  - e.g. deciding whether a particular observation comes from a population
  - we can use probability to quantify uncertainty
- **Toolkit 2**: descriptive statistics and data visualizations
  - before running a model, we can get a long way by looking at key summary stats and charts
  - e.g. means by group tells us something about differences / similarities
  - e.g. key charts to visualize distributions and relationships

Where are we going

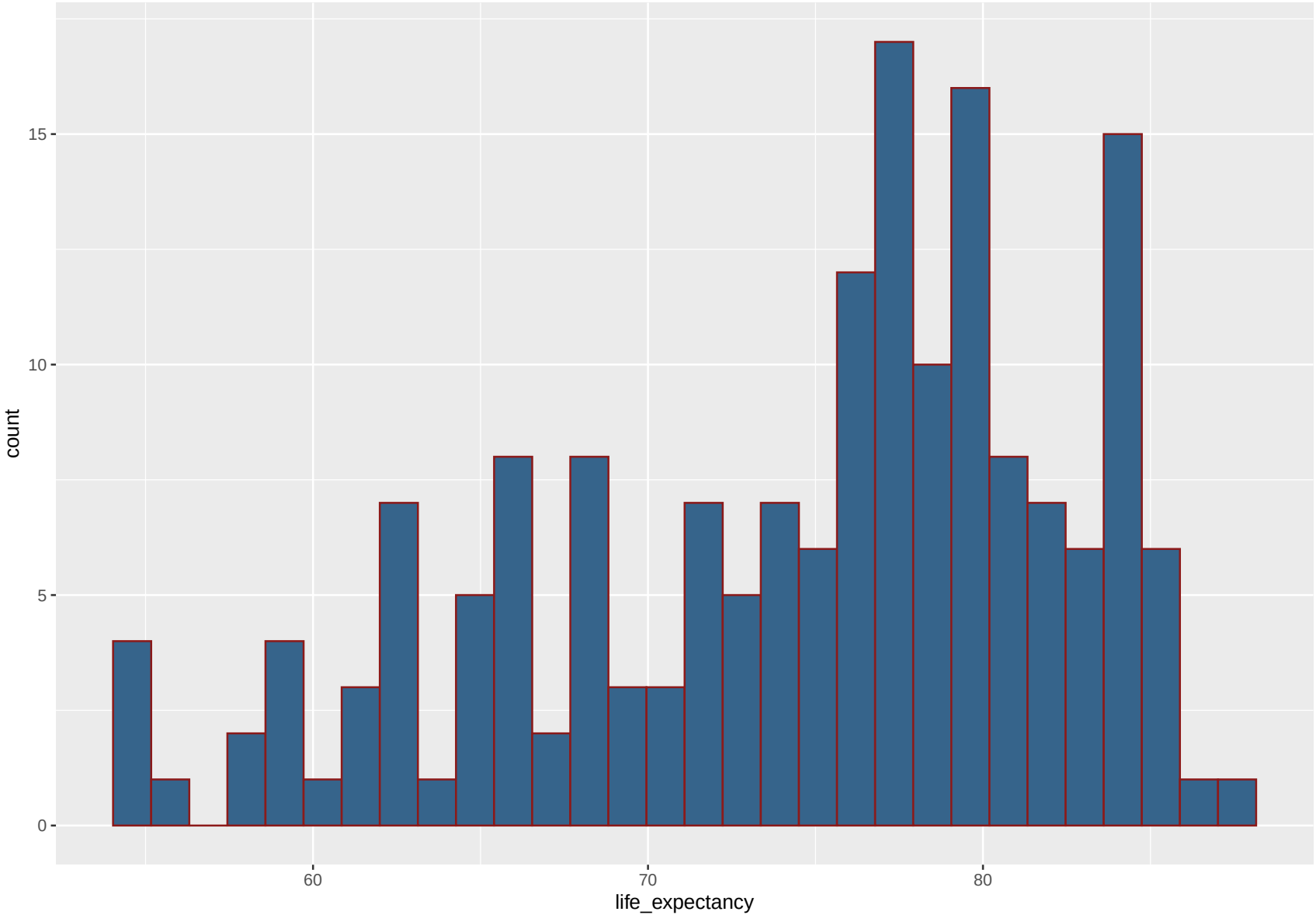
- We are interested in explaining patterns in an outcome of interest (dependent variable) Y in relation to one or more explanatory variables $X_1, X_2, \ldots$
- i.e. how does Y vary with different levels of X_1?
- We could explore this with graphs/ summary statistics!
- But **regression models** allow us to quantify relationships, taking into account **uncertainty** based on the data that we observe

Running example

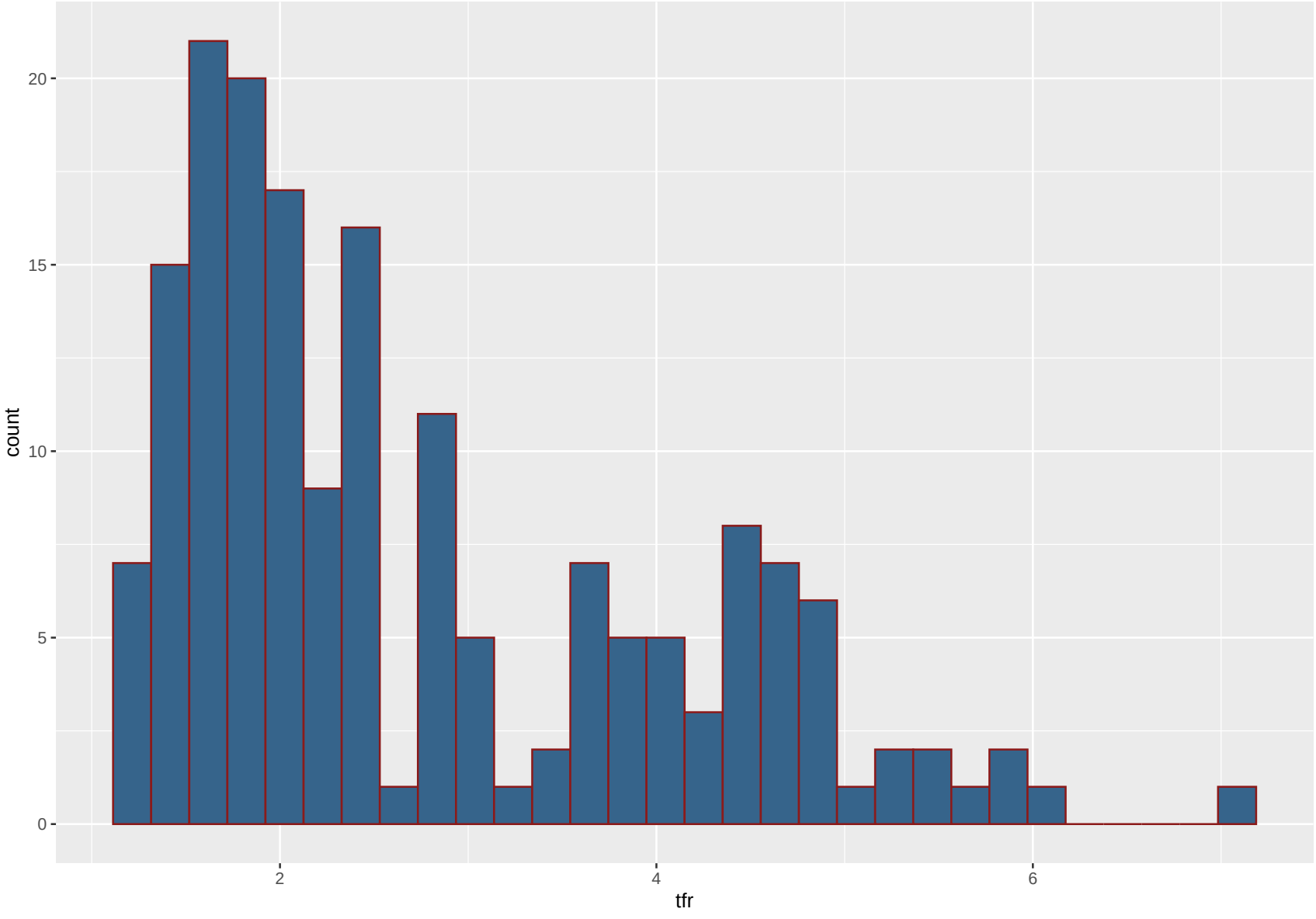
- Back to the `country_indicators` dataset.
- Research question: In 2017, how does the expected value of life expectancy differ or change across countries with different levels of fertility?
- In other words, is life expectancy associated with fertility, and if so, how?

How could we explore this graphically?

Histogram of life expectancy



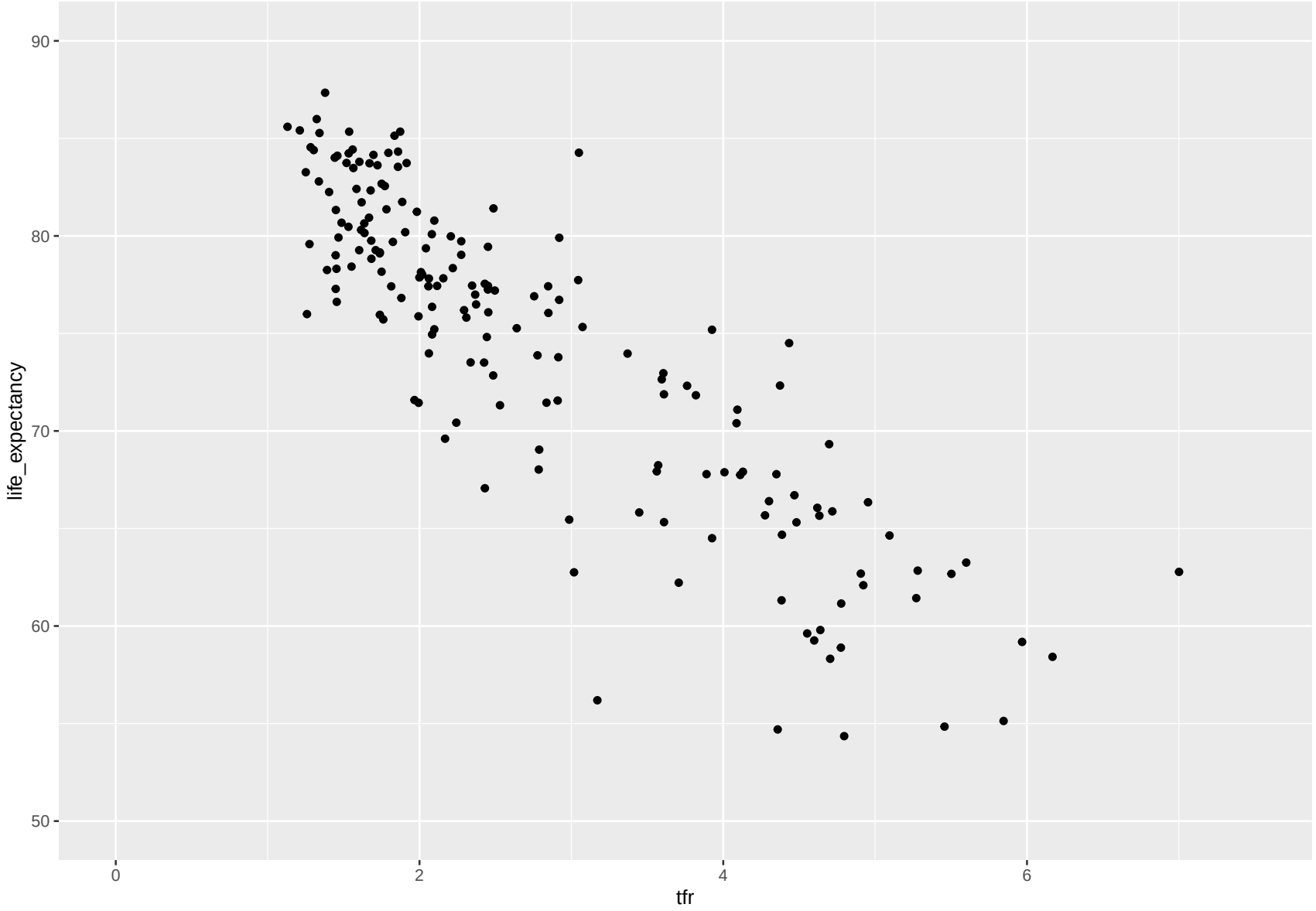
Histogram of fertility



Summarief s by TFR level

tfr_level	mean_life_expectancy
Low TFR	81.15
Mid TFR	74.08
High TFR	63.73
Very high TFR	60.60

Scatterplot of life expectancy versus TFR



The story so far

Our question:

- In 2017, how does the expected value of life expectancy differ or change across countries with different levels of fertility?

We can already kind of answer this based on summary statistics and graphs of our data. So what's missing?

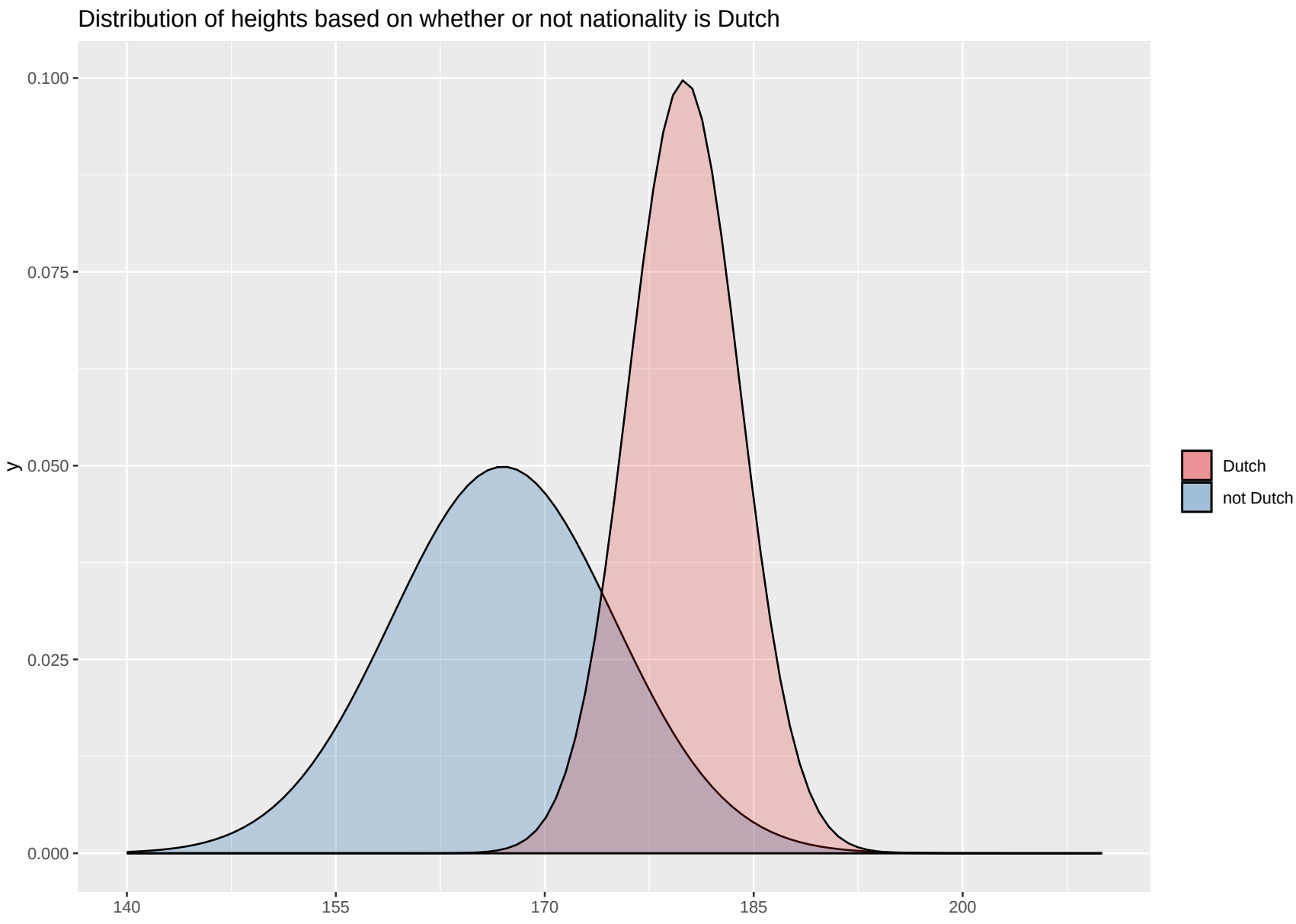
- Ideally, want to account for uncertainty in the data we observe and make some statements about how sure (or otherwise) we are that the outcome of interest (life expectancy) differs by covariate of interest (fertility)
- Linear regression is a model that allows us to do this

Conditioning

Conditioning on covariates

- We are considering our outcome of interest (life expectancy) by different values of our covariate
- That is, we are **conditioning** on values of fertility and considering different characteristics of our outcome
- This is trying to study at the **conditional distribution** of life expectancies

Conditional distributions more generally



Conditional expectations

- As well as looking at conditional distributions, we can look at other measures conditioning on covariates of interest
- Of particular interest is the **conditional expectation**, which is the (weighted) average of all possible values of the outcome of interest Y, given a particular value of covariate of interest X.
- This is essentially a group mean; a measure of central tendency of a conditional distribution.

Expectations: notation

Expected values (not conditioned!) are written

$$E(Y)$$

Expected values: calculations

$$E(Y) = \sum_y yp(y)$$

- I toss an unbiased coin 3 times. What's the expected number of heads?

Expected values: calculations

Conditional expectations

Conditional expected values are written

$$E(Y \mid X = x)$$

or for more than one X

$$E(Y \mid X_1 = x_1, X_2 = x_2, \dots)$$

Conditional expectation: calculation

$$E(Y \mid X = x) = \sum_y y p_{Y|X}(y \mid x)$$

- I toss a coin 3 times, and my technique is such that H/T are equally likely
- My toddler tosses a coin 3 time, and his technique is such that it always comes up heads

What is $E(Y|X = \text{Monica})$? What is $E(Y|X = \text{toddler})$?

Conditional expectations with outcomes more interesting than coins

- In our example, one measure that we are interested in is life expectancy, conditioning on fertility levels
- $E(\text{life expectancy} \mid \text{fertility}) = E(Y_i|X_i = x_i)$ is a good summary measure that allows us to quantify differences across subgroups of interest
- In our example we don't know (for sure) the underlying probabilities of all possible outcomes (which are infinite!), but can still estimate $E(Y \mid X_1 = x_1, X_2 = x_2, \dots)$ from the data

How does $E(Y_i|X_i = x_i)$ relate to the data we observe, Y_i?

The conditional expectation decomposition property

Any outcome Y_i can be decomposed into the following

$$Y_i = E\left(Y_i \mid \mathbf{X_i}\right) + \varepsilon_i$$

One way to interpret the this is that Y_i can be decomposed into two independent components: a component "explained by X_i" and a component "unexplained by X_i"

The simple linear regression model

Back to our example

- Research question: In 2017, how does the expected value of life expectancy differ or change across countries with different levels of fertility?
- In other words, is life expectancy associated with fertility, and if so, how?

SLR set-up and notation

- Y_i is the response variable, and X_i is the explanatory variable

Questions:

- In our example, what is Y and what is X?
- In our example, what does i refer to?
- In a example using GSS, what would i refer to?

The simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

SLR models Y_i as a simple linear function of X_i with two parameters, β_0 and β_1

- β_0 and β_1 are **regression coefficients**
- β_0 is called the **intercept**
- β_1 is called the **slope**

The simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- β_0 is the is the expected value, or population mean, of Y_i given X_i is equal to zero.
- β_1 is the change in the expected value, or population mean, of Y_i associated with a one unit increase in X_i

Side note: remember this?

The conditional expectation decomposition property

$$Y_i = E\left(Y_i \mid X_i\right) + \varepsilon_i$$

So the SLR is a model for the conditional expectation

$$\begin{aligned} Y_i &= E\left(Y_i \mid X_i\right) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_i + \varepsilon_i \end{aligned}$$

- We are interested in estimating $E\left(Y_i \mid X_i\right)$ from the data
- One reasonable model to do this is SLR
- Hence why the interpretation of β_0 etc is the expected value or population mean.

Estimated SLR model for life expectancy / TFR

$$Y_i = 89.2 - 5.35X_i + \varepsilon_i$$

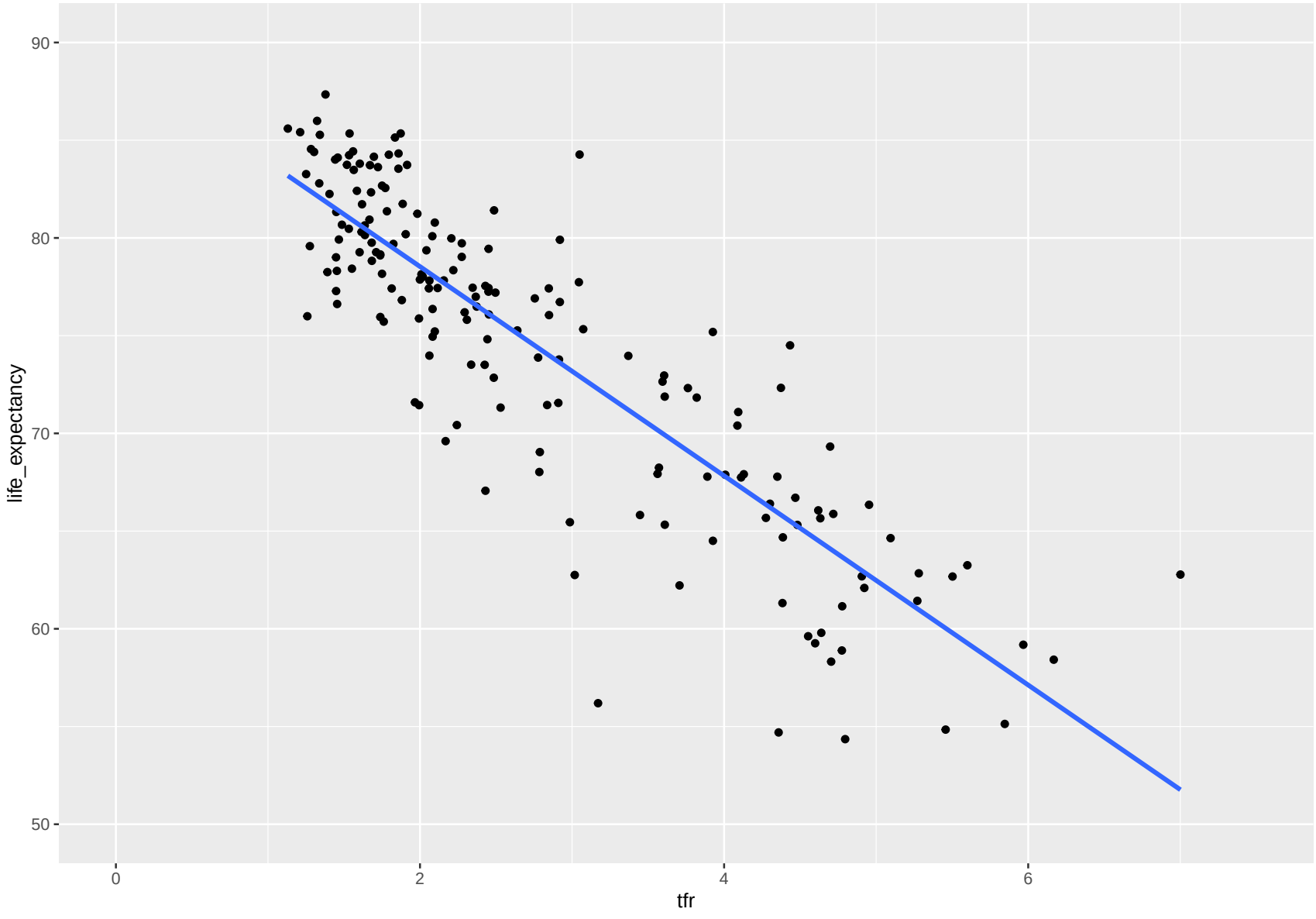
- $\hat{\beta}_0 = 89.2$
- $\hat{\beta}_1 = -5.35$

Notice that the regression coefficients get little hats!

Notation:

- β_0, β_1 are estimands (parameters of interest)
- $\hat{\beta}_0$, $\hat{\beta}_1$ are estimators (functions/methods of getting a value of the parameters)
- $\hat{\beta}_0 = 89.2$ and $\hat{\beta}_1 = -5.35$ are estimates (values calculated from observed data)

Interpretation of the SLR model



Why is it called regression?

- The term regression was first used by Francis Galton in the 19th century
- Comes from the concept of "regression to the mean"
- If a sample point of a random variable is extreme, a future point will tend to be closer to the mean
- Galton observed that extreme characteristics (e.g., height) in parents are not passed on completely to their offspring. Rather, the characteristics in the offspring regress towards the mean.

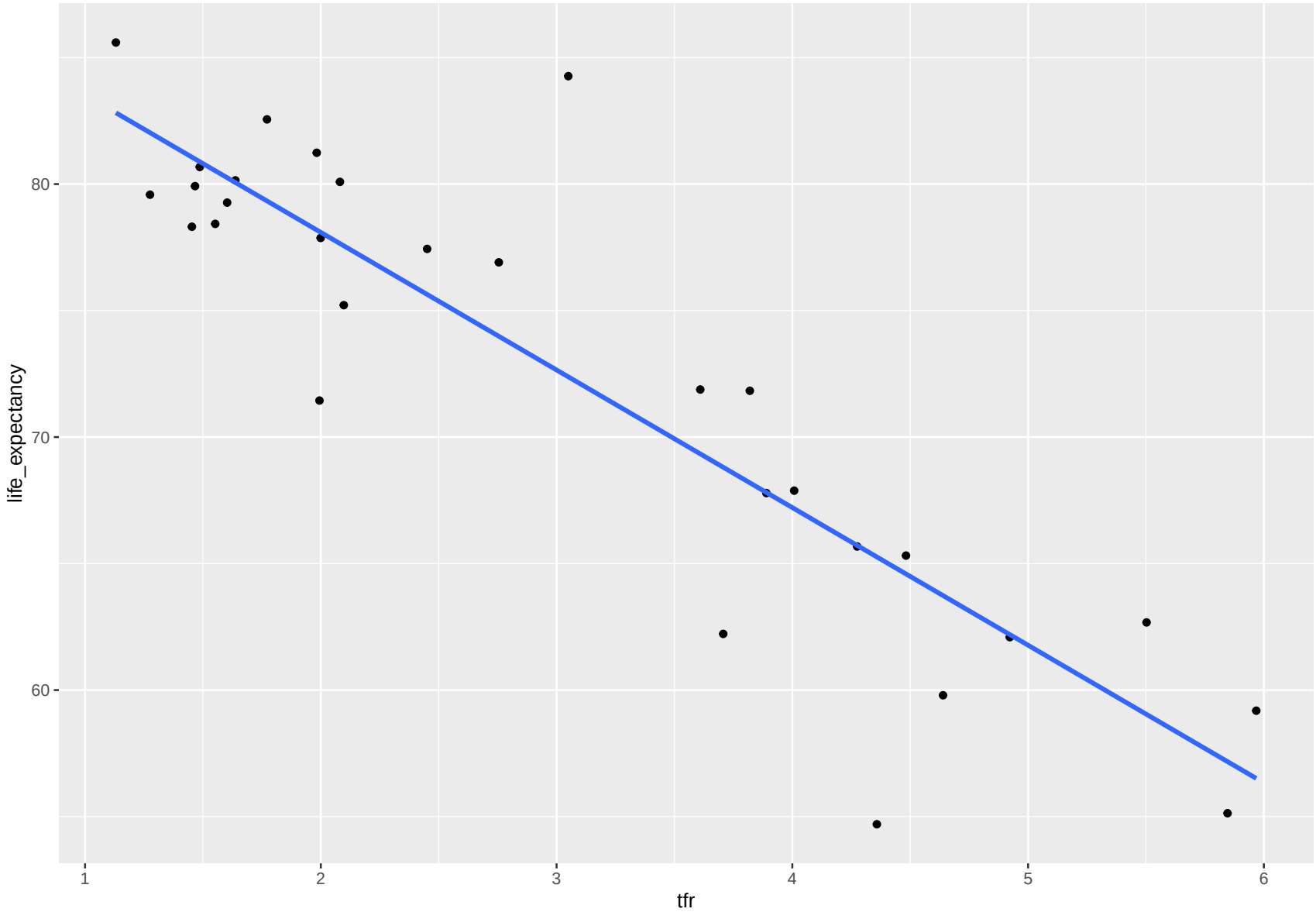
Estimation and the method of least squares

Estimating the regression coefficients

$$Y_i = \beta_0 + \beta_1 X_i$$

- In order to estimate our regression coefficients, we want to find the **line of best fit**
- Which line, of all possible lines, results in the least amount of difference between the observed data points and the line
- Looking at the **residuals** (vertical distances) between what the model predicted and each data point

Draw on the residuals



Minimizing the residuals

- Positive and negative residuals tend to cancel each other out, so we look at the squared residuals

$$(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

- We want to find the line results in the lowest sum of squared residuals

$$SSR = \sum_i (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 = \sum_i \hat{\varepsilon}_i^2$$

- **The method of ordinary least squares (OLS)** finds the line that has the lowest sum of squared residuals

Method of least squares estimators

- The OLS estimators for the SLR model parameters are:

$$\hat{\beta}_1 = \frac{\sum_i \left(Y_i - \frac{1}{n}\sum_i Y_i\right)\left(X_i - \frac{1}{n}\sum_i X_i\right)}{\sum_i \left(X_i - \frac{1}{n}\sum_i X_i\right)^2} = \frac{\sum_i \left(Y_i - \bar{Y}_i\right)\left(X_i - \bar{X}_i\right)}{\sum_i \left(X_i - \bar{X}_i\right)^2}$$

$$\hat{\beta}_0 = \frac{1}{n}\sum_i Y_i - \widehat{\beta}_1 \left(\frac{1}{n}\sum_i X_i\right) = \bar{Y}_i - \widehat{\beta}_1 \bar{X}_i$$

OLS estimation of the SLR model

Sample of 4 points only

Country	X (TFR)	Y (life expectancy)
1	1.6	80.6
2	2.2	69.6
3	5.0	66.3
4	2.0	79.4

$$\hat{\beta}_1 = \frac{\sum_i \left(Y_i - \bar{Y}_i\right)\left(X_i - \bar{X}_i\right)}{\sum_i \left(X_i - \bar{X}_i\right)^2}$$

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

What components do we need to calculate?

OLS estimation of the SLR model

Country	X (TFR)	Y (life expectancy)
1	1.6	80.6
2	2.2	69.6
3	5.0	66.3
4	2.0	79.4

- $\bar{X}_i = ?$
- $\bar{Y}_i = ?$

$$\hat{\beta}_1 = \frac{\sum_i \left(Y_i - \bar{Y}_i\right)\left(X_i - \bar{X}_i\right)}{\sum_i \left(X_i - \bar{X}_i\right)^2}$$

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

OLS estimation of the SLR model

Country	X	Y	X - Xbar	Y - Ybar	(X - Xbar)^2	(X - Xbar)(Y - Ybar)
1	1.6	80.6	-1.1	6.6	1.2	-7.3
2	2.2	69.6	-0.5	-4.4	0.2	2.2
3	5.0	66.3	2.3	-7.7	5.3	-17.7
4	2.0	79.4	-0.7	5.4	0.5	-3.8

- $\sum_i \left(Y_i - \bar{Y}_i\right)\left(X_i - \bar{X}_i\right) = ?$
- $\sum_i \left(X_i - \bar{X}_i\right)^2 = ?$

$$\hat{\beta}_1 = \frac{\sum_i \left(Y_i - \bar{Y}_i\right)\left(X_i - \bar{X}_i\right)}{\sum_i \left(X_i - \bar{X}_i\right)^2}$$

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

OLS estimation of the SLR model

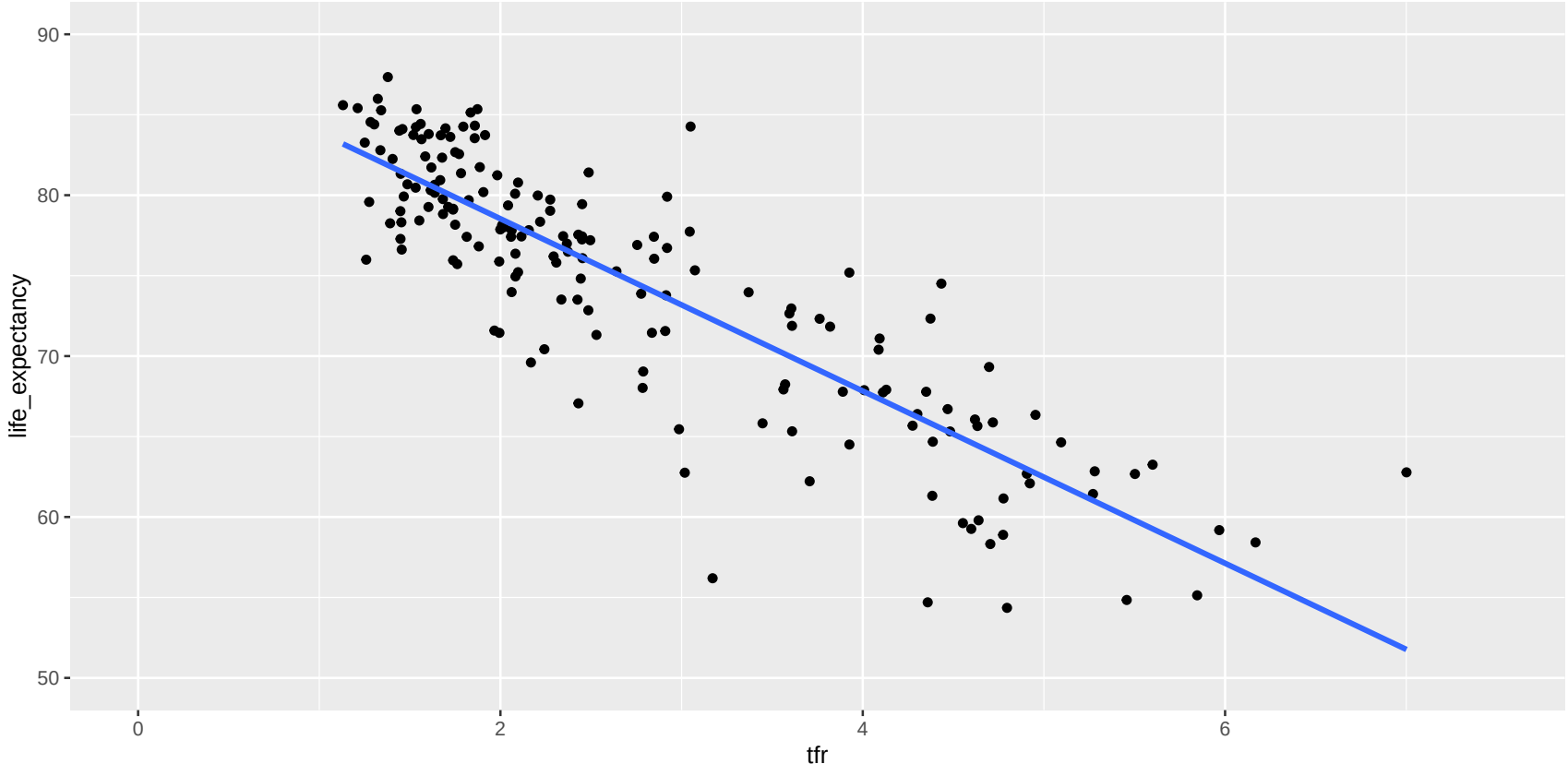
Country	X	Y	X - Xbar	Y - Ybar	(X - Xbar)^2	(X - Xbar)(Y - Ybar)
1	1.6	80.6	-1.1	6.6	1.2	-7.3
2	2.2	69.6	-0.5	-4.4	0.2	2.2
3	5.0	66.3	2.3	-7.7	5.3	-17.7
4	2.0	79.4	-0.7	5.4	0.5	-3.8

- $\hat{\beta}_1 = ?$
- $\hat{\beta}_0 = ?$

$$\hat{\beta}_1 = \frac{\sum_i \left(Y_i - \bar{Y}_i\right)\left(X_i - \bar{X}_i\right)}{\sum_i \left(X_i - \bar{X}_i\right)^2}$$

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

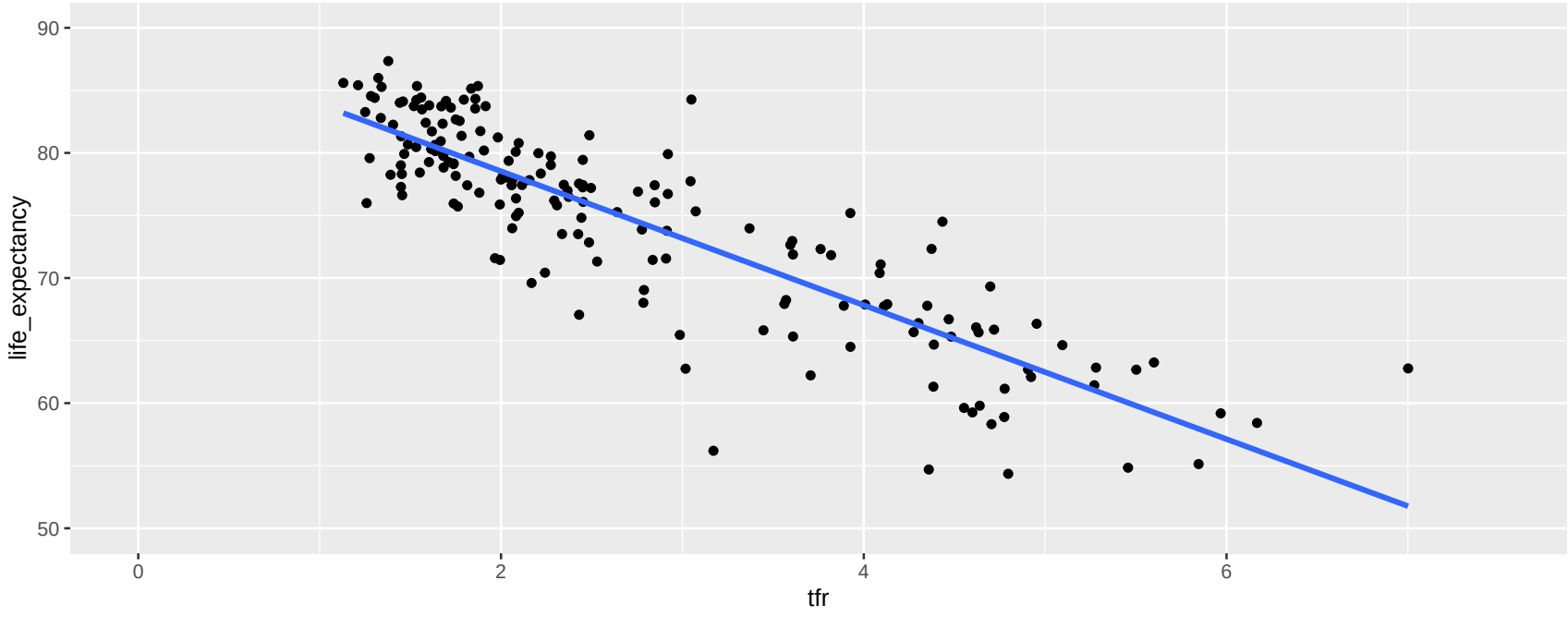
Back to full country dataset



$$Y_i = 89.2 - 5.35X_i + \varepsilon_i$$

- $\hat{\beta}_0 = 89.2$
- $\hat{\beta}_1 = -5.35$

TFR and life expectancy



Focus on one data point: Niger ($i = 176$)

$$Y_{i=176} = \hat{E}(Y_i|X_i = 7) + \hat{\varepsilon}_{i=176}$$

$$Y_{i=176} = \hat{Y}_{i=176} + \hat{\varepsilon}_{i=176}$$

$$Y_{i=176} = 89.2 + 7 \times -5.35 + 11.05$$

$$\textit{observed} = \textit{estimated} + \textit{residual}$$

OLS estimation of the SLR model

Back to using 4 sample points only

Country	X	Y	Yhat
1	1.6	80.6	78.08
2	2.2	69.6	75.86
3	5.0	66.3	65.50
4	2.0	79.4	76.60

$$\hat{E}\left(Y_i \mid X_i\right) = \widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 84 - 3.7 X_i$$

OLS estimation of the SLR model

Country	X	Y	Yhat	ehat
1	1.6	80.6	78.08	2.52
2	2.2	69.6	75.86	-6.26
3	5.0	66.3	65.50	0.80
4	2.0	79.4	76.60	2.80

$$\hat{E}\left(Y_i \mid X_i\right) = \widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 84 - 3.7 X_i$$

$$\hat{\varepsilon} = Y_i - \widehat{E}\left(Y_i \mid X_i\right) = Y_i - \widehat{Y}_i = Y_i - (84 - 3.7 X_i)$$

SLR in R

```
country_ind_2017 <- country_ind |>
  filter(year==2017)
mod2 <- lm(life_expectancy~tfr, data = country_ind_2017)
summary(mod2)
```

```
##
## Call:
## lm(formula = life_expectancy ~ tfr, data = country_ind_2017)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0718  -2.3864   0.3132   2.6537  11.3498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  89.2394     0.7085  125.95   <2e-16 ***
## tfr          -5.3526     0.2326  -23.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.994 on 174 degrees of freedom
## Multiple R-squared:  0.7527, Adjusted R-squared:  0.7513
## F-statistic: 529.7 on 1 and 174 DF,  p-value: < 2.2e-16
```

Summary

- We are interested in the relationships between two variables
- On average, if X changes, how much to we expect Y to change?
- Simple linear regression gives us the tools to study relationships between two variables, accounting for uncertainty in data

How much variation does our model explain: R^2

Thinking about variation

- So far we've been mostly concerned about conditional expectations, that is, population means for different subgroups/populations of different characteristics
- Let's think about variation in Y_i around measures of central tendency for a moment

What sorts of variation may we be interested in?

- Variation of data Y_i around the observed mean $\bar{Y}_i$
- Variation of fitted values $\hat{Y}_i$ around observed mean $\bar{Y}_i$
- Variation of data Y_i around fitted values $\hat{Y}_i$

Sums of squares

- Variation of data Y_i around the observed mean $\bar{Y}_i$
  - Total sum of squares SST: $(Y_i - \bar{Y}_i)^2$
- Variation of fitted values $\hat{Y}_i$ around observed mean $\bar{Y}_i$
  - Model sum of squares SSM: $(\hat{Y}_i - \bar{Y}_i)^2$
- Variation of data Y_i around fitted values $\hat{Y}_i$
  - Residual sum of squares SSR: $(Y_i - \hat{Y}_i)^2$

Sums of squares

- Variation of data Y_i around the observed mean $\bar{Y}_i$
  - Total sum of squares SST: $(Y_i - \bar{Y}_i)^2$
  - Total variation in Y_i
- Variation of fitted values $\hat{Y}_i$ around observed mean $\bar{Y}_i$
  - Model sum of squares SSM: $(\hat{Y}_i - \bar{Y}_i)^2$
  - Variation explained by our X's
- Variation of data Y_i around fitted values $\hat{Y}_i$
  - Residual sum of squares SSR: $(Y_i - \hat{Y}_i)^2$
  - Variation not explained by X's

$$SST = SSM + SSR$$

R^2

$$SST = SSM + SSR$$

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSR}{SST}$$

The proportion of total variation in Y_i explained by covariates X_i.

Simple Linear Regression with a categorical covariate

Running example

Using GSS data

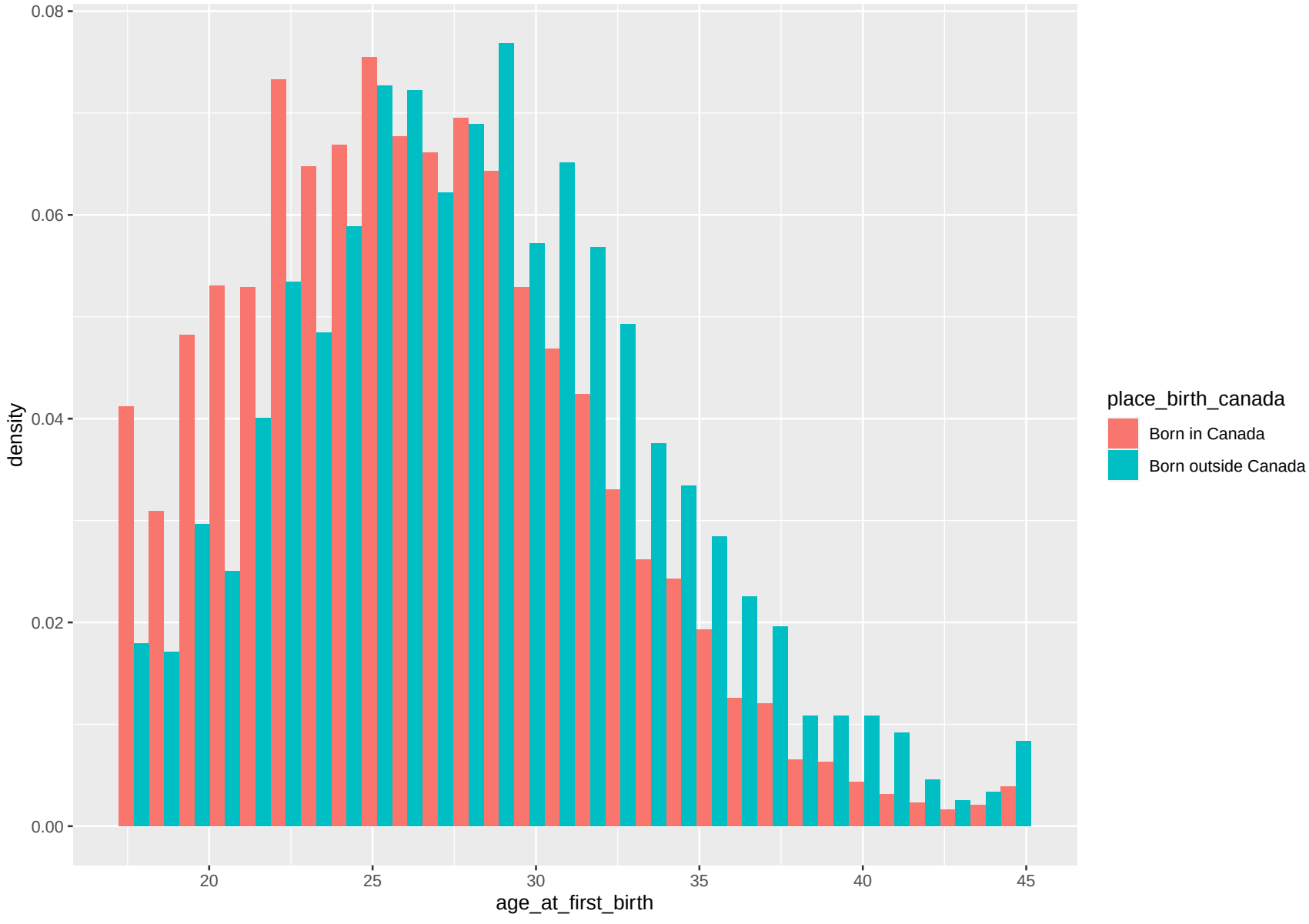
- Question: are people born outside of Canada more likely to start having children later compared to those born in Canada?

Variables:

- Age at first birth
- Place of birth (Canada, outside Canada)

How could we explore this graphically? Or with summary stats?

Looking at conditional distributions



Estimated SLR model for age at first birth and place of birth

In this case, the control for X_i is born in Canada.

$$Y_i = 26.5 + 1.82X_i + \varepsilon_i$$

- $\hat{\beta}_0 = 26.5$
- $\hat{\beta}_1 = 1.82$

Interpretation of $\hat{\beta}_1$ is **compared to the baseline**. In this case, it is the difference in age of first birth of people born outside Canada compared to born in Canada.

SLR in R

```
# filter out the don't knows
gss <- gss %>% filter(place_birth_canada!="Don't know")
# run the regression
mod <- lm(age_at_first_birth ~ place_birth_canada, data = gss)
```

SLR in R

```
summary(mod)
```

```
##
## Call:
## lm(formula = age_at_first_birth ~ place_birth_canada, data = gss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3113  -4.0060  -0.4113   3.4097  18.5097
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         26.49032    0.05362  494.05   <2e-16 ***
## place_birth_canadaBorn outside Canada  1.82096    0.11773   15.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.37 on 12652 degrees of freedom
##   (7798 observations deleted due to missingness)
## Multiple R-squared:  0.01856,    Adjusted R-squared:  0.01848
## F-statistic: 239.3 on 1 and 12652 DF,  p-value: < 2.2e-16
```

More than one category

Baseline is Atlantic region

```
##
## Call:
## lm(formula = age_at_first_birth ~ region, data = gss)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.622 -4.078 -0.422  3.393 19.122
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              25.87768    0.09949 260.095  < 2e-16 ***
## regionBritish Columbia    1.66734    0.17168   9.712  < 2e-16 ***
## regionOntario             1.74428    0.13531  12.891  < 2e-16 ***
## regionPrairie region      0.58190    0.14652   3.971 7.19e-05 ***
## regionQuebec              1.12922    0.14956   7.550 4.64e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.377 on 12649 degrees of freedom
##   (7798 observations deleted due to missingness)
## Multiple R-squared:  0.01603,    Adjusted R-squared:  0.01572
## F-statistic: 51.51 on 4 and 12649 DF,  p-value: < 2.2e-16
```

Summary

- In linear regression, the dependent variable has to be continuous
- But the covariate/ indepedent variable does not have to be
- If the covariate is categorial, the interpretation of the coefficient β_1 becomes a comparison
- Can be a binary covariate (i.e. two categories) but extends to more categories as well
- In R, `lm` automatically drops one of the categories (by default, this is done alphabetically)