SOC6707 Intermediate Data Analysis

Monica Alexander

Week 3: Linear Regression

Announcements

- Assignment 2 out soon (on linear regression)

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?
  - If we consider X_2 as well, does Y still vary with 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

Using GSS data

- Question: are people born outside of Canada more likely to start having children later compared to those born in Canada?
- Does the answer to this question persist after we take into account education?

Variables:

- Age at first birth
- Place of birth (Canada, outside Canada)
- Bachelor or higher (yes/no)

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

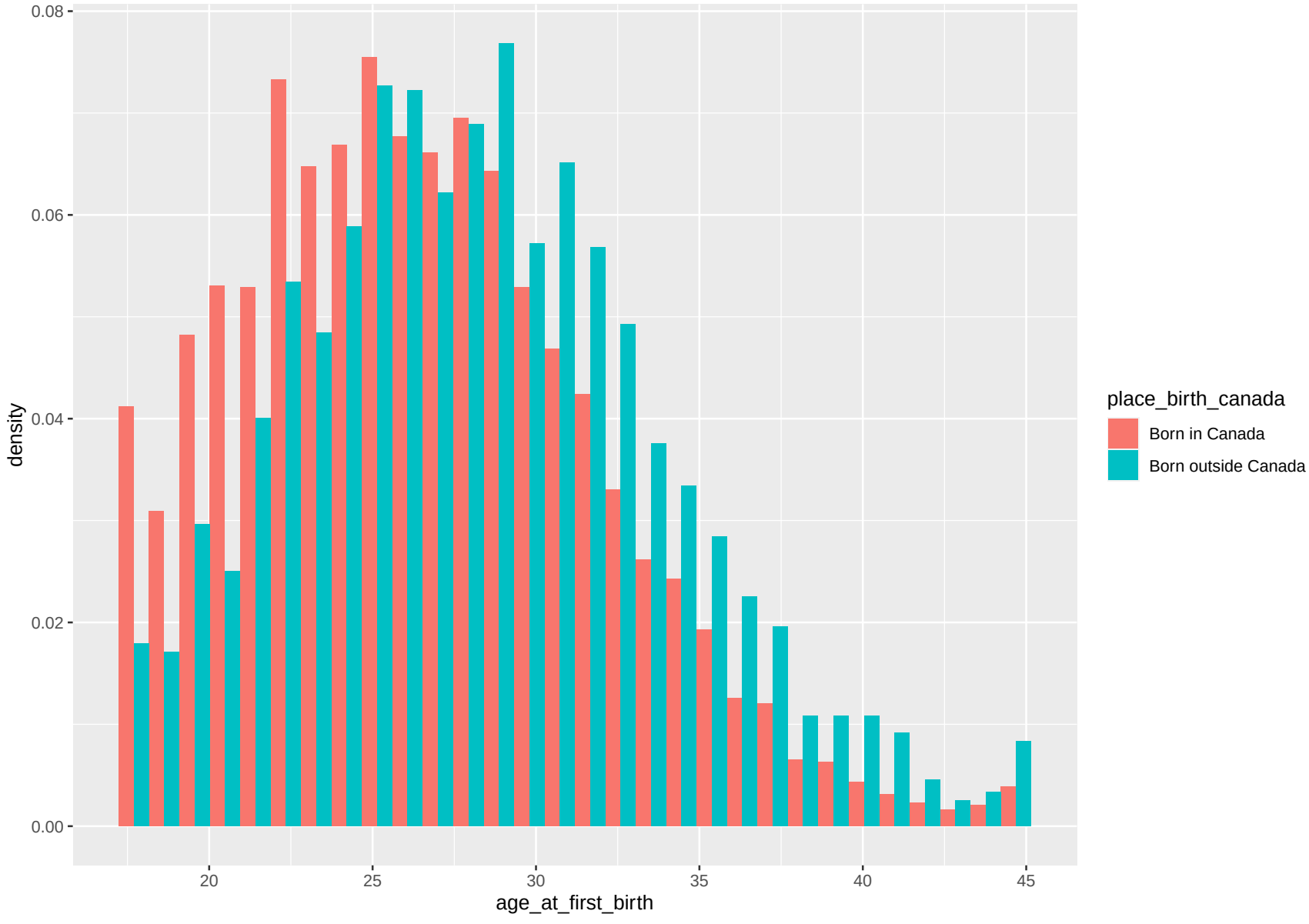
Summary stats

place of birth	n	mean	SD
Born in Canada	10029	26.49	5.32
Born outside Canada	2625	28.31	5.56

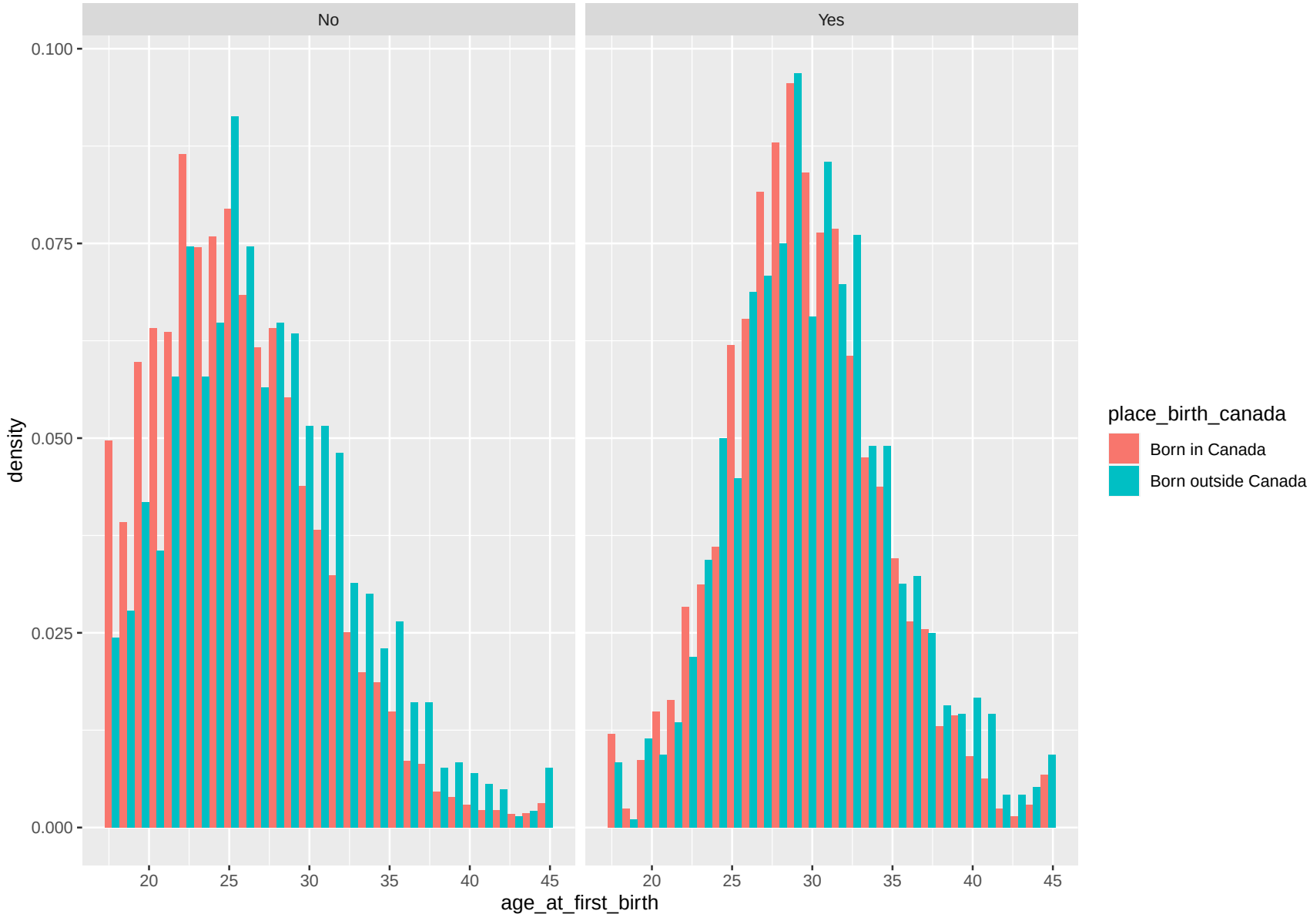
Or with more info:

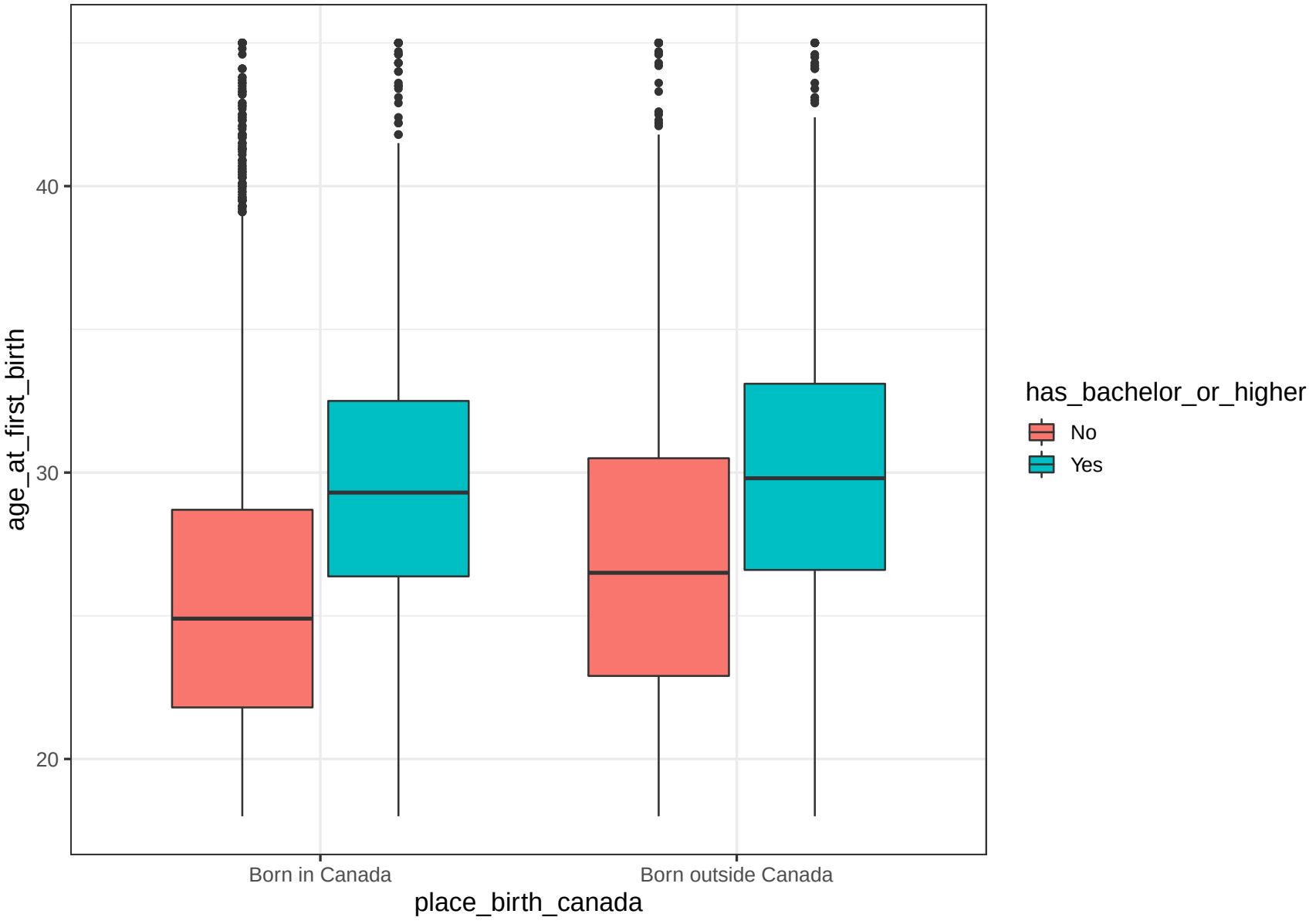
place of birth	educ	n	mean	SD
Born in Canada	No	7669	25.60	5.11
Born in Canada	Yes	2236	29.57	4.83
Born outside Canada	No	1541	27.13	5.51
Born outside Canada	Yes	1031	30.11	5.11

Looking at distributions



Looking at distributions




age_at_first_birth
40
30
20
Born in Canada
Born outside Canada
place_birth_canada
has_bachelor_or_higher
No
Yes

The story so far

Our question:

- are people born outside of Canada more likely to start having children later compared to those born in Canada?
- Does the answer to this question persist after we take into account education?

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

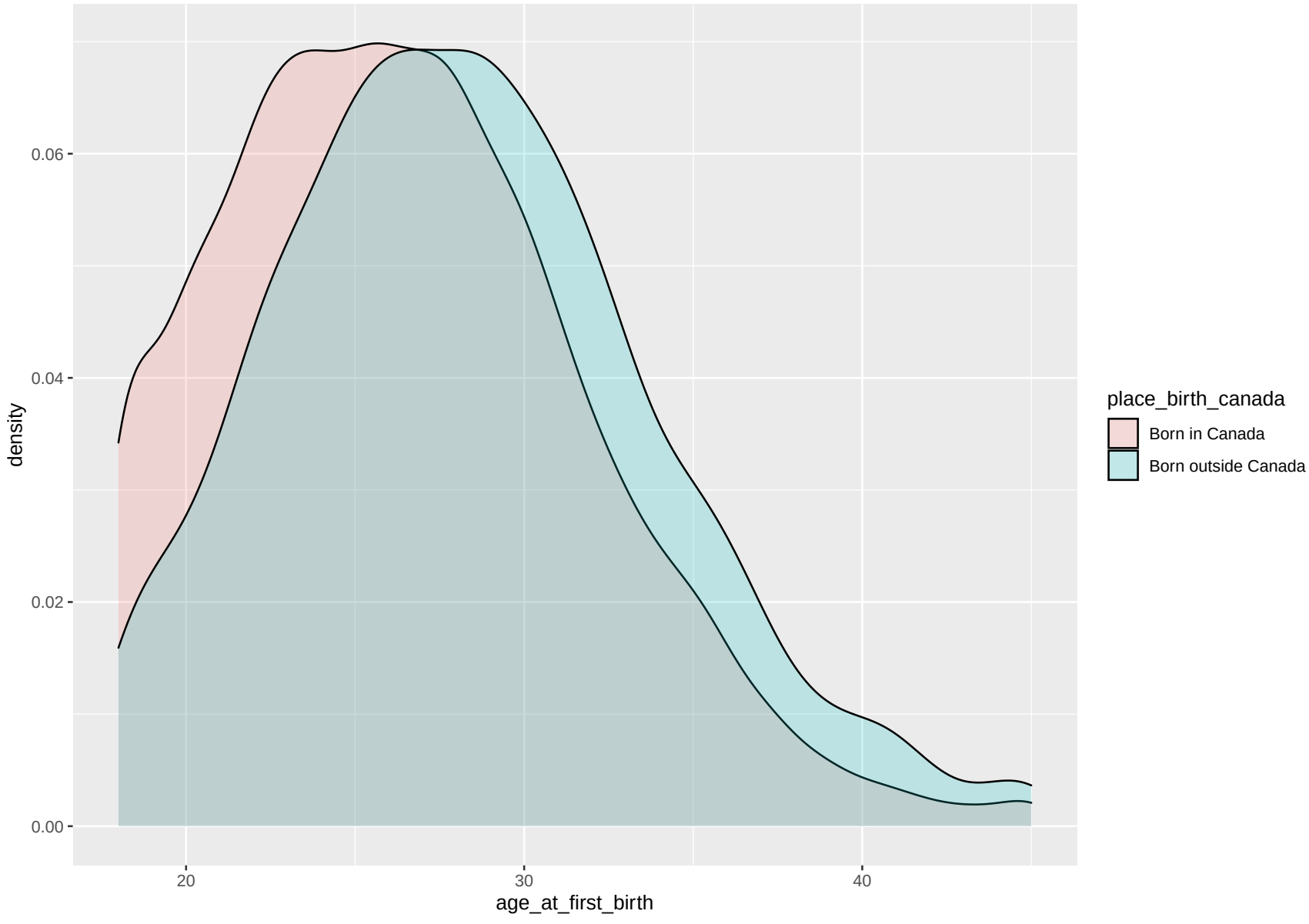
- Ideally, want to account for uncertainty in the data we observe and make some statements about how sure (or otherwise) we are that the outcome of interest (age at first marriage) differs by covariates of interest (place of birth, education)
- Linear regression is a model that allows us to do this

Side note: Conditioning

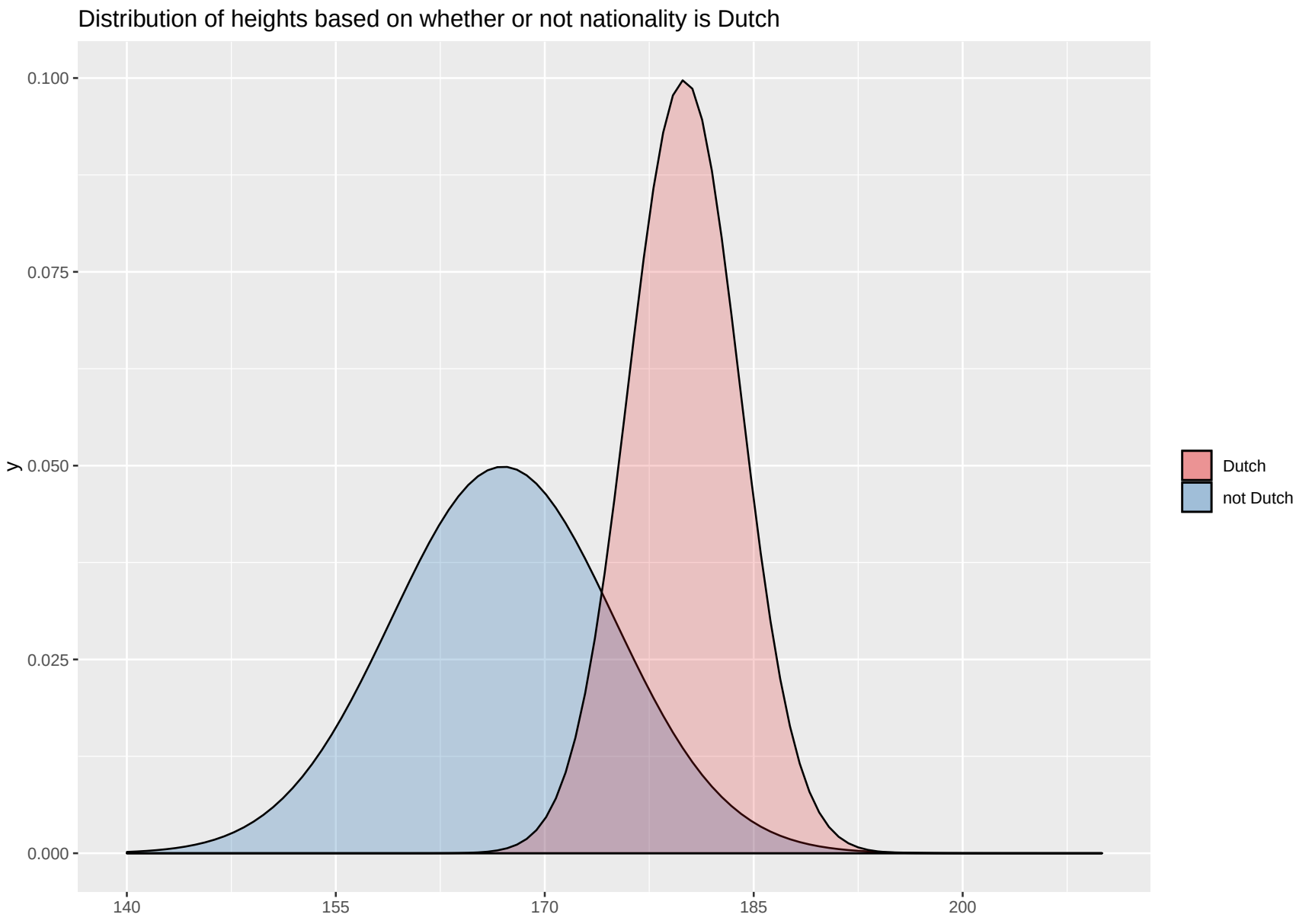
Conditioning on covariates

- We are considering our outcome of interest (age at first birth) by groups of interest (e.g. place of birth)
- That is, we are conditioning on place of birth and considering different characteristics of our outcome
- For example, plotting histograms by group is looking at (an estimate of) the **conditional distribution** of age at first birth

Conditional distributions of age at first birth



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 the expectation of age at first birth, conditioning on place of birth
- $E(\text{age at first birth} \mid \text{place of birth}) = E(Y_i|X_i = x_i)$ 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

- Question: are people born outside of Canada more likely to start having children later compared to those born in Canada?
- Does the answer to this question persist after we take into account education?

Initially, let's focus on the first question. We have one outcome and one covariate of interest. Introducing some notation:

- Y_i is the response variable (dependent variable)
- X_i is the explanatory variable (covariate)

Questions:

- In our example, what is Y and what is X?
- In our example, what does 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

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

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

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

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

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)
- $\hat{\beta}_0 = 26.5$ and $\hat{\beta}_1 = 1.82$ are estimates (values calculated from observed data)

Interpretation

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

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.

Fancy averages

- Conditional expectations are just fancy averages
- Multiple linear regression is a model to estimate conditional expectations
- So regression is just estimating fancy averages

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
```

Multiple linear regression

Back to our example

- Question: are people born outside of Canada more likely to start having children later compared to those born in Canada?
- Does the answer to this question persist after we take into account education?

Now let's consider the second part of our question, by including another variable in our regression model.

- Y_i is the dependent variable or response variable
- X_{i1} and X_{i2} are the independent variables, explanatory variables or predictors

We have

- Y_i is age at first birth
- X_{i1} is place of birth (in Canada or outside Canada)
- X_{i2} is education (has bachelor + or not)

MLR model

With two covariates, the MLR model is

$$\begin{aligned} Y_i &= E\left(Y_i \mid X_{i1}, X_{i2}\right) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \end{aligned}$$

Specifically, the most basic MLR model is a simple linear function of X_{i1} and X_{i2}, and three parameters, β_0, β_1 and β_2.

Interpretation

The MLR model: $E(Y_i \mid X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$

- What is β_0?
- β_0 is the is the expected value, or population mean, of Y_i given both X_{i1} and X_{i2} equal zero.

Interpretation

The MLR model: $E(Y_i \mid X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$

- What is β_1?
- β_1 is the change in the expected value, or population mean, of Y_i associated with a one unit increase in X_{i1}, **holding** X_{i2} **constant at any value**

Same idea for β_2.

Interpretation

The MLR model: $E(Y_i \mid X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$

- In general $\beta_1 (x_1^* - x_1)$ is the change in the expected value of Y_i associated with a $(x_1^* - x_1)$ change in X_{i1}, holding X_{i2} constant
- $\beta_2 (x_2^* - x_2)$ is the change in the expected value of Y_i associated with a $(x_2^* - x_2)$ change in X_{i2}, holding X_{i1} constant

Our example

```
mod2 <- lm(age_at_first_birth ~ place_birth_canada+ has_bachelor_or_higher, data = gss)
summary(mod2)
```

```
##
## Call:
## lm(formula = age_at_first_birth ~ place_birth_canada + has_bachelor_or_higher,
##     data = gss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5503  -3.7545  -0.5503   3.1455  19.3455
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        25.65453    0.05665  452.86   <2e-16 ***
## place_birth_canadaBorn outside Canada  1.17891    0.11470   10.28   <2e-16 ***
## has_bachelor_or_higherYes           3.71687    0.10554   35.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.115 on 12474 degrees of freedom
##   (7975 observations deleted due to missingness)
## Multiple R-squared:  0.1074, Adjusted R-squared:  0.1073
## F-statistic: 750.8 on 2 and 12474 DF,  p-value: < 2.2e-16
```

Interpretation

$$Y_i = 25.7 + 1.18X_{i1} + 3.71X_{i2}$$

What else can we do with regression?

- So far we've been interpreting coefficients: inferring relationships between variables
- We can also look at the **fitted values** of our outcome of interest to see how well our model represent observed data Y_i
- We can also **predict new values** for new individuals with particular characteristics (more later)

Fitted values

Fitted values get the symbol $\hat{Y}_i$: a hat because they're estimated. For example, let's look at the second respondent:

```
gss %>%
  drop_na(age_at_first_birth) %>%
  select(age_at_first_birth,
         place_birth_canada,
         has_bachelor_or_higher) %>%
  slice(2) # get second row
```

```
## # A tibble: 1 x 3
##   age_at_first_birth place_birth_canada has_bachelor_or_higher
##                <dbl> <chr>              <chr>
## 1               23.2 Born in Canada     Yes
```

So $Y_2 = 23.2$. What is $\hat{Y}_2$?

Fitted values

```
coefficients <- coef(mod2)
# have a look at coefficients
coefficients
```

```
##                             (Intercept) place_birth_canadaBorn outside Canada
##                               25.654527                                1.178915
##             has_bachelor_or_higherYes
##                                3.716866
```

```
# calculate fitted value for third person
yhat_2 <- coefficients[[1]] + coefficients[[2]]*0 + coefficients[[3]]
yhat_2
```

```
## [1] 29.37139
```

$\hat{Y}_2 = 29.4.$

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.