SOC6707 Intermediate Data Analysis

Monica Alexander

Week 3: Linear Regression

Overview

- Random variables, probability essentials
- Conditional expectation function
- Simple linear regression
- Multiple linear regression
- Estimation
- Inference

It is assumed that this is review.

Random variables

A **random variable** is a variable whose values depend on the outcomes of a random process.

Coin toss example

Imagine tossing a coin 4 times. Say we are interested in the number of heads that turns up. The observed outcomes are:

```
## [1] "T" "H" "T" "H"
```

So the number of heads is 2. But we can toss it another 4 times. The second set of observed outcomes are

```
## [1] "T" "T" "H" "T"
```

So the number of heads is 1.

The number of heads is a **random variable** that depends on the random process of flipping a coin.

Heights example

Say we are interested in heights of people in Canada. We take a random sample of 6 people. Their heights are (in cm)

```
## [1] 177.16 169.96 181.95 188.82 175.19 177.94
```

We sample another 6 people. Their heights are

```
## [1] 164.53 180.91 175.36 166.61 188.82 165.57
```

So height is a random variable that depends on the random process of sampling the population

Notation

- Call our random variable of interest X
  - in coin example $X =$ number of heads
  - in heights example $X =$ height
- After we observe values we denote these with lower case x
  - coin example $x = 2$ and $x = 1$
  - heights example $\{x_1 = 177.16, x_2 = 169.96, x_3 = 181.95, x_4 = 188.82, x_5 = 175.19, x_6 = 177.94\}$ etc

Probability distributions

Back to coin flipping example

- The process of tossing a coin four times qualifies as an experiment
- We can observe the outcome of each toss, and the outcome is uncertain.
- Our random variable of interest was the number of heads

First, let's look at possibilities. On the first toss, we could observe an outcome of heads (H) or tails (T). On each of the remaining three tosses, we could observe an H or a T. Thus, the possibilities for four tosses can be enumerated as follows:

- HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, and TTTT.

We can see that there are 16 different possible outcomes when listed as simple events.

Flipping a coin

We can enumerate these possible outcomes in a table with the associated probability and observed number of heads

event	probability	number of heads
HHHH	0.0625	4
HHHT	0.0625	3
HHTH	0.0625	3
HHTT	0.0625	2
HTHH	0.0625	3
HTHT	0.0625	2
HTTH	0.0625	2
HTTT	0.0625	1
THHH	0.0625	3
THHT	0.0625	2
THTH	0.0625	2
THTT	0.0625	1
TTHH	0.0625	2
TTHT	0.0625	1
TTTH	0.0625	1
TTTT	0.0625	0

Probability distribution for the number of heads

Given our RV of interest is the number of heads and that all events are mutually exclusive, we can summarize the table as

Number of heads (X)	P(X)
4	1/16
3	4/16
2	6/16
1	4/16
0	1/16

We have a **probability distribution** for the number of heads. That is, a rule or function that associates the probability of observing that particular value with each value of a random variable. The probability distribution for a **discrete** RV (like # heads) is called a **probability mass function**

The expected value of a random variable

For a discrete random variable, X, with a known probability distribution $P(X_i)$ and where X_i is the ith outcome in the set of k simple events:

$$E(X) = X_1 \times P(X_1) + X_2 \times P(X_2) + \ldots + X_k \times P(X_k) = \sum_{i=1}^{k} X_i \times P(X_i) = \mu$$

The expected value is a weighted mean of all the possible values of the RV, weighted by their probabilities. It is given the symbol μ.

Calculate the expected value for the number of heads in four coin flips.

The variance of a random variable

The definition of expected value to derive the variance, given the probability distribution

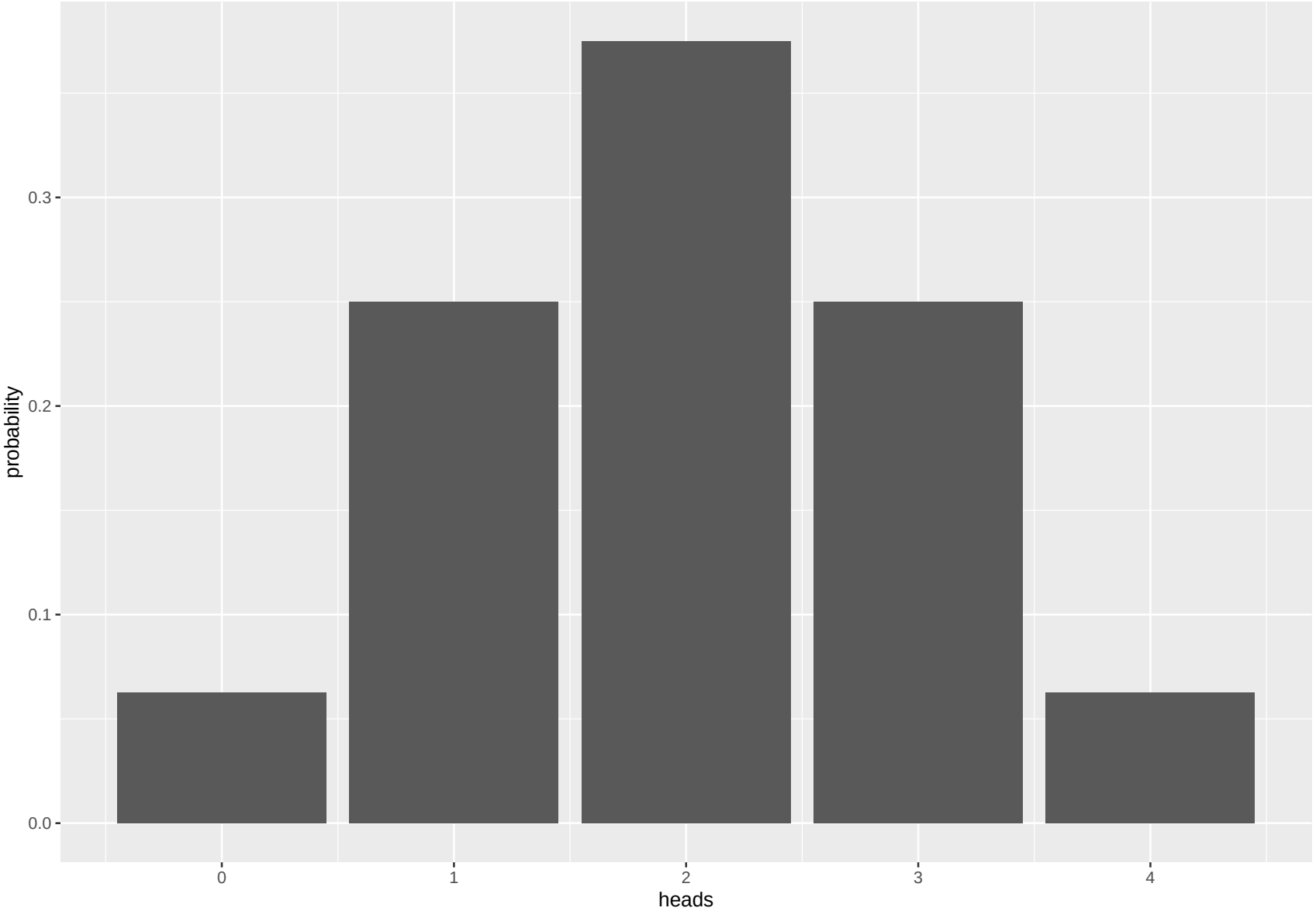
$$\begin{aligned}\sigma^2 &= E\left[(X-\mu)^2\right] \\ &= [X_1 - E(X)]^2 \times P(X_1) + \ldots + [X_k - E(X)]^2 \times P(X_k) \\ &= \sum_{i=1}^{k} [X_i - E(X)]^2 \times P(X_i)\end{aligned}$$

Calculate the variance for the number of heads in four coin flips.

Summary

- If we know the probability distribution of a discrete random variable, we know the mean and variance of the random variable, and hence, the standard deviation of the random variable.
- Thus, we can make predictions about where the values should center and how spread out they should be.
- If the random variable X is a continuous variable, then the idea is the same, but the sums $\sum$ need to be replaced with integrals $\int$ and we would need some calculus.

Probabilities as areas



Probabilities as areas

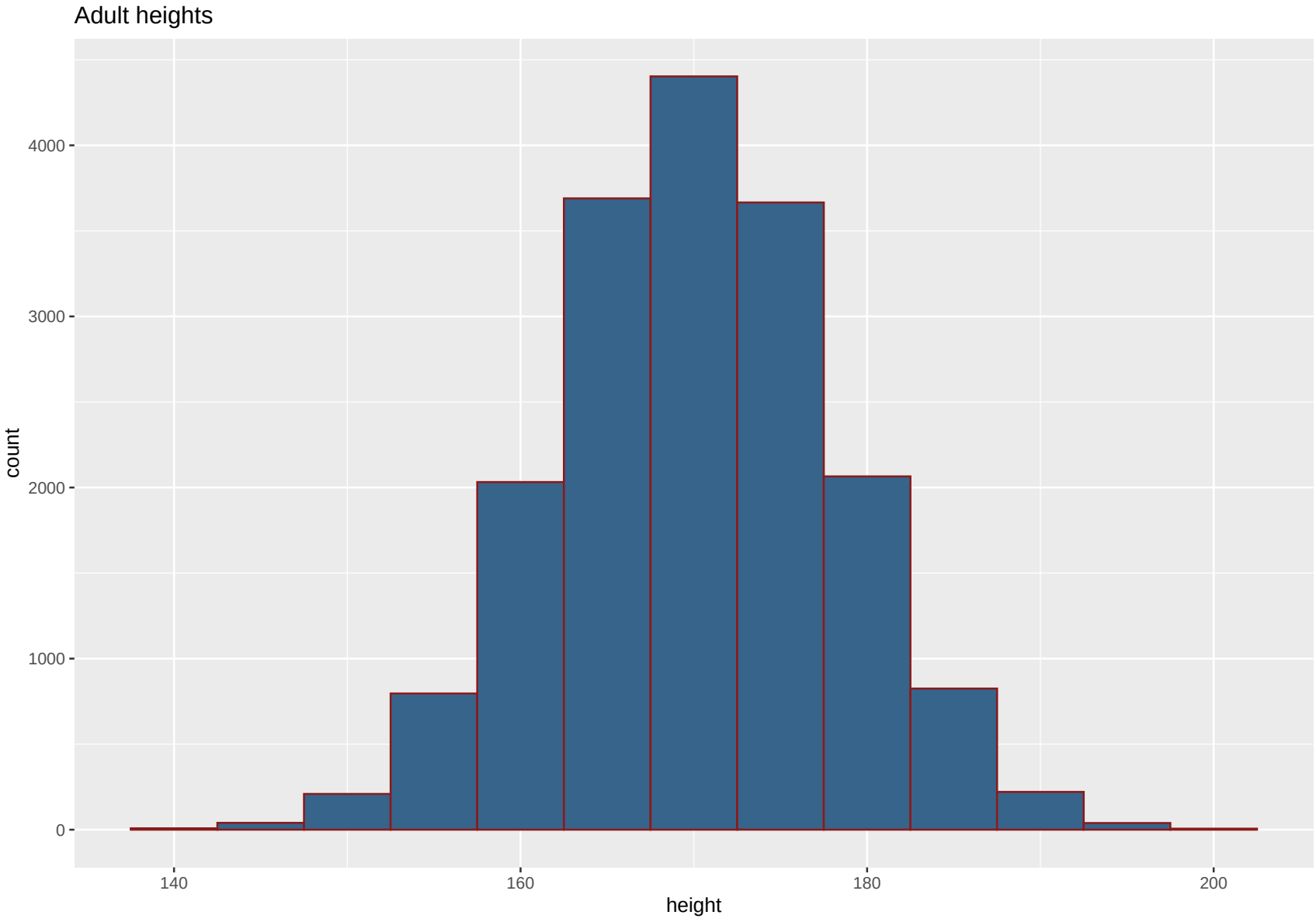
- To calculate probabilities, can sum up the area of the rectangles
- E.g. $P(X \geq 3)$ would be the sum of the right two rectangles
- What is $P(1 \leq X \leq 3)$?
- What is $P(1 \leq X < 3)$?

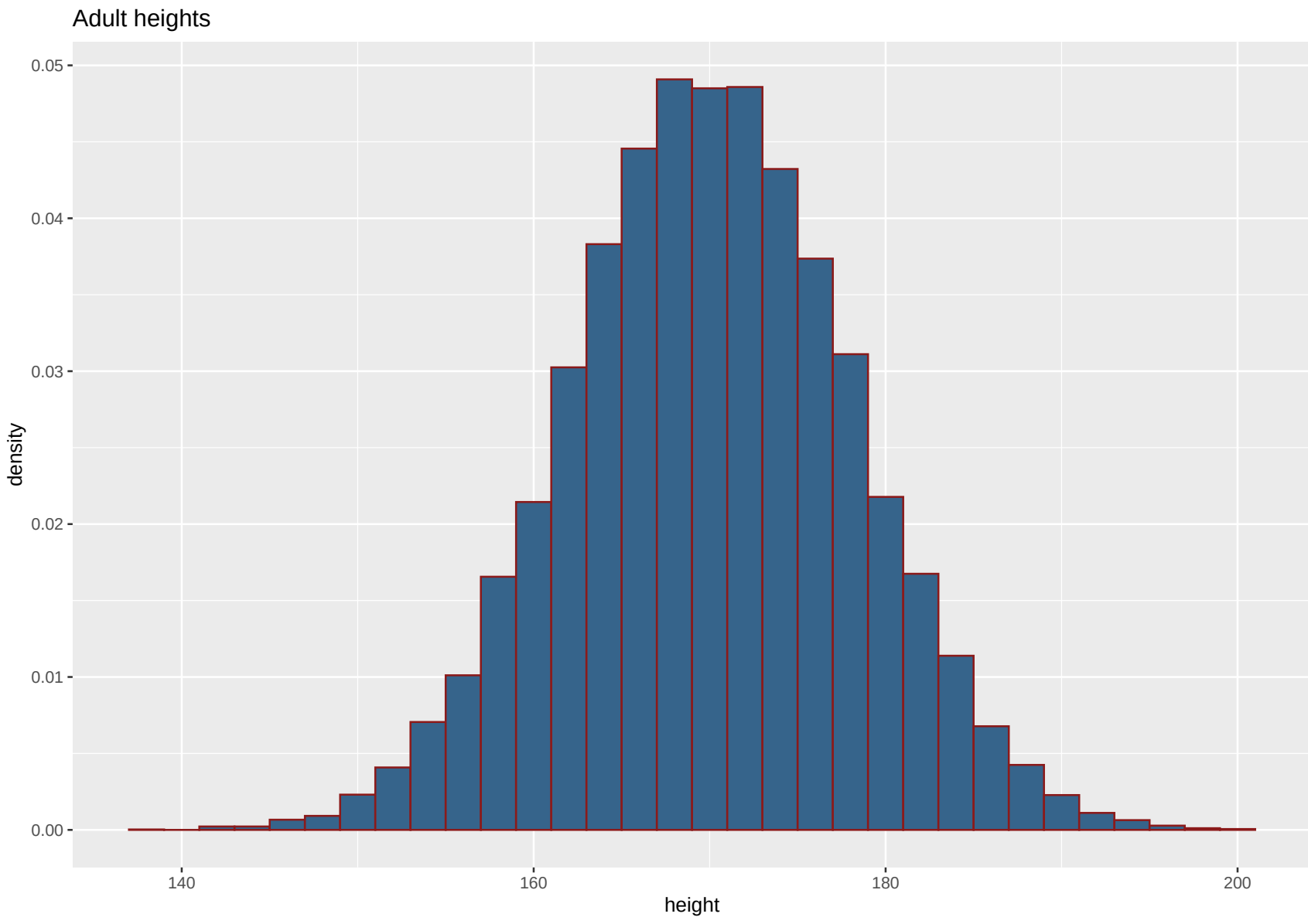
Continuous random variables and probability distributions

What if our variable is continuous?

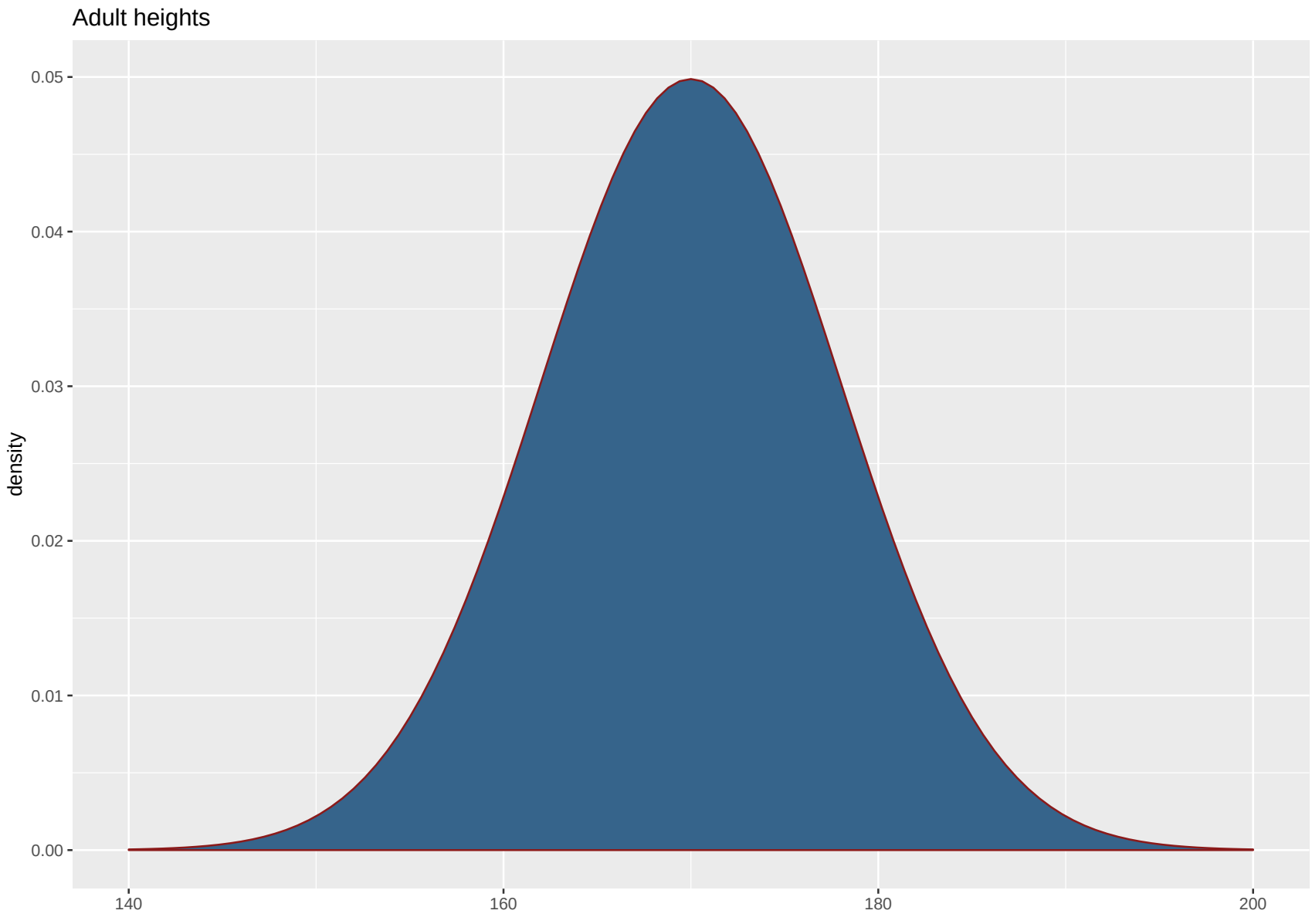
- Can think about in the same way (defining probability distributions, expected values, etc)
- Instead of having a table of values making up the probability distribution (or pmf), we have a mathematically defined function
- A probability distribution for a continuous RV is called a **probability density function**

A continuous probability distribution is just a histogram with infinitely small bins

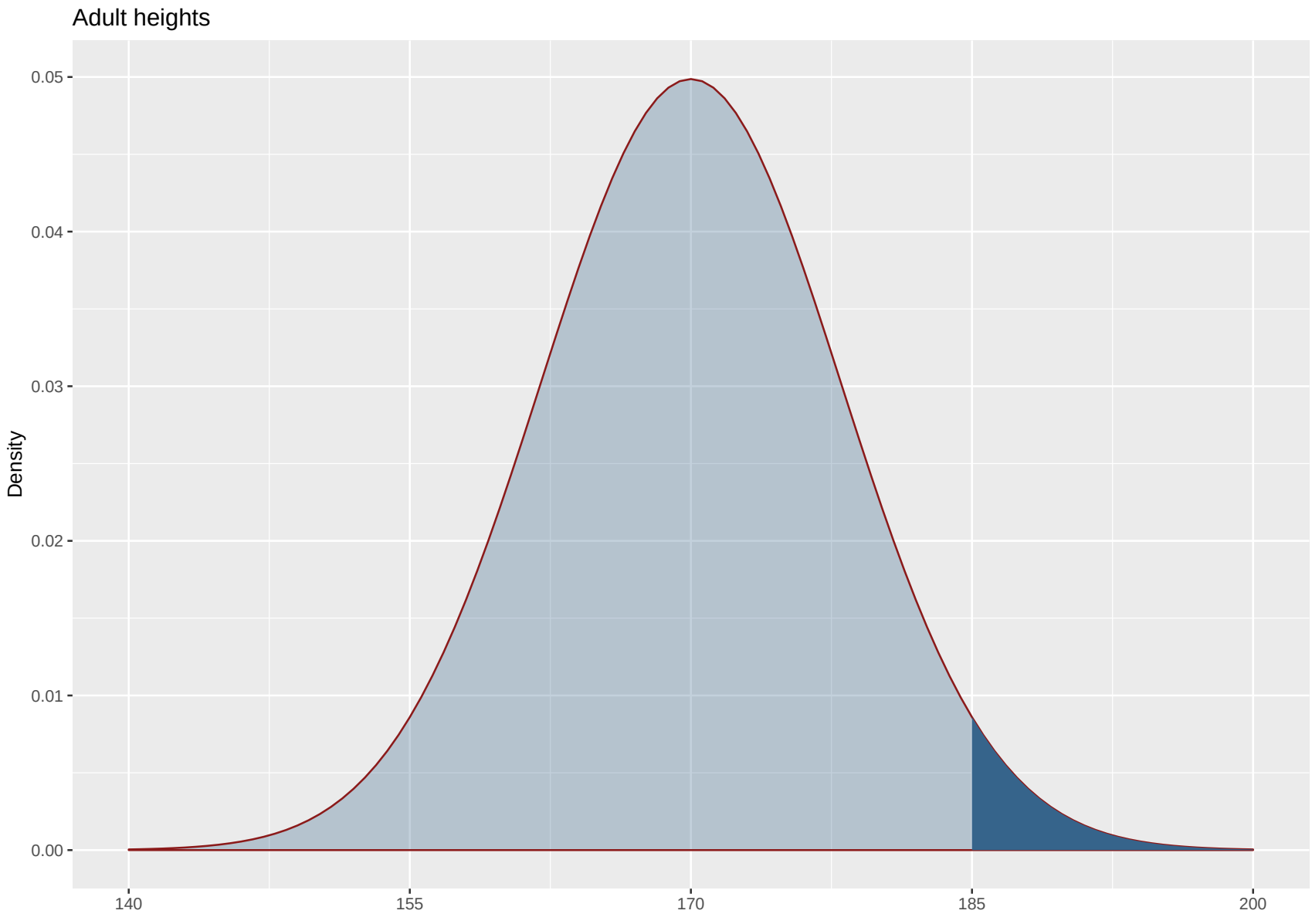



Adult heights
density
0.05
0.04
0.03
0.02
0.01
0.00
140
160
180
200
height

Probability density function



What probability does this represent?



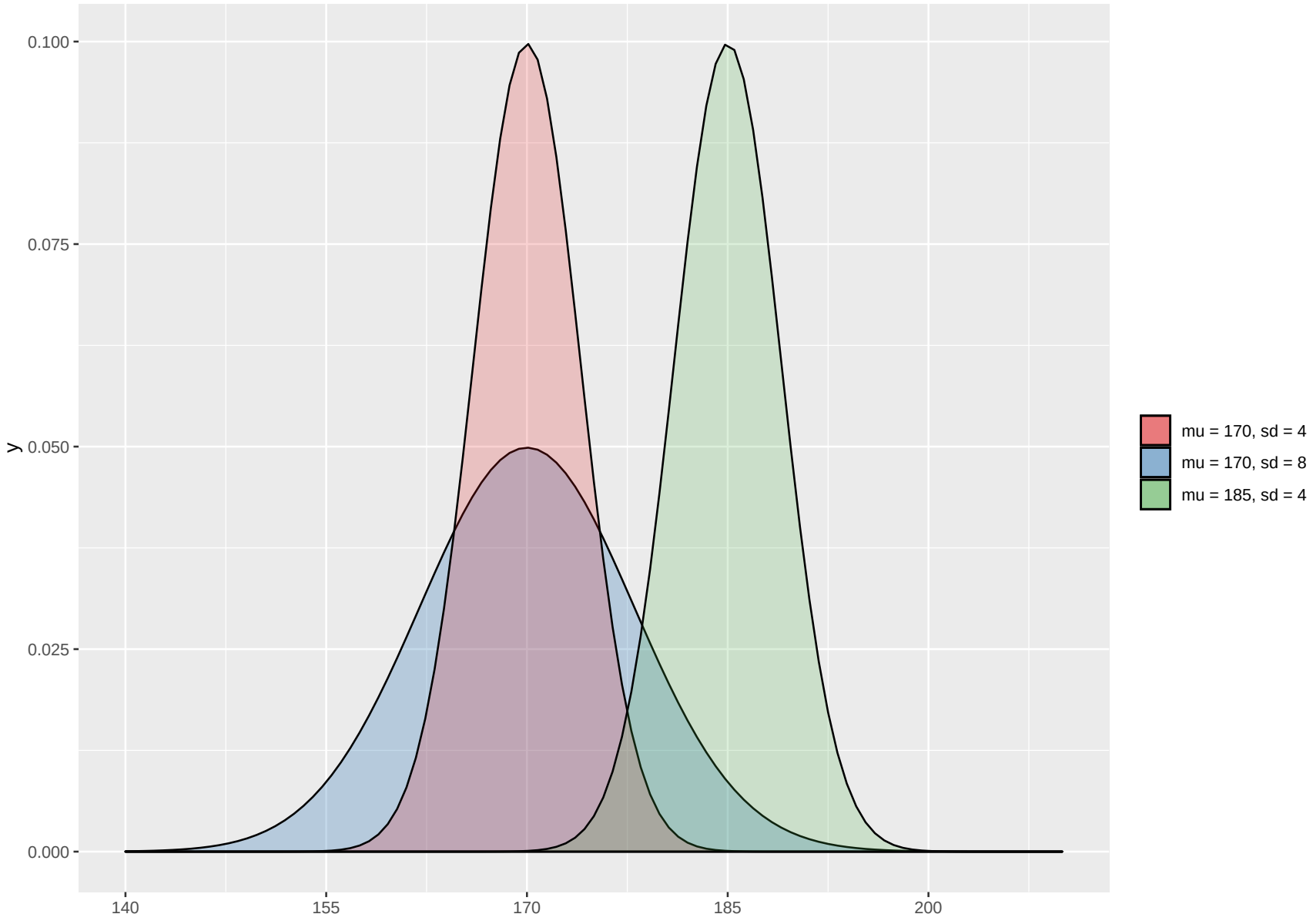
The normal distribution

The normal distribution

- One of the most important continuous probability distributions
- Is described by the formula

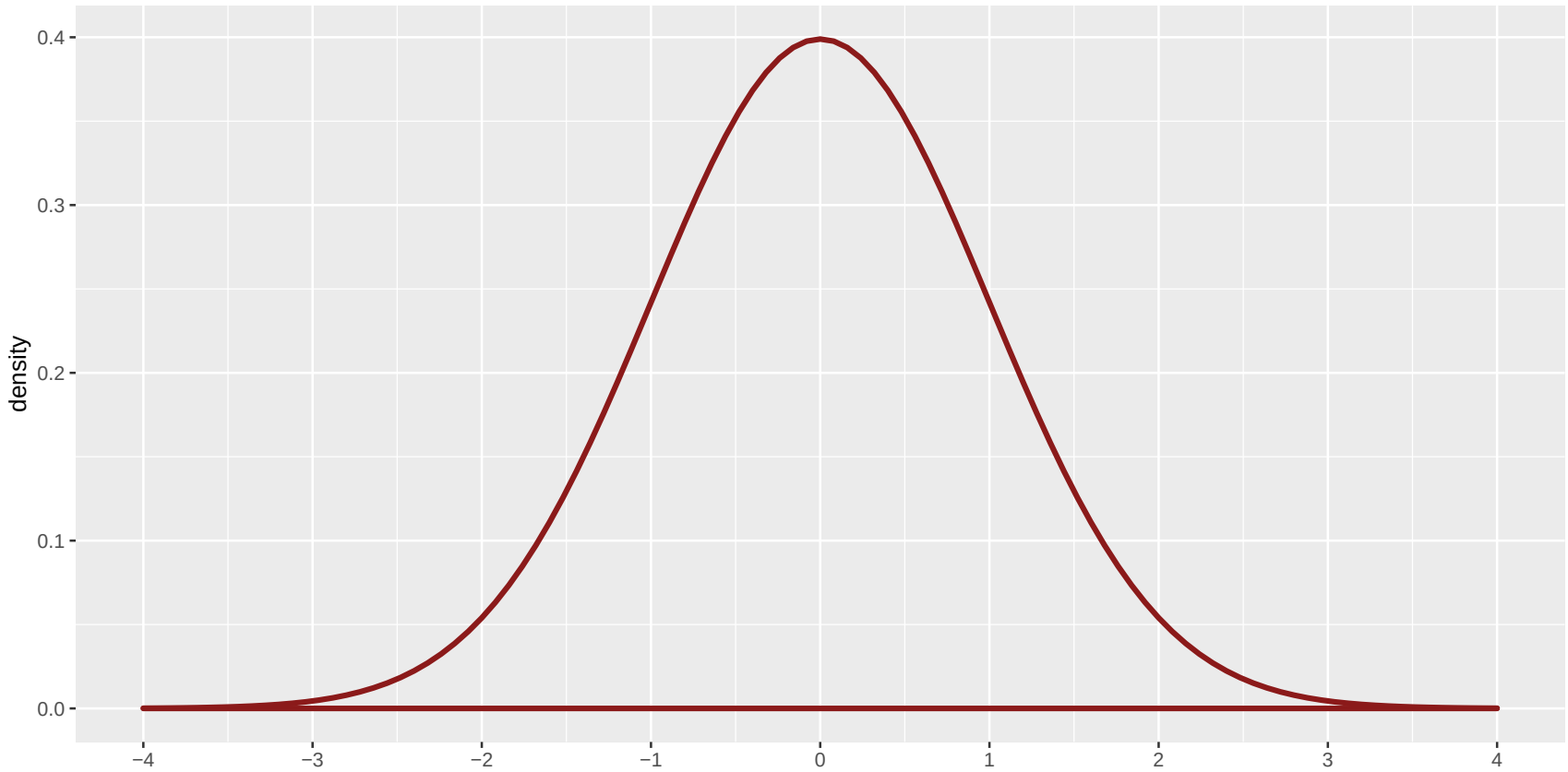
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

- The shape is determined by two **parameters**, μ and σ
- If we were to plot $f(x)$ as a function of x, we would obtain a normal distribution that would be centered at whatever value of μ we specified, and it would have a standard deviation equal to σ.


0.100
0.075
0.050
0.025
0.000
y
140
155
170
185
200
mu = 170, sd = 4
mu = 170, sd = 8
mu = 185, sd = 4

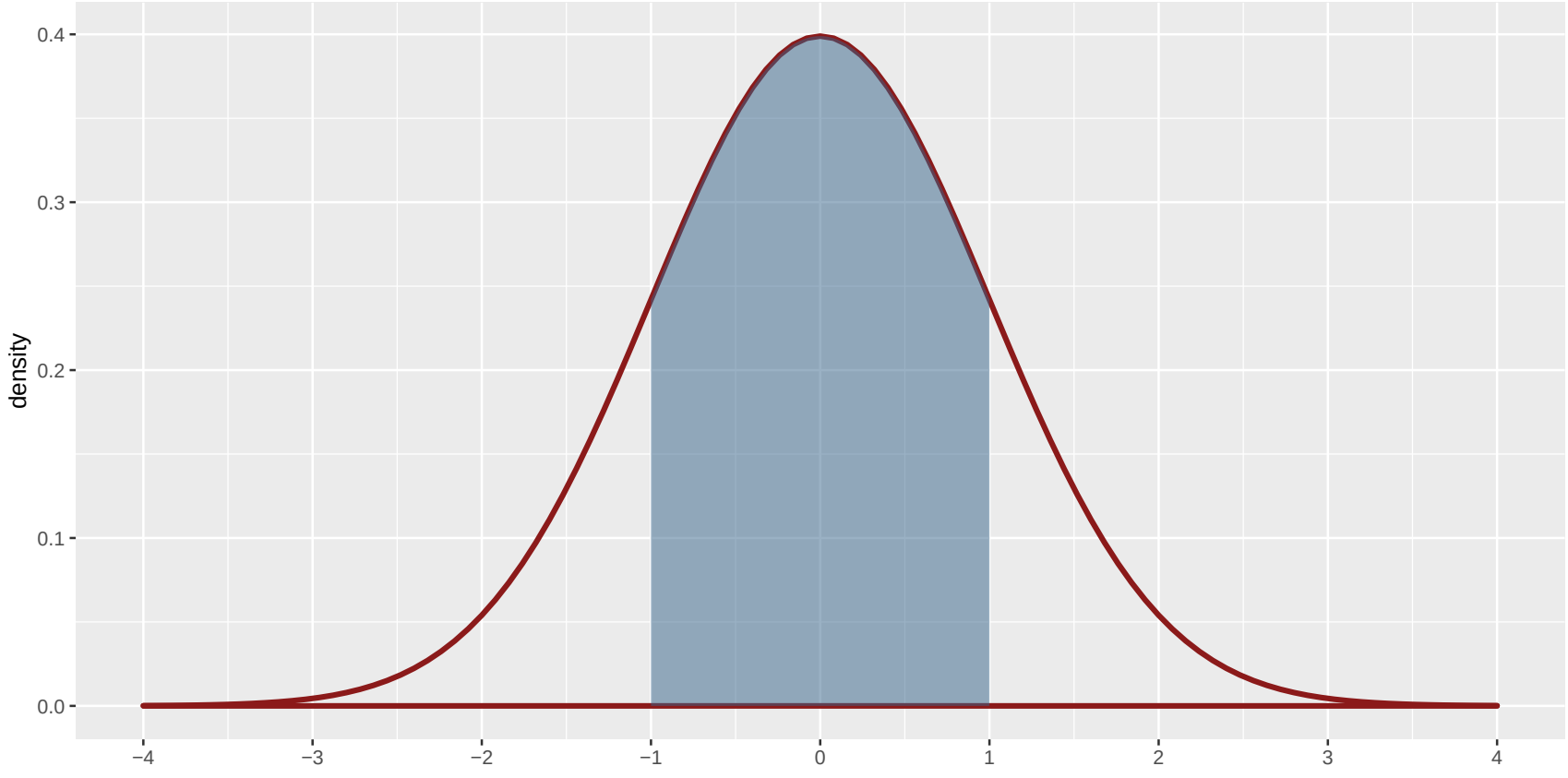
The standard normal distribution

A special case of the normal distribution with $\mu = 0$ and $\sigma = 1$.



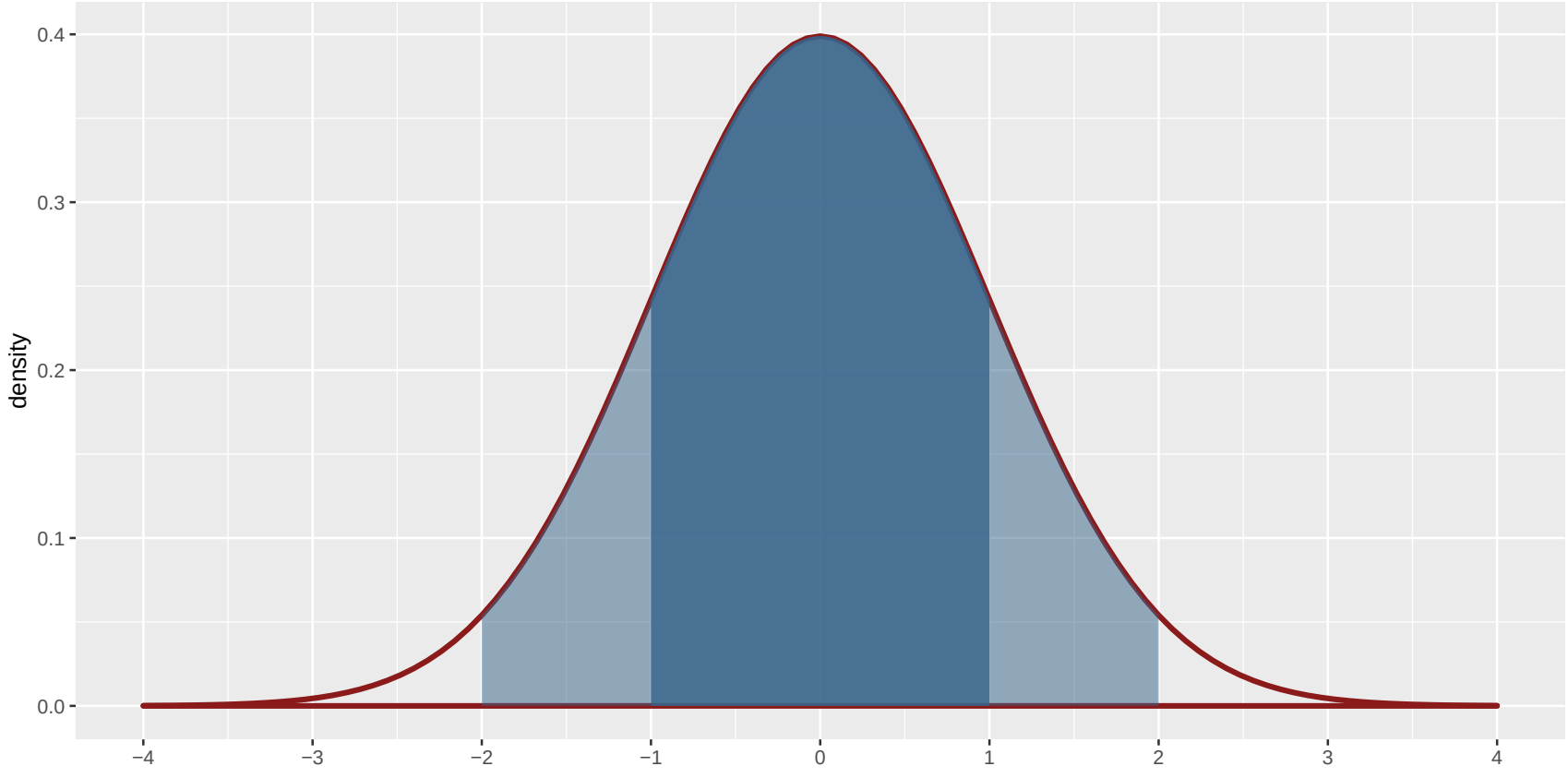
The standard normal distribution

- ~68% of area within 1 standard deviation



The standard normal distribution

- ~95% of the area within 2 standard deviations



Any normal distribution can be transformed into the standard normal

Say X is normally distributed with mean μ and variance σ^2. We can write this as

$$X \sim N(\mu, \sigma^2)$$

We can transform X using the z-**transformation**

$$\frac{X - \mu}{\sigma}$$

Call this transformed version Z i.e. $Z = \frac{X-\mu}{\sigma}$. Then

$$Z \sim N(0, 1)$$

we can refer to the transformed version as **Z-scores**.

Z-scores

- Z-scores tell you the number of standard deviations by which the value of a raw score is above or below the mean value.
- In the heights example, the mean $\mu = 170$ and standard deviation $\sigma = 8$.

Rohan is 180cm. What is his Z-score?

$$Z = \frac{180 - 170}{8} = 1.25$$

So Rohan is 1.25 standard deviations above the mean height.

- Monica is 168cm, so her Z-score is -0.25. So she is 0.25 standard deviations below the mean height.

Conditional expectation

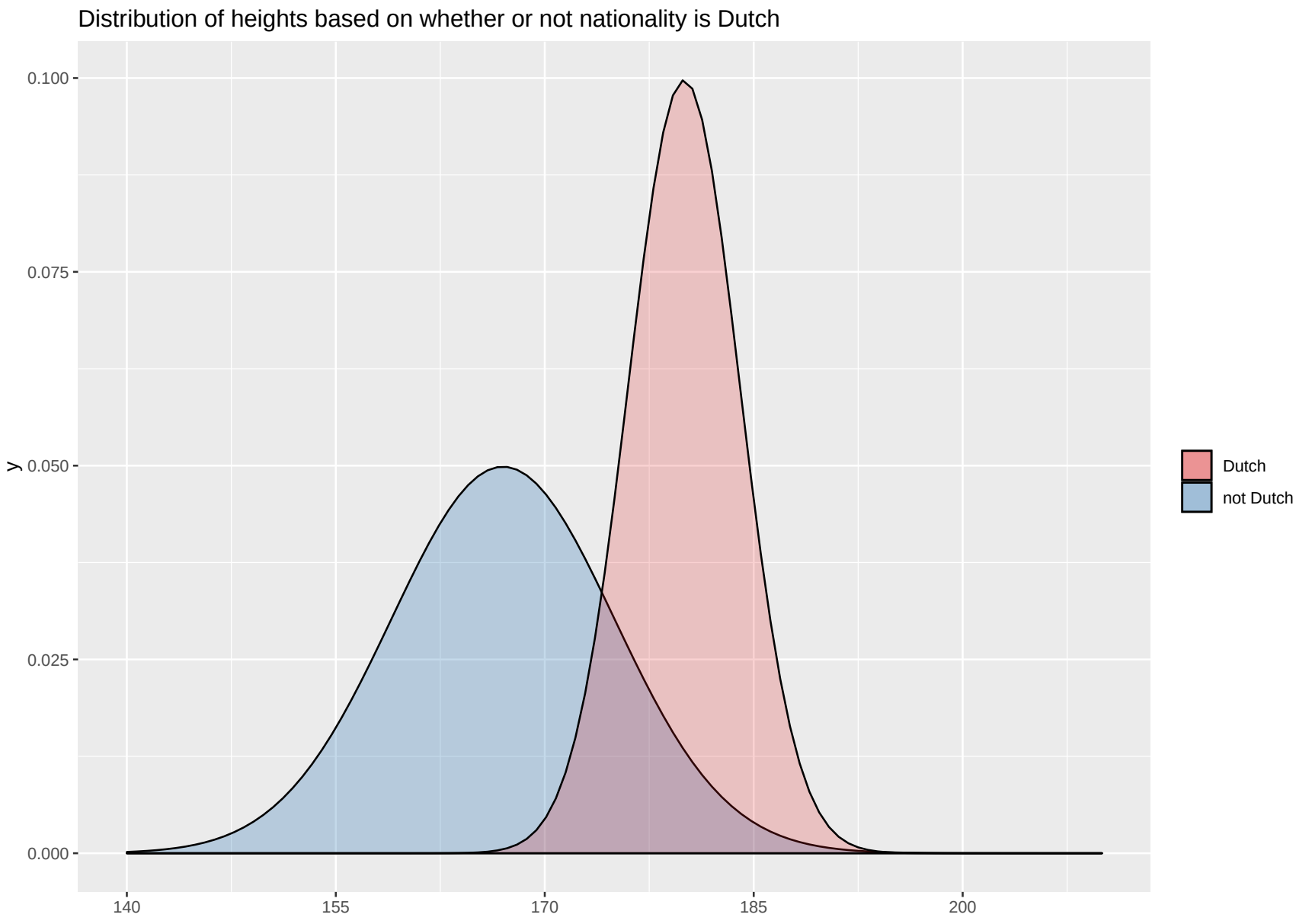
Conditional expectation

- The conditional expected value of a random variable, Y, is the probability- weighted average of all possible values of Y given that another random variable, X, is equal to some specific value

$$E(Y \mid X = x) = \sum_{y} y f_{Y|X}(y \mid x)$$

A conditional expected value is essentially just a sub-population mean—it is a measure of central tendency for a conditional probability distribution

Conditional distributions and expectations



The conditional expectation function (CEF) decomposition property

Any outcome Y_i can be decomposed into the following

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

One way to interpret the CEF decomposition property 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

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?
- 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?

The SLR model

In the case of SLR, the model is:

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

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**

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)

Recall the CEF decomposition property

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

So the SLR is a model for the CEF

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

Hence why the interpretation of β_0 etc is the expected value or population mean.

SLR in R

```
# filter dataset to just be 2017
country_ind_2017 <- country_ind %>% filter(year==2017)
# run the regression
mod <- lm(life_expectancy ~ tfr, data = country_ind_2017)
```

SLR in R

```r
summary(mod)
```

```
##
## 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
```

Multiple linear regression

Multiple Linear Regression

Let's look at the case of two independent variables

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

Example:

- $\{Y_1, Y_2, \ldots, Y_{176}\}$ is life expectancy by country in 2017
- $\{X_{1,1}, X_{2,1}, \ldots, X_{176,1}\}$ is TFR by country in 2017
- $\{X_{1,2}, X_{2,2}, \ldots, X_{176,2}\}$ is child mortality by country in 2017

Research question:

- How does life expectancy differ across different levels of fertility and child mortality
- In other words, is life expectancy associated with fertility and child mortality, and if so, how?

MLR model

In a similar way to SLR, MLR is a model for the CEF:

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

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?

$$\begin{aligned} E(Y_i \mid X_{i1} = 0, X_{i2} = 0) &= \beta_0 + \beta_1(0) + \beta_2(0) \\ &= \beta_0 \end{aligned}$$

- 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?

$$\begin{aligned} E(Y_i \mid X_{i1} &= x_1 + 1, X_{i2} = x_2) - E(Y_i \mid X_{i1} = x_1, X_{i2} = x_2) \\ &= (\beta_0 + \beta_1 (x_1 + 1) + \beta_2 x_2) - (\beta_0 + \beta_1 x_1 + \beta_2 x_2) \\ &= (\beta_0 + \beta_1 x_1 + \beta_1 + \beta_2 x_2) - (\beta_0 + \beta_1 x_1 + \beta_2 x_2) \\ &= \beta_1 \end{aligned}$$

- 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

MLR in R

Simple extension of SLR:

```
mod <- lm(life_expectancy~tfr+child_mort, data = country_ind_2017)
summary(mod)
```

```
##
## Call:
## lm(formula = life_expectancy ~ tfr + child_mort, data = country_ind_2017)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7103 -1.6787 -0.1197  1.7379  5.7605
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.80622    0.56028 149.578  < 2e-16 ***
## tfr         -1.07102    0.30293  -3.536 0.000522 ***
## child_mort  -0.21031    0.01301 -16.171  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.527 on 173 degrees of freedom
## Multiple R-squared:  0.9015, Adjusted R-squared:  0.9004
## F-statistic: 792.1 on 2 and 173 DF,  p-value: < 2.2e-16
```

How should we interpret?

Variance decomposition

The variance of Y_i can be decomposed into two components: a component 'explained by X_{i1} and X_{i2}' and a component 'unexplained by X_{i1} and X_{i2}'.

$$\text{total sum of squares} = \text{model sum of squares} + \text{reisdual sum of squares}$$

$$SST = SSM + SSR$$

$$\Sigma_i \left(Y_i - \widehat{E}\left(Y_i\right) \right)^2 = \sum_i \left(\widehat{E}\left(Y_i \mid X_{i1}, X_{i2}\right) - \widehat{E}\left(Y_i\right) \right)^2 + \sum_i \left(Y_i - \widehat{E}\left(Y_i \mid X_{i1}, X_{i2}\right) \right)$$

$$\sum_i \left(Y_i - \bar{Y}_i\right)^2 = \sum_i \left(\widehat{Y}_i - \bar{Y}_i \right)^2 + \quad \sum_i \left(Y_i - \widehat{Y}_i \right)^2$$

Variance decomposition

We can use this to assess model fit, through the R^2:

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

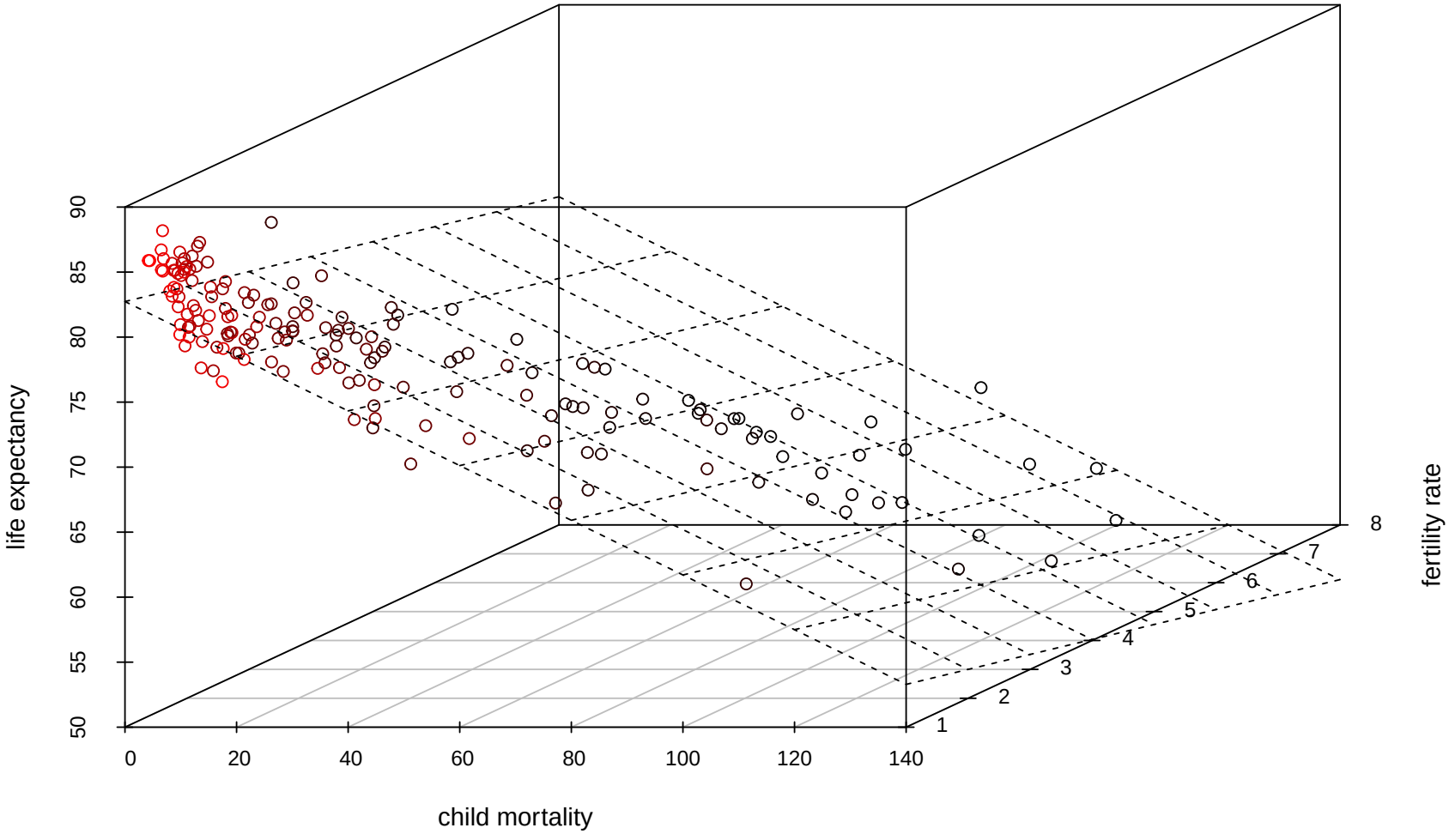
Estimation

OLS Estimation

- $E(Y_i \mid X_{i1}, X_{i2})$, and by extension, β_0, β_1 and β_2 are unknown population quantities, so we need a way of estimating the MLR from sample data
- We use ordinary least squares (OLS) to choose estimators for $\{\beta_0, \beta_1, \beta_2\}$, denoted $\left\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right\}$, that minimize the sum of squared residuals. This can be written as

$$\begin{aligned}\sum_i \hat{\varepsilon}_i^2 &= \Sigma_i \left(Y_i - \hat{E}(Y_i \mid X_{i1}, X_{i2})\right)^2 \\ &= \sum_i \left(Y_i - \left(\hat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2}\right)\right)^2\end{aligned}$$

OLS Estimation: minimizing square residuals



OLS Estimation

The OLS estimators for the MLR model parameters are:

$$\hat{\beta}_1 = \frac{\sum_i \left(\tilde{Y}_i \tilde{X}_{i1}\right) \sum_i \left(\tilde{X}_{i2} \tilde{X}_{i2}\right) - \sum_i \left(\tilde{Y}_i \tilde{X}_{i2}\right) \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}{\sum_i \left(\tilde{X}_{i1} \tilde{X}_{i1}\right) \sum_i \left(\tilde{X}_{i2} \tilde{X}_{i2}\right) - \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right) \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}$$

$$\hat{\beta}_2 = \frac{\sum_i \left(\tilde{Y}_i \tilde{X}_{i2}\right) \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i1}\right) - \sum_i \left(\tilde{Y}_i \tilde{X}_{i1}\right) \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}{\sum_i \left(\tilde{X}_{i1} \tilde{X}_{i1}\right) \sum_i \left(\tilde{X}_{i2} \tilde{X}_{i2}\right) - \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right) \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_i Y_i - \hat{\beta}_1 \left(\frac{1}{n} \sum_i X_{i1}\right) - \hat{\beta}_2 \left(\frac{1}{n} \sum_i X_{i2}\right) = \bar{Y}_i - \hat{\beta}_1 \bar{X}_{i1} - \hat{\beta}_2 \bar{X}_{i2}$$

where $\tilde{Y}_i = Y_i - \bar{Y}_i$, $\tilde{X}_{i1} = X_{i1} - \bar{X}_{i1}$, and $\tilde{X}_{i2} = X_{i2} - \bar{X}_{i2}$.

OLS Estimation

$$\hat{\beta}_1 = \frac{\sum_i \left(\tilde{Y}_i \tilde{X}_{i1}\right) \sum_i \left(\tilde{X}_{i2} \tilde{X}_{i2}\right) - \sum_i \left(\tilde{Y}_i \tilde{X}_{i2}\right) \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}{\sum_i \left(\tilde{X}_{i1} \tilde{X}_{i1}\right) \sum_i \left(\tilde{X}_{i2} \tilde{X}_{i2}\right) - \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right) \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}$$

Covariation between Y_i and X_{i1} that is independent of X_{i2} divided by variation in X_{i1} that is independent of X_{i2}.

(Similarly for $\hat{\beta}_2$, but it is the covariation between Y_i and X_{i2} that is independent of X_{i1} divided by variation in X_{i2} that is independent of X_{i1}.)

OLS Estimation

The 'partial effect' estimators can also be expressed as:

$$\hat{\beta}_1 = \frac{\sum_i \left(Y_i - \frac{1}{n}\Sigma_i Y_i\right)\left(X_{i1}^r - \frac{1}{n}\Sigma_i X_{i1}^r\right)}{\Sigma_i \left(X_{i1}^r - \frac{1}{n}\Sigma_i X_{i1}^r\right)^2}$$

where $X_{i1}^r = X_{i1} - \hat{E}\left(X_{i1} \mid X_{i2}\right)$ are the residuals from an SLR of X_{i1} on X_{i2}.

In a similar way, $\hat{\beta}_2$ can be expressed in terms of the residuals from an SLR of X_{i2} on X_{i1}.

OLS estimation of the MLR model: general case

The OLS estimators for the MLR model parameters are:

$$\hat{\beta}_k = \frac{\sum_i \left(Y_i - \frac{1}{n}\Sigma_i Y_i\right)\left(X_{ik}^r - \frac{1}{n}\Sigma_i X_{ik}^r\right)}{\Sigma_i \left(X_{ik}^r - \frac{1}{n}\Sigma_i X_{ik}^r\right)^2}$$

$$\begin{aligned}\widehat{\beta}_0 &= \frac{1}{n}\sum_i Y_i - \widehat{\beta}_1\left(\frac{1}{n}\sum_i X_{i1}\right) - \cdots - \widehat{\beta}_k\left(\frac{1}{n}\sum_i X_{ik}\right) \\ &= \bar{Y}_i - \widehat{\beta}_1\bar{X}_{i1} - \cdots - \hat{\beta}_k\bar{X}_{ik}\end{aligned}$$

where X_{ik}^r are the residuals from a MLR of X_{ik} on all the other explanatory variables in the model

The MLR assumptions

Recall the five assumptions of MLR:

1. no model misspecification
2. there is independent variation in all of the explanatory variables
   - In other words, none of the explanatory variables are constants, and there are no perfect linear relationships among the explanatory variables
   - e.g. can't have $X_{i1} = X_{i2} + X_{i3}$
3. All variables are from a simple random sample
   - This assumption implies that all members of a population have an equal probability of selection, that all possible samples of size n have an equal probability of selection, and that each observation is independent of all the others

The MLR assumptions

4. The variance of $\varepsilon_i = Y_i - E(Y_i \mid X_{i1}, X_{i2}, \ldots, X_{ik})$ is the same across all values of the explanatory variables
   i.e. $\text{Var}(\varepsilon_i \mid X_{i1}, X_{i2}, \ldots, X_{ik}) = \sigma^2$
   - This is called homoskedasticity
5. The normality assumption $\varepsilon_i = Y_i - E(Y_i \mid X_{i1}, X_{i2}, \ldots, X_{ik})$ is normally distributed

Sampling distribution of the MLR-OLS estimator

- Under the MLR model assumptions, the OLS estimator, $\hat{\beta}_k$ is normally distributed with a mean equal to

$$E\left(\hat{\beta}_k\right) = \beta_k$$

and variance

$$\text{Var}\left(\hat{\beta}_k\right) = \frac{\sigma^2}{\sum_i \left(X_{ik} - \bar{X}_{ik}\right)^2 \left(1 - R_k^2\right)}$$

We use information about the probability distribution of $\hat{\beta}_k$ to make inferences about β_k.

Sampling distribution of the MLR-OLS estimator

The standard deviation of $\hat{\beta}_k$

$$sd\left(\hat{\beta}_k\right) = \sqrt{\frac{\sigma^2}{\sum_i \left(X_{ik} - \bar{X}_{ik}\right)^2 \left(1 - R_k^2\right)}}$$

Under ideal conditions, statistical inferences about MLR parameters would be based on the fact that the standardized MLR-OLS estimator follows the z-distribution (i.e., the standard normal distribution)

$$Z_{\widehat{\beta}_k} = \frac{\widehat{\beta}_k - \beta_k}{sd\left(\widehat{\beta}_k\right)} \sim N(0,1)$$

But we don't know know the true value for σ^2, so we have to estimate.

Standard error of the MLR-OLS estimator

- Because σ^2 is an unknown population quantity, and thus $sd\left(\widehat{\beta}_1\right)$ is unknown, we have to estimate them
- An estimator for the error variance $\sigma^2 = \text{Var}\left(\varepsilon_i \mid X_i\right)$ is

$$\hat{\sigma}^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n - (k+1)} = \frac{SSR}{df}$$

And the standard error of $\hat{\beta}_k$, which is an estimator of $sd\left(\widehat{\beta}_1\right)$, is

$$\text{se}\left(\hat{\beta}_k\right) = \sqrt{\frac{\widehat{\sigma}^2}{\sum_i \left(X_{ik} - \bar{X}_{ik}\right)^2 \left(1 - R_k^2\right)}}$$

Sampling distribution of the SE- standardized MLR-OLS estimator

Under the five assumption discussed, the SE-standardized $\hat{\beta}_k$

$$T_{\widehat{\beta}_k} = \frac{\widehat{\beta}_k - \beta_k}{se\left(\widehat{\beta}_k\right)}$$

follows a t-distribution with $n - (k + 1)$ degrees of freedom.

Example R output

```r
summary(lm(life_expectancy~tfr+child_mort+maternal_mort, data = country_ind))
```

```
##
## Call:
## lm(formula = life_expectancy ~ tfr + child_mort + maternal_mort,
##     data = country_ind)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1077  -1.8229  -0.0268   1.9726   7.9812
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   82.6668192  0.2067649 399.811  < 2e-16 ***
## tfr           -0.7195526  0.1073323  -6.704 2.81e-11 ***
## child_mort    -0.2068586  0.0058846 -35.153  < 2e-16 ***
## maternal_mort -0.0003844  0.0006366  -0.604    0.546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.837 on 1580 degrees of freedom
## Multiple R-squared:  0.8954, Adjusted R-squared:  0.8952
## F-statistic:  4510 on 3 and 1580 DF,  p-value: < 2.2e-16
```

Interval estimation

We can calculate the confidence intervals for MLR parameters, which provide a range of values that contain the true value of the parameter with known probability in repeated sampling.

Interval estimation steps

1. Choose your confidence level (i.e., the probability that the interval estimate will cover the parameter of interest in repeated sampling)

- Usually choose $\nu = 1 - \alpha$ with $\alpha = 0.05$ so the confidence level is $\nu = 0.95$ or 95%

2. find the critical value, t_α, of the t-distribution with $n - 2$ degrees of freedom for which $P(|T| > t_\alpha) = 1 - \nu = \alpha$

- In words, the probability of the absolute value of our T statistic of interest being greater than the critical value (i.e. outside the bounds defined by t_α) is α (e.g. 0.05 or 5%)

Interval estimation steps (ctd)

3. Compute the limits of the confidence interval

- Lower limit: $\hat{\beta}_k - \left(t_\alpha \times se\left(\hat{\beta}_k\right)\right)$
- Upper limit: $\hat{\beta}_k + \left(t_\alpha \times se\left(\hat{\beta}_k\right)\right)$

4. Interpret.

- if random samples were repeatedly collected and confidence intervals were computed as outlined above for each sample, the true value of the parameter, β_k, would lie in the confidence interval in $\nu \times 100$ percent of the samples

```r
mlr_mod <- lm(life_expectancy~tfr+child_mort+maternal_mort, data = country_ind)
n <- nrow(country_ind)
k <- 3
# extract beta1 hat and se
b1_hat <- summary(mlr_mod)$coefficients[2,1]
se_b1_hat <- summary(mlr_mod)$coefficients[2,2]
# choose a confidence level
alpha <- 0.05
v <- 1-alpha
# calculate critical value
t_alpha <- abs(qt(p = alpha/2, df = n-(k+1)))
# calculate confidence interval
# lower
b1_hat - t_alpha*se_b1_hat
```

```
## [1] -0.9300813
```

```r
# upper
b1_hat + t_alpha*se_b1_hat
```

```
## [1] -0.5090239
```

Summary

- Linear regression is a model for the conditional expectation function
- R^2 is a summary of model fit
- Parameters β are estimated used ordinary least squares
- Assuming five MLR assumptions hold, the standardized MLR estimator has a t distribution with $n - k$ degrees of freedom
- Can use this to do hypothesis tests (normally testing whether $\beta = 0$) and confidence intervals

Lab

Practice with linear regression in R!