Python Data Products

Course 2: Design thinking and predictive pipelines

Lecture: Features from categorical data

Learning objectives

In this lecture we will...

- Demonstrate how to incorporate binary and categorical features into regressors
- Compare the benefits of various feature representation strategies

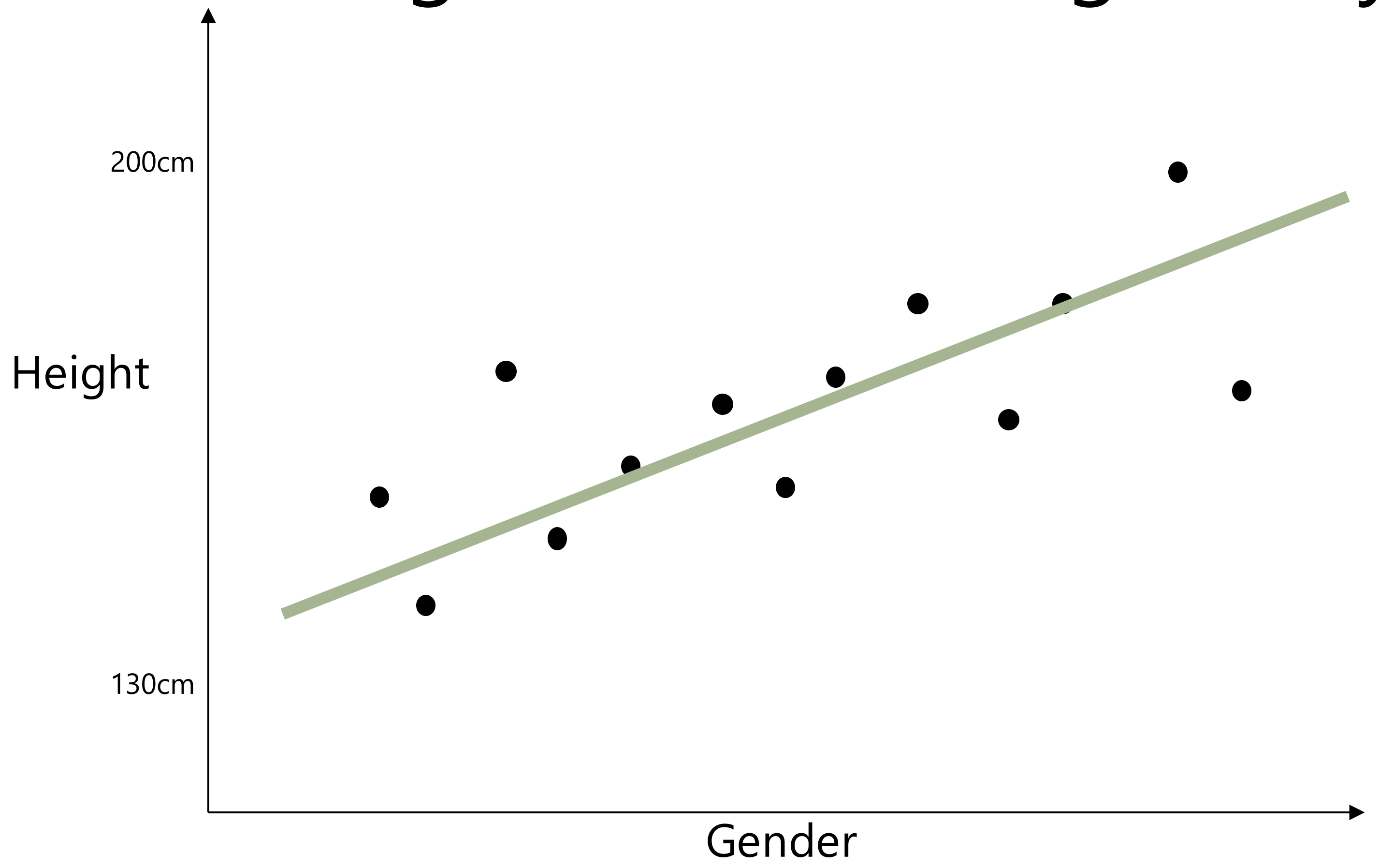
Motivating examples

How would we build regression models that incorporate features like:

- How does height vary with **gender?**
- How do preferences vary with **geographical region?**
- How does product demand change during different **seasons?**

Motivating examples

E.g. How does height vary with **gender?**



Motivating examples

E.g. How does height vary with **gender?**

- Previous picture doesn't quite make sense: we're unlikely to have a dataset including a continuum of gender values, so fitting a "line" doesn't seem to fit
- So how can we deal with this type of data using a linear regression framework?

Motivating examples

E.g. How does height vary with **gender?**

- Presumably our gender values might look more like
  {"male", "female", "other", "not specified"}
- **Let's first start with a binary problem where we just have {"male", "female"}**

Motivating examples

What should our **model equation** look like?

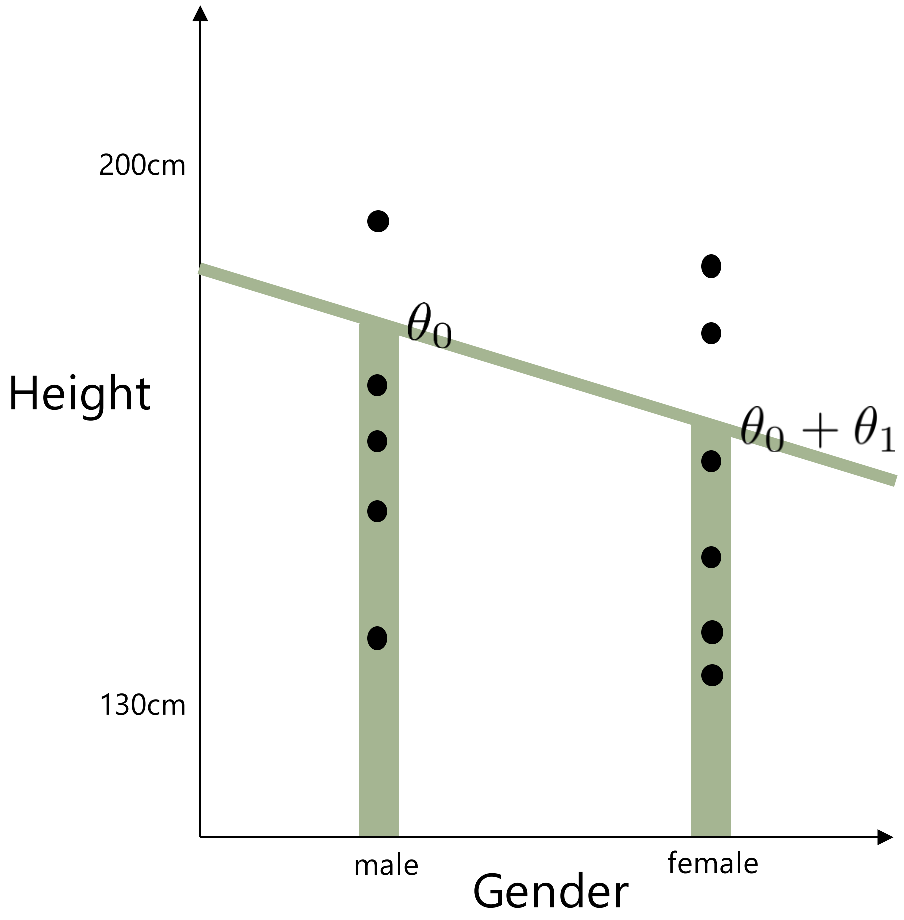
$$\text{Height} = \theta_0 + \theta_1 \times \text{gender}$$

gender = **0 if male, 1 if female**

$$\text{Height} = \theta_0$$ **if male**

$$\text{Height} = \theta_0 + \theta_1$$ **if female**

Motivating examples



θ_0 is the (predicted/average) height for males

θ_1 is the **how much taller** females are than males (in this case a negative number)

We're really still fitting a line though!

Motivating examples

What if we had more than two values?
(e.g {“male”, “female”, “other”, “not specified”})

Could we apply the same approach?

$$\text{Height} = \theta_0 + \theta_1 \times \text{gender}$$

gender = **0 if “male”, 1 if “female”, 2 if “other”, 3 if “not specified”**

$\text{Height} = \theta_0$ **if male**

$\text{Height} = \theta_0 + \theta_1$ **if female**

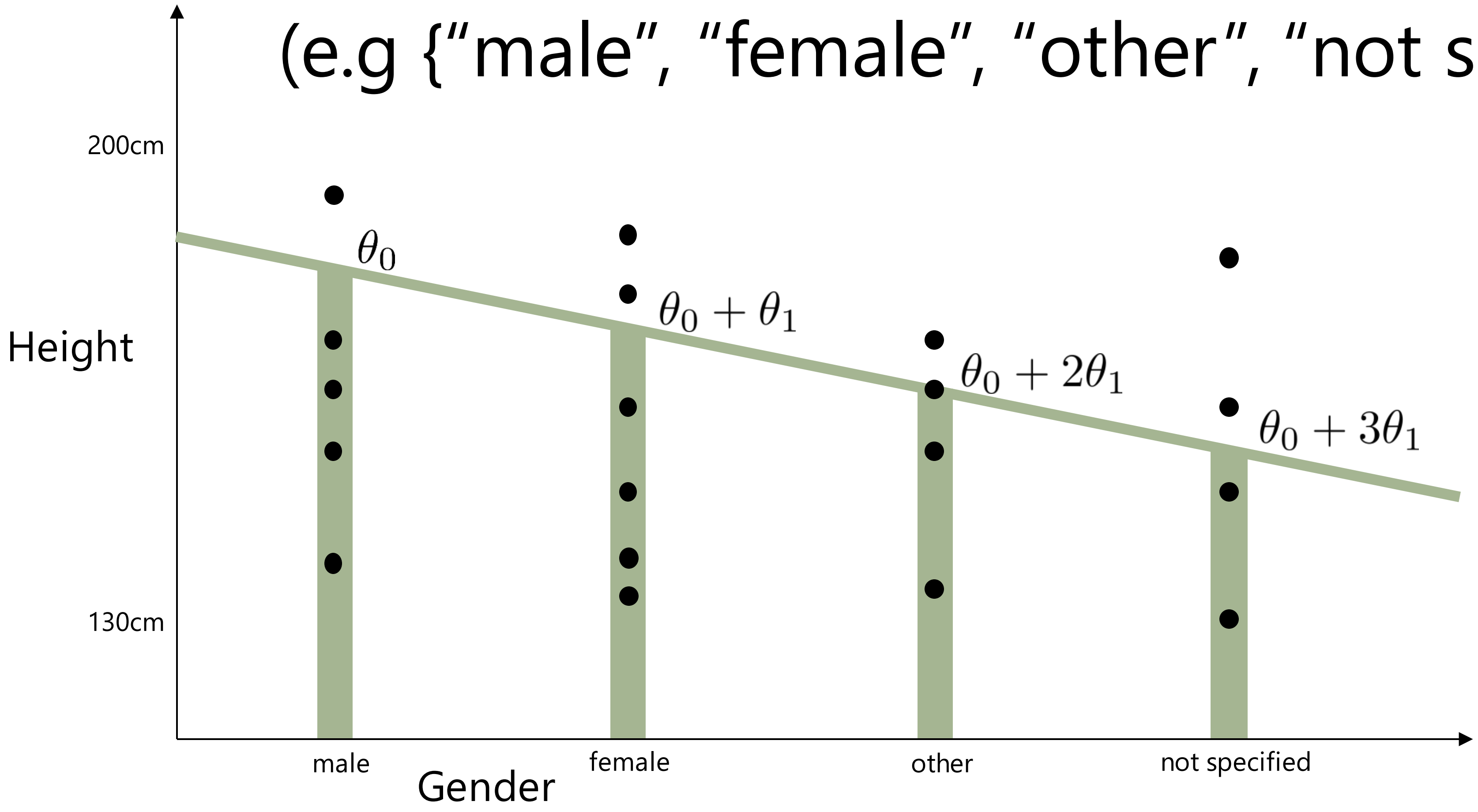
$\text{Height} = \theta_0 + 2\theta_1$ **if other**

$\text{Height} = \theta_0 + 3\theta_1$ **if not specified**

Motivating examples

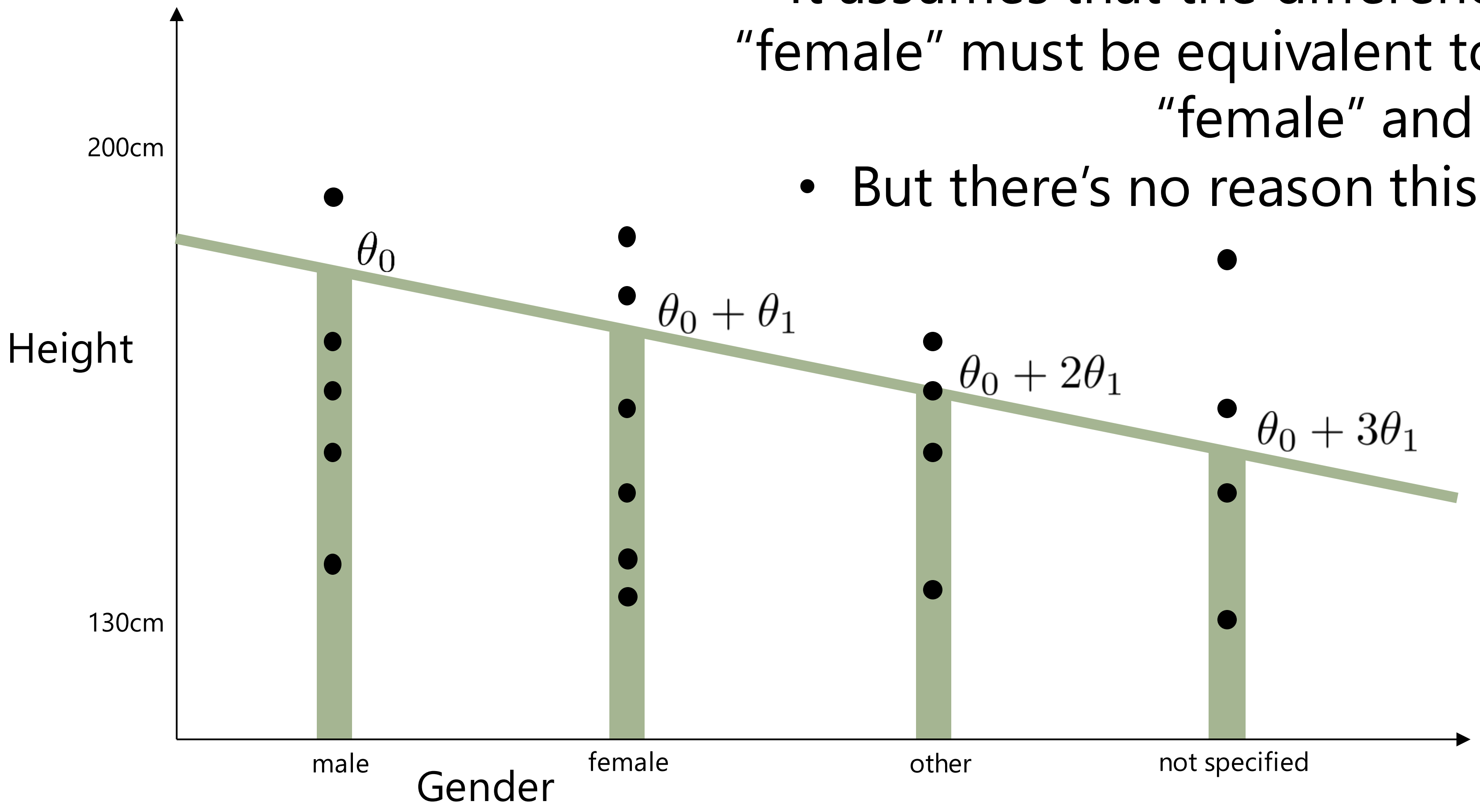
What if we had more than two values?

(e.g {“male”, “female”, “other”, “not specified”})



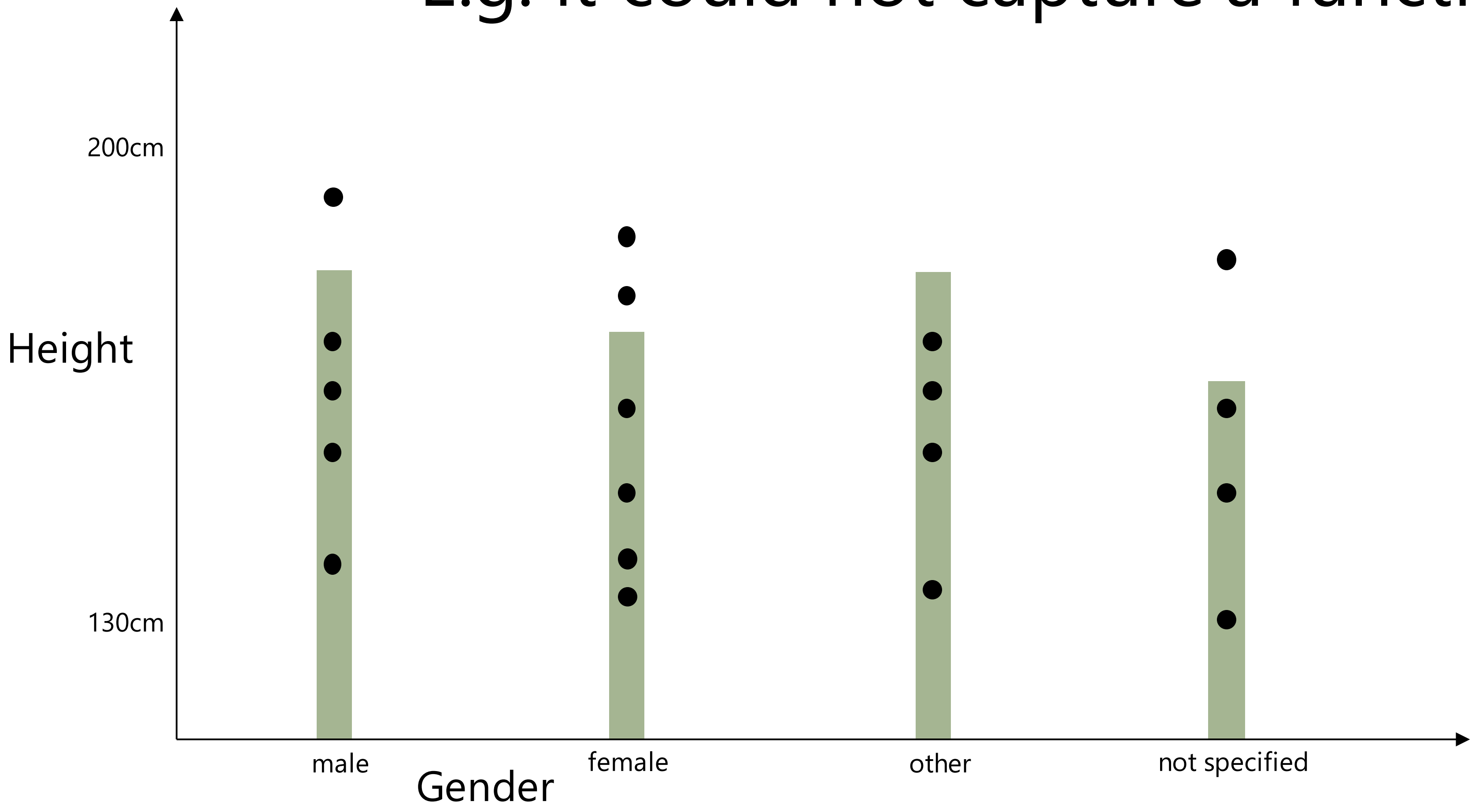
Motivating examples

- This model is **valid,** but won't be very **effective**
- It assumes that the difference between "male" and "female" must be equivalent to the difference between "female" and "other"
- But there's no reason this should be the case!



Motivating examples

E.g. it could not capture a function like:



Motivating examples

Instead we need something like:

$\text{Height} = \theta_0$ **if male**

$\text{Height} = \theta_0 + \theta_1$ **if female**

$\text{Height} = \theta_0 + \theta_2$ **if other**

$\text{Height} = \theta_0 + \theta_3$ **if not specified**

Motivating examples

This is equivalent to:

$$(\theta_0, \theta_1, \theta_2, \theta_3) \cdot (1; \text{feature})$$

where feature = [1, 0, 0] for “female”
feature = [0, 1, 0] for “other”
feature = [0, 0, 1] for “not specified”

Concept: One-hot encodings

feature = [1, 0, 0] for "female"
feature = [0, 1, 0] for "other"
feature = [0, 0, 1] for "not specified"

- This type of encoding is called a **one-hot encoding** (because we have a feature vector with only a single "1" entry
- Note that to capture 4 possible categories, we only need three dimensions (a dimension for "male" would be redundant)
- This approach can be used to capture a variety of categorical feature types, as well as objects that belong to multiple categories

Summary of concepts

- Described how to capture binary and categorical features within linear regression models
- Introduced the concept of a "one-hot" encoding

On your own...

- Think how you would encode different categorical features, e.g. the set of categories a business belongs to, or the set of a user's friends on a social network