Adding Categorical Predictors

Stat 230: Applied Regression Analysis

Example

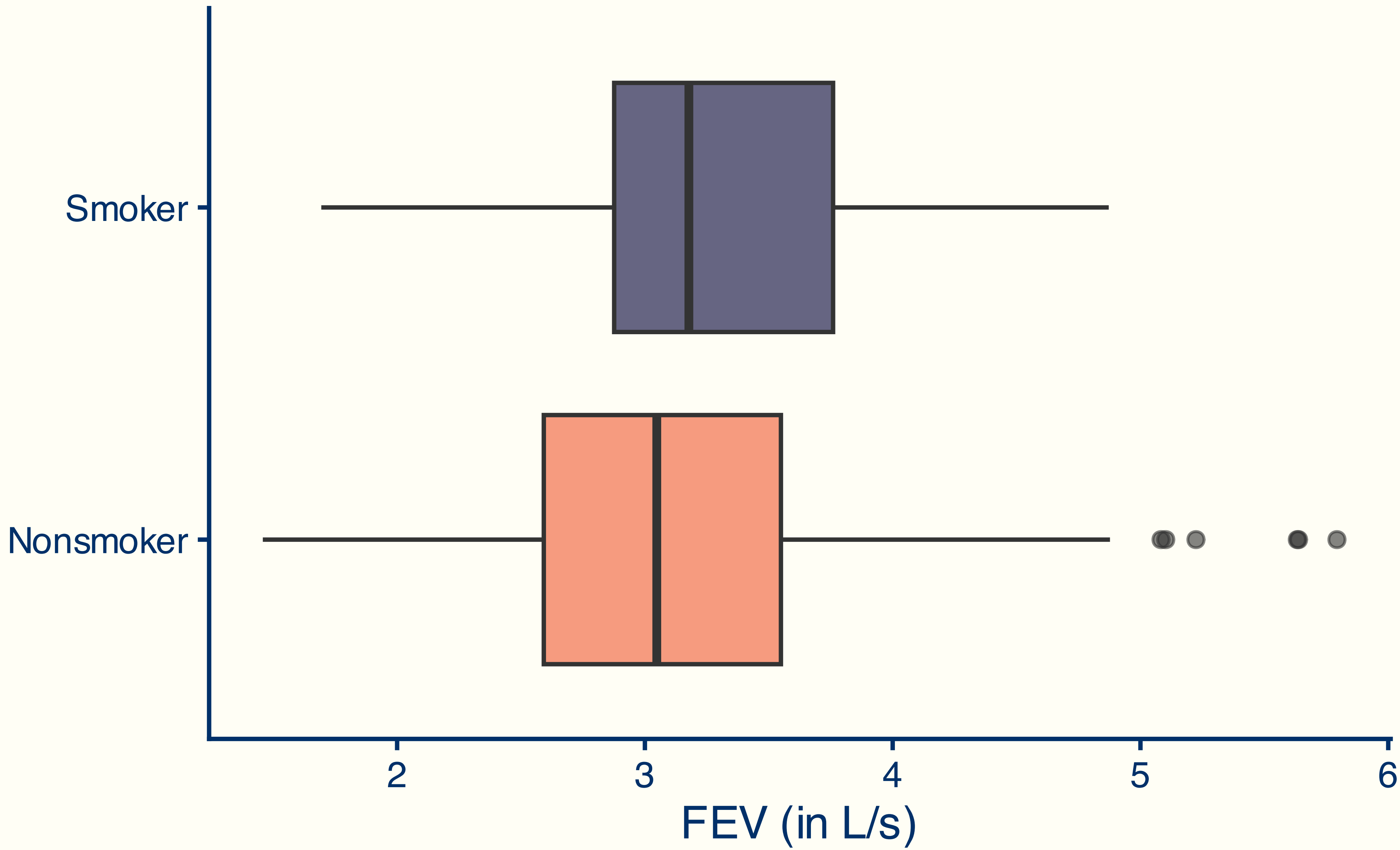
Goal: investigate the association between smoking and lung capacity using data from 345 adolescents between the ages of 10 and 19

Wrinkle: Lung function is expected to increase during adolescence, but smoking may slow it's progression

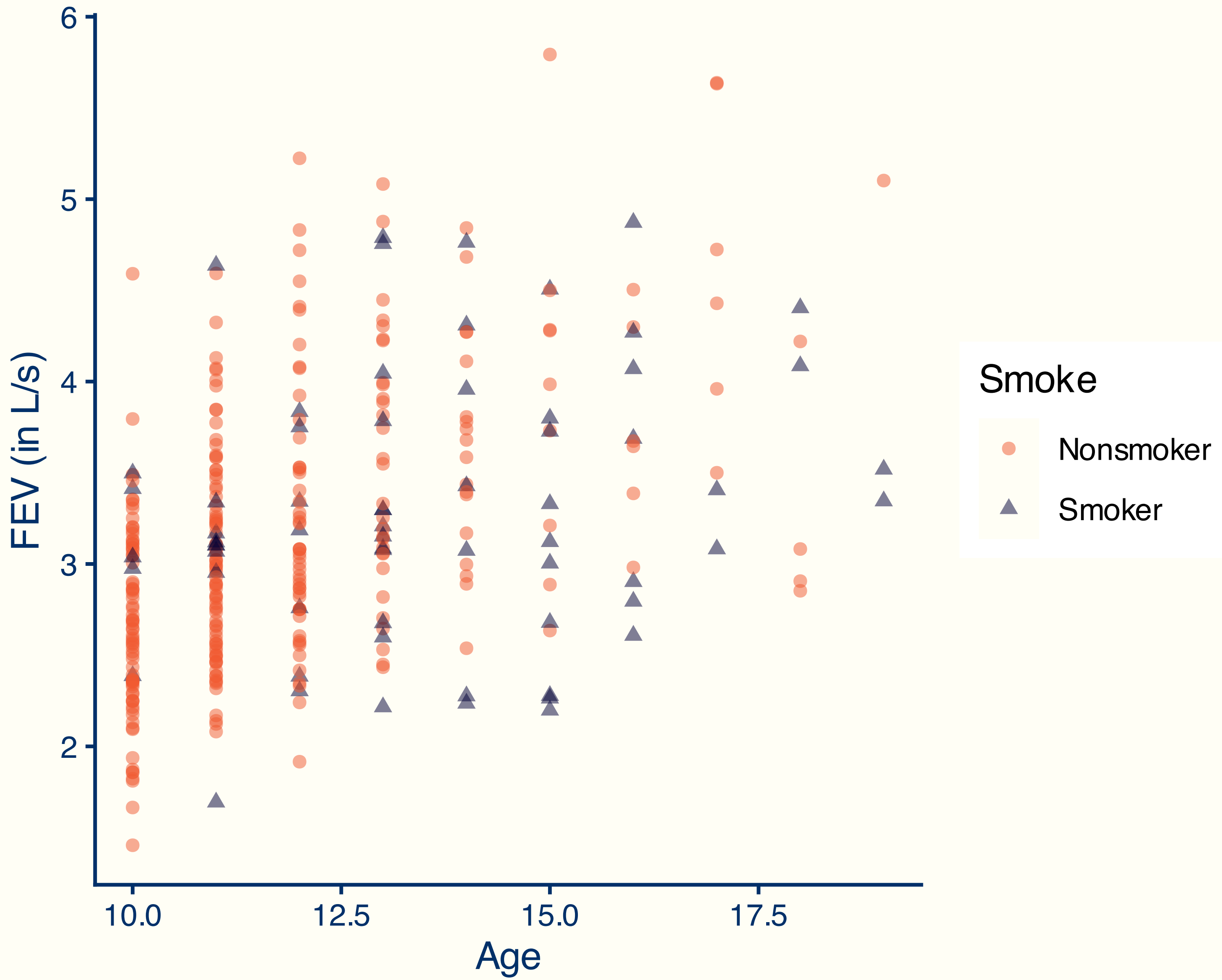
Data:

Variable	Description
`FEV`	forced expiratory volume (in liters per second)
`Age`	age in years
`Smoke`	`Smoker` or `Nonsmoker`

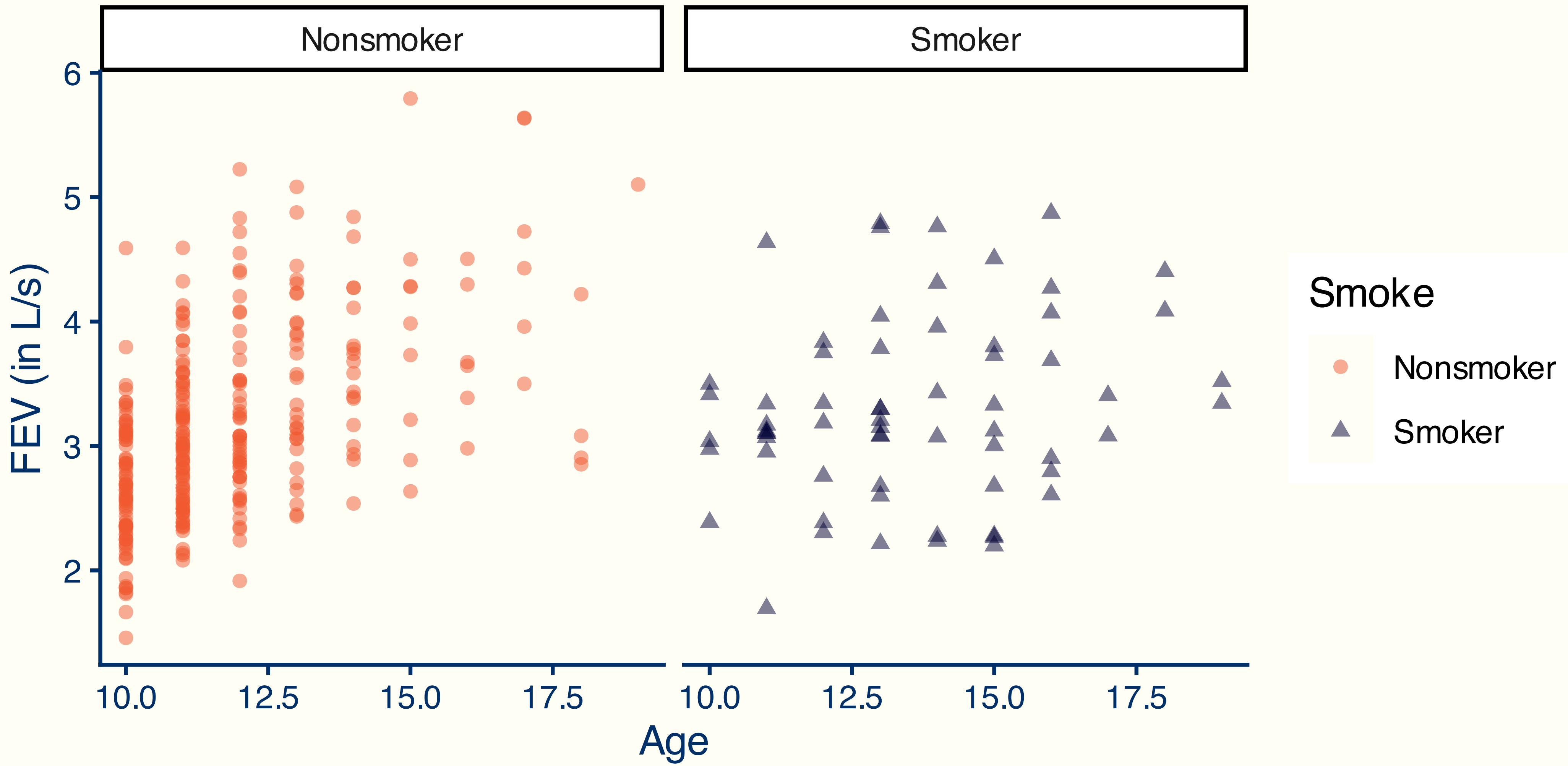
EDA



EDA



EDA



Nonsmoker
Smoker
6
5
4
3
2
FEV (in L/s)
10.0
12.5
15.0
17.5
10.0
12.5
15.0
17.5
Age
Smoke
Nonsmoker
Smoker

Indicator variable

Regression requires a numeric representation of all variables

Create a `Smoker` indicator variable:

- `Smoker` $= 1$
- `Nonsmoker` $= 0$

Example 1

- For each regression model on the handout, sketch the fitted model on the whiteboard
- Each fitted model will have two lines: one for smokers, one for nonsmokers
- Work with your neighbors

Model 1

$$\mu(y|x) = 10 + 1\texttt{age} - 2\texttt{smoker}$$

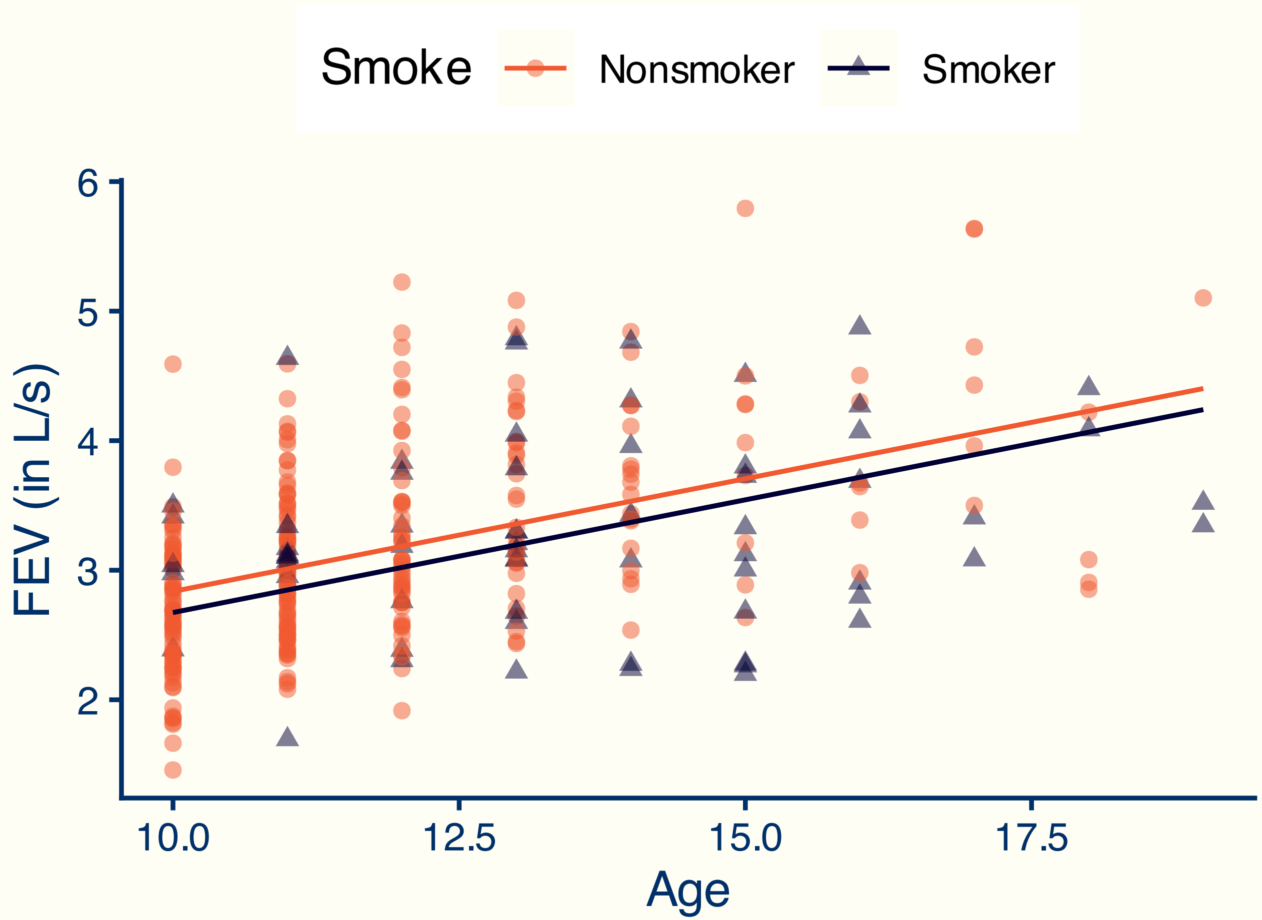
Model 2

$$\mu(y|x) = 5 + 1\text{age} - 0.5\text{age} \times \text{smoker}$$

Model 3

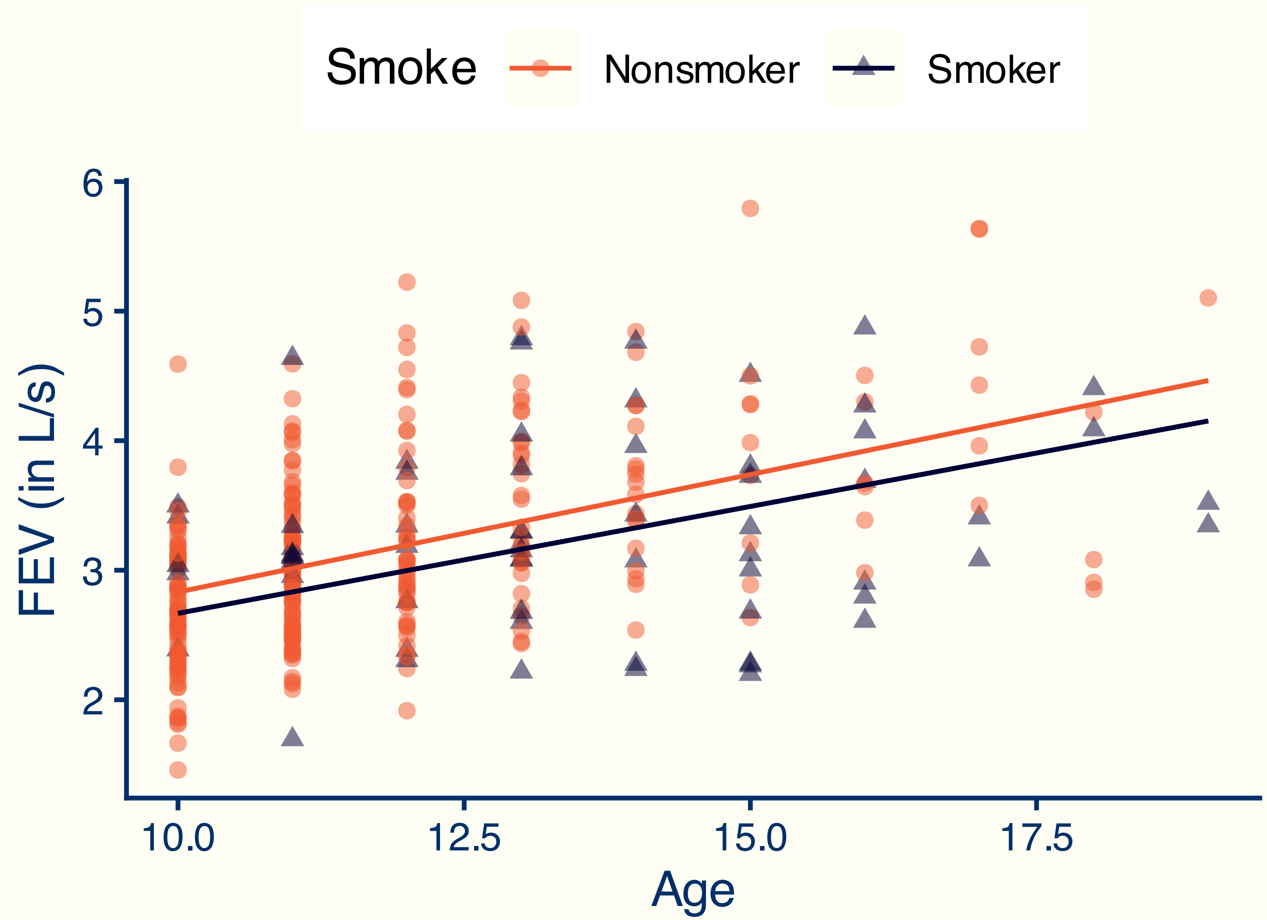
$$\mu(y|x) = 4 + 0.5\mathtt{age} + 3\mathtt{smoker} - 0.5\mathtt{age} \times \mathtt{smoker}$$

Parallel lines model



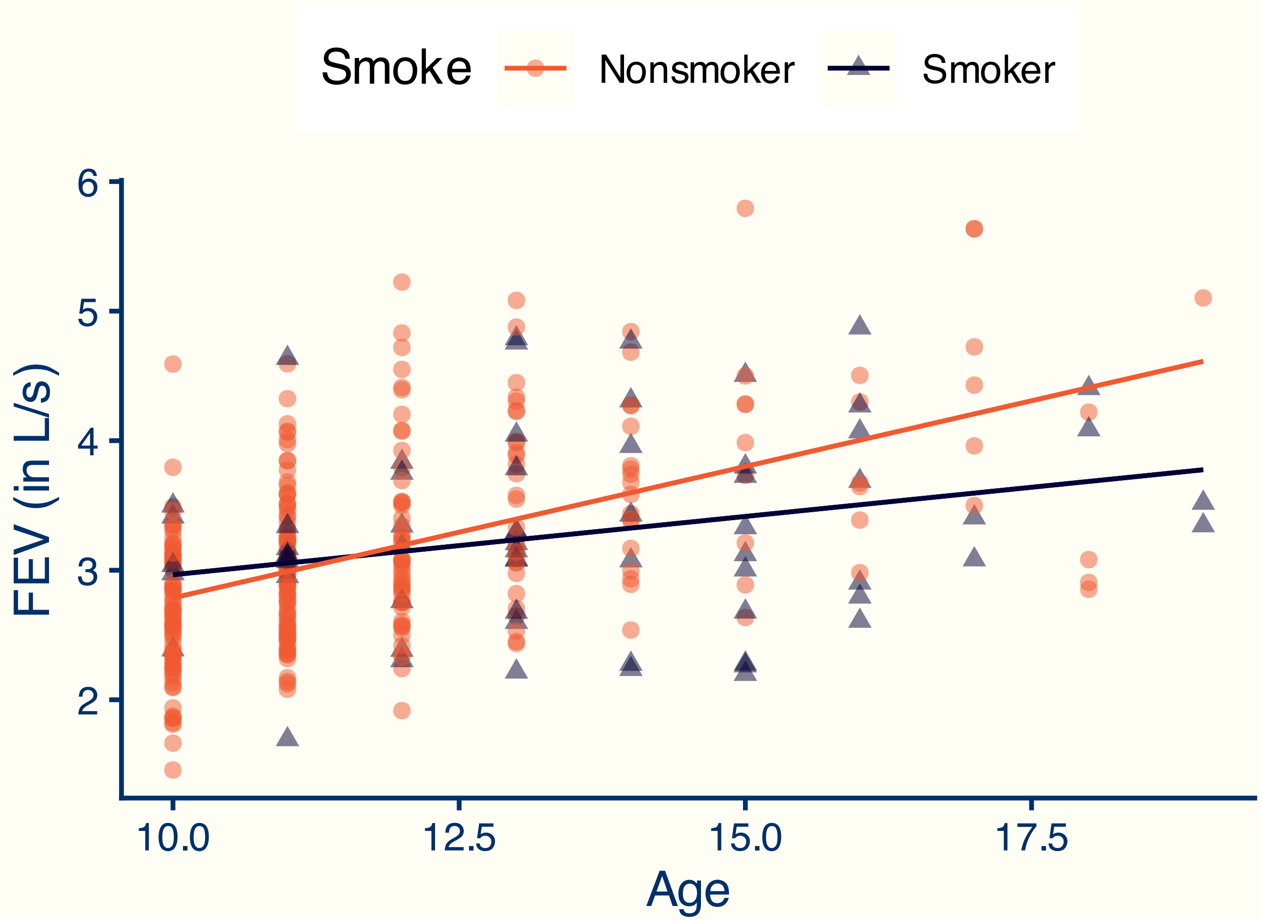
Lung function develops at the same pace, but always lower for smokers

Different slopes model



Adolescents start with similar lung capacity, but smokers develop at a slower rate

Separate lines model



The smokers/non-smokers have different starting lung capacities and develop at different rates

Parallel lines model

$$\mu(y|x) = \beta_0 + \beta_1 \texttt{smoker} + \beta_2 \texttt{age}$$

Interpretations

β_0 = y-intercept for non-smokers

$\beta_0 + \beta_1$ = y-intercept for smokers

β_2 = expected rate of change for both groups

Different slopes model

$$\mu(y|x) = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{age} \times \texttt{smoker}$$

Interpretations

β_0 = y-intercept for both groups

β_1 = expected rate of change (slope) for non-smokers

$\beta_1 + \beta_2$ = expected rate of change (slope) for smokers

Separate lines model

$$\mu(y|x) = \beta_0 + \beta_1 \texttt{smoker} + \beta_2 \texttt{age} + \beta_3 \texttt{age} \times \texttt{smoker}$$

Interpretations

β_0 = y-intercept for non-smokers

$\beta_0 + \beta_1$ = y-intercept for smokers

β_2 = expected rate of change (slope) for non-smokers

$\beta_2 + \beta_3$ = expected rate of change (slope) for smokers

More than 3 categories

Suppose we have student survey data and one of the columns records the year in school:

- First year, Sophomore, Junior, and Senior.

How can we include this variable in a multiple regression model?

A potentially bad idea

We could convert the column to numeric

- First year $\rightarrow$ 1
- Sophomore $\rightarrow$ 2
- Junior $\rightarrow$ 3
- Senior $\rightarrow$ 4

$$\mu(Y|\texttt{classyear}) = \beta_0 + \beta_1 \texttt{classyear}$$

A good idea

We can create a series of indicator (dummy) variables to represent the four categories

Original
First year
Sophomore
Senior
Senior
Junior
Sophomore

A good idea

We can create a series of indicator (dummy) variables to represent the four categories

Original	**FY**
First year	1
Sophomore	0
Senior	0
Senior	0
Junior	0
Sophomore	0

A good idea

We can create a series of indicator (dummy) variables to represent the four categories

Original	FY	Soph
First year	1	0
Sophomore	0	1
Senior	0	0
Senior	0	0
Junior	0	0
Sophomore	0	1

A good idea

We can create a series of indicator (dummy) variables to represent the four categories

Original	**FY**	**So**	**Ju**
First year	1	0	0
Sophomore	0	1	0
Senior	0	0	0
Senior	0	0	0
Junior	0	0	1
Sophomore	0	1	0

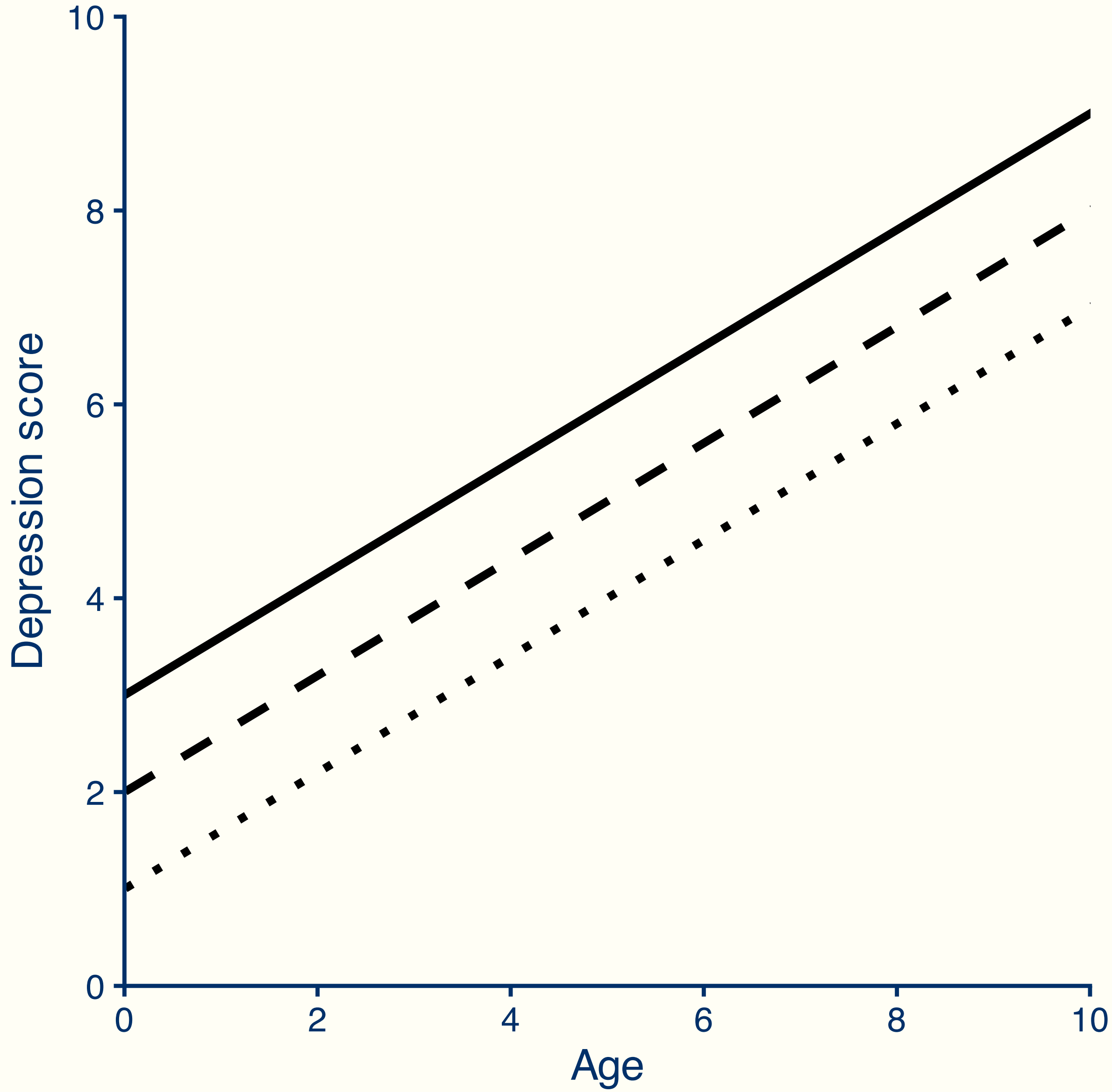
Key idea

For a categorical variable with k levels, $k - 1$ indicator variables are needed.

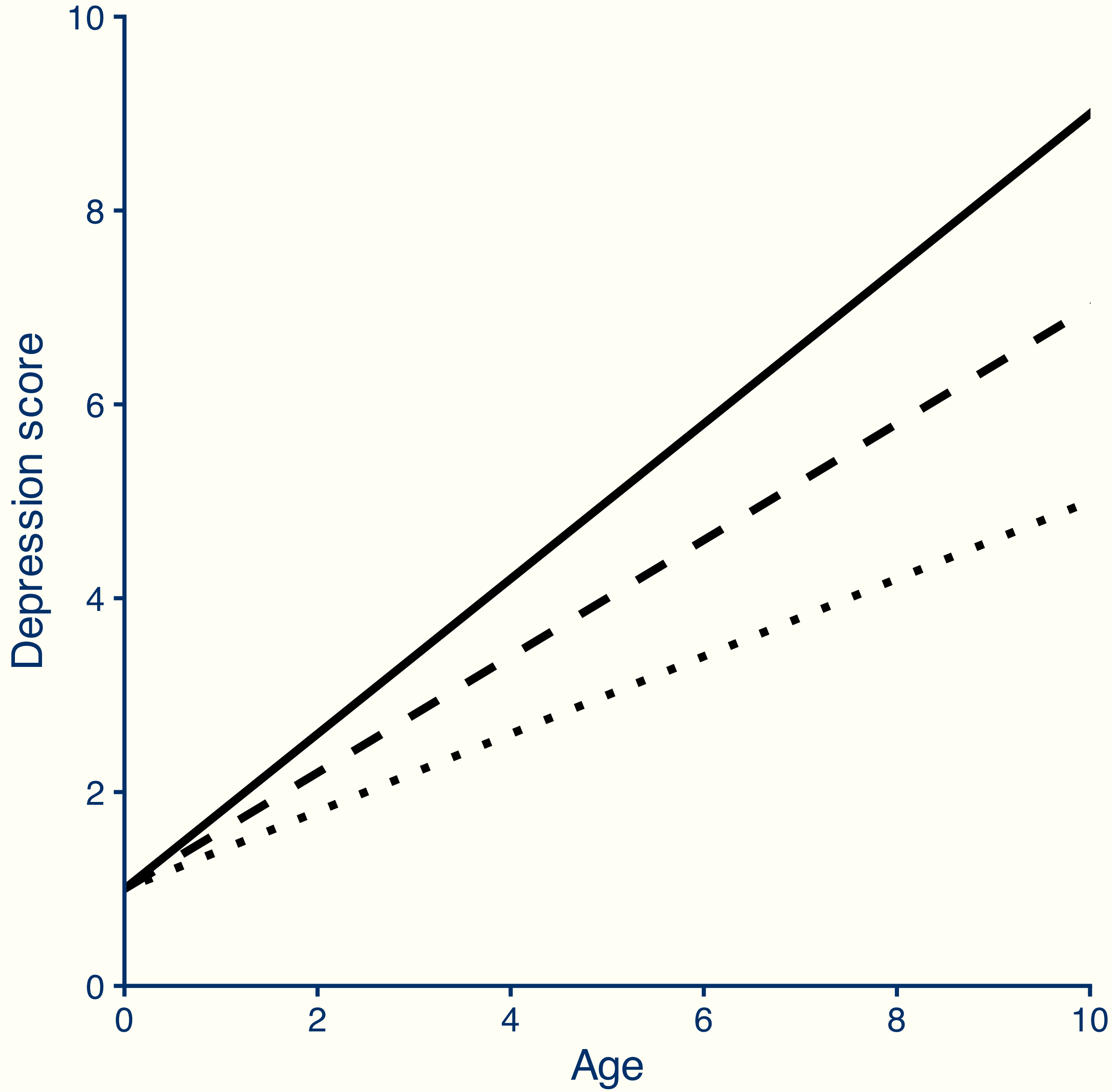
Example 2

- For each sketched regression model, write the mean function for a regression model that matches on the whiteboard
- Note that line type = education level
- Work with your neighbors

Model 1



Model 2



Model 3