Multiple Linear Regression

Applied Regression in R

Aleš Vomáčka

15. 02. 2026

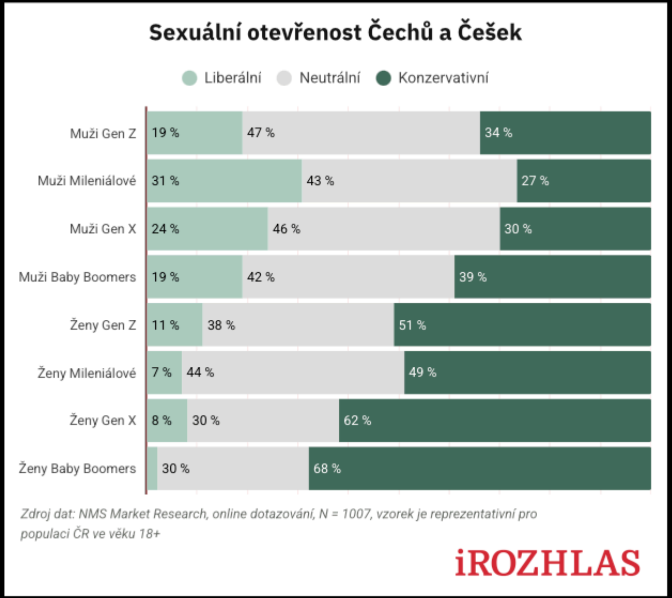
Faculty of Arts, Charles University

When a Line Is Not Enough

Multiple Regression

How sexually liberal are Czechs?

Jak Češi přistupují k sexu nebo OnlyFans? Generace Z je konzervativnější než mileniálové, ukázal průzkum

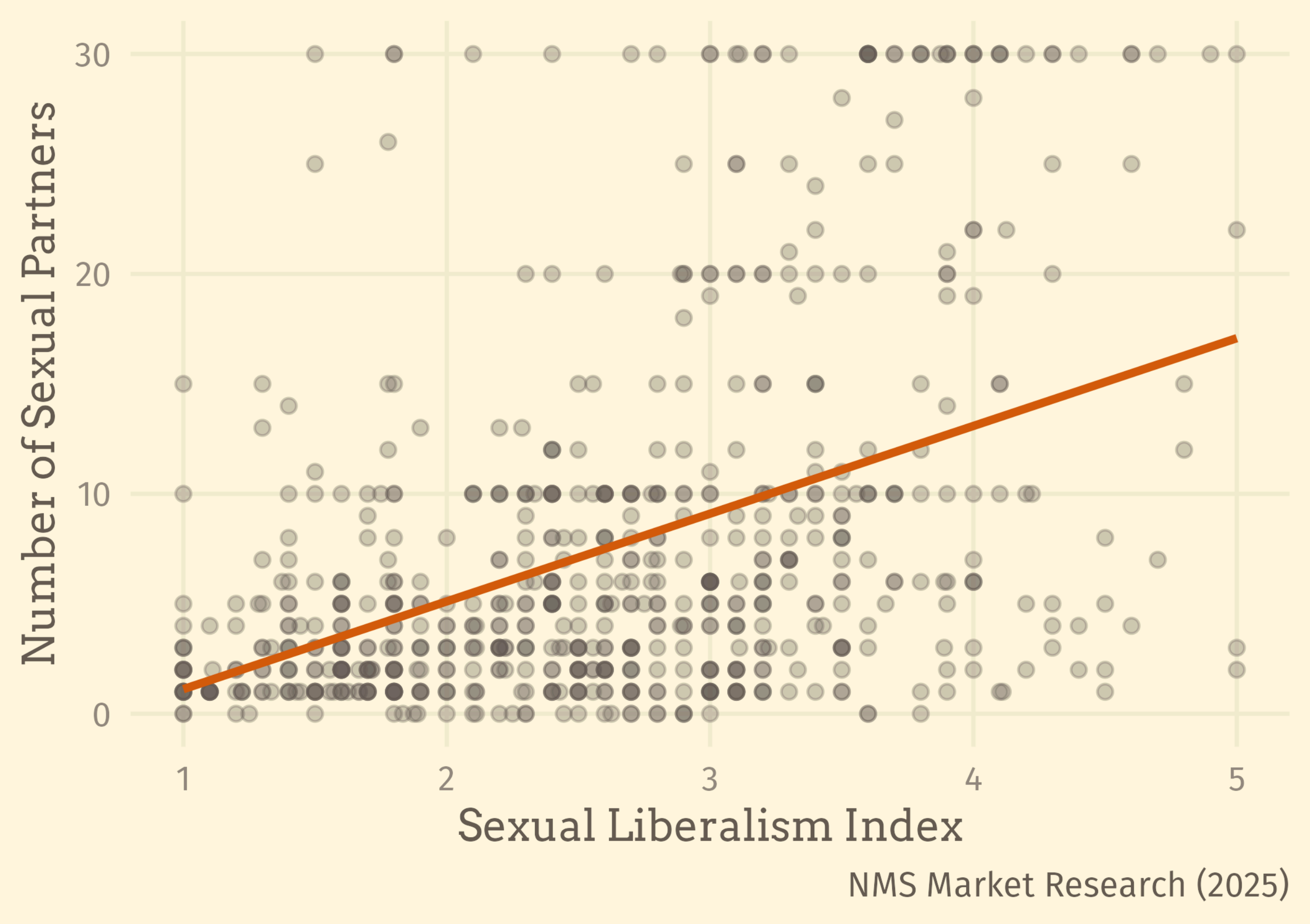


Do kin přichází dokument Virtuální přítelkyně, který odhaluje zákoutí erotické platformy OnlyFans (OF). Společnost NMS Market Research při této příležitosti vypracovala výzkum o sexuálním chování Čechů. Odhalil například, že generace Z je ohledně sexuality konzervativnější než mileniálové. Podle dat také každá jedenáctá žena spadající do generace Z někdy zvažovala tvorbu obsahu na OnlyFans.

PRŮZKUM Praha 8:00 9. listopadu 2025

Multiple Regression

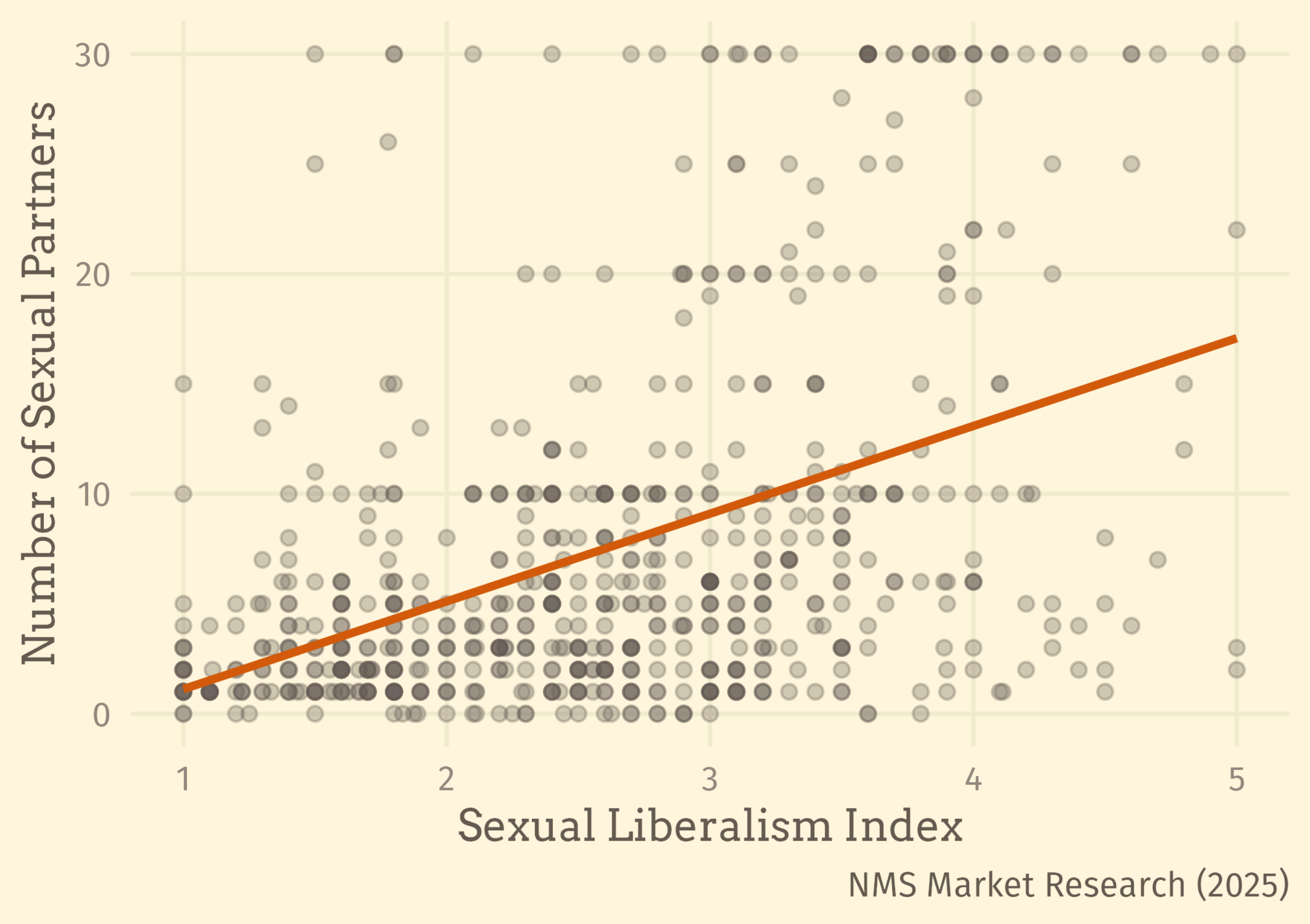
More sexually liberal people have on average more sexual partners.



Multiple Regression

But isn’t that just an effect of age?

We need a way to compare sexual liberalism and number of sex partners for **people of the same age**.



Multiple Regression

We already know how to create a simple regression.

```
lm(sex_partners ~ sexlib_index)
```

Multiple Regression

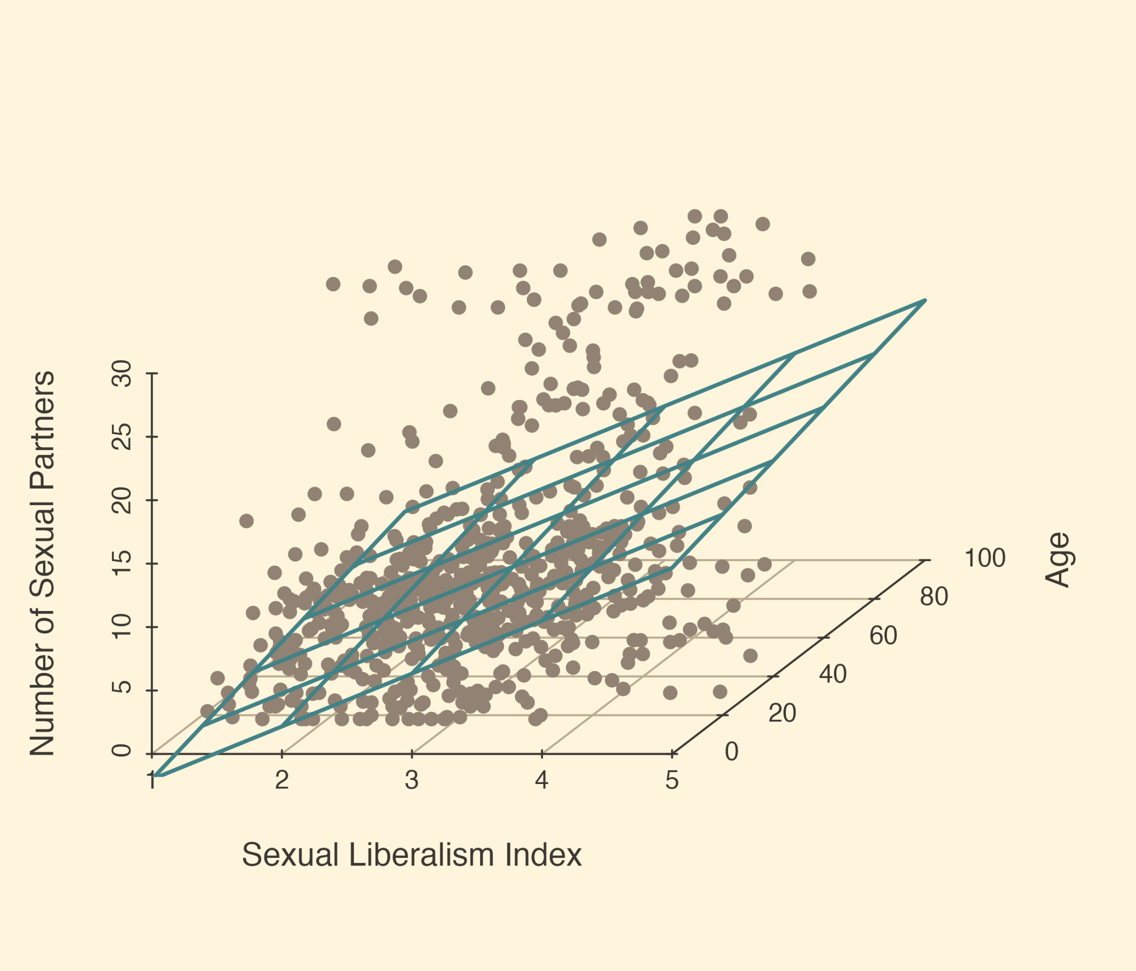
Controlling for age is easy, just add it in!

```
lm(sex_partners ~ sexlib_index + age)
```

This is called **multiple regression**.

Multiple Regression

Geometrically, we went from line to a **plane**.



When a Line Is Not Enough

Mathematically, we've added another parameter.

From this:

$$\text{E}(\text{Sexual Partners}) = \beta_0 + \beta_1 \cdot \text{sex liberalism}$$

to this:

$$\text{E}(\text{Sexual Partners}) = \beta_0 + \beta_1 \cdot \text{sex liberalism} \; + \beta_2 \cdot \text{age}$$

Multiple Regression

When not controlling for age, people with 1 unit higher sex index have on average **3.99** more sexual partners.

Parameter	Coefficient
(Intercept)	-2.87
sexlib_index	3.99

Multiple Regression

After, controlling for age, people with 1 unit higher sex index have on average **4.17** more sexual partners.

We have evidence, **the relationship isn't just due to age**.

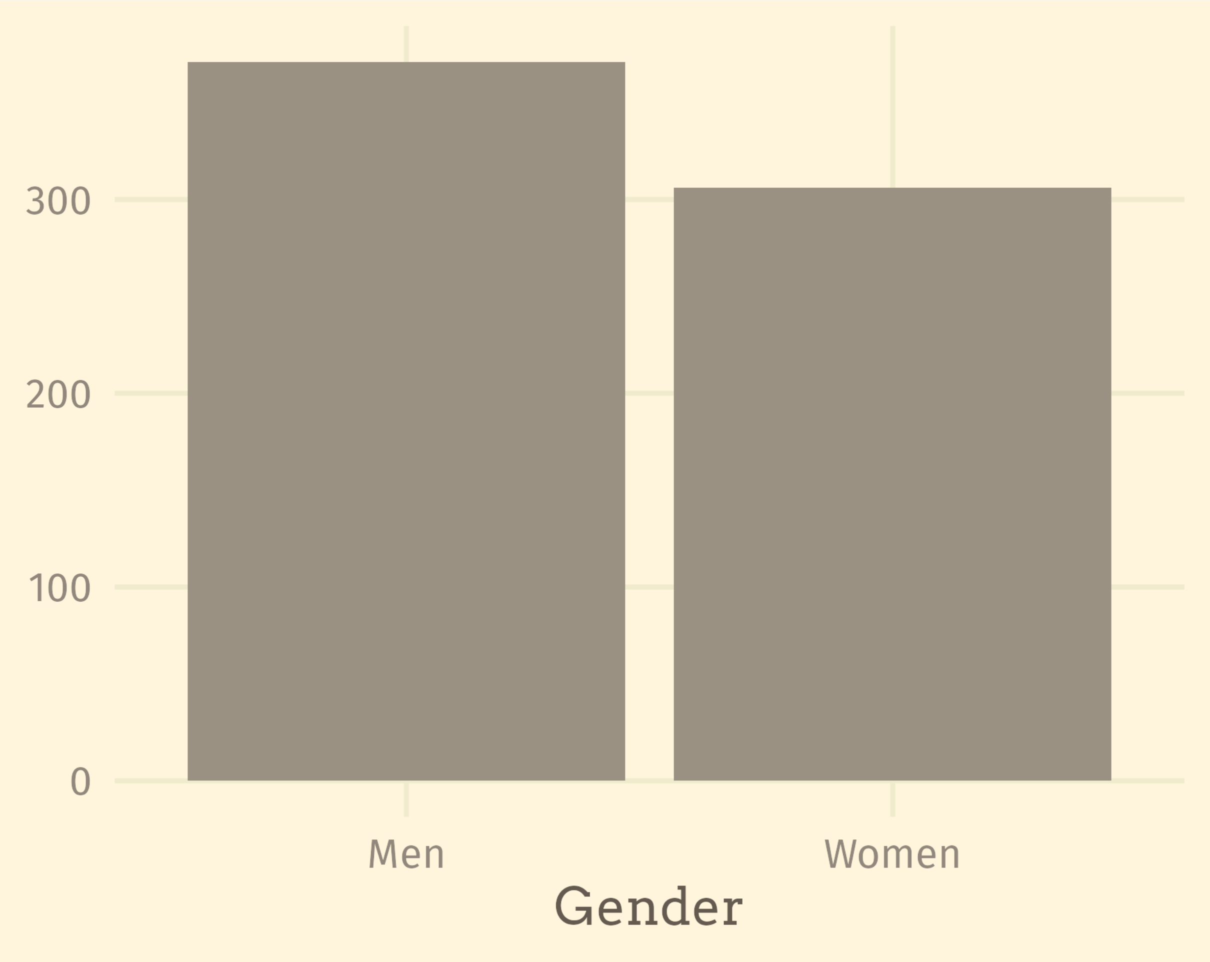
Parameter	Coefficient
(Intercept)	-6.1677
sexlib_index	4.1665
age	0.0581

Questions?

Categorical Predictors

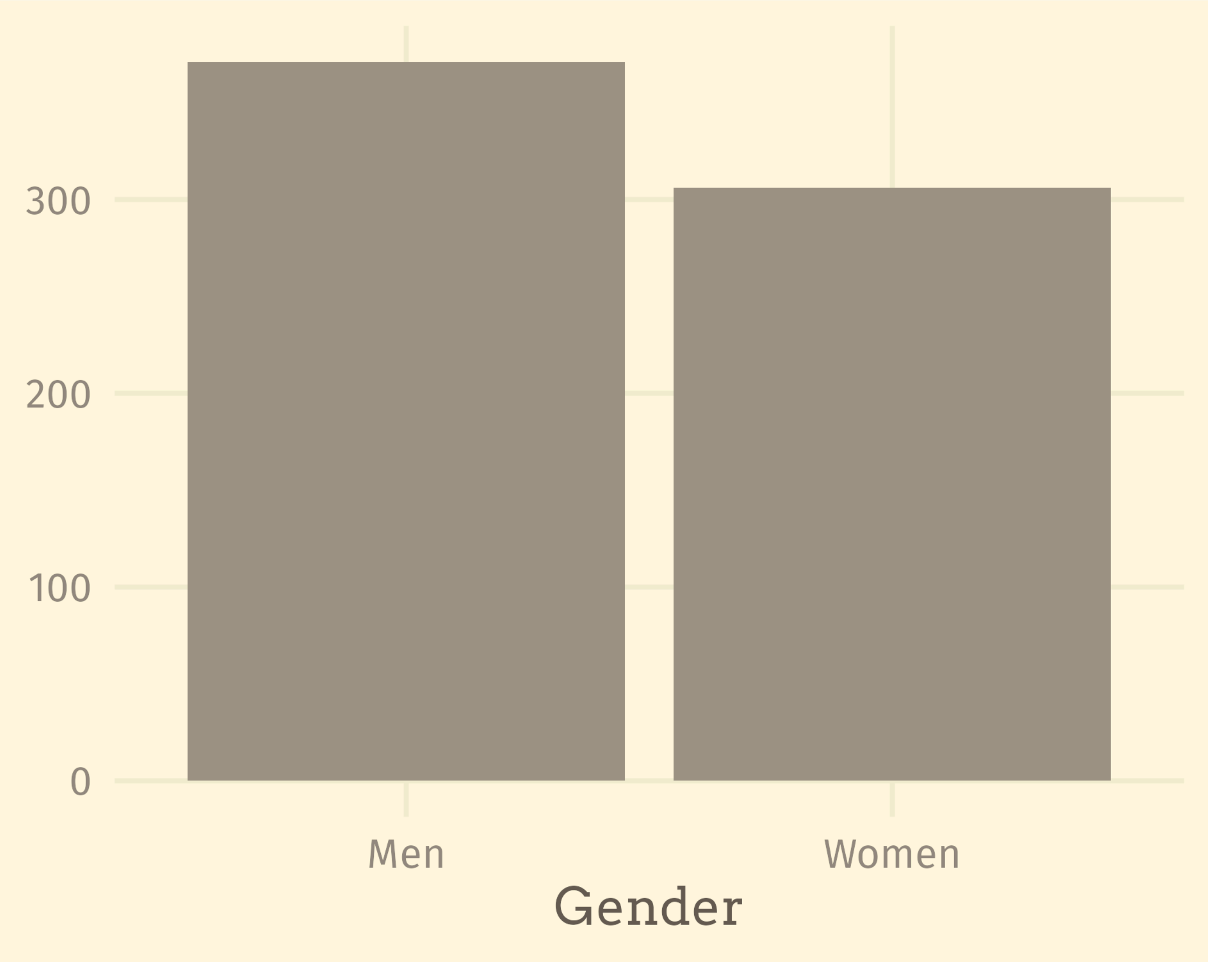
Categorical Predictors

The question: Do men have more sexual partners than women?



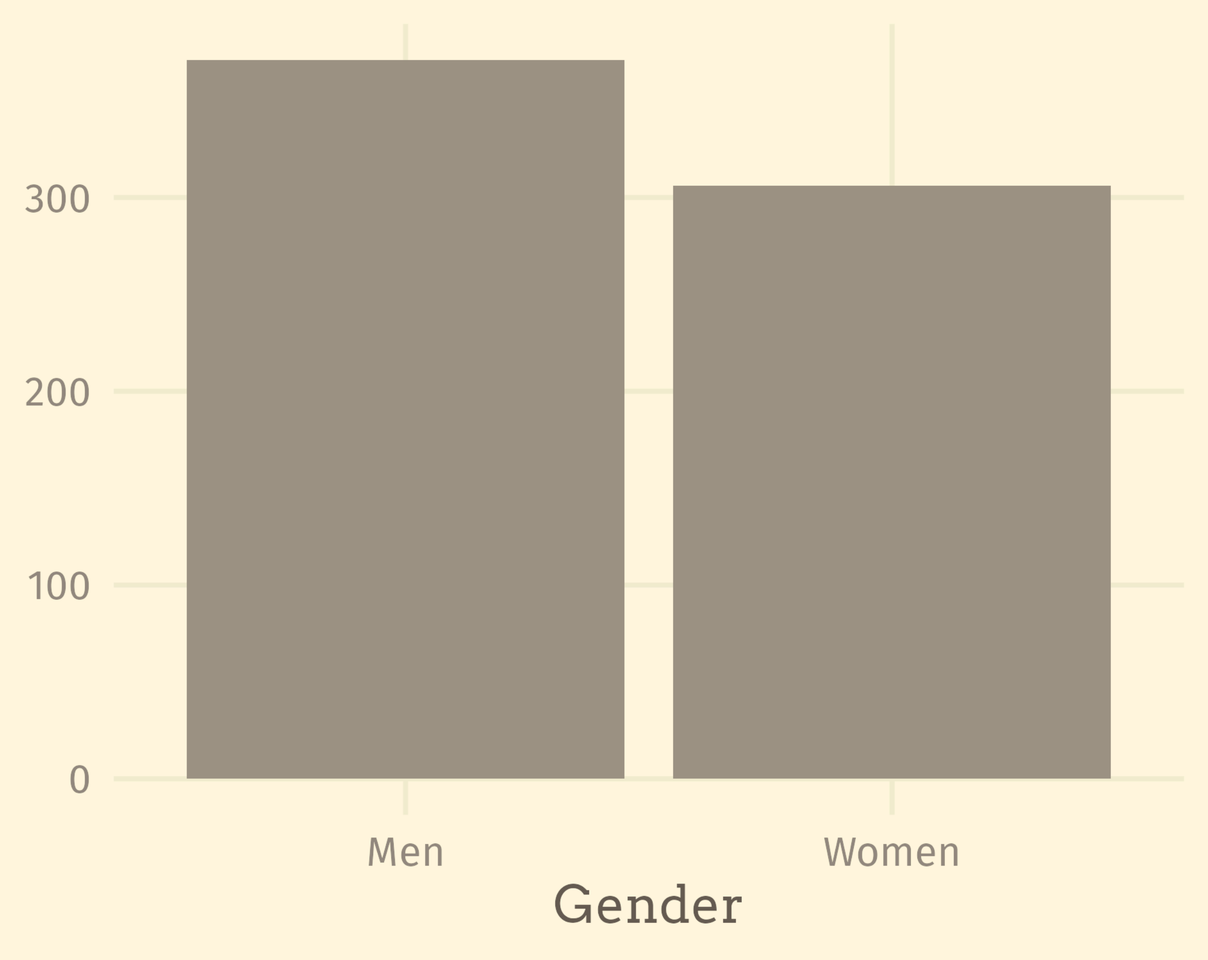
Categorical Predictors

The problem: Regression only works with numbers. How can we use it with categorical variables?



Categorical Predictors

The solution: Transform categorical variables into a set of numerical **dummy variables**.



Categorical Predictors

Dummy variables are **binary variables** indicating whether the observation belongs to the specific group or not.

Gender	Man	Woman
Man	1	0
Woman	0	1

Categorical Predictors

Dummy variables can be inserted into our model.

However, we'll run into a problem...

$$\mathrm{E}(\text{Sexual Partners}) = \beta_0 + \beta_1 \cdot \text{Men} + \beta_2 \cdot \text{Women}$$

Categorical Predictors

We have 3 regression parameters, but only 2 groups. Not enough information to estimate the model!

This is called the **dummy variable trap**.

$$\mathrm{E}(\text{Sexual Partners}) = \underbrace{\beta_0}_{??} + \underbrace{\beta_1 \cdot \text{Men}}_{\text{Information about men}} + \underbrace{\beta_2 \cdot \text{Women}}_{\text{Information about women}}$$

Categorical Predictors

Solution 1: The Reference category approach

Drop one of the category parameters (e.g. men). This will become the **reference category** for others.

$$\text{E(Sexual Partners)} = \beta_0 + \beta_1 \cdot \text{Women}$$

Difference between expected (average) number of sex. partners between women and men

Expected (average) number of sexual partners for men

Categorical Predictors

Solution 2: The Group intercepts approach

We drop the global intercept (β_0). The other parameters become **intercepts for each group**.

$$\mathrm{E}(\text{Sexual Partners}) = \boxed{\beta_1 \cdot \text{Men}} + \boxed{\beta_2 \cdot \text{Women}}$$

Expected (average) number of sex. partners for women

Expected (average) number of sexual partners for men

Categorical Predictors

Reference category approach

Parameter	Coefficient
(Intercept)	8.81
genderWomen	-2.81

Group intercepts approach

Parameter	Coefficient
genderMen	8.81
genderWomen	6

Categorical Predictors

R will do the dummy encoding for you.

```
Reference category approach:  lm(sex_partners ~ gender)

Group Intercepts approach:    lm(sex_partners ~ 0 + gender)
```

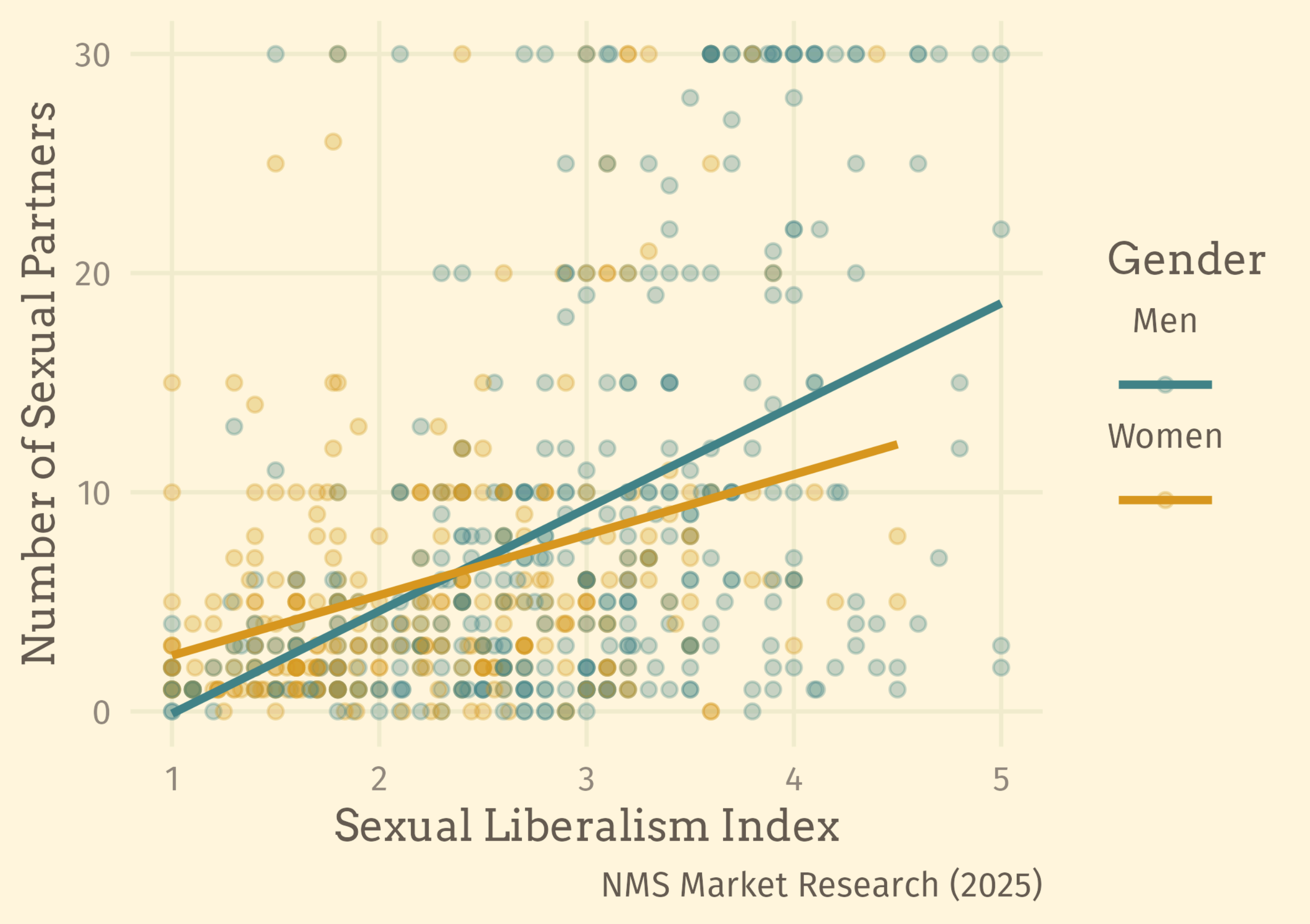
Questions?

Interactions

Interactions

Sometimes, we are interested how relationship between variables differs across subpopulations.

We can do this, using **interactions**.



Interactions

Interactions are **products** of two variables. They tell us how relationship between dependent and indepent variable changes based on third variable.

$$\mathrm{E}(\text{Partners}) = \beta_0 + \beta_1 \cdot \text{sex liberalism} + \beta_2 \cdot \text{gender} + \beta_3 \cdot (\text{sex liberalism} \cdot \text{gender})$$

Interactions

The relationship between Sex Liberalism and number of sexual partners is stronger among men than women.

Parameter	Coefficient	
(Intercept)	-4.78	
sexlib_index	4.68	The slope of Sex Liberalism for men is 4.68
genderWomen	4.58	
sexlib_index:genderWomen	-1.93	The slope of Sex Liberalism for women is 4.68 - 1.93 = 2.75

Interactions

In R, you can use * to create interactions between variables.

```
lm(sex_partners ~ sexlib_index * gender)
```

Or alternatively more explicit (same result)

```
lm(sex_partners ~ sexlib_index + gender + sexlib_index:gender)
```

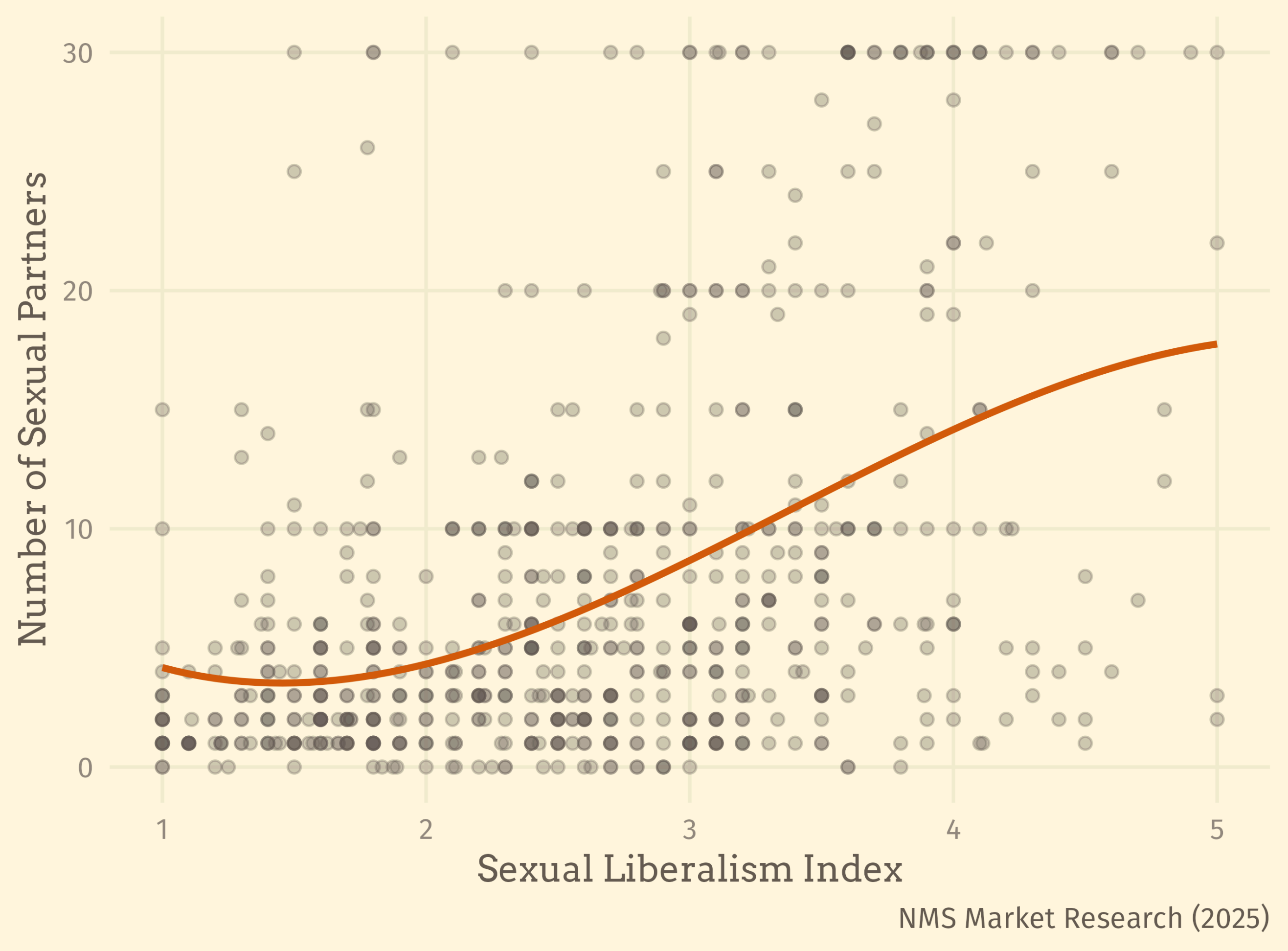
Questions?

Nonlinear Relationships

Nonlinear Relationships

Real-life relationships are messy. This is also true in statistics.

Linearity can be a good approximation, but sometimes we need to account for **nonlinearity**.



Nonlinear Relationships

There are many different ways to model nonlinear relationships.
We'll talk about them in more detail later.

The most popular are **simple polynomials**, e.g. x^2, x^3,x^4...

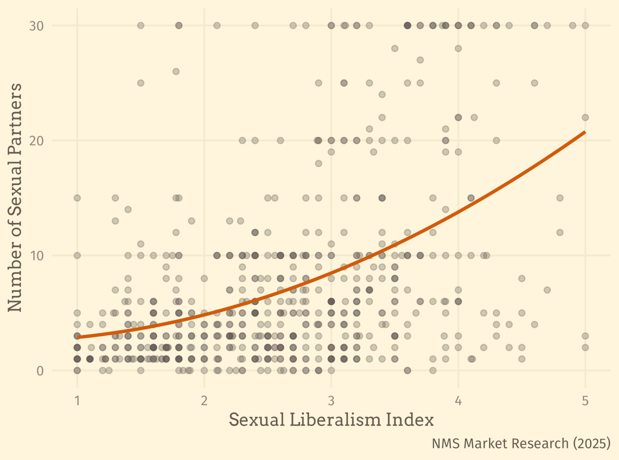
Nonlinear Relationships

To model nonlinear relationships using simple polynomials, we simply power transform the original parameters.

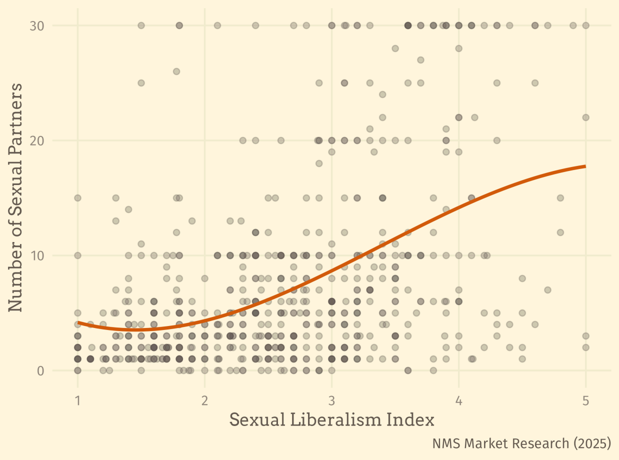
$$\mathrm{E}(\text{Sexual Partners}) = \beta_0 + \beta_1 \cdot \text{sex liberalism} \boxed{+ \beta_2 \cdot \text{sex liberalism}^2}$$

Nonlinear Relationships

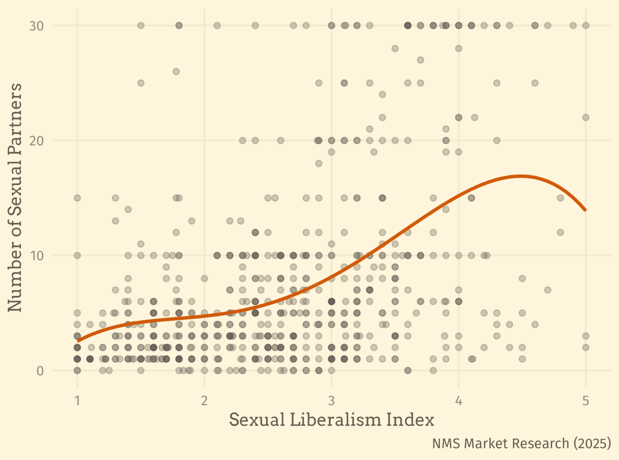
sex index^2



sex index^3



sex index^4



Nonlinear Relationships

In R, we can use the `poly()` function to create polynomials.

```
lm(sex_partners) ~ poly(sex_index, 2))
```

is equivalent to

$$\mathrm{E}(\text{Sexual Partners}) = \beta_0 + \beta_1 \cdot \text{sex liberalism} + \beta_2 \cdot \text{sex liberalism}^2$$

Nonlinear Relationships

Once we move to nonlinear relationships, models can no longer be interpreted using regression coefficients.

Don't worry, we'll solve this problem soon!

Parameter	Coefficient
(Intercept)	7.54
sex_index^1	93.9
sex_index^2	20.74

Questions?

Main Takeaways

- We can have **multiple predictors** in our models. This allows us to ask „*What is the relationship between variables among people who are the same in other characteristics?*“
- Categorical predictors are possible using **dummy coding**.
- We can explore how relationships change across subpopulations using **interactions**.
- We don’t need to assume straight lines. Modeling **nonlinear** relationships are possible using linear regression.

InteRmezzo!