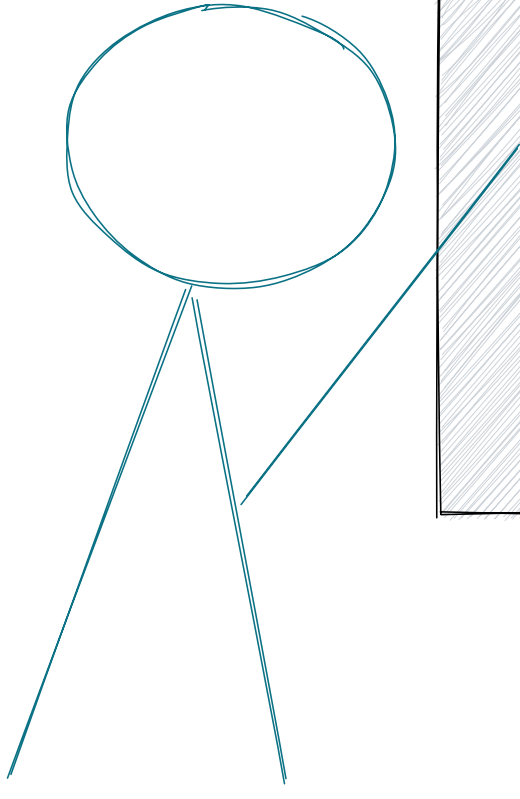


Categorical variables

EBA3500: Data analysis with programming

Lecturer: Jonas Moss
Contact: jonas.moss@bi.no

Office hours: 10-11 on Tuesdays.

Types of data

1.) Continuous / real-valued / decimal valued / numeric.

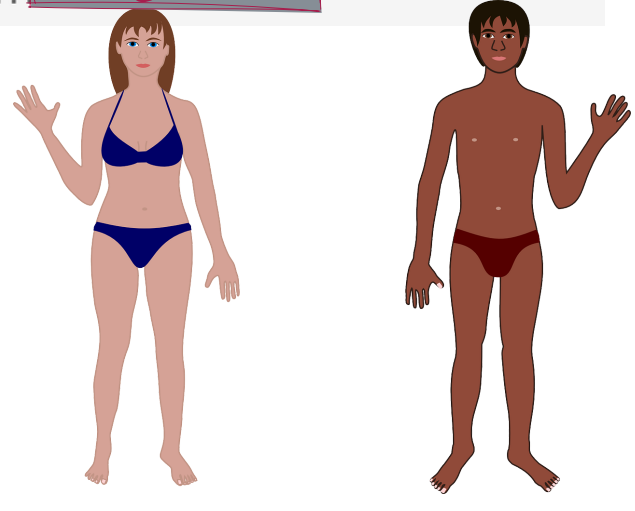
2.) Ordered / ordinal data

3.) Proper categorical data.

Continuous / real-valued / decimal-valued / numeric. These are `float` or `float64` in Python. A common example is height and weight. Most of our numeric data has been unbounded, i.e., it can take on arbitrarily large or small values.

Ordered / ordinal data. This is data equipped with an *order*. For instance, a gold medal is better than a silver medal, and a silver medal is better than a bronze medal. Education level is also ordinal. Having finished high school obviously scores as more education than having finished junior high, but it's unclear by how much. Continuous data is ordered, but not all ordered data is continuous!

Categorical data. Data with no order that's not naturally associated with a decimal number. Examples include gender, ethnicity, and major in college (sociology, economics, etc.). We often handle ordinal data as categorical data, just because it's easier. Categorical data can often be worked with using `set` objects and the `Counter` function from collections



We have until now looked at the case of:

Y : continuous

X : continuous

Covariate

Response!

Ordered response type is somewhat uncommon, but the other variants are very common!

A rough categorization of methods

A rough delineation of the different regression types are in this table, where ANOVA is an abbreviation of "analysis of variance".

	Response type		
Covariate type	Numeric	Categorical or binary	Ordered
Numeric	Linear regression	(Multivariate) Logistic regression	Ordered logit
Categorical	Linear regression / ANOVA	(Multivariate) logistic regression	Ordered logit
Ordinal	Constrained linear regression	Constrained logistic regression	Constrained ordered logit

The key message: It's the response type that primarily determines the regression method, not the covariate type! As we'll see, there is a general method for handling categorical data that can be used in any regression model that supports decimal data. This matters quite a bit, as it allows you to mix and match covariate types in a single model.

Definition: Categorical regression with one category

> Let A be a set. Then the linear regression model on this set of categories is
>
> $$y = \sum_{a \in A} \beta_a 1[x = a] + \epsilon, \quad x \text{ in } A.$$

Let A be a set of categories, for instance $A = \{\text{male}, \text{female}\}$. Recall the indicator function $1[x \text{ in } A]$, which equals 1 if x in A and 0 otherwise.

Interpreting the definition.

We can write the model

$$y = \sum_{a \in A} \beta_a 1[x = a] + \epsilon, \quad x \text{ in } A.$$

in a way that might be easier to interpret. Let $x = a'$ for some a' in A. By the definition of the indicator function, $1[x = a] = 0$ for all a in A except for a', where $1[x = a'] = 1$. Plugging this into the definiton of y, we find that $y = \beta_{a'} + \epsilon$.

For instance, if $a = \text{female}$, then $y = \beta_{\text{female}} + \epsilon$.

Live coding!

With Penguins!

The omnibus test

* Omnibus sounds scary, but is latin for "all".
* It also an archaic way of saying "bus".
* An omnibus test tests a bunch of coefficients at the same time!
* Needed if you have more than one category.

Formally, we are dealing with problems of the sort:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$$
$$H_1 : \text{some of the } \beta\text{s are not equal}$$

Live coding!

The F-test

- An F-test is a generic term for a testing problem where the distribution of the test statistic is F-distributed.
- We care about the F-distributed statistic above because it equals 0 if and only if R^2 is equal to 0. The distribution of R^2 is not commonly used, for reason I do not fathom.
- Instead, everyone uses the F-test. You can find its output in e.g. `statsmodels`.

We can write down the test statistic for the F distribution in several ways. This one is especially convenient for us:

$$\frac{n-K}{K-1}\frac{R^2}{1-R^2} \sim F(K-1, n-K)$$

Number of covariates (including intercept)

$$\frac{n-K}{K-1}\frac{R^2}{1-R^2} \sim F(K-1, n-K)$$

The F tests takes two parameters, d_1 and d_2, called the degrees of freedom. In this case, $d_1 = K - 1$ and $d_2 = n - K$. The parameter d_1 is sometimes called the numerator degree of freedom, d_2 the denominator degree of freedom.

Live coding!

With 'students' example.

Summary

1. There are roughly three kinds of data: Decimal, ordinal, and categorical.
2. The interpretation of categories variables in a regression model uses the indicator function:

$$y = \sum_{a \in A} \beta_a 1[x = a] + \epsilon, \quad x \text{ in } A.$$

3. `statsmodels` automatically fits a regression model of the kind above when given categorical covariates.
4. We can test if at least one of the coefficients $\beta \neq 0$ using the F-test.
5. Using the F-test is equivalent to testing if the R^2 is greater than 0.