Plan for today

Why go beyond the linear regression (OLS)?

Structure of the class

Statistical models beyond linear regression

Silje Synnøve Lyder Hermansen

09 februar 2023

Plan for today

Plan for today

1st hour

- why you should care about our topic
- practicalities:
  - how we work
  - the exam

2nd hour

- intro to R, our statistics program

Why go beyond the linear regression (OLS)?

Assumptions of the linear model

Assumptions of the linear model

Linear models (OLS) rely on two assumptions that are often violated

- outcomes are continuous and unbounded
- observations are independent and identically distributed (iid)

$\Rightarrow$ *this class: alternative models when these are not satisfied.*

Our research topics don't fit the OLS

- **Most phenomena in political science are not continuous**
  - (re)election, vote choice, degree of satisfaction, civil war, difficulty of negotiations, labor force participation...
- **...nor are they independent of each other**
  - same MP has an increased probability of reelection in several elections
  - several civil wars happen in the same country

Assumptions of the linear model

Assumptions of the linear model

Linear models (OLS) rely on two assumptions that are often violated

- **outcomes are continuous and unbounded (second part of class)**
- observations are independent and identically distributed (iid) (third part of class)

Assumption 1: continuous and unbounded outcomes

Assumption 1: continuous and unbounded outcomes

Outcomes are continuous and unbounded ("asymptotic")

- political science theories imply a relationship between two phenomena: x and y
  - $y = \alpha + \beta x$
- for each unit increase in x, y increases with β units

$\Rightarrow$ *this relationship is linear*

Violations of assumption 1

What happens if...

- β has digits, while y does not?
- x increases so that we pass what is feasible for y?
- the relationship between x and y is not linear?

Example: Outcome is continuous, but predictions unrealistic

- Theory: We may model life expectancy as a function of income: $age = \alpha + \beta * income$
- Data:

age (y)	income (x)
0	0
50	50.000
?	200.000

- Results: $age = 0 + 0.001 * income$
- Prediction (scenario): $200 = 0 + 0.001 * 200.000$

What is the problem?

Predictions are unrealistic because the relationship between x and y is not linear

- **Why is this a problem?**
  - predictions are wrong (least of our problems)
  - β is wrong (kind of sad)
  - standard error is wrong (catastrophy!)
- **How do we fix it?**
  - we can recode x: e.g. log-transformation, truncation, etc.
  - we can recode y
  - we can recode y and its relationship with $x \rightarrow$ *generalized linear models* (GLMs)

Our research strategy: GLMs

- **The model we choose depends on**
  - the "data generating process" (probability distribution)
  - the measurement-level of the dependent variable (a mental short-cut)
- **The GLM does:**
  - a recoding of the dependent variable to become continuous and unbounded
  - draws from a probability distribution

$\Rightarrow$ *We end up with a linear statistical relationship*

Examples

Prospecting for relevant models often looks something like this

Theoretical concept	Operationalization	Measurement level	Model choice
(re)election	are MPs in period 1 observed in period 2?	binary	logit
vote choice	party names	categorical	multinomial
degree of satisfaction	dissatisfied, OK, satisfied. . .	ordinal	ordered
civil war	# of dead people	count	poisson
difficulty of negotiations	length of proceedings	# days to conclusion	event-history
labor force participation	time to employment	# days in unemployment	event-history

Your turn

What kind of phenomenon are you interested in for your BA/MA/secred dreams?

- your name
- your topic

Assumptions of the linear model (recap)

Linear models (OLS) rely on two assumptions that are often violated

- outcomes are continuous and unbounded (second part of class)
- **observations are independent and identically distributed (iid)** (third part of class)

Assumption 2: observations are iid

Assumption 2: observations are iid

Observations are independent and identically distributed (iid)

- **independent**
  - the probability of observing one unit is not dependent on observing another
- **identically distributed:**
  - they come from the same *probability distribution*:
  - describes the *data generating process*
    - the shape of the relationship between x and y
    - the probability of an event (e.g. standard error)

Independent observations

Observations are not independent when they share characteristics (x) that may affect the outcome (y)

- missing data: may lead to a biased sample
- nested data: observations are correlated

$\Rightarrow$ *our* β *and standard error might be wrong*

Missing data

When we lack observations, and these observations are non-random, our sample is not representative

- **Diagnostics of problem**
  - Missing completely at random (MCAR): absence is not related to the observation
  - Missing at random (MAR): absence is related to observation, but not outcome
  - Missing not at random (MNAR): absence is related to observation + outcome → *problem!*
- **Solving the problem:**
  - Collect the data? Ignore it?
  - Impute the data?

⇒ *second part of the class*

Nested observations

We have nested observations when they belong to a group/share features

- e.g.: any panel data, civil wars in country, job-seekers in a locality, MPs in parties/committees/legislative periods...
- shared variation on x: a way to cluster standard errors
- relation to y: controlling for unobserved confounders

$\Rightarrow$ *some resemblance with MAR/MNAR*

GLMs in context

GLMs in context

There are other ways to approach statistics than what we will learn here:

- y-centred approaches
- x-centred approaches

$\Rightarrow$ *. . . but regressions remain the bread and butter of statistical analysis*

Y-centred/prediction approaches

Some statistical models are primarily predictive or descriptive

- machine learning: aim to predict outcomes at all costs
- text anlaysis: categorizations, scaling. . .
- network analysis: description of networks

What's in it for us?

- often use GLMs "under the hood"
- create variables we can use in a regression

X-centred/causal inference approaches

Some statistical models are primarily predictive or descriptive

- rely on one or two linear models:
  - diff-in-diff, RDD, matching + OLS
  - instrumental variable/ fuzzy RDD
- focus on theory; statistics are often very simple

What's in it for us?

- understanding regressions helps us understand causal inference
- often very narrow applicability

Structure of the class

Flow

Flow

We will progress through the semester in cycles

- We start with 3-4 calm weeks (learn R), then pick up pace (learn models)
- 2-week cycle:
  - week 1: lecture + reading
  - week 2: theory recap + seminar + homework assignment
  - before next lecture:
    - assignment is due. . .
    - . . . if you want feedback
- Final portfolio due end of May

Aim for the class

No magic, just hard work

My aim is to push you out of your comfort zone, and keep you there

- if you do the work. . .
  - readings
  - class activities
  - exercises
- . . . you will succeed

$\Rightarrow$ *A work-intensive class, but you don't have to be a genius*

Three aims

We will go through a series of models and learn

- when to use them
- how to use them + limitations
- how to understand the results

$\Rightarrow$ *The portfolio exam tests these learning outcomes. Class activities help you acquire them*

Aim 1: When to use a model

A mental map over data structures, different outcomes and what models to use

- Structure of class:
  - 2-week cycle on a *family of models*
- Group work
  - Twitter (@beyond_LM): “ID-card” on family
  - Presentation: theoretical “highlights” of family
- Exam:
  - executive summary of the class

$\Rightarrow$ *When you see data in the future, you know where you are and where to look for more info.*

Aim 2: How to use a model

Intuitive understanding of the models: estimation (in R) and assumptions

- Structure of class:
  - week 1: lecture on theory
  - week 2: R seminar
- Group work
  - Twitter (@beyond_LM): results from replication + R-codes on Absalon
  - Presentation: theoretical "highlights" of family
- Exam:
  - 2 replication exercises + critical assessment
  - you can hand in a draft for feedback beforehand

⇒ *Once you know some of these models, you have the intuition for regressions in general.*

Aim 3: How to understand the results

Interpretation and communication of results

- Structure of class:
  - week 1: what goes into the model (recoding + propabiliy distribution)
  - week 2: what comes out of the model (results)
- Group work
  - Twitter on week 2 (@beyond_LM):
    - text
    - numbers
    - visuals
- Exam:
  - take the model results seriously
  - go beyond the authors

$\Rightarrow$ *Communication == understanding, but also a superpower.*

Peer learning

Peer learning

This is a class designed for peer learning, because we learn much more

- **Group responsibility**: each group is responsible for a topic (Thu-Thu)
  - Captains of the class/"first responders": you help your peers, I help you
  - Twitter
  - Presentation (theory)/R-codes (replication)
- **Group exam**
  - you can coauthor the portfolio
- **Colloquiums**
  - meet up and exchange (codes, insights, feelings...)

A few hacks and other advice

A few hacks and other advice

- **Use your calendar:**
  - your group week is going to be busy
  - assignments are due after 6 days (if you want feedback)
  - replication group must share codes early
- **Group work prepares you for the exam**
  - tweets → mental map → executive summary
  - R-codes → how to → portfolio
- **Coauthor the exam**
- **Keep faith (in yourself)**
  - if you do the assignments, you pass the exam

Practicalities

Practicalities

- The final hand-in of the portfolio end of May
- Group weeks:
  - put your name down on the spreadsheet on Absalon
  - if not equal spread in people/groups, I will reshuffle