Week 5: Supervised Scale Measurement II: Regression

POLS0013 Measurement in Data Science

Dr. Julia de Romémont

Academic Year 24-25

UCL Departement of Political Science

Last lecture..

… we looked at competition data

- Cases where we do not have data directly on the the target concept, but ***comparisons*** (e.g. indvidual matches) **which *depend* on the target concept** (i.e. the better side won)[1]
- We discussed how the results of a pairwise comparison can be **modelled** as depending on the (difference in the) *latent* quantities of each side
- Introduced Bradley-Terry models (a special case of logistic regression) as a way to estimate these latent abilities
- We also talked about how, just because our **model *assumes*** a 'latent' variable, doesn't mean that it really exists!

[1] Ideally... Unless the ref was biased!

This lecture is about...

Training models

- Connecting an already existing measure of a target concept to a set of indicators (i.e. run a regression)
- In order to learn about the relationship between the measure and the indicators (i.e. get a regression equation)
- To then calculate the scores of measure for data where we have the indicators, but not the measure (i.e. calculate fitted/predicted values)

This is an appropriate measurement strategy ***only*** when you already have **some** data for which the measure exists!

This lecture

From models to measures

Making sensible models

Applications

From models to measures

Indicators

Indicator

$\equiv$ An already measured quantity that provides evidence (i.e. indicates something) regarding the concept that we aim to measure.

- An indicator is one we believe **indicates** something about the presence/absence of the target concept.
  - E.g. the distribution of incomes tells us something about economic inequality
- Generally, we think of indicators as *partial* and/or *noisy* signals of the target concept.
  - E.g. the better side might not always win (a *noisy* signal), or winning might depend on other factors other than being good at the game (a *partial* signal)

Connecting indicators and target concept

But **how** and **to what extent** does an indicator ***translate*** into the target concept? **How** can we ***connect*** the target quantity to the indicators?

These are questions about the *relationship* between the two!

- Two weeks ago, we considered cases where we had theoretical arguments linking indicators to the target concept.
- One week ago, we considered cases where single indicators (competitions) were directly dependent on the target concept.
- This week, we are considering cases where we *estimate* the indicator-concept relationship.

Training data

- To do this, we need "gold standard" measurements m from some pre-existing measurement procedure for the concept of interest μ
  - "Gold standard" implies that the chosen m should be the best available approximation of μ (i.e. with low ϵ_m)
  - This is often called the *training* data, which we are using to *calibrate* a new measurement procedure.
- Our goal is to determine how to most effectively use one or more indicators I (I_1, I_2, etc) to approximate μ
  - Given the indicator variables I that we have
  - Using the information contained in m about how they relate to μ.

Nothing new

To fit any regression model we need:

1. Data on the dependent variable Y
2. Data on the independent variable(s) X_k

$$Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$$

Translated to the context of measurement:

1. The dependent variable is our measure m
2. The independent variables are the indicators I_k

$$m_i = \alpha + \beta_1 I_{1i} + ... + \beta_k I_{ki} + \epsilon_i$$

Applying the trained model

We can then use the *trained* model (i.e. the $\hat{\beta}'s$) that we estimated in this calibration exercise...

- ...combined with the indicator values I for the new units for which we want to measure μ...
- ...to do the actual **measuring**.

Our new **measure** of μ for a given unit i is then the fitted value

$$\hat{m}_i = \hat{\beta}_0 + \hat{\beta}_1 I_{1i} + ... + \hat{\beta}_2 I_{2i}$$

The errors/residuals $\epsilon_i = m_i - \hat{m}_i$ from this regression are the **measurement error** $\epsilon_m = m - \mu$

Requirements

There are two key things that **must be true** to make this approach useful:

1. We have a gold standard measure m of the target concept μ for some units, but lack that measure for other units for which we want to measure μ
2. We have one or more indicators I that predict/indicate something about the target concept μ for *all* units

Applicability

1. When the training data is costly to construct.
   - i.e. Situations where humans can provide gold standard evaluation, but it is “expensive”
   - e.g. Medical diagnosis (e.g. heart disease example from POLS0010), quantitative text analysis
2. When the training data is only available for the past.
   - i.e. Decisions which can be evaluated in retrospect
   - e.g. loan/mortgage granting, school/university admissions
3. When the training data is only available for a different population of unit than the one you are interested in
   - i.e. Situations where the population of interest is hard to access directly
   - e.g Multilevel Regression and Post-Stratification (MRP), Leave vote estimates by constituency

Making sensible models

Decisions, Decisions...

When constructing a measure this way, there are three elements on which one needs to make decisions.

1. Training data m
2. Indicator data I
3. Model $f()$

Making sensible choices in all these three domains is important to generate useful measures.

Training data m

1. The gold standard measure should be of *high quality* (obviously).
   - Deviations of the gold standard measure m from the target concept μ should be small…
   - …and not associated with quantities relevant to the intended application
2. The training data set should be *representative* of the population where you want to apply the measurement procedure.
   - The relationships between the indicators and target concept among the training units should be *transferable* to the other units.
   - Differences in the parameters (i.e. the model) for the units in the calibration (training) set versus the target population should be small…
   - …and not associated with quantities relevant to the intended application.

Indicator data I

- The indicator data should be of *high quality* (obviously).
- In other words, the indicators need to be sufficiently predictive of the gold standard measure such that
  - the residual error of the regression is small and
  - is not associated with quantities relevant to the intended application.
- Note that there is - as always - a bit of a tension here between more or less supervision of the measurement procedure
  - More *supervised* selection of indicators might help ensure that the indicators we choose make 'substantive' sense (to us)
  - More *unsupervised* selection of indicators might help ensure that the chosen indicators are (statistically) predictive of m

Model choice for $f()$

- $f()$ denotes the (generally unknown) function which connects the inputs (here: indicators) to the output (here: measure).
- *Statistical learning* (or: machine learning) then refers to a set of approaches for estimating (i.e. learning) f
  - For more on Statistical Learning, see James et al
- Basic linear regression is only *one* of the many tools available to connect indicators and measure
  - Many more parametric, semi-parametric or non-parametric models available
    - E.g. Interactions, non-linearities, random effects, generalised linear models, Support Vector Machines, Neural Networks etc
  - Regularisation methods to deal with many indicators, e.g. ridge, lasso
  - Machine learning methods for model *assessment* and model *selection*, i.e. cross-validation

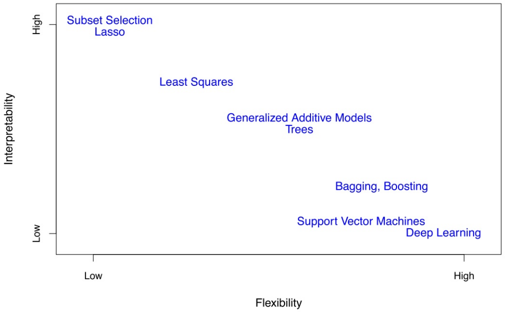

FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

James et al. (2022)

Yet another trade-off

- More complex/flexible models generally **perform better** (i.e. are more predictive of the outcome) than less complex ones
  - *Provided the variables/indicators are high quality!*
  - Think about how the R^2 increases as the number of independent variables X increase
- More complex/flexible models are **harder to understand/interpret** than less complex ones
  - It becomes increasingly hard to know what the relative weight of each individual X variable is

Connections to machine learning

- The goal here is the generic goal of all regression methods:
  - to use the indicators to best approximate the target concept
  - i.e. to minimise the (test) mean square error $\frac{1}{N}\sum_i (\hat{m}_i - m)^2$
- We are interested in **predictive performance**, not coefficients
  - We care about estimating the concept of interest $\hat{\mu}$ rather than about the estimates of any parameters like $\hat{\beta}$
- We are interested in **out-of-sample** predictive performance
  - The focus is on the target population of units where you want to apply the measurement strategy, not the training set.
  - Use adjusted R^2 not R^2
  - More generally, use machine learning tools for model assessment like cross-validation

Applications

Example 1: Transport Misery

Indicator data I

- Timetable data on trains and buses and delays / cancellations
- User data from tapping in / out
- Some ability to calculate crowding of carriages / buses

Training Data m

- Some kind of user survey of how people evaluated their transit “today”
- Linked to indicator data about their journey and on what happened on it

Model $f()$

- If you can predict evaluations $\hat{m}$ using features of journeys I where you collected m, you can predict the average subjective evaluations for all other journeys on the system too.

Design considerations

Training Data Quality

- Important that the subjective journey evaluations in your survey are really what you wanted to measure!
- The form of the survey prompt would need to be carefully considered.

Representative Training Set

- Do you want a representative sample of *journeys* or of system *users*?
- How could TfL get this?

Indicator Quality

- Is the timing/crowding data sufficiently high quality to find a clear signal?
- I have no idea, this is a made up example!

Model Choice

- Given the indicators, did we choose the model that yields the most accurate predictions?
- Can try different models and compare predictive performance

Recall from Lecture 1

- "The Global Health Security (GHS) Index is the first comprehensive assessment and benchmarking of health security and related capabilities across the 195 countries ..."
- "... the GHS Index will spur measurable changes in national health security and improve international capability to address one of the world's most omnipresent risks: infectious disease outbreaks that can lead to international epidemics and pandemics."
- "... a detailed and comprehensive framework of 140 questions, organized across 6 categories, 34 indicators, and 85 subindicators to assess a country's capability to prevent and mitigate epidemics and pandemics."

Can we learn something from COVID?

- The original measurement strategy combined these indicators in a way that doesn't predict COVID death rates in a useful way
  - In fact, better preparedness scores predicted *higher* death rates, not lower death rates!
- But can we use the data we now had on which countries actually performed well to figure out which indicators mattered?
  - Well, we *can*, but it turns out this is *not* a good application for this approach, for reasons that will become clear as we try to do it!

What are the indicators?

X1.2.1c..Cross.ministerial.department.agency.unit.for.zoonotic.disease
X1.2.2..Surveillance.systems.for.zoonotic.diseases.pathogens
X1.2.2a..Surveillance.reporting.mechanism.for.zoonotic.disease.for.livestock.owners
X1.2.2c..Wildlife.zoonotic.disease.surveillance
X1.2.3..International.reporting.of.animal.disease.outbreaks
X1.2.4a..Number.of.veterinarians.per.100.000.people
X1.2.4b..Number.of.veterinary.para.professionals.per.100.000.people
X1.2.5..Private.sector.and.zoonotic.disease
X1.3..Biosecurity
X1.3.1..Whole.of.government.biosecurity.systems
X1.3.1a..Updated.national.records.of.especially.dangerous.pathogen.toxin.inventories
X1.3.1c..Agency.for.enforcement.of.biosecurity.laws.regulations
X1.3.3..Personnel.vetting..regulating.access.to.sensitive.locations
X1.3.4a..National.transport.regulations.for.Category.A.and.B.infectious.substances
X1.3.5..Cross.border.transfer.and.end.user.screening
X1.3.5a..Laws.regulations.on.cross.border.transfer.and.end.user.screening
X1.4..Biosafety
X1.4.1..Whole.of.government.biosafety.systems
X1.4.1b..Agency.for.enforcement.of.biosafety.laws.regulations
X1.4.2a..Biosafety.training.using.a.standardised..required.approach
X1.5.1..Oversight.of.dual.use.research
X1.5.1b..National.law.regulation.on.oversight.of.dual.use.research
X1.5.2a..Requirement.to.screen.synthesised.DNA.against.list.prior.to.sale
X1.6..Immunisation
X1.6.1a..Immunisation.rate.for.humans..measles.MCV1.

X2.1..Laboratory.systems
X2.1.1..Lab.capacity.for.detecting.priority.diseases
X2.1.1a..Capacity.of.national.lab.system.to.conduct.5.or.more.WHO.core.tests
X2.1.2..Specimen.referral.and.transport.system
X2.1.3..Laboratory.quality.systems
X2.1.3a..Existence.of.an.accredited.national.lab.serving.as.a.reference.facility
X2.2.1a..Evidence.of.ongoing.event.based.surveillance.and.analysis
X2.2.1b..Evidence.of.reporting.a.potential.PHEIC.to.the.WHO..last.2.years.
X2.2.2a..Electronic.national.and.sub.national.reporting.surveillance.system
X2.2.2b..Collection.of.ongoing.real.time.lab.data.by.electronic.surveillance.system
X2.2.3..Transparency.of.surveillance.data
X2.2.3a..Availability.of.de.identified.health.surveillance.data.on.disease.outbreaks
X2.2.4b..Inclusion.of.cyber.protections.in.health.data.confidentiality.law.regulation
X2.2.5..Coverage.and.use.of.electronic.health.records
X2.2.5b..Public.health.system.access.to.individual.electronic.health.records
X2.2.5c..Existence.of.data.standards.for.health.record.data.comparability
X2.3..Epidemiology.workforce
X2.3.1a..Access.to.field.epidemiology.training.program.in.country.and.or.abroad
X2.3.2..Epidemiology.workforce.capacity
X2.3.2a..Evidence.of.at.least.1.trained.field.epidemiologist.per.200.000.people
X2.4.1..Data.integration.between.human.animal.environmental.health.sectors
X2.4.1a..Mechanisms.for.ministries.to.share.animal.human.wildlife.surveillance.data
X3..RAPID.RESPONSE.TO.AND.MITIGATION.OF.THE.SPREAD.OF.AN.EPIDEMIC
X3.1.1a..National.emergency.response.plan.for.diseases.with.pandemic.potential
X3.1.1b..National.public.health.emergency.response.plan.updated.in.past.3.years
X3.1.1d..Existence.of.public.pandemic.influenza.preparedness.plan.updated.since.2009

An example model

- We could decide which variables to use based on theoretical reasoning
- Here's an example of a linear regression with some indicator variables that are possibly predictive of our concept of interest

```
lm_fit <- lm(log(deaths_per_1m) ~
               BG4..Human.Development.Index..2018. +
               X3.7.1b..Alignment.of.movement.restrictions.with.WHO.OIE.regulations.r
               X6.2.4a..Public.confidence.in.government +
               X6.2.5a..Robust..open..diverse.local.media.and.reporting,
             data=deaths)

# could use packages stargazer or modelsummary creating regression table
```

An example model

	Total Deaths per m (log)
HDI	7.519***
	(0.700)
Movement Restrictions Aligned with WHO	0.568
	(0.433)
Public Confidence in Government	−0.519***
	(0.165)
Robust, open and diverse local media	0.521***
	(0.149)
Intercept	−0.229
	(0.637)
Observations	179
Adjusted R^2	0.462
Note:	*p<0.1; **p<0.05; ***p<0.01

Choosing the model

- Alternatively we could automate the process by using R figure out for us which variables are most predictive of the outcome variable
- This can be done with with **regularisation** methods, of which an example is LASSO (Least Absolute Selection and Shrinkage Operator)
- LASSO tries to find the model that predicts the most variance with the least possible number of covariates
- Specifically, it estimates parameters that minimise the sum of squared errors with a penalty for complexity (which shrinks coefficients of less important variables to 0)

$$\beta_{lasso} = \arg\min_{\beta} \left[\sum_{i=1}^{n} (Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}))^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

The important question is which λ to choose!

Regression with LASSO

```r
# load package and prepare the inputs
library(glmnet)
x <- as.matrix(deaths_complete[,-1])
y <- log(deaths_complete$deaths_per_1m)
```

```r
# run least square lasso with cross validation to choose lambda
set.seed(1234)
lasso.fit <- cv.glmnet(x,y)
```

```r
# lambda values and number of coefficients
lasso.fit
```

```
##
## Call:  cv.glmnet(x = x, y = y)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure     SE Nonzero
## min 0.0518    35   1.703 0.2091      49
## 1se 0.3330    15   1.900 0.2201       7
```

Regression with LASSO

```r
# extract the coefficients the LASSO has identified
lasso.coef <- as.matrix(coef(lasso.fit)[coef(lasso.fit)[,1]!=0,])
lasso.coef <- as.matrix(lasso.coef[order(lasso.coef[,1],decreasing=T),])
lasso.coef
```

```
##                                                                              [,1]
## (Intercept)                                                                 1.355
## X5.6.1b..Evidence.of.non.compliance.with.sample.sharing.element.of.PIP.framework  0.671
## X6.2.5a..Robust..open..diverse.local.media.and.reporting                    0.093
## X6.5.2a..Access.to.potable.water                                            0.020
## X6.2.1a..Adult.literacy.rate..15..years.old..both.sexes.                    0.016
## X6.5.2b..Access.to.at.least.basic.sanitation.facilities                     0.009
## X4.1.2a..Hospital.beds.per.100.000.people                                   0.000
## X6.2.3a..Poverty.headcount.ratio.at..1.90.a.day..2011.PPP.....of.population.  -0.007
```

```r
# use those variables for regression
deaths_lasso <- deaths[,c("deaths_per_1m",row.names(lasso.coef)[-1])]
lm_fit2 <- lm(log(deaths_per_1m) ~ ., data=deaths_lasso)
```

Regression with LASSO

	Total Deaths per m (log)
PIP non-compliance	4.941***
	(1.204)
Robust, open and diverse local media	0.398***
	(0.138)
Access to potable water	0.018
	(0.012)
Literacy rate	0.022***
	(0.008)
Access to basic sanitation	0.007
	(0.008)
Hospital beds per 100thds	0.001*
	(0.001)
Poverty	−0.031*
	(0.016)
Intercept	−3.587**
	(1.475)
Observations	178
Adjusted R^2	0.574

Too many indicators, not enough data points

- We only have about 170 countries to work with in these data, but there are 132 indicator variables.
  - How do we figure out which of the indicators might have been important?
  - There are some machine learning techniques that are a bit helpful here, but not helpful enough.
  - Difficult to generate even a moderate adjusted R^2.
- This is actually more of a causal inference problem than a measurement problem!
  - Pandemic preparedness simply did not seem to have a strong effect on deaths from Covid-19
  - It may well be that a useful measure of pandemic preparedness can be constructed from these variables, but Covid-19 deaths are not a good 'gold-standard' measurement.

Design considerations

Training Data Quality

- Not great: there are many things besides pandemic preparedness that have had effects on death rates.

Representative Training Set

- Not amazing: The training data are representative with respect to countries, but probably not with respect to other possibly pandemic diseases.

Indicator Quality

- Not very good: none of the indicators are very predictive and there are too many of them relative to the number of observations in the training set.

Model Choice

- Meh: There is little what we could have done by ways of more complex modelling here, given the low training and indicator quality
- Fancy modelling will not save you if the trainig data is low quality! (garbage in, garbage out)

Compare to textbook example

- In the "what is a curry" example in the textbook, there were many indicators (about 70), but many more data points (about 8000)
  - This meant there was a relatively strong signal about which indicators were associated with being called a curry.
- In this example, we have
  - More indicators (132) and fewer observations (170).
  - Generally weak bivariate relationships between the indicators and the outcome (death rates)
  - Not enough information!

Summing up

- When we don't have good theoretical intuitions about how several, different indicators should be aggregated into a measure...
- ... we can try to **estimate** that relationship instead, as long as we have some already pre-existing **gold-standard measurements** available.
- In that case, we can **train a measurement model**[2] with the indicators as independent variables and the gold-standard measurements as dependent variables.
- The model can then be used to **predict** the outcome (i.e. create measurement estimates) for other units for which we have indicator data but no gold-standard measurements available.

[2] i.e. run a linear regression, or some other more or less complicated model