Week 4: Supervised Scale Measurement I: Comparison Data

POLS0013 Measurement in Data Science

Dr. Julia de Romémont

Academic Year 24-25

UCL Departement of Political Science

Last lecture..

...we translated concepts into formulas

- Where the concept we have in mind is translatable (by us!) into a mathematical function through which the relevant indicator(s) aggregated into the measure
- Such cases are necessarily concept-specific, but there are two strategies to check the internal logic of the function/equation:
    1. Dimensional Analysis[1]
    2. Axiomatic Analysis[2]
- Possible when the indicator data is 'close' to the target concept

[1] Are we combining things that we can combine?

[2] Does the measure behave in ways that we want it to behave?

This lecture is about...

Competition as Measurement

- The **measurement problem** we want to solve is *scoring/ranking*
  - That is, assessing the *relative* degree to which units have some concept of interest.
- The **kind of data** we will use to solve that problem is *competition/comparison data*
  - That is, direct comparisons between the units that depend on the concept of interest.
- The indicator is still 'close' to the target concept, but not in a way that we can come up with a a good mathematical aggregation ourselves
  - $\longrightarrow$ We will need a *model* to *estimate* the aggregation formula

This Lecture

Scoring wins and losses

Quick logistic regression recap

Bradley-Terry models

Bradley-Terry Model implementation

Designing competition data collection

Interpreting latent variables

Scoring wins and losses

Natsu Basho, May 2012, Tokyo

Kyokutenho, winner of the Natsu Basho, Tokyo, May 2012.

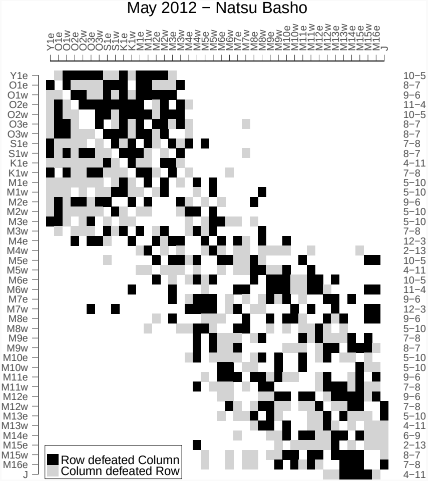
In May 2012, wrestler Kyokutenhō won the Summer Tournament of the Honbasho by defeating Tochiōzan in a playoff.

He had been in professional sumo for over twenty years and was 37 years and 8 months at the time, a record winning age in modern sumo history.

He retired in July 2015, holding the all time record for most bouts (1445) in the top division.

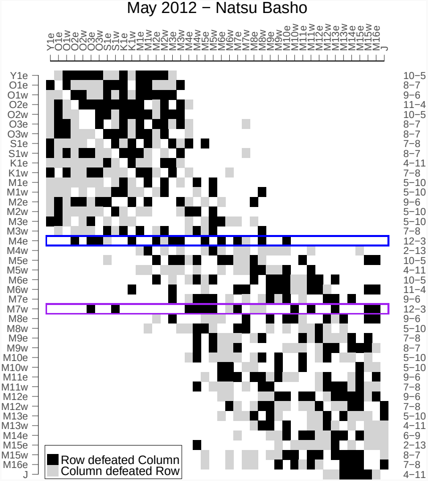
His career record was a mediocre 927 wins and 944 losses.

Sumo tournaments are badly designed



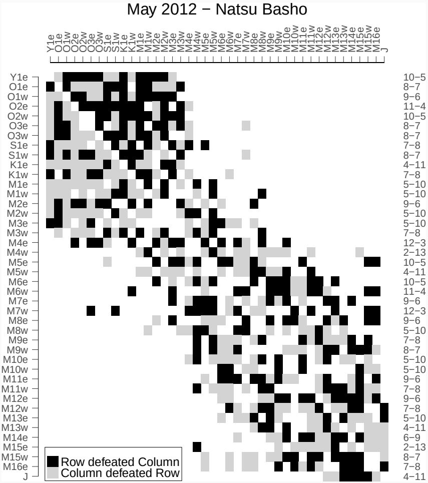
- Results of sumo matches in May 2012 top division Natsu Basho, sorted by pre-tournament wrestler ranks.
- Black (win) and grey (loss) squares indicate bouts. Win-loss records are shown in the right margin.
- Generally wrestlers face other wrestlers near their own rank.

Sumo tournaments are badly designed



- Kyokutenhō (purple) and Tochiōzan (blue) tied with 12-3 records. Kyokutenhō won the playoff bout and the tournament.
- Wrestlers in the middling ranks almost never face top ranked ones, even though they are competing for the ‘best’ record.
- Kyokutenhō only faced two of the top fifteen wrestlers, based on the pre-tournament ranking (Banzuke).

Sumo tournaments are badly designed



Tournament victories by low ranked wrestlers seem to happen about once or twice a decade.

- Takatōriki, ranked Maegashira 14 East, in March 2000
- Asanoyama, ranked Maegashira 8 West, in May 2019.
- Tokushōryū, ranked Maegashira 17 West, in January 2020.

Scoring by total wins

- In many competitions, the winner is determined simply by the number of wins accumulated by each side.
    - The underlying assumption being that the stronger side at doing the object of the competition will win.
- Our measure for competitor j simply could be:

$$\text{strength}_j = \sum \text{Wins}_j$$

→ When does this model of "scoring" make sense as a means of measuring which units have more of **some underlying quality or concept of interest**?

Who wins, who loses?

- Underlying quality or concept α_j, where j indexes individuals/ teams/ units.
- Let $W_{jk} = 1$ if j defeats their opponent k, and 0 if k defeats j.
- Then the total number of wins we would expect side j to receive in a series of competitions is:

$$E\left[\text{Wins}_j\right] = \sum_{k=1}^{n_j} p\left(W_{jk} | \alpha_j, \alpha_k\right)$$

→ Under which conditions will Wins_j be a good measure of α_j?

Winning at winning

For Wins$_j$ to be a (good) measure of α_j, better individuals/teams/units (those with higher α_j) must be more likely to win (obviously!).

Assumption 1

Individuals/teams/units that have more of the concept of interest will be more likely to succeed in the pairwise competitions:

$$\frac{\partial E\left[W_{jk}\right]}{\partial \alpha_j} > 0 \quad \text{and} \quad \frac{\partial E\left[W_{jk}\right]}{\partial \alpha_k} < 0$$

- This means that as α_j increases, a win of j over k becomes more likely, and as α_k increases, a win of k over j becomes more likely
- Crucially, we want $E[W_{jk}]$ to only depend on α_j and α_k, hence the next two assumptions

A well-structured competition

One way to guarantee that $E\left[\text{Wins}_j\right]$ is increasing in α_j is to impose a number of requirements on the structure of the competitions.

Assumption 2

Every individual/team/unit has the same number of matches n_j.

Assumption 3

Every individual/team/unit has opponents with the same distribution of strengths $f(\alpha_k)$.

- Both of these mean that all individuals/teams/units need to have similar opportunities to succeed $=$ strict fairness
- We need assumption 2 because the expected number of wins for j is going to increase in the number of competitions n_j, *regardless* of α_j
- We need assumption 3 because the expected number of wins for j is going to decrease with increasingly strong opponents α_k, *regardless* of α_j

What sumo gets wrong

- Honbasho sumo tournaments fail to meet assumption 3.
  - Not all wrestlers in the competition for the top division prize face similarly strong competition.
- Sports league competitions are often structured with balanced schedules so that the competition is *strictly* fair.
  - In the Premier League, for example, every team faces every other team twice, once at home and once away.
  - In the rugby union Six Nations, all six teams face each other.
- Strict fairness of the schedule is difficult to achieve without small divisions.
  - There are too many sumo wrestlers in the top division, given the length of the tournament!
  - Most world cups (football, rugby etc) need all-play-all and knock-out phases

Scoring draws

- Many competitions also can have draws/ties in individual match-ups
  - Obvious solution: count draws as intermediate between a win and a loss.
- In practice, it is common to define a point system so no one has to cope with fractions
  - Obvious solution: win is worth 2 points, a draw is worth 1 point, and a loss is worth 0 points.
- Can redefine W_{jk} as the number of points received instead of the number of wins and then sum of *points* is the measure of α_j
  - The same conditions for fair competition still need to hold!

Varieties of points systems

Many sports leagues **do not** count a draw as intermediate between a win and a loss.

- In domestic and international football competitions, it is standard to award 3 points for a win, 1 point for a draw, and 0 points for a loss.
- In the National Hockey League in the US and Canada, two points are awarded for a win, one point for losing in overtime or in a shootout, and zero points for a loss in regulation time.
- These systems (and all scoring systens more broadly!) are designed to incentivise certain strategies and disincentivise others.

Limitations of scoring wins/draws/losses

- Neither strict fairness nor fairness in distribution are achievable in all contexts.
- Consider the problem of determining who are the strongest tennis or chess players.
    - Too many individuals to have everyone play everyone else regularly.
    - If you selected opponents randomly from a very large pool, most of the competitions would be uninteresting as competitions.
- How can we measure which individuals/teams/units are strongest if we observe a very unbalanced set of competitions?

Rating transfer systems

- Chess ratings use a rating system called Elo, named after their inventor Árpád Imre Élő.
- Core idea:
  - Whenever you face an opponent in a match, some rating points are at stake.
  - Depending on the pre-match ratings and the result of the match, some number of points are transfered between the two opponents.
  - You gain more points for a better result, and against a better opponent.
- Rating can improve rapidly by defeating highly rated opponents, but the highly rated cannot gain much rating by defeating weak competition.

Advantages and Disadvantages of Elo

Advantages

- Reasonably simple, good mathematical properties for the most part
- Decentralised calculation is possible, so long as everyone is honest or you have a reliable register of results.
  - You could use a blockchain! “[Blockchains allow] mutually mistrusting entities to exchange financial value and interact without relying on a trusted third party.”

Disadvantages

- Sensitive to grade inflation over time as new entrants introduce more points to be redistributed.

Bottom line

- Not all wins are equally impressive.
- *If you want to measure something from data that involve an imbalanced schedule of competitions, you need to take into account not just the result, but also the strength of the opposition.*
- Without being able to come up with an obvious way to aggregate[3] the indicator data (wins/losses) into a measure, we will try to *model* the relationship

[3] I.e. a mathematical formula.

Quick logistic regression recap

Modelling Binary Outcomes

Logistic regression

$$log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \beta_0 + \beta_1 X + \beta_2 X_2 + \ldots$$

- Remember, in order to restrict the regression output to be between 0 and one, we apply the following transformation:

$$p(Y=1) = \frac{e^{(\beta_0+\beta_1 X+\beta_2 X_2+\ldots)}}{1+e^{(\beta_0+\beta_1 X+\beta_2 X_2+\ldots)}}$$

- With some rearranging, we 'free' the right hand side and the left hand side is equal to the **log odds of $Y = 1$**
- The *units* of the dependent variable are therefore in terms of **log odds**
- The β coefficients are **log odds ratio**

Logistic Regression in R

Let’s look at data from the wreck of the Titanic.

```r
# OLS
m0 <- lm(survive ~ adult + male + thirdclass,
         data=titanic)

# Logistic
m1 <- glm(survive ~ adult + male + thirdclass,
          data=titanic, family = binomial(link = "logit"))
```

Logistic Regression in R

Coefficients

	OLS	Logit
(Intercept)	0.958^{***}	2.493^{***}
	(0.044)	(0.276)
adult	-0.171^{***}	-1.050^{***}
	(0.041)	(0.245)
male	-0.531^{***}	-2.538^{***}
	(0.021)	(0.131)
thirdclass	-0.170^{***}	-1.112^{***}
	(0.019)	(0.131)
Num.Obs.	2201	2201

Logistic Regression in R

Predicted outcomes

```r
# In Logg odds
predict(m1,
        newdata = data.frame(adult=1,male=1,thirdclass=1))
```

```
##         1
## -2.207968
```

```r
# In predicted probabilities
predict(m1,
        newdata = data.frame(adult=1,male=1,thirdclass=1),
        type="response")
```

```
##          1
## 0.09903722
```

Bradley-Terry models

Towards a model

We have

1. An *unobserved* ('latent') quantity we want to measure
   - 'strength" or 'propensity to win'
2. *observed* and relevant indicator data
   - wins, losses and draws

→ We want to connect the two.

Why Model?

- If you have a perfectly balanced schedule of competitions, point systems work very well.
- And even if that is the case, there are advantages to fitting a measurement *model*.
  - Can quantify uncertainty (confidence intervals)
  - Can calculate predictions for (unobserved) comparisons

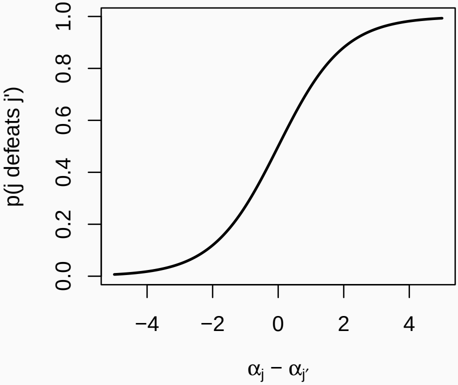
Bradley-Terry Model

- The model we will be using was first described by Bradley & Terry (1952)[4]
- We assume that each team/individual/unit j has a strength in competition that is described by a single parameter α_j
- We then assume that the log-odds of the competition results are determined by the difference between the parameters for the two sides:

$$log\left(\frac{p(j \text{ defeats } j')}{p(j' \text{ defeats } j)}\right) \;=\; \alpha_j - \alpha_{j'}$$

[4] *Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons.* Biometrika, 39(3/4), 324–345. https://doi.org/10.2307/2334029

Bradley-Terry model predictions



If $\alpha_j = 3$ and $\alpha_{j'} = 0.3$, then
$p(j \text{ defeats } j') = \frac{e^{3-0.3}}{1+e^{3-0.3}} = \frac{e^{2.7}}{1+e^{2.7}} = 0.94$

If $\alpha_j = -0.5$ and $\alpha_{j'} = 0.3$, then
$p(j \text{ defeats } j') =$

- The more *positive* the difference between α_j and $\alpha_{j'}$, the greater the probability that j wins.
- The more *negative* the difference between α_j and $\alpha_{j'}$, the greater the probability that j' wins.
- If $\alpha_j = \alpha_{j'}$, both j and j' are equally likely to win.

Connection to logistic regression

Remember the definition of logistic regression as the log of the odds of $Y = 1$ conditional on X:

$$log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ...$$

- Define $Y = 1$ to correspond to a victory of j over j' and $Y = 0$ to correspond to a victory of j' over j.
- Define a set of indicator variables X, one for each unit, which equal 1 when that unit is j and -1 when that unit is j'. Then:

$$log\left(\frac{p(j \text{ defeats } j')}{p(j' \text{ defeats } j)}\right) = \alpha_j - \alpha_{j'}$$

$\rightarrow$ The Bradley-Terry model is a special case of logistic regression.

Baseline

- We need to exclude the the indicator variable for one of the units
  - In the same way as we need to choose and exclude from the model the reference category for categorical variables, whose value is then included in the intercept
- This means that we are estimating the strength of all other units *relative* to the baseline.
- This means the absolute levels of the α parameters are arbitrary, *only the differences between them matter*. There is no meaningful “zero” for this scale.
  - You can add any constant number to all the α parameters without changing any of the model predictions, because all that matters are the *differences*![5]

[5] See two slides ago.

Intercept

- In addition to excluding one of the units, it can make sense to exclude the logistic regression intercept.
    - ‘Excluding’ here means that you force the intercept to be 0 (which also means you force the baseline to be zero)
- Reasons to include:
    - In sports competitions, there is often a home side advantage. If you include the intercept in the logistic regression, and always code the “home” side as the j rather than j', then the intercept will estimate the extent of home side advantage.
- Reasons to exclude:
    - If the choice of j and j' is arbitrary, there is no reason to have a systematic advantage/disadvantage for j versus j'.
    - If $\alpha_j = \alpha_{j'}$, then the interceptless regression equation logit$[p(j \text{ defeats } j')] = 0$ and $p(j \text{ defeats } j') =$

Extensions

If we want a model that can cope with draws, we can use ***ordinal* logistic regression**, a small extension to binary logistic regression.

Ordinal Logistic Regression

$$log\left(\frac{p(Y >= k)}{p(Y < k)}\right) \quad = \quad \alpha_k + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

where α_k are the log-odds that $Y >= k$, when all X's are 0.

- Instead of Y = 0,1, we have ordered Y = 0,1,...,k
- Functions like a binary logistic regression for every **different threshold** between levels of Y.
- Each threshold has its own intercept, but all have the same set of coefficients.

Ordinal Bradley-Terry model

- The same logic can be applied to the Bradley-Terry model:

$$\begin{aligned} log\left(\frac{p(j \text{ defeats or draws} j')}{p(j' \text{ defeats } j)}\right) &= \gamma_0 + \alpha_j - \alpha_{j'} \\ log\left(\frac{p(j \text{ defeats } j')}{p(j' \text{ defeats or draws} j)}\right) &= \gamma_1 + \alpha_j - \alpha_{j'} \end{aligned}$$

where, when $\alpha_j = \alpha_{j'}$:

- γ_0 are the log-odds of a win **or** a draw ***over*** a loss
- γ_1 are the log-odds of a win ***over*** a draw **or** a loss.

Covariates

- If we want to include covariates, we can do that by adding them to the binary/ordinal logistic regression
- Must code these in a way that makes sense given how we have selected to code the outcome, j and j'!
  - Covariates specific to each *match-up* can be included as ‘usual’ in a multivariatie (logistic) regression
  - Covariates specific to each *player* should be included within a multilevel model where the α’s themselves are a function of the player-covariates

Units?

- It is obvious how we talk about the units of wins or of points, but what are the units of the Bradley-Terry estimates?
    - The straightforward answer is the correct one: they are *log-odds ratios of better results versus worse results*
- When you fit *this* model, you are deciding to measure the strength of each side according to their log-odds of getting better, as opposed to worse, outcomes in competition with one another.
    - Log-odds are a reasonable unit of account for competition data for all the same reasons they are a reasonable basis for a limited dependent variable model with binary or ordered categorical outcomes.

Measurement error in Bradley-Terry estimates

Variance

= to what extent will our estimates vary around the value we aimed to measure?

- Think of target as the results, were we able to run an infinite number of competitions.
- Less data on individuals/teams/units will mean more imprecise estimates.

Bias

= to what extent, and in what ways, will the estimates tend to deviate from the thing we actually wanted to measure?

- Data generated by a different process than the one we intended can cause bias.
- If we ask people to code which of two political candidates is more charismatic, but they actually just code which one they would vote for, we will measure something about the latter rather than the former.

Bradley-Terry Model implementation

Bradley-Terry with the Sumo data

Structure of the data

```
##   wrestler1_win Y1e O1e O1w O2e O2w O3e O3w S1e S1w
## 1             0   0   0   0   0   0   0   1   0   0
## 2             1   0   0   0   0   0   0   0   1  -1
## 3             0   0   1   0   0  -1   0   0   0   0
## 4             1   1   0   0   0   0  -1   0   0   0
## 5             0   0   1   0  -1   0   0   0   0   0
```

```
summary(sumo_data$wrestler1_win)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.5641  1.0000  1.0000
```

```
summary(sumo_data$M7w)
```

```
##     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1.00000  0.00000  0.00000 -0.00641  0.00000  1.00000
```

Fitting the model with glm

```r
# + 0 to force the intercept to 0 (can also do '- 1')
# "." means to use all other variables in data
# remove one of the wrestlers as baseline with subset()
bt_model <- glm(wrestler1_win ~ . + 0,
                data = subset(sumo_data, select = -Y1e),
                family = binomial(link = "logit"))

# Y1e is our baseline and therefore 0
bt_estimates <- c("Y1e"=0, coef(bt_model))

# standard errors
bt_ses <- c("Y1e"=0, summary(bt_model)$coefficients[,2])
```

- Note that, if we didn't exclude one of the wrestler indicator variables ourselves, one of the coefficients would not be estimatable and therefore NA

Fitting the model with BTm( )

The **same** can be achieved with the `BTm( )` function from the `BradleyTerry2` package.

```r
# Data structure needs to be a bit different
sumo_data2[1:3,]
```

```
##   wrestler1_rank wrestler2_rank wrestler1_win wrestler2_win
## 1           M16e              J             1             0
## 2           M15e           M15w             0             1
## 3           M13w           M14e             0             1
```

```r
# install.packages("BradleyTerry2")
library(BradleyTerry2)
bt_model2 <- BTm(outcome = cbind(rikishi1_win, rikishi2_win),
                 player1 = rikishi1_rank,
                 player2 = rikishi2_rank,
                 data = natsu)
# running BTabilities(bt_model2) gives the estimates and SEs
# Note that BTm() chooses the first factor level as baseline
```

Looking at the α's

```
summary(bt_model) # Note I am not showing the full output here
```

```
...
##
## Call:
## glm(formula = wrestler1_win ~ . + 0, family = binomial(link = "logit"),
##     data = subset(sumo_data, select = -Y1e))
##
## Coefficients:
##      Estimate Std. Error z value Pr(>|z|)
## O1e  -0.61611    0.77493  -0.795 0.426582
## O1w  -0.29810    0.77416  -0.385 0.700192
## O2e   0.49457    0.80908   0.611 0.541020
## O2w  -0.04813    0.78683  -0.061 0.951223
## O3e  -0.39701    0.78131  -0.508 0.611358
## O3w  -0.45445    0.77484  -0.587 0.557535
## S1e  -0.63550    0.77633  -0.819 0.413012
## S1w  -0.52845    0.77840  -0.679 0.497211
## K1e  -1.86007    0.83185  -2.236 0.025348 *
...
```

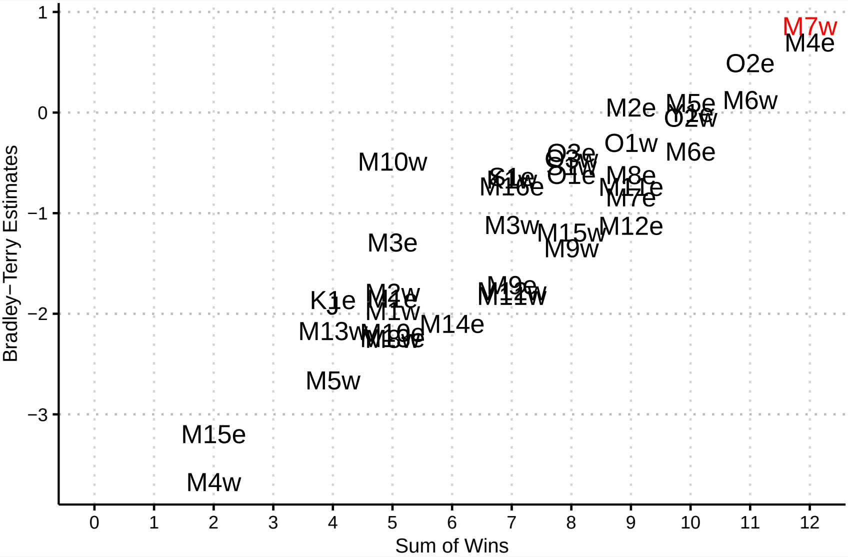
```
# look at the 5 highest ranked, according to the model
head(coef(bt_model)[order(coef(bt_model),decreasing = T)])
```

```
##        M7w        M4e        O2e        M6w        M5e        M2e
## 0.86221718 0.69976506 0.49456659 0.12790976 0.09921473 0.05383042
```

Code to combine the results

```r
# Let's combine it all in a data frame for convenience
results <- data.frame(
  "rank" = names(bt_estimates),
  "estimate" = bt_estimates,
  "se" = bt_ses,
  "conf.low" = bt_estimates - 1.96*bt_ses,
  "conf.high" = bt_estimates + 1.96*bt_ses,
  "total_wins" = wrestler_wins # sum of wins by wrestler
)
```

Total wins vs Bradley-Terry estimates



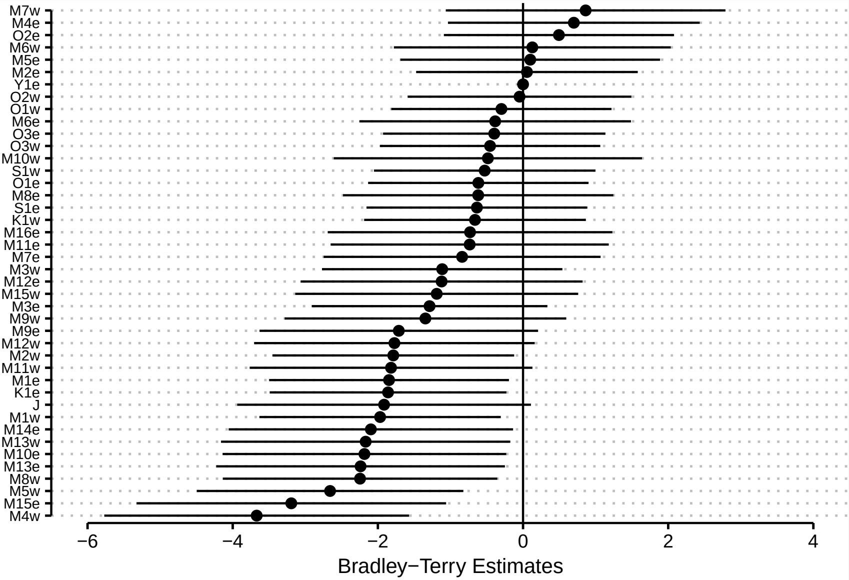
Code for figure

```r
library(ggplot2)
library(ggthemes)

results$winner <- results$rank=="M7w"

ggplot(results,aes(x=total_wins,y=estimate)) +
  geom_text(aes(label = rank,colour = winner,size=4)) +
  scale_x_continuous(breaks = 0:12, limits =c(0,12)) +
  scale_color_manual(values = c("black","red")) +
  labs(x="Sum of Wins",y="Bradley-Terry Estimates") +
  theme_clean() +
  theme(plot.background = element_rect(color=NA),
        panel.grid.major.x = element_line(linetype = "dotted",
                                          color = "lightgray"),
        legend.position = "none")
```

Bradley-Terry Model for Natsu Basho 2012

- The relative ordering of other wrestlers changes compared to simply counting the wins because of the imbalanced schedule
- Not all wrestlers had 15 opportunities to win
  - M10w (Chiyotairyu) was a wrestler who was doing well, got injured, and then dropped out of the tournament, winning 5 of 9 completed bouts.
- However, imbalanced competition format does not “explain away” Kyokutenhō’s victory
  - But we only used data from this one tournament!

Measurement Uncertainty



Code for figure

```r
# install.packages("lemon")

ggplot(results,aes(x=estimate,y=reorder(rank,estimate))) +
  geom_point(size=2) +
  geom_linerange(aes(xmin = conf.low, xmax = conf.high)) +
  geom_vline(xintercept = 0) +
  labs(y="",x="Bradley-Terry Estimates") +
  xlim(-6,4) +
  lemon::coord_capped_cart(bottom = "both",left = "both") +
  theme_clean() +
  theme(plot.background = element_rect(color=NA),
        legend.position = "none")
```

Measurement uncertainty

- Do we have enough data to be confident that Kyokutenhō was the best wrestler in Sumo in May 2012?
  - Not even close!
  - 15 bouts per wrestler is **far from enough** to be confident that the best wrestler will win any given tournament
- Indeed, most sporting competitions do not last nearly long enough to ensure the best side wins.
  - Sport would be less interesting to watch if they did!

Designing competition data collection

Is this kind of model useful for social science?

Sometimes the best way to measure an unobserved quantity is to setup comparisons that are responsive to that quantity.

- Imagine you want to figure out which political parties are further to the right and which are further to the left, across Europe.
- One thing you might do is ask some experts on European political parties to rank party positions on a 0 - 10 left-right scale.
- This is really difficult to do though, especially to do consistently across countries.

Pairwise comparison is often easier for eliciting knowledge

- What if we asked for pairwise comparison instead?
    - Possibly easier to answer whether the UK Labour party is to the left of the UK Conservative party
    - Perhaps also possible to assess whether the UK Labour party is to the left of the Irish Labour party or the German Social Democrats?
- Experts might be able to make meaningful binary comparisons without being able to generate valid 0-10 scores
- If your experts cannot make these kinds of binary comparisons, their 0-10 scale scores are definitely useless

Other examples

- Loewen, Rubenson and Spriling (2012)[6] studied which arguments are more persuasive in a Canadian referendum about a possible electoral reform.
  - You could ask people to directly rate the strength of the arguments
  - But it is probably easier to have them make pairwise comparisons between arguments.
- Blumenau and Lauderdale (2022)[7] study how much political arguments vary in effectiveness depending on different rhetorical elements.
  - Short answer: they vary in effectiveness a lot, but not in very systematic ways.

[6] *Testing the power of arguments in referendums: A Bradley-Terry approach.* Electoral Studies, 31(1), 212–221. https://doi.org/10.1016/j.electstud.2011.07.003

[7] *The Variable Persuasiveness of Political Rhetoric.* American Journal of Political Science, 68(1), 255–270. https://doi.org/10.1111/ajps.12703

Other examples

- Barnes, de Romémont and Lauderdale (2024)[8] study which taxes are more or less "popular"
  - Survey experiment where respondents had to choose between two revenue-equal changes to the UK tax system
  - Of note: the Truss government tried to cut the least popular taxes to cut!
- Zucco, Batista and Power (2019)[9] want to assess which ministerial roles in government are valued more highly in Brazil
  - Scoring 37 different ministerial roles would be difficult
  - Asked legislators and experts to "choose the ministry they thought a typical politician would prefer to obtain" for his/her party in a coalition negotiation, based on a randomly generated pair of portfolios.

[8] *Public Preferences Over Changes to the Composition of Government Tax Revenue.* British Journal of Political Science, 1–11. https://doi.org/10.1017/S0007123424000127

[9] *Measuring portfolio salience using the Bradley–Terry model: An illustration with data from Brazil.* Research and Politics, 6(1). https://doi.org/10.1177/2053168019832089

Yet more possible examples

- Competition datasets you could generate easily by surveying one another:
  - Which coffee shops around UCL are better?
  - Which quant methids modules were taught better?
  - Many further and more horrible possibilities!
- If there is common knowledge among a population that can be the basis of ranking a set of units, you can measure it.

Or you can study sports data and pretend it is social science

Title: “Contest competition and men’s facial hair: beards may not provide advantages in combat” *Evolution and Human Behaviour*

Abstract: … Hypotheses have been advanced that beards provide advantages in intra-sexual combat, as protective organs and honest signals of fighting ability. Here we provide the first test of these hypotheses using data from professional mixed martial arts fighters competing in the Ultimate Fighting Championship…. **We found no evidence that beardedness was associated with fewer losses by knock-out or greater fighter ability**…

How much data do you need?

This depends on...

1. ... How “big” the differences between units are:
   - If the “stronger” units almost always “win”, you do not need as much data as if the stronger units only win slightly more often
2. ... The number of comparisons involving each unit of interest.
   - Total number of comparisons needed will be proportional to the number of units n and not to the number of possible pairwise (ordered) comparisons $n^2 - n$.
   - The smaller the actual differences in the underlying strength between units, the more comparisons inolving each unit you’ll need.

How should the competitions be structured?

- Balanced competitions are good.
  - Ideal is to observe all pairwise comparisons equally frequently.
  - If too many are possible, select them at random.
  - If you can just barely generate enough comparisons for the number of units, some kind of adaptive testing to avoid re-running the “obvious” match-ups
- You cannot use this model to assess the relative strengths of two groups of units that could never face one another (eg two different sports leagues).
  - If you have only a few “bridging observations” between two groups, you may be very uncertain about their relative strengths.

How can I be sure I am measuring the right thing?

When setting up competitions to solve a measurement problem[10] **you need to make sure that people are doing comparisons that you intended them to do**.

1. Is the prompt clear?
   - Does the prompt ask respondents to make a choice based on the target concept, not something else?
   - What elements that you are *not* trying to measure could influence the choices made? I.e. could the prompt be misinterpreted?
2. Are you asking the right people?
   - Do the people you are asking (experts? general population?) know anything relevant about the target concept? Also, do you care whether their view of concept is 'correct' or not?
   - Are the people you are asking respresentative of the population you want to make claims about? (this is just basic population inference)

[10] For instance as a proposed improvement in your essay!!

Interpreting latent variables

Latent variable models

- The Bradley-Terry model is our first example of the broader class of "latent variable models".
    - We will see many more!
- What makes it a "latent variable model" is that we have *hypothesized* a variable – "quality" or "strength" or "propensity to win competitions" – that describes each individual/team/unit.
    - That variable is not observable directly; it is latent.
    - But we *assume* that it predicts the wins and losses (and draws) that we do observe.

Interpreting latent variable models

- Is there is a real thing in the world that we are calling the “strength” or “quality” of the individual/team/unit, and that thing is determining the outcomes?
  - No!
- This is a common conceptual error that people make when interpreting and talking about these kinds of models.
- Even though the Bradley-Terry model itself belongs to the *family of generative measures*, we should still understand what we are doing when we apply the model as a case of *pragmatic* measurement
  - We should not assume we are representing anything ‘real’ when we fit this model

Interpreting latent variable models

- The model acknowledges the “stronger” unit does not always win the competition.
- The model only attempts to measure one “factor”, but other things must also matter.
  - In future weeks: we will try to measure one (or two or three) “factors” that might predict an outcome, and then treat everything else as noise.
- Just because you hypothesize a monocausal explanation for something, that does not make it true.

→ The model presents a **useful summary** of factors that contribute to success in competitions

#nerdymemes

And those “latent variables,” are they in the room with us right now?

8:37 pm · 8 Oct 2021 · Twitter Web App

Summing up

- When we have relevant competition data we can make the case that the data is quite ‘close’ to the target concept
    - If the target concept is “which units tend to win competitions like the ones that we observe”, then observing data about who wins is really the best data we can hope for
    - However, we saw that there was not necessarily a straightforward way to *theoretically derive* a mathematical formula to aggregate the data
- Therefore, we discussed how we could **model** (i.e. create estimates of) the relationship between data and target concept
- Bradley-Terry models as a special case of logistic regression where the difference between two competitors’ coefficients is equal to the log-odds of one winning over the other
- If the data you have is not competition data, this approach is useless!