Who has won the Rugby Union World Cup, and why ?

A fundamentals-based empirical analysis of past outcomes in an international team-based sequential elimination tournament.

Vincenzo Verardi

joint with Brian O'Rourke

21st Stata users meeting, Cass Business School, London

Stata meeting

September 2015

Roadmap

Structure of the presentation

- Introduction
- RUWC 2015
- Probability tree
- Empirical model
- Specification
- Stata command
- Conclusion
- References

Rugby Union World Cup Tournament

History

- **First held in 1987** co-hosted by New Zealand and Australia
- The winners are awarded the **William Webb Ellis Cup**. "In 1823, William Webb Ellis **first picked up the ball in his arms and ran** with it. And for the **next 156 years forwards have been trying to work out why**." – Sir Tasker Watkins (1979)
- **Australia, New Zealand,** and **South Africa** have won the title **twice** while **England once**.
- Sixteen teams were invited to participate in the inaugural tournament in 1987. **Since 1999 twenty teams have taken part**.
- **England will host the 2015 World Cup**, while Japan will host the event in 2019.
- The current format allows for twelve of the twenty available positions to be filled by automatic qualification, as the **teams who finish third or better in the pools stages qualify for the subsequent** edition

Rugby Union World Cup Tournament

Past world cup Facts

- There have only been **7 Rugby World Cup so far**
- The **2003** Rugby World Cup had a global **cumulative audience** of **3.5 billion**, and was **broadcast in 205 countries** around the world.
- **No team has won Tri-Nations** tournament and a **Rugby World Cup** in the **same year**
- **Winners** of 5 or **6 Nations** tournaments **have reached the semi finals at least** of the Rugby World Cup happened in the same year.
- **Ireland is the only host nation** which has **not reached the semi finals** of World Cup

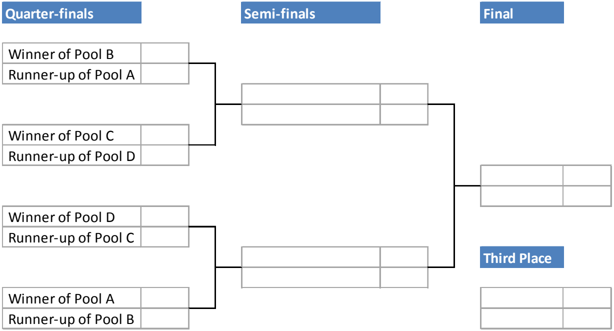
England 2015

- Global television is expected to reach over **4 billion people**.
- Potential **economic impact** to the UK of approx **£2.1 billion**
- **13 venues** in **11 cities**

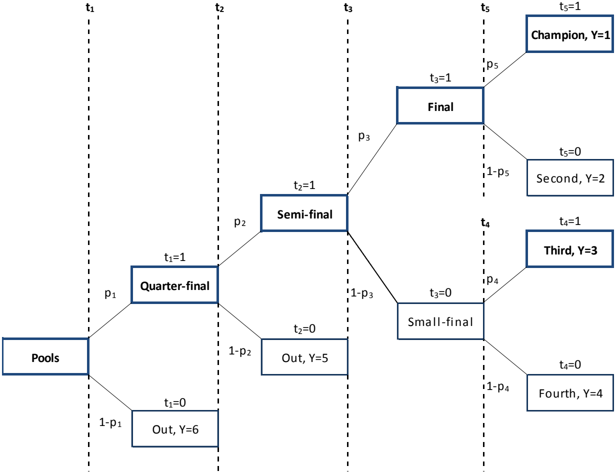
Rugby Union World Cup Tournament

England 2015

Pool A	Pool B	Pool C	Pool D
Australia	South Africa	New Zealand	France
England	Samoa	Argentina	Ireland
Wales	Scotland	Tonga	Italy
Fiji	Japan	Georgia	Canada
Uruguay	United States	Namibia	Romania



Probability tree for a country/year



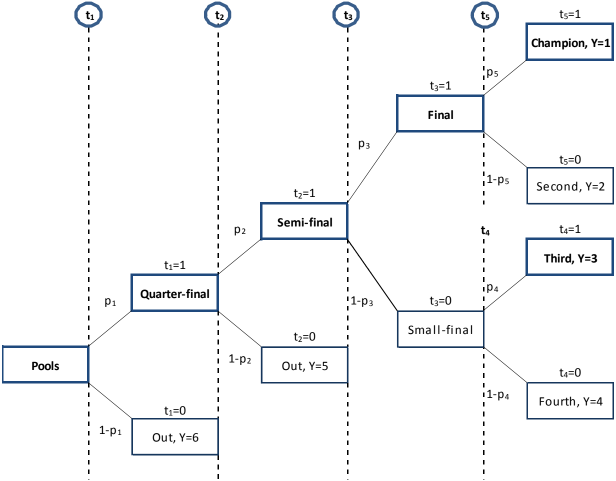
Probability tree

Definitions

- t_j **is the transition indicator from stage** $j-1$ **to stage** j**. It will be** 1 **in case of success and** 0 **otherwise.**
- p_j is the probability of transition from stage $j-1$ to stage j,i.e. $P(t_j = 1 | \textit{sequence to reach } j-1)$
- Y is the outcome variable, $Y \in \{1, 2, 3, 4, 5, 6\}$
- i is the country indicator
- t is the time indicator
- *Pools* indicates the pool stage, *QF* quarter-finals, *SF* semi-finals, *F* the final, *FT* the final for the third place, *T* the third place and *W* winning of the tournament

Probability tree

Probability tree for a country/year

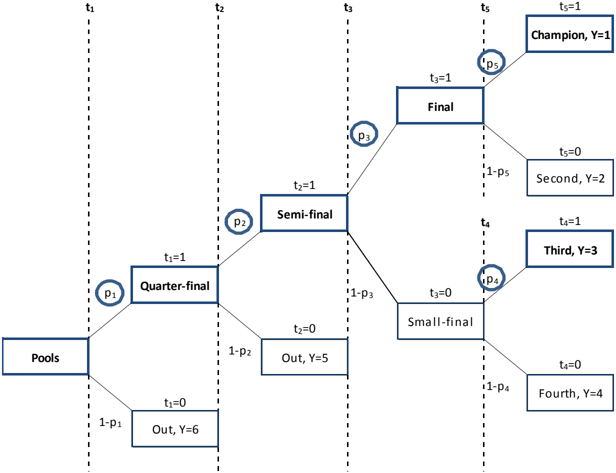


Probability tree

Definitions

- t_j is the transition indicator from stage $j-1$ to stage j. It will be 1 in case of success and 0 otherwise.
- p_j **is the probability of transition from stage** $j-1$ **to stage** j**,i.e.** $P(t_j = 1|$**sequence to reach** $j-1)$
- Y is the outcome variable, $Y \in \{1, 2, 3, 4, 5, 6\}$
- i is the country indicator
- t is the time indicator
- *Pools* indicates the pool stage, *QF* quarter-finals, *SF* semi-finals, *F* the final, *FT* the final for the third place, *T* the third place and *W* winning of the tournament

Probability tree for a country/year



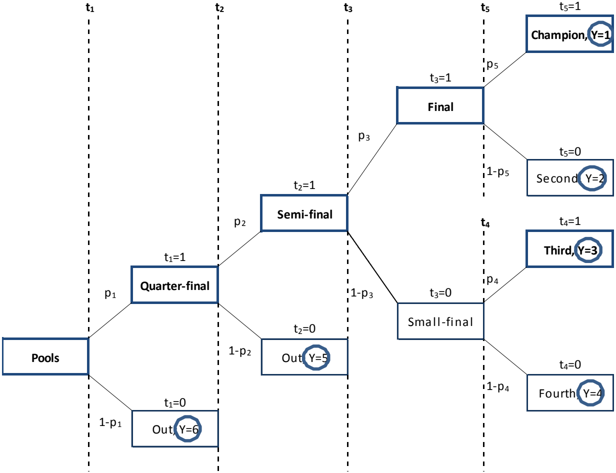
Probability tree

Definitions

- t_j is the transition indicator from stage $j-1$ to stage j. It will be 1 in case of success and 0 otherwise.
- p_j is the probability of transition from stage $j-1$ to stage j,i.e. $P(t_j = 1|\text{sequence to reach } j-1)$
- Y **is the outcome variable,** $Y \in \{1, 2, 3, 4, 5, 6\}$
- i is the country indicator
- t is the time indicator
- *Pools* indicates the pool stage, *QF* quarter-finals, *SF* semi-finals, *F* the final, *FT* the final for the third place, *T* the third place and *W* winning of the tournament

Probability tree

Probability tree for a country/year



Probability tree

Probability of outcomes

- $P(Y=6)=P(t_1=0)$
- $P(Y=5)=P(t_1=1, t_2=0)$
- $P(Y=4)=P(t_1=1, t_2=1, t_3=0, t_4=0)$
- *etc...*

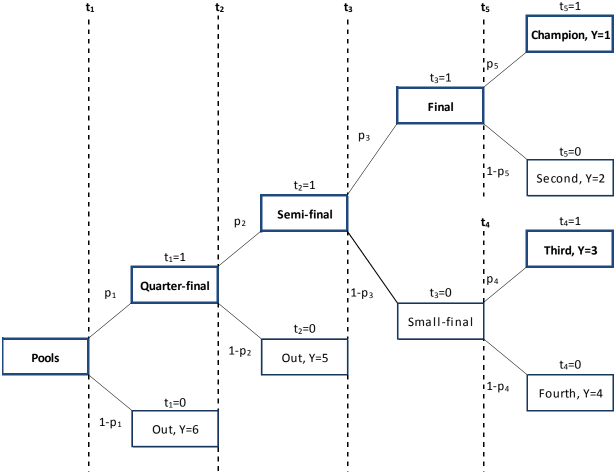
Probability of reaching given stages

- $P(QF)=P(t_1=1)$
- $P(SF)=P(t_1=1, t_2=1)$
- $P(F)=P(t_1=1, t_2=1, t_3=1)$
- *etc...*

Transition probabilities

- $P(SF|QF)=\frac{P(SF\cap QF)}{P(QF)}=\frac{P(SF)}{P(QF)}=\frac{P(t_1=1,t_2=1)}{P(t_1=1)}$
- $P(F|SF)=\frac{P(F\cap SF)}{P(SF)}=\frac{P(F)}{P(SF)}=\frac{P(t_1=1,t_2=1,t_3=1)}{P(t_1=1,t_2=1)}$
- *etc...*

Probability tree for a country/year



Empirical model

Probability of outcomes

We assume that success in each stage j for individual i at time t is associated to a latent variable $t^*_{j,i,t}$ such that

$$\begin{cases} t_{j,i,t} = 0 & \text{if} \quad t^*_{j,i,t} \leq 0 \\ t_{j,i,t} = 1 & \text{if} \quad t^*_{j,i,t} > 0 \end{cases}$$

We model the latent variable by $t^*_{j,i,t} = x'_{i,t}\beta_j + \varepsilon_{j,i,t}$; $i \in N_j$ where:

- $x'_{i,t}$ is the vector of **explanatory variables**. Here we assume the variables are the same for all stages
- β_j is the vector of **parameters** to be estimated at stage j
- $\varepsilon_{j,i,t}$ are the **unobservables** for stage j. They are assumed to be multivariate normally distributed with mean zeros and covariance Σ
- N_j is the set of **individuals still at risk** (i.e. still in the competition) at stage j.

Empirical model

Probability of outcomes (individual i, period t)

- $P(Y_{it} = 6) = P(t_{1it} = 0) = P\left(\varepsilon_{1it} \leq -x'_{it}\beta_1\right)$
- $P(Y_{it} = 5) = P(t_{1it} = 1, t_{2it} = 0) =$
  $P\left(\varepsilon_{1it} > -x'_{it}\beta_{it}, \varepsilon_{2it} \leq -x'_{it}\beta_2\right)$
- *etc* ...

Probability of reaching given stages (individual i, period t)

- $P(QF_{it}) = P(t_{1it} = 1) = P\left(\varepsilon_{1it} > -x'_{1it}\beta_1\right)$
- $P(SF_{it}) = P(t_{1it} = 1, t_{2it} = 1) =$
  $P\left(\varepsilon_{1it} > -x'_{1it}\beta_1, \varepsilon_{2it} > -x_{2it}\beta_2\right)$
- *etc*...

Transition probability (individual i, period t)

- $P(SF_{it}|QF_{it}) = \frac{P(t_{1it}=1, t_{2it}=1)}{P(t_{1it}=1)} = \frac{P(\varepsilon_{1it}>-x'_{1it}\beta_1, \varepsilon_{2it}>-x_{2it}\beta_2)}{P(\varepsilon_{1it}>-x'_{1it}\beta_1)}$
- *etc* ...

Empirical model

Independence of the unobservables

- If the ε**s are independent**, the **joint probability is the product of the individual probabilities** and the model becomes a simultaneous estimation of probits of successes at each stages (considering only those individuals still at risk)
- This is assumption is **standard in sequential logit/probit** models
- Under this assumption the **likelihood function is easy to write**

Problems

- This simple procedure is however **unrealistic** in our setup as it is difficult to consider the unobservable variables to be **uncorrelated between stages**.
- Ignoring these correlations would most probably create **biases since the selection rules** of each stage would be neglected.

Empirical model

Solution

- Should tackle the problem from a **different perspective**
- We should start by estimating the **probability of reaching the 6 possible observed modalities** which is the same as the probability of observing specific sequences of successes and failures in transitions
- What we should estimate is $P(Y_{it} = k) = x'_{i,t}\beta_k + \varepsilon_{k,i,t}$ where $k \in \{1, 2, 3, 4, 5, 6\}$, $i = \{1, ..., 25\}$, $t = \{1987, 1991, ..., 2011\}$ using a **multinomial probit (asmprobit in Stata)**
- We should **then calculate the marginal effects associated to an expression**. For example $\frac{\partial(1-P(Y_{it}=6))}{\partial x_{\ell,i,t}}$ will tell us how the probability of going to the quarter-finals is affected by a change in the ℓ variable
- Would be easy to do using `expression`(*pnl_exp*) in **Stata 13 if post-estimation command** was available.
- Easy trick to have it

Empirical model

Was work in progress **mprobit2.ado,** bacame useless with **Stata 14**

- Calls on **asmprobit.ado** with only *casevar* with the desired correlation structure to estimate the parameters
- Saves the coefficients matrix and covariance matrix
- Quetly runs a standard **mprobit.ado** with only one iteration
- Reposts matrices b and V using the ones estimated in **asmprobit.ado**
- The **marginal effects** associated to the **desired expression** are now available. For example $\frac{\partial(1-P(Y_{it}=6))}{\partial x_{\ell,i,t}}$ will tell us how the probability of going to the quarter-finals is affected by a change in the ℓ variable
- For the illustration we assume independent latent variable errors (128 cases to estimate 104 parameters with unstructured correlation structure)

Model specification

Team Variables

- Percentage of **points scored** in the prior 4 years by foot
- Percentage **wins** in the prior 4 years
- Mean **scrum weight**
- Mean **second row height**
- Mean **number of caps**
- Mean **experience**
- **Debut** year
- WRU **ranking**

Model specification

Socio-economic Variables

- **Southern hemisphere** dummy
- Number of **affiliated** players
- Total **population**
- Percentage **land in geographical tropics**
- **Mortality rate**, infant (per 1,000 live births)
- **Arable land** (% of land area)
- Population ages **65 and above** (% of total)
- **GDP growth** (annual %)
- **GDP per capita** (constant 2005 US$)
- Percentage of **catholics in total population**

Results team variables

Marginal effects *(d(y)/d(lnx) for population and affiliated)*

	Outcome-Multinomial Probit					Level-Average marginal effect			
VARIABLES	1	2	3	4	5	Quarter	Semi	Final	Champion
Affiliated players	4.06e-05**	4.79e-05**	3.62e-05*	3.72e-05*	4.04e-05**	0.194*	0.068	0.103**	-0.006
	(2.06e-05)	(2.16e-05)	(2.07e-05)	(2.05e-05)	(2.05e-05)	(0.109)	(0.045)	(0.045)	(0.029)
Total Population	-8.29e-07**	-2.81e-07**	7.15e-08	1.20e-07***	-7.52e-08*	-0.078	-0.367	-0.991***	-0.824**
	(3.53e-07)	(1.11e-07)	(5.97e-08)	(3.87e-08)	(4.14e-08)	(0.574)	(0.233)	(0.315)	(0.344)
Percentage of points scored by foot (4 prior years)	15.68	-4.122	0.258	-8.777	-1.362	-0.127	0.196	0.563*	0.652**
	(9.854)	(6.101)	(7.191)	(7.484)	(3.839)	(0.271)	(0.332)	(0.312)	(0.290)
Percentage wins (4 prior years)	15.20*	40.69***	21.78**	30.36***	12.45***	1.057	0.877**	0.304	-0.176
	(8.370)	(9.939)	(10.10)	(10.25)	(4.620)	(0.695)	(0.402)	(0.230)	(0.209)
Mean scrum weight	0.465*	0.147	-0.357	0.371	0.136	0.008	0.002	0.016*	0.015**
	(0.248)	(0.175)	(0.218)	(0.292)	(0.103)	(0.007)	(0.009)	(0.009)	(0.007)
Mean second row height	12.59	-48.79***	35.12***	92.16***	-9.059*	0.154	2.135**	-1.680	-0.117
	(29.02)	(17.05)	(13.50)	(29.78)	(5.248)	(0.331)	(0.834)	(1.062)	(0.936)
Mean number of caps	-0.161*	0.114*	-0.256***	-0.223**	-0.0482	-0.005**	-0.008**	0.004	-0.002
	(0.0927)	(0.0671)	(0.0839)	(0.102)	(0.0338)	(0.002)	(0.003)	(0.003)	(0.003)
Debut	-2.080**	-0.231	-0.257***	-0.310***	-0.157***	-0.013***	-0.041***	-0.062**	-0.067**
	(0.858)	(0.184)	(0.0749)	(0.0877)	(0.0468)	(0.003)	(0.015)	(0.026)	(0.028)
Ranking	0.908**	-1.013***	0.522*	0.240	-0.0706	-0.000	0.024	0.003	0.031**
	(0.386)	(0.297)	(0.306)	(0.316)	(0.0996)	(0.007)	(0.035)	(0.014)	(0.014)
Pseudo-R^2			67.92%				67.92%		
Observations			128				128		

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Coefficients associated to Affliliated and Population are semi-elasticities

Results country variables

Marginal effects

VARIABLES	Outcome-Multinomial Probit 1	2	3	4	5	Level-Average marginal effect Quarter	Semi	Final	Champion
Southern hemisphere	60.60***	56.56***	43.70***	56.01***	4.129	0.383***	0.555***	0.310*	0.147
	(16.23)	(11.39)	(9.634)	(15.48)	(2.911)	(0.052)	(0.054)	(0.178)	(0.105)
Percentage land in geographical tropics	-0.591	-22.70***	-22.74***	-29.44***	-10.51	-0.927*	-0.552*	0.460	0.616*
	(12.83)	(8.648)	(8.817)	(10.29)	(6.779)	(0.527)	(0.328)	(0.363)	(0.363)
Mortality rate, infant (per 1,000 live births)	-0.452*	-0.997**	-0.251	-1.303**	-0.334*	-0.029	-0.024**	-0.008	0.003
	(0.259)	(0.442)	(0.167)	(0.514)	(0.182)	(0.024)	(0.011)	(0.008)	(0.008)
Arable land (% of land area)	1.526***	1.705***	0.965***	1.701***	-0.223	0.005	0.094***	0.056***	0.035***
	(0.496)	(0.455)	(0.304)	(0.512)	(0.148)	(0.009)	(0.016)	(0.016)	(0.013)
Population ages 65 and above (% of total)	0.106	-0.446	1.576**	0.220	-0.233	-0.003	0.042	-0.023	-0.006
	(1.317)	(0.932)	(0.613)	(1.251)	(0.255)	(0.020)	(0.040)	(0.048)	(0.045)
GDP growth (annual %)	-0.841***	-0.252	-0.244	-0.641*	-0.162*	-0.015***	-0.023*	-0.018	-0.020*
	(0.324)	(0.229)	(0.167)	(0.343)	(0.0834)	(0.005)	(0.013)	(0.012)	(0.011)
GDP per capita (constant 2005 US$)	-3.60E-04	5.63E-04	-5.45e-05	2.08E-04	-1.40E-04	-0.000	0.000	0.000	-0.000
	(3.24e-04)	(2.16e-04)	(1.66e-04)	(1.63e-04)	(8.90e-05)	(0.000)	(0.000)	(0.000)	(0.000)
Percentage of catholics in total population	-0.371	-0.0637	-0.209**	-0.280***	-0.128*	-0.010*	-0.009*	-0.004	-0.007
	(0.273)	(0.0937)	(0.0873)	(0.107)	(0.0746)	(0.005)	(0.005)	(0.009)	(0.009)
Constant	3,787**	498.9	367.9***	306.4*	321.1***	-	-	-	-
	(1,617)	(346.2)	(140.2)	(166.4)	(99.43)				
Pseudo-R^2			67.92%				67.92%		
Observations			128				128		

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Coefficients associated to Affliliated and Population are semi-elasticities

Illustrative example

Estimated result

A	B	C	D
Australia	**South Africa**	**New Zealand**	**France**
England	Scotland	Argentina	Ireland
Wales	Samoa	Tonga	Canada
Fidji	Japan	Georgia	Romania
Uruguay	USA	Namibia	Italy

Quarter Finals	Semi finals	Final
South Africa / **England**	England / **New Zealand**	**New Zealand** / France
New Zealand / Ireland	**France** / Australia	
France / Argentina		
Australia / Scotland		

Third Place

England	Australia

Commands in Stata

Declaring variables

- global national "south tropicar data785 data83 data1036 data454 data455 first catho80 "
- global team "feet perc mscrumweight msecond mcaps debut ranking"
- global exp "$team $national"

Estimating multinomial Probit (ideally with **mprobit2**)

- xi: mprobit2 score $exp, baseoutcome(6) robust correlation(ind)

Calculating marginal effects

- margins, dydx(*) exp(1-(predict(outcome(6))))
- margins, dydx(*) exp(1-predict(outcome(5))-predict(outcome(6)))
- margins, dydx(*) exp(predict(outcome(1))+predict(outcome(2)))
- margins, dydx(*) expression((1-predict(outcome(5))-predict(outcome(6)))/(1-(predict(outcome(6)))))

Conclusion

Key elements for success in RWU

- Have, a long tradition in rugby and many affiliated players
- Be in a good form period
- Come from the southern hemisphere
- Have a large share of coutry area outside the tropics
- Have a low infant mortality rate
- Have a large share of arable land
- Have a high second row and heavy scrum

Selected references

- **Bernard A, and Busse M**, (2004), '*Who wins the Olympic games : Economics resources and medal totals*', Review of Economics and Statistics, 86, 413 - 417.
- **Lazear E P, Rosen S**, (1981), '*Rank Order Tournaments as Optimum Labor Contracts*', Journal of Political Economy, vol. 89, 841 - 864.
- **Ehrenberg R, Bognanno M**, (1990), '*Do tournaments have incentive effects* ?,' Journal of Political Economy, Vol. 98, No. 6, 1,307 - 1,324
- **Ferrall C, Smith A**, (1999), '*A sequential game model of sports championship series : theory and estimation*', The Review of Economics and Statistics, Vol. 81, No. 4, 704 - 719.
- **Koning R H**, (2009), '*Home Advantage in Professional Tennis*', Journal of Sport Sciences, Vol 29(1), 19 - 27.
- **Szymanski S**, (2003), '*The economic design of sporting contests*', Journal of Economic Literature, Vol XLI, December, 1137 - 1187.

Legal disclaimer

This presentation is intended for academic research purposes only. No representations are made as to the accuracy, reliability or exhaustiveness of any information, description, comment, or projection contained herein. No third party should represent, use or rely upon it for any other purpose, including but not limited to any form of speculative or gaming activity.