MATH 60604A
Statistical modelling
§ 4h - Logistic model for proportions

Léo Belzile

HEC Montréal
Department of Decision Sciences

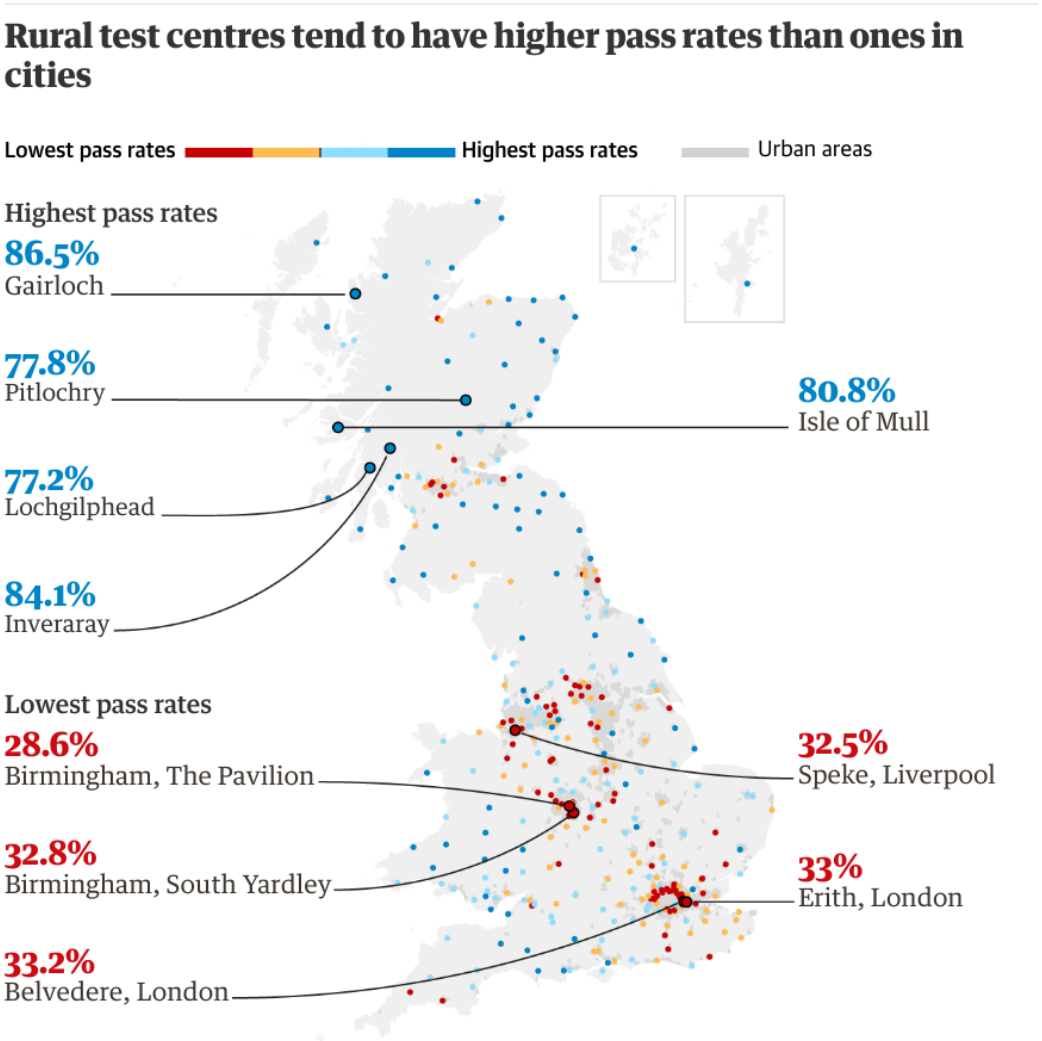
Logistic model for proportions

- Sometimes, we don't have access to individual records, but rather to aggregated counts such as the number of successes (out of m trials).
- We may use a binomial model instead by simply specifying the total number of trials associated to each number of successes.
- The parameter interpretation remains the same.

We consider the pass rate for all 346 Great-Britain driving license practical testing sites; the data are from 2018.

- 761 750 people succeeded in their exam out of 1 663 897 attempts.
- A news article from *The Guardian* hinted that exam takers in rural areas got an easy ride. Since we do not have a classification of urban/rural centers, we use the number of tests conducted as proxy.
- Other covariates are `sex` and the `region` for England; all of Scotland and Wales are pooled.

Binomial model for driving license pass rate in Great-Britain



Source: The Guardian.

SAS code to fit a logistic regression for binomial data

```
data gbdriving;
set statmod.gbdriving;
if(total < 500) then size="small";
else if (total < 1000) then size="medium";
else size = "large";
run;

proc logistic data=gbdriving;
class sex(ref="women") region(ref="London")
    size / param=glm;
model pass/total = sex region size /
    plrl plcl expb;
run;
```

Size of center per region

	size		
	large	mediu	small
	N	N	N
region			
East Midlands	40	3	3
East of England	54	.	.
London	48	4	6
North East England	29	5	8
North West England	63	3	2
Scotland	41	17	94
South East England	78	.	.
South West England	44	6	.
Wales	30	9	9
West Midlands	54	6	2
Yorkshire and the Hu	32	.	2

Scotland boasts the largest number of small centers (fewer than 500 exams per year).

Model specification for Great-Britain driving licenses

Model Information	
Data Set	WORK.GBDRIVING
Distribution	Binomial
Link Function	Logit
Response Variable (Events)	pass
Response Variable (Trials)	total

Number of Observations Read	692
Number of Observations Used	692
Number of Events	761750
Number of Trials	1663897

Model Fit Statistics

		Intercept and Covariates	
Criterion	**Intercept Only**	**Log Likelihood**	**Full Log Likelihood**
AIC	2294792.5	2278217.4	26619.303
SC	2294804.8	2278390.0	26791.848
-2 Log L	2294790.5	2278189.4	26591.303

Type 3 Analysis of Effects

Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
sex	1	8510.4974	<.0001
region	10	5565.9869	<.0001
size	2	1537.2919	<.0001

Odds estimates for Great-Britain driving licenses data

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
sex men vs women	1.0000	1.335	1.327	1.343
region East Midlands vs London	1.0000	1.279	1.262	1.297
region East of England vs London	1.0000	1.241	1.225	1.257
region North East England vs London	1.0000	1.500	1.475	1.524
region North West England vs London	1.0000	1.231	1.216	1.246
region Scotland vs London	1.0000	1.261	1.243	1.280
region South East England vs London	1.0000	1.257	1.243	1.271
region South West England vs London	1.0000	1.405	1.385	1.425
region Wales vs London	1.0000	1.447	1.423	1.472
region West Midlands vs London	1.0000	1.046	1.033	1.060
region Yorkshire and the Hu vs London	1.0000	1.094	1.078	1.110
size large vs small	1.0000	0.614	0.597	0.631
size mediu vs small	1.0000	0.766	0.741	0.792

Parameter interpretation for Great-Britain driving license

All other things being constant,

- The odds of men are 33% higher than women of obtaining a driver license;
- Greater London is the region with the lowest success rate after accounting for the site volume; the odds of success are 50% higher in North East England and 44.7% higher in Wales, etc.
- The odds of success are 63% higher in small center than in large centers $(1/0.614)$.
- All parameters are statistically significant.

Remark on models for Bernoulli/binomial data

- While the deviance and Pearson X^2 statistics are reported for logistic binomial model, their distribution depends on the unknown parameter vector β.
- As such, the deviance is approximately χ^2_{n-p-1} only when the number of trials m is in the several thousands.
- Comparisons of deviance, which amount to likelihood ratio tests, are however valid.

Revisiting the US road casualties example

We can fit a binomial model for the `crash` where the "event" is death.

Parameter Estimates and Profile-Likelihood Confidence Intervals				
Parameter		**Estimate**	**95% Confidence Limits**	
Intercept		-10.8702	-10.8913	-10.8495
time	**night**	0.2593	0.2372	0.2815
year	**2018**	0.2322	0.2101	0.2544

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
time night vs day	1.0000	1.296	1.268	1.325
year 2018 vs 2010	1.0000	1.261	1.234	1.290

- The estimated rate of death dying on the road during the day in 2010 is $\widehat{\pi} = \exp(\widehat{\beta}_0)/\{1 + \exp(\widehat{\beta}_0)\} = 0.000019016$, so a death rate of 1.9 per 100000 inhabitants. This estimate is slightly higher than the one from the negative binomial model.
- The odds of dying during nightime (relative to daytime) increase by 29.6%, whereas the odds for 2018 (relative to 2010) increase by 26.1%.