MATH 60604
Modélisation statistique
§ 4h - Modèle binomiale et proportions

Léo Belzile

HEC Montréal
Département de sciences de la décision

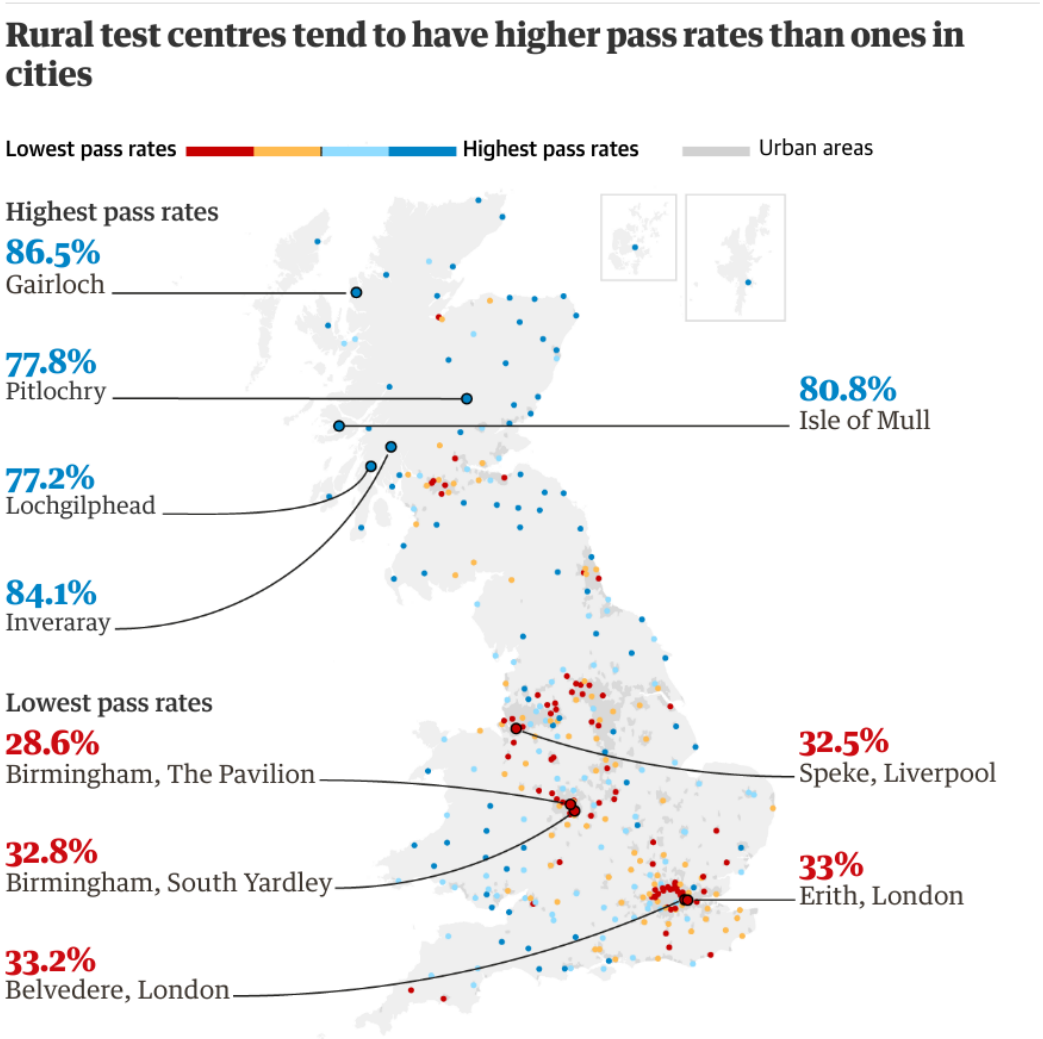
Régression logistique pour des proportions

- Souvent, seules les données aggrégées sont disponibles et on a accès au nombre de réussites (sur m essais).
- On peut utiliser la loi binomiale pour le modèle logistique en ajoutant le nombre d'essais au modèle.
- L'interprétation des paramètres reste identique.

On considère le taux de succès dans 346 centres pour examens de conduite pratique en Grande-Bretagne; les données sont pour 2018.

- 761 750 personnes on réussi l'examen pratique, pour 1 663 897 essais.
- Un article du journal *The Guardian* a suggéré que le taux de réussite était significativement plus élevé dans la campagne écossaise. Puisque nous n'avons pas de classification des centres de test ruraux/urbains, on utilise le nombre d'examens passés comme proxy.
- Les autres variables explicatives sont le `sexe` et la région en Angleterre; toutes les régions d'Écosse et du Pays de Galles sont regroupées.

Modèle binomial logistique pour examens de conduite britanniques



Source: The Guardian.

Code SAS pour régression logistique binomiale

```
data gbconduite;
set modstat.gbconduite;
if(total < 500) then volume="petit";
else if (total < 1000) then volume="moyen";
else volume = "grand";
run;

proc logistic data=gbconduite;
class sexe(ref="femme") region(ref="London")
    volume / param=glm;
model reussite/total = sexe region volume /
     plrl plcl expb;
run;
```

Taille des centres par région

	volume		
	grand	moyen	petit
	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

La plupart des petits centres (moins de 500 examens par année) sont situés en Écosse.

Spécification du modèle

Informations sur le modèle	
Table	WORK.GBCONDUITE
Distribution	Binomial
Fonction Link	Logit
Variable de réponse (Evénements)	reussite
Variable de réponse (Expériences)	total

Nombre d'observations lues	692
Nombre d'observations utilisées	692
Nombre d'événements	761750
Nombre d'expériences	1663897

Statistique d'ajustement du modèle

		Constante et Covariables	
Critère	**Constante uniquement**	**Log-vraisemblance**	**Log-vraisemblance complète**
AIC	2294792.5	2278217.4	26619.303
SC	2294804.8	2278390.0	26791.848
-2 Log L	2294790.5	2278189.4	26591.303

Statistique LR pour Analyse de Type 3

Source	**DDL**	**Khi-2**	**Pr > khi-2**
sexe	1	8529.02	<.0001
region	10	5589.96	<.0001
volume	2	1559.03	<.0001

Estimés des cotes pour les examens de conduite britanniques

Estimations du rapport de cotes et intervalle de confiance de vraisemblance de profil				
Effet	**Unité**	**Estimation**	**Intervalle de confiance à95%**	
sexe homme vs femme	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
volume grand vs petit	1.0000	0.614	0.597	0.631
volume moyen vs petit	1.0000	0.766	0.741	0.792

Interprétation des paramètres pour les examens de conduite britanniques

Toute autre chose étant égale par ailleurs,

- La cote des hommes pour la réussite de l'examen est 33% plus élevée que celle des femmes;
- Londres est la région avec le plus faible taux de succès, même en prenant en compte le volume des centres de test; la cote pour la réussite est 50% plus élevée en Angleterre du Nord-Est et 44.7% plus élevée au Pays de Galles, etc.
- La cote du taux de réussite est 63% plus grande dans les petits centres que dans les grands ($1/0.614$).
- Tous les paramètres sont statistiquement significatifs.

Remarque pour les modèles pour données binomiales/Bernoulli

- Si la déviance et la statistique du X^2 de Pearson sont parfois rapportées, leur loi nulle dépend des paramètres inconnus β.
- Ainsi, la déviance ne suit qu'approximativement une loi χ^2_{n-p-1} et ce même quand le nombre d'essais m est de l'ordre de plusieurs milliers.
- Les comparaisons de déviance entre modèle (ce qui revient à faire des tests de rapport de vraisemblance) sont néanmoins toujours valides.

Accidents de la route aux États-Unis, prise deux

On peut ajuster un modèle binomial aux données `accident` où l'événement d'intérêt est le décès.

Valeurs estimées du paramètre et intervalle de confiance de vraisemblance de profil

Paramètre		Estimation	Intervalle de confiance à95%	
Intercept		-10.8702	-10.8913	-10.8495
moment	nuit	0.2593	0.2372	0.2815
annee	2018	0.2322	0.2101	0.2544

Estimations du rapport de cotes et intervalle de confiance de vraisemblance de profil

Effet	Unité	Estimation	Intervalle de confiance à95%	
moment nuit vs jour	1.0000	1.296	1.268	1.325
annee 2018 vs 2010	1.0000	1.261	1.234	1.290

- Le taux estimé de décès sur la route le jour en 2010 est $\widehat{\pi} = \exp(\widehat{\beta}_0)/\{1 + \exp(\widehat{\beta}_0)\} = 0{,}0000019016$, soit 1,9 morts par 100000 habitants. Cet estimé du taux est légèrement plus élevé que celui du modèle de régression binomiale négative.
- La cote pour la probabilité de mourir augmente de 29,6% la nuit par rapport au jour, tandis que la cote pour 2018 (relativement à 2010) a augmenté de 26,1%.