MATH 60604
Modélisation statistique
§ 4c - Exemple de régression logistique

Léo Belzile

HEC Montréal
Département de sciences de la décision

Achat suite au visionnement d'une publicité

Dans le cadre d'une étude, des sujets ont navigué sur un site web qui contenait, entre autres, une publicité pour des bonbons. Pendant la navigation, un dispositif de suivi occulaire enregistrait l'endroit où se posait le regard du sujet. On a ainsi pu mesurer le temps de fixation du sujet. Un logiciel d'analyse des expressions faciales (FaceReader) a aussi servi à mesurer l'émotion du sujet pendant le visionnement de la publicité.

Supposons qu'au lieu de mesurer l'intention d'achat par questionnaire au moment où nos participants était au laboratoire, nous les avions plutôt contacté un mois plus tard pour voir s'ils avaient acheté le produit depuis leur visite.

Nombre d'achats

Le jeu de données `intention` contient deux autres variables:

- `achat`: variable binaire, `1` si le sujet à acheté des bonbons, `0` sinon.
- `nachat`: nombres d'achats de l'item

nachat	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	62	51.67	62	51.67
1	13	10.83	75	62.50
2	16	13.33	91	75.83
3	7	5.83	98	81.67
4	8	6.67	106	88.33
5	2	1.67	108	90.00
6	2	1.67	110	91.67
7	2	1.67	112	93.33
8	4	3.33	116	96.67
9	1	0.83	117	97.50
10	3	2.50	120	100.00

La variable `nachat` varie de 0 à 10 avec une moyenne de 1,71 achat par participant(e). 51,7% des sujets n'ont fait aucun achat.

Nombre d'achats suite au visionnement d'une publicité

On se concentre sur `acaht` comme variable réponse. Les variables explicatives sont les même que précédemment, à savoir

- `fixation`: durée totale de fixation de la publicité (en secondes).
- `emotion`: une mesure de la valence durant la fixation, soit le ratio de la probabilité d'une émotion positive sur la probabilité d'une émotion négative
- `sexe`: sexe du sujet, soit homme (`0`) ou femme (`1`).
- `age`: âge (en années).
- `revenu`: variable catégorique indiquant le revenu annuel du sujet; un parmi
  1. $[0, 20\ 000]$;
  2. $[20\ 000, 60\ 000]$;
  3. 60 000 et plus.
- `educ`: variable catégorique indiquant le niveau d'éducation, soit le plus haut grade obtenu
  1. secondaire ou moindre;
  2. collégial;
  3. universitaire.

Modèle logistique pour les achats

On dénote $\pi = \mathrm{P}\left(Y = 1 \mid \mathbf{X}\right)$ la probabilité d'acheter le produit dans le mois suivant l'étude conditionnelles aux variables explicatives, selon le modèle

$$\begin{aligned}\mathrm{logit}(\pi) = \beta_0 &+ \beta_1 \texttt{sexe} + \beta_2 \texttt{age} + \beta_3 \texttt{revenu}_1 + \beta_4 \texttt{revenu}_2 \\ &+ \beta_5 \texttt{educ}_1 + \beta_6 \texttt{educ}_2 + \beta_7 \texttt{statut} \\ &+ \beta_8 \texttt{fixation} + \beta_9 \texttt{emotion}.\end{aligned}$$

Code SAS pour la régression logistique

- Les paramètres β ne sont vraiment interprétables qu'à l'échelle exponentielle.
- La procédure `logistic` avec les options `plcl plrl expb` permet d'obtenir les estimés $\exp(\widehat{\beta})$ et des intervalles de confiance basés sur la vraisemblance pour ces paramètres.
- Avec `proc logistic`, la paramétrisation habituelle pour les variables catégorielles est obtenue avec l'option `param=glm`.

Code SAS pour `proc logistic`

```
proc logistic data=modstat.intention;
class educ revenu / param=glm;
model achat(ref="0")=sexe age revenu educ statut
    fixation emotion / plcl plrl expb;
run;
```

Sortie SAS de la procédure `logistic`

Statistique d'ajustement du modèle

Critère	Constante uniquement	Constante et Covariables
AIC	168.222	134.514
SC	171.009	162.389
-2 Log L	166.222	114.514

Test de l'hypothèse nulle globale : BETA=0

Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	51.7077	9	<.0001
Score	43.3703	9	<.0001
Wald	29.4146	9	0.0006

Analyse des effets Type 3

Effet	DDL	Khi-2 de Wald	Pr > khi-2
sexe	1	0.9841	0.3212
age	1	1.2875	0.2565
revenu	2	7.8853	0.0194
educ	2	0.0374	0.9815
statut	1	3.9751	0.0462
fixation	1	15.9150	<.0001
emotion	1	8.4149	0.0037

- Des diagnostics sur la qualité de l'ajustement sont rapportés pour le modèle sans covariables qui inclut uniquement l'ordonnée à l'origine (probabilité constante de succcès) et le modèle ajusté.
- En plus des critères d'information, la sortie contient les statistiques de tests (Wald, score, rapport de vraisemblance) pour l'hypothèse nulle, $\mathscr{H}_0 : \beta_1 = \cdots = \beta_p = 0$.
- Le tableau de significativité des paramètres (effets type III) est basé sur des statistiques de Wald (pour des tests de rapport de vraisemblance, utiliser la procédure `genmod` avec l'option `type3`).

Estimés des paramètres

Analyse des valeurs estimées du maximum de vraisemblance

Paramètre		DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2	Exp(Est)
Intercept		1	-1.3325	1.8603	0.5131	0.4738	0.264
sexe		1	0.4894	0.4934	0.9841	0.3212	1.631
age		1	-0.0624	0.0550	1.2875	0.2565	0.940
revenu	**1**	1	1.2923	0.6788	3.6245	0.0569	3.641
revenu	**2**	1	-0.4326	0.6198	0.4872	0.4852	0.649
revenu	**3**	0	0	.	.	.	.
educ	**1**	1	-0.0989	0.7198	0.0189	0.8907	0.906
educ	**2**	1	-0.1126	0.5907	0.0363	0.8488	0.894
educ	**3**	0	0	.	.	.	.
statut		1	-1.0199	0.5115	3.9751	0.0462	0.361
fixation		1	1.1694	0.2931	15.9150	<.0001	3.220
emotion		1	1.4460	0.4985	8.4149	0.0037	4.246

- La dernière colonne du tableau contient $\exp(\widehat{\beta_j})$ (option `expb`).
- Les tests de significativité pour $\beta_i = 0$ sont basés sur des statistiques de Wald.

Tableau des estimés des coefficients

Valeurs estimées du paramètre et intervalle de confiance de vraisemblance de profil				
Paramètre		**Estimation**	**Intervalle de confiance à95%**	
Intercept		-1.3325	-5.0270	2.3387
sexe		0.4894	-0.4771	1.4723
age		-0.0624	-0.1745	0.0432
revenu	**1**	1.2923	-0.0176	2.6649
revenu	**2**	-0.4326	-1.6846	0.7685
educ	**1**	-0.0989	-1.5526	1.2990
educ	**2**	-0.1126	-1.2910	1.0458
statut		-1.0199	-2.0572	-0.0342
fixation		1.1694	0.6506	1.8074
emotion		1.4460	0.5186	2.4897

Tableau des estimés des coefficients à l'échelle exponentielle

Estimations du rapport de cotes et intervalle de confiance de vraisemblance de profil				
Effet	**Unité**	**Estimation**	**Intervalle de confiance à95%**	
sexe	1.0000	1.631	0.621	4.359
age	1.0000	0.940	0.840	1.044
revenu 1 vs 3	1.0000	3.641	0.983	14.367
revenu 2 vs 3	1.0000	0.649	0.186	2.157
educ 1 vs 3	1.0000	0.906	0.212	3.665
educ 2 vs 3	1.0000	0.894	0.275	2.846
statut	1.0000	0.361	0.128	0.966
fixation	1.0000	3.220	1.917	6.095
emotion	1.0000	4.246	1.680	12.058

- À l'échelle exponentielle, les paramètres ne sont pas significatifs à niveau 5% si 1 est inclut dans l'intervalle de confiance.
- Pour obtenir les intervalles de confiance basés sur les rapports de vraisemblance, spécifier l'option `plrl`.

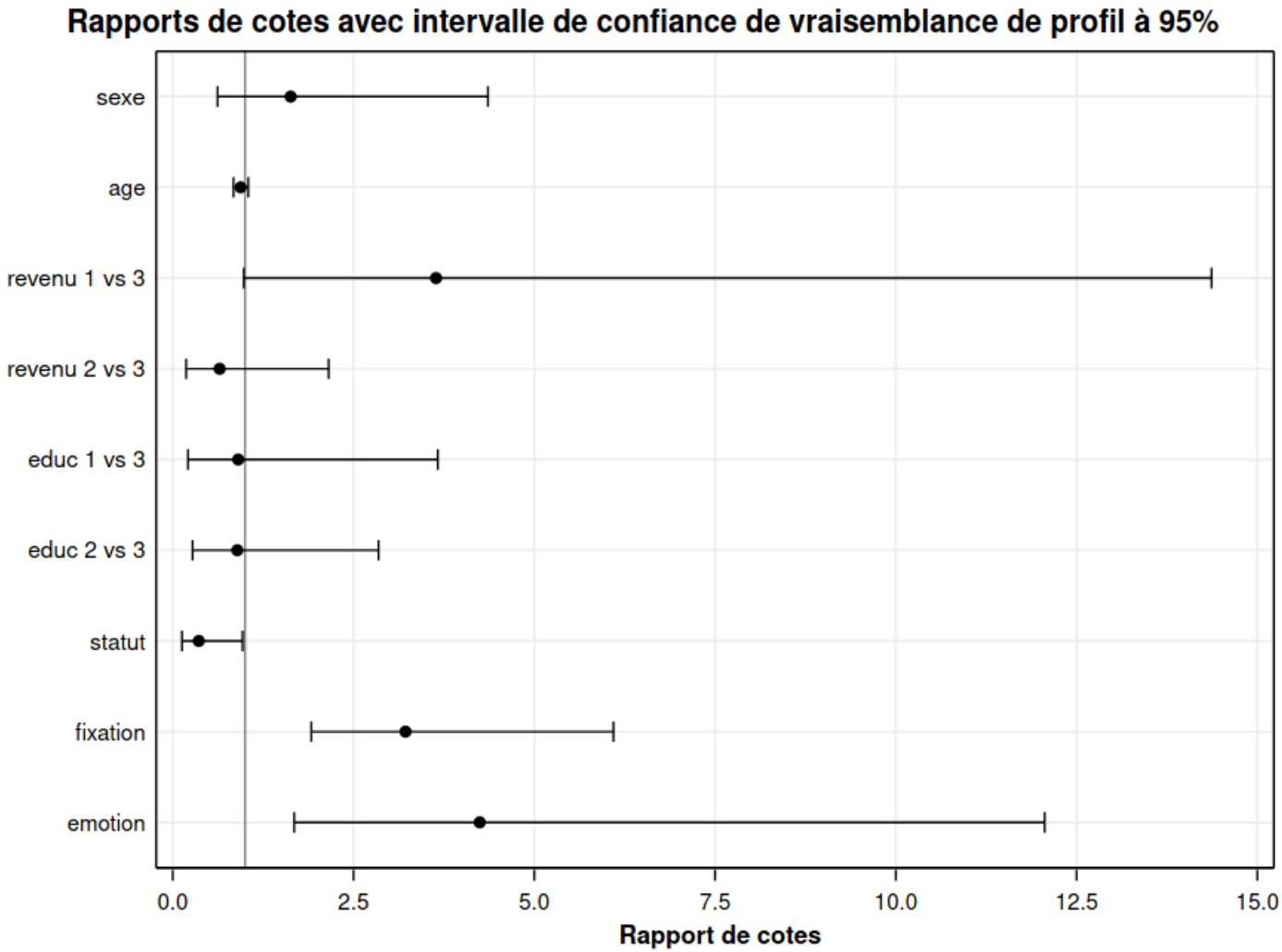
Interprétation des paramètres

- $\exp(\widehat{\beta}_{\texttt{sex}}) = 1.631$: la cote d'achat pour les femmes (`sexe=1`) est 1.631 fois plus élevée que pour les hommes, toute autre chose étant égale par ailleurs. La probabilité que les femmes achètent au moins un item est plus élevée que celle des hommes après ajustement.
- $\exp(\widehat{\beta}_{\texttt{age}}) = 0.94$: pour deux personnes qui ne différent que d'un an, la cote pour l'achat de la personne plus âgée est 0.94 celle de la personne plus jeune, une diminution de 6%, *ceteris paribus*.
- Toutes choses étant égales par ailleurs, la cote pour l'achat augmente de $\exp(\widehat{\beta}_{\texttt{fixation}}) = 3.22$ pour toute augmentation du temps de fixation de une seconde.

Comparaison des niveaux de `revenu`

- Le coefficient pour $\texttt{revenu}_1$ est relatif au niveau `revenu=3` et $\exp(\widehat{\beta}_{\texttt{revenu}_1}) = 3{,}641$: la cote d'achat pour les sujets à faible revenu (`revenue=1`) est $3{,}641$ fois plus élevée que celle des personnes à revenu élevé (`revenu=3`), **toutes choses étant égales par ailleurs**.
- Pour obtenir le rapport de cotes entre les niveaux 1 et 2 de `revenu`, on pourrait réajuster le modèle en changeant la catégorie de référence.
- On peut également calculer le rapport de rapports de cotes de $3{,}641/0{,}649 = 5{,}61$; on a donc que la cote de succès pour le niveau de revenu 1 est 461% plus élevée que celle pour la classe de revenu moyen 2, toute autre chose étant égale par ailleurs.

Représentation visuelle des rapports de cote



Les intervalles de confiance calculés à partir de la vraisemblance profilée sont **invariants aux reparamétrisations**: on obtient un intervalle de confiance pour $\exp(\beta_k)$ en appliquant la fonction exponentielle à chaque borne de l'intervalle pour β_k.

Prédiction

- La régression logistique est fréquemment utilisée en apprentissage machine pour la classification.
  - si on veut prédire la valeur de Y (zéro ou un) pour de nouvelles observations pour lesquelles on a observé les covariables.
- On peut aisément **prédire** à partir du modèle de régression logistique ajusté, sachant que

$$\pi_i = \mathrm{P}\left(Y_i = 1 \mid \mathbf{X}_i\right) = \frac{\exp(\beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip})}{1 + \exp(\beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip})}.$$

- Si on substitue $\widehat{\beta}$ en lieu et place des paramètres inconnus, on obtient un estimé de la probabilité de succès, $\widehat{\pi}_i = \widehat{\mathrm{P}}(Y_i = 1 \mid \mathbf{X}_i)$.
- On peut utiliser cet estimé $\widehat{\pi}_i$ pour prédire la valeur de Y_i avec un point de coupure c,
  - si $\widehat{\pi}_i > c$, on prédit $\widehat{Y}_i = 1$.
  - si $\widehat{\pi}_i \leq c$, on prédit $\widehat{Y}_i = 0$.

Séparation (quasi)-complète de variables

- Dans certaines applications, il existe des combinations de variables explicatives qui permettent de prédire exactement certaines ou toutes les valeurs de $Y \in \{0, 1\}$.
  - Par exemple, un sondage à la sortie des urnes dans lequel chaque vote correspond à l'affiliation politique rapportée.
- Bien que les prédictions soient toujours valides, la séparation (quasi)-complète est problématique pour l'interprétation (une bonne analogie est l'impact de la collinéarité).
- L'estimaté du maximum de vraisemblance est infini pour certains paramètres (cas limite) ou encore n'est pas unique — cela empêche le calcul d'erreurs-type, etc.

Détection et remèdes pour la séparation (quasi)-complète

- Il est facile de diagnostiquer les problèmes dans les logiciels.
- Par exemple, R imprime le message

```
Warning message: glm.fit: des probabilités ont été ajustées
numériquement à 0 ou 1
```

dans la console, tandis que dans SAS, on obtient

```
Complete separation of data points detected.
WARNING: The maximum likelihood estimate does not exist.
```

- On peut restaurer l'identifiabilité des paramètres en pénalisation la fonction de log-vraisemblance. La correction de Firth est une solution populaire qui permet d'obtenir des coefficients finis et uniques.
  - option `firth` dans la procédure SAS `logistic`
  - la fonction `logistf` du paquetage éponyme dans R