LNG-1100 : Méthodes expérimentales et analyse de données

Les variables binaires et le nettoyage des données

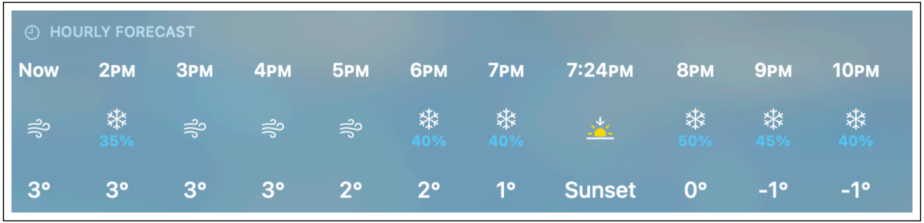
Guilherme D. Garcia

fr.gdgarcia.ca

11

Université Laval

La météo



- Comment interpréter 40 % (p. ex., à 18h)?
- D’où vient cette prévision?

Un dé

1. Quelle est la probabilité de lancer un 4? Et de lancer un nombre pair?
2. Quelles sont les cotes pour lancer un 4? Et de lancer un nombre pair?

Un dé

1. Quelle est la probabilité de lancer un 4? Et de lancer un nombre pair?

   $P(4) = \frac{1}{6} \approx \mathbf{0.17}$ $\qquad P(pair) = \frac{3}{6} = \frac{1}{2} = \mathbf{0.5}$

2. Quelles sont les cotes pour lancer un 4?

   $$\frac{P(4)}{P(\neg 4)} \rightarrow \frac{\frac{1}{6}}{\frac{5}{6}} = \frac{1}{6} \times \frac{6}{5} = \mathbf{\frac{1}{5}}$$

   Pour chaque fois qu'on lance un 4, il y a 5 fois où on ne lance pas

   Et les cotes pour lancer un nombre pair?

   $$\frac{\frac{3}{6}}{\frac{3}{6}} = \frac{3}{6} \times \frac{6}{3} = \mathbf{1}$$

   Les cotes pour lancer un nombre pair et de ne le lancer pas sont égales : 1

Terminologie de la séance

Deux concepts importants

- En statistique, la **probabilité** est une mesure allant de 0 à 1, représentant la fréquence attendue d'un événement.
- Les **cotes** (ou le rapport des cotes) représentent le rapport entre la probabilité que l'événement ait lieu et la probabilité qu'il ne se produise pas : $\frac{P(x)}{P(\neg x)}$

☞ La probabilité est **partout** : la météo, les primes d'assurance, la sélection de candidats, la santé publique (épidémiologie), les jeux de hasard, les recommandations de films/produits, les systèmes de transport, la gestion des risques en entreprise, etc.

Plan de la séance

1. Révision : régression linéaire
2. Les réponses binaires : log-odds et probabilité

Révision

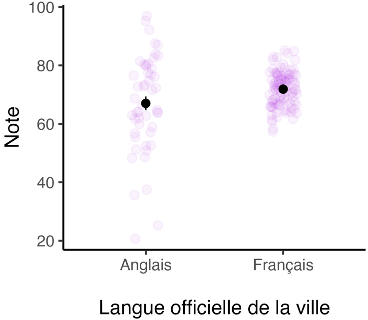
Régression linéaire

Villes

1. Importez `villes2.csv` et ajoutez une nouvelle colonne : `francophone` (`0/1`)
2. Créez une figure qui examine l'effet de la francophonie sur la note des apprenants
3. Créez un modèle aligné à la figure
4. Rapportez les résultats

Révision

Régression linéaire



- Effet positif : $\hat{\beta} = 4.9$, $p = 0.01$
- IC 95 % $= [1.186, 8.552]$

☞ Donc, la langue officielle de la ville affecte positivement la note moyenne des apprenants

Problème : les variances sont très différentes[1]

[1] On ignore ce problème dans notre cours.

Révision

Régression linéaire

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    67.009      1.522  44.034  < 2e-16 ***
francophone1    4.869      1.864   2.613  0.00991 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1
```

- $\hat{\beta}_0 \rightarrow$ la note moyenne quand `francophone == 0` (anglais)
- $\hat{\beta}_1 \rightarrow$ le slope (la différence entre français et anglais)

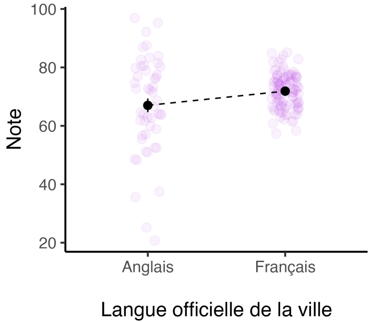
☞ La régression estime **les moyennes**

Révision

Régression linéaire : `note ~ francophone`

- On n'utilise pas des droites dans un graphique dont la variable `x` est catégorielle

☞ Mais c'est utile de l'ajouter au graphique ici :



- La droite ici représente notre $\hat{\beta}_1$

Les hypothèses nulles :

1. Du modèle : pas d'effet de `francophone`
2. De l'intercept : $\beta_0 = 0$
3. Du slope : $\beta_1 = 0$ (pas de différence entre β_0 et β_1)

☞ On rejette ces hypothèses ici :

notre modèle confirme un effet significatif

Un autre type de modèle

- On vient d'analyser `note ~ francophone`, où la variable de réponse est continue[2]

☞ Et si on voulait analyser la relation contraire?

$$\boxed{\texttt{francophone ~ note}}$$

- Si la réponse est binaire (`0/1`), il faudra penser aux **probabilités** :

P. ex. *Prenons la note 79 : quelle est la **probabilité** que l'apprenant soit dans une ville francophone?*

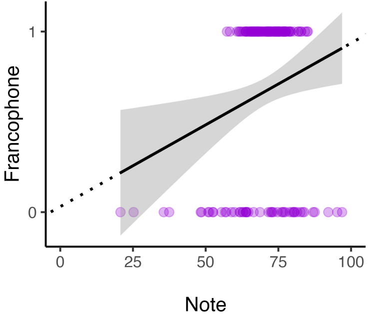
$$P(francophone = 1 | note = 79) \tag{1}$$

[2] Et idéalement normale!

Un autre type de modèle

- **Pourquoi ne pas utiliser la même méthode?**

```
lm(francophone ~ note, data = d)
```



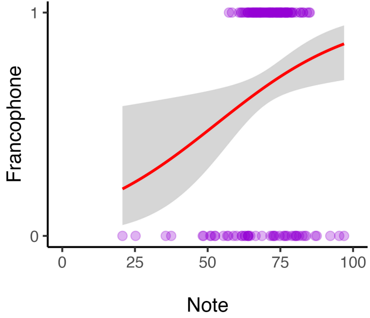
- Une régression linéaire ne nous donne pas la probabilité d'un résultat
- Probabilité : **toujours** limitée à l'intervalle [0,1]

☞ Donc, il faut « adapter » notre modèle

Un autre type de modèle

- **Une régression logistique**

```r
glm(francophone ~ note, data = d, family = "binomial")
```



- Probabilité : **toujours** limitée à l'intervalle [0,1]
- Notez que la ligne est **courbe**

☞ Notre interprétation des résultats sera différente

Visualiser un exemple interactif ici

Un autre type de modèle

La logique

1. On veut utiliser la même architecture de modèle : une droite (changement constant)
2. Mais cela n'est pas compatible avec la notion de probabilité
3. On pourrait utiliser les cotes (*odds*), qui sont linéaires; mais elles sont asymétriques
4. **Solution** : on utilise les **logs de cotes** (*log-odds*) = c'est le **logit de la probabilité**

Régression linéaire : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{e}$

Régression logistique : $logit(P) = \hat{\beta}_0 + \hat{\beta}_1 X$

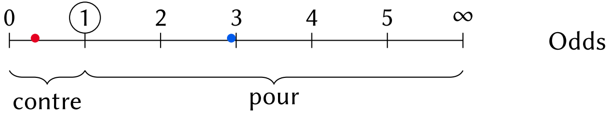
☞ Le modèle nous donnera des coefficients en **log-odds** (moins intuitif)

On utilisera la fonction `glm( ... , family = "binomial")` en R

Odds et log-odds

Supposez la note 79. On verra plus tard que :

- Les cotes pour être dans une ville francophone = $\boxed{2.9}$
- Les cotes pour **n**'être **pas** dans une ville francophone = $\boxed{0.3}$

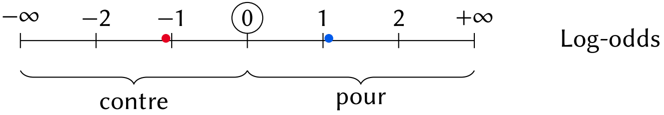


☞ Les cotes (*odds*) ne sont pas symétriques, ce qui les rend moins intuitives

Odds et log-odds

Supposez la note 79. On verra plus tard que :

- Le log des cotes (*log-odds*) pour être dans une ville francophone = $\boxed{\mathbf{1.08}}$
- Le log des cotes (***log-odds***) pour **n**'être **pas** dans une ville francophone = $\boxed{\mathbf{-1.08}}$



☞ Les logs des cotes (***log-odds***) sont symétriques, ce qui les rend **plus** intuitifs

La probabilité

- Nos résultats seront **toujours** en log-odds
- Comment calculer $P(francophone = 1|note = 79)$ en R?

```r
# Manuellement :
exp(1.08) / (1 + exp(1.08))

# Automatiquement :
library(arm) # vous devez installer l'extension
invlogit(1.08)
```

☞ Pour les détails mathématiques, consultez (Garcia, 2021, p. 147)

Probabilités, odds et log-odds

- Voici un tableau de conversion rapide

(Garcia 2021, p. 146)

☞ L'hypothèse nulle est toujours la même : $\hat{\beta} = 0$, c'est-à-dire $P = 0.5$

P	**Odds**	**ln(odds)** = $\hat{\beta}$
0.10	0.11	−2.20
0.20	0.25	−1.39
0.30	0.43	−0.85
0.40	0.67	−0.41
0.50	1.00	0.00
0.60	1.50	0.41
0.70	2.33	0.85
0.80	4.00	1.39
0.90	9.00	2.20

Pratique I

Villes

Dans le fichier `villes2.csv` :

1. Quel type de graphique pourrait être utilisé ici?

   La figure déjà créée est-elle suffisante?
2. Créez un modèle logistique
3. Rapportez les résultats

Pratique I

Correction

Villes : solution

```r
mod1 = glm(francophone ~ note, data = d, family = "binomial")
summary(mod1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.17676    1.18675  -1.834   0.0666 .
note         0.04119    0.01692   2.435   0.0149 *
```

- L'intercept ($\hat{\beta}_0$) : log-odds de `francophone = 1` quand la `note = 0`
- Le slope ($\hat{\beta}_1$) : le changement en log-odds de `francophone = 1` pour chaque unité de `note`

P. ex. Si la `note = 79` $\boxed{-2.17676 + 79 \times 0.04119 \approx 1.08 \text{ (log-odds)} \approx 75\% \text{ (probabilité)}}$

Pratique I

Bonus

- Voici comment rapporter les résultats d'une façon détaillée :

Notre modèle confirme un effet significatif de la variable `note` *sur la variable* `francophone` *($\hat{\beta} = 0,04$, IC 95 % = $[0,0094; 0,0076]$, $p = 0,0149$). Ces résultats indiquent que, pour chaque augmentation de 10 points dans la note d'un apprenant, les log-odds d'être dans une ville francophone augmente de 0,4. Autrement dit, les cotes pour être dans une ville francophone augmentent d'un facteur de 1,5.*[3]

☞ On évite l'interprétation avec des probabilités car sa ligne de tendance n'est pas constante. Donc, le degré de changement dépende de l'intervalle considéré dans l'axe `x`. Si vous voulez utiliser des probabilités, considérez quelques notes spécifiques et calculez ses probabilités.

[3] C'est-à-dire `exp(0.4)`

Pratique I

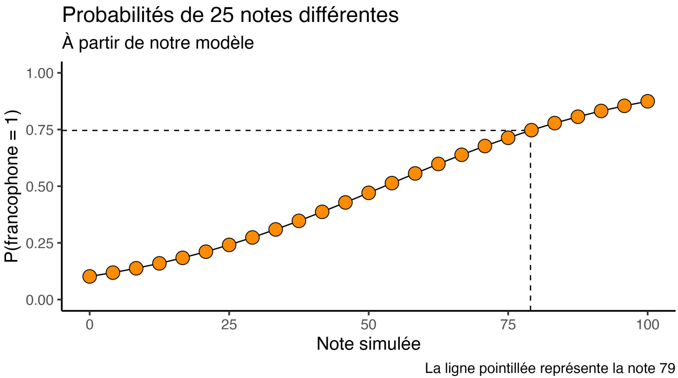
- Voici une façon rapide de prévoir une probabilité (ou plusieurs probabilités) :

```
1 predict(mod1, newdata = tibble(note = 79), type = "response")
2 1
3 0.7459201
4
5 predict(mod1, newdata = tibble(note = c(50, 60, 70, 80)), type = "response")
6 1         2         3         4
7 0.4706742 0.5730795 0.6695812 0.7536465
```

☞ Veuillez noter que les distances entre les probabilités dans la ligne 7 ne sont pas identiques

Prévision de nouvelles valeurs

- **Simulation** : les probabilités prévues pour 25 nouvelles notes à partir de notre modèle



Pratique II

Questionnaire 2

1. Chargez le fichier `francais.csv`
2. Créez une figure et exécutez une régression logistique pour analyser l'effet des variables. Quelle pourrait être la question de recherche?

Extra

Calculs de transformations

```r
# a) Probabilités à partir de log(cotes) :
log_odds <- 1.2 # exemple
prob <- exp(log_odds) / (1 + exp(log_odds))

# b) Log(cotes) à partir de probabilités :
p <- 0.75 # une probabilité
log_odds <- log(p / (1 - p))

# c) Cotes à partir de log-odds :
log_odds <- 1.2
odds <- exp(log_odds)

# d) Cotes à partir de probabilités :
p <- 0.75 # une probabilité
odds <- p / (1 - p)

# e) Probabilités à partir de cotes :
odds <- 3
prob <- odds / (1 + odds)
```

- Voici le code pour jouer avec les transformations et mieux comprendre leurs relations

Extra

Comment créer la figure des prédictions de la séance

```r
# Créer un tibble avec quelques notes hypothétiques :
nd = tibble(note = seq(0, 100, length.out = 25))
pred = tibble(note = seq(0, 100, length.out = 25),
              pred = predict(fit3, newdata = nd, type = "response"))

# Figure :
ggplot(data = pred, aes(x = note, y = pred)) +
  geom_line() +
  theme_classic(base_size = 15) +
  scale_y_continuous(limits = c(0, 1)) +
  annotate("segment", x = 79, xend = 79, y = -Inf, yend = 0.7459201,
           linetype = "dashed") +
  annotate("segment", x = -Inf, xend = 79, y = 0.7459201, yend = 0.7459201,
           linetype = "dashed") +
  geom_point(shape = 21, fill = "darkorange", size = 5) +
  labs(y = "P(francophone = 1)",
       x = "Note simulée",
       title = "Probabilités de 25 notes différentes",
       subtitle = "À partir de notre modèle",
       caption = "La ligne pointillée représente la note 79")
```

Références I

Garcia, G. D. (2021). *Data visualization and analysis in second language research.* Routledge, New York NY.