Université Paris Nanterre
M2 SES-IES Économétrie des variables qualitatives

Examen du 12 avril 2019

Exercice 1

Nous utilisons ici le jeu de données hdv2003 extrait de l'enquête **histoire de vie** réalisée par l'INSEE en 2003, concernant un étude sociologique des configurations familiales. Il contient 2000 individus et 20 variables parmi lesquelles : d'une part des variables décrivant les caractéristiques socio-démographiques des individus (age, sexe, nivetud, etc.), et d'autre part des variables décrivant leurs pratiques de loisirs (hard.rock, lecture.bd, peche.chasse, etc.)

Dans cet exercice, on s'intéresse à savoir si il existe un lien entre les variables : lecture de bandes dessinées (`lecture.bd`) et cinéma au cours des 12 derniers mois (`cinema`) .

1. TABLE 1 – Table des valeurs observées

	lecture.bd	
cinema	Non	Oui
Non	1156	18
Oui	797	29

(a) Préciser les hypothèses nulle et alternative du test d'indépendance.

(b) Nous avons utilisé le logiciel R pour calculer avec les données le tableau des effectifs théoriques (voir tableau ci-dessous)

TABLE 2 – Table des valeurs attendues (effectifs théoriques)

	lecture.bd	
cinema	Non	Oui
Non	1146.411	27.589
Oui	806.589	19.411

Expliquer pourquoi on peut appliquer le test du chi 2 entre ces deux variables ?

(c) Nous avons effectué le test d'indépendance du chi 2 entre les deux variables (voir les sorties de R ci-dessous). Combien vaut la réalisation de la statistique sur les données ? Que pouvez-vous conclure au seuil $\alpha = 5\%$?

```
Pearson's Chi-squared test with Yates' continuity correction

data:  hdv2003.bis$cinema and hdv2003.bis$lecture.bd
X-squared = 7.4246, df = 1, p-value = 0.006434
```

Exercice 2

Cet exercice est tiré de l'article *Prediction for the 2012 United States presidential election using multiple regression model*, écrit par trois chercheurs de l'Université de Delhi. On cherche dans un premier temps à expliquer et dans un second temps à prévoir les résultats des élections présidentielles américaines, entre 1948 et 2012.

On s'intéresse à la variable `Vote`, la proportion des voix obtenue par le parti au pouvoir aux élections ayant lieu à la fin du mandat présidentiel (par rapport au total des voix obtenues par les deux partis ultra-majoritaires, républicain et démocrate), que l'on veut expliquer en fonction de deux types de variables explicatives : des variables économiques et des variables politiques.

Les variables explicatives économiques sont les suivantes :

— `UnemplRate` : taux de chômage annuel moyen lors du mandat écoulé (en %) ;
— `Infl` : inflation annuelle moyenne lors des trois premières années et demie du mandat écoulé (en %) ;
— `GrowthRate` : croissance mesurée lors des trois premiers trimestres de l'année électorale (en %) ;
— . . . , etc.

Les variables explicatives politiques sont quant à elles :
— `Scandal` : variable prenant trois valeurs 0, 1 ou 2, selon qu'au cours du mandat écoulé, il n'y a eu aucun scandale politique, au moins un scandale n'ayant pas soulevé la question d'un *impeachment*, ou un scandale retentissant ayant mené ou ayant failli mener à un *impeachment*
— `RatingJune` : une mesure de popularité de l'action présidentielle effectuée au mois de juin de l'année électorale (précédant donc l'élection de quelques mois) ;
— (`MidTerm`) : variable prenant deux valeurs +1 et -1 selon que le parti du président a conservé ou non la majorité à la chambre des représentants lors des élections de mi-mandat

Year	Vote	UnemplRate	Infl	GrowthRate	Scandal	RatingJune	MidTerm
1944	53.774	NA	0.000	4.279	NA	NA	NA
1948	52.370	3.8	0.000	3.579	1	39.5	-1
1952	44.595	3.0	2.362	0.691	1	31.5	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2000	50.265	4.0	1.605	1.219	2	57.5	-1
2004	51.233	5.5	2.325	2.690	0	48.5	1
2008	46.600	5.8	3.052	0.220	1	29.0	-1

Il faut remarquer que dans cet étude, les auteurs considèrent 10 variables économiques.

1. Analyse descriptive

   (a) Dans cet étude les auteurs considèrent 10 variables économiques. Pensez-vous qu'il y en ait trop ? Aurait-on pu en trouver d'autres ? Justifier votre réponse.

   (b) En 2000, George W. Bush a succédé à Bill Clinton. Que pensez-vous de la valeur de la variable Vote en 2000 ?

2. On cherche à expliquer la variable `Vote` en fonction de quelques variables explicatives. On a ajusté quelques modèles de régression linéaires multiples avec R (parmis les modèles testés, on vous présente seulement deux, voir les sorties de R ci-dessous).

   Modèle 1 : (`model1`)

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.0680     5.0513   6.744 1.37e-05 ***
UnemplRate    0.2628     0.6213   0.423 0.679235
RatingJune    0.3386     0.0680   4.979 0.000252 ***
---
Residual standard error: 3.523 on 13 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared:  0.6561,Adjusted R-squared:  0.6032
F-statistic:  12.4 on 2 and 13 DF,  p-value: 0.0009707
---
AIC(model1):
90.37759
---
RMSE (Racine de l'erreur quadratique moyen par validation croisée)
3.228723
```

Modèle 2 : (`model2`)

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.59925    2.71824  14.936 4.08e-09 ***
GrowthRate   0.75041    0.21905   3.426 0.005024 **
Scandal     -2.74302    0.94565  -2.901 0.013310 *
RatingJune   0.24454    0.04616   5.298 0.000189 ***
---
Residual standard error: 2.147 on 12 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared:  0.882,Adjusted R-squared:  0.8525
F-statistic: 29.91 on 3 and 12 DF,  p-value: 7.494e-06
---
AIC(model2) :
75.25725
---
RMSE (Racine de l'erreur quadratique moyen par validation croisée):
2.088588
```

(a) Dans le modèle 1, Est-ce que la variable taux de chômage `UnemplRate` est significative ? Justifier votre réponse.

(b) Quel modèle choisissez-vous ? Justifier votre réponse ? Justifier que ce modèle est valide statistiquement (ici on vous demande de faire un test de coefficients et un test global, sans oublier de donner les hypothèses nulle et alternatives dans chaque cas) et écrire la relation linéaire proposée avec les coefficients estimés.

(c) Interpréter les coefficients du modèle choisi.

(d) En septembre 2012, la mesure de popularité de Barack Obama était de 46.4%, le taux de croissance sur les trois premiers trimestres était de 1.62%, et il n'y avait pas eu de scandale pendant le mandat 2009-2012. Montrer que la prévision du score de Barack Obama lors des élections de novembre 2012 est de 53.2%.

(e) Comparer la prévision du score de Barack Obama avec le score réalisé de 51.9%. Le modèle choisi prévoyait-il de manière nette le vainqueur de l'élection ?

Exercice 3

Nous traitons un problème de défaut bancaire. La variable `default` est la variable à expliquer. Nous disposons ici d'un échantillon de taille 10000 et deux variables explicatives `student` et `balance`.

— `default` : `Yes` (ou 1) si le client fait défaut sur sa dette et `No` (ou 0) sinon.
— `student` : `Yes` (ou 1) si le client est un étudiant et `No` (ou 0) sinon
— `balance` : montant moyen mensuel d'utilisation de la carte de crédit

Nous avons ajusté deux modèles logistique (simple et multiple)

```
Model1 = glm(formula = default~student, family = binomial(link = "logit"),
data = Default)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
---
AIC: 1600.5
---
10-Fold cross validation: 0.0274
```

```
Model2 = glm(formula = default ~ student + balance,
family = binomial(link = "logit"), data = Default)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.749496  3.692e-01 -29.116  < 2e-16 ***
studentYes  -0.7148776  1.475e-01  -4.846 1.26e-06 ***
balance      0.0057381  2.318e-04  24.750  < 2e-16 ***
---
AIC: 1577.7
---
10-Fold cross validation: 0.0261
```

1. Quel modèle choisissez-vous ? Justifier votre réponse.
2. Donner l'équation du modèle logistique choisi avec les coefficients estimés. Donner également le *classifieur* ou *prédicteur*.
3. Nous avons relevé les valeurs estimées de la proportion de `default` selon les caractéristiques de trois clients au hasard. Est-ce qu'on peut dire si les clients 1, 137 et 9999 feront `default` ? Justifier.

```
predict(Modele2,newdata=Default[c(1,137,9999),c("student","balance")],
type="response")
          1         137        9999
0.001409096 0.050602655 0.148507089
```

Exercice 4

Notons $X \in \mathbb{R}^d$ l'ensemble des variables explicatives et $Y \in \mathbb{R}$ la variable à prédire (la cible). La distribution (loi) de (X, Y) est inconnue. Nous disposons d'un échantillon $\mathsf{D} = \{(x_i, y_i)\}_{i=1}^n$ de n copies indépendantes de (X, Y).

1. Expliquer brièvement le but de la régression.
2. Soit f un estimateur de régression, à quoi correspond la quantité $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$? Est-ce un bon estimateur de la moyenne de l'erreur de prédiction pour un nouveau point x ? Expliquer pourquoi ?
3. Pourquoi la validation croisée est-elle une meilleure méthode ? Expliquer au passage la méthode de validation croisée (brièvement).

Bon courage !