M1SID — UE Modélisation

31 Mars 2021, Durée : 2h

Exercice 1 : Cours

— Décrire le modèle linéaire classique sous sa forme matricielle. On prendra le soin de préciser les dimensions des différentes matrices en question et leur signification.
— Donner la formule de l'estimateur MCO (Moindres Carrés Ordinaires) – OLS (Ordinary Least Squares) en anglais.
— Sous quelle hypothèse est-ce que cela coincide-t-il avec l'estimateur de maximum de vraisemblance ?

Exercice 2 : Système de recommandations

Vous êtes data-analyst dans l'entreprise de commerce en ligne "NOZAMA" et vous êtes en charge de mettre en place le système de recommandation automatique. Votre supérieur vous demande de modéliser un client type par une chaine de Markov.

I. Préambule Grâce à l'aide des ingénieurs en charge du système d'information, vous récoltez les historiques de navigation d'un même individu type sur votre site de vente en ligne. Vous remarquez que l'individu en question visite systématiquement les 4 articles suivants : Whisky 12 ans, Whisky 18 ans, Maillot PSG, Ecran TV 43 pouces. Aussi, vous estimez les transitions suivantes, d'article en article :

— A partir de l'article "Whisky 12 ans", le client clique sur "Whisky 18 ans" avec probabilité 0.75 et "Ecran TV 43 pouces" avec probabilité 0.25.
— A partir de l'article "Whisky 18 ans", le client clique sur "Whisky 12 ans" avec probabilité 0.75 et "Ecran TV 43 pouces" avec probabilité 0.25.
— A partir de l'article "Maillot PSG", le client clique sur "Whisky 12 ans" avec probabilité 0.5 et "Whisky 18 ans" avec probabilité 0.5.
— A partir de l'article "Ecran TV 43 pouces", le client clique sur "Maillot PSG" avec probabilité 1.

En assimilant chaque article à un état, décrire la chaine de Markov associée $X = (X_n \ ; \ n \in \mathbb{N})$. On donnera une description par un graphe orienté ainsi que la matrice de transition.

II. Structure de la chaine de Markov

1. Est-ce que cette chaine de Markov X est irréductible ?
2. Est-elle apériodique ?
3. Calculer la mesure invariante π.

III. Interprétation

— On suppose de le client type a visité les pages de T articles. Expliquer pourquoi la proportion du temps passé sur l'état i est donné par la formule :

$$\frac{1}{T}\sum_{k=0}^{T} \mathbb{1}_{\{X_k=i\}} \ .$$

— Donner cette limite lorsque $T \to \infty$, pour chacun des 4 articles. On justifiera soigneusement la réponse en invoquant un théorème du cours.
— En déduire le classement résultant du système de recommandation. L'idée est qu'il faut recommander les articles qui sont les plus de chances d'être achetés. Et un client est d'autant plus intéressé à acheter qu'il visite souvent un article.

Exercice 3 : Régression logistique et COVID-19

Considérons une étude médicale où l'objectif est d'évaluer l'impact de l'âge et du sexe sur le taux de survie de la COVID-19, pour des patients hospitalisés. Plus précisément, les variables sont :

— la survie (notée Y et codée par 1 s'il y a survie, 0 s'il y a décès).
— l'âge (noté $X = X_1$ et exprimé en années).
— le sexe biologique (noté $Z = X_2$ et codé par 0 si femme, 1 si homme).

Sur 1319 hospitalisations, il y a 224 décès. Pour répondre à la problématique, un modèle de régression logistique a été mis en oeuvre sous la forme :

$$\text{logit}\, \mathbb{P}(Y = 1 \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z \ .$$

Pour un individu i $(i = 1, \ldots, n)$, nous notons les réalisations des variables par y_i, x_i, z_i. De plus :

$$p_i = \mathbb{P}(Y = y_i \mid X = x_i, Z = z_i) \ .$$

I. Modèle blanc

1. Que vaut la probabilité de survie en cas d'hospitalisation.
2. Démontrer qu'en l'absence de facteurs, la log-vraisemblance d'un modèle logistique (sans facteurs) est de la forme :

$$\log L(y; p) = \sum_{i=1}^{n} y_i \log(p) + (1 - y_i) \log(1 - p) \ .$$

3. Calculer la log-vraisemblance du modèle blanc.

II. Modèle de régression logistique. La mise en oeuvre du modèle a permis d'obtenir les résultats suivants :

In [70]:
```
import statsmodels.api as sm

model=sm.Logit(y,factors)
result=model.fit()
result.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.378204
         Iterations 8
```

Out[70]:

Logit Regression Results

Dep. Variable:	y	**No. Observations:**	1319
Model:	Logit	**Df Residuals:**	1316
Method:	MLE	**Df Model:**	2
Date:	Wed, 31 Mar 2021	**Pseudo R-squ.:**	0.1699
Time:	09:07:53	**Log-Likelihood:**	-498.85
converged:	True	**LL-Null:**	-600.95
Covariance Type:	nonrobust	**LLR p-value:**	4.565e-45

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.7302	0.445	12.876	0.000	4.858	6.602
x1	-0.0525	0.005	-10.140	0.000	-0.063	-0.042
x2	-0.7532	0.165	-4.574	0.000	-1.076	-0.430

1. D'après le modèle de regression logistique, écrivez explicitement la probabilité de décès des facteurs et des 3 paramètres β_0, β_1 et β_2. L'expression ne doit utiliser que les fonctions usuelles comme exp et log.
2. On rappelle que la statistique du rapport de vraisemblance sous l'hypothèse H_0 : "$\beta_1 = \beta_2 = 0$" s'écrit :

$$\Lambda_n = 2 \log \frac{L(y; H_1)}{L(y; H_0)} \ ,$$

et est asymptotiquement donné par une loi du $\chi^2(1)$.
Faire le test de cette hypothèse grâce au test du rapport de vraisemblance. Est-ce que les variables explicatives sont pertinentes ?

3. Calculer l'intervalle de confiance du paramètre β_1 de niveau 95%. De même pour β_2. Concluez.
4. Commentez les valeurs estimées de β_1 et β_2 : Quel risque de mortalité pour 10 ans de plus ? Quel risque de mortalité quand on est un homme plutôt qu'une femme ?

   Par risque, on donnera la variation de la probabilité dans chaque cas.