Université Paris Dauphine Année 2001-2002
U.F.R. Mathématiques de la Décision
MD4 - Modèles Linéaires et généralisations

Feuille de Travaux Dirigés 5

Modèles linéaires généralisés

Exercice 1 Nous disposons de deux qualités de papier. Le papier de type 1 a un poids de β_1 grammes par feuille et le papier de type 2 a un poids de β_2 grammes par feuille. Les feuilles arrivent par paquet de 100 et ne doivent pas être séparées. Nous recevons 3 paquets. Dans le premier, il y a 50 feuilles de chaque type. Dans le second, il y 40 feuilles de type 1 et 60 feuilles de type 2, et dans le troisième 60 de type 1 et 40 de type 2. Nous pesons les paquets sur une balance l'un après l'autre. Nous admettons que les erreurs de la balance sont indépendantes et distribuées comme $N(0, \sigma^2)$. Les poids observés sont y_1, y_2 et y_3. Ils sont la réalisation des variables aléatoires Y_1, Y_2 et Y_3 correspondant au poids réel plus une erreur aléatoire.

1 Écrire, sous forme matricielle, le modèle linéaire décrivant ces données.

2 Donner l'estimateur MCO du paramètre inconnu β_1.

Exercice 2 Les modèles catalogués dans la classe des modèles linéaires généralisés sont caractérisés par trois composantes.

La composante aléatoire identifie la distribution de probabilités de la variable à expliquer.

Nous supposons ici que l'échantillon statistique est constitué de n variables aléatoires Y_i indépendantes admettant des distributions issues d'une structure exponentielle particulière. Les lois de ces variables aléatoires sont dominées par une même mesure dite de référence et la famille de leurs densités par rapport à cette mesure se met sous la forme :

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - v(\theta_i)}{\phi} + w(y_i, \phi)\right\}.$$

ϕ est appelé paramètre de dispersion et θ_i est appelé paramètre naturel.
L'expression de la structure exponentielle précédente peut se mettre sous la forme canonique. En posant, $Q(\theta_i) = \left(\frac{\theta_i}{\phi}\right)$, $a(\theta_i) = \exp\left\{\frac{-v(\theta_i)}{\phi}\right\}$, $b(y_i) = \exp\{w(y_i, \phi)\}$. Nous obtenons $f(y_i, \theta_i) = a(\theta_i)\, b(y_i) \exp\{y_i Q(\theta_i)\}$.

La deuxième composante d'un modèle linéaire généralisé est constituée pour chaque Y_i d'une observation planifiée x_i de p variables explicatives. x_i est un vecteur de dimension p.

La dernière composante exprime une relation fonctionnelle entre Y_i et x_i.
Soit $\mathbb{E}(Y_i) = \mu_i$ et β un vecteur de p paramètres inconnus.
Nous supposons que

$$g(\mu_i) = x_i'\beta = \eta_i, \quad \forall i \in \{1, \ldots, n\}$$

où g, appelée fonction de lien, est supposée monotone et différentiable.
Ceci revient alors à écrire un modèle dans lequel une fonction de l'espérance mathématique μ_i appartient au sous-espace vectoriel engendré par les variables explicatives.
La fonction de lien qui associe la moyenne μ_i au paramètre naturel θ_i est appelée fonction de lien canonique, soit $g(\mu_i) = \theta_i = x_i'\beta$.

1 Donner la fonction de lien canonique lorsque, pour tout $i \in \{1, \ldots, n\}$, $Y_i \sim N(\mu_i, \sigma^2)$, $Y_i \sim \mathcal{B}(\pi_i)$ et $Y_i \sim \mathcal{P}(\mu_i)$.

L'estimation des paramètres β_j est calculée en maximisant la log-vraisemblance du modèle linéaire généralisé.

2 Dans le cas général, donner la log-vraisemblance ainsi que ses dérivées premières et secondes par rapport à θ_i. En déduire de nouvelles écritures $\mathbb{E}(Y_i)$ et $\mathbb{V}(Y_i)$.

3 Donner les équations de vraisemblance. Que donnent-elles lorsque la fonction de lien est la fonction de lien canonique ?

4 Appliquer au cas où $Y_i \sim \mathcal{B}(\pi_i)$.

Exercice 3 Nous avons exposé des insectes à différentes concentrations d'un gaz nouveau. Le tableau donné ci-dessous présente les résultats de l'expérience.

Dose x_i	Nombre d'insectes exposés, n_i	Nombre d'insectes morts, y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	59

Pour tout $i \in \{1, \ldots, 8\}$, y_i est la réalisation de la variable aléatoire Y_i suivant une loi binomiale $\mathcal{B}(n_i, \pi_i)$. Les variables aléatoires Y_i sont supposées indépendantes. Par ailleurs, nous supposons que $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$ où β_0 et β_1 sont des paramètres réels inconnus.

1 Calculer la log-vraisemblance de $y_1, \ldots, y_8$.

2 Donner les équations de vraisemblance. Pouvons-nous calculer les expressions analytiques des estimateurs du maximum de vraisemblance des paramètres β_0 et β_1 ?

3 Montrer que la matrice d'information de Fisher peut s'écrire

$$I(\beta_0, \beta_1) = \begin{pmatrix} \sum_{i=1}^{8} n_i \pi_i (1-\pi_i) & \sum_{i=1}^{8} n_i x_i \pi_i (1-\pi_i) \\ \sum_{i=1}^{8} n_i x_i \pi_i (1-\pi_i) & \sum_{i=1}^{8} n_i x_i^2 \pi_i (1-\pi_i) \end{pmatrix}.$$

2 Commenter la sortie SAS donnée ci-dessous.

```
    The GENMOD Procedure

        Model Information

Data Set                    WORK.INSECTES
Distribution                Binomial
Link Function               Logit
```

Response Variable (Events)	Tues
Response Variable (Trials)	Exposes
Observations Used	8
Number Of Events	291
Number Of Trials	481

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	6	11.2322	1.8720
Scaled Deviance	6	11.2322	1.8720
Pearson Chi-Square	6	10.0268	1.6711
Scaled Pearson X2	6	10.0268	1.6711
Log Likelihood		-186.2354	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95 % Confidence Limits	
Intercept	1	-60.7175	5.1807	-70.8715	-50.5634
Dose	1	34.2703	2.9121	28.5626	39.9780
Scale	0	1.0000	0.0000	1.0000	1.0000

NOTE: The scale parameter was held fixed.

Observation Statistics

Obs.	Tues	Expo.	dose	Pred	Xbeta	Std
1	6	59	1.6907	0.058601	-2.776615	0.2870224
2	13	60	1.7242	0.1640279	-1.628559	0.2050525
3	18	62	1.7552	0.362119	-0.566179	0.1472353
4	28	56	1.7842	0.6053149	0.4276606	0.1318339
5	52	63	1.8113	0.7951718	1.3563864	0.1620395
6	53	59	1.8369	0.9032358	2.2337068	0.2146704
7	61	62	1.861	0.9551961	3.0596216	0.2736897
8	60	60	1.8839	0.9790493	3.8444121	0.3337981