Exercice 1 : réponse

chap. 3 p.14 :

Calcul de la vraisemblance.

Poser le calcul de la vraisemblance dans le cas suivant : 4 lots de 20 insectes sont soumis à un insecticide (dose 1, 2, 4 et 8 respectivement). On dénombre les insectes morts dans chaque lot. On trouve 1, 3, 9, 10 respectivement.

$\mathcal{C}_{20}^{1}\ p_1^1\ (1-p_1)^{19}\ \ .\ \ \mathcal{C}_{20}^{3}\ p_2^3\ (1-p_2)^{17}\ \ .\ \ \mathcal{C}_{20}^{9}\ p_3^9\ (1-p_3)^{11}\ \ .\ \ \mathcal{C}_{20}^{10}\ p_4^{10}\ (1-p_4)^{10}$
où p_i est la probabilité de réponse pour la dose i.

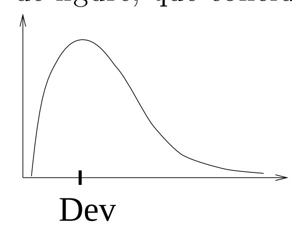
Exercice 2 : réponse

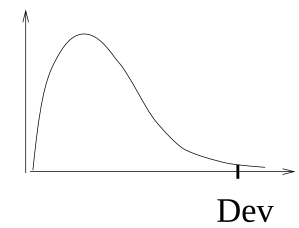
chap. 3 p.17 :

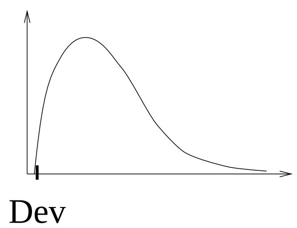
Loi de la déviance.

On étudie l'effet d'un insecticide sur des guêpes. On constitue 10 groupes de 12 guêpes. On applique une dose différente à chacun des groupes, et on observe le nombre de guêpes survivantes dans chacun des groupes. On choisit le lien logit, le logarithme de la dose comme régresseur, et on ajuste un modèle de régression logistique, à deux paramètres, la pente et l'ordonnée à l'origine. *On a 8 degrés de liberté résiduels.*

La déviance obtenue est reportée sur le graphe présentant la distribution d'un $\chi^2(8)$. Trois cas de figure caricaturaux sont proposés. Dans chacun des cas de figure, que concluez-vous ?







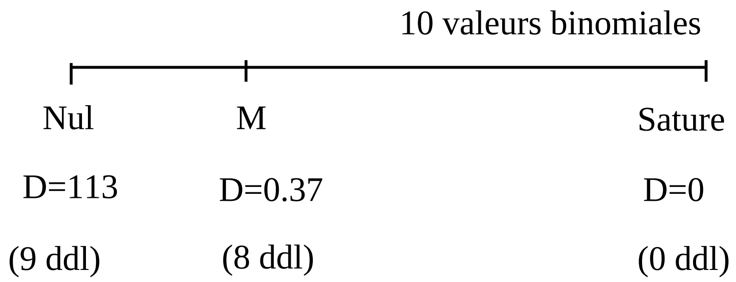
Conclusion :
Dans le premier cas, le modèle est adéquat. Dans le second cas, le modèle n'explique pas tout. Dans le dernier cas, la déviance est suspecte. Les données ont été truquées, ou bien l'expérimentateur est particulièrement chanceux . . .

Exercice 3 : réponse

chap. 3. p.20 :

Test du rapport de vraisemblance.

Exemple “Aircraft”. L’exemple est traité avec SAS et Splus, dans le document “Collett, mise en œuvre avec SAS et Splus”, chapitre 3, Models for binary and binomial data.
A partir des sorties numériques de SAS ou de Splus, représenter sur un schéma le modèle nul, le modèle saturé, le modèle estimé, les déviances, et les degrés de liberté associés. Réaliser un test d’adéquation du modèle, puis un test d’explication. *Conclusion :*



Test d’adéquation : $0.37 \sim \chi^2(8)$ *(p=0.9999579)*

```
> 1-pchisq(0.37,8)
[1] 0.9999579
```

La déviance est bien petite, elle est suspecte.

Test d’explication : $112.83 - 0.37 \sim \chi^2(1)$
Le modèle apporte une explication significative.

Exercice 4 : réponse

chap. 3 p.22 :

Données groupées et non groupées.

exemple “cancer de la prostate”. (Collett, mise en œuvre avec SAS et Splus, chap. 3, Models for binary and binomial data, a futher example on model selection.)

“The data are from the observational study on nodal involvement in patients presenting with prostatic cancer. In the data base there are observations on six variables for each of 53 patients, the variables where as follows.

Age :	age of patient in years,
Acid :	level of serum acid phosphatase,
X-ray :	result of an X-ray examination (0=negative, 1=positive), ´’
Size :	size of tumour (0 = small, 1= large),
Grade :	grade of tumour (0=less serious, 1= more serious),
Nodes :	nodal involvement (0 = no involvement, 1= involvement).

Pour chacun des modèles ci-dessous, dire s’il s’agit de données groupées ou pas. (on suppose que l’on regroupe les données chaque fois que le modèle le permet).

Modèle :	Xray	*Oui*
Modèle :	Xray + Size	*Oui*
Modèle :	Xray + Size + Xray . Size	*Oui*
Modèle :	Xray + Size + Xray . Size + Age	*Non*
Modèle :	Age	*Non*
Modèle :	cut(Age,4)	*Oui*

Exercice 5 : réponse

Exercice indispensable
chap. 3 p.23 :

Interprétation du terme constant

Exemple “Aircraft”. L’exemple est traité avec SAS et Splus, dans le document “Collett, mise en œuvre avec SAS et Splus”, chapitre 3, Models for binary and binomial data.
Tester la nullité du terme constant, dans le cas des deux paramétrisations suivantes du même modèle :

1. $\text{logit}(p) = \alpha + \beta \text{ load}$
2. $\text{logit}(p) = \alpha' + \beta \text{ (load} - \overline{\text{load}})$

Dans le cas des deux paramétrisations suivantes du même modèle, quelle est l’hypothèse nulle correspondant à la nullité du terme constant ?

1. *α est significativement négatif.*
   $logit(p) = 0$, pour $load = 0 \Leftrightarrow \log(\frac{p}{1-p}) = 0 \Leftrightarrow \frac{p}{1-p} = 1 \Leftrightarrow p = \frac{1}{2}$

   H_0 : “$p = \frac{1}{2}$”, pour $load = 0$.

   H_0 : “lorsque la pression est nulle, la probabilité vaut 1/2 que la sangle résiste”.
   Cette hypothèse n’a pas de sens ici. On trouve bien évidemment un $\alpha < 0$, c’est à dire une probabilité significativement inférieure à 1/2.

2. *$logit(p) = 0$, pour $load = \overline{load} \leftrightarrow p = \frac{1}{2}$ pour $load = \overline{load}$*

   H_0 : “$p = \frac{1}{2}$”, pour $load = \overline{load}$.

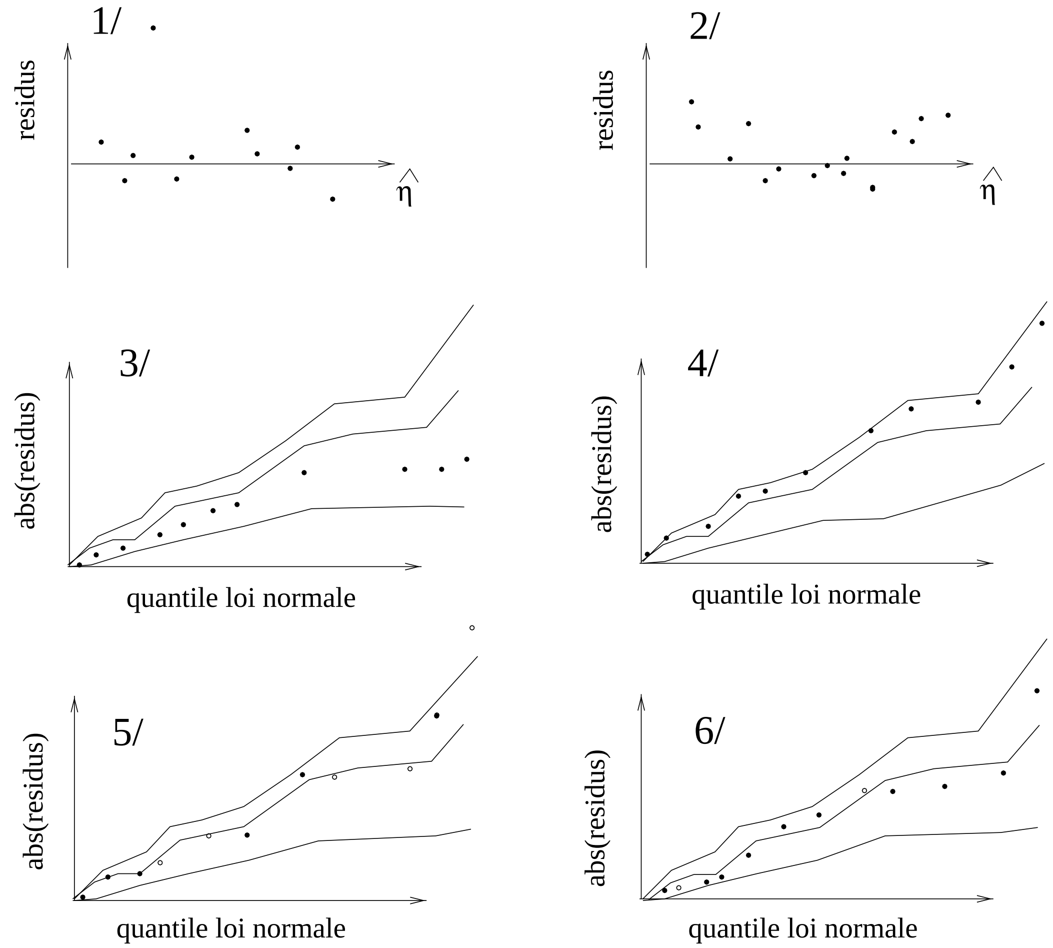
   H_0 : “lorsque la pression est égale à la moyenne des pressions expérimentées, la probabilité vaut 1/2.”

Exercice 6 : réponse

chap. 4 p.13 :

graphe des résidus

Comment interpréter chacun des graphes suivants ? proposer une solution, une amélioration du modèle le cas échéant.



De gauche à droite, et de haut en bas :

1/ un point extérieur. erreur ? se poser des questions.

2/ ajouter un terme quadratique, ou transformer la variable explicative.

3/ modèle surparamétré, modifier le lien.

4/ le modèle n'explique pas tout. Peut-on ajouter des régresseurs, des facteurs, ou des termes d'interaction ?

5/un point extérieur, à grand résidu.

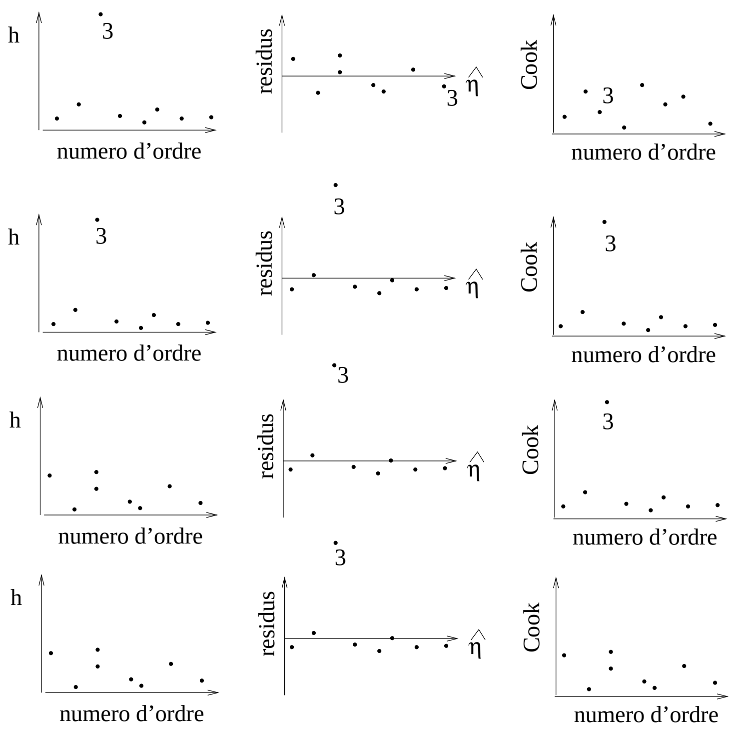
6/ modèle OK.

Exercice 7 : réponse

à la fin du chapitre 4 :

Points influents.

Quel diagnostic, pour chacun des groupes de 3 graphes ?



1/ le point 3 est potentiellement influent, mais il est bien ajusté, et en réalité, il n'est pas influent.
2/ le point 3 diffère des autres, par les valeurs des variables explicatives. Il a un grand résidu, il est donc mal ajusté, et il influe sur les paramètres. Se poser des questions sur cette observation. Réajuster le modèle sans elle, pour voir si les conclusions changent.
3/ le point 3 n'est pas différent des autres, de part les valeurs prises par les variables explicatives. Il a un grand résidu, et il est influent. Cette valeur observée n'est pas compatible avec le modèle ajusté, c'est la seule.
4/ le point 3 n'a pas des valeurs des variables explicatives différentes des autres points. Il a un grand résidu, mais il n'est pas influent. Si on l'enlève, on obtiendra à peu près les mêmes valeurs des paramètres.

Exercice 8 : réponse

à la fin du chapitre 4 :

Choix du régresseur.

- *On choisit la dose comme régresseur. La déviance est plus faible.*
- *C'est un exemple où le* χ^2 *de Pearson est assez différent de la déviance, surtout lorsqu'on choisit le régresseur log(dose).*
- *C'est un exemple où les résidus de Pearson sont assez différents des résidus de déviance.*
- *On a des* $n_i = 7$*, pas tout petits, mais pas très grands quand même. Ce qui explique l'écart entre* χ^2 *de Pearson et déviance, et le fait que les résidus de Pearson ne sont pas distribués selon une loi normale. On semble "assez loin de l'asymptotique".*

Exercice 9 : réponse

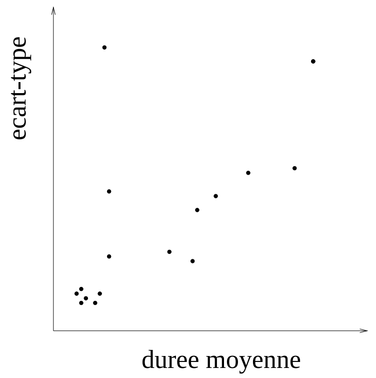
Exercice indispensable
à la fin du chapitre 8 :

Choix de la loi.
A partir de la description des problèmes ci-dessous, choisir une loi succeptible de convenir.

1/ Le nombre total de nouveaux cas de SIDA, déclarés chaque mois, en Angleterre, à partir de novembre 85 est reporté. On souhaite modéliser l'évolution du nombre de nouveaux cas.
0 0 3 0 1 1 1 2 2 4 2 8 0 3 4 5 2 2 2 5 4 3 15 12 7 14 6 10 14 8 19 10 7 20 10 19
Réponse : une loi de poisson peut-être appropriée : la variable à expliquer est discrète, elle correspond à des évènements rares, et indépendants les uns des autres.

2/ L'effet de la température sur la durée du développement embryonnaire de la drosophile est étudié. Voici un extrait des données, et une représentation graphique.

numéro d'expérience	Temp.	Durée moyenne	Taille du lot	Ecart-type
1	14.95	67.5	54	2.41
2	16.16	57.10	182	2.28
...	...	...	...	...
21	30.05	16.81	148	0.39
22	30.80	16.97	195	0.39
23	32.00	18.20	58	2.23



Réponse : la variable à expliquer est continue. Sa variance est proportionnelle au carré de l'espérance. Une loi gamma peut convenir.

3/ On étudie le cancer de la prostate : “The data are from the observational study on nodal involvement in patients presenting with prostatic cancer. In the data base there are observations on six variables for each of 53 patients, the variables where as follows.
Age : age of patient in years,
Acid : level of serum acid phosphatase,
X-ray : result of an X-ray examination (0=negative, 1=positive),
Size : size of tumour (0 = small, 1= large),
Grade : grade of tumour (0=less serious, 1= more serious),
Nodes : nodal involvement (0 = no involvement, 1= involvement).” On cherche à savoir quels sont les variables et facteurs explicatifs de la variable Nodes.
Réponse : la variable à expliquer est discrète, binaire. Les données sont non groupées. Les observations sont indépendantes les unes des autres. Une loi binomiale peut convenir (ou une loi de Bernoulli, et c'est équivalent puisque les n_i valent 1).

4/ On étudie la liaison entre le nombre d'accidents de la route, le volume du trafic, le type de rue, et l'heure de la journée. Les nombres d'accidents reportés sont de 11 9 4 4 20 4. Les 3 premières valeurs correspondent à une première rue, les 3 suivantes à une autre rue. Les heures de la journée sont 7h-9h30, 9h30-15h, 15h-18h30 pour les deux rues. Le volume du trafic associé à chaque observation est 2206 3276 19999 1399 2276 1417.
Réponse : une loi de poisson peut convenir. La variable à expliquer est discrète, elle correspond à un comptage d'évènements rares et indépendants les uns des autres.

5/ On cherche à comparer la proportion de plants de pruniers qui “prennent” ou pas selon qu'ils sont plantés de façon précoce, ou au printemps, et selon que les racines sont coupées court ou long.

	precoce.court	precoce.long	tardif.court	tardif.long
vivant	107	156	31	84
mort	133	84	209	156

Réponse : la variable à expliquer est binaire, les données sont groupées. On suppose que les observations sont indépendantes les unes des autres. Une loi binomiale peut convenir.