Statistique asymptotique : le modèle linéaire généralisé

Agathe Guilloux
Professeure au LaMME - Université d'Évry - Paris Saclay

References I

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin, 2001.

Ludwig Fahrmeir and Heinz Kaufmann. "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models". In: *The Annals of Statistics* (1985), pp. 342–368.

Peter J Green. "Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.2 (1984), pp. 149–170.

Gareth James et al. *An introduction to statistical learning*. Vol. 6. Springer, 2013.

James K Lindsey. *Applying generalized linear models*. Springer Science & Business Media, 1997.

Peter McCullagh and John A Nelder. *Generalized linear models*. Vol. 37. CRC press, 1989.

Introduction

Les jeux de données

"Default" : téléchargement, voir aussi James et al. 2013

On veut expliquer le défaut de paiement (`default=Yes`) ou non (`default=No`) de 833 clients avec les variables

- `student` : `Yes` ou `No`
- `balance` : le solde du compte client
- `income` : les revenus du client

"crab" : téléchargement

On veut expliquer le nombre de mâles "satellites" (`Sa`) autour de $n = 173$ crabes femelles par

- la couleur de la femelle (`C`)
- l'état de sa colonne ("spine condition") (`S`)
- le poids (`Wt`)
- la largeur de sa carapace (`W`)

Modèles de régression et famille exponentielle

- Dans le cas des données `Default`, $Y_i =$`Yes` ou `No` suit une loi de Bernoulli
- dans le cas des données `crab`, $Y_i \in \mathbb{N}$, on pense à la loi de Poisson

dans les deux cas, on veut lier l'espérance de Y_i aux covariables X_i.

Définition et estimation

Modèles de régression et famille exponentielle

On considère donc que les Y_i une densité de la famille exponentielle :

$$f(y_i) = exp\Big(\frac{y_i\theta_i^\star - b(\theta_i^\star)}{\phi^\star} + c(y_i, \phi^\star)\Big)$$

où $\theta_i^\star$ et $\phi^\star$ sont des paramètres inconnus et b, c sont des fonctions déterministes connues.

Espérance et variance dans le modèle exponentiel

On montre alors que

$$\mathbb{E}(Y_i) = b'(\theta_i^\star)$$
$$\mathbb{V}(Y_i) = b''(\theta_i^\star)\phi^\star$$

Fonction de lien (link function)

On introduit alors une fonction de lien g inversible et continument différentiable telle que

$$\eta_i^\star = X_i\beta^\star = g(\mu_i) = g(\mathbb{E}(Y_i)).$$

Quand on choisit la fonction g de telle sorte que $\eta_i^\star = \theta_i^\star$, on parle alors de fonction de lien canonique.

Exercice

Quelles sont les fonction de lien canoniques pour les lois de Bernoulli et de Poisson ?

Vraisemblance et estimation

A partir de l'échantillon (Y_i, X_i), on forme alors la log-vraisemblance (on prend ici le lien canonique)

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{n} \log(f(Y_i)) = \sum_{i=1}^{n} \Big\{ \frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi) \Big\}$$
$$= \sum_{i=1}^{n} \Big\{ \frac{Y_i\eta_i - b(\eta_i)}{\phi} + c(Y_i, \phi) \Big\},$$

la dernière égalité vient des simplifications.

Vraisemblance et estimation

On peut alors définir l'estimateur au maximum de vraisemblance de β par

$$\hat{\beta} = \underset{\beta}{\text{argmax}} \sum_{i=1}^{n} \left\{ \frac{Y_i \eta_i - b(\eta_i)}{\phi} + c(Y_i, \phi) \right\}$$

$$= \underset{\beta}{\text{argmin}} -\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{Y_i \eta_i - b(\eta_i)}{\phi} + c(Y_i, \phi) \right\},$$

ou de façon équivalente (en admettant la convexité) $\hat{\beta}$ est solution de

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i X_i - b'(\eta_i)X_i\right\} = 0.$$

Exercice

Se convaincre avec la régression de Poisson que l'on ne sait pas forcément résoudre cette équation.

On définit également la prédiction $\hat{Y}_i$ de Y_i comme

$$\hat{Y}_i = g^{-1}(X_i\hat{\beta}).$$

IRLS : iterated reweighted least squares

Pour calculer (approcher) l'estimateur au maximum de vraisemblance, on utilise un algorithme de type Newton-Raphson. A l'étape k, on note $\hat{\beta}^k$ la solution courante. On approxime $-\frac{1}{n}\log\mathcal{L} = \ell_n$ par une fonction quadratique :

$$\ell_n(\hat{\beta}^k + h) = \ell_n(\hat{\beta}^k) + \nabla\ell_n(\hat{\beta}^k)^\top h + \frac{1}{2}h^\top\nabla^2\ell_n(\hat{\beta}^k)h$$

puis on minimise cette approximation pour obtenir h^* et on pose

$$\hat{\beta}^{k+1} = \hat{\beta}^k + h^*$$

puis on itère.

Exercice

Comprendre sur la régression de Poisson le nom "IRLS".

Loi asymptotique

Loi asymptotique des estimateurs

On note $\mathcal{I}(\gamma) = \mathbb{E}(\nabla^2 \ell_n(\gamma))$.

Consistance et normalité asymptotique

Sous certaines conditions (cf. Fahrmeir and Kaufman - 1985), on peut montrer que, pour tout vrai paramètre β,

- $|\hat{\beta} - \beta| \xrightarrow{\mathbb{P}} 0$
- $\hat{\beta}$ est asymptotiquement gaussien

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\beta)^{-1})$$

- et même

$$\sqrt{n}(\mathcal{I}(\hat{\beta}))^{1/2}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{Id})$$

C'est en particulier vrai pour les modèles considérés (à fonction de lien canonique) quand les covariables sont bornées et si $\lambda_{\min}(X^\top X) \to \infty$.

Loi asymptotique du vecteur des scores

On appelle vecteur des scores le vecteur

$$s(\hat{\beta}) = -n\nabla\ell_n(\hat{\beta}) = \sum_{i=1}^{n}\left\{Y_iX_i - b'(X_i\hat{\beta})X_i\right\} = X^\top(Y - \vec{b'}(X\hat{\beta})).$$

Le point essentiel pour montrer la normalité asympototique de $\hat{\beta}$ est le résultat suivant.

Normalité asymptotique du vecteur des scores

Sous les mêmes conditions (cf. Fahrmeir and Kaufman - 1985), on peut montrer que

$$\sqrt{n}(\mathcal{I}(\hat{\beta}))^{-1/2}s(\hat{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{Id})$$

Modèles saturé et null

Modèle saturé

Le modèle saturé est le modèle à n paramètres où chaque moyenne de Y_i est remplacée par Y_i. En se rappelant que $\theta_i = g(\mathbb{E}(Y_i))$, le modèle saturé a alors la log-vraissemblance

$$\log \mathcal{L}^{\text{sat}} = \sum_{i=1}^{n} \Big\{ \frac{Y_i g(Y_i) - b(g(Y_i))}{\phi} + c(Y_i, \phi) \Big\}.$$

Modèle `null`

Le modèle `null` est le modèle à 1 paramètre : l'intercept seul. On note $\log \mathcal{L}^{\text{null}}$ sa log-vraisemblance.

Déviance / déviance résiduelle

On définit alors la déviance (ou déviance résiduelle) pour une estimation $\hat{\beta}$ comme

$$D(\hat{\beta}) = 2\Big\{ \log \mathcal{L}^{\text{sat}} - \log \mathcal{L}(\hat{\beta})\Big\} = 2\Big\{ \sum_{i=1}^{n} \Big\{ \frac{Y_i\Big(g(Y_i) - X_i\hat{\beta}\Big) - \Big(b(g(Y_i)) - b(X_i\hat{\beta})\Big)}{\phi} \Big\}$$

Exercice

- Calculer la déviance dans le modèle linéaire gaussien et donner sa loi asymptotique.
- Faire de même dans le modèle logistique

Tests

Tests sur les coefficients (1)

Test de Wald

Puisque

$$\sqrt{n}(\mathcal{I}(\hat{\beta}))^{1/2}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{Id}),$$

on peut tester la nullité d'un coefficient ($H_0 : \beta_j = 0$) grâce à la statistique

$$\sqrt{n}(\mathcal{I}(\hat{\beta}))^{1/2}_{jj}(\hat{\beta}_j)$$

que l'on compare à un fractile de la loi $\mathcal{N}(0, 1)$.

Exercice

Expliciter le test dans le modèle poissonnien.

Tests sur les coefficients (2)

Pour tester une hypothèse $H_0 : \beta_{k_1} = \ldots = \beta_{k_l} = 0$, on utilise en modèle linéaire gaussien la statistique de Fisher

$$\frac{(n-p)\left(\|Y - X\tilde{\beta}\|^2 - \|Y - X\hat{\beta}\|^2\right)}{l\|Y - X\hat{\beta}\|^2} \underset{H_0}{\sim} \mathcal{F}(l, n-p).$$

On rappelle que la loi de $\frac{1}{\sigma^2}\left(\|Y - X\tilde{\beta}\|^2 - \|Y - X\hat{\beta}\|^2\right)$ est une $\chi^2(l)$ sous H_0.

Exercice

Montrer que l'équivalent de $\left(\|Y - X\tilde{\beta}\|^2 - \|Y - X\hat{\beta}\|^2\right)$ dans un modèle linéaire généralisé est

$$D(\tilde{\beta}) - D(\hat{\beta}) \tag{1}$$

Théorème de Wilks

$D(\tilde{\beta}) - D(\hat{\beta}) \xrightarrow{\mathcal{L}} \chi^2(l)$ sous H_0. On teste donc $H_0 : \beta_{k_1} = \ldots = \beta_{k_l} = 0$ à partir de la statistique du rapport de vraisemblance (LRT : likilihood ratio test) $D(\tilde{\beta}) - D(\hat{\beta})$.

Résidus (de déviance)

A nouveau par analogie avec le modèle linéaire gaussien, on construit les résidus de déviance, en identifiant

$$D(\hat{\beta}) = 2\Big\{ \sum_{i=1}^{n} \Big\{ \frac{Y_i\big(g(Y_i) - X_i\hat{\beta}\big) - \big(b(g(Y_i)) - X_i\hat{\beta}\big)}{\phi} \Big\} = \sum_{i=1}^{n} r_i^2.$$

On définit alors

$$r_i = sign(Y_i - \hat{Y}_i)\sqrt{r_i^2}$$

Exercice

Expliciter les résidus dans les modèles logisitique et poissonnien.

Modèle logistique

Modèle logistique

On observe pour $i = 1, \ldots, n$

- des variables explicatives X_i en dimension $p + 1$ (en comptant l'intercept)
- une variable Y_i de loi de Bernoulli $\mathcal{B}\left(1, \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}\right)$.

On définit $\hat{\beta}$ comme l'estimateur au maximum de vraisemblance.

Rapport de côtes ou odd-ratios

Définition : odds ou côte

La quantité $\pi(X_i)/1-\pi(X_i)$ est appelé odds ou côte.

Dans le modèle logistique, on a définit l'odds (ou la côte) par

$$\frac{\pi(X_i)}{1-\pi(X_i)} = \exp(X_i\beta) = \exp(\beta_0 + \beta_1 X_i^1 + \ldots + \beta_p X_i^p).$$

On considère deux individus i_1 et i_2 dont la valeur des covariables ne diffère que pour la j-ième covariable avec $X_{i_1}^j - X_{i_2}^j = 1$, on calcule l'odds-ratio (ou le rapport des côtes)

$$\frac{\pi(X_{i_1})}{1-\pi(X_{i_1})} \Big/ \frac{\pi(X_{i_2})}{1-\pi(X_{i_2})} = \exp(\beta_j)$$

On dira alors qu'une augmentation de 1 de la variable j entraîne une multiplication de l'odds ratio de $\exp(\beta_j)$.

Prédiction

On prédit en régression logistique en calculant pour un nouvel individu avec les covariables X_+

$$\hat{\pi}(X_+) = \frac{\exp(X_+\hat{\beta})}{\exp(X_+\hat{\beta}) + 1},$$

cela nous donne une valeur entre 0 et 1, si on a besoin de prédire 0 ou 1, on compare $\hat{\pi}(X_+)$ à $1/2$, si

$$\hat{\pi}(X_+) > 1/2,$$

on prédit $Y_+^P = 1$ et 0 sinon.

Intervalle de prédiction

On peut également définir un intervalle de confiance pour $\pi(X_i)$ au niveau 0.95 par

$$\left[\frac{\exp(X_i\hat{\beta} - 1.96\hat{s})}{1 + \exp(X_i\hat{\beta} + 1.96\hat{s})}; \frac{\exp(X_i\hat{\beta} + 1.96\hat{s})}{1 + \exp(X_i\hat{\beta} - 1.96\hat{s})}\right]$$

où $\hat{s}$ est un estimateur de l'écart-type de $X_i\hat{\beta}$.

Matrice de confusion

Définitions : matrice de confusion

Pour chaque individu $i = 1, \ldots, n$ de notre échantillon, on note Y_i^P la prédiction de Y_i, on peut construire une matrice de confusion

		Valeurs observées	
		$Y_i = 0$	$Y_i = 1$
Valeurs prédites	$Y_i^P = 0$	TN	FN
	$Y_i^P = 1$	FP	TP
	total	N	P

où P=POSITIVE, N=NEGATIVE, F=FALSE, T=TRUE.

On définit alors

- le **true positive rate ou sensibilité** comme TP/P
- le **false discovery rate** comme FP/(FP+TP)
- le **true negative rate ou spécificité** comme TN/N
- le **false positive rate** comme FP/(FP+TN)=FP/N = 1 - spécificité

Dans notre exemple

		Valeurs observées	
		$Y_i = 0$	$Y_i = 1$
Valeurs prédites	$Y_i^P = 0$	451	52
	$Y_i^P = 1$	49	281
	total	500	333

donc

- le **true positive rate ou sensibilité** vaut environ 0.84
- le **false positive rate** vaut environ 0.1
- le **false discovery rate** vaut environ 0.15
- le **true negative rate ou spécificité** vaut environ 0.9

Courbe ROC

Pour construire les prédictions (Y_i^P), nous avons pris un seuil $1/2$. Si, maintenant, nous faisons varier ce seuil, nous obtenons de nouvelles prédictions définies par

$$\text{si } \hat{\pi}(X_+) > s, \text{ on prédit } Y_+^{P,s} = 1 \text{ et 0 sinon.}$$

Si $s = 0$

		0bs.	
		$Y_i = 0$	$Y_i = 1$
Pred.	$Y_i^P = 0$	0	0
	$Y_i^P = 1$	500	333

donc la sensibilité vaut 1 et la spécificité vaut 0.

Si $s = 1$

		Obs.	
		$Y_i = 0$	$Y_i = 1$
Pred.	$Y_i^P = 0$	500	333
	$Y_i^P = 1$	0	0

donc la sensibilité vaut 0 et la spécificité vaut 1.

Définition : la courbe ROC et l'AUC

La courbe ROC (receiver operating characteristic) représente la sensibilité (qui vaut TP/P) contre 1− la spécificité (qui vaut FP/N) pour toutes les valeurs du seuil entre 0 et 1.

L'AUC (area under the ROC curve) est l'aire sous la courbe ROC.

Une courbe ROC idéale sera collée au coin supérieur gauche, donc plus l'AUC est grande meilleur est le classifieur. Une règle de classification au hasard aura un AUC d'environ 0.5.

Modèle poissonnien

Modèle poissonnien

On observe pour $i = 1, \ldots, n$

- des variables explicatives X_i en dimension $p + 1$ (en comptant l'intercept).
- une variable Y_i de loi de Poisson $\mathcal{P}\left(\exp(X_i\beta)\right)$ soit

$$\log\left(\mathbb{E}(Y_i)\right) = X_i\beta$$

On définit $\hat{\beta}$ comme l'estimateur au maximum de vraisemblance.

Taux relatifs

On considère deux individus i_1 et i_2 dont la valeur des covariables ne diffère que pour la j-ième covariable avec $X^j_{i_1} - X^j_{i_2} = 1$, on calcule alors les espérances

$$\mathbb{E}(Y_{i_1}) = exp(X_{i_1}\beta)$$
$$\mathbb{E}(Y_{i_2}) = exp(X_{i_2}\beta)$$

et leur rapport

$$\frac{\mathbb{E}(Y_{i_1})}{\mathbb{E}(Y_{i_2})} = exp(\beta_j)$$

ainsi $\exp(\beta_j)$ est la valeur par laquelle est multipliée l'espérance de la variable à expliquée quand X^j augmente d'une unité, on l'appelle taux relatif.

Prédiction

On prédit en régression poissonnienne en calculant pour un nouvel individu avec les covariables X_+

$$\hat{Y}_+^P = \hat{\lambda}(X_+) = \exp(X_+\hat{\beta}).$$

Intervalle de prédiction

On peut également définir un intervalle de confiance pour $\pi(X_i)$ au niveau 0.95 par

$$\left[\exp(X_i\hat{\beta} - 1.96\hat{s}); \exp(X_i\hat{\beta} + 1.96\hat{s})\right]$$

où $\hat{s}$ est un estimateur de l'écart-type de $X_i\hat{\beta}$.

Mesure de la qualité d'un modèle

Pour mesure la qualité d'un modèle, on utilisera une erreur absolue en pourcent moyenne (mean absolute percentage error)

$$\text{MAPE} = \sum_i \left| \frac{\hat{Y}_i^P - Y_i}{Y_i + 1} \right|$$

Variable d'exposition et offset

Les modèles pour les comptes incluent souvent une variable d'exposition (`exposure`), qui indique combien de temps a duré l'exposition ou le nombre de fois où l'événement pourrait avoir eu lieu. Dans ce cas, on veut plutôt modéliser le rapport entre le nombre observé et cette variable d'exposition

$$\log\big(\mathbb{E}(Y_i)/\text{exposure}_i\big) = X_i\beta$$

ce revient à inclure la variable d'exposition dans le modèle

$$\log\big(\mathbb{E}(Y_i)\big) = X_i\beta + \log\big(\text{exposure}_i\big)$$

En R, il faudra mettre `offset(log(exposure))`.

Sélection de variables ℓ_0

Test pour modèles emboîtés

Pour tester $H_0 : \beta_{k_1} = \ldots = \beta_{k_l} = 0$, on utilise le test de Fisher.

- On note $W = \text{vect}\{(\mathbf{1}, X_1, \ldots, X_p)/(X_{k_1}, \ldots, X_{k_l})\}$ de dimension $p + 1 - l$.
- On note $X\tilde{\beta} = \text{proj}_W^\top(Y)$ et on a toujours $X\hat{\beta} = \text{proj}_V^\top(Y)$.

Statistique de Fisher

On a

$$\frac{(n-p-1)\big(\|Y - X\tilde{\beta}\|^2 - \|Y - X\hat{\beta}\|^2\big)}{l\|Y - X\hat{\beta}\|^2} \underset{H_0}{\sim} \mathcal{F}(l, n-p-1).$$

Test du rapport de vraisemblance (LRT)

Pour tester $H_0 : \beta_{k_1} = \ldots = \beta_{k_l} = 0$, dans un modèle linéaire généralisé,on utilise le test du rapport de vraisemblance (LRT).

- On note $\hat{\beta}$ l'estimateur du maximum de vraisemblance dans le modèle complet
- et $\tilde{\beta}$ l'estimateur du maximum de vraisemblance dans le modèle sans les variables $X^{k_1}, \ldots, X^{k_l}$

Statistique du LRT

- On sait que

$$D^0(\hat{\beta}) - D^0(\tilde{\beta}) = -2\log(\mathcal{L}(\tilde{\beta})) + 2\log(\mathcal{L}(\hat{\beta}))$$

  suit, sous $\mathcal{H}_0$, si n est grand, une loi du $\chi^2(l)$
- On rejette $\mathcal{H}_0$ au niveau α si $D^0(\hat{\beta}) - D^0(\tilde{\beta}) > z'$ où z' est un fractile de la loi du $\chi^2(l)$.

Modèles, vrai modèle

On se donne une famille de modèles $\mathcal{M}$, par exemple

$$\mathcal{M} = \mathcal{P}\{1, \ldots, p\} = 1, (1, X^1), (1, X^2), \ldots, (1, X^1, X^2 n \ldots, X^p).$$

On suppose qu'il existe un vrai modèle $m^* \in \mathcal{M}$ tel que :

$$\mathbb{E}(Y) = g^{-1}(X^{(m^*)}\beta^{(m^*)}).$$

On veut retrouver m^*.

- **Attention :** le R^2 ou la log-vraisemablence ne sont pas des bons critères pour ce problème car ils choisiront toujours le modèle complet (avec toutes les covariables)
- On note $m^{\text{full}} = (1, X^1, X^2, \ldots, X^p)$ le modèle complet.

Estimation dans le modèle m

Dans le modèle m, on note $|m|$ le nombre de covariables qu'il contient

$$\hat{\beta}^{(m)}$$

l'estimateur au maximum de vraisemblance dans ce modèle.

Pour les glm

AIC/BIC

On choisit $\hat{m}_{AIC} \in \mathcal{M}$ et $\hat{m}_{BIC} \in \mathcal{M}$tel que :

$$\hat{m}_{AIC} = \underset{m \in \mathcal{M}}{\text{argmin}}\, AIC(m),$$
$$\hat{m}_{BIC} = \underset{m \in \mathcal{M}}{\text{argmin}}\, BIC(m),$$

avec

$$AIC(m) = -\frac{2}{n} \log \mathcal{L}(\hat{\beta}^{(m)}) + 2\frac{|m|}{n}$$
$$BIC(m) = -\frac{2}{n} \log \mathcal{L}(\hat{\beta}^{(m)}) + \frac{\log(n)|m|}{n}$$

où $\log \mathcal{L}(\hat{\beta}^{(m)})$ est la log-vraisemblance dans le modèle m.

Régressions pénalisées

Régression Ridge

Pénalité ridge

$$\hat{\beta}_{\lambda}^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} -\frac{1}{n} \log \mathcal{L}(\hat{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2$$

Le lasso

Introduit en 1996 par Tibshirani, la lasso peut être vu comme un intermédiaire entre la régression ridge et la sélection ℓ_0.

Pénalité lasso

$$\begin{aligned}\hat{\beta}_\lambda^{\text{lasso}} &= \underset{\beta \in \mathbb{R}^p}{\text{argmin}} -\frac{1}{n} \log \mathcal{L}(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \\ &= \underset{\beta \in \mathbb{R}^p}{\text{argmin}} -\frac{1}{n} \log \mathcal{L}(\hat{\beta}) + \lambda \|\beta\|_1.\end{aligned}$$

L'elastic net

Elastic net

$$\hat{\beta}_\lambda^{\text{en}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} -\frac{1}{n} \log \mathcal{L}(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

Attention dans `glmnet`, l'elastic-net est défini par

$$\hat{\beta}_\lambda^{\text{en}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} -\frac{1}{n} \log \mathcal{L}(\beta) + \lambda \Big(\alpha \|\beta\|_1 + \frac{1-\alpha}{2} \|\beta\|_2^2 \Big)$$

Sur-dispersion

Retour sur la famille exponentielle

Nous avons considéré que les Y_i ont une densité de la famille exponentielle :

$$f(y_i) = exp\Big(\frac{y_i\theta_i^\star - b(\theta_i^\star)}{\phi^\star} + c(y_i, \phi^\star)\Big)$$

où

- $\theta_i^\star$ et a_i, b, c sont des fonctions déterministes connues.
- et $\phi^\star = 1$ pour les modèles Bernoulli et Poisson

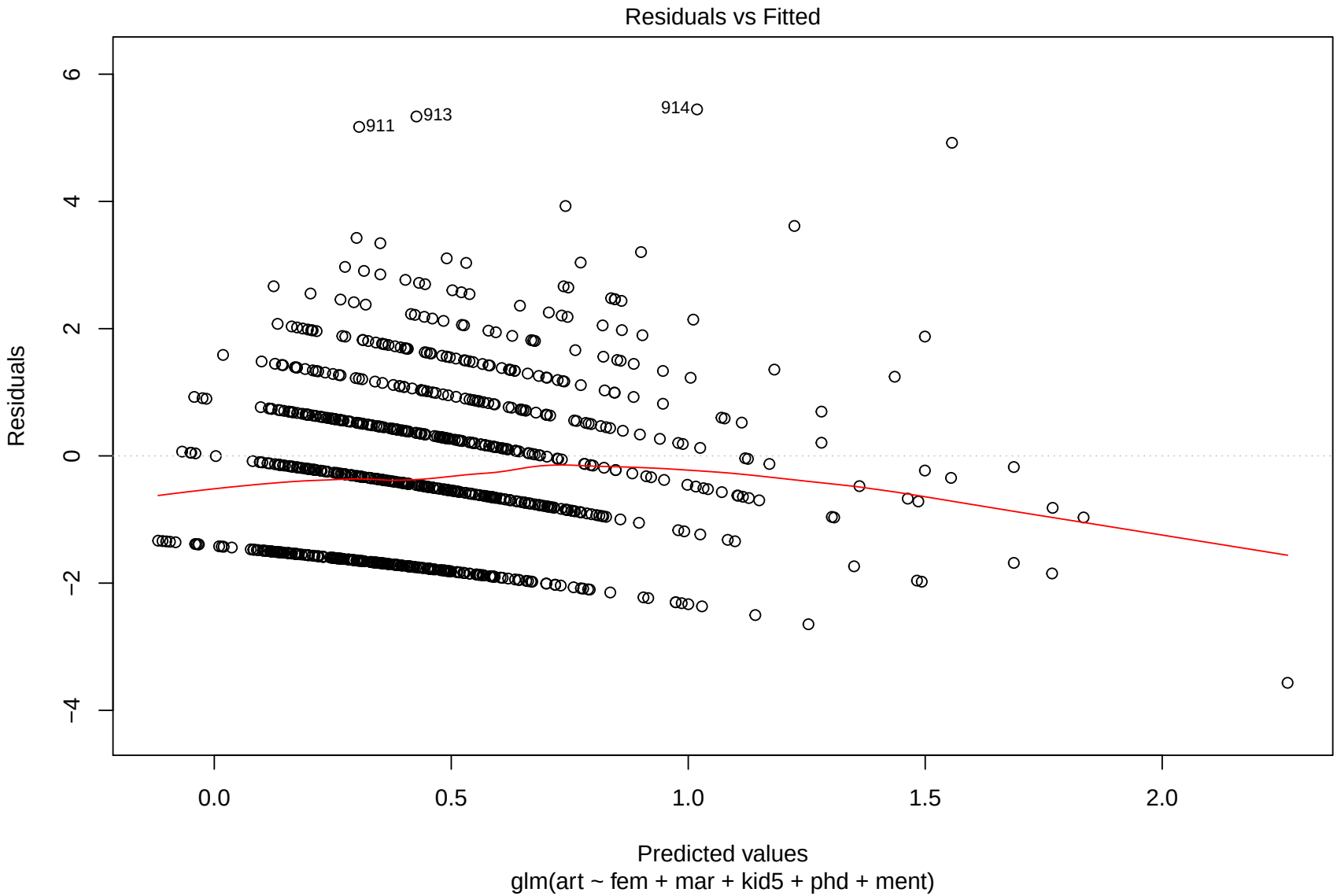
En conséquence

$$\mathbb{E}(Y_i) = b'(\theta_i^\star)$$
$$\mathbb{V}(Y_i) = b''(\theta_i^\star)\phi^\star = b''(\theta_i^\star).$$

```r
fit_poisson = glm(art ~fem + mar + kid5 + phd + ment,
                  family = "poisson", data = couart2)
summary(fit_poisson)
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5672  -1.5398  -0.3660   0.5722   5.4467
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.459860   0.093335   4.927 8.35e-07 ***
## femWomen    -0.224594   0.054613  -4.112 3.92e-05 ***
## marSingle   -0.155243   0.061374  -2.529   0.0114 *
## kid5        -0.184883   0.040127  -4.607 4.08e-06 ***
## phd          0.012823   0.026397   0.486   0.6271
## ment         0.025543   0.002006  12.733  < 2e-16 ***
## ---
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1634.4  on 909  degrees of freedom
## AIC: 3314.1
```

```r
plot(fit_poisson,which=1)
```



Approche par quasi-vraisemblance

Une première solution consiste à ne plus considérer $\phi^\star = 1$. La (quasi) log-vraisemblance est alors donnée par

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{n} \log(f(Y_i)) = \sum_{i=1}^{n} \Big\{ \frac{Y_i \eta_i - b(\eta_i)}{\phi} + c(Y_i, \phi) \Big\},$$

pour le lien canonique $\theta_i = \eta_i = X_i \beta$.

On garde l'équation vérifiée par l'estimateur au maximum de vraisemblance $\hat{\beta}$ de $\beta^\star$

$$-\frac{1}{n} \sum_{i=1}^{n} \Big\{ Y_i X_i - b'(\hat{\beta} X_i) X_i \Big\} = 0.$$

Estimation de $\phi^\star$

On remarque ensuite que

$$\mathbb{E}\left(\frac{(Y_i - b'(X_i\beta^\star))^2}{b''(X_i\beta^\star)}\right) = \phi^\star.$$

On définit donc un estimateur de $\phi^\star$ par

$$\hat{\phi} = \frac{1}{n-p-1}\sum_{i=1}^{n}\left(\frac{(Y_i - b'(X_i\hat{\beta}))^2}{b''(X_i\hat{\beta})}\right).$$

Le dénominateur $n-p-1$ est pris par analogie avec l'estimation de σ^2 dans le modèle linéaire.

quasi-likelihood

```
fit_poisson_quasi = glm(art ~fem + mar + kid5 + phd + ment,
                        family = "quasipoisson", data = couart2)
summary(fit_poisson_quasi)
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5672  -1.5398  -0.3660   0.5722   5.4467
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.459860   0.126227   3.643 0.000285 ***
## femWomen    -0.224594   0.073860  -3.041 0.002427 **
## marSingle   -0.155243   0.083003  -1.870 0.061759 .
## kid5        -0.184883   0.054268  -3.407 0.000686 ***
## phd          0.012823   0.035700   0.359 0.719544
## ment         0.025543   0.002713   9.415  < 2e-16 ***
## ---
## (Dispersion parameter for quasipoisson family taken to be 1.829006)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1634.4  on 909  degrees of freedom
```

Approche par mélange : glm hiérarchique

Une autre approche consiste à considérer que les paramètres $\theta_i^\star$ sont eux-mêmes aléatoires.

On choisit des lois conjuguées avec la loi des Y_i dans le modèle considéré, c'est-à-dire

- la famille des lois γ pour le modèle poissonnien
- la famille des lois β pour le modèle logistique
- etc.

Modèle Poisson-gamma ou négative binomiale

Considérons que

- la loi de Y_i conditionnellement à $\tilde{\theta}_i$ est $\mathcal{P}(\tilde{\theta}_i)$
- la loi de $\tilde{\theta}_i$ est $\gamma\left(\eta^\star, \eta^\star \theta_i^{\star,-1}\right)$.

Dans ce cas, on sait que

- Y_i a une loi binomiale-négative $NB\left(\eta^\star, \eta^\star/(\eta^\star + \theta_i^\star)\right)$et
- $\mathbb{E}(Y_i) = \theta_i^\star$
- $\mathbb{V}(Y_i) = \theta_i^\star + \theta_i^{\star,2}/\eta^\star = \theta_i^\star \phi^\star$ en posant $\phi^\star = (1 + \theta_i^\star/\eta^\star)$

Negative binomial

```
fit_negative_binomial = glm.nb(art ~fem + mar + kid5 + phd + ment,
                               data = couart2)
summary(fit_negative_binomial)
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1678  -1.3617  -0.2806   0.4476   3.4524
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.406633   0.125778   3.233 0.001225 **
## femWomen    -0.216418   0.072636  -2.979 0.002887 **
## marSingle   -0.150489   0.082097  -1.833 0.066791 .
## kid5        -0.176415   0.052813  -3.340 0.000837 ***
## phd          0.015271   0.035873   0.426 0.670326
## ment         0.029082   0.003214   9.048  < 2e-16 ***
## ---
##     Null deviance: 1109.0  on 914  degrees of freedom
## Residual deviance: 1004.3  on 909  degrees of freedom
## AIC: 3135.9
##
##               Theta:  2.264
##           Std. Err.:  0.271
```

Outline