Université de Bourgogne 2021-2022

Construction de modèles linéaires généralisés

MONA Raphaël SAUCE Justine

9 Mai 2022

Table des matières

Introduction

Les modèles linéaires généralisés, abrégés par GLM (Generalized Linear Models), permettent d'étudier la liaison entre la variable à expliquer, que l'on note Y, et les variables explicatives $(X_1, \ldots, X_n)$.

Les GLM sont formés à partir de trois composantes :

- Y : variable à expliquer.
- $(X_1, \ldots, X_n)$: variables explicatives.
- la fonction lien (link function), définissant la relation entre l'espérance de la variable à expliquer, notée μ dans la suite, et les variables explicatives.

Qu'est-ce qu'un modèle linéaire généralisé ?

Une structure basique de GLM est la suivante :

$$g(\mu_i) = X_i \, \beta \tag{1}$$

où $\mu = \mathbb{E}(Y_i)$, g est une fonction lien monotone et régulière, X_i est la i$^{\text{ème}}$ ligne d'une matrice X, formée des variables explicatives, enfin β un vecteur de $\mathbb{R}^n$ de paramètres inconnus.

Motivation. La différence avec la régression linéaire basique où l'on modélise l'espérance μ est que, dans le cas des GLM, on modélise une fonction de l'espérance $g(\mu)$. Les GLM sont utiles notamment dans le cas où la variable à expliquer n'est pas continue (par exemple pour une variable de comptage).

En général, le modèle linéaire est fondé sur l'hypothèse que les résidus $\epsilon_i = Y_i - X_i\beta$ ainsi que la variable Y sont distribués selon une loi normale. Cela suppose que le lien entre X et Y est linéaire, ce qui n'est souvent pas le cas (par exemple l'évolution des températures au cours d'une journée, ou encore l'évolution du nombre de cas de sida par an). L'intérêt des GLM est de tenir compte de cette non-normalité des résidus. Le but étant choisir une distribution plus adaptée pour la modélisation.

Les GLM sont donc une extension des modèles classiques qui s'appliquent au cas où la variable à expliquer est qualitative. Ils sont utilisés essentiellement lorsque Y est une variable de type comptage, ordinale ou non, ou encore binaire.

Hypothèses. Les modèles linéaires généralisés reposent sur les hypothèses suivantes :

$\mathbf{H_0}$: La distribution de la variable Y_i appartient à la famille exponentielle, pour $i = 1, \ldots, n$.
$\mathbf{H_1}$: $Y_1, \ldots, Y_n$ sont indépendantes.

Modèles linéaires généralisés

1 Comment construire un GLM ?

1.1 Quelques propriétés de la famille exponentielle

Il y a plusieurs façons équivalentes de définir l'appartenance à la famille exponentielle, ici on la définira de la même façon que dans le livre de Wood, en introduisant le paramètre ϕ.

Définition. Une famille exponentielle est un ensemble de lois (de v.a. discrète ou continue) caractérisées par une densité qui peut s'écrire sous la forme suivante,

$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{2}$$

où, Y est la variable aléatoire (et y une observation de Y), a, b et c sont des fonctions arbitraires, θ le paramètre canonique et ϕ un paramètre arbitraire (appelé de dispersion).

Sous les hypothèses $\mathbf{H_0}$ et $\mathbf{H_1}$, on peut noter des liens entre les fonctions a, b, le paramètre θ et l'espérance de Y.

La log-vraisemblance pour une seule observation y de la variable de Y est donnée par,

$$\ln(L(y,\theta)) = \ln(f_\theta(y)) = \left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

En s'intéressant maintenant à la fonction score, on a,

$$S(y,\theta) = \frac{\partial \ln(L(y,\theta))}{\partial\theta} = \frac{y - b'(\theta)}{a(\phi)}$$

En considèrant $\ln(L(y,\theta))$ comme une variable aléatoire et en remplaçant la réalisation y par sa variable aléatoire, cela nous donne,

$$\mathbb{E}\left(\frac{\partial \ln(L(y,\theta))}{\partial\theta}\right) = \frac{\mathbb{E}(Y) - b'(\theta)}{a(\phi)}$$

De plus, sous certaines hypothèses de régularités [1], $\mathbb{E}\left(S(y,\theta)\right) = 0$ donc,

$$\mathbb{E}(Y) = b'(\theta) \tag{3}$$

Ainsi, la moyenne de toute variable aléatoire qui appartient à la famille exponentielle est égale à la dérivée première de b. Cette propriété nous permets alors de faire le lien l'espérance μ de Y et le paramètre canonique β de la famille exponentielle.

1. Les dérivées $1^{\text{ère}}$ et 2^{nd} de la fonction de vraisemblance existent en tout point y, et vues comme des fonctions de y elle sont intégrables quelque-soit le paramètre θ. On suppose aussi l'intervertion intégrales et dérivées.

De la même façon, on peut déterminer une propriété sur la variance de Y qui nous sera utile pour la suite.

En effet, en dérivant une seconde fois par rapport à θ on obtient,

$$\frac{\partial^2 \ln(L(y,\theta))}{\partial\theta^2} = \frac{-b''(\theta)}{a(\phi)}$$

De plus, sous certaines conditions de réguralités [2], on a,

$$\mathrm{Var}(S(y,\theta)) = \mathbb{E}\left(\left(\frac{\partial \ln(L(y,\theta))}{\partial\theta}\right)^2\right) = -\mathbb{E}\left(\frac{\partial^2 \ln(L(y,\theta))}{\partial\theta^2}\right)$$

Ce qui nous livre que,

$$-\frac{b''(\theta)}{a(\phi)} = -\frac{\mathbb{E}((Y-b'(\theta))^2)}{a(\phi)^2}$$

$$\Leftrightarrow -b''(\theta)a(\phi) = -\mathbb{E}((Y-b'(\theta))^2)$$

Ainsi,

$$\mathrm{Var}(Y) = b''(\theta)a(\theta)$$

En pratique, dans le cas où le paramètre ϕ est inconnu, la forme de a peut être définie comme suit,

$$a(\phi) = \frac{\phi}{\omega}$$

où ω est un réel fixé. Même si cette forme est restrictive, elle est adaptée dans la plupart des cas pour un GLM, et souvent on aura $\omega = 1$. Ce qui nous donne,

$$\mathrm{Var}(Y) = \frac{b''(\theta)\phi}{\omega} \tag{4}$$

Pour la suite on aura tendance à exprimer la variance à partir d'une fonction de μ, on posera,

$$V(\mu) = \frac{\mathrm{Var}(Y)}{\phi}$$

1.2 Fonction de lien

La fonction de lien dans les GLM :

$$g(\mu_i) = X_i\,\beta \tag{5}$$

Pour retrouver la prédiction moyenne, il est nécessaire d'appliquer la fonction de lien inverse, on obtiendra alors, $\mu_i = g^{-1}(X_i\,\beta)$.

Le principe sous l'introduction d'une fonction lien est de contraindre les valeurs que l'on va prédire à l'aide du modèle à être dans l'échelle des valeurs observées.

2. Les conditions de la note [1] et $\mathbb{E}(S(y,\theta)^2) < \infty$

Ainsi on fournira des prédictions bien plus cohérentes.

Remarque. Les β sont estimés après application de la fonction lien aux variables explicatives.

En général, la fonction de lien utilisée est la fonction lien dite canonique, déterminée par le paramètre canonique de la famille exponentielle. Par exemple, lorsque Y suit une loi de poisson, on a,

$$f(y) = \frac{\mu^y \exp(-\mu)}{y!} = \exp(y \ln(\mu) - \mu - \ln(y!))$$

Où,

$$\begin{aligned} \theta &= \ln(\mu) \\ b(\theta) &= \mu = \exp(\theta) \\ a(\phi) &= 1 \\ c(y, \theta) &= -\ln(y!) \end{aligned}$$

Ainsi,

$$\begin{aligned} \mu = \mathbb{E}(Y) = b'(\theta) = \exp(\theta) \\ \Leftrightarrow \theta = \ln(\mu) \end{aligned}$$

On obtient donc la fonction de lien canonique g suivante,

$$g(\mu) = \ln(\mu)$$

1.3 Estimation du paramètre β par maximum de vraisemblance

1.3.1 Fonction de vraisemblance

L'estimation par maximum de vraisemblance, abrégé par emv, consiste à maximiser la vraisemblance définie par,

$$\begin{array}{rccc} L & : & \mathbb{R}^n & \to & \mathbb{R} \\ & & \beta & \mapsto & \prod_i^n f_\theta(y_i) \end{array}$$

sous $\mathbf{H_0}$ et $\mathbf{H_1}$.

La vraisemblance s'exprime comme une fonction de β puisque les θ_i pour $i = 1, \ldots, n$ sont dépendants du paramètre inconnu β.
La log-vraisemblance sera donc donnée par,

$$\ln(L(y_1, \ldots, y_n, \beta) = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b_i(\theta_i)}{a_i(\phi)} + c_i(\phi, y_i) \right)$$

Remarque. Le paramètre ϕ est supposé être identique $\forall i \in \{1, \ldots, n\}$, contrairement aux fonction a, b et c qui elles peuvent varier.

En faisant la même supposition que pour 4, on peut écrire,

$$\ln(L(y_1,\ldots,y_n,\beta) = \sum_{i=1}^{n}\left(\frac{y_i\theta_i - b_i(\theta_i)}{\phi}\omega_i + c_i(\phi,y_i)\right)$$

Pour maximiser cette quantité, on différentie partiellement cette dernière par rapport à chaque β_j pour ensuite résoudre l'équation de vraisemblance pour β. On obtient donc :

$$\frac{\partial ln(L(y_1,\ldots,y_n,\beta)}{\partial\beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\left(y_i\frac{\partial\theta_i}{\partial\beta_j} - b_i'(\theta_i)\frac{\partial\theta_i}{\partial\beta_j}\right)$$

et avec l'aide de la règle de la chaîne, et en différentiant (3), on obtient :

$$\frac{\partial\mu_i}{\partial\theta_i} = b_i''(\theta_i) \Rightarrow \frac{\partial\theta_i}{\partial\mu_i} = \frac{1}{b_i''(\theta_i)}$$

Ce qui implique que,

$$\frac{\partial ln(L(y_1,\ldots,y_n,\beta)}{\partial\beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\left(\frac{(y_i - b_i'(\theta_i))}{b_i''(\theta_i)/\omega_i}\frac{\partial\mu_i}{\partial\beta_j}\right)$$

et donc en substituant (3) et (4), on doit donc résoudre pour β :

$$\forall j, \sum_{i=1}^{n}\left(\frac{(y_i-\mu_i)}{V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_j}\right) = 0$$

Remarque. On peut noter que ces équations sont les mêmes que celles qu'il faut résoudre pour trouver β dans le problème des moindres carrés si les valeurs $V(\mu_i)$ étaient connues et indépendantes de β.

R commande. Sous le logiciel R, la recherche des β se fait via la fonction `glm()`. Elle utilise un algorithme appelé Fisher Scoring basé sur une méthode de Newton pour résoudre numériquement les équations de vraisemblance.

1.3.2 Propriétés de l'estimateur du maximum de vraisemblance

- Pour les GLM, on a montré que l'EMV existe et qu'il est unique.
- Il est consistent, asymptotiquement efficace et asymptotiquement normal.

2 Adéquation du modèle aux données

2.1 Deviance

Quand on travaille avec des GLM, il est pratique d'avoir une quantité pouvant être interprétée un peu comme la somme résiduelle des carrés. On définit donc la déviance de la façon suivante :

$$D = 2(\ln(L(\hat{\beta}_{\text{emv}}) - \ln(L(\hat{\beta}))\phi$$

où, $\ln(L(\hat{\beta}_{\text{emv}}))$ indique le maximum de vraisemblance du modèle saturé.

Ce qu'il faut retenir est que la déviance est une mesure d'écart entre les valeurs attendues et les observations calculée à partir de la vraisemblance du modèle pour les paramètres estimés.

2.2 Résidus

Dans le cas des modèles linéaires ordinaires, on étudie les résidus bruts du modèle pour vérifier la cohérence du modèle. Cette étude est plus complexe dans le cas des GLM étant donné qu'il est difficile de vérifier la validité de la relation supposée entre la moyenne et la variance à partir des résidus bruts. Pour faciliter la compréhension, on peut prendre pour exemple un modèle de Poisson : la variance des résidus devrait augmenter proportionnellement avec la taille des valeurs ajustées. Pour cela, dans le cas des GLM, on normalise les résidus et grâce à cette manoeuvre, si les hyphotèses du modèle sont correctes, les résidus normalisés doivent avoir une variance approximativement égale.

2.2.1 Résidus de Pearson

La façon la plus intuitive de normaliser les résidus est de les diviser par une valeur proportionnel à l'écart-type attendu de la réponse . C'est ainsi que sont définit les résidus de Pearson :

$$\epsilon_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

Si le modèle est correctement ajusté, les résidus doivent avoir une moyenne approximativement nulle et une variance ϕ. Ces résidus devraient donc avoir une variance plus homogène que les résidus bruts.

2.2.2 Résidus déviance

Dans la pratique, la distribution des résidus de Pearson n'est pas symétrique autour de zéro et donc le résultat n'est pas aussi proche des résidus du modèle linéaire ordinaire que l'on pourrait espérer. C'est pourquoi on introduit les résidus de déviance. Ces derniers sont obtenus en supposant que la déviance, joue le même rôle pour les GLM que la somme résiduelle des carrés pour les modèles linéaires ordinaires. Ainsi, si on écrit d_i la composante de la déviance du $i^{\text{ème}}$ terme de la somme on obtient :

$$D = \sum_{i=1}^{n} d_i$$

et donc, en analogie avec le modèle linéaire ordinaire, on définit :

$$\hat{\epsilon_i^d} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

2.3 Pseudo-R^2

On introduit alors le pseudo-R^2, définit à partir de la déviance du modèle :

$$R^2 = \frac{\text{Déviance nulle} - \text{Déviance}}{\text{Déviance nulle}}$$

Où la déviance nulle d'un modèle correspond à la déviance du modèle nul ne comptant aucun prédicteur.

3 Qualité d'estimation des coefficients

3.1 Critère AIC

Le critère d'Akaike (AIC) est définit par la formule suivante,

$$AIC = -2\log(\hat{\beta}) + 2p$$

où $\hat{\beta}$ est le maximum de la fonction de vraisemblance du modèle et p le nombre de paramètres à estimer du modèle.

L'AIC représente un compromis entre le biais (qui diminue avec le nombre de paramètres) et la parcimonie (nécessité de décrire les données avec le plus petit nombre de paramètre possible). Ainsi, le retrait ou l'ajout de paramètres dans un modèle pourra être fait sur la base de l'AIC. Si l'ajout d'un parmètre augmente l'AIC c'est que le compromis n'est pas bon.

Exemple d'application : Considérons le jeu de données étudié en section 5.2, sur la propagation du SIDA. On a les deux modèles suivants avec $y_i \sim \mathcal{P}(\mu_i)$

- Le premier modèle : $\text{mod}_0 =$ `glm(y~ t, family = poisson)`
  où $g(\mu_i) = \log(\mu_i) = \log(c) + bx_i = \beta_0 + x_i\beta_1$.

  Avec la commande `AIC(mod0)`, on obtient un critère AIC d'une valeur de 166,4 qui est assez élevée.

- Le second modèle : $\text{mod}_1 =$ `glm(y~ t+I(t^2), family = poisson)`
  où $g(\mu_i) = \log(\mu_i) = \log(c) + bx_i + bx_i^2 = \beta_0 + x_i\beta_1 + x_i^2\beta_2$.

  On obtient un critère AIC de l'ordre de 96,92 qui est une valeur beaucoup plus petite que la valeur du modèle 0.

En résumé, quand on étudie un jeu de données avec des modèles de types GLM, on essaie de choisir le modèle avec le critère AIC le plus faible.

3.2 Tests statistiques

3.2.1 Test entre modèles emboîtés

Le but de cette partie est de déterminer quel est le meilleur modèle parmi p modèles. On note $\text{mod}_1, \ldots, \text{mod}_p$ les p différents modèles.

Soit deux modèles mod_0 et mod_1. On va supposer que mod_0 est emboîté dans mod_1 et nous considérons l'exemple ci-dessous.

On pose,

$$\begin{aligned} \text{mod}_0 &: g(\mu) = \beta_1 x_1 + \beta_2 x_2 \\ \text{mod}_1 &: g(\mu) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \end{aligned}$$

Pour tester quel modèle est meilleur que l'autre par rapport à un certain critère serait de poser,

$$H_0 : \beta_3 = \beta_4 = 0$$
$$\text{contre}$$
$$H_1 : \beta_3 \neq \beta_4 \neq 0$$

En réalité cela revient à tester la nullité des coefficients présent dans un modèle et non dans l'autre. On met alors en place un test statitisque sous une loi du $\chi^2_{p_2-p_1}$.

3.2.2 Anova sous R

Pour confronter deux modèles on peut faire une Anova, le principe est de déterminer à l'aide d'un test d'hypothèse le modèle le plus adapté. Cela n'a de sens statistique que si les modèles sont emboîtés comme vu dans la sous-section précédente. Le test effectué est celui décrit en section 3.2.1 il considère pour hypothèse nulle que tous les coefficients en plus d'un modèle par rapport à l'autre sont égaux à 0, pour hypothèse concurrente que les coefficients du second modèle sont non nuls. Dans cet exemple, on effectue un test du χ^2 car le paramètre de dispersion ϕ est connu pour ce modèle (loi de Poisson).

Sous R on utilisera la commande `anova()` qui prend en argument les modèles à confronter et en dernier argument le test à considérer. Lorsque l'on spécifie plus d'un modèle dans la fonction `anova()`, le tableau en sortie contient une ligne pour les degrés de liberté et de déviance résiduels pour chaque modèle. Ce tableau contiendra aussi les résultats des tests statistiques comparant la réduction de la déviance par rapport aux résidus.

Pour les modèles avec dispersion connue (par exemple, loi binomiale, de poisson), le test du khi-deux est le plus approprié. Le test de Fisher est utilisé dans un cas particulier des GLM, lorsque la variable Y suit une loi normale et où la fonction de lien sera l'identité, on retombera alors sur le modèle linéaire de base.

Exemple d'application : Cas de la propagation du Sida :

La commande : `anova(mod0,mod1,test="Chisq")` compare le modèle 0 au modèle 1.

```
Model 1: y ~ t
Model 2: y ~ t + I(t^2)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        11     80.686
2        10      9.240  1   71.446 < 2.2e-16
```

On choisira le modèle avec le moins de résidus de déviance. Ici, le modèle 1 est réellement meilleur en comparaison du modèle 0. La p-valeur du test est extrêmement petite donc le test est très significatif. Cela signifie que les paramètres supplémentaire du modèle 1 ne sont clairement pas nuls.

Régression logistique

4 Introduction à la régression logistique

Lorsque la variable à expliquer Y est une variable binaire (Succès/Échec, Homme/-Femme, ...), il paraît évident que les erreurs ne peuvent pas suivre une loi normale de moyenne nulle et de variance constante. On utilisera alors un cas particuler de GLM appelée la régression logistique.

Ce type de données suit une distribution Binomiale, de paramètres n et p. Dans ce cas, la fonction de lien canonique est appelée logit et définie par,

$$g(\mu_i) = \text{logit} = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$$

où $g^{-1}(x) = \dfrac{1}{1+e^x}$, souvent appelée fonction sigmoide, est bijective et définie sur $[0,1]$.

Alors, on aura,

$$\mu_i = \frac{1}{1+e^{X_i\beta}}$$

5 Expériences numériques sous R

5.1 Prédiction d'une attaque cardiaque

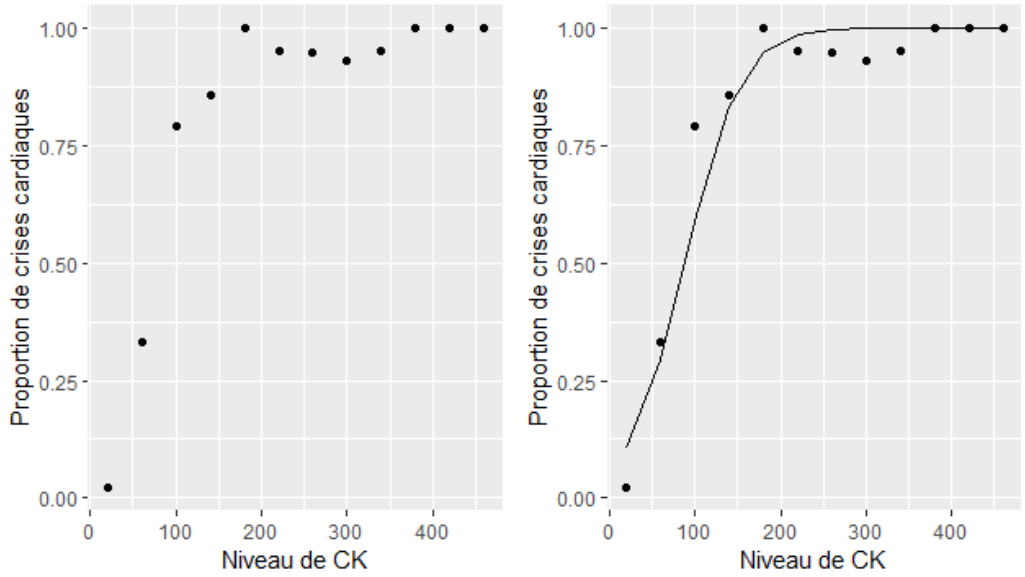
Une aide suggérée au diagnostic précoce de la crise cardiaque est le niveau de l'enzyme créatinine kinase (CK) dans la circulation sanguine. On s'intéresse donc à des données sur les attaques cardiaques des patients en fonction de leur niveau de CK.

v_CK	v_HA	v_OK
20	2	88
60	13	26
100	30	8
140	30	5
180	21	0
220	19	1
260	18	1
300	13	1
340	19	1
380	15	0
420	7	0
460	8	0

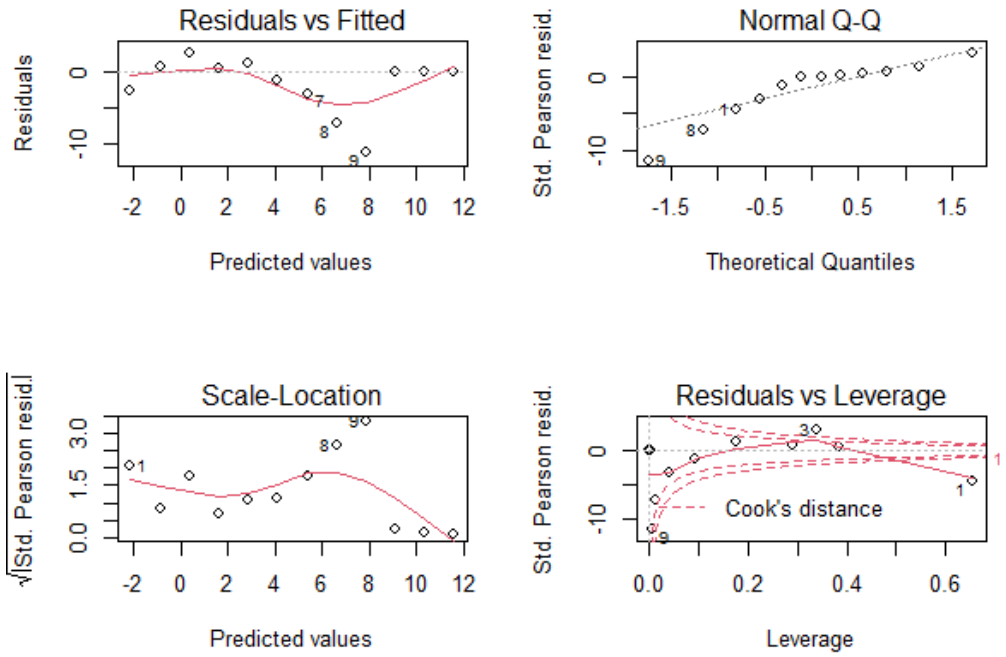
On contruit alors le GLM suivant, basé sur la proportion de patients diagnostiqués avec attaque cardiaque :

```
pCC <- Crise_Card$v_HA/(Crise_Card$v_HA+Crise_Card$v_OK)
modCC.0 <- glm(cbind(v_HA, v_OK)~v_CK, family = binomial(link = "logit"),
               data=Crise_Card)
```

```
Degrees of Freedom: 11 Total (i.e. Null);  10 Residual
Null Deviance:      271.7
Residual Deviance: 36.93    AIC: 62.33
```



On trace alors les graphiques associés au modèle 0.



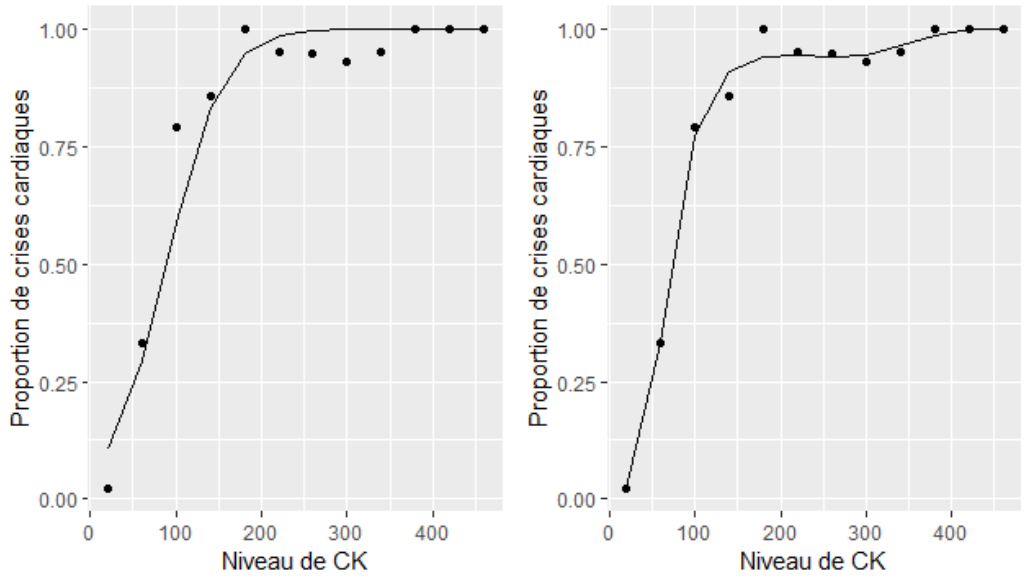
Puisque la réponse n'est pas supposée suivre une distribution normale, nous ne nous intéressons pas vraiment au diagramme quantile-quantile (`Normal Q-Q`). Le graphique `Residuals vs Fitted` permet de vérifier l'absence de tendance dans les résidus et le graphique `Residuals vs Leverage` permet de détecter des points avec un gros impact sur la régression. Notons que trois des graphiques utilisent les résidus de Pearson.

Pour mieux comprendre l'effet non-linéaire des prédicteurs, nous pouvons visualiser les prédictions du modèle pour différentes combinaisons de la valeur de CK. Ici on propose d'essayer un prédicteur linéaire cubique.

```
modCC.1 <- glm(cbind(v_HA, v_OK)~v_CK+I(v_CK^2)+I(v_CK^3),
               family = binomial(link = "logit"),
               data=Crise_Card)
```

```
Degrees of Freedom: 11 Total (i.e. Null);  8 Residual
Null Deviance:      271.7
Residual Deviance: 4.252    AIC: 33.66
```

On a nettement diminuer la déviance et l'AIC, on obtient l'ajustement suivant,



Le modèle 0 est représenté à gauche et à droite le modèle 1.

A propos du pseudo-R^2, pour le modèle 0 on avait, $R_0^2 = \dfrac{271.7 - 36.93}{271.7} = 0.864$

Pour le modèle 1 on a, $R_1^2 = \dfrac{271.7 - 4.252}{271.7} = 0.985$

Ce qui signifie qu'avec le modèle 1, plus de 98% de la variable réponse est expliquée.

Enfin les résultats de l'anova entre les deux modèles sont les suivants :

```
anova(modCC.0, modCC.1, test="Chisq")
```

```
Model 1: cbind(v_HA, v_OK) ~ v_CK
Model 2: cbind(v_HA, v_OK) ~ v_CK + I(v_CK^2) + I(v_CK^3)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        10     36.929
2         8      4.252  2   32.676 8.025e-08
```

Ce qui confirme que le modèle 1 est nettement meilleur par rapport au modèle 0.

5.2 Propagation du Sida

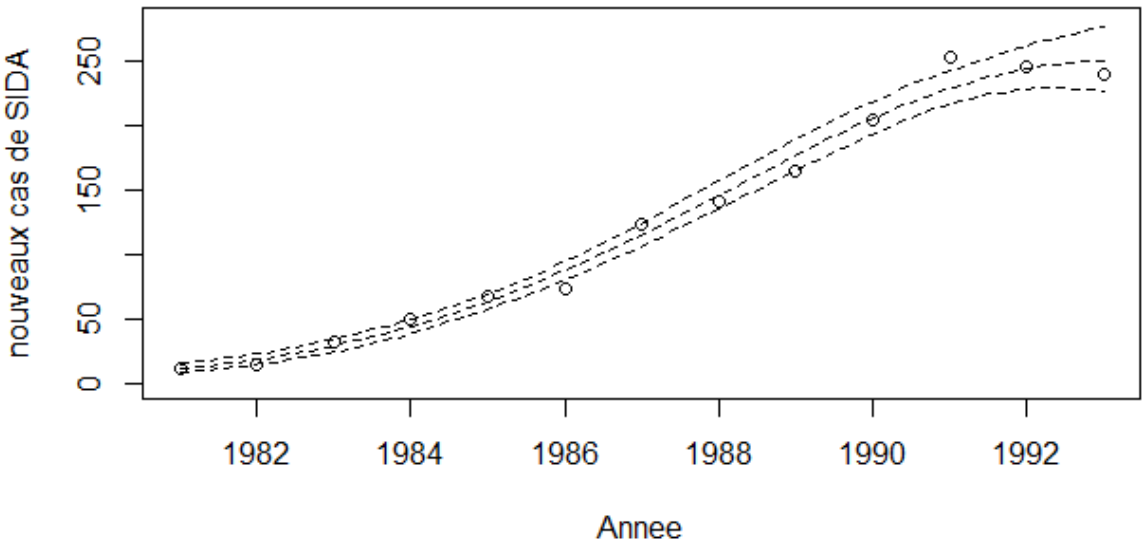
On travaille avec le jeu de données suivant, qui décrit le nombre de cas de sida en fonction de l'année.

Annee	Nb_cas
1981	12
1982	14
1983	33
1984	50
1985	67
1986	74
1987	123
1988	141
1989	165
1990	204
1991	253
1992	246
1993	240

On propose d'étudier les trois modèles suivants,

```
mS0 <- glm(Nb_cas~Annee, family = poisson)
mS1 <- glm(Nb_cas~Annee+I(Annee^2), family = poisson)
mS2 <- glm(Nb_cas~Annee+I(Annee^2)+I(Annee^3), family = poisson)
mS3 <- glm(Nb_cas~Annee+I(Annee^2)+I(Annee^3)+I(Annee^4),
           family = poisson)
```

Le modèle 1 `mS1` ayant le plus petit AIC, est retenu après un test d'Anova.
On peut alors tracer un intervalle de confiance sur l'estimation des coefficients en utilisant la commande `predict()`, on obtient le graphique suivant,

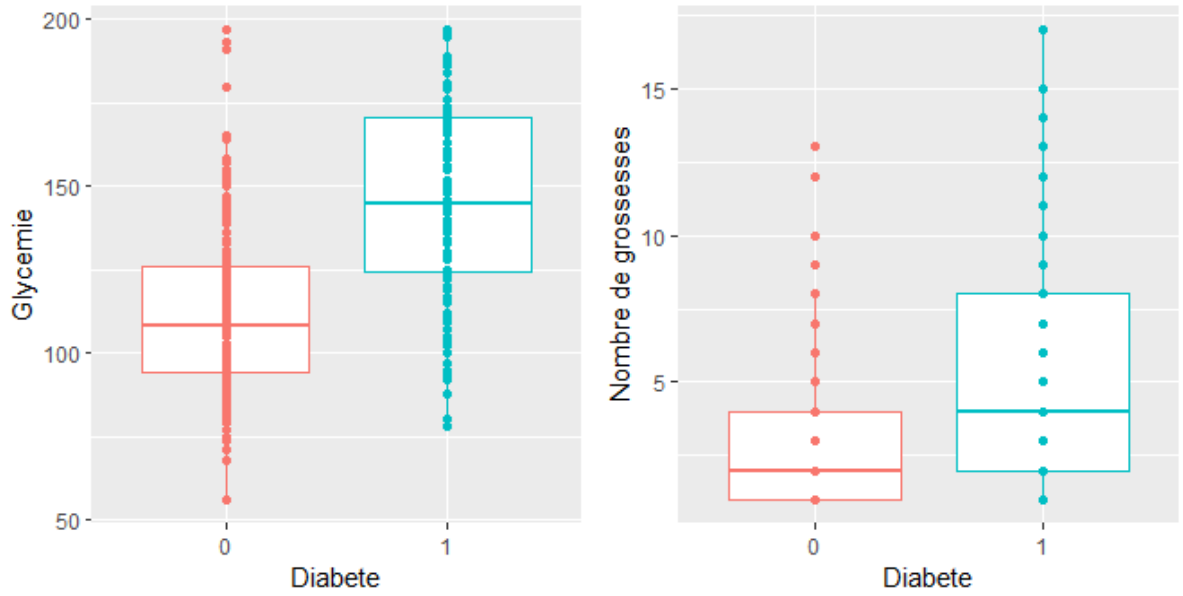


5.3 Prédiction du diabète

Le jeu de données `Diabete` est composé de 336 observations, dont 8 variables explicatives $X_1, \ldots, X_8$ et une variable binaire à expliquer Y décrivant si le patient est atteint de diabète ou non.

Contrairement au jeu de données précédents, nous avons des mesures précises pour différentes variables sur chaque individu. Ici, il faut d'abord s'intérresser aux variables qui permettront au mieux d'expliquer le diabète.

On trace les boxplots suivants :



On choisit alors de construire un modèle sur la base du taux de glycémie malgré les quelques valeurs atypiques, pour cela on créé un nouveau data de la façon suivante :

GlycemieMax	PatientsDiab	PatientsNonDiab
60	0	1
70	0	3
80	1	9
90	2	34
100	4	35
110	10	39
120	6	25
130	16	36
140	9	11
150	12	13
160	13	13
170	12	2
180	10	0
190	12	1
200	4	3

Où le nombre de patients diabétiques ayant un taux de glycémie entre 90 et 100 est de 4 et 35 patients avec un taux de cet ordre sont non-diabétiques.

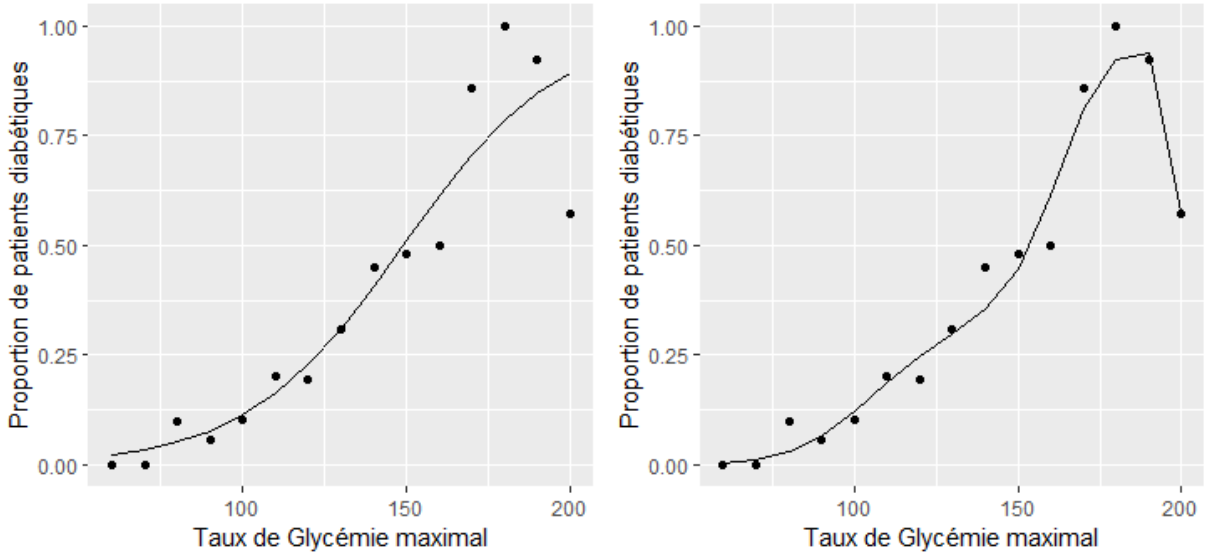
On considère alors deux modèles linéaires généralisés,

```
mod.0<-glm(cbind(PatientsDiab,PatientsNonDiab)~GlycemieMax,
           family = binomial(link = "logit"), data=DiabeteGlyc)

mod.1<-glm(cbind(PatientsDiab,PatientsNonDiab)~GlycemieMax +
           I(GlycemieMax^7) + I(GlycemieMax^8) + I(GlycemieMax^10),
           family = binomial(link="logit"),data=DiabeteGlyc)
```

Le pseudo-R^2 pour le modèle 1 vaut 0.943, ce qui est très bon. L'AIC du modèle 1 est légèrement plus faible.

Graphiquement cela nous donne,



Le modèle 0 est représenté à gauche, le modèle 1 à droite.
On réalise alors une Anova entre les deux modèles et on obtient,

```
anova(mod.0,mod.1,test="Chisq")
```

```
Analysis of Deviance Table

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        13    15.4297
2        10     6.1369  3   9.2928  0.02564
```

La p-valeur du test d'anova n'est pas aussi petite que pour les autres modèles, mais elle reste raisonnable.

On peut observer un comportement atypique des patients non-diabétiques qui ont cependant un très fort taux de glycémie > 190, en particulier on observe que le modèle 1 prédit ce comportement ce qui n'est pas forcément une bonne chose. La prédiction du modèle reste relativement bonne.

Conclusion

L'essentiel à retenir c'est que lorsque la variable réponse Y, i.e. la variable que l'on cherche à expliquer est de type binaire, ou encore de comptage il est nécessaire d'utiliser un GLM et non un modèle linéaire classique.

Résumé :

- Un modèle linéaire généralisé est de la forme $g(\mu_i) = X_i\beta$, il modélise une fonction de l'espérance de la variables Y. Il se compose donc d'une fonction de lien g pour la valeur moyenne à expliquer, et d'une distribution de la valeur à expliquer en fonction de sa moyenne.
- Un cas particulier de GLM est la régression logistique, qui sert à modéliser des réponses binaires (0 ou 1) ou binomiales. Elle utilise en général une fonction de lien canonique appelé logit et une distribution binomiale pour Y.
- La fonction logit transforme une probabilité $p \in [0, 1]$ en un réel. Un logit négatif correspond à une probabilité $p \in [0, \frac{1}{2}]$, un logit positif correspond à une probabilité $p \in [\frac{1}{2}, 1]$.
- Lorsqu'on construit un GLM sur la base d'une régression logistique, l'effet de la combinaison des $X_i\beta$ sur Y n'est pas linéaire et dépend de la valeur de chaque combinaison. Il est donc intéressant de visualiser les prédictions du modèle pour différentes combinaisons des variables.
- Lorsqu'on construit un GLM sous R, il y a trois critères important à prendre en compte pour vérifier la cohérence de notre modèle : l'AIC, la déviance et le pseudo-R^2.
- Enfin, dans le cas où plusieurs modèles sont construits, imbriqués les uns dans les autres, c'est l'anova, avec un test du khi-deux dans le cas d'une régression logistique, qui nous permettra d'orienter notre choix.