# Régression logistique

Introduction à R pour la recherche biomédicale

http://www.aliquote.org/cours/2012_biomed

# Objectifs

- Dans ce cours, on aborde la régression logistique (simple et multiple) qui sert à modéliser la relation entre des prédicteurs catégoriels ou continus et une variable réponse binaire.
- Sans rentrer dans les détails des modèles linéaires généralisés, on insistera sur la correspondance entre la régression logistique et la régression linéaire sur le plan de la modélisation, tout en soulignant les différences les plus notables : interprétation des coefficients, qualité d'ajustement du modèle, analyse des résidus.

Lectures conseillées : Vittinghoff, Glidden, Shiboski, & McCulloch (2005), Hosmer & Lemeshow (1989).

# Régression logistique

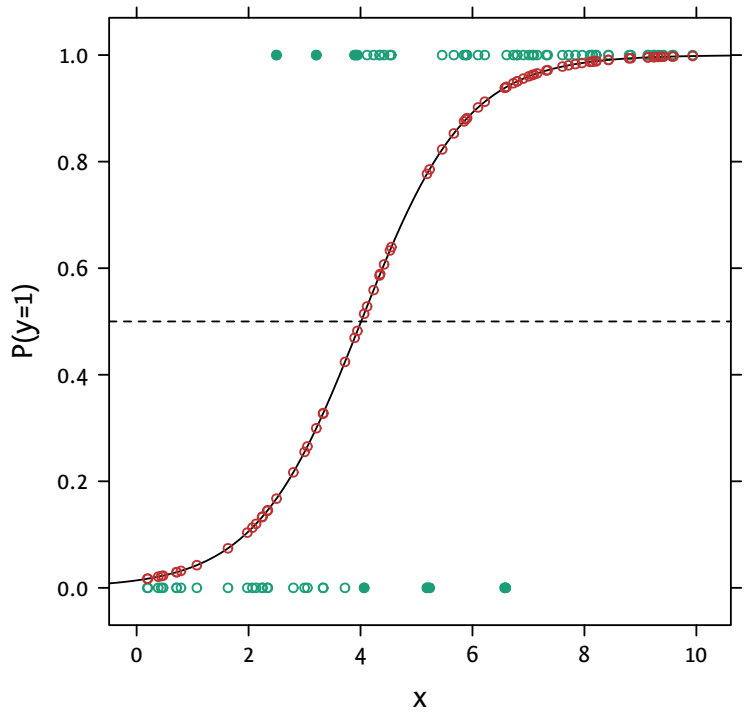
La régression logistique permet de traiter le cas où la variable réponse est de type binaire (oui/non, malade/pas malade, etc.), et non pas continu comme dans le modèle de régression linéaire. Tout en relaxant certaines des hypothèses du modèle de régression multiple, on maintient quand même l'idée d'une relation linéaire entre la réponse et les prédicteurs.

Dans le cas d'une variable binaire, sa "moyenne" correspond à la proportion d'individus possédant la caractéristique étudiée ou répondant positivement à l'événement, d'où l'idée de modéliser la probabilité de succès, comprise entre 0 et 1, en fonction d'un certain nombre de prédicteurs.

Dans les enquêtes épidémiologiques cas-témoin (avec ou sans matching) où l'incidence de la maladie n'est pas connue, la régression logistique fournit une estimation de l'odds-ratio ajusté sur les co-facteurs d'intérêt (âge, sexe, etc.). D'autre part, lorsque la prévalence de la maladie est faible, l'OR fournit une bonne approximation du risque relatif.

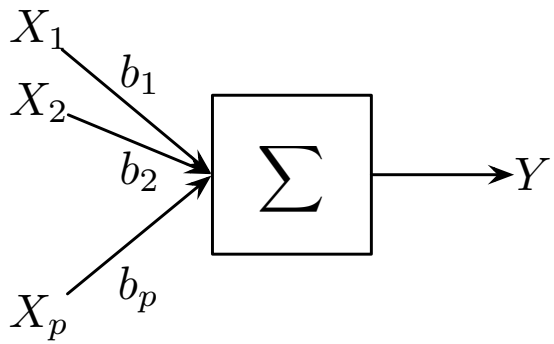
# Illustration

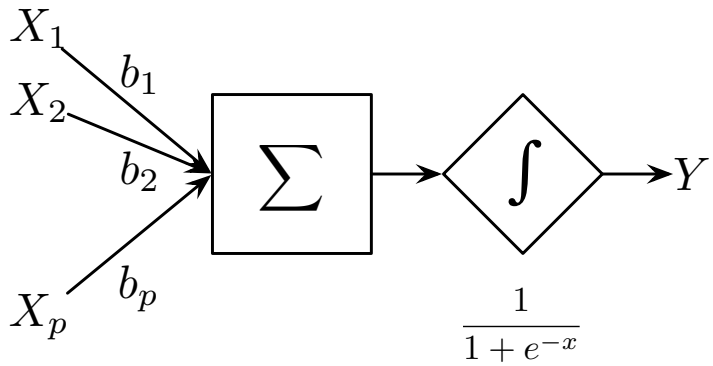
L'exemple ci-dessous montre les réponses observées (0/1) et les probabilités individuelles correspondantes, telles qu'estimées à partir d'une régression logistique. En considérant $\tilde{y}_i = \mathbb{I}(P(x) \geq 0.5)$, on dénombre 8 mal classés (10 %).



# Parallèle avec la régression linéaire

Comme dans la régression linéaire, on cherche la meilleure combinaison linéaire des données d'entrée pour modéliser la réponse, à ceci près que c'est une transformation de cette combinaison (on parle d'une fonction de lien) qui est utilisée en sortie.



# En détails

## Le modèle de régression logistique

Si l'on note $\pi$ la probabilité d'observer l'événement $y = 1$ (vs. 0), alors le log odds (transformation *logit*) peut s'exprimer comme une fonction linéaire des paramètres du modèle à $p$ prédicteurs :

$$g(x) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

et la probabilité prédite s'écrit alors

$$P(y = 1 \mid x_1, x_2, \ldots, x_p) = \hat{y}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)}.$$

Dans ce type de modèle, on fait l'hypothèse que $y_i$ suit une distribution binomiale et que les observations sont indépendantes (aucune hypothèse sur la variance qui n'est pas un paramètre dans ce cas de figure). Notons également l'absence de terme d'erreur. L'estimation d'un tel type de modèle se fait par la méthode du maximum de vraisemblance. D'autres fonctions de lien existent (probit, log-log).

# Différences avec le modèle linéaire

L'analyse des résidus du modèle permet de vérifier si celui-ci est satisfaisant (en terme de spécification et de qualité d'ajustement).
Les différences principales avec la régression linéaire sont les suivantes :

- On ne parle plus de sommes de carrés (OLS, résidus, variance) mais de déviance (dans le cas gaussien, elle est équivalente à la somme de carrés de la résiduelle), mais cette dernière reflète toujours l'écart entre les données et le modèle.
- En raison de la nature binaire de la variable réponse, l'analyse classique des résidus en fonction des valeurs prédites ou la notion d'hétéroskédasticité ne font plus sens ; en revanche, on s'intéresse toujours à la qualité d'ajustement du modèle, et à la comparaison de modèles emboîtés qui permettent d'évaluer l'apport d'un ou plusieurs prédicteurs par rapport à un modèle de base. Pour cela, on utilise des tests de rapport de vraisemblance.

# Interprétation des coefficients

Le terme d'intercept s'interprète comme un odds, et les coefficients de régression comme des odds-ratio : lorsque $X_j$ augmente de $d = 1$ unité, l'odds de $y = 1$ augmente de $\exp(\beta_j d)$ (de manière équivalente, le log-odds augmente de $\beta_j d$). Dans le cas où l'on a un seul prédicteur, binaire, on peut vérifier à partir de la relation $\frac{P(x)}{1-P(x)} = \exp(\beta_0 + \beta_1 x)$ que $\frac{P(1)/[1-P(1)]}{P(0)/[1-P(0)]} = \exp(\beta_1)$.

```
glm(formula = chd69 ~ age, family = binomial, data = wcgs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6209  -0.4545  -0.3669  -0.3292   2.4835

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.93952    0.54932 -10.813  < 2e-16 ***
age          0.07442    0.01130   6.585 4.56e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\log\left[\frac{P(56)}{1-P(56)}\right] - \log\left[\frac{P(55)}{1-P(55)}\right] = (-5.940 + 0.074 \times 56) - (-5.940 + 0.074 \times 55) = 0.074.$$

D'où un odds-ratio de $\exp(0.074) = 1.077$ associé à une augmentation d'âge d'un an sur le risque d'infarctus, c'est-à-dire une augmentation du risque de 8 %. Pour une variation de 10 ans, l'OR est de $\exp(0.074 \times 10) = 2.105$.

# Données groupées vs. non groupées

Dans le cas où on dispose des données individuelles, soit n réponses binaires codées sous forme numérique 0/1 ou à l'aide d'un facteur dont le premier niveau code l'échec (0), le modèle de régression logistique s'écrit :

```
glm(y ~ x, family=binomial)
```

Dans le cas des données "groupées" (aggrégées par niveaux d'un ou plusieurs facteurs), on utilise une matrice à deux colonnes représentant les effectifs 0/1 (n0=n-n1) :

```
glm(cbind(n1, n0) ~ x, family=binomial)
```

ou directement la fréquence empirique et le nombre d'observations :

```
glm(n1/n ~ x, family=binomial, weights=n)
```

On peut toujours passer d'un format à l'autre avec `aggregate` ou `untable` (package `reshape`).

# Qualité d'ajustement du modèle et diagnostics

La déviance et le $\chi^2$ généralisé de Pearson sont les deux outils principaux pour l'évaluation et la comparaison de modèles logistiques. La déviance se définit comme deux fois la différence entre les log vraisemblances de deux modèles, tandis que $\chi^2 = \sum_i (y_i - \hat{y}_i)^2 / V(\hat{y})$. (On peut aussi définir des résidus basés sur la déviance.)

Le test de Hosmer-Lemeshow consiste à évaluer la concordance entre les valeurs prédites et observées des observations regroupées en quantiles, typiquement des déciles. Ce test dépend du nombre de groupes fixés *a priori*, et il est peu puissant en cas de mauvaise spécification. Les techniques de lissage non-paramétrique ou d'estimations stratifiées sont utiles pour identifier des déviations locales ou globales (Harrell, 2001).

Il existe des mesures de type pseudo-$R^2$ (Cox-Snell, Nagelkerke) qui permettent d'évaluer la qualité du modèle global (par rapport au modèle nul). Concernant la capacité prédictive du modèle, on utilise généralement le score de Brier ou l'index $C$ (probabilité de concordance), en lien avec la courbe ROC.

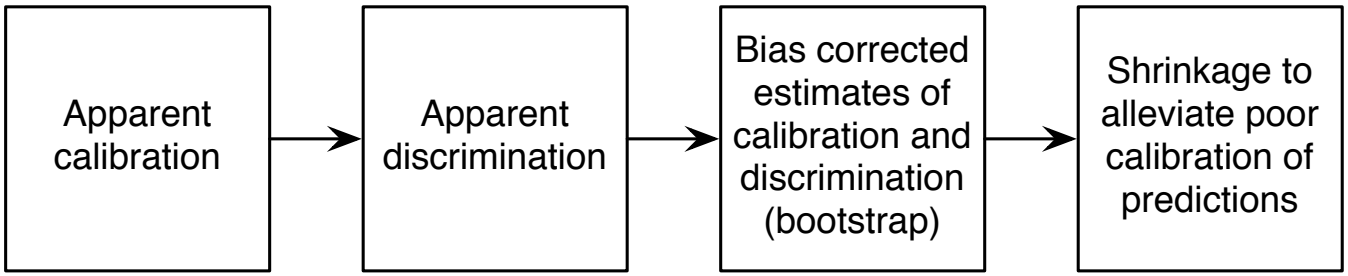
# Qualité d'ajustement du modèle et diagnostics (2)

Comme dans le cas du modèle linéaire, on s'intéressera aux points influents et aux "outliers", tous deux dérivés à partir des résidus standardisés de Pearson. Le problème est que dans ce cas, ces résidus appartiennent à deux classes ce qui rend délicat, voire inutile, les représentations graphiques de type résidus *vs.* valeurs prédites. On se concentrera donc plutôt sur les mesures d'influence, comme les DFBETAS. On peut représenter ces dernières en fonction des probabilités prédites. Voir aussi `residual.plots` (package `alr3`). Il est également possible d'examiner les variations dans le $\chi^2$ (ou la déviance) en fonction du leverage ou des probabilités prédites, voir par exemple `http://bit.ly/JMtaho`.

Concernant la linéarité de la relation entre les prédicteurs et le log-odds, on peut utiliser de simple diagramme de dispersion, ou utiliser ce que l'on appelle les "marginal model plot" (par ailleurs plus utile dans le cas des données non groupées), et on peut juger de la linéarité de la relation à partir d'une courbe lowess. Voir `mmps` (package `car`) ainsi que le package `marginalmodelplots`.

# Sélection de variables, modèles prédictifs

La construction et la validation de modèles prédictifs dans le domaine clinique ont fait l'objet de nombreux ouvrages, en particulier Harrell (2001) et Steyerberg (2009) (voir aussi Steyerberg et al. (2001)). Le site RMS fournit de nombreuses ressources sur ce sujet : `http://biostat.mc.vanderbilt.edu/wiki/Main/RmS`




**Internal validation of predictive models**

Pour un tour d'horizon rapide des enjeux de la modélisation à partir d'enquêtes épidémiologiques, voir Greenberg & Kleinbaum (1985). De nombreux autres articles fournissent des recommendations pour le reporting des résultats (Bagley, White, & Golomb, 2001; Bouwmeester et al., 2012; K. J. Ottenbacher, Ottenbacher, Tooth, & Ostir, 2004).

# Application 1

Heart disease and blood pressure. (Everitt & Rabe-Hesketh, 2001, p. 208)

```
bp <- read.table("hdis.dat", header=TRUE)
blab <- c("<117","117-126","127-136","137-146",
          "147-156","157-166","167-186",">186")
clab <- c("<200","200-209","210-219","220-244",
          "245-259","260-284",">284")
bp <- within(bp, {
  bpress <- factor(bpress, labels=blab)
  chol <- factor(chol, labels=clab)
})
round(xtabs(hdis/total ~ bpress + chol, data=bp), 2)
```

Ici, les données sont groupées et on cherche à modéliser la probabilité de survenue d'une crise cardiaque en fonction de la tension artérielle, que l'on peut considérer comme une variable ordinale, ou numérique en considérant les valeurs centrales des catégories.

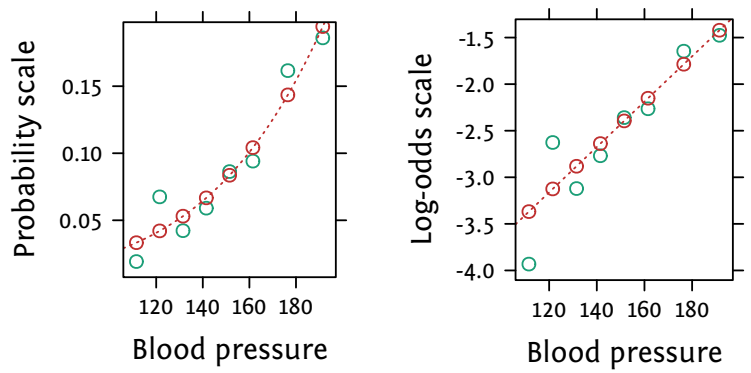
# Résultats

```
midpoint <- function(x) {
  x <- as.numeric(unlist(strsplit(x, "-")))
  return(sum(x)/2)
}
val <- sapply(levels(bp$bpress)[-c(1,8)], midpoint)
dfrm <- aggregate(bp[,3:4], list(bpress=bp[,1]), sum)
dfrm$bpress <- c(val[1]-10, val, val[6]+15)
mod1 <- glm(cbind(hdis, total-hdis) ~ bpress,
            data=dfrm, family=binomial)
summary(mod1)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.082033   0.724320  -8.397  < 2e-16 ***
bpress       0.024338   0.004843   5.025 5.03e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 30.0226  on 7  degrees of freedom
Residual deviance:  5.9092  on 6  degrees of freedom
AIC: 42.61
```

# Illustration

Ci-dessous sont représentées les valeurs observées et prédites, sous forme de proportion ou de logit. Dans le deuxième cas (droite), on voit clairement la linéarité entre le log odds, $\hat{g}(x)$, et $x$.




16/271 = 0.059

log(0.059/(1-0.059)) = -2.769

1/(1+exp(-(-6.082+0.024*141.5))) = 0.067

| | bpress | hdis | total | prop | pred | ologit | plogit |
|---|---|---|---|---|---|---|---|
| 1 | 111.5 | 3 | 156 | 0.01923077 | 0.03330037 | -3.931826 | -3.368319 |
| 2 | 121.5 | 17 | 252 | 0.06746032 | 0.04209028 | -2.626372 | -3.124937 |
| 3 | 131.5 | 12 | 284 | 0.04225352 | 0.05307297 | -3.120895 | -2.881554 |
| 4 | 141.5 | 16 | 271 | 0.05904059 | 0.06672179 | -2.768675 | -2.638172 |
| 5 | 151.5 | 12 | 139 | 0.08633094 | 0.08357090 | -2.359280 | -2.394789 |
| 6 | 161.5 | 8 | 85 | 0.09411765 | 0.10419982 | -2.264364 | -2.151407 |
| 7 | 176.5 | 16 | 99 | 0.16161616 | 0.14352287 | -1.646252 | -1.786333 |
| 8 | 191.5 | 8 | 43 | 0.18604651 | 0.19446420 | -1.475907 | -1.421260 |

# Application 2

The low birth weight study. (Hosmer & Lemeshow, 1989)

```
data(birthwt, package="MASS")
ethn <- c("White","Black","Other")
birthwt$race <- factor(birthwt$race, labels=ethn)
fm <- low ~ age + lwt + race + ftv
glm1 <- glm(fm, data=birthwt, family=binomial)
summary(glm1)
confint(glm1)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.295366   1.071443   1.209   0.2267
age         -0.023823   0.033730  -0.706   0.4800
lwt         -0.014245   0.006541  -2.178   0.0294 *
raceBlack    1.003898   0.497859   2.016   0.0438 *
raceOther    0.433108   0.362240   1.196   0.2318
ftv         -0.049308   0.167239  -0.295   0.7681
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

L’âge de la mère ne semble pas être un prédicteur intéressant, mais son poids oui, et on retrouve des différences entre les mères White vs. Black.

# Interprétation des coefficients

Le logit estimé s'écrit (Hosmer & Lemeshow, 1989 p. 36)

$$\begin{aligned}\hat{g}(x) = \quad & 1.295 - 0.024 \times \text{age} - 0.014 \times \text{lwt} \\ & + 1.004 \times \mathbb{I}(\text{race} = \text{Black}) + 0.433 \times \mathbb{I}(\text{race} = \text{Other}) \\ & - 0.049 \times \text{ftv},\end{aligned}$$

ce qui permet d'obtenir les valeurs prédites ("probabilité d'un enfant avec un poids inférieur à 2,3 *kg* à la naissance") facilement. Pour un enfant dont la mère possède les caractéristiques moyennes ou modales de l'échantillon (âge 23,2 ans, poids 59 *kg*, `White`, aucune visite chez le gynécologue durant le premier trimestre de grossesse), on obtient :

$$\begin{aligned}P(\text{low} = 1) \quad & = 1/\big(1 + \exp(-(1.30 - 0.02 \times 23.24 - 0.01 \times 129.82))\big) \\ & = 0.253.\end{aligned}$$

```
log.odds <- predict(glm1, data.frame(age=mean(birthwt$age),
                                     lwt=mean(birthwt$lwt),
                                     race="White", ftv=0))
exp(log.odds)/(1+exp(log.odds))  # 0.2483566
```

# Significativité des termes du modèle

Les facteurs age et ftv apparaissent non significatifs. On peut *a priori* simplifier le modèle comme suit :

```
glm2 <- update(glm1, . ~ . - age - ftv)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.805753   0.845167   0.953   0.3404
lwt         -0.015223   0.006439  -2.364   0.0181 *
raceBlack    1.081066   0.488052   2.215   0.0268 *
raceOther    0.480603   0.356674   1.347   0.1778
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ces deux modèles sont bien emboîtés (glm2 opère juste une restriction dans l'espace des prédicteurs).

À partir de l'analyse des déviances résiduelles, on pourrait confirmer que ce modèle n'est pas "moins bon" que le précédent ($\chi^2(2) = 0.686$, $p = 0.710$) :

```
anova(glm2, glm1, test="Chisq")
```

# Significativité des termes du modèle (2)

En raison des multiples degrés de liberté, l'évaluation des tests de significativité (Wald) associés aux coefficients de régression des variables catégorielles est délicat.
Pour évaluer la significativité du facteur `race`, on peut procéder par comparaison de modèles également et le test du rapport de vraisemblance s'obtient comme suit :

```
anova(update(glm2, . ~ . - race), glm2)
```

D'où un test non significatif à 5 % :
(2 degrés de liberté)

```
pchisq(5.4316, 2, lower.tail=FALSE)
```

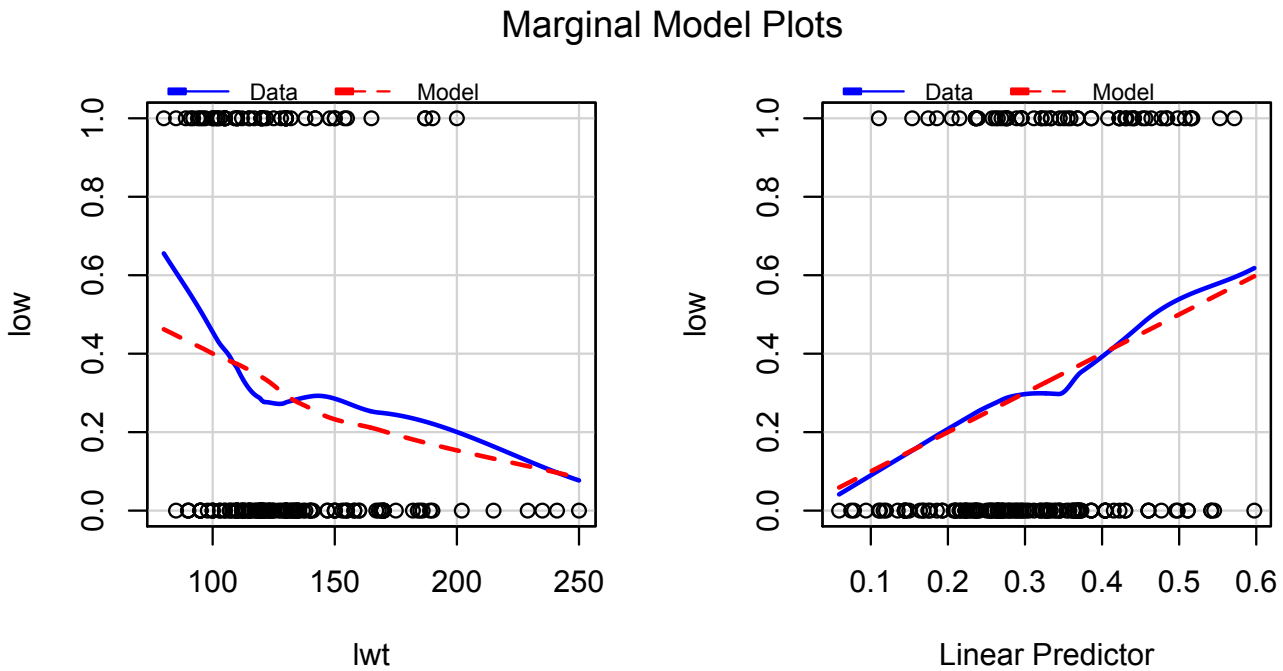
```
Analysis of Deviance Table

Model 1: low ~ lwt
Model 2: low ~ lwt + race
  Resid. Df Resid. Dev Df Deviance
1       187     228.69
2       185     223.26  2   5.4316
```

Toutefois, ce facteur étant connu pour être cliniquement important, on peut préférer le conserver dans le modèle final même si à l'évidence il ne change pas significativement la qualité explicative ou prédictive du modèle.

# Vérification des hypothèses du modèle

Pour la linéarité, on peut examiner les prédictions (probabilités) en fonction de $g(x)$, ou des variables continues.

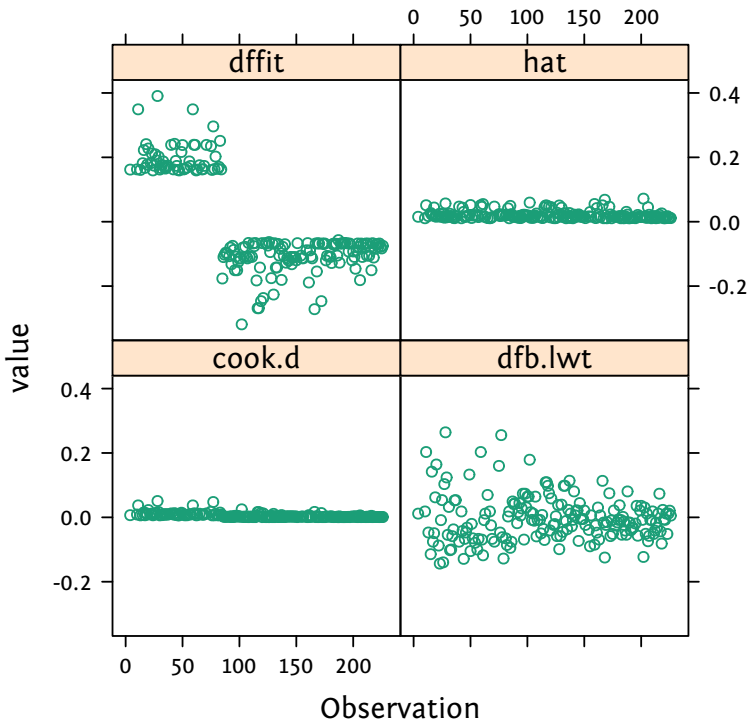
```
library(car)
mmps(glm2, terms=~lwt)
```



# Vérification des hypothèses du modèle (2)

Les résidus (Pearson ou déviance) peuvent être obtenus à partir de la fonction `rstandard`. Quant aux mesures d'influence, il s'agit du même principe que pour le modèle linéaire :

```
influence.measures(glm2)
```



# Fonctionnalités avancées

La fonction `lrm` du package `rms` (Harrell, 2001) est beaucoup plus informative que `glm`, en particulier en ce qui concerne la qualité d'ajustement du modèle et son pouvoir discriminant.

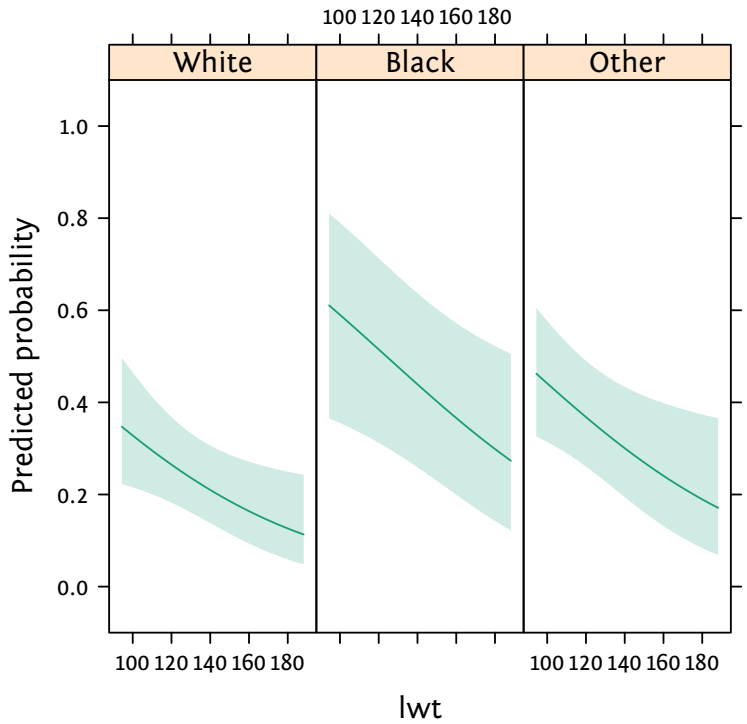
```
library(rms)
ddist <- datadist(birthwt)
options(datadist="ddist")
glm2b <- lrm(low ~ lwt + race, data=birthwt)
glm2b
exp(coef(glm2b))
```

```
                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes          Indexes
Obs           189    LR chi2     11.41    R2        0.082    C         0.647
 yes          130    d.f.            3    g         0.643    Dxy       0.293
 no            59    Pr(> chi2) 0.0097    gr        1.903    gamma     0.296
max |deriv| 2e-07                         gp        0.128    tau-a     0.127
                                          Brier     0.202

           Coef    S.E.   Wald Z Pr(>|Z|)
Intercept   0.8058 0.8452  0.95  0.3404
lwt        -0.0152 0.0064 -2.36  0.0181
race=Black  1.0811 0.4881  2.22  0.0268
race=Other  0.4806 0.3567  1.35  0.1778
```

# Prédictions

Le package `rms` fournit également tout un ensemble de fontions additionnelles pour les prédictions et leur représentation graphique (à l'image du package `effects`).

```
Predict(glm2b, lwt, race, fun=plogis, conf.type="mean")
```



# Régression logistique et diagnostique médical

Le modèle de régression logistique est idéal pour fournir des probabilités (au niveau individuel, ou en moyenne conditionnellement à certains co-facteurs). Dans un contexte de classification, on réduit ces prédictions à deux classes, 0/1, à partir d'un "cut-off" optimisé sur la base d'un compromis entre sensibilité et spécificité, sans prendre en considération un coût différentiel pour les faux-positifs et les faux-négatifs. D'autre part, transformer une probabilité (continue dans [0;1]) en une variable binaire, sans autoriser de zone d'incertitude, est risqué.

Pour plus d'informations sur la présentation des résultats d'un modèle de régression logistique dans les études diagnostiques, voir Steyerberg (2009).

> *There is a reason that the speedometer in your car doesn't just read "slow" and "fast". Frank Harrell on R-help, February 2011*

Voir aussi *Direct Measures of Diagnostic Utility Based on Diagnostic Risk Models*, `http://1.usa.gov/Mba7xK`.

# Comparaison avec d'autres méthodes (classification)

06-logistic-extras.r

Ci-dessous figurent quelques résultats concernant l'utilisation de différents modèles de classification et la comparaison de leur capacité prédictive. Les taux de classification correcte sont estimés par validation croisée (25 x 10 fold), avec le package `caret` (Kuhn, 2008).

Sur les données d'origine, le taux de classification correcte est de 68.8 % :

```
table(predict(glm1, type="resp")>=.5, birthwt$low)
```

```
Models: LR, RF, SVM, KNN
Number of resamples: 250

Accuracy
      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
LR  0.4737  0.6316 0.6842 0.6688  0.6842 0.7895
RF  0.3684  0.5789 0.6316 0.6401  0.6842 0.8421
SVM 0.6842  0.6842 0.6842 0.6880  0.6842 0.7222
KNN 0.4737  0.6316 0.6842 0.6605  0.7332 0.8889
```

Attention, il s'agit dans tous les cas d'un taux de classification "apparent" !

# Index

# Bibliographie


Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*, *54*, 979–985.

Bouwmeester, W., Zuithoff, N. P. A., Mallett, S., Geerlings, M. I., Vergouwe, Y., Steyerberg, E. W., Altman, D. G., et al. (2012). Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLoS Medicine*, *9*, 1001221.

Everitt, B., & Rabe-Hesketh, S. (2001). *Analyzing Medical Data Using S-PLUS*. Springer.

Greenberg, R. S., & Kleinbaum, D. G. (1985). Mathematical modeling strategies for the analysis of epidemiological research. *Annual Review of Public Health*, *6*, 223–245.

Harrell, F. E. (2001). *Regression modeling strategies with applications to linear models, logistic regression and survival analysis*. Springer.

Hosmer, D., & Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*.

Ottenbacher, K. J., Ottenbacher, H. R., Tooth, L., & Ostir, G. V. (2004). A review of two journals found that articles using multivariable logistic regression frequently did not report commonly recommended assumptions. *Journal of Clinical Epidemiology*, *57*, 1147–1152.

Steyerberg, E. W. (2009). *Clinical prediction models: a practical approach to development, validation, and updating*. Springer.

Steyerberg, E. W., Harrell, F. E., Borsboom, G. J. J. M., Eijkemans, M. J. C., Vergouwe, Y., & Habbema, J. D. F. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, *54*, 774–781.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch. (2005). *Regression Methods in Biostatistics. Linear, Logistic, Survival, and Repeated Measures Models*. Springer.