

Régression sur données non-linéaires

Certificat Data Analyst

6-7 février 2017

Martin Chevalier
Insee



1 / 109

Régression sur données non-linéaires

Objectifs du module

Dresser un **panorama raisonné** des méthodes de régression sur **données non-linéaires**.

Insister sur le modèle de **régression logistique dichotomique** et son **interprétation**.

Mettre la **mise en œuvre pratique avec R** au cœur du module : nombreux exemples dans le support, exercices corrigés.



2 / 109

Introduction : Modéliser des données non-linéaires



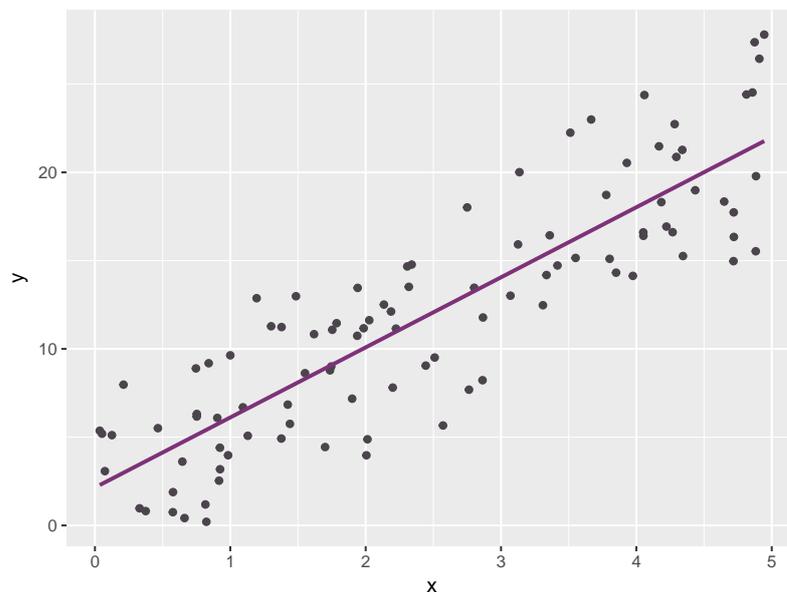
3 / 109

Introduction : Modéliser des données non-linéaires Régression linéaire classique

Variable expliquée Y Quantitative

Variable(s) explicative(s) X

- ▶ quantitatives ou qualitatives ;
- ▶ en relation linéaire avec la variable expliquée.

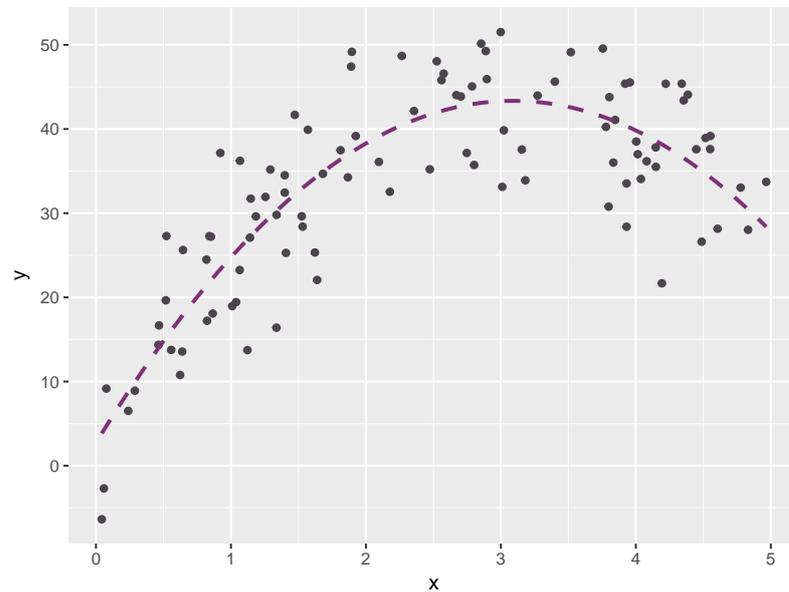


4 / 109

Introduction : Modéliser des données non-linéaires

Régression linéaire classique

Le modèle de régression linéaire classique peut capter **certaines relations non-linéaires**.



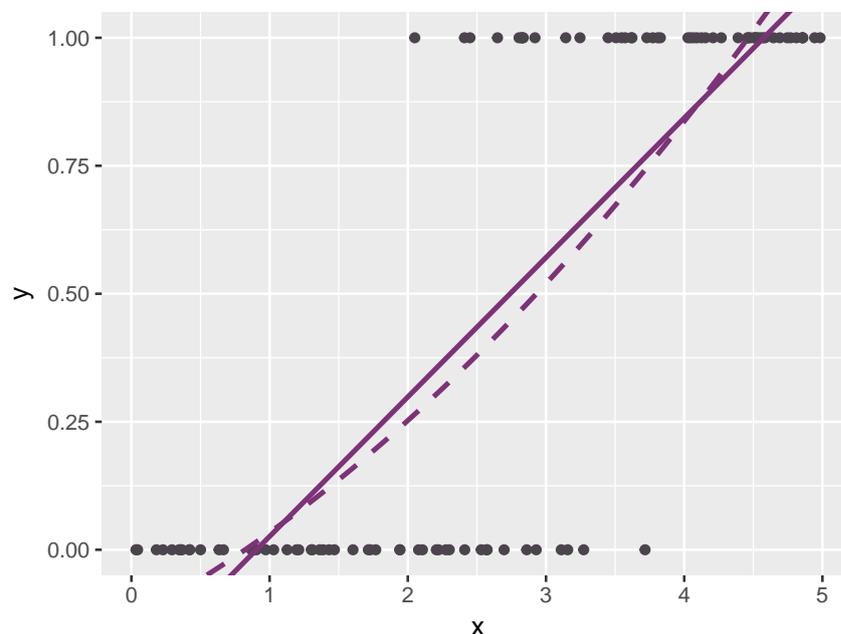
$$y_i = \beta_0 + \beta_1 \times x_i + \beta_2 \times x_i^2 + \varepsilon_i$$



Introduction : Modéliser des données non-linéaires

Limites de la régression linéaire classique

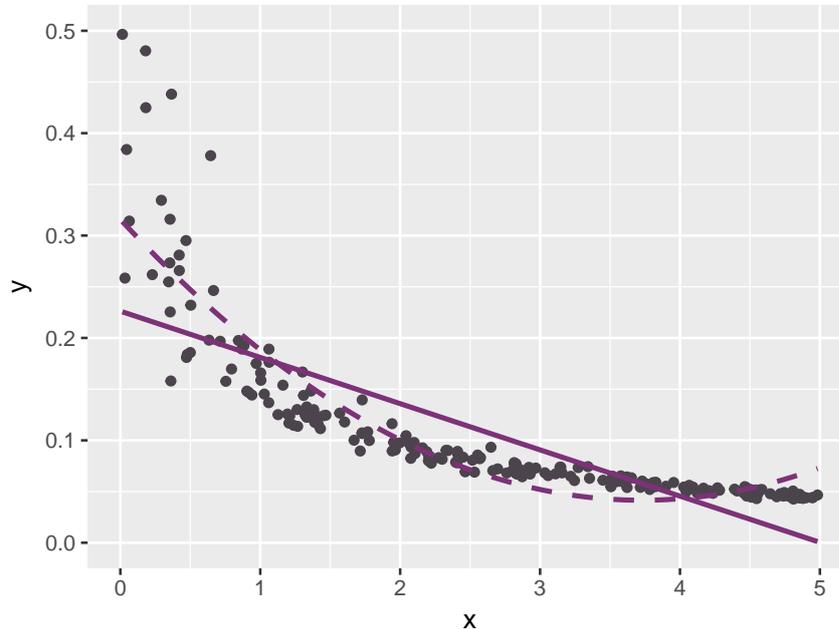
Dans certains cas cependant, la distribution de la variable dépendante Y semble **trop particulière** pour être modélisée avec la régression linéaire classique.



Introduction : Modéliser des données non-linéaires

Limites de la régression linéaire classique

Dans certains cas cependant, la distribution de la variable dépendante Y semble **trop particulière** pour être modélisée avec la régression linéaire classique.



6 / 109

Introduction : Modéliser des données non-linéaires

Solution : Généraliser le modèle linéaire

Pour modéliser des données particulièrement **non-linéaires**, on utilise le **modèle linéaire général** du type :

$$f(y_i) = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \varepsilon_i$$

Linéaire ?

- ▶ Ce modèle est « **linéaire en ses coefficients** » : les opérations entre les X et β sont uniquement des **sommes ou des multiplications**...
- ▶ ... mais il peut modéliser des relations *non-linéaires* grâce notamment à la **fonction de lien** f .

Exemples de fonctions de lien Identité, logarithme, inverse, **logit**, etc.

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

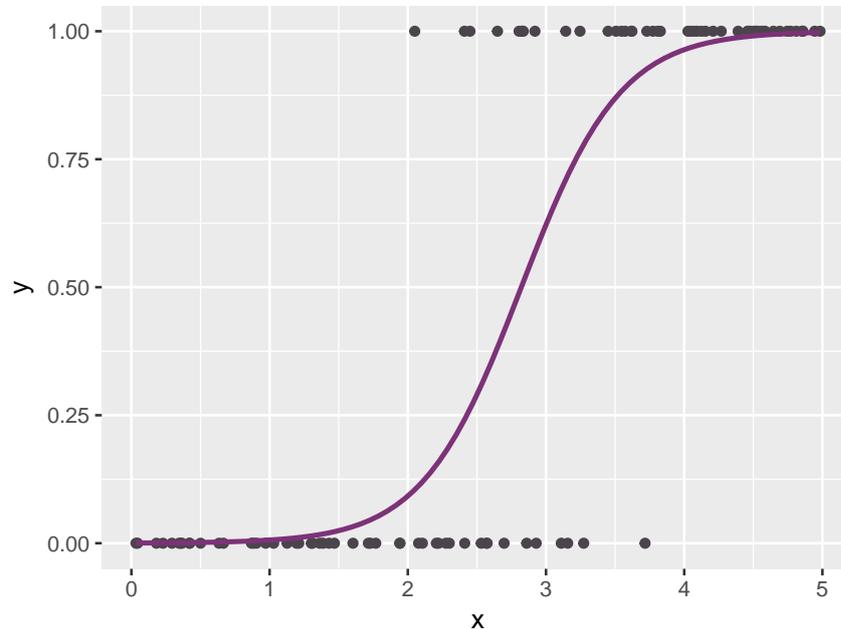


7 / 109

Introduction : Modéliser des données non-linéaires

Exemple : Régression logistique dichotomique

La régression logistique est la méthode la plus utilisée pour modéliser des **données dichotomiques** (deux modalités distinctes exactement).

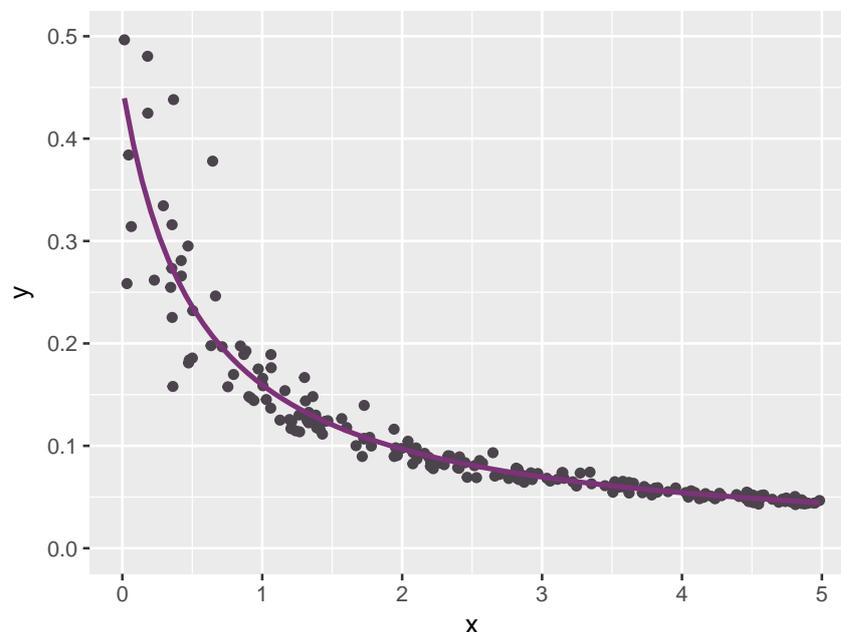


8 / 109

Introduction : Modéliser des données non-linéaires

Exemple : Régression gamma

La régression gamma permet de modéliser des variables présentant une distribution **très asymétrique** (actifs financiers par exemple).



9 / 109

Introduction : Modéliser des données non-linéaires

Démarche générale

Partie 1 Présenter le modèle linéaire général et son application à la régression logistique dichotomique.

Partie 2 Insister sur l'interprétation de la régression logistique dichotomique : indicateurs de qualité, interprétation des coefficients et tests.

Partie 3 Introduire d'autres spécifications du modèle linéaire général : modèles pour données polytomiques, modèles pour données asymétriques.



10 / 109

Introduction : Modéliser des données non-linéaires

Données utilisées pour les exemples

La plupart des exemples sont construits à partir de l'enquête **Emploi en continu** (EEC) de l'Insee :

- ▶ environ 100 000 personnes de 15 ans ou plus interrogées chaque trimestre ;
- ▶ questionnaire de 50 pages, mesure du chômage selon la définition du Bureau international du travail (BIT) ;
- ▶ fichier complet accessible jusqu'au millésime 2012.

De nombreuses variables issues de cette enquête gagnent à être modélisées avec le **modèle linéaire général** :

- ▶ avoir un emploi stable → **logit dichotomique**
- ▶ être au chômage ou inactif plutôt qu'en emploi → **logit polytomique non-ordonné**
- ▶ intensité du temps partiel → **logit polytomique ordonné**
- ▶ salaire mensuel → **régression gamma**



11 / 109

Introduction : Modéliser des données non-linéaires

Quelques références utiles en ligne

Le modèle Logit. Théorie et application (Cédric Afsa, Document de travail de l'Insee)

<https://www.insee.fr/fr/statistiques/fichier/2022139/Le-modele-Logit-CB.pdf>

Régression logistique avec R (Luc Rouvière, Université Rennes 2, UFR Sciences sociales)

perso.univ-rennes2.fr/system/files/users/rouviere_l/poly_logistique_web.pdf

Tutoriels de UCLA

<http://www.ats.ucla.edu/stat/dae/>



12 / 109

Estimer un modèle logistique dichotomique



13 / 109

Estimer un modèle logistique dichotomique

Objectifs de l'estimation

$$f(y_i) = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \varepsilon_i$$

1. Trouver les paramètres β_0, \dots, β_p qui maximisent l'ajustement du modèle aux données.
2. Pouvoir quantifier la qualité de cet ajustement et ses conséquences sur l'estimation.

Mais contrairement au modèle linéaire classique, il n'existe aucune formule qui donne directement la valeur de

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$



14 / 109

Estimer un modèle logistique dichotomique

Principe d'estimation

On utilise donc des **algorithmes d'optimisation** pour maximiser une certaine **fonction objectif**, la **(log-)vraisemblance** du modèle.

La **forme de la log-vraisemblance** dépend de la **spécification du modèle** :

- ▶ la **distribution supposée** de Y : gaussienne, binomiale, gamma, poissonienne, etc.
- ▶ la **fonction de lien** utilisée : identité, logarithme, inverse, logit, etc.

En pratique dans **R**, ces opérations sont menées par la fonction `glm()` (pour *generalized linear model*) :

- ▶ paramètre `family` : distribution supposée de Y ;
- ▶ paramètre `link` (de `family`) : fonction de lien.

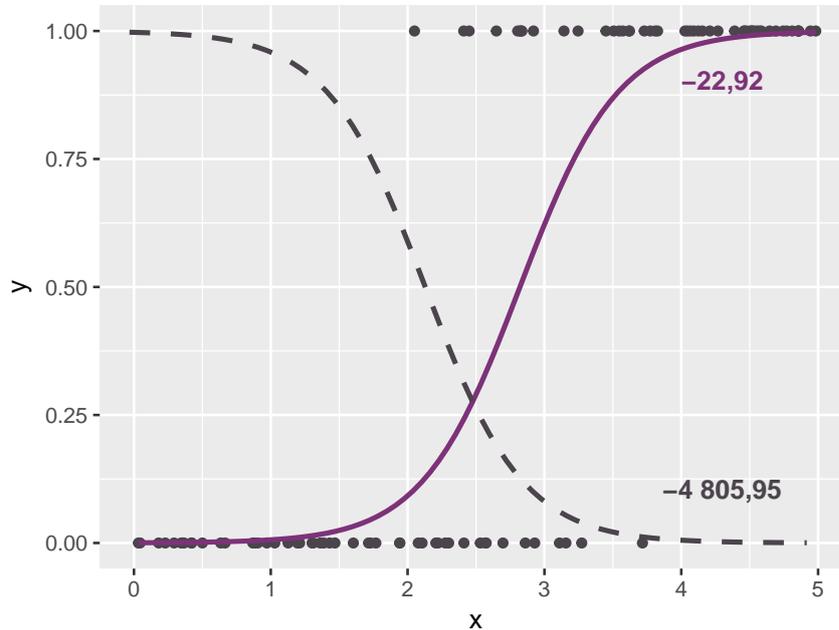


15 / 109

Estimation par maximum de vraisemblance

Vraisemblance et ajustement aux données

Plus la vraisemblance est élevée, meilleur est l'ajustement du modèle aux données.



16 / 109

Estimation par maximum de vraisemblance

Algorithme itératif

L'algorithme utilisé est **itératif** : il **maximise la log-vraisemblance** ℓ_n jusqu'à atteindre un **maximum**.

Itération 1 $\ell_n = -28,62$ $\beta_0 = -3,63$ $\beta_1 = 1,33$

Itération 2 $\ell_n = -24,16$ $\beta_0 = -5,55$ $\beta_1 = 1,98$

Itération 3 $\ell_n = -23,04$ $\beta_0 = -7,05$ $\beta_1 = 2,50$

Itération 4 $\ell_n = -22,92$ $\beta_0 = -7,75$ $\beta_1 = 2,75$

Itération 5 $\ell_n = -22,92$ $\beta_0 = -7,86$ $\beta_1 = 2,79$

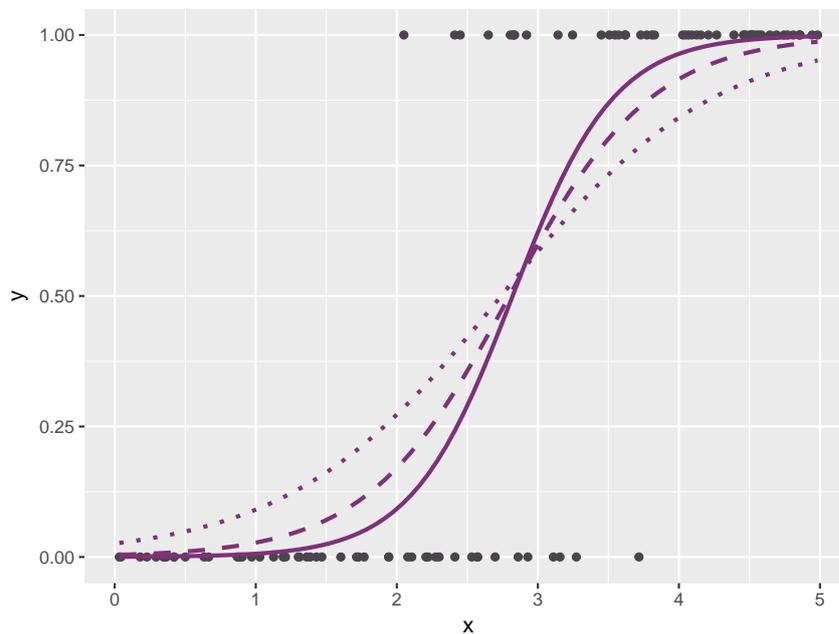


17 / 109

Estimation par maximum de vraisemblance

Algorithme itératif

L'algorithme utilisé est **itératif** : il **maximise la log-vraisemblance** ℓ_n jusqu'à atteindre un **maximum**.



Itérations 1, 2 et 5



17 / 109

Estimation par maximum de vraisemblance

Régression logistique dichotomique

Le modèle logistique dichotomique est une **spécification** du modèle linéaire général, que l'on peut réécrire :

$$\text{logit} [\mathbb{P}(y_i = 1 | X_i)] = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \varepsilon_i$$

Ses caractéristiques sont les suivantes :

- ▶ la quantité modélisée est la **probabilité** que la variable Y prenne la modalité 1 plutôt que la modalité 0 ;
- ▶ il appartient à la **famille binomiale** au sein des modèles linéaires généralisés ;
- ▶ sa **fonction de lien** est la fonction *logit* :

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

Complément Dérivation de la vraisemblance du modèle logistique dichotomique.



18 / 109

Estimation par maximum de vraisemblance

Parenthèse : La fonction *logit*

La fonction *logit* est historiquement utilisée pour exprimer sur \mathbb{R} une proportion p définie sur $]0; 1[$.

1. $\frac{p}{1-p}$ est la **cote** associée à la proportion p (comme dans les paris hippiques). Elle est à valeurs dans \mathbb{R}^+ .

Exemple Une probabilité de succès de 20 % correspond à une cote de 0,25 soit 1 succès pour 4 échecs. Dans les paris hippiques, on retourne le rapport et on dira « 4 contre 1 ».

2. $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$ est donc bien à valeurs dans \mathbb{R} .



19 / 109

Estimation par maximum de vraisemblance

Probabilités prédites par le modèle

L'estimation produit un vecteur $\hat{\beta}$ de paramètres estimés :

1. Dans le modèle linéaire, il suffit de **multiplier ces coefficients par les variables explicatives** pour obtenir la prédiction du modèle.
2. Dans le modèle logistique dichotomique, il faut de surcroît **appliquer la fonction réciproque de la fonction *logit***.

$$\begin{aligned}\hat{p}_i &= \mathbb{P}(\widehat{y_i = 1} | X_i) = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_p x_{p,i}) \\ &= \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_p x_{p,i}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_p x_{p,i}}} \\ &= \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_p x_{p,i})}}\end{aligned}$$



20 / 109

Estimation par maximum de vraisemblance

Inférence

Comme en régression linéaire classique, les paramètres du modèle sont estimés avec une certaine **imprécision**.

En plus de la valeur de $\hat{\beta}$, l'algorithme produit la matrice de variance-covariance dont on extrait les **erreurs standards** des coefficients.

Pour déterminer si un coefficient β_k est statistiquement différent de 0, on pose le **test statistique** :

$$H_0 : \beta_k = 0 \quad \text{contre} \quad H_1 : \beta_k \neq 0$$

On peut alors montrer que sous H_0 :

$$z = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

avec $se(\hat{\beta}_k)$ l'**erreur-standard** de $\hat{\beta}_k$.



Estimation par maximum de vraisemblance

Inférence

Il est dès lors possible de **tester la significativité** du coefficient β_k pour un risque de première espèce α donné (5 % ou 1 % en général) :

- ▶ en comparant la statistique de test au quantile à $1 - \alpha/2$ % d'une loi normale centrée réduite.

Rappel	90%	95%	97,5%	99%	99,5%
$q_{\gamma}^{\mathcal{N}(0,1)}$	1,28	1,64	1,96	2,33	2,58

- ▶ en interprétant la **p-valeur** : on peut rejeter H_0 au seuil α si la p-valeur est inférieure à α ;
- ▶ en construisant l'intervalle de confiance au seuil $1 - \alpha$:

$$IC_{1-\alpha} \%(\hat{\beta}_k) = \left[\hat{\beta}_k - q_{1-\alpha/2}^{\mathcal{N}(0,1)} se(\hat{\beta}_k); \hat{\beta}_k + q_{1-\alpha/2}^{\mathcal{N}(0,1)} se(\hat{\beta}_k) \right]$$

Exemple : Probabilité d'être en emploi stable

Mesure de la stabilité de l'emploi dans l'EEC

Pour illustrer l'estimation d'un modèle de régression logistique dichotomique, on mène l'étude de la probabilité d'être en emploi stable à partir de l'EEC.

Plusieurs questions de l'enquête Emploi en continu permettent de déterminer la stabilité du contrat de travail :

- ▶ **type de contrat** (CONTRA) : CDI, CDD, contrat saisonnier, intérim, apprentissage ou alternance ;
- ▶ **appartenance à la fonction publique** (CHPUB) ;
- ▶ **statut au sein de la fonction publique** (TITC) : titulaire, stagiaire ou contractuel.

On considère comme **en contrat stable les individus** :

- ▶ soit sous contrat de droit privé (y compris contractuels du public) en CDI ;
- ▶ soit fonctionnaires titulaires.



23 / 109

Exemple : Probabilité d'être en emploi stable

Mesure de la stabilité de l'emploi dans l'EEC

```
# Lecture du sous-échantillon
e <- readRDS("ee_da.rds")

# Restriction aux actifs occupés
# (et restriction aux moins de 70 ans)
e <- e[e$ACTEU == "1", ]
e$age <- as.numeric(e$AGE)
e <- e[e$age < 70, ]

# Création de la variable stable
e$stable <- 1 * (e$CONTRA %in% "1" |
  (e$CHPUB %in% c("1", "2", "3") & e$TITC %in% "2"))
```

Nombre de contrats considérés comme stables	751
---	-----

Nombre de contrats considérés comme instables	257
---	-----



24 / 109

Exemple : Probabilité d'être en emploi stable

Variables explicatives potentielles

On cherche tout particulièrement à mesurer la relation entre stabilité du contrat et variables **socio-démographiques** :

- ▶ **âge** : les plus jeunes sont-ils surreprésentés parmi les titulaires d'un contrat de travail instable ?
- ▶ **sexe** : constate-t-on un écart entre hommes et femmes en matière de stabilité du contrat de travail ?
- ▶ **diplôme** : un diplôme élevé protège-t-il de la précarité associée à un contrat de travail instable ?

On souhaite autant que possible prendre également en compte le **secteur d'activité agrégé** de l'entreprise (primaire, industrie, construction ou tertiaire).



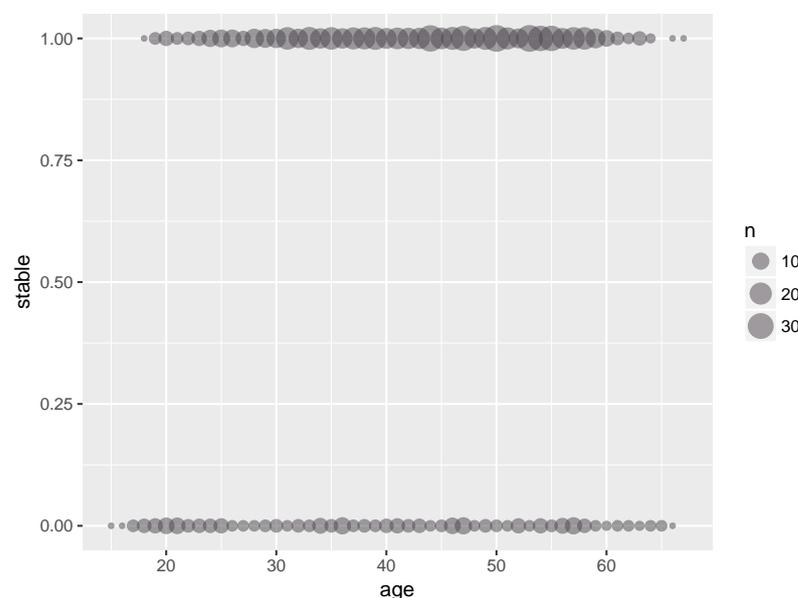
25 / 109

Exemple : Probabilité d'être en emploi stable

Relation entre âge et stabilité du contrat

On s'intéresse tout d'abord à la relation entre **âge et stabilité du contrat** :

$$\text{logit} [\mathbb{P}(\text{stable}_i = 1 | \text{age}_i)] = \beta_0 + \beta_1 \times \text{age}_i + \varepsilon_i$$



26 / 109

Exemple : Probabilité d'être en emploi stable

Relation entre âge et stabilité du contrat

```
# Régression logistique dichotomique simple
glm(
  formula = stable ~ age
  , data = e
  , family = binomial(link = "logit")
)
##
## Call:  glm(formula = stable ~ age, family = binomial(link = "logit"),
##      data = e)
##
## Coefficients:
## (Intercept)          age
##   -0.07058         0.02762
##
## Degrees of Freedom: 1007 Total (i.e. Null); 1006 Residual
## Null Deviance:      1145
## Residual Deviance: 1124  AIC: 1128

# Stockage des résultats dans une liste
m1 <- glm(formula = stable ~ age, data = e, family = binomial)
typeof(m1)
## [1] "list"
```

27 / 109

Exemple : Probabilité d'être en emploi stable

Relation entre âge et stabilité du contrat

```
# Affichage des principaux résultats
summary(m1)
##
## Call:
## glm(formula = stable ~ age, family = binomial, data = e)
##
## Deviance Residuals:
##   Min       1Q   Median       3Q      Max
## -1.9555 -1.3508  0.6911  0.7987  1.0025
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.07058    0.26021  -0.271    0.786
## age         0.02762    0.00617   4.476 0.0000076 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##   Null deviance: 1144.5  on 1007  degrees of freedom
## Residual deviance: 1124.2  on 1006  degrees of freedom
## AIC: 1128.2
##
## Number of Fisher Scoring iterations: 4
```

28 / 109

Exemple : Probabilité d'être en emploi stable

Relation entre âge et stabilité du contrat

```
# Résultats stockés dans m1
names(m1)
## [1] "coefficients"      "residuals"
## [3] "fitted.values"    "effects"
## [5] "R"                 "rank"
## [7] "qr"                "family"
## [9] "linear.predictors" "deviance"
## [11] "aic"               "null.deviance"
## [13] "iter"              "weights"
## [15] "prior.weights"    "df.residual"
## [17] "df.null"           "y"
## [19] "converged"         "boundary"
## [21] "model"             "call"
## [23] "formula"           "terms"
## [25] "data"              "offset"
## [27] "control"           "method"
## [29] "contrasts"         "xlevels"
```

29 / 109

Exemple : Probabilité d'être en emploi stable

Relation entre âge et stabilité du contrat

```
# Extraction de la log-vraisemblance
logLik(m1)
## 'log Lik.' -562.0967 (df=2)

# Extraction des coefficients
coef(m1)
## (Intercept)      age
## -0.07057609  0.02761678

# Extraction des coefficients
# et de leur p-valeur
coef(summary(m1))
##              Estimate Std. Error   z value    Pr(>|z|)
## (Intercept) -0.07057609 0.260212735 -0.2712246 0.786218318654
## age          0.02761678 0.006169563  4.4762936 0.000007595001
```

Exemple : Probabilité d'être en emploi stable Relation entre âge et stabilité du contrat

```
# Calcul des intervalles de confiance
# avec hypothèse gaussienne
confint.default(m1)
##           2.5 %      97.5 %
## (Intercept) -0.58058368 0.4394315
## age          0.01552466 0.0397089
# Note : confint() (sans .default) n'utilise
# pas l'hypothèse de normalité pour construire
# les intervalles de confiance.

# À la main !
alpha <- 0.05
m1_coef <- coef(summary(m1))
cbind(
  m1_coef[, 1] - qnorm(1 - alpha/2)*m1_coef[, 2]
  , m1_coef[, 1] + qnorm(1 - alpha/2)*m1_coef[, 2]
)
##           [,1]      [,2]
## (Intercept) -0.58058368 0.4394315
## age          0.01552466 0.0397089
```



31 / 109

Exemple : Probabilité d'être en emploi stable Relation entre âge et stabilité du contrat

```
# Probabilités prédites par le modèle
# avec predict()
p1 <- predict(m1, type = "response")
p1[1:3]
##      315000      315062      315299
## 0.7781692 0.7733654 0.6687847

# Probabilités prédites par le modèle
# avec fitted.values
p1b <- m1$fitted.values
p1b[1:3]
##      315000      315062      315299
## 0.7781692 0.7733654 0.6687847
identical(p1, p1b)
## [1] TRUE
```



32 / 109

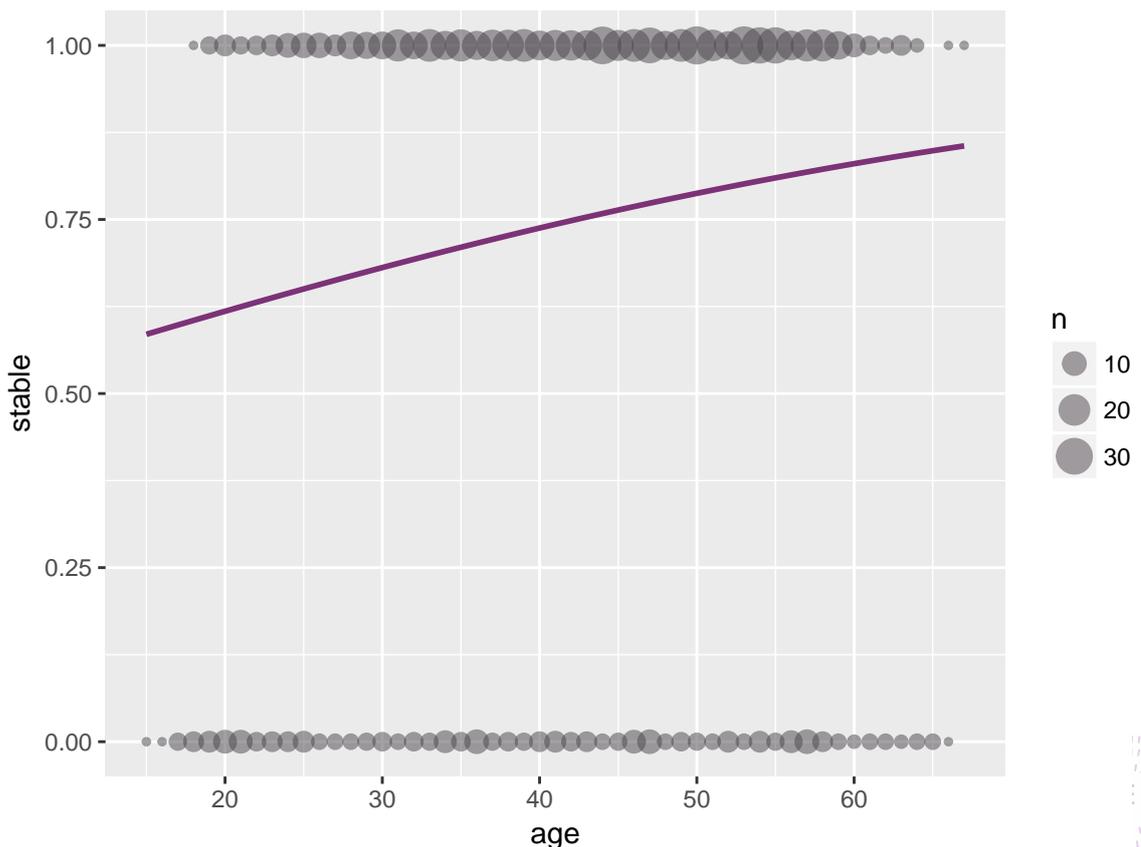
Exemple : Probabilité d'être en emploi stable Relation entre âge et stabilité du contrat

```
# Représentation de la relation avec base R
# - graphique initial
stable_age <- aggregate(
  e$stable, list(e$stable, e$age), length
)
names(stable_age) <- c("stable", "age", "n")
plot(stable_age$age, stable_age$stable, type="n")
symbols(
  x = stable_age$age, y = stable_age$stable
  , circles = sqrt(stable_age$n), bg = "purple"
  , inches = 1/4, ann = F, fg = NULL, add = T
)
# - Ajout de la courbe associée au modèle m1
curve(
  plogis(coef(m1)[1] + coef(m1)[2]*x)
  , add = TRUE
)
```



33 / 109

Exemple : Probabilité d'être en emploi stable Relation entre âge et stabilité du contrat



34 / 109

Exemple : Probabilité d'être en emploi stable

Relation entre âge et stabilité du contrat

Contrairement à ce qu'il se passe en régression linéaire, la valeur des coefficients ne peut **pas être interprétée directement**.

Ici le coefficient associé à l'âge est positif, aussi la relation entre âge et stabilité de l'emploi est **positive**, ce qu'illustrent les probabilités prédites.

Pour déterminer si β_1 est significativement différent de 0, on pose le **test statistique** :

$$H_0 : \beta_1 = 0 \quad \text{contre} \quad H_1 : \beta_1 \neq 0$$

La statistique de test vaut $4,48 > 1,96$ donc on peut **rejeter l'hypothèse H_0 au seuil de 5 %** ($0 \notin IC_{95} \%$).

La p-valeur est même inférieure à 0,01 donc on peut **rejeter H_0 au seuil de 1 %**.



Exemple : Probabilité d'être en emploi stable

Dichotomisation des variables qualitatives

Les autres variables explicatives du modèles sont des **variables qualitatives** :

- ▶ elles doivent être **dichotomisées** pour être intégrées au modèle ;
- ▶ toutes les indicatrices associées à une variable doivent être intégrées sauf une : la **modalité de référence**.

Exemple Pour intégrer la variable `sexe` dans le modèle :

1. on la dichotomise en deux variables indicatrices (homme et femme) ;
2. on intègre l'une ou l'autre au modèle.

Dès lors que la variable a **plus de deux modalités**, le choix de la modalité de référence n'est **pas neutre**.



Exemple : Probabilité d'être en emploi stable

Dichotomisation des variables qualitatives

```
# Dichotomisation des variables explicatives
e <- within(e, {
  homme <- (SEXE == "1") * 1
  femme <- (SEXE == "2") * 1
  supbac <- (DDIPL %in% c("1", "3")) * 1
  bac <- (DDIPL == "4") * 1
  infbac <- (DDIPL %in% c("5", "6", "7")) * 1
  agri <- (NAFG4N == "ES") * 1
  indus <- (NAFG4N == "ET") * 1
  cons <- (NAFG4N == "EU") * 1
  tert <- (NAFG4N == "EV") * 1
})
```



37 / 109

Exemple : Probabilité d'être en emploi stable

Dichotomisation des variables qualitatives

```
# Choix de la modalité de référence pour le sexe
coef(glm(stable ~ homme, data = e, family = binomial))
## (Intercept)      homme
##  1.2884811  -0.4152825

coef(glm(stable ~ femme, data = e, family = binomial))
## (Intercept)      femme
##  0.8731986   0.4152825

# Choix de la modalité de référence pour le diplôme
coef(glm(stable ~ bac + supbac, data = e, family = binomial))
## (Intercept)      bac      supbac
##  0.9972945  -0.1077740   0.3019885

coef(glm(stable ~ infbac + supbac, data = e, family = binomial))
## (Intercept)      infbac      supbac
##  0.8708284   0.1264662   0.4284546

coef(glm(stable ~ infbac + bac, data = e, family = binomial))
## (Intercept)      infbac      bac
##  1.2992830  -0.3019885  -0.4097625
```

38 / 109

Exemple : Probabilité d'être en emploi stable

Estimation du modèle complet

Formulation du modèle

$$\begin{aligned} \text{logit} [\mathbb{P}(\text{stable}_i = 1 | \text{age}_i)] &= \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{femme}_i + \beta_3 \text{infbac}_i \\ &+ \beta_4 \text{supbac}_i + \beta_5 \text{agri}_i + \beta_6 \text{cons}_i \\ &+ \beta_7 \text{tert}_i + \varepsilon_i \end{aligned}$$

```
m2 <- glm(stable ~ age + femme + infbac
          + supbac + agri + cons + tert
          , data = e , family = binomial
          )
coef(summary(m2))
##              Estimate Std. Error   z value    Pr(>|z|)
## (Intercept)  0.33083930 0.373150080  0.8866119 3.752879e-01
## age          0.03152330 0.006558967  4.8061371 1.538744e-06
## femme       0.35261130 0.161064115  2.1892604 2.857792e-02
## infbac      -0.08431424 0.208650357 -0.4040935 6.861440e-01
## supbac       0.21288742 0.219811305  0.9685008 3.327943e-01
## agri        -3.14835955 0.473026619 -6.6557767 2.818078e-11
## cons        -0.76040610 0.350850198 -2.1673241 3.021015e-02
## tert        -0.74364553 0.259468760 -2.8660311 4.156535e-03
```

39 / 109

Exemple : Probabilité d'être en emploi stable

Estimation du modèle complet

Les coefficients d'un modèle de régression logistique multiple rendent compte d'effets « **toutes choses égales par ailleurs** » ou plutôt :

tous les autres paramètres du modèle constants par ailleurs

Exemple

1. Le coefficient associé à la variable `cons` est négatif.
2. On interprète alors : « À âge, sexe et diplômes **égaux par ailleurs**, le fait de travailler dans la construction est associé à une **probabilité plus faible** d'être en emploi stable **par rapport** aux salariés de l'industrie (modalité de référence) ».

Exemple : Probabilité d'être en emploi stable

Estimation du modèle complet

À nouveau **la valeur des coefficients n'est pas interprétable en tant que telle**. Il est néanmoins possible d'interpréter :

- ▶ le signe des coefficients : relation positive s'ils sont positifs, négative sinon ;
- ▶ **au sein d'un même modèle**, l'amplitude relative des coefficients.

Exemple

1. En valeur absolue, le coefficient associé à la variable *femme* est supérieur à celui associé à la variable *tert*.
2. On interprète alors : « L'**effet propre** du sexe (à âge, diplôme et secteurs égaux par ailleurs) sur la probabilité d'être en emploi stable est **moindre** que celui associé au fait de travailler dans la construction **plutôt que** dans l'industrie (modalité de référence) ».



41 / 109

Estimer un modèle logistique dichotomique

En guise de conclusion

Le modèle de régression logistique est adapté pour modéliser des **données dichotomiques**.

Comme la plupart des spécifications du modèle linéaire général, son estimation est effectuée par **maximum de vraisemblance** et avec la fonction `glm()` de **R**.

Cette méthodologie fournit l'ensemble des éléments présents en régression linéaire classique, **coefficients** β_0, \dots, β_p et **erreurs standard** notamment.

Néanmoins l'**interprétation des coefficients est plus complexe**.



42 / 109

Compléments

Le modèle probit dichotomique

C'est la **fonction de lien** f qui différencie le modèle probit dichotomique du modèle logistique dichotomique.

Dans le modèle probit dichotomique, f est telle que

$$p_i = f^{-1}(X\beta) = \Phi(X\beta)$$

où $\Phi(x)$ est la **fonction de répartition de la loi normale centrée réduite**.

Ses coefficients diffèrent mais qualitativement **ses résultats sont proches** de ceux d'un modèle logistique dichotomique.

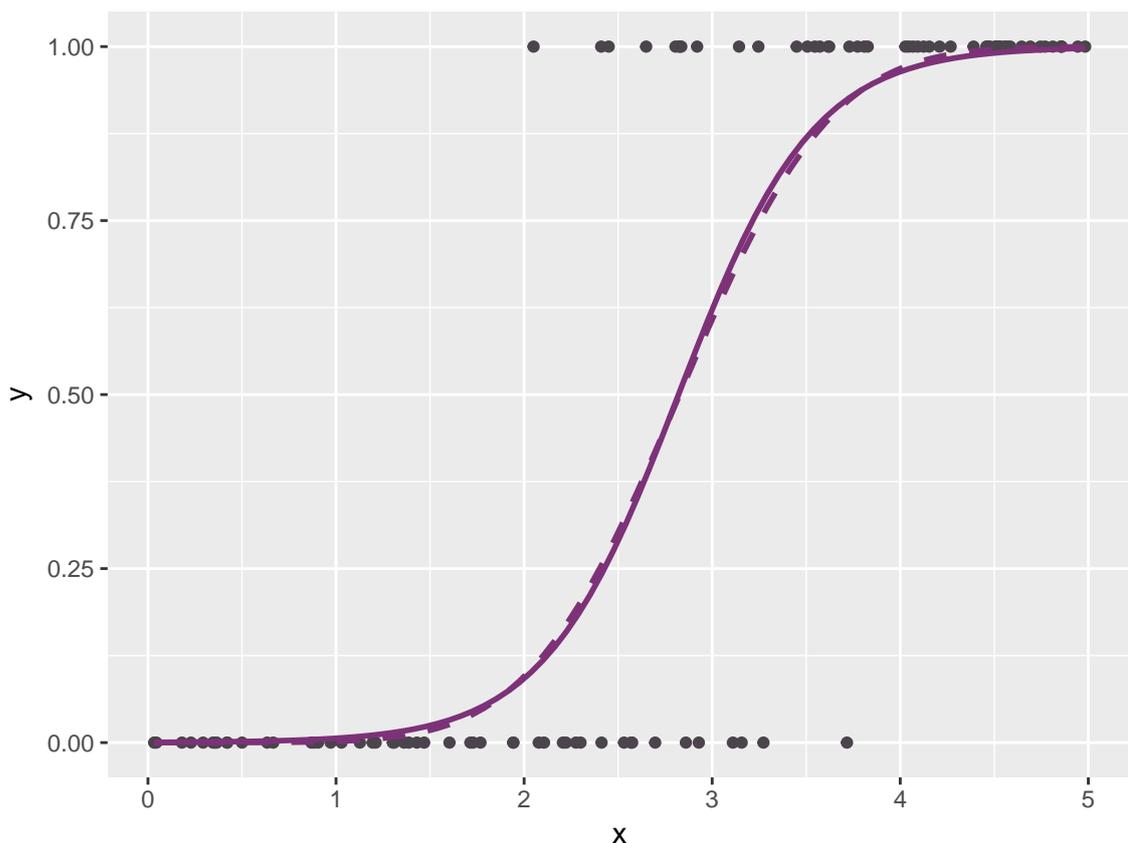
```
probit <- glm(z ~ x, data = sim_dicho
, family = binomial(link = "probit")
)
summary(probit)
```



43 / 109

Compléments

Le modèle probit dichotomique



44 / 109

Retour au modèle linéaire classique

Le modèle linéaire classique peut être vu comme un cas particulier de modèle linéaire général :

1. de la **famille gaussienne** ;
2. avec la **fonction de lien identité**.

```
# Avec la fonction lm()
lin_lm <- lm(y ~ x, data = sim_lin)

# Avec la fonction glm()
lin_glm <- glm(y ~ x, data = sim_lin
, family = gaussian(link = "identity")
)

# Comparaison des coefficients
identical(lin_lm$coefficients, lin_glm$coefficients)
## [1] TRUE
```

Vraisemblance du modèle logistique dichotomique

L'objectif de cette annexe est de déterminer l'**expression de la (log-)vraisemblance** dans le cas d'une régression logistique dichotomique.

Au-delà de son contenu théorique, elle doit permettre de **mieux comprendre les paramètres** à indiquer au logiciel pour effectuer l'estimation.

En toute généralité, la vraisemblance d'une variable Y sachant les observations X est définie par

$$L_n = \mathbb{P}(y_1, \dots, y_n | X_1, \dots, X_n)$$

Il s'agit de la **probabilité d'observer les valeurs** (y_1, \dots, y_n) **sachant les valeurs** (X_1, \dots, X_n) .

Compléments

Vraisemblance du modèle logistique dichotomique

Sous les hypothèses que les observations sont indépendantes les unes des autres et qu'elles suivent une même distribution, L_n devient :

$$L_n = \mathbb{P}(y_1|X_1) \times \dots \times \mathbb{P}(y_n|X_n) = \prod_{i=1}^n \mathbb{P}(y_i|X_i)$$

Pour faciliter les manipulations et le fonctionnement des algorithmes, on travaille en général sur la

log-vraisemblance ℓ_n :

$$\ell_n = \ln(L_n) = \ln \left[\prod_{i=1}^n \mathbb{P}(y_i|X_i) \right] = \sum_{i=1}^n \ln [\mathbb{P}(y_i|X_i)]$$

Pour déterminer l'expression de la vraisemblance dans le cas d'un modèle logistique dichotomique, on réexprime $\mathbb{P}(y_i|X_i)$.



47 / 109

Compléments

Vraisemblance du modèle logistique dichotomique

Dans le cas d'un modèle dichotomique, Y ne peut prendre que deux valeurs (0 ou 1), aussi :

$$\mathbb{P}(y_i|X_i) = \mathbb{P}(y_i = 1|X_i)^{y_i} \times \mathbb{P}(y_i = 0|X_i)^{1-y_i}$$

On dit que les modèles dichotomiques correspondent à la **famille binomiale** de modèles linéaires généraux.

$p_i = \mathbb{P}(y_i = 1|X_i)$ représente la **probabilité de succès**.

Comme

$$\mathbb{P}(y_i = 0|X_i) = 1 - \mathbb{P}(y_i = 1|X_i) = 1 - p_i$$

on peut réécrire :

$$\mathbb{P}(y_i|X_i) = p_i^{y_i} \times (1 - p_i)^{1-y_i}$$



48 / 109

Compléments

Vraisemblance du modèle logistique dichotomique

La **log-vraisemblance d'un modèle dichotomique**

s'écrit ainsi :

$$\begin{aligned}\ell_n &= \sum_{i=1}^n \ln [\mathbb{P}(y_i | X_i)] \\ &= \sum_{i=1}^n \ln [p_i^{y_i} \times (1 - p_i)^{1-y_i}] \\ &= \sum_{i=1}^n [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]\end{aligned}$$

où p_i est la probabilité de $y_i = 1$ sachant les variables explicatives X_i .

C'est la manière de **relier** p_i aux variables explicatives X_i qui distingue les différents modèles de régression pour variable dichotomique.



49 / 109

Compléments

Vraisemblance du modèle logistique dichotomique

Le modèle logistique dichotomique est le modèle tel que :

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i}$$

Ainsi

$$p_i = \text{logit}^{-1}(X_i \beta) = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}}$$

et donc

$$\begin{aligned}\ell_n &= \sum_{i=1}^n [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)] \\ &= \sum_{i=1}^n \left[y_i \ln \left(\frac{e^{X_i \beta}}{1 + e^{X_i \beta}} \right) + (1 - y_i) \ln \left(1 - \frac{e^{X_i \beta}}{1 + e^{X_i \beta}} \right) \right]\end{aligned}$$

L'estimateur du maximum de vraisemblance $\hat{\beta}$ est obtenu en maximisant la quantité ℓ_n .



50 / 109

Interpréter un modèle logistique dichotomique



51 / 109

Interpréter un modèle logistique dichotomique De l'importance de l'interprétation

L'interprétation d'un modèle renvoie à plusieurs opérations essentielles :

1. **Evaluer sa pertinence et sa qualité** : comme toute tentative de modélisation, un modèle de régression logistique dichotomique présente des **limites**.
2. **Expliciter la signification des coefficients** : plus encore que dans le modèle de régression linéaire classique, l'interprétation des coefficients est complexe.
3. **Confronter la modélisation aux questions que l'on se pose** avec des tests d'hypothèses complexes.



52 / 109

Interpréter un modèle logistique dichotomique

Exemple : Stabilité du contrat de travail

Comme dans la partie précédente, la plupart des exemples sont tirés de l'étude sur la stabilité du contrat de travail.

Pour rappel, ont été estimés le modèle logistique simple $m1$:

$$\text{logit} [\mathbb{P}(\text{stable}_i = 1 | \text{age}_i)] = \beta_0 + \beta_1 \times \text{age}_i + \varepsilon_i$$

et le modèle logistique multiple $m2$:

$$\begin{aligned} \text{logit} [\mathbb{P}(\text{stable}_i = 1 | \text{age}_i)] = & \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{femme}_i + \beta_3 \text{infbac}_i \\ & + \beta_4 \text{supbac}_i + \beta_5 \text{agri}_i + \beta_6 \text{cons}_i \\ & + \beta_7 \text{tert}_i + \varepsilon_i \end{aligned}$$



53 / 109

Indicateurs de qualité du modèle

Statistiques construites à partir de ℓ_n

Pour comparer deux modèles portant sur la même variable expliquée, on peut comparer les valeurs de leur vraisemblance : **on privilégie le modèle présentant la plus grande vraisemblance.**

Cependant, quand un modèle comporte davantage de variables explicatives, son pouvoir prédictif **augmente mécaniquement** (comme pour le R^2).

On peut alors utiliser des indicateurs qui **pénalisent la vraisemblance par le nombre de variables (p)** :

- ▶ *Akaike information criterion* : $AIC = -2\ell_n + 2(p + 1)$
- ▶ *Bayesian information criterion* (ou critère de Schwartz) : $BIC = -2\ell_n + \ln(n)(p + 1)$



54 / 109

Indicateurs de qualité du modèle

Exemple : Stabilité du contrat de travail

```
# La "deviance" du modèle correspond à  $-2 \times l_n$ 
m2$deviance
## [1] 1056.144
-2*logLik(m2)
## 'log Lik.' 1056.144 (df=8)

# Calcul manuel de l'AIC et du BIC
m2$deviance + 2*(m2$rank) #AIC
## [1] 1072.144
m2$deviance + log(length(m2$y))*m2$rank #BIC
## [1] 1111.47

# Calcul avec les fonctions AIC() et BIC()
AIC(m2)
## [1] 1072.144
BIC(m2)
## [1] 1111.47
```

55 / 109

Indicateurs de qualité du modèle

Exemple : Stabilité du contrat de travail

```
# Comparaison des log-vraisemblance
logLik(m1)
## 'log Lik.' -562.0967 (df=2)
logLik(m2)
## 'log Lik.' -528.0719 (df=8)

# Comparaison des AIC
AIC(m1)
## [1] 1128.193
AIC(m2)
## [1] 1072.144

# Comparaison des BIC
BIC(m1)
## [1] 1138.025
BIC(m2)
## [1] 1111.47
```

56 / 109

Indicateurs de qualité du modèle

Test de significativité globale

Pour évaluer le **pouvoir explicatif** d'un modèle, on peut comparer sa vraisemblance à celle du modèle ne comportant que la constante.

Il est possible de formaliser cette comparaison dans le cadre du test du **ratio de vraisemblance**.

On peut en effet montrer que sous l'hypothèse H_0 d'égalité des deux vraisemblances,

$$LR = -2 \ln \left(\frac{L^0}{L_n} \right) = (-2\ell^0) - (-2\ell_n) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi_p^2$$

avec ℓ^0 la log-vraisemblance du modèle ne comportant que la constante.



57 / 109

Indicateurs de qualité du modèle

Exemple : Stabilité du contrat de travail

```
# Calcul de la statistique de test à partir de la deviance
LR <- m2$null.deviance - m2$deviance
LR
## [1] 88.37818

# P-valeur du test
pchisq(LR, df = m2$rank - 1, lower.tail = FALSE)
## [1] 2.662749e-16

# Calcul automatique avec le package lmtest
library(lmtest)
lrtest(m2)
## Likelihood ratio test
##
## Model 1: stable ~ age + femme + infbac + supbac + agri + cons + tert
## Model 2: stable ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1    8 -528.07
## 2    1 -572.26 -7 88.378  2.663e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

58 / 109

Indicateurs de qualité du modèle

Pourcentage de concordance

Le modèle de régression permet d'obtenir, pour chaque individu de l'échantillon, une probabilité prédite \hat{p}_i sur la base des variables explicatives.

On peut alors classer chaque paire d'observations selon trois catégories :

- ▶ **concordante** : $y_1 = 0, y_2 = 1$ et $\hat{p}_1 < \hat{p}_2$ ou $y_1 = 1, y_2 = 0$ et $\hat{p}_1 > \hat{p}_2$
- ▶ **discordante** : $y_1 = 0, y_2 = 1$ et $\hat{p}_1 > \hat{p}_2$ ou $y_1 = 1, y_2 = 0$ et $\hat{p}_1 < \hat{p}_2$
- ▶ **ex-aequo** : $\hat{p}_1 = \hat{p}_2$.

On peut alors calculer un **pourcentage de paires concordantes** rapporté au nombre de paires total.



59 / 109

Indicateurs de qualité du modèle

Exemple : Stabilité du contrat de travail

```
# La fonction de calcul conc() est définie
# dans le support d'exercices
conc(m2)
## Pct concordant Pct discordant Pct ex-aequo
## 66.9649287 32.8029553 0.2321159
```



60 / 109

Indicateurs de qualité du modèle

Performance de la classification et courbe ROC

Bien souvent, l'objectif d'un modèle est d'aboutir à une **classification binaire**.

Exemples Le radar détecte-t-il un avion ennemi? Le message reçu est-il un *spam*?

Mais en sortie du modèle, on obtient pour chaque individu la probabilité \hat{p}_i , et non une valeur 0 ou 1.

Question Où placer la probabilité seuil p^* entre les cas à classer comme positifs ($\hat{p}_i > p^*$) et les cas à classer comme négatifs ($\hat{p}_i < p^*$)?



Indicateurs de qualité du modèle

Performance de la classification et courbe ROC

Le meilleur modèle serait celui qui ne conduirait à **aucun faux négatif et aucun faux positif**.

Mais on a en fait affaire à un arbitrage :

- ▶ **Si le seuil est trop haut**, certains individus positifs risquent d'être classés comme négatifs (faux négatifs).
- ▶ **Si le seuil est trop bas**, certains individus négatifs risquent d'être classés comme positifs (faux positifs).

Moralité Afin de limiter le risque de faux négatifs on est amené à tolérer un certain nombre de faux positifs, et inversement.

La courbe ROC (*Receiver operating characteristics*) représente cet arbitrage.



Construction de la courbe ROC

1. Estimer le modèle et classer les observations par **probabilités prédites \hat{p}_i croissantes** ;
2. Pour chaque observation i :
 - ▶ calculer la part des **positifs classés positifs** (**sensibilité**) si \hat{p}_i constitue le seuil entre positif et négatif ;
 - ▶ calculer la part des **négatifs classés négatifs** (**spécificité**) si \hat{p}_i constitue le seuil entre positif et négatif ;
3. La courbe ROC est la représentation de la **sensibilité en fonction de la spécificité** (axe inversé).

L'**aire sous la courbe** (*Area under the curve* ou AUC) est un **indicateur synthétique de la performance** de classification du modèle.



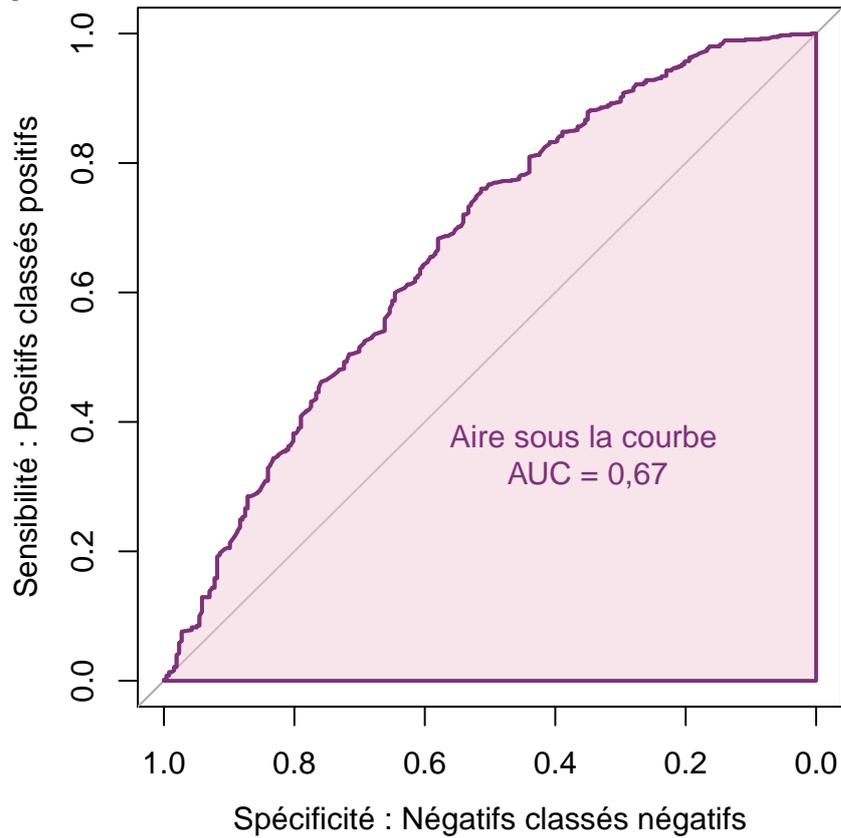
Exemple : Stabilité du contrat de travail

```
# Calcul de la courbe ROC avec le package pROC
library(pROC)
m2_roc <- roc(m2$y ~ m2$fitted.values)
m2_roc
##
## Call:
## roc.formula(formula = m2$y ~ m2$fitted.values)
##
## Data: m2$fitted.values in 257 controls (m2$y 0) < 751 cases (m2$y 1)
## Area under the curve: 0.6708
```

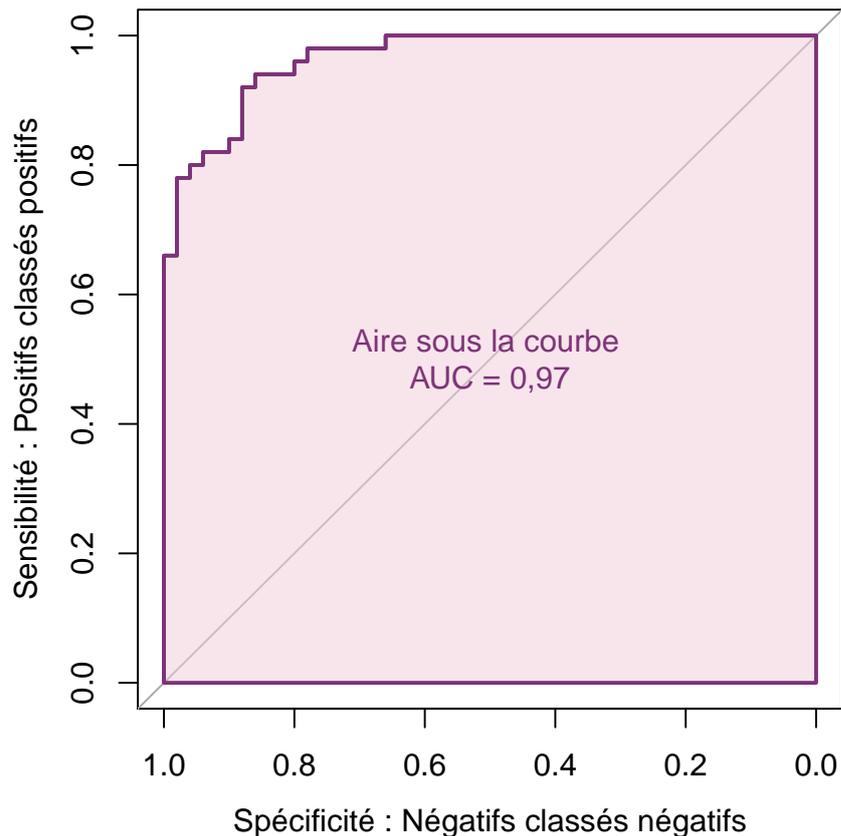
```
# Représentation avec la fonction plot()
par(pty="s")
plot(m2_roc
, xlab = "Spécificité : Négatifs classés négatifs"
, ylab = "Sensibilité : Positifs classés positifs"
, asp = TRUE
)
```



Exemple : Stabilité du contrat de travail



Comparaison : Données simulées



Odds-ratio et effets marginaux

Définition de l'odds-ratio

On rappelle que la « **cote** » (ou *odd*) d'une proportion p est le rapport

$$odd_p = \frac{p}{1-p}$$

Exemple Pour une proportion de 25 %, la cote est de 1/3 (ou 3 contre 1 dans les paris hippiques).

On appelle alors « **rapport des cotes** » (ou *odds-ratio*) des proportions p et q :

$$OR_{p|q} = \frac{\frac{p}{1-p}}{\frac{q}{1-q}}$$

Interprétation Si $p > q$ alors $OR_{p|q} > 1$.



67 / 109

Odds-ratio et effets marginaux

Les odds-ratio dans une régression logistique

Mathématiquement, les *odds-ratio* d'un modèle de régression logistique correspondent à l'exponentielle de la valeur des coefficients :

$$OR_{k|ref} = e^{\beta_k} = \exp(\beta_k)$$

```
cbind(coef=coef(m2), or = exp(coef(m2)))
##              coef              or
## (Intercept) 0.33083930 1.39213605
## age         0.03152330 1.03202542
## femme      0.35261130 1.42277800
## infbac     -0.08431424 0.91914238
## supbac     0.21288742 1.23724535
## agri       -3.14835955 0.04292248
## cons       -0.76040610 0.46747655
## tert       -0.74364553 0.47537775
```

Remarque Quand le coefficient est positif, l'*odds-ratio* est supérieur à 1 et inversement.



68 / 109

Odds-ratio et effets marginaux

Inférence à partir des *odds-ratio*

Pour réaliser une inférence à partir des *odds-ratio*, il suffit de **recalculer les bornes de l'intervalle de confiance au niveau souhaité**.

```
cbind(OR = exp(coef(m2)), exp(confint.default(m2)))
##                OR          2.5 %          97.5 %
## (Intercept) 1.39213605 0.66997062 2.8927280
## age         1.03202542 1.01884329 1.0453781
## femme      1.42277800 1.03762187 1.9509007
## infbac     0.91914238 0.61063161 1.3835227
## supbac     1.23724535 0.80417785 1.9035292
## agri       0.04292248 0.01698415 0.1084740
## cons       0.46747655 0.23502583 0.9298311
## tert       0.47537775 0.28587662 0.7904949
```



69 / 109

Odds-ratio et effets marginaux

Odds-ratio et risque relatif

Le terme « **risque relatif** » des proportions p et q désigne le rapport : $RR_{p|q} = \frac{p}{q}$.

Les confusions entre risque relatif et *odds-ratio* sont fréquentes.

Si pour les proportions rares les deux quantités sont proches, pour les proportions fréquentes ce n'est pas du tout le cas.

Exemple $p = 0,70$, $q = 0,40$

- ▶ $RR_{p|q} = \frac{0,70}{0,40} = 1,75$
- ▶ $OR_{p|q} = \frac{0,70/0,30}{0,40/0,60} = 3,5$

Pour aller plus loin Le site de la sociologue Carina Mood
<http://logisticregression.su.se/> .



70 / 109

Définition de l'effet marginal

Dans un modèle logistique dichotomique, l'**effet marginal** est un moyen simple pour réexprimer la relation entre une variable explicative et la variable d'intérêt en termes de **points de pourcentages**.

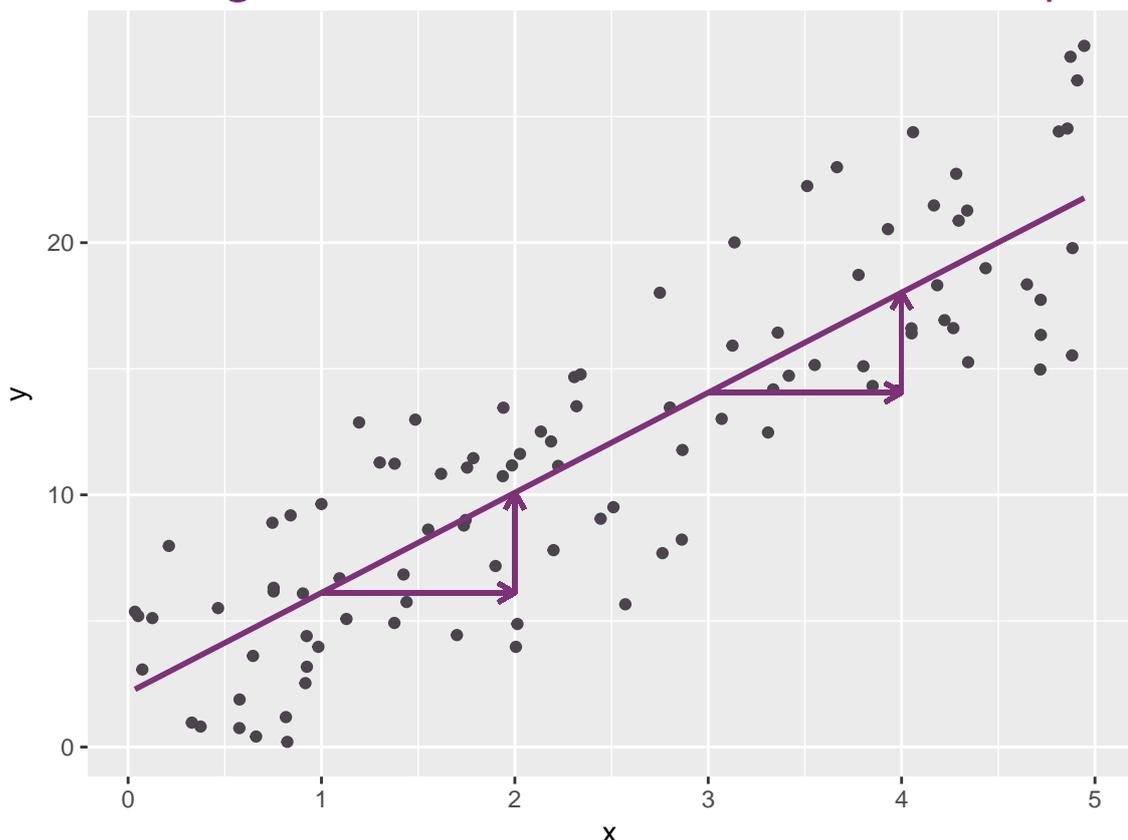
Exemple En moyenne dans l'échantillon et à âge, sexe et diplôme égaux par ailleurs, le fait de travailler dans la construction plutôt que dans l'industrie est associé à une probabilité inférieure d'avoir un emploi stable de l'ordre de 15 points de pourcentage.

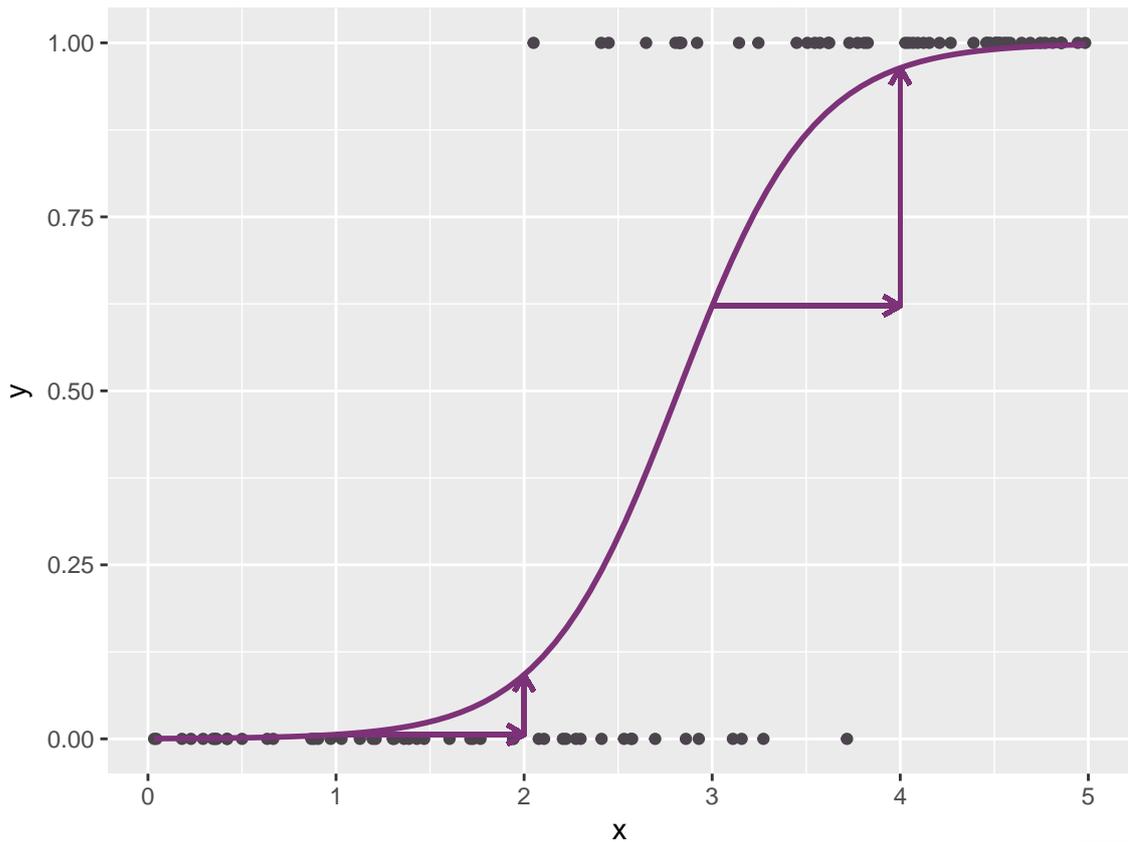
Dans le **modèle linéaire classique**, l'effet marginal de la variable x_j sur Y est tout simplement $\hat{\beta}_j$.

Exemple Ainsi dans $salairer_i = \beta_0 + \beta_1 age_i + \varepsilon_i$ l'effet marginal de la variable age est **constant et égal à $\hat{\beta}_1$** .



Effet marginal dans un modèle linéaire classique





Dans un modèle de régression logistique dichotomique, l'effet marginal de la variable x_j sur Y peut **varier d'un individu à l'autre**.

Quand la variable x_j est **dichotomique**, le calcul de l'effet marginal de la variable x_j pour l'individu i $\delta_i(x_j)$ est effectué de la façon suivante :

1. on calcule la probabilité de i prédite par le modèle $\hat{p}_{i|x_j=1}$ si x_j **était égale à 1** ;
2. on calcule la probabilité de i prédite par le modèle $\hat{p}_{i|x_j=0}$ si x_j **était égale à 0** ;
3. on calcule l'effet marginal avec :

$$\delta_i(x_j) = \hat{p}_{i|x_j=1} - \hat{p}_{i|x_j=0}$$

Odds-ratio et effets marginaux

Effet marginal dans un modèle logistique

Exemple Dans le modèle

$$\mathbb{P}(\text{stable}_i = 1 | \text{age}_i, \text{femme}_i) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{femme}_i + \varepsilon_i$$

on calcule l'effet marginal du sexe sur la stabilité de l'emploi pour un individu i $\delta_i(\text{femme})$ de la façon suivante :

1. on calcule $\hat{p}_{i|\text{femme}=1} = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 \text{age}_i + \hat{\beta}_2)$;
2. on calcule $\hat{p}_{i|\text{femme}=0} = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 \text{age}_i)$;
3. l'effet marginal est alors

$$\delta_i(\text{femme}) = \hat{p}_{i|\text{femme}=1} - \hat{p}_{i|\text{femme}=0}$$

Si la relation entre stabilité de l'emploi et le fait d'être une femme est **positive** ($\hat{\beta}_2 > 0$), $\delta_i(\text{femme}) > 0$, et inversement.



75 / 109

Odds-ratio et effets marginaux

Effet marginal moyen

L'effet marginal moyen est directement calculé comme la moyenne sur l'échantillon des effets marginaux individuels :

$$\begin{aligned} \bar{\delta}(x_j) &= \frac{1}{n} \sum_{i=1}^n \delta_i(x_j) = \frac{1}{n} \sum_{i=1}^n \hat{p}_{i|x_j=1} - \frac{1}{n} \sum_{i=1}^n \hat{p}_{i|x_j=0} \\ &= \bar{p}_{|x_j=1} - \bar{p}_{|x_j=0} \end{aligned}$$

Interprétation L'effet marginal moyen correspond à l'**augmentation moyenne dans l'échantillon** de la probabilité $\mathbb{P}(Y = 1)$ quand x_j passe de 0 à 1.

Exemple Si dans le modèle de la diapositive précédente $\bar{\delta}(\text{femme}) = 0,10$, on dira qu'en moyenne dans l'échantillon et à âge égal par ailleurs, le fait d'être une femme est associé à une probabilité d'être en contrat stable supérieure **de 10 points de pourcentage**.



76 / 109

Exemple : Stabilité du contrat de travail

```
# La fonction de calcul margins() est définie
# dans le support d'exercices
margins(m2) [1:3]
## $femme
##           Average MFX Std. Error   z value   P>|z|
## femme = 0   0.71396457         NA         NA         NA
## femme = 1   0.77498811         NA         NA         NA
## Diff        0.06102354 0.02789538  2.187586 0.02869976
##
## $infbac
##           Average MFX Std. Error   z value   P>|z|
## infbac = 0   0.75188391         NA         NA         NA
## infbac = 1   0.73736314         NA         NA         NA
## Diff        -0.01452077 0.03599009 -0.4034658 0.6866056
##
## $supbac
##           Average MFX Std. Error   z value   P>|z|
## supbac = 0   0.73252143         NA         NA         NA
## supbac = 1   0.76868772         NA         NA         NA
## Diff         0.03616628 0.03683972  0.9817197 0.3262379
```



Intérêt de l'effet marginal moyen

L'effet marginal moyen présente plusieurs avantages :

1. Il s'exprime en **termes de probabilités**, ce qui le rend extrêmement intuitif et facile à utiliser.
2. Des **erreurs standards** peuvent être obtenues pour l'effet marginal moyen, ce qui permet de juger de la significativité de l'écart en termes de probabilité.
3. La comparaison d'effets marginaux moyens entre plusieurs modèles emboîtés semble **plus robuste** que la comparaison des *odds-ratio* à l'hétérogénéité inobservée.

Pour aller plus loin MOOD C. (2010)

<https://doi.org/10.1093/esr/jcp006>



Test d'hypothèses complexes

Motivation et exemples

Il est fréquent que certaines hypothèses ne puissent pas être testées à l'aide d'un test sur **un seul paramètre** :

- ▶ Quelle est la relation entre le niveau de diplôme pris dans son ensemble et stabilité du contrat ?
- ▶ Quelle est la relation entre le secteur d'activité pris dans son ensemble et la stabilité du contrat ?

Dans le **cas du diplôme** par exemple, on pose ce test de la façon suivante :

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

Rappel du modèle

$$\begin{aligned} \text{stable}_i = & \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{femme}_i + \beta_3 \text{infbac}_i \\ & + \beta_4 \text{supbac}_i + \beta_5 \text{agri}_i + \beta_6 \text{cons}_i + \beta_7 \text{tert}_i + \varepsilon_i \end{aligned}$$



Test d'hypothèses complexes

Test du ratio de vraisemblance

Principe Comparer la log-vraisemblance de **deux modèles emboîtés** :

- ▶ d'une part le modèle **complet** ou **non-contraint** qui comporte tous les paramètres ;
- ▶ d'autre part le modèle **contraint** qui correspond au cas où l'hypothèse nulle est vérifiée.

Dans le **cas du diplôme**, le modèle contraint est :

$$\begin{aligned} \text{stable}_i = & \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{femme}_i + \beta_5 \text{agri}_i \\ & + \beta_6 \text{cons}_i + \beta_7 \text{tert}_i + \varepsilon_i \end{aligned}$$

Intuition Si le modèle non-contraint est **beaucoup plus vraisemblable** que le modèle contraint, alors on a tendance à **rejeter la contrainte**, c'est-à-dire l'hypothèse nulle.



Statistique de test

On peut montrer que sous H_0 :

$$LR = -2 \ln \left(\frac{L_c}{L_{nc}} \right) = (-2\ell_c) - (-2\ell_{nc}) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi_q^2$$

où ℓ_{nc} est la log-vraisemblance du modèle non-contraint, ℓ_c la log-vraisemblance du modèle contraint et q le nombre de restrictions.

Dans le **cas du diplôme** :

1. $\ell_{nc} = -528,07$ et $\ell_c = -529,55$ et ainsi $LR = 2,95$;
2. La p-valeur du test du ratio de vraisemblance vaut $1 - F_{\chi^2_2}(LR) = 0,2291$;
3. On ne peut pas rejeter l'hypothèse nulle au seuil de 5 %.



Exemple : Stabilité du contrat de travail

```
# On teste la significativité jointe des coefficients
# associés au secteur d'activité de l'entreprise
m4 <- glm(stable ~ age + femme + infbac + supbac
, data = e, family = binomial(link = "logit")
)

# Test du ratio de vraisemblance avec le package lmtest
library(lmtest)
lrtest(m2, m4)
## Likelihood ratio test
##
## Model 1: stable ~ age + femme + infbac + supbac + agri + cons + tert
## Model 2: stable ~ age + femme + infbac + supbac
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1    8 -528.07
## 2    5 -556.10 -3  56.065  4.069e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Le secteur d'activité semble très significatif.
```



Adapter la spécification du modèle aux données



83 / 109

Adapter la spécification du modèle aux données Au-delà des données dichotomiques

Les deux premières parties ont permis d'introduire le modèle linéaire général et de développer son application aux données de nature dichotomique.

Néanmoins, de très nombreuses autres formes de **non-linéarité** dans la relation entre Y et les variables explicatives peuvent survenir en pratique :

- ▶ variables **polytomiques** : information non-ordonnée ou ordonnée ;
- ▶ variables **asymétriques** : variable asymétrique continue ou discrète (données de comptage).

Le même principe s'applique ici à ces nouveaux types de données : **chercher la spécification du modèle linéaire général qui permette de les modéliser au mieux.**



84 / 109

Définition et exemples

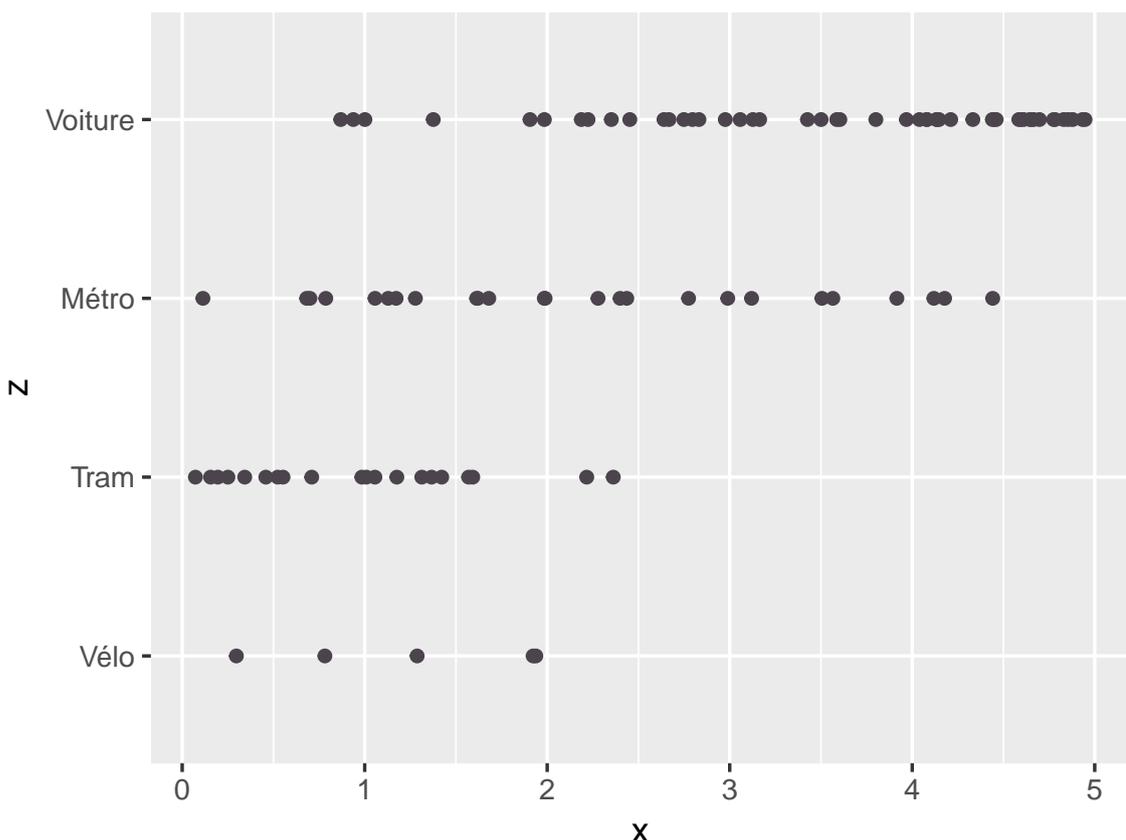
On parle de données polytomiques non-ordonnées dès lors que la variable d'intérêt Y correspond à une **alternative** parmi plus de deux possibilités (*discrete choice model*).

Exemples

- ▶ déterminants de l'**orientation des élèves** après le bac ;
- ▶ relation entre **santé et position sur le marché du travail** (activité, chômage ou inactivité dont incapacité) ;
- ▶ choix opéré par un client parmi un ensemble d'offres avec des **positionnements qualitatifs différents** (modèles ou catégorie de voiture par exemple) ;
- ▶ **distance domicile-travail** et choix du mode de transport.



Illustration : Choix du mode de transport



Données polytomique non-ordonnées

Principes de modélisation

Idée On se ramène à une série de modèles logistiques dichotomiques en **choisissant une modalité de la variable à expliquer comme référence**.

Exemple Dans le modèle sur le choix de mode de transport domicile-travail, la voiture s'impose comme la modalité de référence :

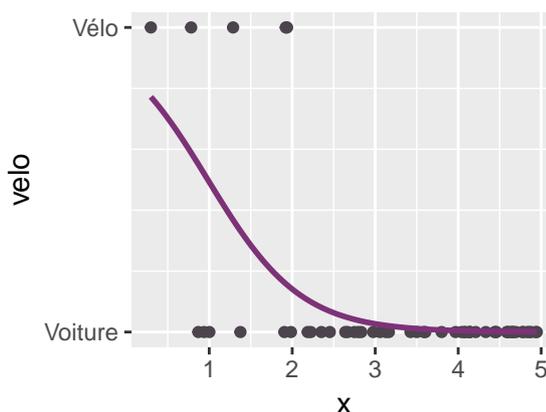
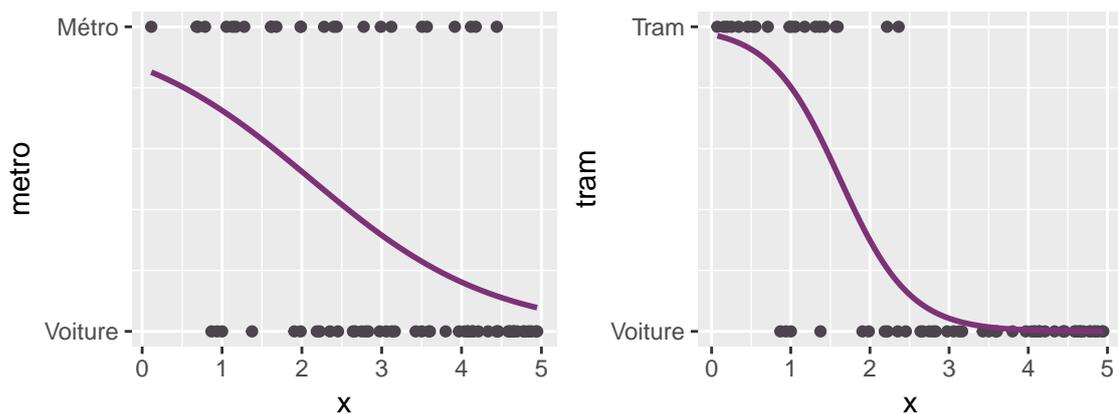
- ▶ elle est **ultra-dominante** quelle que soit la distance domicile-travail ;
- ▶ la limitation de son utilisation constitue un **enjeu** de la politique des transports de plusieurs grandes agglomérations.

On **modélise donc les probabilités** : (1) de prendre le métro *plutôt que la voiture*, (2) de prendre le tram *plutôt que la voiture*, (3) de prendre le vélo *plutôt que la voiture*.

87 / 109

Données polytomique non-ordonnées

Principes de modélisation



Données polytomique non-ordonnées

Principes de modélisation

Les paramètres de ce modèle logistique **multinomial** sont estimés **simultanément** : il y a une seule vraisemblance pour les trois alternatives à la voiture.

Note Ce modèle n'est donc pas équivalent à trois régressions logistiques dichotomiques menées indépendamment sur chaque sous-échantillon.

De plus, on peut dans ce type de modèle intégrer des variables dont **la valeur dépend de l'alternative choisie** (*alternative-specific variable*).

Exemple Pour le mode de transport, il est possible d'évaluer le **coût moyen** associé à chacun des quatre modes envisagés même si en définitive un seul est choisi.



89 / 109

Données polytomique non-ordonnées

Interprétation du modèle

Toutes les interprétations sont à effectuer **par rapport à la modalité de référence de la variable dépendante**.

En définitive, on est donc amené à **interpréter les coefficients** comme dans le modèle logistique dichotomique :

- ▶ la valeur des coefficients n'est pas interprétable en tant que telle ;
- ▶ des odds-ratio peuvent être calculés avec $OR_k = e^{\beta_k}$.

Les **probabilités bilatérales** estimées par le modèle peuvent être **combinées pour déterminer la probabilité globale** d'utiliser l'un ou l'autre des modes de transport.

Comme dans le modèle logistique dichotomique, il est donc possible de calculer et d'interpréter des **effets marginaux moyens** pour les variables qualitatives du modèle.

90 / 109

Données polytomique non-ordonnées

Hypothèse de validité

Cette modélisation repose sur l'hypothèse dite d'« **indépendance vis-à-vis des alternatives non-pertinentes** ».

En pratique, cela signifie que pour que, le modèle « tienne », il faut que le mécanisme de choix considère les alternatives **deux-à-deux uniquement**.

Exemple Le fait de choisir le vélo plutôt que la voiture doit être indépendant du fait de choisir le tram.

Le **test de Hausman et MacFadden** permet de tester cette hypothèse en comparant le modèle complet aux modèles dans lesquels on exclut explicitement les alternatives non-pertinentes.



Données polytomique non-ordonnées

Exemple : Position sur le marché du travail

La **position sur le marché du travail** est une variable polytomique : actif occupé, chômeur, inactif.

Certaines études portant par exemple sur les transitions entre vie active et retraite s'appuient sur une **modélisation multinomiale** de la position sur le marché du travail.

On prend en général le fait d'être **actif occupé** comme référence et on modélise donc (simultanément) :

1. la probabilité d'être au chômage *par rapport à en activité*;
2. la probabilité d'être inactif *par rapport à en activité*.

Les **variables explicatives** envisagées ici sont l'âge, le sexe et le diplôme.



Données polytomique non-ordonnées

Exemple : Position sur le marché du travail

```
# Chargement du package mlogit
library(mlogit)

# Passage de la variable ACTEU en facteur
e$acteu <- factor(
  as.integer(e$ACTEU)
  , labels = c("Actifs occupés", "Chômeurs", "Inactifs")
)

# Reformatage des données
# 1 ligne par alternative (donc 3 par individu)
e_long <- mlogit.data(e, shape = "wide", choice = "acteu")
nrow(e)
## [1] 2128
nrow(e_long)
## [1] 6384

# Estimation du modèle avec la fonction mlogit()
m_multi <- mlogit(
  acteu ~ 0 | age + infbac + supbac + femme
  , data = e_long, refllevel = "Actifs occupés"
)
```

93 / 109

Données polytomique non-ordonnées

Exemple : Position sur le marché du travail

```
summary(m_multi)
##
## Call:
## mlogit(formula = acteu ~ 0 | age + infbac + supbac + femme, data = e_long,
##        refllevel = "Actifs occupés", method = "nr", print.level = 0)
##
## Frequencies of alternatives:
## Actifs occupés      Chômeurs      Inactifs
##      0.475094      0.047462      0.477444
##
## nr method
## 6 iterations, 0h:0m:0s
## g'(-H)^-1g = 2.56E-07
## gradient close to zero
##
## Coefficients :
##              Estimate Std. Error t-value Pr(>|t|)
## Chômeurs:(intercept) -1.3595738  0.3108728  -4.3734 1.223e-05 ***
## Inactifs:(intercept) -1.9421422  0.1720676 -11.2871 < 2.2e-16 ***
## Chômeurs:age         -0.0176837  0.0061715  -2.8654  0.004165 **
## Inactifs:age         0.0403059  0.0027489  14.6626 < 2.2e-16 ***
## Chômeurs:infbac      0.0133736  0.2635336   0.0507  0.959527
## Inactifs:infbac      0.1016392  0.1335026   0.7613  0.446462
## Chômeurs:supbac     -0.6422571  0.3169754  -2.0262  0.042744 *
## Inactifs:supbac     -1.1914041  0.1592108  -7.4832  7.261e-14 ***
## Chômeurs:femme      -0.1577550  0.2115544  -0.7457  0.455852
## Inactifs:femme       0.2857907  0.0991584   2.8822  0.003950 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -1574.5
## McFadden R^2: 0.13076
## Likelihood ratio test : chisq = 473.73 (p.value = < 2.22e-16)
```

94 / 109

Données polytomique non-ordonnées

Exemple : Position sur le marché du travail

```
# Calcul des effets marginaux
margins_mlogit(m_multi)
## $infbac
##          P(acteu = Actifs occupés) P(acteu = Chômeurs) P(acteu = Inactifs)
## infbac = 0          0.43258112      0.084195812      0.48322307
## infbac = 1          0.41092861      0.085217737      0.50385366
## Diff              -0.02165252      0.001021926      0.02063059
##
## $supbac
##          P(acteu = Actifs occupés) P(acteu = Chômeurs) P(acteu = Inactifs)
## supbac = 0          0.3510642      0.09502313      0.5539126
## supbac = 1          0.6407880      0.05261860      0.3065934
## Diff              0.2897237      -0.04240454      -0.2473192
##
## $femme
##          P(acteu = Actifs occupés) P(acteu = Chômeurs) P(acteu = Inactifs)
## femme = 0          0.44332551      0.09115675      0.46551775
## femme = 1          0.39781042      0.07905990      0.52312968
## Diff              -0.04551509      -0.01209685      0.05761194
```



95 / 109

Données polytomique non-ordonnées

Exemple : Position sur le marché du travail

```
# Test de Hausman-MacFadden
# 1) On réestime le modèle
# en excluant l'alternative 3 (inactivité)
m_multi12 <- mlogit(
  acteu ~ 0 | age + infbac + supbac + femme
  , data = e_long[e_long$ACTEU != "3", ]
)
# 2) On mène le test
hmftest(m_multi, m_multi12)
##
## Hausman-McFadden test
##
## data: e_long
## chisq = 27.893, df = 10, p-value = 0.001878
## alternative hypothesis: IIA is rejected
```



96 / 109

Données particulièrement asymétriques

Définition et exemples

Des données sont considérées comme particulièrement asymétriques quand un **modèle polynomial** en les variables explicatives **ne suffit pas** pour modéliser correctement les données.

Plus spécifiquement, ce type de données est caractérisé par une très forte **hétéroscédasticité** : leur degré de variabilité n'est pas constant mais s'intensifie avec leur valeur.

Exemples

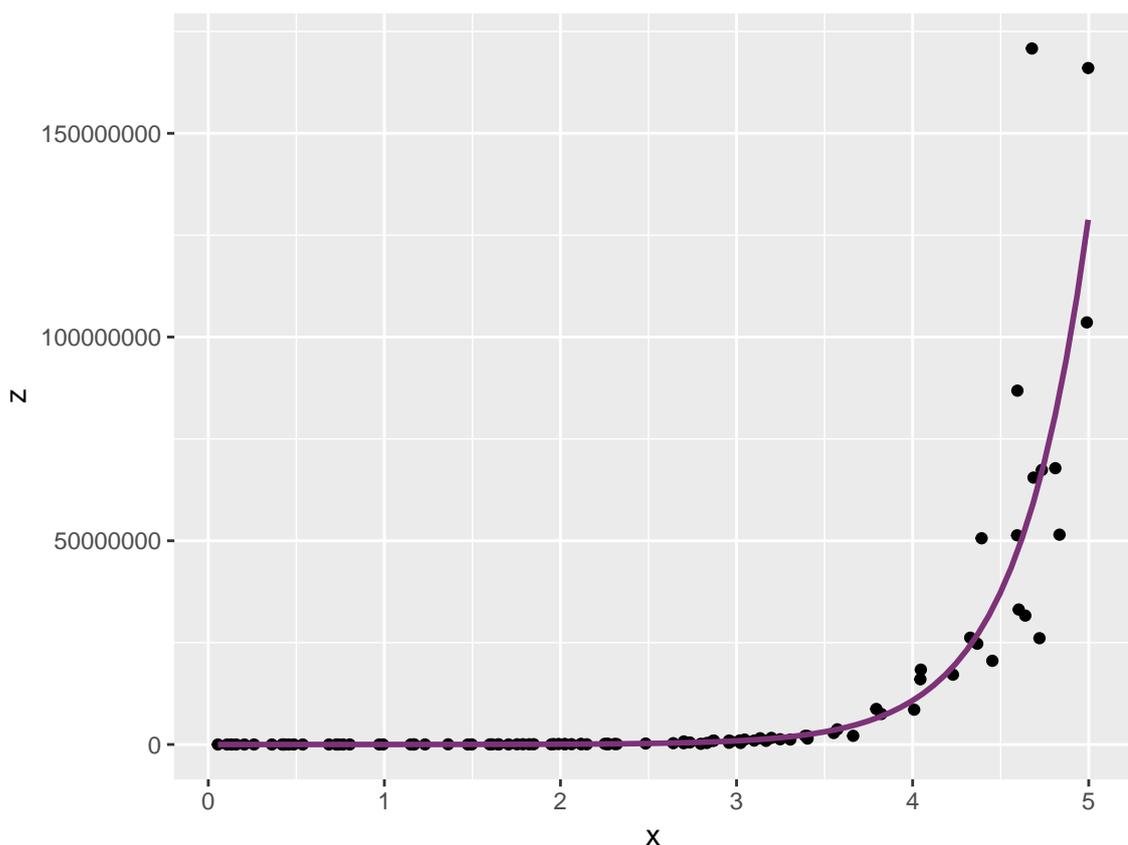
- ▶ données continues : actifs financiers, patrimoine, valeur boursière, etc.
- ▶ données discrètes : données de comptage affectées par un effet taille (par exemple nombre d'accidents du travail par entreprise, etc.)



97 / 109

Données particulièrement asymétriques

Illustration : Distribution d'un actif financier



98 / 109

Données particulièrement asymétriques

Principes de modélisation

Les modèles à mettre en œuvre diffèrent selon que la variable d'intérêt est continue ou discrète.

- ▶ **régression gamma** pour les données continues et positives ;
- ▶ **régression de Poisson** ou négatives-binomiales (éventuellement inflatées en zéro) pour les données discrètes.

On ne développe ici que le cas de la **régression gamma**.

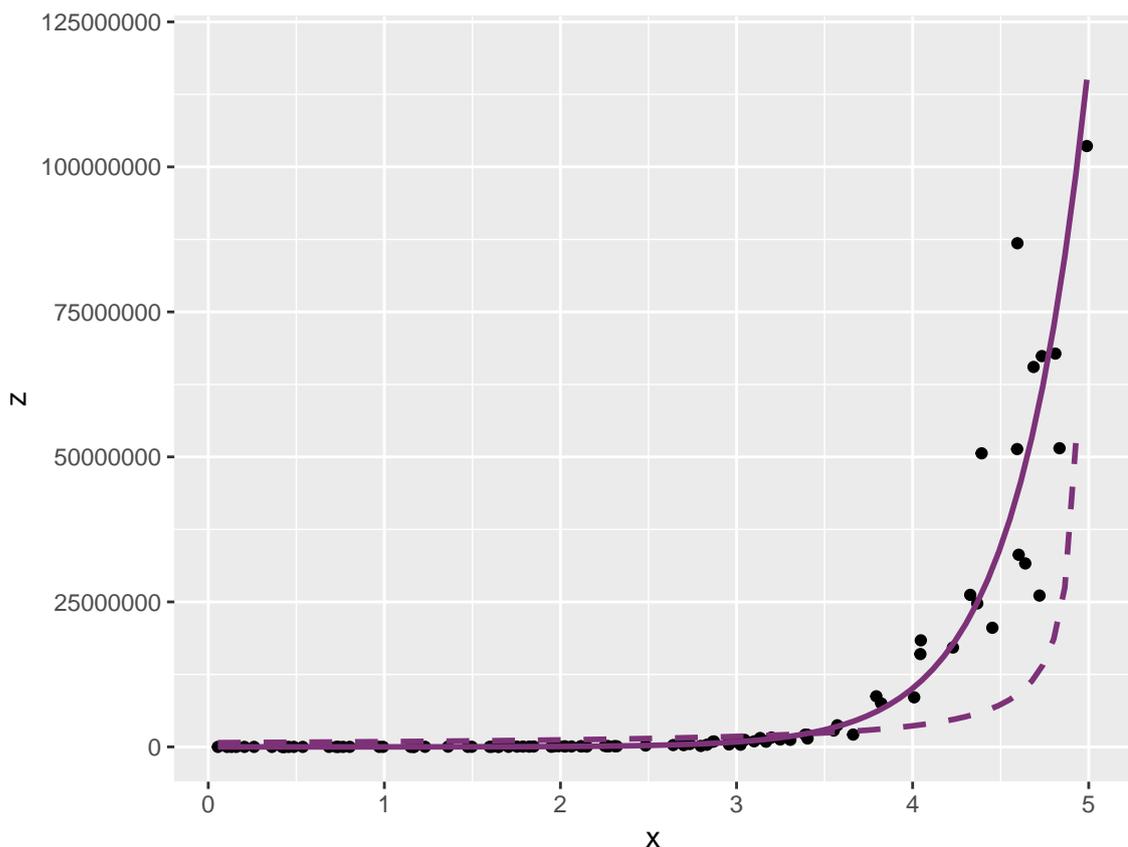
La régression gamma est une spécification du modèle linéaire général où :

- ▶ la variable dépendante est supposée suivre une **loi Gamma** ;
- ▶ la **fonction de lien** est soit la **fonction inverse**, soit la **fonction logarithme**.

99 / 109

Données particulièrement asymétriques

Illustration : Distribution d'un actif financier



100 / 109

Données particulièrement asymétriques

Interprétation du modèle

Quand la fonction de lien utilisée est le **logarithme**, il est possible d'**interpréter directement** les coefficients du modèle.

En effet, on peut montrer que la valeur du coefficient β_k correspond à l'**augmentation moyenne en pourcentage** associée à une augmentation de 1 unité de la variable x_k .

Exemple Si dans un modèle de régression gamma sur le patrimoine $\hat{\beta}_{age} = 0,005$, alors cela signifie que toutes les autres variables du modèle égales par ailleurs, en moyenne dans l'échantillon à une année supplémentaire est associé un patrimoine supérieur de l'ordre de 0,5 %.

Il est bien entendu également possible de calculer des **effets marginaux moyens**.

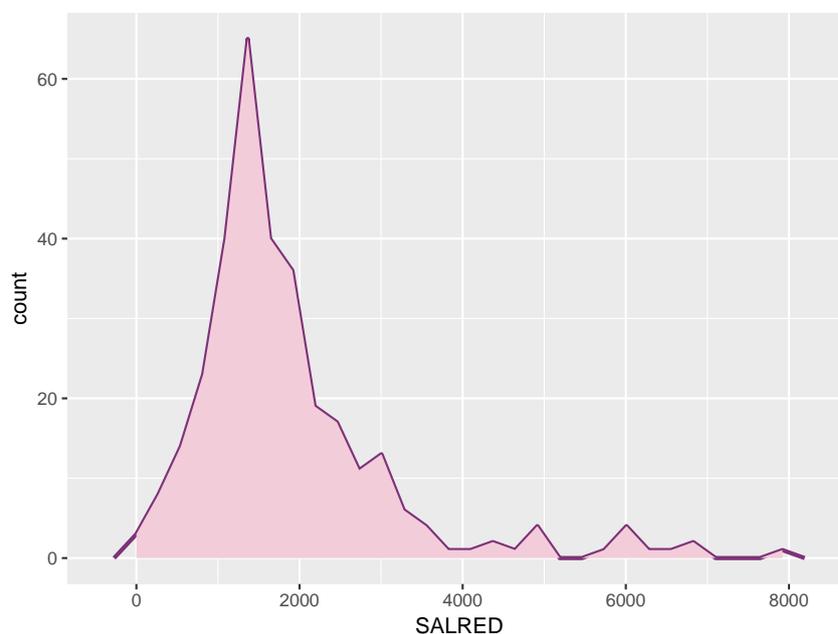


101 / 109

Données particulièrement asymétriques

Exemple : Salaire mensuel

Le salaire est une variable en général particulièrement asymétrique, et c'est le cas (dans une certaine mesure) dans l'échantillon considéré.



102 / 109

Données particulièrement asymétriques

Exemple : Salaire mensuel

```
# Création d'une variable supplémentaire
# sur le temps partiel
e$tpp <- (e$TPP == "2") * 1

# Estimation d'un modèle assez complet
m_gamma <- glm(
  SALRED ~ age + I(age^2) + infbac + supbac + femme + tpp
  , data = e, family = Gamma(link = "log")
)

# Paramètres
summary(m_gamma)$coefficients
```

##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	6.0274757394	0.3108794892	19.388464	1.866570e-55
##	age	0.0673411073	0.0160294721	4.201081	3.474124e-05
##	I(age^2)	-0.0006439487	0.0001943887	-3.312686	1.033012e-03
##	infbac	-0.2198516241	0.0799609106	-2.749489	6.317982e-03
##	supbac	0.3213982761	0.0814504063	3.945938	9.830986e-05
##	femme	-0.1596633305	0.0595913812	-2.679302	7.770213e-03
##	tpp	-0.5634592264	0.0764008866	-7.375035	1.501731e-12

103 / 109

Données particulièrement asymétriques

Exemple : Salaire mensuel

```
# Effets marginaux
margins(m_gamma)
```

##	\$infbac	Average MFX	Std. Error	z value	P> z
##	infbac = 0	2007.7649	NA	NA	NA
##	infbac = 1	1611.5082	NA	NA	NA
##	Diff	-396.2567	142.3548	-2.783585	0.005376174
##	\$supbac	Average MFX	Std. Error	z value	P> z
##	supbac = 0	1606.8699	NA	NA	NA
##	supbac = 1	2215.9615	NA	NA	NA
##	Diff	609.0916	162.004	3.759733	0.000170095
##	\$femme	Average MFX	Std. Error	z value	P> z
##	femme = 0	2004.6631	NA	NA	NA
##	femme = 1	1708.8365	NA	NA	NA
##	Diff	-295.8267	111.7223	-2.647874	0.008099959
##	\$tpp	Average MFX	Std. Error	z value	P> z
##	tpp = 0	1991.3790	NA	NA	NA
##	tpp = 1	1133.5657	NA	NA	NA
##	Diff	-857.8133	101.0009	-8.493127	2.011508e-17

CEPE

104 / 109

Compléments : Données polytomiques ordonnées

Définition et exemples

Les données polytomiques ordonnées correspondent à une information qui n'est **pas quantitative** mais qui présente néanmoins un **ordre naturel**.

Exemples

- ▶ **appréciation** : un peu, beaucoup, passionnément, à la folie.
- ▶ **fréquences** : jamais, rarement, parfois, souvent, tous les jours

Même si souvent ces variables sont codées par des nombres (1 pour « un peu », 2 pour « beaucoup », etc.), ceux-ci sont conventionnels et n'ont **aucune signification** (sinon leur ordre).

On ne peut **pas intégrer ce type de données dans un modèle linéaire classique**.



105 / 109

Compléments : Données polytomiques ordonnées

Principes de modélisation

On fait l'hypothèse que la variable qualitative ordonnée modélisée Y est obtenue à partir d'une **variable latente** continue et d'un **ensemble de seuils**.

C'est cette hypothèse qui guide la dérivation de la log-vraisemblance et donc l'estimation du modèle.

Interprétation Comme dans le modèle logistique dichotomique, la quantité modélisée est une **probabilité** et la fonction de lien la **fonction logit**.

Néanmoins, il ne s'agit pas de la probabilité d'une valeur particulière de Y mais de **toutes les valeurs de Y supérieures à la valeur considérée**.

Pour aller plus loin Le [site](#) de la bibliothèque de l'Université de Virginie et le [blog](#) `doingbayesiandataanalysis`.



106 / 109

Compléments : Données polytomiques ordonnées

Exemple : Intensité du temps partiel

```
# Restriction aux individus pour lesquels
# la question du temps partiel fait sens
e <- e[e$TPP %in% c("1", "2"), ]

# Définition de la variable TXTPPB comme facteur ordonné
e$TXTPPB[e$tp == 0] <- "0"
e$TXTPPB <- factor(e$TXTPPB, ordered = TRUE)

# Création de variables supplémentaires relatives aux enfants
e <- within(e, {
  nbenf1 <- (NBAGENF %in% c("1", "2", "3")) * 1
  nbenf2 <- (NBAGENF %in% c("4", "5", "6")) * 1
  nbenf3p <- (NBAGENF %in% c("7", "8", "9")) * 1
  enf3ans <- (NBAGENF %in% c("3", "6", "9")) * 1
})

# Chargement du package MASS et estimation du modèle avec polr()
library(MASS)
m_ordered <- polr(
  TXTPPB ~ age + femme + infbac + supbac +
    enf3ans + nbenf1 + nbenf2 + nbenf3p
  , data = e, Hess=TRUE
)
```

107 / 109

Compléments : Données polytomiques ordonnées

Exemple : Intensité du temps partiel

```
summary(m_ordered)
## Call:
## polr(formula = TXTPPB ~ age + femme + infbac + supbac + enf3ans +
##       nbenf1 + nbenf2 + nbenf3p, data = e, Hess = TRUE)
##
## Coefficients:
##           Value Std. Error t value
## age         0.0006941  0.007624  0.09105
## femme       1.7191963  0.199227  8.62933
## infbac      0.0038150  0.234330  0.01628
## supbac     -0.2312823  0.242264 -0.95467
## enf3ans     0.1257900  0.334713  0.37581
## nbenf1     -0.1979754  0.242205 -0.81739
## nbenf2      0.2460879  0.238670  1.03108
## nbenf3p     0.5545493  0.319711  1.73454
##
## Intercepts:
##      Value Std. Error t value
## 0|1  2.5619  0.4061  6.3092
## 1|2  2.9605  0.4094  7.2311
## 2|3  3.3563  0.4136  8.1152
## 3|4  4.0924  0.4263  9.5997
## 4|5  5.0838  0.4621 11.0025
##
## Residual Deviance: 1436.77
## AIC: 1462.77
```

CEPE

108 / 109

Compléments : Données polytomiques ordonnées

Exemple : Intensité du temps partiel

```
# Calcul des p-valeurs
pnorm(abs(coef(summary(m_ordered))[, "t value"]), lower.tail = FALSE) * 2
##          age          femme          infbac          supbac          enf3ans
## 9.274540e-01 6.171324e-18 9.870107e-01 3.397450e-01 7.070545e-01
##          nbenf1          nbenf2          nbenf3p          0|1          1|2
## 4.137061e-01 3.025038e-01 8.282319e-02 2.804323e-10 4.790431e-13
##          2|3          3|4          4|5
## 4.848627e-16 8.019502e-22 3.716418e-28

# Calcul des effets marginaux
margins_pplr(m_ordered)[c("femme", "enf3ans")]
## $femme
##          P(TXTPPB = 0) P(TXTPPB = 1) P(TXTPPB = 2) P(TXTPPB = 3)
## femme = 0          0.9264626      0.02288421      0.01595055      0.01775698
## femme = 1          0.6957767      0.07658370      0.06161504      0.07861309
## Diff              -0.2306859      0.05369948      0.04566449      0.06085612
##          P(TXTPPB = 4) P(TXTPPB = 5)
## femme = 0          0.01058629      0.006359394
## femme = 1          0.05298674      0.034424721
## Diff              0.04240045      0.028065327
##
## $enf3ans
##          P(TXTPPB = 0) P(TXTPPB = 1) P(TXTPPB = 2) P(TXTPPB = 3)
## enf3ans = 0          0.81283484      0.049617386      0.038544695      0.047678797
## enf3ans = 1          0.79485656      0.053140211      0.041851939      0.052538494
## Diff              -0.01797827      0.003522826      0.003307244      0.004859697
##          P(TXTPPB = 4) P(TXTPPB = 5)
## enf3ans = 0          0.031307289      0.020016997
## enf3ans = 1          0.035004742      0.022608049
## Diff              0.003697453      0.002591052
```

