Régression sur données non-linéaires

Certificat Chargé d'études statistiques

1-3 mars 2017

Martin Chevalier
Insee

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 SAS** au cœur du module : nombreux exemples dans le support, exercices corrigés.

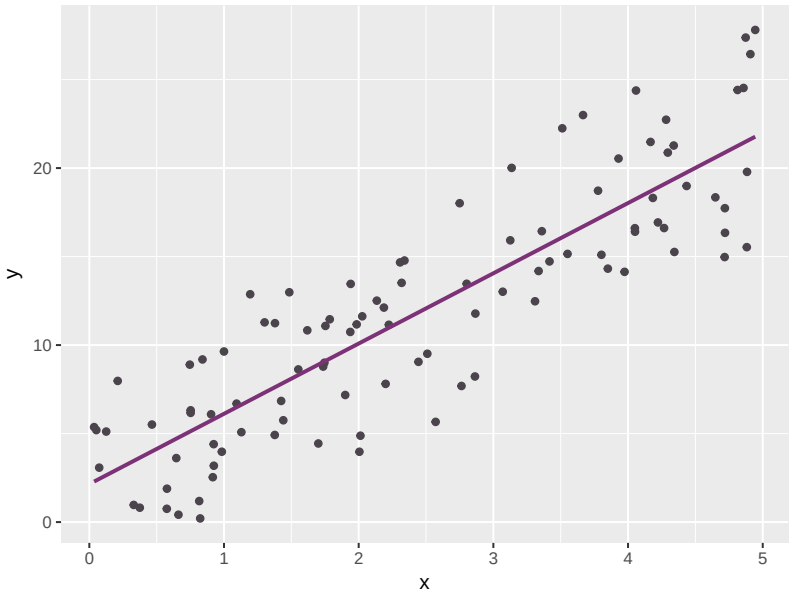
Introduction : Modéliser des données non-linéaires

Régression linéaire classique

Variable expliquée Y Quantitative

Variable(s) explicative(s) X

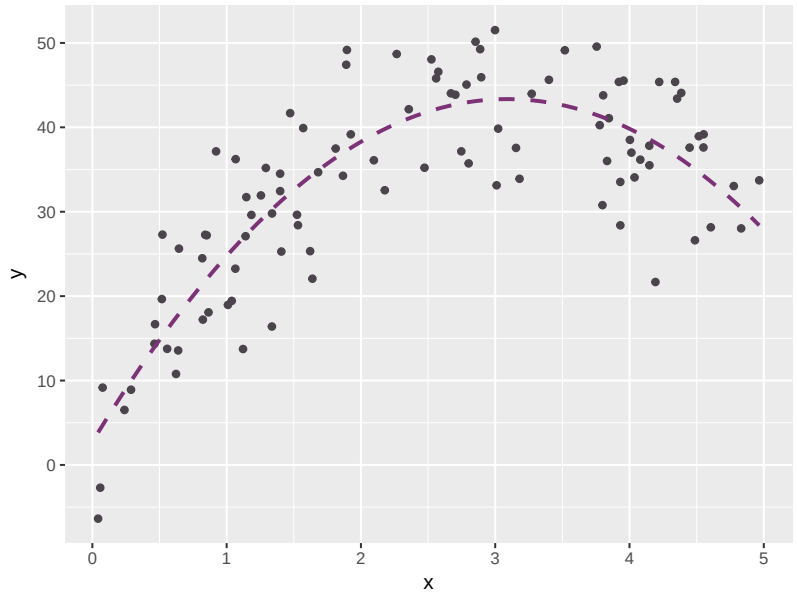
- quantitatives ou qualitatives ;
- en relation linéaire avec la variable expliquée.



Introduction : Modéliser des données non-linéaires

Régression linéaire classique

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

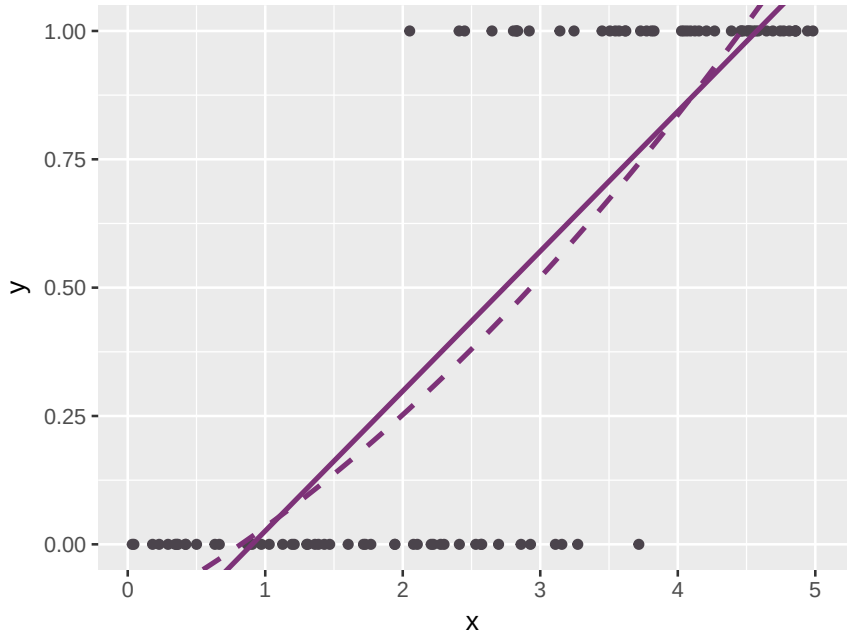


$$y_i = \beta_0 + \beta_1 \times x_i + \beta_2 \times x_i^2 + \varepsilon_i$$

Introduction : Modéliser des données non-linéaires

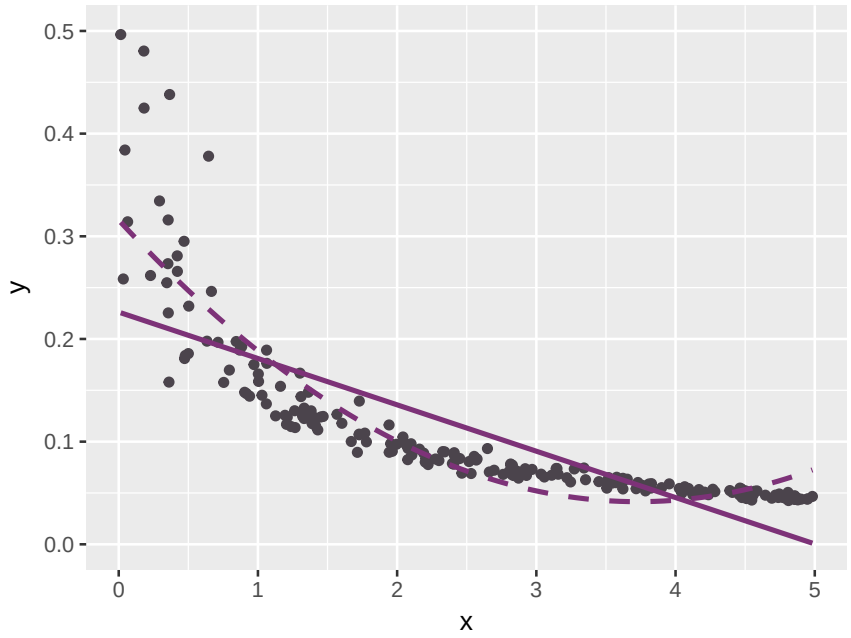
Limites de la régression linéaire classique

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



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.



Solution : Généraliser le modèle linéaire

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

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

Linéaire ?

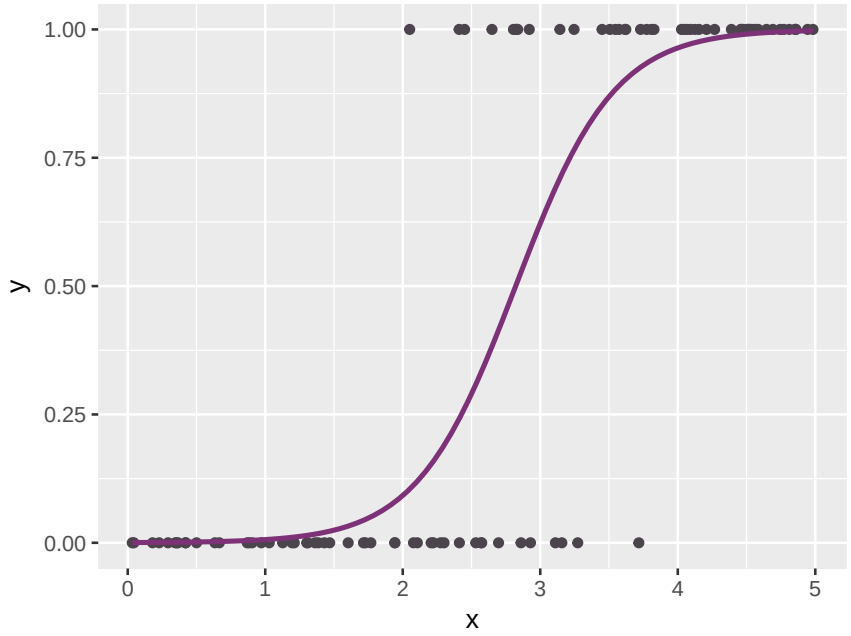
- Ce modèle est « **linéaire en ses coefficients** » : les opérations entre les X et β sont uniquement des **sommes ou des multiplications**...
- ... mais il peut modéliser des relations *non-linéaires* grâce notamment à la **fonction de lien** f.

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

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

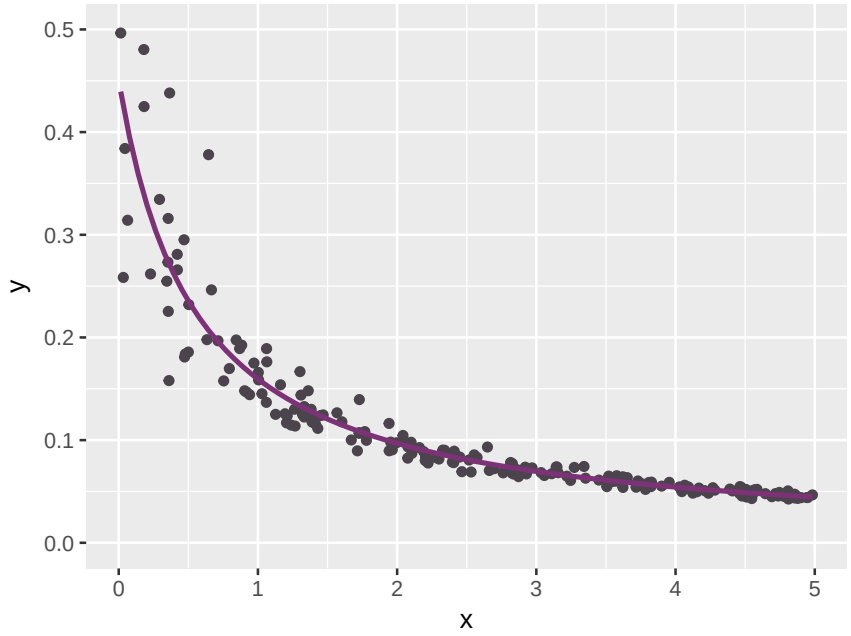
Introduction : Modéliser des données non-linéaires
Exemple : Régression logistique dichotomique

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



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).



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 asymétriques, modèles pour données polytomiques.

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

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`

Tutoriels de UCLA
`http://www.ats.ucla.edu/stat/dae/`

Estimer un modèle logistique dichotomique

Modèle linéaire général

Formulation du modèle linéaire général

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

- Y la variable expliquée ;
- $X = (1\ X_1 \cdots X_p)$ la matrice de variables explicatives ;
- ε le résidu ;
- β le vecteur de coefficients à estimer ;
- $f()$ la **fonction de lien**.

Objectifs de l'estimation

1. Trouver les paramètres $\beta_0, \ldots, \beta_p$ qui maximisent l'ajustement du modèle aux données.
2. Pouvoir quantifier la qualité de cet ajustement et ses conséquences sur l'estimation en termes d'inférence.

Principe d'estimation

Contrairement au modèle linéaire classique, il n'existe **aucune formule qui donne directement** la valeur de $\hat{\beta}$.

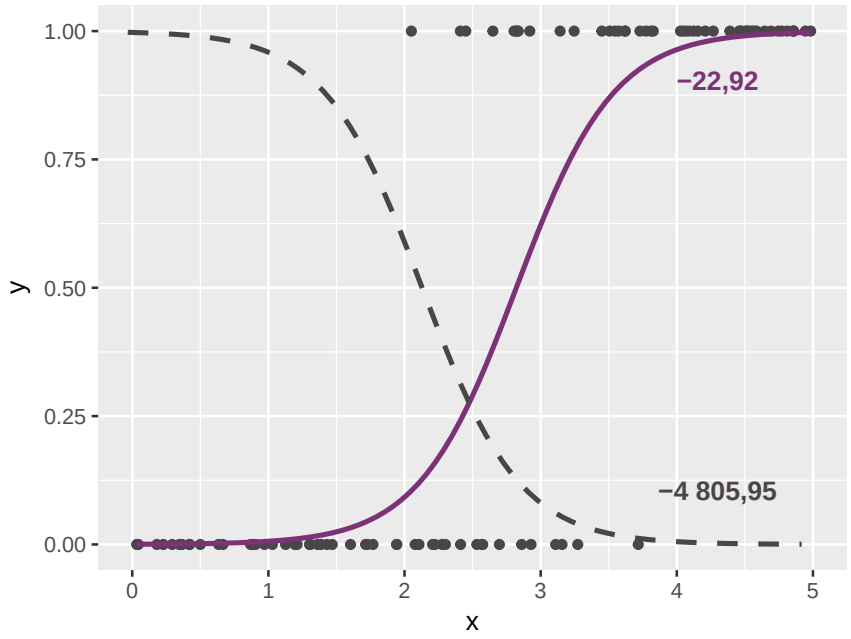
On utilise donc des **algorithmes d'optimisation** pour maximiser une certaine **fonction objectif**, la **(log-)vraisemblance** du modèle.

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.

Estimation par maximum de vraisemblance
Vraisemblance et ajustement aux données

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



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

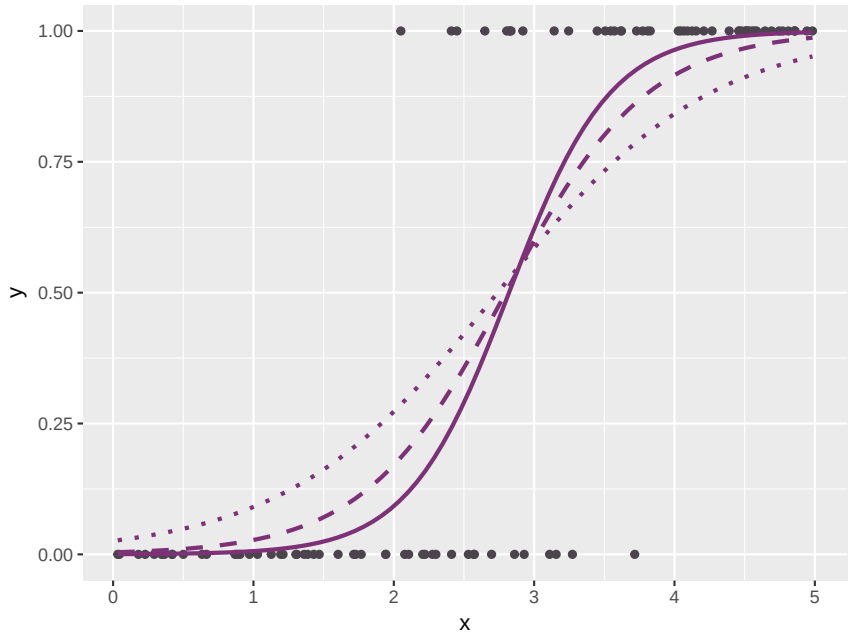
Itération 3 $\ell_n =$ -23,04 $\beta_0 =$ -7,05 $\beta_1 =$ 2,50

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**

Algorithme itératif

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



Itérations 1, 2 et 5

Régression logistique dichotomique

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

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

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* :

$$\textit{logit}(q) = \ln\left(\frac{q}{1-q}\right)$$

Complément Dérivation de la vraisemblance du modèle logistique dichotomique.

Parenthèse : La fonction *logit*

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

1. $\dfrac{q}{1-q}$ est la **cote** associée à la proportion q (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. $logit(q) = \ln\left(\dfrac{q}{1-q}\right)$ est donc bien à valeurs dans $\mathbb{R}$.

Probabilités prédites par le modèle

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

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

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

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 :

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

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

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.

Exemple : Probabilité d'être en emploi stable
Mesure de la stabilité de l'emploi dans l'EEC

```
/*Import et travail sur les données*/
DATA ee;
    SET d.ee_ces;
    agenum = INPUT(AGE, 2.);
    IF acteu = "1" AND agenum < 70;
    stable = (contra = '1' OR (chpub IN("1",
       "2", "3") AND titc = "2"));
RUN;
/*Statistique descriptive sur stable*/
PROC FREQ DATA = ee;
    TABLES stable;
RUN;
```

stable	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	257	25.50	257	25.50
1	751	74.50	1008	100.00

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 ?

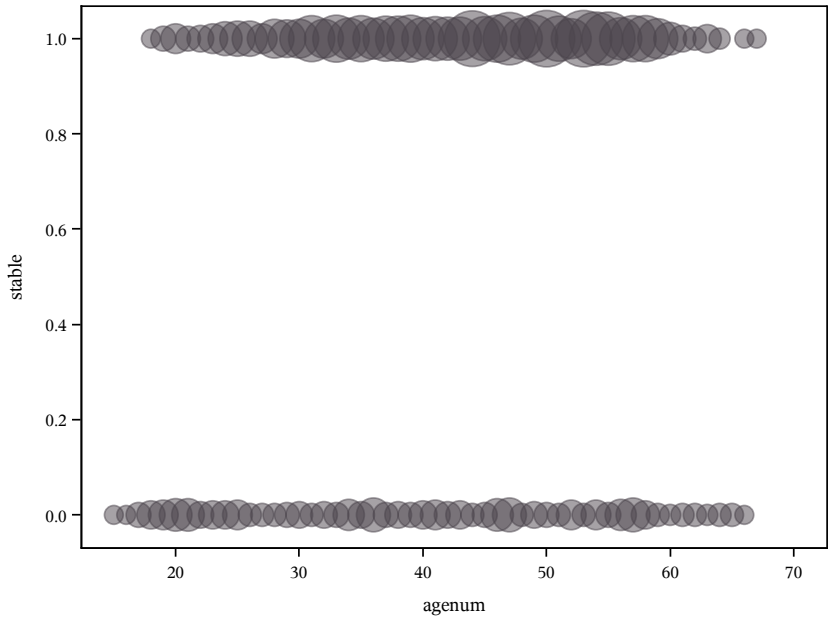
On souhaite autant que possible prendre également en compte le **secteur d'activité agrégé** de l'entreprise (primaire, industrie, construction ou tertiaire).

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** :

$$logit\left[\mathbb{P}(stable_i = 1 | age_i)\right] = \beta_0 + \beta_1 \times age_i + \varepsilon_i$$



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

```
/*Agrégation par stable x agenum*/
PROC SQL;
    CREATE TABLE stable_age AS
        SELECT stable, agenum, COUNT(*) AS n
        FROM ee GROUP BY stable, agenum
    ;
QUIT;

/*Construction du bubble plot*/
PROC SGPLOT DATA = stable_age;
  BUBBLE X = agenum Y = stable SIZE = n /
     TRANSPARENCY = 0.50;
RUN;
```


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

```
/*Régression logistique de stable sur agenum*/
PROC LOGISTIC DATA = ee;
    MODEL stable = agenum;
RUN;
```

Model Information	
Data Set	WORK.EE
Response Variable	stable
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile		
Ordered Value	**stable**	**Total Frequency**
1	0	257
2	1	751

Probability modeled is stable=0.

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

```
/*Passage de 1 comme modalité modélisée*/
PROC LOGISTIC DATA = ee;
    MODEL stable (DESC) = agenum;
RUN;
```

Model Information	
Data Set	WORK.EE
Response Variable	stable
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile		
Ordered Value	stable	Total Frequency
1	1	751
2	0	257

Probability modeled is stable=1.

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

Coefficients estimés

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0706	0.2602	0.0736	0.7861
agenum	1	0.0276	0.00617	20.0342	<.0001

Intervalles de confiance à 95 %

```
PROC LOGISTIC DATA = ee;
    MODEL stable (DESC) = agenum /  CLPARM = WALD;
RUN;
```

Parameter Estimates and Wald Confidence Intervals			
Parameter	Estimate	95% Confidence Limits	
Intercept	-0.0706	-0.5806	0.4394
agenum	0.0276	0.0155	0.0397

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

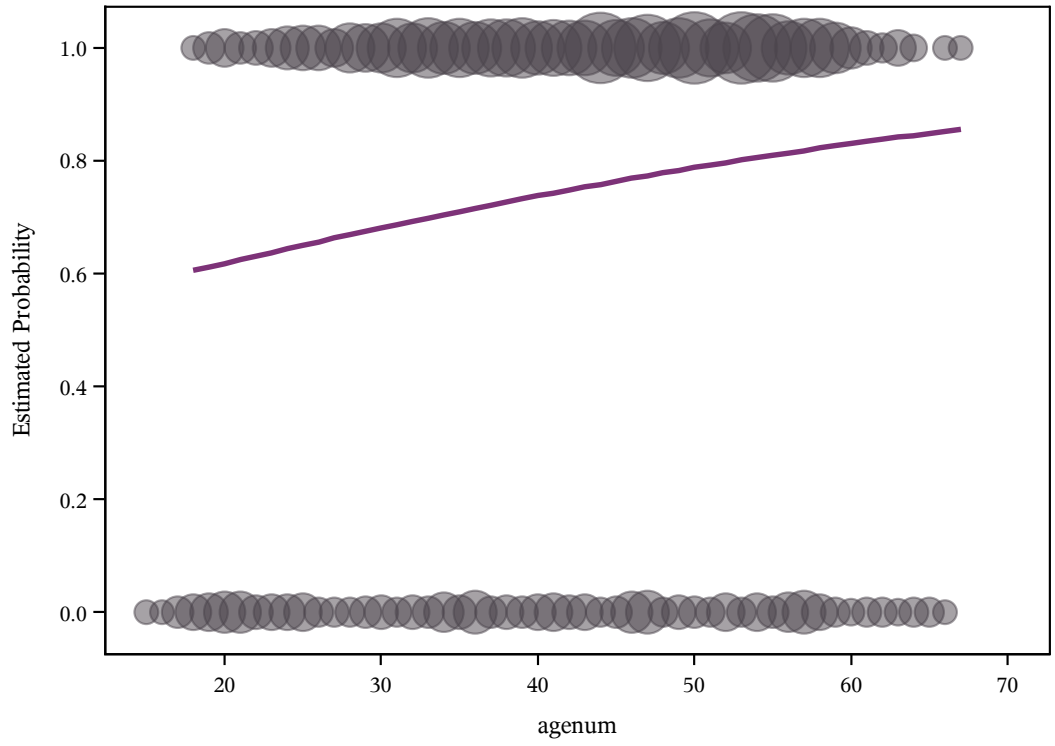
Probabilités prédites par le modèle

```
PROC LOGISTIC DATA = ee;
     MODEL stable (DESC) = agenum;
     OUTPUT OUT = ee P = prob_m1;
RUN;
PROC MEANS DATA = ee;
     VAR prob_m1;
RUN;
```

Analysis Variable : prob_m1 Estimated Probability				
N	Mean	Std Dev	Minimum	Maximum
1008	0.7450116	0.0622335	0.5850658	0.8556368

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

Probabilités prédites par le modèle



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

Probabilités prédites par le modèle

```
/*Agrégation par stable x agenum x prob_m1*/
PROC SQL;
    CREATE TABLE stable_age_prob AS
        SELECT stable, agenum, prob_m1, COUNT(*) AS n
        FROM ee GROUP BY stable, agenum, prob_m1
    ;
QUIT;

/*Suppression des valeurs redondantes de prob_m1*/
DATA stable_age_prob;
    SET stable_age_prob;
    IF stable = 0 THEN prob_m1 = .;
RUN;

/*Création du graphique*/
PROC SGPLOT DATA = stable_age_prob;
  BUBBLE X = agenum Y = stable SIZE = n / TRANSPARENCY =
      0.50;
  SERIES X = agenum Y = prob_m1;
RUN;
```


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 $20,03 > 1,96$ 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

Relation entre âge et stabilité du contrat

Statistiques d'ajustement

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1146.522	1128.193
SC	1151.438	1138.025
-2 Log L	1144.522	1124.193

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	20.3287	1	<.0001
Score	20.4072	1	<.0001
Wald	20.0342	1	<.0001

Exemple : Probabilité d'être en emploi stable

Relation entre âge et stabilité du contrat

Statistiques d'ajustement

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	57.1	Somers' D	0.162
Percent Discordant	40.9	Gamma	0.166
Percent Tied	2.1	Tau-a	0.062
Pairs	193007	c	0.581

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 manuelle de la variable
    sexe*/
DATA ee;
  SET ee;
  homme = (sexe = "1");
  femme = (sexe = "2");
RUN;

/*Régression de stable sur les indicatrices
    femme et homme*/
PROC LOGISTIC DATA = ee;
  MODEL stable (DESC) = agenum femme;
RUN;
PROC LOGISTIC DATA = ee;
  MODEL stable (DESC) = agenum homme;
RUN;
```

Exemple : Probabilité d'être en emploi stable
Dichotomisation des variables qualitatives

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-0.2352	0.2679	0.7706	0.3800
agenum	1	0.0270	0.00618	19.0995	<.0001
femme	1	0.3944	0.1472	7.1781	0.0074

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.1592	0.2747	0.3358	0.5623
agenum	1	0.0270	0.00618	19.0995	<.0001
homme	1	-0.3944	0.1472	7.1781	0.0074

Exemple : Probabilité d'être en emploi stable
Dichotomisation des variables qualitatives

```
/*Intégration des deux variables
    simultanément*/
PROC LOGISTIC DATA = ee;
  MODEL stable (DESC) = agenum femme homme;
RUN;
```

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

homme =	Intercept - femme

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-0.2352	0.2679	0.7706	0.3800
agenum	1	0.0270	0.00618	19.0995	<.0001
femme	1	0.3944	0.1472	7.1781	0.0074
homme	0	0	.	.	.

Exemple : Probabilité d'être en emploi stable
Dichotomisation des variables qualitatives

```
/*Utilisation de l'instruction CLASS*/
PROC LOGISTIC DATA = ee;
  CLASS sexe (REF = "1") / PARAM = REF;
  MODEL stable (DESC) = agenum sexe;
RUN;
```

Class Level Information		
Class	Value	Design Variables
SEXE	1	0
	2	1

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2352	0.2679	0.7706	0.3800
agenum		1	0.0270	0.00618	19.0995	<.0001
SEXE	2	1	0.3944	0.1472	7.1781	0.0074

Exemple : Probabilité d'être en emploi stable
Dichotomisation des variables qualitatives

```
/*NE PAS OUBLIER PARAM = REF !*/
PROC LOGISTIC DATA = ee;
  CLASS sexe (REF = "1");
  MODEL stable (DESC) = agenum sexe;
RUN;
```

Class Level Information		
Class	Value	Design Variables
SEXE	1	-1
	2	1

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.0380	0.2612	0.0212	0.8843
agenum		1	0.0270	0.00618	19.0995	<.0001
SEXE	2	1	0.1972	0.0736	7.1781	0.0074

Exemple : Probabilité d'être en emploi stable

Dichotomisation des variables qualitatives

```
/*Recodage du diplôme en trois modalités*/
DATA ee;
  SET ee;
  LENGTH dipl3 $ 20;
  IF ddipl IN("1", "3") THEN dipl3 = "Supérieur au bac";
  IF ddipl = "4" THEN dipl3 = "Bac";
  IF ddipl IN("5", "6", "7") THEN dipl3 = "Inférieur au
      bac";
RUN;

/*Impact du choix de la modalité de référence*/
PROC LOGISTIC DATA = ee;
  CLASS dipl3 (REF = "Inférieur au bac") / PARAM = REF;
  MODEL stable (DESC)  = dipl3;
RUN;
PROC LOGISTIC DATA = ee;
  CLASS dipl3 (REF = "Bac") / PARAM = REF;
  MODEL stable (DESC)  = dipl3;
RUN;
PROC LOGISTIC DATA = ee;
  CLASS dipl3 (REF = "Supérieur au bac") / PARAM = REF;
  MODEL stable (DESC) = dipl3;
RUN;
```


Exemple : Probabilité d'être en emploi stable

Dichotomisation des variables qualitatives

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.9973	0.1034	93.0020	<.0001
dipl3	Bac	1	-0.1078	0.1932	0.3114	0.5768
dipl3	Supérieur au bac	1	0.3019	0.1663	3.2957	0.0695

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.8895	0.1632	29.7231	<.0001
dipl3	Inférieur au bac	1	0.1078	0.1932	0.3114	0.5768
dipl3	Supérieur au bac	1	0.4097	0.2088	3.8515	0.0497

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.2992	0.1303	99.4752	<.0001
dipl3	Bac	1	-0.4097	0.2088	3.8515	0.0497
dipl3	Inférieur au bac	1	-0.3019	0.1663	3.2957	0.0695

Exemple : Probabilité d'être en emploi stable

Dichotomisation des variables qualitatives

Significativité jointe des coefficients

Comme en analyse de variancen, SAS teste la significativité jointe de l'**ensemble des paramètres associés à une même variable**.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
dipl3	2	4.8149	0.0900

Exemple : Probabilité d'être en emploi stable

Estimation du modèle complet

Formulation du modèle

$$stable_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \beta_3 infbac_i + \beta_4 supbac_i + \beta_5 agri_i + \beta_6 cons_i + \beta_7 indus_i + \varepsilon_i$$

```
/*Recodage de la variable de secteur*/
DATA ee;
  SET ee;
  IF NAFG4N = "ES" THEN secteur = "Agriculture";
  IF NAFG4N = "ET" THEN secteur = "Industrie";
  IF NAFG4N = "EU" THEN secteur = "Construction";
  IF NAFG4N = "EV" THEN secteur = "Tertiaire";
RUN;

/*Estimation du modèle complet*/
PROC LOGISTIC DATA = ee;
  CLASS sexe (REF = "1") dipl3 (REF = "Bac") secteur (REF
      = "Tertiaire") / PARAM = REF;
  MODEL stable (DESC)  = agenum sexe dipl3 secteur;
RUN;
```

Estimation du modèle complet

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.3507	0.3119	1.2643	0.2608
agenum		1	0.0308	0.00660	21.8009	<.0001
SEXE	**2**	1	0.3439	0.1619	4.5095	0.0337
dipl3	**Inférieur au bac**	1	-0.1159	0.2112	0.3011	0.5832
dipl3	**Supérieur au bac**	1	0.1861	0.2225	0.6990	0.4031
secteur	**Agriculture**	1	-2.4070	0.4197	32.8967	<.0001
secteur	**Construction**	1	-0.0279	0.2804	0.0099	0.9207
secteur	**Industrie**	1	0.8605	0.2750	9.7939	0.0018

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 modalité « Agriculture » est négatif.
2. On interprète alors : « À âge, sexe et diplômes **égaux par ailleurs**, le fait de travailler dans le secteur agricole est associé à une **probabilité plus faible** d'être en emploi stable **par rapport** aux salariés du secteur tertiaire (modalité de référence) ».

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 modalité « Femme » est inférieur à celui associé à la modalité « Industrie ».
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 l'industrie **plutôt que** dans le secteur tertiaire (modalité de référence) ».

En guise de conclusion

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

Son estimation est effectuée par **maximum de vraisemblance** en utilisant la `PROC LOGISTIC` de SAS.

Cette méthodologie fournit l'ensemble des éléments présents en régression linéaire classique, **coefficients** $\beta_0, \ldots, \beta_p$ et **erreurs standard** notamment.

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

Le modèle probit dichotomique

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

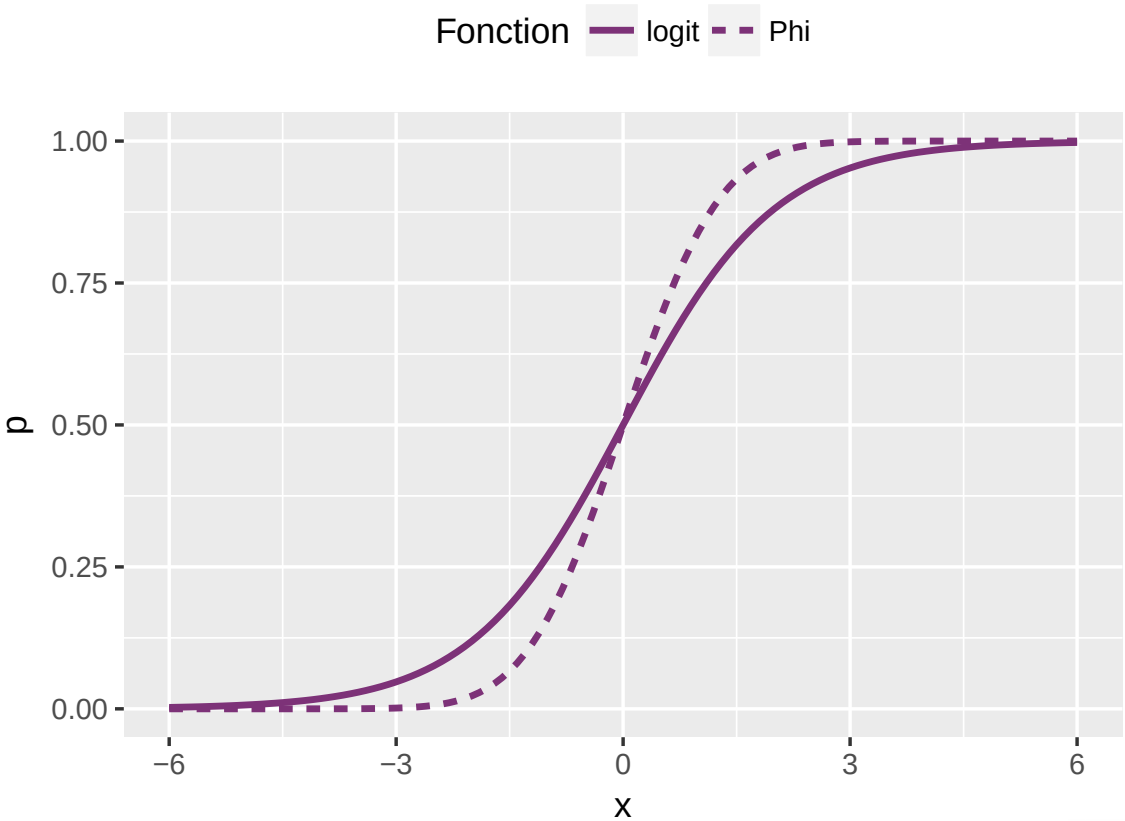
Dans le modèle probit dichotomique, f est telle que

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

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.

Le modèle probit dichotomique



Le modèle probit dichotomique

```
PROC LOGISTIC DATA = ee;
  CLASS sexe (REF = "1") dipl3 (REF = "Bac") secteur (REF
      = "Tertiaire") / PARAM = REF;
  MODEL stable (DESC)  = agenum sexe dipl3 secteur / LINK
      = PROBIT;
RUN;
```

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.1590	0.1853	0.7356	0.3911
agenum		1	0.0173	0.00384	20.3101	<.0001
SEXE	2	1	0.2087	0.0943	4.9028	0.0268
dipl3	Inférieur au bac	1	-0.0572	0.1240	0.2129	0.6445
dipl3	Supérieur au bac	1	0.1081	0.1297	0.6941	0.4048
secteur	Agriculture	1	-1.4515	0.2421	35.9499	<.0001
secteur	Construction	1	-0.0214	0.1681	0.0162	0.8987
secteur	Industrie	1	0.4969	0.1504	10.9138	0.0010

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 le fonctionnement du logiciel** lorsqu'il effectue l'estimation.

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

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

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

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 :

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

Pour faciliter les manipulations et le fonctionnement des algorithmes, on travaille en général sur la **log-vraisemblance** ℓ_n :

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

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

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 :

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

Vraisemblance du modèle logistique dichotomique

La **log-vraisemblance d'un modèle dichotomique** s'écrit ainsi :

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

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.

Vraisemblance du modèle logistique dichotomique

Le modèle logistique dichotomique est le modèle tel que :

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

Ainsi

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

et donc

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

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

Interpréter un modèle logistique dichotomique

De l'importance de l'interprétation

L'interprétation d'un modèle renvoie à deux 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.

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, le **modèle logistique complet** estimé à la partie précédente est :

$$stable_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \beta_3 infbac_i + \beta_4 supbac_i + \beta_5 agri_i + \beta_6 cons_i + \beta_7 indus_i + \varepsilon_i$$

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* : $AIC = -2\ell_n + 2(p+1)$
- *Schwartz criterion* (ou encore *Bayesian information criterion*) : $SC = -2\ell_n + \ln(n)(p+1)$

Exemple : Stabilité du contrat de travail

Rappel du modèle

$$stable_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \beta_3 infbac_i + \beta_4 supbac_i + \beta_5 agri_i + \beta_6 cons_i + \beta_7 indus_i + \varepsilon_i$$

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1133.214	1058.267
SC	1138.122	1097.529
-2 Log L	1131.214	1042.267

$AIC = 1042,267 + 2 \times (7 + 1) = 1\ 058,267$
$SC = 1042,267 + \ln(1000) \times (7 + 1) = 1\ 097,529$

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,

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

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

Exemple : Stabilité du contrat de travail

Rappel du modèle

$$stable_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \beta_3 infbac_i + \beta_4 supbac_i + \beta_5 agri_i + \beta_6 cons_i + \beta_7 indus_i + \varepsilon_i$$

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	88.9471	7	<.0001
Score	95.2768	7	<.0001
Wald	71.8109	7	<.0001

$LR = 1131,214 - 1042,267 = 88,947$

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** (n_c) rapporté au nombre de paires total (n).

Exemple : Stabilité du contrat de travail

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	67.1	**Somers' D**	0.345
Percent Discordant	32.6	**Gamma**	0.346
Percent Tied	0.2	**Tau-a**	0.131
Pairs	188991	**c**	0.673

Indicateurs Plus ils sont proches de 1, mieux c'est !

$$D \text{ (Somer)} = \frac{n_c - n_d}{n} = 0{,}668 - 0{,}323 = 0{,}345$$

$$\gamma \text{ (Goodman-Kruskal)} = \frac{n_c - n_d}{n_c + n_d} = \frac{0,668 - 0,323}{0,668 + 0,323} = 0,348$$

$$\tau_a \text{ (Kendall)} = \frac{n_c - n_d}{0,5N(N-1)} = \frac{(0,668 - 0,323) \times 188\ 991}{0,5 \times 1000 \times 999} = 0,131$$

$$\text{c} = \frac{n_c + 0,5(n - n_c - n_d)}{n} = 0,668 + 0,5(1 - 0,668 - 0,323) = 0,673$$

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 ($\hat{p}_i < p^*$) ?

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.

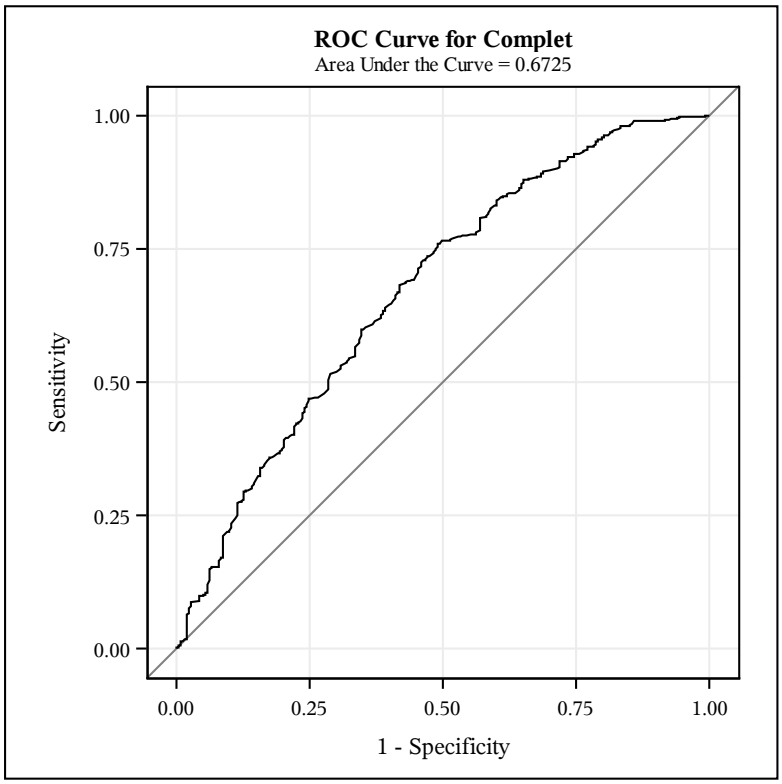
Indicateurs de qualité du modèle

Exemple : Stabilité du contrat de travail

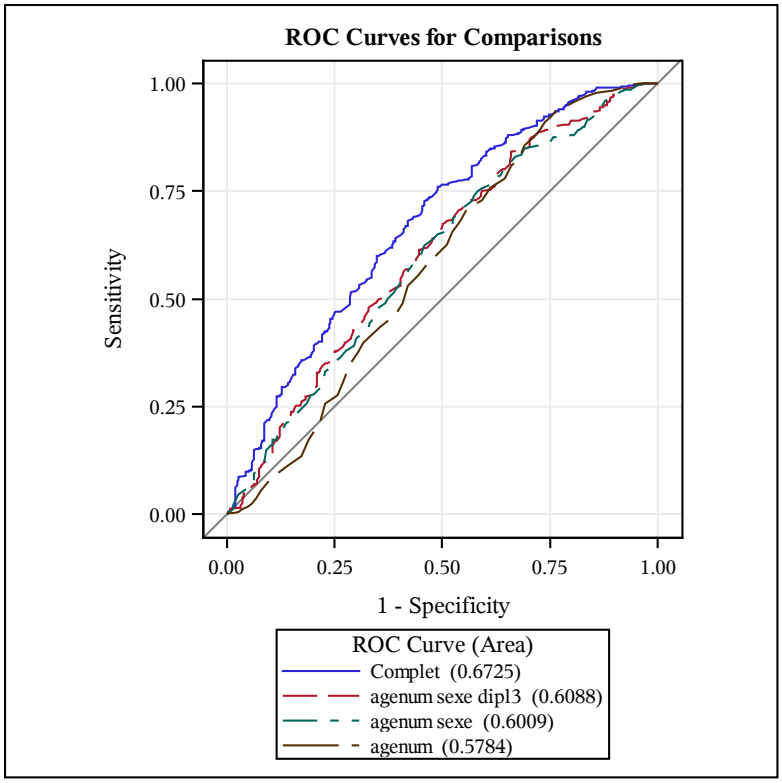
```
/*Activation des graphiques si nécessaire*/
ODS GRAPHICS ON;

/*Affichage de la courbe ROC du modèle complet*/
PROC LOGISTIC DATA = ee PLOTS(ONLY) = ROC;
  CLASS sexe (REF = "1") dipl3 (REF = "Bac") secteur (REF
      = "Construction") / PARAM = REF;
  MODEL stable (DESC)  = agenum sexe dipl3 secteur;
RUN;

/*Comparaison de plusieurs courbes ROC*/
PROC LOGISTIC DATA = ee PLOTS(ONLY) = ROC;
  CLASS sexe (REF = "1") dipl3 (REF = "Bac") secteur (REF
      = "Construction") / PARAM = REF;
  Complet: MODEL stable (DESC)  = agenum sexe dipl3
      secteur;
  ROC "agenum sexe dipl3" agenum sexe dipl3;
  ROC "agenum sexe" agenum sexe;
  ROC "agenum" agenum;
RUN;
```


Indicateurs de qualité du modèle

Exemple : Stabilité du contrat de travail



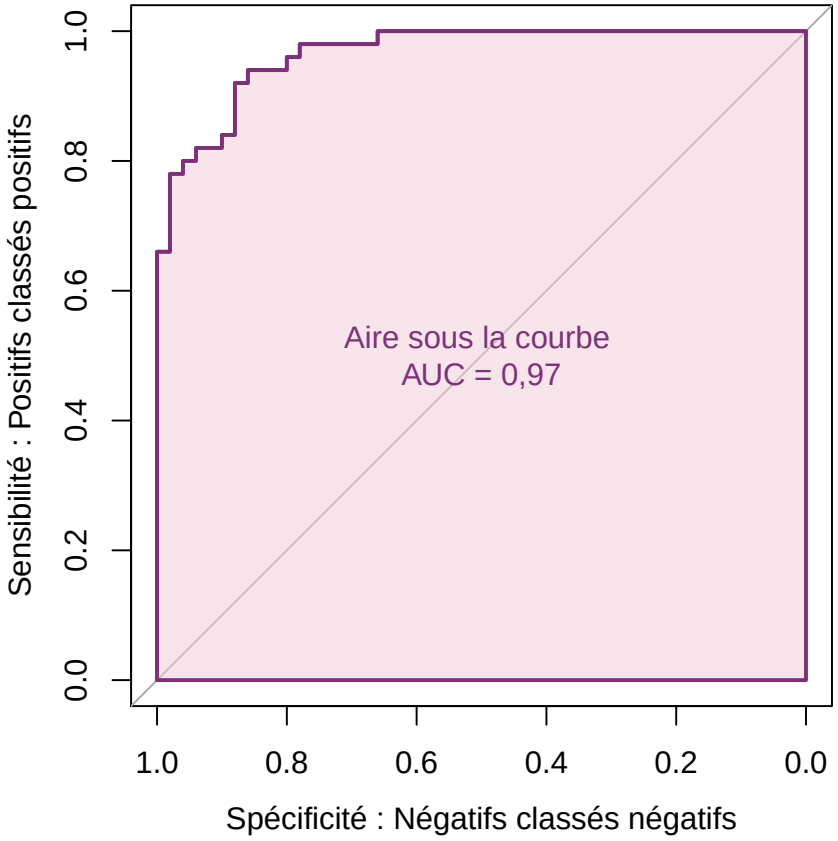
Exemple : Stabilité du contrat de travail



Indicateurs de qualité du modèle
Exemple : Stabilité du contrat de travail

ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D (Gini)	Gamma	Tau-a
Complet	0.6725	0.0202	0.6329	0.7122	0.3451	0.3459	0.1306
agenum sexe dipl3	0.6088	0.0213	0.5672	0.6505	0.2177	0.2186	0.0824
agenum sexe	0.6009	0.0212	0.5593	0.6424	0.2017	0.2038	0.0763
agenum	0.5784	0.0228	0.5337	0.6231	0.1568	0.1601	0.0593

Comparaison : Données simulées



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 20 %, la cote est de 1/4 (ou 4 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 $OR_{p|q} > 1$.

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 :

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

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
agenum	1.031	1.018	1.045
SEXE 2 vs 1	1.410	1.027	1.937
dipl3 Inférieur au bac vs Bac	0.891	0.589	1.347
dipl3 Supérieur au bac vs Bac	1.205	0.779	1.863
secteur Agriculture vs Tertiaire	0.090	0.040	0.205
secteur Construction vs Tertiaire	0.972	0.561	1.685
secteur Industrie vs Tertiaire	2.364	1.379	4.053

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

Interprétation courante des *odds-ratio*

En règle générale, on interprète l'*odds-ratio* associé à une modalité d'une variable qualitative comme le **rapport des chances** pour les individus présentant cette modalité d'être dans la situation modélisée **par rapport aux individus présentant la modalité de référence**.

Exemple L'*odds-ratio* associé au fait d'être une femme est de 1,410 : à âge, diplôme et secteur égaux par ailleurs, les femmes ont **1,410 fois plus de chances** d'être en contrat stables que les hommes.

Néanmoins, **cette interprétation très courante assimile *odds-ratio* et risque relatif**, ce qui pose problème quand l'*odds-ratio* est proche de 1.

Odds-ratio et risque relatif

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

En toute rigueur, **c'est ce risque relatif qui s'interprète comme un « rapport de chances »**, et non l'*odds-ratio*.

Qui plus est, quand p et q sont proches, risque relatif et *odds-ratio* diffèrent sensiblement.

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

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

Définition de l'effet marginal

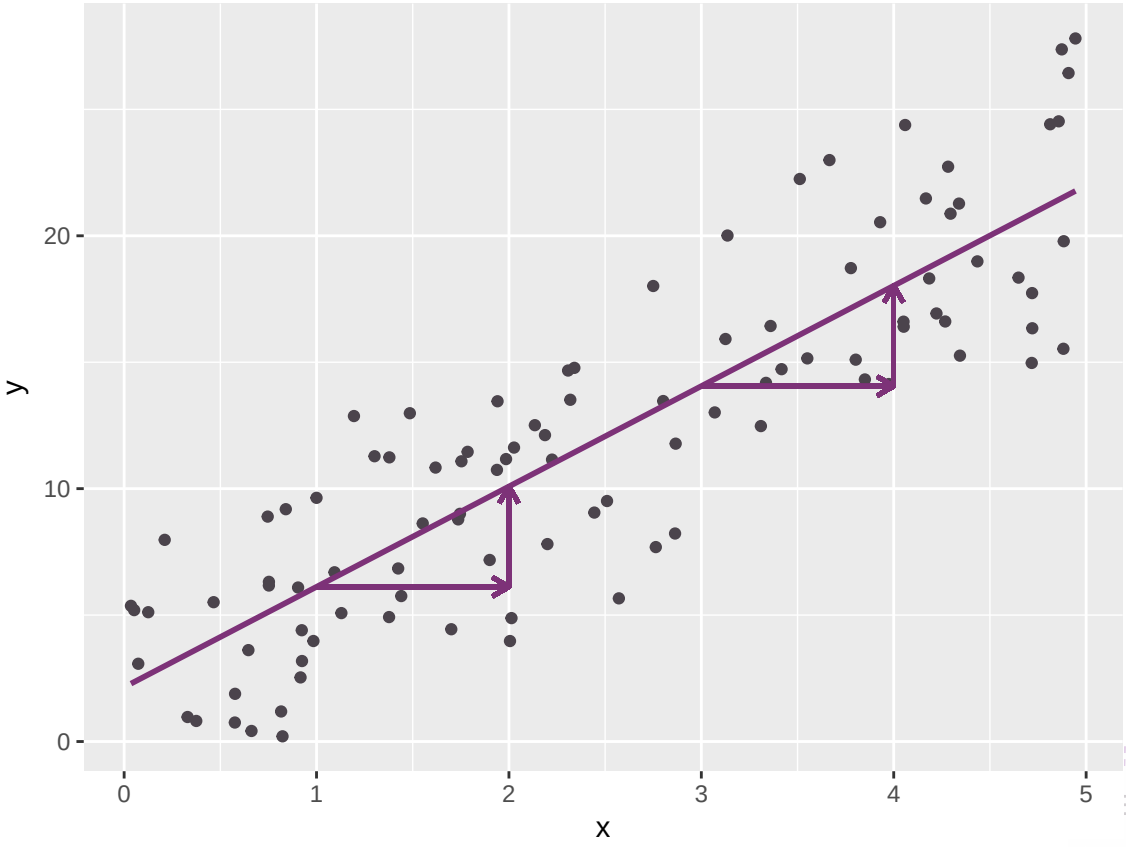
Dans un modèle logistique dichotomique, l'**effet marginal** est un moyen simple pour réexprimer la relation entre une variable explicative et la variable d'intérêt en termes de **points de pourcentages**.

Exemple En moyenne dans l'échantillon et à âge, sexe et diplôme **égaux par ailleurs**, le fait de **travailler dans l'industrie** plutôt que dans le secteur tertiaire est associé à une probabilité d'avoir un emploi stable **supérieure** de l'ordre de **13,7 points de pourcentage**.

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

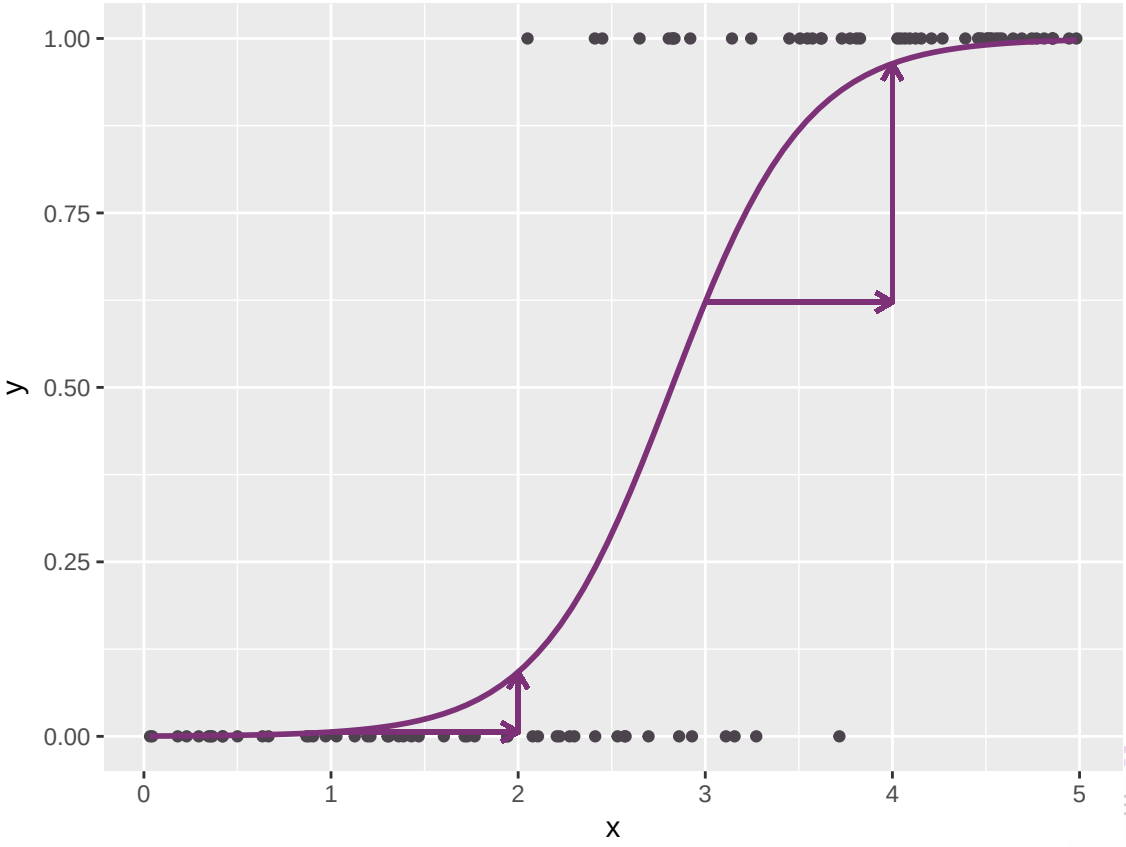
Odds-ratio et effets marginaux

Effet marginal dans un modèle linéaire classique



Odds-ratio et effets marginaux

Effet marginal dans un modèle logistique



Effet marginal dans un modèle logistique

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 :

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

Effet marginal dans un modèle logistique

Exemple Dans le modèle

$$stable_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \varepsilon_i$$

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

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

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

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

Effet marginal moyen

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

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

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 $\bar{\delta}(femme) = 0{,}10$, 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**.

Exemple : Stabilité du contrat de travail

```
/*Code de la macro %logistic_marginal*/
%INCLUDE "C:\logistic_marginal.sas";

/*Estimation des effets marginaux moyens
avec la macro %logistic_marginal*/
%logistic_marginal(
    DATA = ee,
    CLASS = sexe (REF = "1") dipl3(REF = "Bac") secteur
        (REF = "Construction") / PARAM = REF,
    MODEL = stable (DESC)  = agenum sexe dipl3 secteur
);
/*Note: la macro %logistic_marginal s'appuie
sur la macro %marginal de (Afsa, 2016)*/
```

Pour aller plus loin Méthodes alternatives pour calculer l'effet marginal moyen dans SAS : `http://support.sas.com/kb/22/604.html`.

Exemple : Stabilité du contrat de travail

Variable	Modalité	Effet marginal moyen (en points de %)	Ecart-type (en points de %)	P-valeur
SEXE	2	6.3016	3.0649	0.0398
DIPL3	Inférieur au bac	-2.3465	4.2258	0.5787
DIPL3	Supérieur au bac	3.5369	4.2915	0.4098
SECTEUR	Agriculture	-52.0540	6.6526	<.0001
SECTEUR	Construction	-0.5558	5.5993	0.9209
SECTEUR	Industrie	13.7969	3.9807	0.0005

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 est **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

Limites des tests déjà présentés

Les tests déjà présentés sont de deux types :

- les **tests globaux** : tests utilisés en analyse de variance à un facteur, tests de significativité globale des paramètres (statistique F ou ratio de vraisemblance) ;
- les **tests de significativité** : test de Student (régression linéaire) ou test de Wald (modèle linéaire généralisé).

Dans certains cas cependant, il est nécessaire de tester une hypothèse portant sur **plusieurs paramètres d'un modèle mais pas sur tous** :

- influence d'un facteur dans une analyse de variance à plusieurs facteurs ou test de significativité jointe de toutes les modalités d'une variable qualitative (effets de Type III) ;
- tests spécifiques quand une variable est introduite avec son carré ou en présence d'**interactions**.

Formulation d'hypothèses complexes

À des fins d'illustration, on se place dans le modèle :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$$

et on cherche à mener deux tests complexes :

$$H_0 : \beta_2 = 0 \text{ et } \beta_3 = 0 \quad \text{contre} \quad H_1 : \beta_2 \neq 0 \text{ ou } \beta_3 \neq 0$$

et d'autre part :

$$H_0 : \beta_2 = \beta_3 \quad \text{contre} \quad H_1 : \beta_2 \neq \beta_3$$

Remarque Le premier cas correspond au test de significativité globale d'une variable qualitative en régression multiple et le second au test à mener en présence d'interactions.

Complément : Tests d'hypothèses complexes
Modèle contraint et modèle non-contraint (1)

La construction d'une statistique de test générale dans ces deux configurations s'appuie sur les notions de **modèle contraint** et de **modèle non-contraint**.

Le **modèle non-contraint** est le **modèle complet**, celui qui comporte le plus de paramètres distincts. Ici, pour les deux tests il s'agit de :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$$

Le **modèle contraint** est le modèle obtenu **quand l'hypothèse nulle est respectée** : la valeur des paramètres est alors « contrainte » à la valeur que l'on souhaite tester.

Complément : Tests d'hypothèses complexes
Modèle contraint et modèle non-contraint (2)

Pour le test

$$H_0 : \beta_2 = 0 \text{ et } \beta_3 = 0 \quad \text{contre} \quad H_1 : \beta_2 \neq 0 \text{ ou } \beta_3 \neq 0$$

le modèle contraint est donc :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i$$

Pour le test

$$H_0 : \beta_2 = \beta_3 \quad \text{contre} \quad H_1 : \beta_2 \neq \beta_3$$

le modèle contraint est donc :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 (x_{2,i} + x_{3,i}) + \varepsilon_i$$

Intuition du test

Dans la mesure où il comporte plus de paramètres distincts, **le modèle non-contraint conduit mécaniquement à un meilleur ajustement aux données** :

- somme des carrés des résidus (SCR_{nc}) plus faible et donc R^2_{nc} plus élevé en régression linéaire ;
- log-vraisemblance ℓ_{nc} plus élevée dans le modèle linéaire généralisé.

Intuition Si l'ajustement du modèle aux données est bien meilleur dans le modèle non-contraint que dans le modèle contraint ($R^2_{nc} >> R^2_c$ ou $\ell_{nc} >> \ell_c$), alors on aura tendance à rejeter la contrainte, c'est-à-dire H_0.

Statistique de test en régression linéaire

Pour un modèle comportant p variables explicatives (+ la constante) et si le test impose q restrictions, alors on peut montrer que sous H_0 :

$$F = \frac{(SCR_c - SCR_{nc})/q}{SCR_{nc}/(n-(p+1))} = \frac{(R^2_{nc} - R^2_c)/q}{(1-R^2_{nc})/(n-(p+1))} \hookrightarrow F_{q,n-(p+1)}$$

Dans les deux tests présentés $p = 3$. Dans le premier $q = 2$ et dans le second $q = 1$.

Remarque Quand le modèle contraint est le modèle ne comportant que la constante, alors $SCR_c = SCT$, $SCR_{nc} = SCR$ et $q = p$. On retrouve ainsi exactement le test de significativité globale :

$$F = \frac{(SCT - SCR)/p}{SCR/(n-(p+1))} = \frac{SCE/p}{SCR/(n-(p+1))} \hookrightarrow F_{p,n-(p+1)}$$

Statistique de test dans le modèle linéaire généralisé

Dans le modèle linéaire généralisé, ce test peut être posé comme un test du ratio de vraisemblance. On peut en effet montrer que sous H_0 :

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

où ℓ_{nc} est la log-vraisemblance du modèle non-contraint et ℓ_c la log-vraisemblance du modèle contraint.

Remarque Quand le modèle contraint est le modèle ne comportant que la constante, alors $\ell_c = \ell^0$, $\ell_{nc} = \ell_n$ et $q = p$. On retrouve ainsi exactement le test de significativité globale :

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

Exemple : Stabilité du contrat de travail

Les modèles estimiés jusqu'à présent ont permis de constater une **relation positive et significative** entre âge et stabilité du contrat de travail.

On souhaite désormais déterminer si l'intensité de cette relation **varie selon le sexe du salarié**.

Pour ce faire, on estime un modèle avec une **interaction entre âge et sexe** :

$$stable_i = \beta_0 + \beta_1 age_i + \beta_2 femme_i + \beta_3 femme_i \times age_i + \varepsilon_i$$

Exemple : Stabilité du contrat de travail

Dans ce modèle, la relation entre âge et probabilité d'être en emploi stable est captée :

- par β_1 pour les hommes ;
- par $\beta_1 + \beta_3$ pour les femmes.

De ce fait, β_3 représente la différence dans la relation entre âge et stabilité du contrat associé au fait d'être une femme plutôt qu'un homme.

Le test de significativité de β_3 permet donc de **déterminer si cette différence est significative aux seuils statistiques usuels**.

Exemple : Stabilité du contrat de travail

```
/*Modèle avec effet d'interaction*/
PROC LOGISTIC DATA = ee;
    CLASS sexe(REF = "1") / PARAM = REF;
    MODEL stable (DESC) = agenum sexe sexe *
        agenum;
RUN;
```

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.3668	0.3472	1.1160	0.2908
agenum		1	0.0122	0.00810	2.2735	0.1316
SEXE	2	1	-1.0137	0.5275	3.6923	0.0547
agenum*SEXE	2	1	0.0349	0.0126	7.6517	0.0057

Exemple : Stabilité du contrat de travail

Cependant, si dans ce modèle on souhaite tester la **significativité de l'association entre âge et stabilité du contrat de travail pour les femmes**, le test à mener fait intervenir deux coefficients :

$$H_0 : \beta_1 + \beta_3 = 0 \quad \text{contre} \quad H_1 : \beta_1 + \beta_3 \neq 0$$

C'est donc un **test d'hypothèse complexe** dont le modèle contraint correspondant est :

$$\begin{aligned} stable_i &= \beta_0 + \beta_1 age_i + \beta_2 femme_i - \beta_1 femme_i \times age_i + \varepsilon_i \\ &= \beta_0 + \beta_1 (age_i - femme_i \times age_i) + \beta_2 femme_i + \varepsilon_i \end{aligned}$$

car $\beta_1 + \beta_3 = 0 \Leftrightarrow \beta_3 = -\beta_1$

Exemple : Stabilité du contrat de travail

```
/*Test d'hypothèse complexe*/
PROC LOGISTIC DATA = ee;
    CLASS sexe(REF = "1") / PARAM = REF;
    MODEL stable (DESC) = agenum sexe sexe *
       agenum;
    TEST agenum + sexe2agenum = 0;
RUN;
```

Linear Hypotheses Testing Results

	Label	Wald Chi-Square	DF	Pr > ChiSq
	Test 1	23.7471	1	<.0001

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 **asymétriques** : variable asymétrique continue ou discrète ;
- variables **polytomiques** : information non-ordonnée ou ordonnée.

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**.

Données particulièrement asymétriques
Définition et exemples

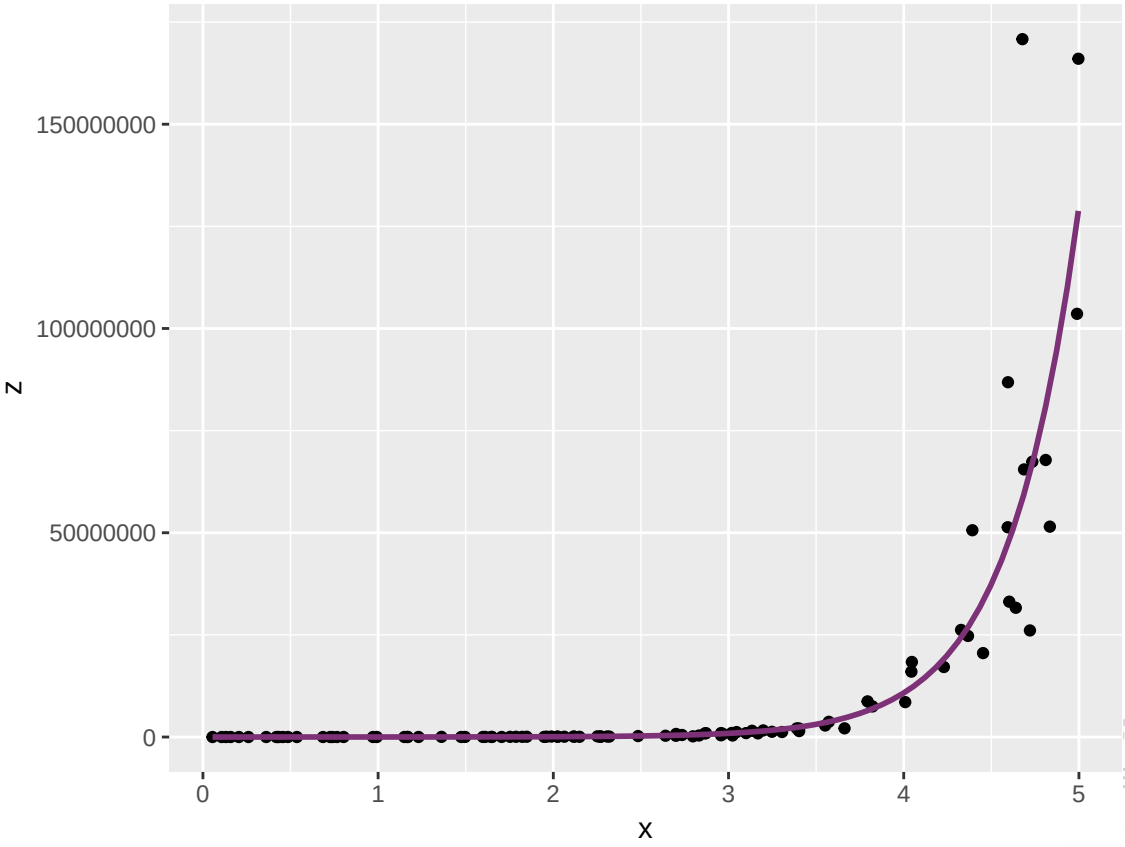
Des données sont considérées comme particulièrement asymétriques quand un **modèle polynomial** en les variables explicatives **ne suffit pas** pour les modéliser correctement.

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.)

Données particulièrement asymétriques
Illustration : Distribution d'un actif financier



Principes de modélisation

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

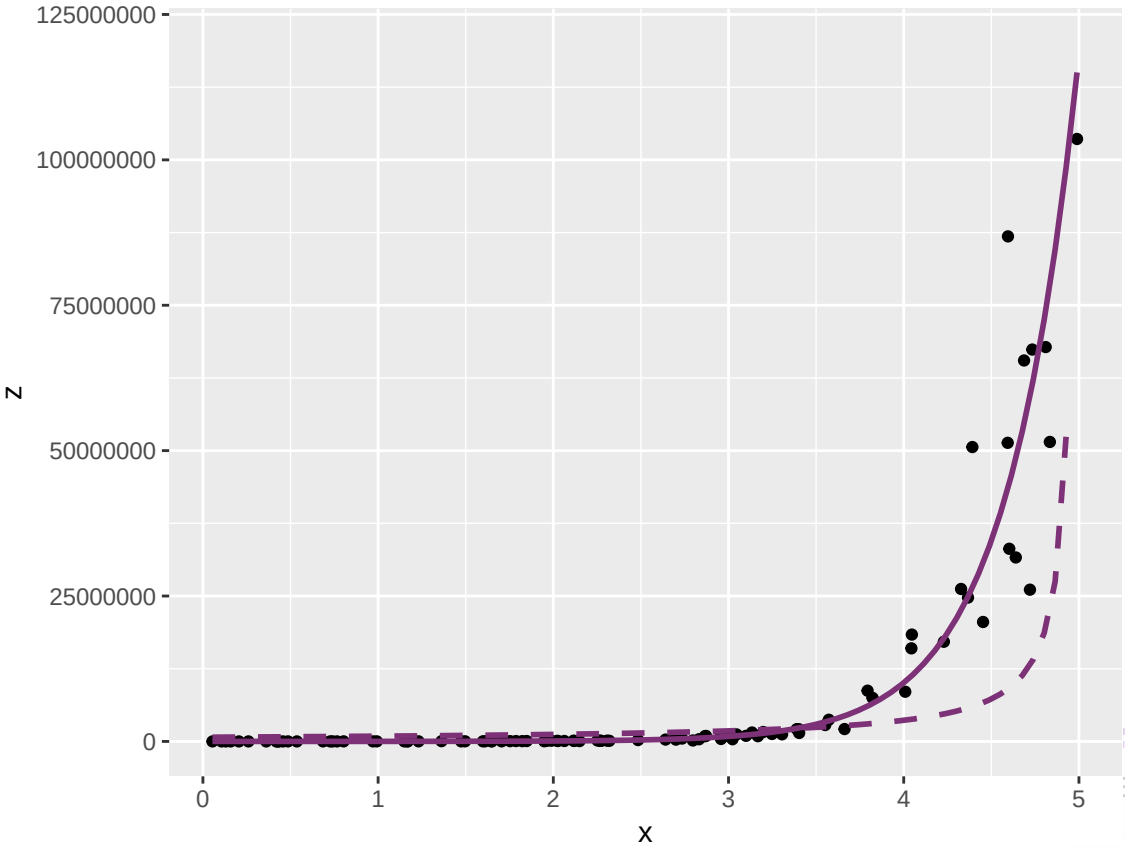
- **régression gamma** pour les données continues et positives ;
- **régression de Poisson** ou négatives-binomiales (éventuellement inflatées en zéro) pour les données discrètes.

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**.

Illustration : Distribution d'un actif financier



Interprétation du modèle

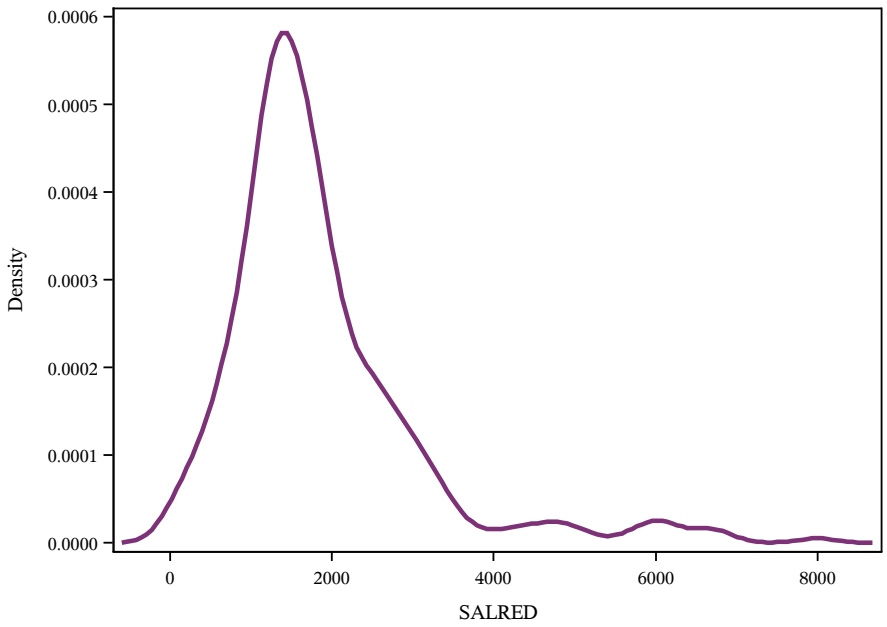
Quand la fonction de lien utilisée est le **logarithme**, il est possible d'**interpréter directement** les coefficients du modèle.

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 %.

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.



Exemple : Salaire mensuel

```
/*Ajout de deux nouvelles variables*/
DATA ee;
  SET ee;
  part = (tpp = "2"); /*Temps partiel*/
  agenum2 = agenum**2; /*Age au carré*/
RUN;

/*Estimation avec la PROC GENMOD*/
PROC GENMOD DATA = ee;
    CLASS dipl3(REF = "Bac") sexe(REF = "1") /
       PARAM = REF;
    MODEL salred = agenum agenum2 dipl3 sexe
       part / DIST = GAMMA LINK = LOG TYPE3;
RUN;
```


Exemple : Salaire mensuel

Model Information		
Data Set	WORK.EE	
Distribution	Gamma	
Link Function	Log	
Dependent Variable	SALRED	SALRED

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
agenum	1	18.90	<.0001
agenum2	1	11.93	0.0006
dipl3	2	75.52	<.0001
SEXE	1	8.16	0.0043
part	1	52.89	<.0001

Exemple : Salaire mensuel

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	6.0275	0.2892	5.4607	6.5943	434.40	<.0001
agenum		1	0.0673	0.0150	0.0380	0.0967	20.22	<.0001
agenum2		1	-0.0006	0.0002	-0.0010	-0.0003	12.51	0.0004
dipl3	Inférieur au bac	1	-0.2199	0.0737	-0.3643	-0.0754	8.90	0.0028
dipl3	Supérieur au bac	1	0.3214	0.0754	0.1737	0.4691	18.18	<.0001
SEXE	2	1	-0.1597	0.0556	-0.2686	-0.0508	8.26	0.0041
part		1	-0.5635	0.0709	-0.7024	-0.4245	63.17	<.0001
Scale		1	4.6350	0.3551	3.9887	5.3860		

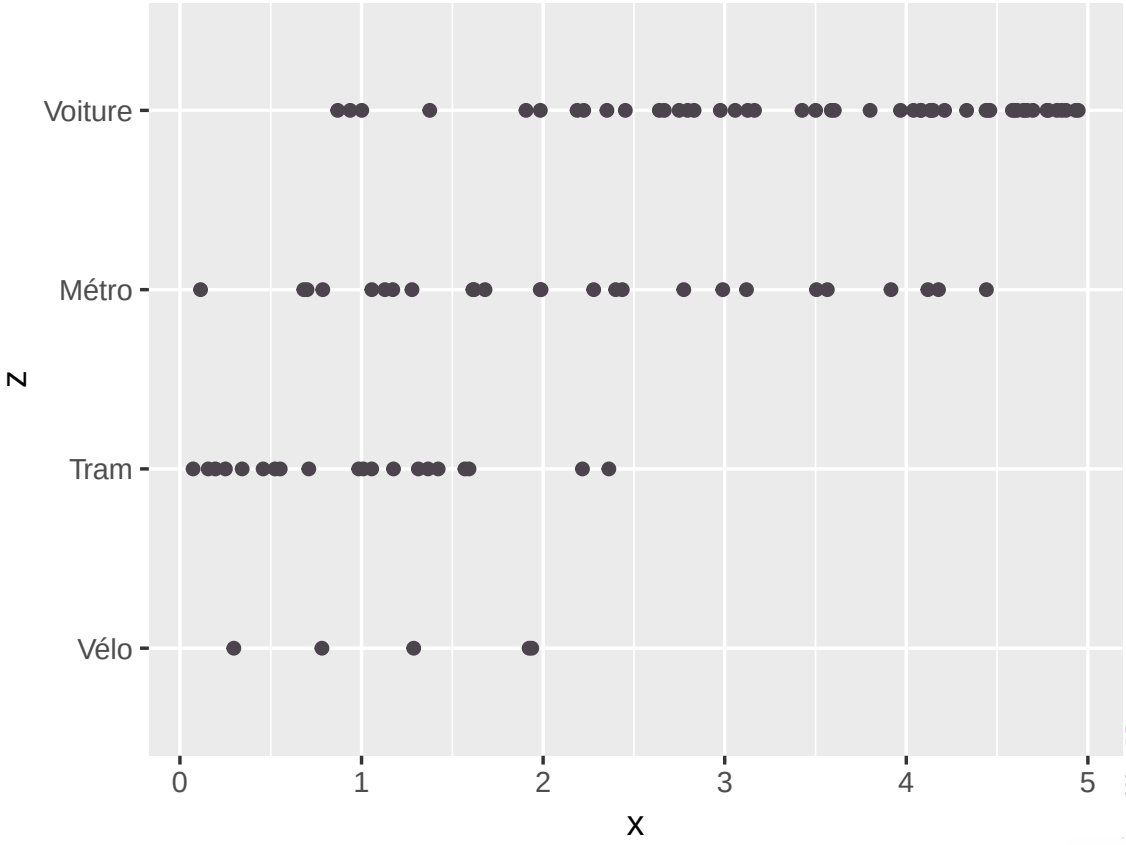
Définition et exemples

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



Principes de modélisation

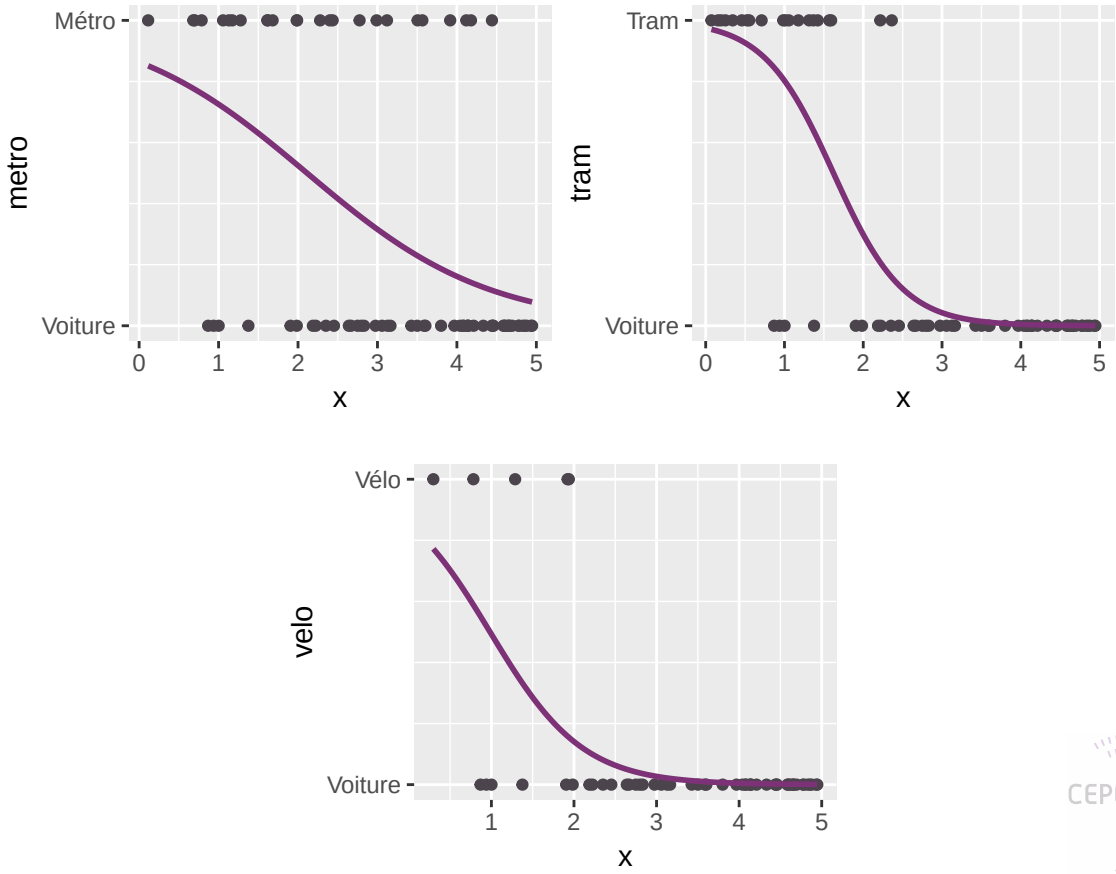
Idée On se ramène à une série de modèles dichotomiques en **choisissant une modalité de la *variable à expliquer* comme référence**.

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*.

Principes de modélisation



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, même si en définitive un seul est choisi.

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 $OR_k = e_k^{\beta}$.

Les **probabilités** 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.

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

```
/*Utilisation d'un format pour clarifier les modalités de
    acteu*/
PROC FORMAT;
  VALUE $acteu
    "1" = "Actifs occupés"
    "2" = "Chômeurs"
    "3" = "Inactifs"
  ;
RUN;

/*Modèle multinomial avec "Actifs occupés" comme
    référence*/
PROC LOGISTIC DATA = ee;
    CLASS dipl3(REF = "Bac") sexe(REF = "1") / PARAM = REF;
    MODEL acteu (REF = "Actifs occupés") = agenum dipl3
        sexe / LINK = GLOGIT;
    FORMAT acteu $acteu.;
RUN;
```


Données polytomique non-ordonnées
Exemple : Position sur le marché du travail

Model Information		
Data Set	WORK.EE	
Response Variable	ACTEU	ACTEU
Number of Response Levels	3	
Model	generalized logit	
Optimization Technique	Newton-Raphson	

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
agenum	2	246.5160	<.0001
dipl3	4	108.3903	<.0001
SEXE	2	10.1120	0.0064

Données polytomique non-ordonnées
Exemple : Position sur le marché du travail

Analysis of Maximum Likelihood Estimates							
Parameter		ACTEU	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		Chômeurs	1	-1.3488	0.3116	18.7366	<.0001
Intercept		Inactifs	1	-1.9814	0.1736	130.3212	<.0001
agenum		Chômeurs	1	-0.0178	0.00617	8.2936	0.0040
agenum		Inactifs	1	0.0405	0.00276	215.9949	<.0001
dipl3	Inférieur au bac	Chômeurs	1	0.00485	0.2634	0.0003	0.9853
dipl3	Inférieur au bac	Inactifs	1	0.1292	0.1342	0.9264	0.3358
dipl3	Supérieur au bac	Chômeurs	1	-0.6506	0.3168	4.2165	0.0400
dipl3	Supérieur au bac	Inactifs	1	-1.1635	0.1598	52.9996	<.0001
SEXE	2	Chômeurs	1	-0.1552	0.2116	0.5383	0.4631
SEXE	2	Inactifs	1	0.2890	0.0994	8.4598	0.0036

Complément : 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**.

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.

Exemple : Intensité du temps partiel

```
/*Ajout de nouvelles variables*/
DATA ee;
    SET ee;
    IF tpp IN("1", "2");
    IF tpp = '1' THEN txtppb = "6";
    nbenf1 = (NBAGENF IN ("1", "2", "3")); /*1 enfant*/
    nbenf2 = (NBAGENF IN ("4", "5", "6")); /*2 enfants*/
    nbenf3p = (NBAGENF IN ("7", "8", "9")); /*3 enfants et
        plus*/
    enf3ans = (NBAGENF in ("3", "6", "9")); /*1 enfant de
        moins de 3 ans*/
RUN;

/*Estimation du modèle*/
PROC LOGISTIC DATA = ee;
    CLASS dipl3(REF = "Bac") sexe(REF = "1") / PARAM = REF;
    MODEL txtppb = agenum sexe dipl3 enf3ans nbenf1 nbenf2
        nbenf3p;
RUN;
```

Exemple : Intensité du temps partiel

Model Information		
Data Set	WORK.EE	
Response Variable	TXTPPB	TXTPPB
Number of Response Levels	6	
Model	cumulative logit	
Optimization Technique	Fisher's scoring	

Response Profile		
Ordered Value	**TXTPPB**	**Total Frequency**
1	1	49
2	2	38
3	3	47
4	4	31
5	5	20
6	6	801

Exemple : Intensité du temps partiel

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**1**	1	-3.9920	0.4264	87.6682	<.0001
Intercept	**2**	1	-3.3581	0.4150	65.4718	<.0001
Intercept	**3**	1	-2.8462	0.4092	48.3755	<.0001
Intercept	**4**	1	-2.5813	0.4069	40.2447	<.0001
Intercept	**5**	1	-2.4281	0.4057	35.8201	<.0001
agenum		1	-0.00076	0.00771	0.0097	0.9214
SEXE	**2**	1	1.6682	0.1992	70.1103	<.0001
dipl3	**Inférieur au bac**	1	0.0751	0.2355	0.1017	0.7498
dipl3	**Supérieur au bac**	1	-0.2361	0.2436	0.9395	0.3324
enf3ans		1	0.0902	0.3406	0.0702	0.7911
nbenf1		1	-0.3319	0.2431	1.8633	0.1722
nbenf2		1	-0.0908	0.2417	0.1411	0.7072
nbenf3p		1	0.6025	0.3216	3.5092	0.0610