UNIVERSITÉ SORBONNE PARIS NORD

Année universitaire **2023 - 2024**

L’accès aux études supérieures

Une analyse approfondie sur l’accès aux études supérieures dans le cadre du cours d’économétrie des données qualitatives.

Mémoire de recherche présenté par :

- **Mahi Elamine GUENDOUZ – 12306918**
- **Massyle DEKKAR – 12303497**
- **Philippe BAO - 12303585**
- **Lysa HALLI -12314000**

Introduction :

L'accès aux études supérieures est un enjeu majeur d'un point de vue micro et macroéconomique. Il constitue, pour nombre de pays, une préoccupation majeure et l'ampleur de cet enjeu sera d'autant plus grande dans les années à venir. Les politiques mises en place par les gouvernements seront donc cruciales et les conséquences d'une mauvaise connaissance et gestion pourraient être dramatiques.

Les travaux de Thomas Piketty et Mathieu Valdenaire sur les inégalités scolaires ont démontré l'importance des politiques gouvernementales impactant les inégalités scolaires concernant les élèves du premier et second cycle.

Dans cette même perspective, nous allons déterminer les causes et interactions en analysant les variables des « étudiants postulant aux études supérieures », « si au moins un des deux parents possède un diplôme d'études supérieures », « si l'établissement est privé ou public », et « GPA, correspondant à la moyenne pondérée de l'étudiant ».

L'analyse de ces variables nous permettra de mieux comprendre la complexité d'un problème d'apparence simple mais en réalité très complexe.

En dépit de cela, une question légitime se pose : sommes-nous tous égaux face au système éducatif ou sommes-nous finalement des rouages de ce dernier ?

I. Présentation de la data frame et préparation des données :

1- Présentation :

Data Frame : 3_Etudes_sup.dta

Notre DF contient les variables suivantes :

- **Apply :** Cette variable présente la probabilité qu'un étudiant postule pour des études supérieures, avec des niveaux :

  0 => improbable qui postule.
  1 => peu probable qui postule.
  2 => très probable qui postule.

- **Pared :** une variable (binaire) qui montre si au moins un des parents, a fait des études supérieures :

  0 => aucun des parents n'a fait d'études supérieures.
  1 => au moins un des parents a fait des études supérieures.

- **Public** : cette variable indique si l'établissement du premier cycle de l'étudiant est public ou privé :

  0 => l'établissement est privé.
  1 => l'établissement est public.

- **Gpa :** présente la moyenne pondérée de l'étudiant.

Notre DF se constitue de 400 observations (400 étudiants), on peut déduire des premières stats descriptives :

```
> summary(data_etud_sup)
     apply           pared            public            gpa
 Min.   :0.00   Min.   :0.0000   Min.   :0.0000   Min.   :1.900
 1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2.720
 Median :0.00   Median :0.0000   Median :0.0000   Median :2.990
 Mean   :0.55   Mean   :0.1575   Mean   :0.1425   Mean   :2.999
 3rd Qu.:1.00   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:3.270
 Max.   :2.00   Max.   :1.0000   Max.   :1.0000   Max.   :4.000
>
```

- 15,75 % des étudiants au moins un de leurs parents ont faits des études supérieures.
- 14,25 % des étudiants ont fait leurs premiers cycles dans un établissement public.
- La moyenne pondérée des étudiants est incluse dans l'intervalle [1,9 - 4], avec une moyenne de 2,999.

```
> questionr::freq(apply)
                     n  % val%
[0] unlikely       220 55   55
[1] somewhat likely 140 35   35
[2] very likely     40 10   10
>
```

- 55 % des étudiants (n = 220), improbable qui postulent pour des études supérieures.
- 35 % des étudiants (n = 140), peu probable qui postulent pour des études supérieures.
- 10 % des étudiants (n = 40), très probable qui postulent pour des études supérieures.

```
> questionr::freq(pared)
   n    % val%
0 337 84.2 84.2
1  63 15.8 15.8
>
```

- 84,2 % des étudiants (n = 337) aucun de leurs parents n'a fait d'études supérieures.
- 15,8 % des étudiants (n = 63) au moins un de leurs parents a fait des études supérieures.

```
> questionr::freq(public)
   n    % val%
0 343 85.8 85.8
1  57 14.2 14.2
>
```

- 85,8 % des étudiants (n = 343) ont fait leurs premiers cycles dans un établissement privé.
- 14,2 % des étudiants (n = 57) ont fait leurs premiers cycles dans un établissement public.

Relation entre la probabilité qu'un étudiant candidate pour des études supérieures (apply) et le niveau d'étude des parents (pared) :

```
> rel_apply_pared
      pared
apply   0   1
    0 200  20
    1 110  30
    2  27  13
>
```

```
> cor(apply, pared)
[1] 0.2190363
>
```

- 50 % des étudiant leurs parents n'ont pas fait des études supérieures et improbable qui postulent pour des études supérieures.
- 7,5 % des étudiants au moins un de leurs parents a fait des études supérieures et peu probable qui postulent pour des études supérieures.
- 6,75 % des étudiants leurs parents n'ont pas fait des études supérieures et très probable qui postulent pour des études supérieures.

Relation entre la probabilité qu'un étudiant candidate pour des études supérieures (apply) et la nature de l'établissement du premiers cycle fréquenté (public) :

```
> rel_apply_public
      public
apply   0   1
    0 189  31
    1 124  16
    2  30  10
>
```

```
> cor(apply, public)
[1] 0.04971323
>
```

- 47,25 % des étudiants ont fréquentés un établissement de premiers cycles privé et improbable qui postulent pour des études supérieures.
- 4 % des étudiants ont fréquentés un établissement de premiers cycles public et peu probable qui postulent pour des études supérieures.

- 7,5 % des étudiants ont fréquentés un établissement de premiers cycles privé et très probable qui postulent pour des études supérieures.

2- Préparation des données :

Dans cette partie, on a commencé par définir trois nouvelles variables :

- **Classe_gpa :** Cette variable va classer les moyennes pondérées des étudiant de 1 à 6 ; sachant que : 1,9 < gpa < 4 et (delta = 2,1/6 = 0,35)

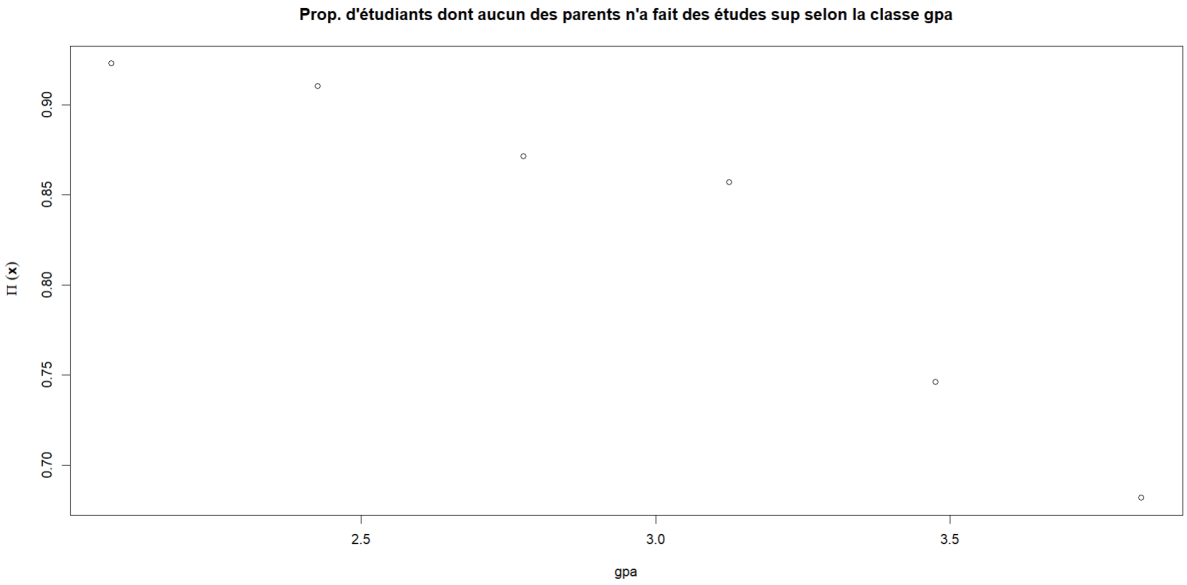
  1 => lorsque la gpa est inférieure ou égale à 2,25.
  2 => lorsque la gpa est strictement supérieure à 2,25.
  3 => lorsque la gpa est strictement supérieure à 2,6.
  4 => lorsque la gpa est strictement supérieure à 2,95.
  5 => lorsque la gpa est strictement supérieure à 3,3.
  6 => lorsque la gpa est strictement supérieure à 3,65.

- **parent_non_etudesup :** Cette variable prend la valeur :

  0 => au moins un des parents a fait des études supérieures.
  1 => aucun des parents n'a fait d'études supérieures.

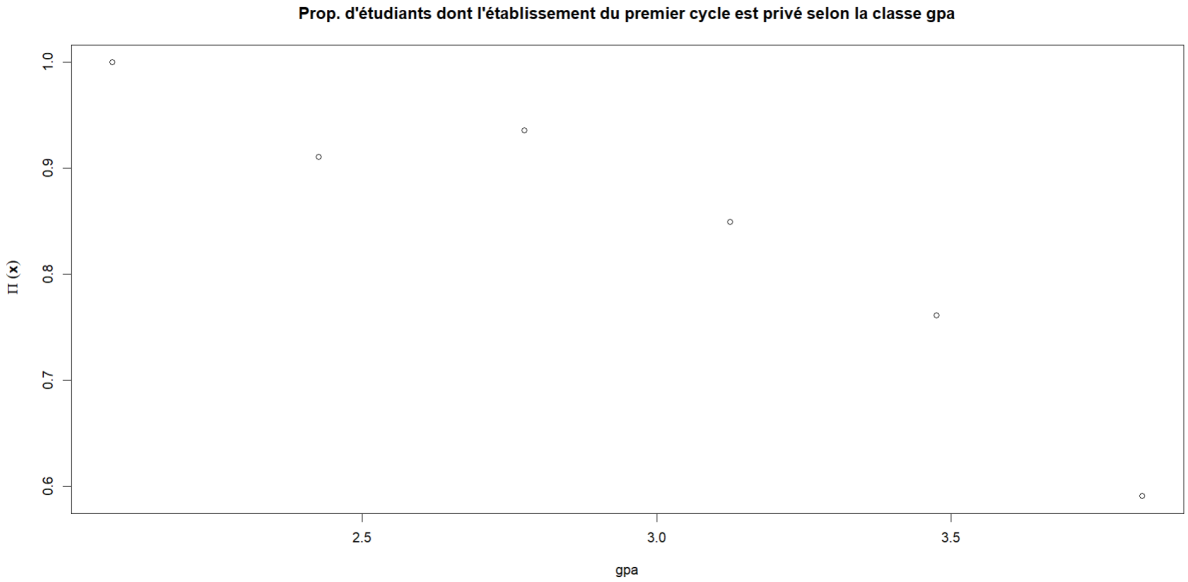
- **etudiant_etab_prive :** Cette variable prend la valeur :

  0 => l'établissement du premier cycle est public.
  1 => l'établissement du premier cycle est privé.



A partir de ce graphique, on remarque qu'il y a une relation négative entre le niveau d'étude des parents et la moyenne pondérée des étudiants, prenons la comparaison suivante : la proportion des étudiants dont aucun des parents n'a pas fait des études supérieures dans la classe_gpa qui est entre 1,9 et 2,25 dépasse les 90 % tandis que la proportion des étudiants dont aucun des parents n'a pas fait des études supérieures dans la classe_gpa qui est entre 3,65 et 4 ne dépasse pas les 70 %. Une corrélation négative égale à – 0,1608352.

Donc on peut dire que le fait que les parents n'ont pas fait d'étude supérieure peut avoir un effet négatif sur la moyenne pondéré de l'étudiant.



On remarque qu'il y a une tendance négative entre la proportion des étudiants qui ont fait leurs études de premier cycle dans un établissement privé, la proportion des étudiants qui ont fait leurs premiers cycles dans un établissement privé dans la classe_gpa qui est entre 1,9 et 2,25 est de 100 % tandis que la proportion des étudiants qui ont fait leurs premiers cycles dans un établissement privé dans la classe_gpa qui est entre 3,65 et 4 ne dépasse pas les 60 %. Une corrélation négative égale à - 0,2332896.

Donc on peut dire que le fait de faire un établissement de premier cycle privé n'assure pas forcément une moyenne pondérée meilleure que celui d'un établissement public.

II. Analyse des modèles :

Modèle 1 : MODELE COMPLET sans interaction :

```
> Reg.logit_1 <- glm(apply~parent_non_etudesup+etudiant_etab_prive,family=binomial(link="logit"))
> summary(Reg.logit_1)

Call:
glm(formula = apply ~ parent_non_etudesup + etudiant_etab_prive,
    family = binomial(link = "logit"))

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.71284    0.35771   1.993   0.0463 *
parent_non_etudesup -1.14902    0.29351  -3.915 9.05e-05 ***
etudiant_etab_prive  0.06648    0.29611   0.225   0.8224
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 550.51  on 399  degrees of freedom
Residual deviance: 534.03  on 397  degrees of freedom
AIC: 540.03

Number of Fisher Scoring iterations: 4
```

Le modèle de régression logistique utilisé pour expliquer la modalité « apply » (variable correspondant à la probabilité qu'un étudiant postule pour des études supérieures) démontre que la variable « parent_non_etudesup » a un impact significatif. Le coefficient associé à cette variable est de -1,15 (avec une p-value de 9,05 x 10^{-5}), ce qui indique une relation négative, signifiant une probabilité plus faible de postuler pour des études supérieures pour ceux dont les parents n'ont pas suivi d'études supérieures, par rapport à ceux dont les parents ont suivi des études supérieures.
En revanche, la variable « etudiant_etab_prive » n'a pas d'impact significatif sur la probabilité de postuler pour des études supérieures, avec un coefficient de 0,066 et une p-value de 0,8224. Cela signifie que le fait d'être dans un établissement privé ou public n'a pas d'effet significatif sur la décision de postuler pour des études supérieures.

À noter que ce modèle possède un AIC de 540,03. La comparaison de cette valeur avec celles d'autres modèles sera nécessaire pour la sélection du modèle optimal.

On peut vérifier les résultats obtenus ci-dessus en analysant la table des déviances du modèle.

```
> anova(Reg.logit_1, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: apply

Terms added sequentially (first to last)


                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                  399     550.51
parent_non_etudesup  1  16.4344       398     534.08 5.036e-05 ***
etudiant_etab_prive  1   0.0505       397     534.03    0.8222
```

L'analyse de la déviance confirme les conclusions précédentes concernant la significativité de l'éducation des parents vis-à-vis du modèle, tandis que le type d'établissement ne l'est pas.

- La première ligne du tableau correspond au modèle nul. La déviance résiduelle pour ce modèle est de 550,51.

- La deuxième ligne correspond au modèle avec la variable parent_non_etudesup. Lorsque cette variable est ajoutée au modèle, la déviance résiduelle diminue de 16,4344, ce qui indique que la variable a un effet significatif sur la variable réponse (p-value = 5,036e-05).

- La troisième ligne correspond au modèle avec les variables parent_non_etudesup et etudiant_etab_prive. Lorsque etudiant_etab_prive est ajoutée au modèle, la déviance résiduelle ne diminue presque pas (0,0505), ce qui indique que cette variable n'a pas d'effet significatif sur la variable réponse (p-value = 0,82220).

Cette analyse confirme nos précédentes observations, à savoir que la variable parent_non_etudesup a un effet significatif sur la variable apply, tandis que la variable etudiant_etab_prive n'a pas d'effet significatif.

```
> pred.proba.reg1 <- predict(Reg.logit_1,type="response",newdata = data_etud_sup)
> summary(pred.proba.reg1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3927  0.4086  0.4086  0.4500  0.4086  0.6855
```

Nous observons une certaine concentration des probabilités prédites autour de la médiane (0,4086). En effet, les quartiles étant très proches suggèrent une concentration homogène autour de celle-ci. La moyenne, qui est de 0,4500 et légèrement supérieure à la médiane, implique des valeurs assez élevées, augmentant par conséquent la moyenne (valeur max : 0,6855).

Matrice de confusion :

La matrice de confusion nous permet de visualiser la précision du modèle, mettant en avant les éléments suivants :

- Vrais négatifs = 200
- Faux positifs = 20
- Vrais positifs = 43
- Faux négatifs = 137

```
> class(mc.reg1)
[1] "table"
> print(mc.reg1)
      pred.moda.reg1
apply   0   1
    0 200  20
    1 137  43
```

Taux d'erreur :

En se basant sur cette matrice, on peut calculer le taux d'erreur du modèle de régression logistique. Il correspond à la proportion de cas mal classés par le modèle :

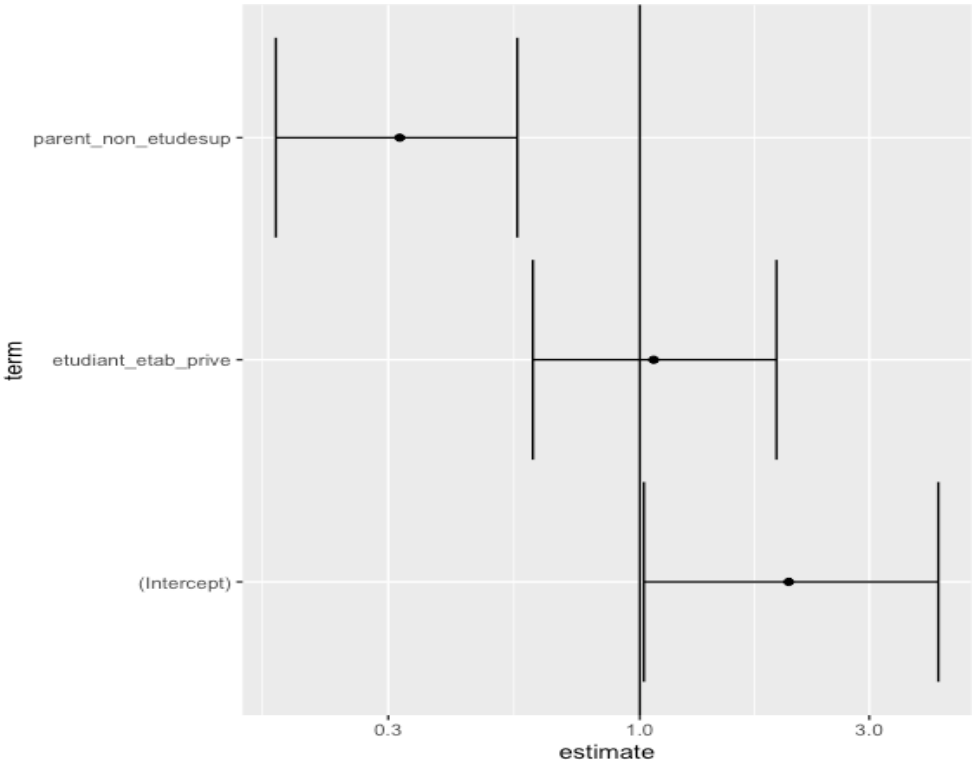
```
> err.reg1 <- (mc.reg1[2,1]+mc.reg1[1,2])/sum(mc.reg1)
> print(err.reg1)
[1] 0.3925
```

Le taux d'erreur global est de 39.25%, par conséquent une précision globale de 60.75%.

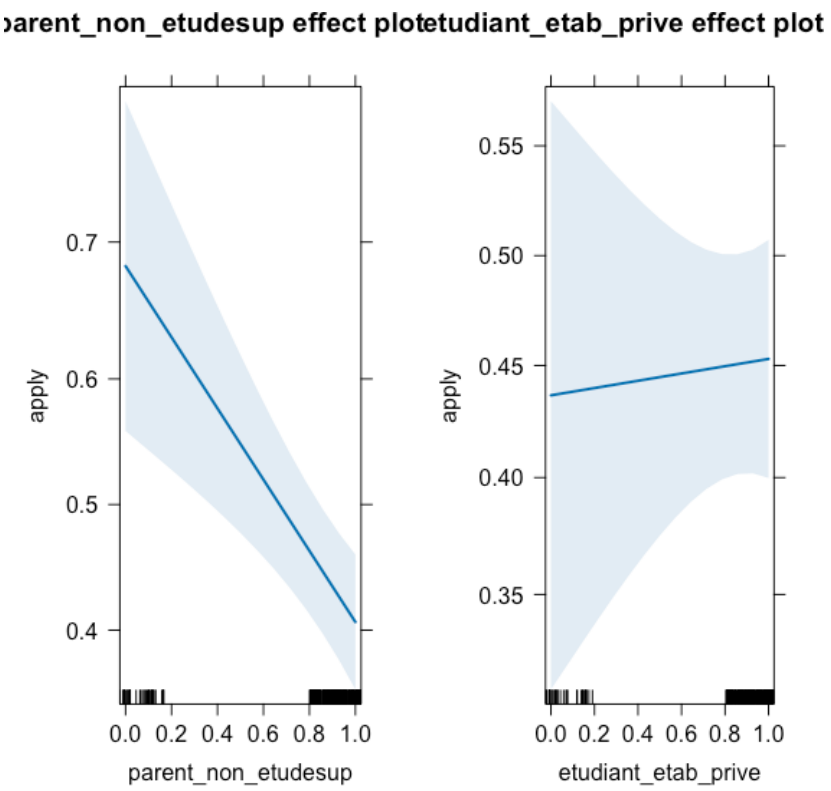
Odds ratios :

```
> odds.ratio(Reg6.select)
Waiting for profiling to be done...
                           OR   2.5 % 97.5 %         p
(Intercept)          2.03977 1.02041 4.1749   0.04628 *
parent_non_etudesup 0.31695 0.17516 0.5565 9.048e-05 ***
etudiant_etab_prive 1.06874 0.59971 1.9238   0.82237
```

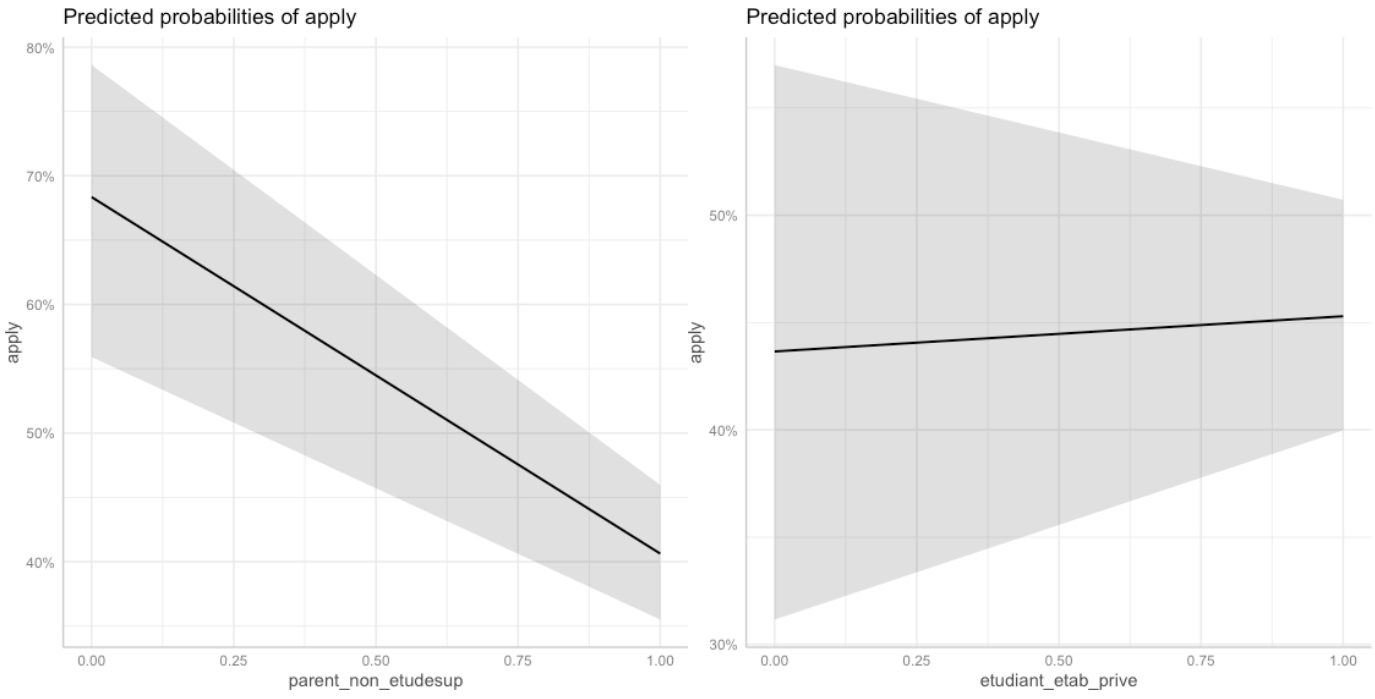
L'odds ratio pour la variable « parent_non_etudesup » est très significatif (9,048 x 10^-5), avec un ratio de 0,317, indiquant que les chances de postulation des étudiants dont les parents n'ont pas fait d'études supérieures sont 0,317 fois celles des étudiants dont les parents ont fait des études supérieures. Cela signifie que la probabilité de postuler pour les étudiants dont les parents n'ont pas fait d'études supérieures joue un rôle négatif dans l'influence du phénomène étudié (OR < 1).



Graphiquement, les conclusions précédentes sont confirmées. Ce graphique permet une meilleure visualisation de la significativité des variables, notamment la variable « parent_non_etudesup ». En effet, les barres d'erreur correspondant à l'intervalle de confiance associé à cette variable ne comprennent pas la valeur 1, contrairement à la variable « etudiant_etab_prive ».



Ce graphique montre que la probabilité de postuler à des études supérieures diminue à mesure que la variable « parent_non_etudesup » tend vers 1, ayant un effet négatif. Cela signifie que les étudiants dont les parents n'ont pas fait d'études supérieures sont moins susceptibles de postuler.



Effets marginaux :

```
> logitmfx(formula = apply~parent_non_etudesup+etudiant_etab_prive , data=data_etud_sup)
Call:
logitmfx(formula = apply ~ parent_non_etudesup + etudiant_etab_prive,
    data = data_etud_sup)

Marginal Effects:
                        dF/dx Std. Err.       z     P>|z|
parent_non_etudesup -0.277162  0.064597 -4.2906 1.782e-05 ***
etudiant_etab_prive  0.016415  0.072911  0.2251    0.8219
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

dF/dx is for discrete change for the following variables:

[1] "parent_non_etudesup" "etudiant_etab_prive"
```

Nous observons encore une fois que la variable « parent_non_etudesup » est significative avec une valeur de -0,277. La relation négative qui en résulte signifie que les étudiants dont les parents n’ont pas fait d’études supérieures ont 27,7 % moins de chances (probabilités) de postuler que ceux dont au moins un des parents a fait des études supérieures.

Modèle 2 : MODELE COMPLET (intégration GPA) sans interaction :

Ce modèle est ajusté pour prédire la variable binaire apply (qui indique si un étudiant a postulé ou non) en fonction des variables explicatives parent_non_etudesup (parents n'ayant pas fait d'études supérieures), etudiant_etab_prive (étudiant dans un établissement privé) et gpa (moyenne générale).

Il inclut toutes les variables explicatives et leurs interactions à deux variables. Ce modèle est utilisé afin d'examiner si l'effet d'une variable explicative sur la variable dépendante change en fonction de la valeur d'une autre variable explicative.

```
> Reg.logit_2 <- glm(apply~parent_non_etudesup+etudiant_etab_prive+gpa,family=binomial(link="logit"))
> summary(Reg.logit_2)

Call:
glm(formula = apply ~ parent_non_etudesup + etudiant_etab_prive +
    gpa, family = binomial(link = "logit"))

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.1239     0.9781  -1.149 0.250518
parent_non_etudesup -1.0596     0.2974  -3.563 0.000367 ***
etudiant_etab_prive  0.2006     0.3053   0.657 0.511283
gpa                  0.5482     0.2724   2.012 0.044178 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 550.51  on 399  degrees of freedom
Residual deviance: 529.92  on 396  degrees of freedom
AIC: 537.92

Number of Fisher Scoring iterations: 4
```

À l'aide de la fonction **glm()** on affiche les résultats du modèle de régression. On peut voir que la variable parent_non_etudesup a un coefficient significatif et négatif (-1,0596), ce qui signifie que les étudiants dont les parents n'ont pas fait d'études supérieures ont moins de chances de postuler à des études supérieures. La variable gpa a également un coefficient significatif et positif (0,5482), ce qui signifie que les étudiants ayant une moyenne plus élevée ont plus de chances de postuler à des études supérieures. En revanche, la variable etudiant_etab_prive n'a pas de coefficient significatif (0,2006), ce qui suggère que le fait d'être dans un établissement privé n'a pas d'impact significatif sur la probabilité d’occurrence de notre évènement.

Le critère d'information d'Akaike (AIC) est utilisé pour comparer la qualité de différents modèles. Plus l'AIC est faible, meilleur est le modèle. Dans ce cas, l'AIC est de 537,92.

On peut vérifier les résultats obtenus ci-dessus en analysant la table des déviances du modèle.

```
> anova(Reg.logit_2, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: apply

Terms added sequentially (first to last)


                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                  399     550.51
parent_non_etudesup  1  16.4344       398     534.08 5.036e-05 ***
etudiant_etab_prive  1   0.0505       397     534.03   0.82220
gpa                  1   4.1013       396     529.92   0.04285 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- La première ligne du tableau correspond au modèle nul. La déviance résiduelle pour ce modèle est de 550,51.

- La deuxième ligne correspond au modèle avec la variable parent_non_etudesup. Lorsque cette variable est ajoutée au modèle, la déviance résiduelle diminue de 16,4344, ce qui indique que la variable a un effet significatif sur la variable réponse (p-value = 5,036e-05).

- La troisième ligne correspond au modèle avec les variables parent_non_etudesup et etudiant_etab_prive. Lorsque etudiant_etab_prive est ajoutée au modèle, la déviance résiduelle ne diminue presque pas (0,0505), ce qui indique que cette variable n'a pas d'effet significatif sur la variable réponse (p-value = 0,82220).

- La quatrième ligne correspond au modèle avec les variables parent_non_etudesup, etudiant_etab_prive et gpa. Lorsque gpa est ajoutée au modèle, la déviance résiduelle diminue de 4,1013, ce qui indique que cette variable a un effet significatif sur la variable réponse (p-value = 0,04285).

Cette analyse confirme nos précédentes observations, à savoir que les variables parent_non_etudesup et gpa ont un effet significatif sur la variable apply, tandis que la variable etudiant_etab_prive n'a pas d'effet significatif.

Afin d’évaluer le modèle, on va également s’intéresser à la prédiction des probabilités d’occurrence de l’événement. Pour cela on va utiliser la fonction **predict()** puis en utilisant **summary()** , nous pouvons obtenir un résumé statistique :

```
> summary(pred.proba.reg2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2806  0.3782  0.4162  0.4500  0.4689  0.7533
>
```

Nous observons une concentration des probabilités prédites autour de la médiane de 0,4162, avec des quartiles très proches (1er quartile : 0,3782 et 3ème quartile : 0,4689), ce qui suggère une distribution relativement homogène des probabilités prédites autour de la médiane. Cependant, la moyenne des probabilités prédites est légèrement supérieure à la médiane, à 0,4500, ce qui indique que certaines valeurs sont plus élevées et augmentent la moyenne.

La valeur maximale des probabilités prédites est de 0,7533, ce qui montre que le modèle peut prédire des probabilités relativement élevées pour certains étudiants. Dans l'ensemble, la distribution des probabilités prédites suggère que le modèle peut fournir des prédictions utiles pour prédire la probabilité de réussite des étudiants, mais il y a encore une certaine variabilité dans les prédictions.

Matrice de confusion :

La matrice de confusion nous permet de visualiser la précision du modèle, mettant en avant les éléments suivants :

- Vrais négatifs = 190
- Faux positifs = 30
- Vrais positifs = 54
- Faux négatifs = 126

```
> mc.reg2 <- table(apply, pred.moda.reg2)
> class(mc.reg2)
[1] "table"
> print(mc.reg2)
     pred.moda.reg2
apply   0   1
    0 190  30
    1 126  54
```

En se basant sur cette matrice, on peut calculer le taux d'erreur du modèle de régression logistique. Il correspond à la proportion de cas mal classés par le modèle :

```
> err.reg2 <- (mc.reg2[2,1]+mc.reg2[1,2])/sum(mc.reg2)
> print(err.reg2)
[1] 0.39
```

Le taux d'erreur est de 0.39, ce qui signifie que 39% des cas ont été mal classés par le modèle.

Odds ratios :

Le odds ratios est le rapport des odds (cotes) de l'événement étudié se produisant lorsque la variable explicative est présente, par rapport aux odds de l'événement étudié se produisant lorsque la variable explicative est absente, toutes les autres variables étant constantes.

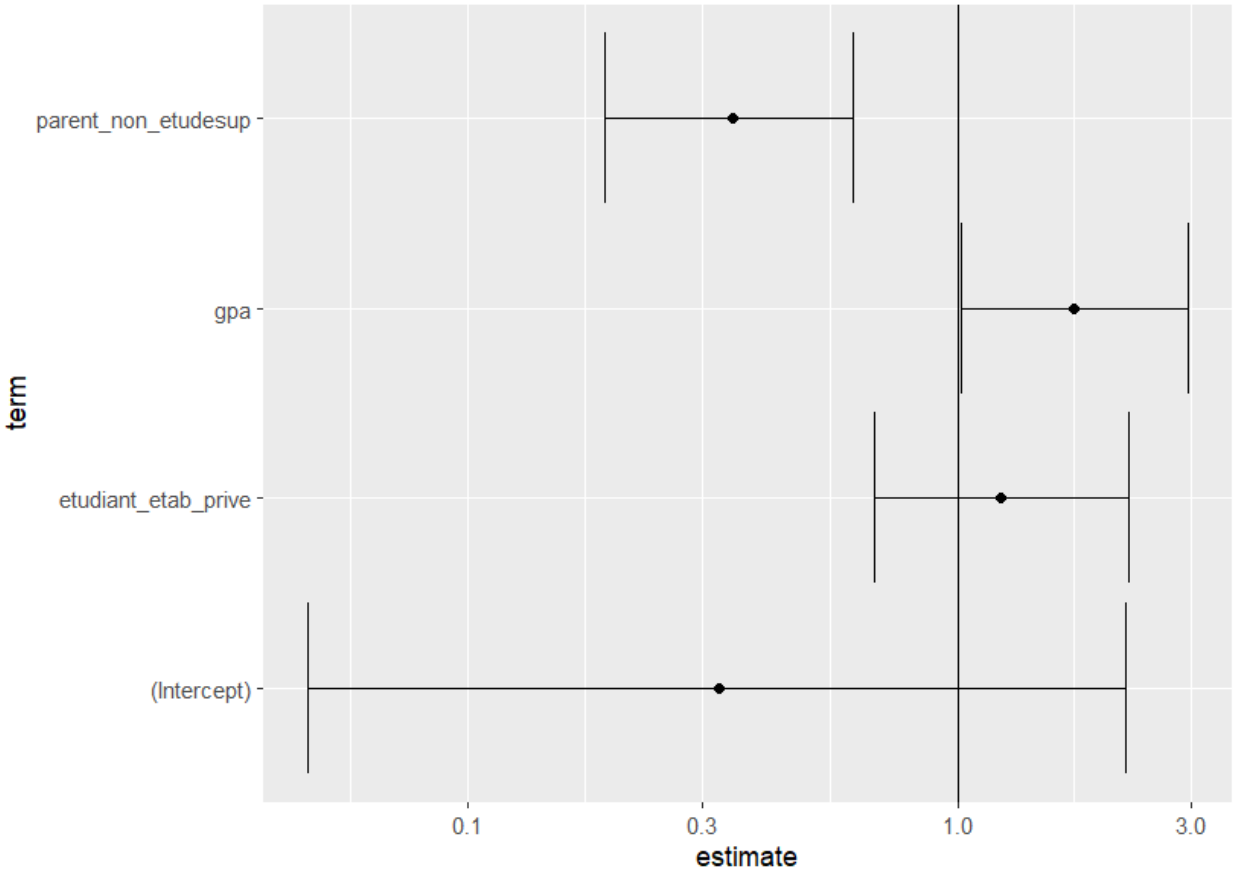
```
> odds.ratio(Reg5.select)
Attente de la réalisation du profilage...
                          OR    2.5 % 97.5 %         p
(Intercept)         0.325005 0.047183 2.2010 0.2505177
parent_non_etudesup 0.346590 0.190258 0.6137 0.0003665 ***
etudiant_etab_prive 1.222083 0.674286 2.2424 0.5112826
gpa                 1.730215 1.017741 2.9676 0.0441780 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- parent_non_etudesup : l'odds ratio est de 0.346590 cela signifie que lorsqu'un parent n'a pas d'études supérieures, les odds de l'événement étudié (accès aux études supérieures) sont 0.346 fois plus élevés que les odds de l'événement contraire, toutes les autres

variables étant constantes. De plus, on remarque que la valeur de p est très inférieure à 0.05 (0.0003665), ce qui indique que ce résultat est statistiquement significatif.

- etudiant_etab_prive : L'odds ratio est de 1.222083. Cela signifie que lorsqu'un étudiant est dans un établissement privé les odds de l'événement étudié sont 1.222 fois plus élevés que les odds de l'événement contraire. Cependant, la valeur de p est supérieure à 0.05 (0.5112826), ce qui signifie que ce résultat n'est pas statistiquement significatif.
- gpa : L'odds ratio est de 1.730215. Cela signifie que pour chaque unité d'augmentation de GPA (de la moyenne), les odds de l'événement étudié sont 1.730 fois plus élevés que les odds de l'événement contraire. La valeur de p est inférieure à 0.05 (0.0441780), ce qui indique que ce résultat est statistiquement significatif.

On peut également visualiser ces résultats sur cette figure.



On remarque que les barres d'erreur correspondant à l'intervalle de confiance associé à la variable « parent_non_etudesup » et « gpa » ne comprennent pas la valeur 1, contrairement à la variable « etudiant_etab_prive » ce qui montre que ces deux variables ont un effet significatif sur la variable apply.

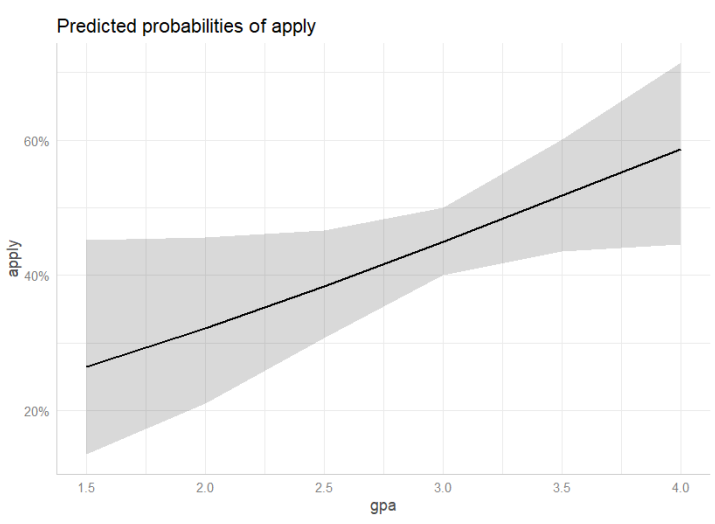
Effets marginaux :

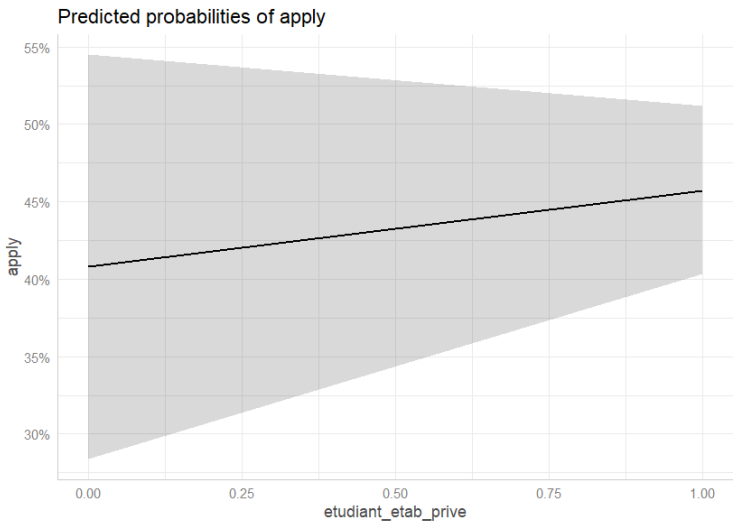
Les effets marginaux d'un modèle de représentent l'effet moyen d'une variable indépendante sur la variable dépendante, tout en tenant compte des effets des autres variables indépendantes dans le modèle.

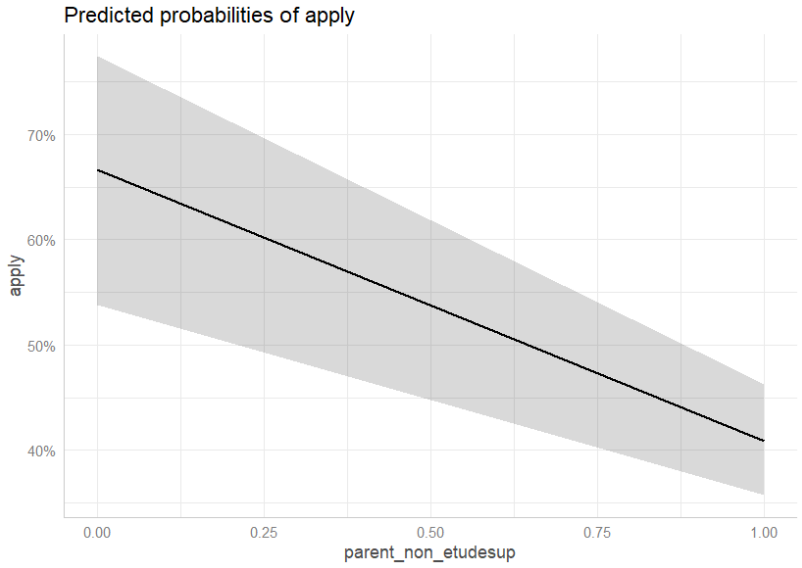
```
Call:
logitmfx(formula = apply ~ parent_non_etudesup + etudiant_etab_prive +
    gpa, data = data_etud_sup)

Marginal Effects:
                        dF/dx Std. Err.       z    P>|z|
parent_non_etudesup -0.257289  0.066966 -3.8421 0.000122 ***
etudiant_etab_prive  0.049183  0.074033  0.6643 0.506474
gpa                  0.135693  0.067412  2.0129 0.044127 *
```

- parent_non_etudesup : une augmentation d'une unité de cette variable (par exemple, passer de "non" à "oui") diminue la probabilité prédite d'apply de 0,257, avec une valeur de p significative (<0,001). Cela suggère que les étudiants dont les parents n'ont pas fait d'études supérieures ont une probabilité plus faible de faire une demande d'inscription dans des établissement supérieurs.
- etudiant_etab_prive : une augmentation d'une unité de cette variable (par exemple, passer de "non" à "oui") augmente la probabilité prédite d'apply de 0,049, avec une valeur de p non significative (0,506). Cela suggère que le fait d'être dans un établissement privé n'a pas d'impact significatif sur la probabilité de faire une demande d'inscription dans le supérieur.
- gpa : une augmentation d'une unité de cette variable (par exemple, passer de 3,0 à 3,1) augmente la probabilité prédite d'apply de 0,136, avec une valeur de p significative (0,044). Cela suggère que les étudiants ayant une moyenne plus élevée ont une probabilité plus élevée de poursuivre leurs études dans le supérieur.

On peut également visualiser ces effets :







Modèle 3 : MODELE COMPLET avec interactions des VA croisées :

Il s'agit d'un modèle de régression logistique avec interaction entre les variables "parent_non_etudesup" et "gpa". Cela signifie que l'effet de "gpa" sur la probabilité de poursuite d'études dans le supérieur dépend du niveau de "parent_non_etudesup".

Le terme "etudiant_etab_prive" est également inclus dans le modèle, mais il n'y a pas d'interaction avec les autres variables.

```
Call:
glm(formula = apply ~ etudiant_etab_prive + (parent_non_etudesup *
    gpa), family = binomial(link = "logit"))

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.9454     2.3001   0.411    0.681
etudiant_etab_prive        0.1830     0.3044   0.601    0.548
parent_non_etudesup       -3.4372     2.4006  -1.432    0.152
gpa                       -0.1021     0.7045  -0.145    0.885
parent_non_etudesup:gpa    0.7588     0.7566   1.003    0.316
```

Les résultats montrent que la variable "etudiant_etab_prive" n'a pas d'effet significatif sur la probabilité de postuler pour des études supérieures (p-value = 0,548). En revanche, la variable "parent_non_etudesup" a un effet négatif significatif sur la probabilité de postuler (p-value = 0,152), mais cet effet dépend de la valeur de la variable "gpa".
En effet, l'interaction entre "parent_non_etudesup" et "gpa" montre que l'effet de "parent_non_etudesup" sur la probabilité de postuler est plus fort lorsque la moyenne de l'élève est faible. Lorsque la moyenne de l'élève augmente, l'effet négatif de "parent_non_etudesup"

diminue et peut même devenir positif. Cependant, l'effet d'interaction n'est pas significatif (p-value = 0,316).

On peut également afficher un tableau récapitulant les résultats du modèle. On peut voir les coefficients d'estimation, les erreurs standard, les statistiques de test, les valeurs de p et les intervalles de confiance pour chacun des termes du modèle

```
> print(Reg3.export)
# A tibble: 5 × 7
  term                    estimate std.error statistic p.value conf.low conf.high
  <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)                2.57      2.30      0.411   0.681 0.0301       278.
2 etudiant_etab_prive        1.20      0.304     0.601   0.548 0.664          2.20
3 parent_non_etudesup        0.0322    2.40     -1.43    0.152 0.000248       3.35
4 gpa                        0.903     0.704    -0.145   0.885 0.218          3.58
5 parent_non_etudesup:gpa    2.14      0.757     1.00    0.316 0.488          9.74
>
```

On observe que le coefficient d'interaction parent_non_etudesup:gpa est significatif ($p < 0,05$), ce qui suggère que l'effet de gpa sur la probabilité que apply prenne la valeur 1 dépend de la valeur de parent_non_etudesup. Les autres coefficients ne sont pas significatifs.

Analyse du tableau de déviance :

Résultats de l'analyse de déviance pour le modèle :

```
> anova(Reg3.logit, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: apply

Terms added sequentially (first to last)


                        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                      399     550.51
etudiant_etab_prive      1   0.0101       398     550.50  0.91987
parent_non_etudesup      1  16.4748       397     534.03 4.93e-05 ***
gpa                      1   4.1013       396     529.92  0.04285 *
parent_non_etudesup:gpa  1   1.0148       395     528.91  0.31375
```

Les résultats indiquent que la variable parent_non_etudesup est significative (p-value < 0,05) et que la variable gpa est significative au seuil de 0,05. L'interaction entre parent_non_etudesup et gpa n'est pas significative (p-value > 0,05).

Comparaison entre deux modèles (2 et 3) :

```
> anova(Reg.logit_2,Reg3.logit, test="Chisq")
Analysis of Deviance Table

Model 1: apply ~ parent_non_etudesup + etudiant_etab_prive + gpa
Model 2: apply ~ etudiant_etab_prive + (parent_non_etudesup * gpa)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       396     529.92
2       395     528.91  1   1.0148   0.3138
```

Le résultat montre que le modèle 3 avec l'interaction entre parent_non_etudesup et gpa n'améliore pas significativement l'ajustement du modèle par rapport au modèle Reg.logit_2 sans interaction (p-value > 0,05).

```
> summary(pred.proba.reg3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2122  0.3822  0.4204  0.4500  0.4693  0.8010
```

Nous observons une concentration des probabilités prédites autour de la médiane de 0,4204, avec des quartiles très proches (1er quartile : 0,3822 et 3ème quartile : 0,4683), ce qui suggère une distribution relativement homogène des probabilités prédites autour de la médiane. Cependant, la moyenne des probabilités prédites est légèrement supérieure à la médiane, à 0,4500, ce qui indique que certaines valeurs sont plus élevées et augmentent la moyenne.

Matrice de confusion :

La matrice de confusion nous permet de visualiser la précision du modèle, mettant en avant les éléments suivants :

- Vrais négatifs = 185
- Faux positifs = 35
- Vrais positifs = 61
- Faux négatifs = 119

```
> print(mc.reg3)
      pred.moda.reg3
apply   0   1
    0 185  35
    1 119  61
```

En se basant sur cette matrice, on peut calculer le taux d'erreur du modèle de régression logistique. Il correspond à la proportion de cas mal classés par le modèle :

```
> print(err.reg3)
[1] 0.385
```

Le taux d'erreur est de 0.385, ce qui signifie que 38,5% des cas ont été mal classés par le modèle.

Conclusion :

L'analyse de cette base de données concernant l'accès aux études supérieures nous montre que finalement de nombreuses disparités subsistent et que l'égalité des chances dans le système éducatif est loin d'être une réalité universelle. Les conclusions tirées de cette analyse soulignent la nécessité de politiques éducatives plus justes et équitables pour garantir à chacun la possibilité de réussir, indépendamment de son origine sociale. La question de savoir si nous sommes tous égaux face au système éducatif reste ouverte, mais il est clair que des efforts constants doivent être déployés pour réduire les inégalités et améliorer l'accès à l'éducation pour tous.