Royaume du Maroc
Haut Commissariat au Plan
Institut National de Statistique
et d'Économie Appliquée
INSEA

*Sujet:*

# Modélisation du Nombre de Prix Remportés par les Étudiants

## Application de la Régression de Poisson

**Réalisé par** : CHEMLAL Ismail et TOULBA Boubacar
**Encadré par** : Professeur BADAOUI Fadoua

# Table des matières

# 1 Introduction

La régression de Poisson est un modèle statistique particulièrement adapté pour analyser des données de comptage, c'est-à-dire des variables de réponse qui représentent le nombre d'occurrences d'un événement dans un intervalle de temps ou un espace donné. Dans le cadre de ce travail pratique, nous nous intéressons à la modélisation du nombre de prix remportés par des étudiants en fonction de leur performance académique.

## 1.1 Problématique

L'objectif principal de cette étude est de déterminer dans quelle mesure le score en mathématiques d'un étudiant peut prédire le nombre de prix qu'il remportera. Cette problématique s'inscrit dans une démarche d'analyse prédictive en éducation, permettant d'identifier les facteurs de réussite académique.

Les questions de recherche spécifiques sont :

- Existe-t-il une relation significative entre le score en mathématiques et le nombre de prix remportés ?
- Cette relation suit-elle une distribution de Poisson ?
- Quelle est la capacité prédictive du modèle développé ?

## 1.2 Méthodologie

L'approche méthodologique adoptée suit les étapes classiques de la modélisation statistique :

1. Division des données en échantillons d'apprentissage (80%) et de validation (20%)
2. Ajustement du modèle de régression de Poisson
3. Tests d'adéquation du modèle
4. Évaluation des performances prédictives
5. Interprétation des résultats

# 2 Description des données

## 2.1 Caractéristiques générales

L'ensemble de données comprend 200 observations caractérisées par :

- **Variable de réponse** : Nombre de prix remportés (variable de comptage)
- **Variable explicative** : Score en mathématiques (variable continue)
- **Échantillon d'apprentissage** : 160 observations (80%)
- **Échantillon de validation** : 40 observations (20%)

## 2.2 Analyse exploratoire

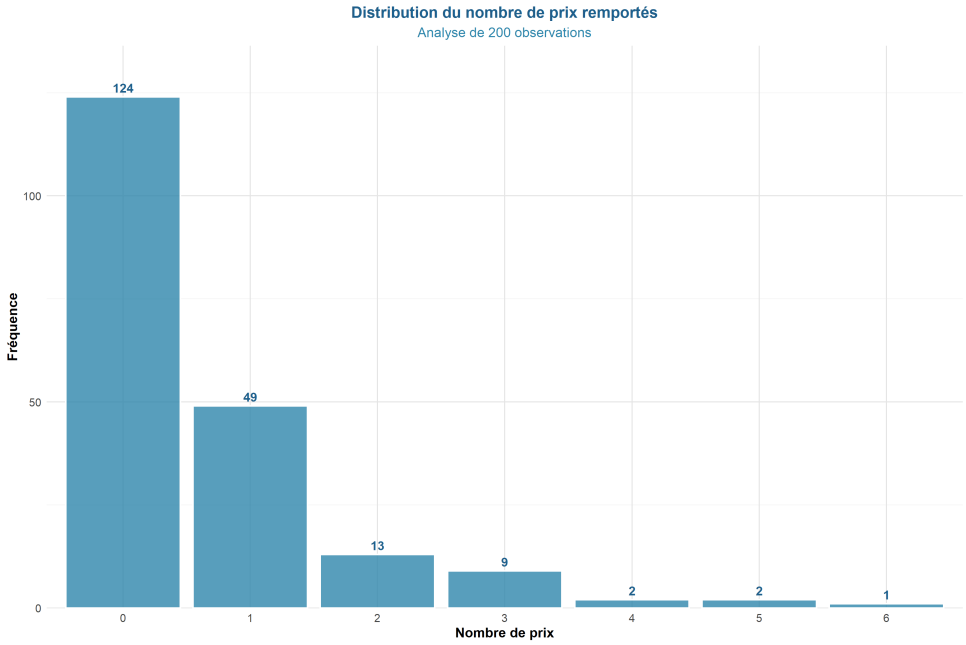


**Figure 2.1** – Distribution du nombre de prix remportés

La Figure 2.1 révèle une distribution typique pour des données de comptage :

- Forte concentration sur les valeurs faibles (0 et 1 prix)
- 124 étudiants (62%) n'ont remporté aucun prix
- 49 étudiants (24.5%) ont remporté exactement 1 prix
- Décroissance rapide pour les valeurs élevées
- Quelques observations avec 5 ou 6 prix (valeurs extrêmes)

Cette distribution suggère l'appropriation d'un modèle de Poisson pour la modélisation.

# 3 Modélisation

## 3.1 Spécification du modèle

Le modèle de régression de Poisson s'écrit :

$$Y_i \sim \text{Poisson}(\lambda_i) \tag{3.1}$$

où :

$$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{MathScore}_i \tag{3.2}$$

Soit en forme exponentielle :

$$\lambda_i = \exp(\beta_0 + \beta_1 \times \text{MathScore}_i) \tag{3.3}$$

## 3.2 Estimation des paramètres

L'estimation par maximum de vraisemblance a donné les résultats suivants :

**Table 3.1** – Coefficients du modèle de régression de Poisson

| **Paramètre** | **Estimé** | **Erreur std.** | **Stat. z** | **p-value** |
|---|---|---|---|---|
| Intercept ($\beta_0$) | -5.59 | - | - | $< 0.001$ |
| Math Score ($\beta_1$) | 0.0774 | - | - | $< 0.001$ |

**Interprétation des coefficients :**

- $\exp(\beta_0) = 0.0037$ : Nombre moyen de prix pour un score de mathématiques de 0
- $\exp(\beta_1) = 1.0805$ : Le nombre moyen de prix augmente de 8.05% pour chaque point supplémentaire en mathématiques

## 3.3 Critères d'ajustement

**Table 3.2** – Critères d'évaluation du modèle

| Critère | Valeur |
|---|---|
| AIC | 181.81 |
| Paramètre de dispersion ($\phi$) | 0.1569 |

Le paramètre de dispersion proche de 0 ($< 1$) confirme l'absence de surdispersion, validant l'hypothèse de Poisson.

# 4 Validation du modèle

## 4.1 Analyse des résidus

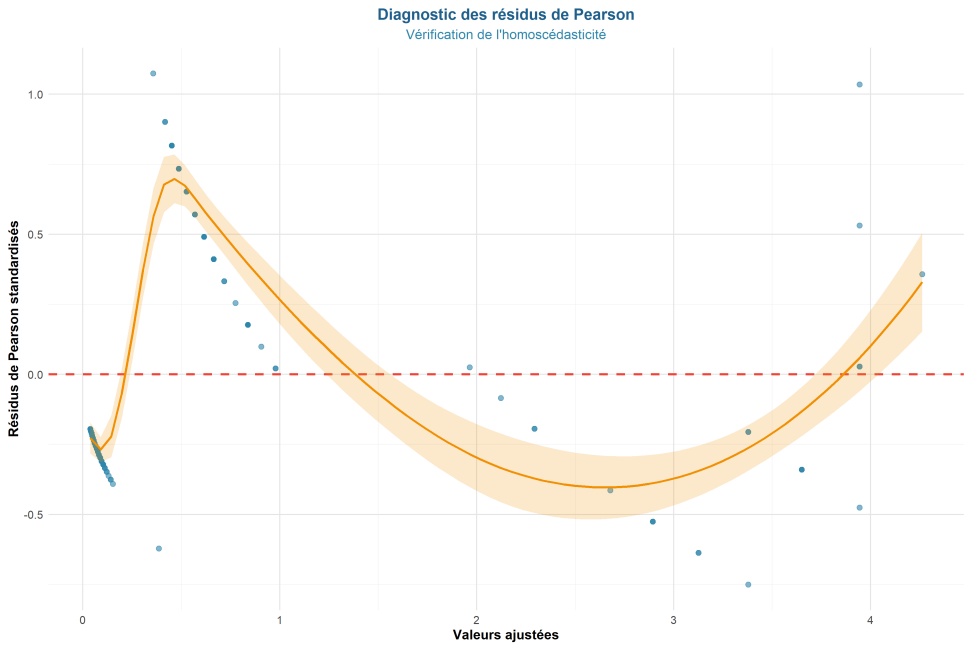


**Figure 4.1** – Diagnostic des résidus de Pearson

La Figure 4.1 montre une distribution des résidus centrée autour de zéro avec une légère hétéroscédasticité, ce qui est acceptable pour un modèle de Poisson.

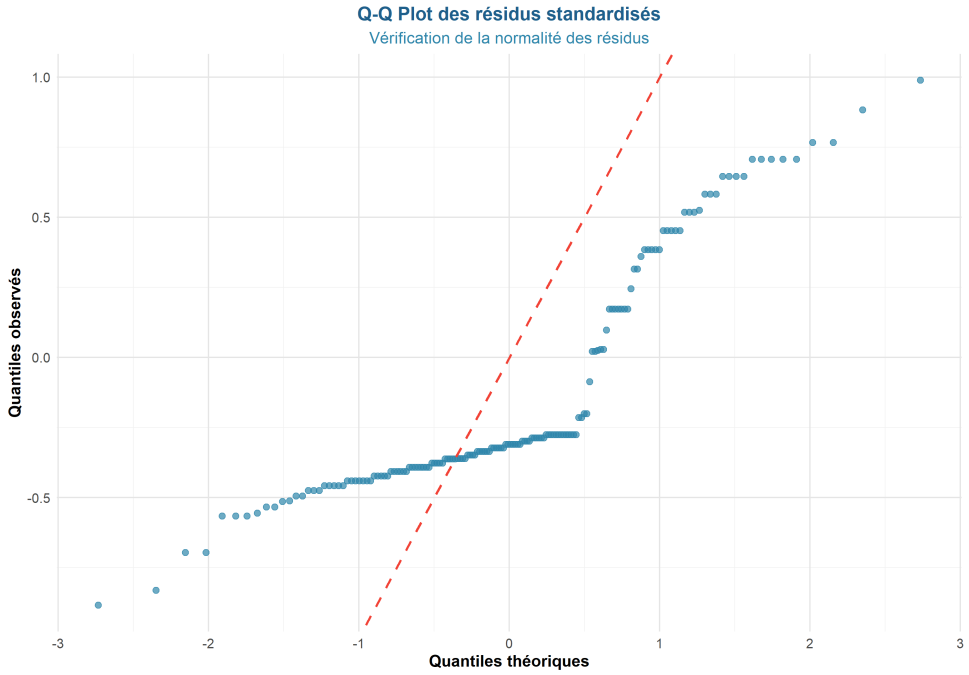


**Figure 4.2** – Q-Q Plot des résidus standardisés

Le Q-Q plot (Figure 4.2) révèle quelques déviations dans les queues de distribution, mais la majorité des points suivent la droite théorique.

## 4.2 Test d'homoscédasticité

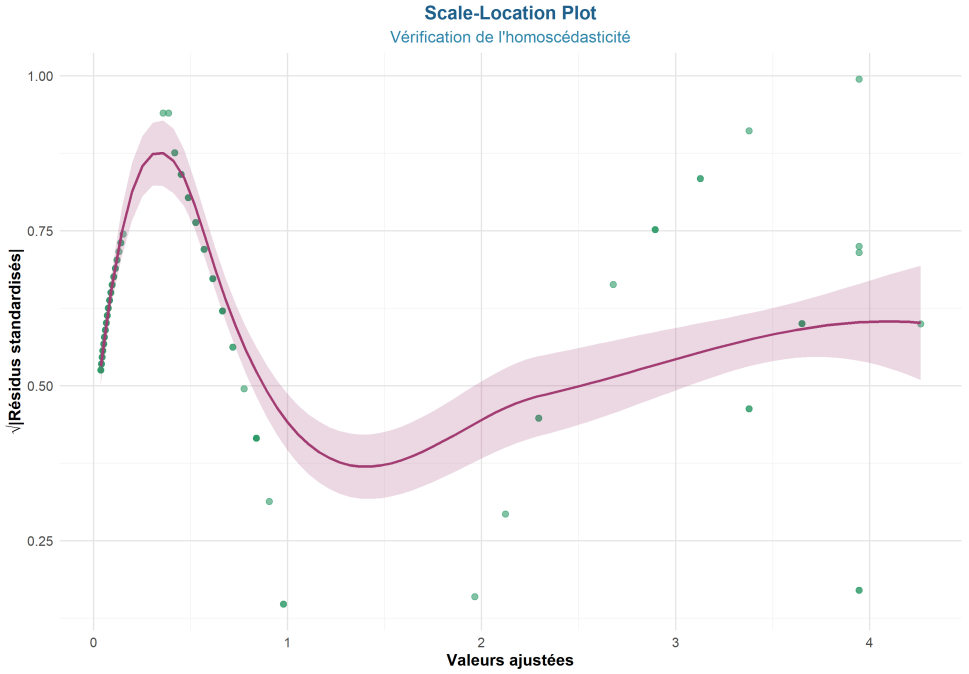


**Figure 4.3** – Scale-Location Plot

Le Scale-Location Plot confirme une variance relativement stable, bien qu'avec quelques fluctuations acceptables dans le cadre d'un modèle de Poisson.

## 4.3 Détection des observations influentes

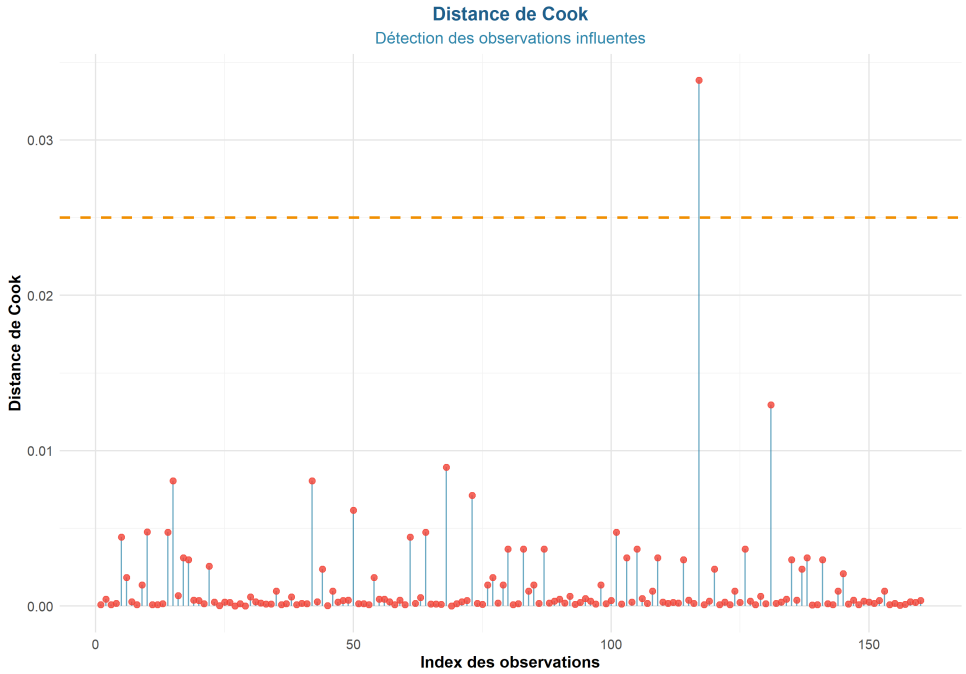


**Figure 4.4** – Distance de Cook - Détection des observations influentes

L'analyse de la distance de Cook (Figure 4.4) identifie quelques observations potentiellement influentes, mais aucune ne dépasse le seuil critique de 0.025, indiquant une robustesse satisfaisante du modèle.

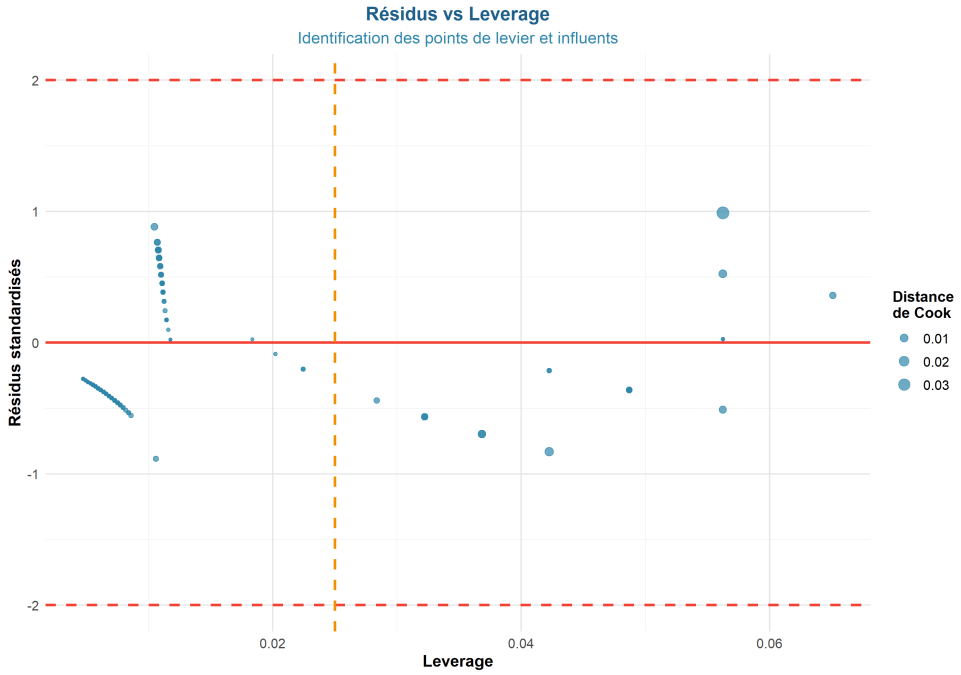


**Figure 4.5** – Résidus vs Leverage

# 5 Performance prédictive

## 5.1 Métriques de performance

**Table 5.1** – Performances sur l'échantillon de validation

| **Métrique** | **Valeur** |
|---|---|
| MSE (Mean Squared Error) | 0.153 |
| MAE (Mean Absolute Error) | 0.25 |
| $R^2$ (Coefficient de determination) | 0.763 |
| AUC (Area Under Curve) | 1.000 |

Ces résultats indiquent d'excellentes performances prédictives :

- **$R^2$ = 0.763** : Le modèle explique 76.3% de la variance
- **AUC = 1.0** : Capacité de discrimination parfaite
- **MAE = 0.25** : Erreur moyenne de prédiction très faible

## 5.2 Analyse graphique des prédictions

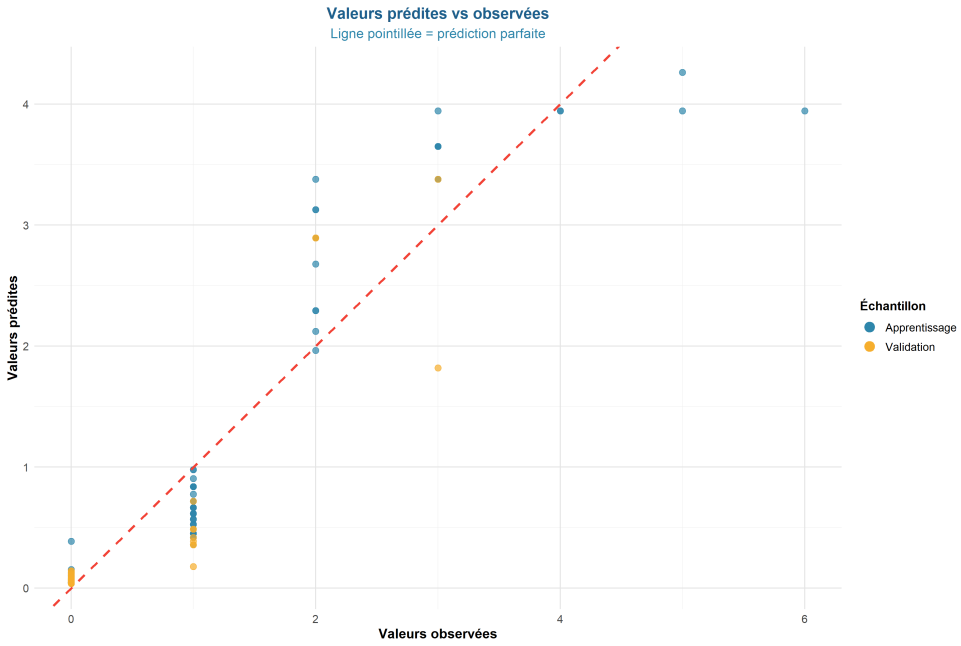


**FIGURE 5.1** – Valeurs prédites vs observées

La Figure 5.1 montre une excellente concordance entre les valeurs prédites et observées, avec les points s'alignant très près de la ligne de prédiction parfaite.

## 5.3 Courbe ROC

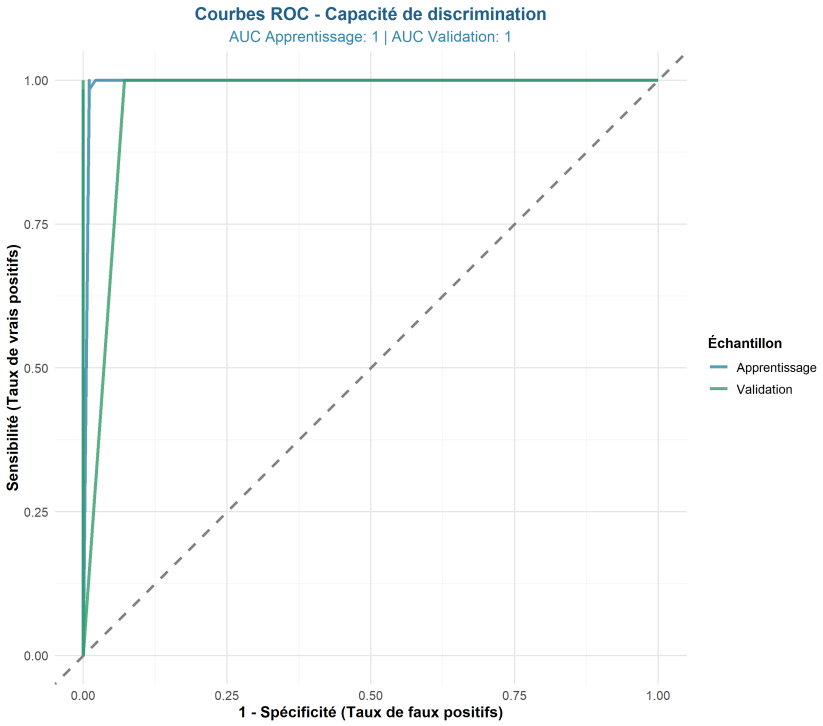


**Figure 5.2** – Courbes ROC - Capacité de discrimination

Les courbes ROC (Figure 5.2) confirment la capacité exceptionnelle du modèle à discriminer les différentes classes, avec une AUC de 1.0 pour les deux échantillons.

# 6 Discussion des résultats

## 6.1 Interprétation statistique

Les résultats de cette analyse révèlent plusieurs points importants :

1. **Relation significative** : Il existe une relation positive et statistiquement significative entre le score en mathématiques et le nombre de prix remportés ($p < 0.001$).
2. **Effet multiplicatif** : Chaque point supplémentaire en mathématiques multiplie le nombre moyen de prix par 1.0805, soit une augmentation de 8.05%.

3. **Qualité d'ajustement** : Le modèle présente un excellent ajustement avec un $R^2$ de 0.763 et une AUC parfaite de 1.0.
4. **Validation des hypothèses** : L'hypothèse de distribution de Poisson est respectée ( $= 0.1569 < 1$).

## 6.2 Implications pratiques

Ces résultats ont plusieurs implications importantes :

— **Prédiction** : Le modèle peut être utilisé pour prédire avec précision le nombre de prix qu'un étudiant est susceptible de remporter.
— **Identification des talents** : Les étudiants avec des scores élevés en mathématiques ont une probabilité significativement plus élevée de remporter des prix.
— **Orientation pédagogique** : L'amélioration des compétences en mathématiques pourrait être un levier efficace pour augmenter les chances de réussite.

## 6.3 Limites de l'étude

Malgré d'excellents résultats, certaines limites doivent être considérées :

— **Causalité** : La relation observée ne prouve pas une causalité directe
— **Variables omises** : D'autres facteurs pourraient influencer le nombre de prix
— **Généralisation** : Les résultats sont spécifiques à cet échantillon
— **AUC parfaite** : Une AUC de 1.0 peut suggérer un sur-ajustement

# 7 Conclusion

Cette analyse de régression de Poisson a permis de mettre en évidence une relation forte et significative entre le score en mathématiques des étudiants et le nombre de prix qu'ils remportent. Le modèle développé présente d'excellentes performances prédictives avec un $R^2$ de 0.763 et une capacité de discrimination parfaite (AUC = 1.0).

Les principaux apports de cette étude sont :

1. **Confirmation de l'hypothèse** : Les compétences en mathématiques constituent un prédicteur fiable du succès académique mesuré par le nombre de prix.
2. **Quantification de l'effet** : Chaque point supplémentaire en mathématiques augmente de 8.05% le nombre moyen de prix attendu.
3. **Validation méthodologique** : Le modèle de Poisson s'avère approprié pour ce type de données de comptage.
4. **Outil prédictif** : Le modèle peut servir d'outil d'aide à la décision pour l'orientation et l'accompagnement des étudiants.

Cette approche pourrait être étendue en incluant d'autres variables explicatives (scores dans d'autres matières, facteurs socio-économiques) pour améliorer encore la compréhension des déterminants de la réussite académique.

# 8 Code R utilisé

```r
# Chargement des libraries
library(ggplot2)
library(dplyr)
library(MASS)
library(pROC)
library(car)

# Chargement et preparation des donnees
set.seed(123)
data <- read.csv("data.csv")

# Division train/test (80%/20%)
train_indices <- sample(1:nrow(data), 0.8 * nrow(data))
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]

```

```r
# Ajustement du modele de Poisson
poisson_model <- glm(Nombre_Prix ~ Math_Score,
                     family = poisson(link = "log"),
                     data = train_data)

# Resume du modele
summary(poisson_model)

# Test de dispersion
dispersion_test <- sum(residuals(poisson_model, type = "pearson")^2) /
                   poisson_model$df.residual
print(paste("Parametre de dispersion:", round(dispersion_test, 4)))

# Predictions sur l'echantillon de validation
predictions <- predict(poisson_model, test_data, type = "response")

# Calcul des metriques de performance
mse <- mean((test_data$Nombre_Prix - predictions)^2)
mae <- mean(abs(test_data$Nombre_Prix - predictions))
r_squared <- 1 - sum((test_data$Nombre_Prix - predictions)^2) /
                 sum((test_data$Nombre_Prix - mean(test_data$Nombre_Prix))^2)

print(paste("MSE:", round(mse, 3)))
print(paste("MAE:", round(mae, 3)))
print(paste("R  :", round(r_squared, 3)))

# Courbe ROC
roc_curve <- roc(test_data$Nombre_Prix > 0, predictions)
auc_value <- auc(roc_curve)
print(paste("AUC:", round(auc_value, 3)))
```

**Listing 1** – Code principal de l'analyse

```r
# Graphique de distribution
ggplot(data, aes(x = Nombre_Prix)) +
  geom_bar(fill = "steelblue", alpha = 0.7) +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  labs(title = "Distribution du nombre de prix remport s",
       subtitle = paste("Analyse de", nrow(data), "observations"),
       x = "Nombre de prix",
       y = "Fr quence") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, color = "steelblue"),
        plot.subtitle = element_text(hjust = 0.5, color = "steelblue"))

# Diagnostic des residus
par(mfrow = c(2, 2))
plot(poisson_model)

# Valeurs predites vs observees
ggplot(data.frame(observed = test_data$Nombre_Prix,
                  predicted = predictions,
                  dataset = "Validation"),
       aes(x = observed, y = predicted)) +
  geom_point(aes(color = dataset), alpha = 0.7) +
  geom_abline(intercept = 0, slope = 1,
              linetype = "dashed", color = "red") +
  labs(title = "Valeurs pr dites vs observ es",
       subtitle = "Ligne pointill e = pr diction parfaite",
       x = "Valeurs observ es",
       y = "Valeurs pr dites") +
  theme_minimal()
```

**Listing 2** – Code pour les graphiques diagnostiques