Régression de Poisson

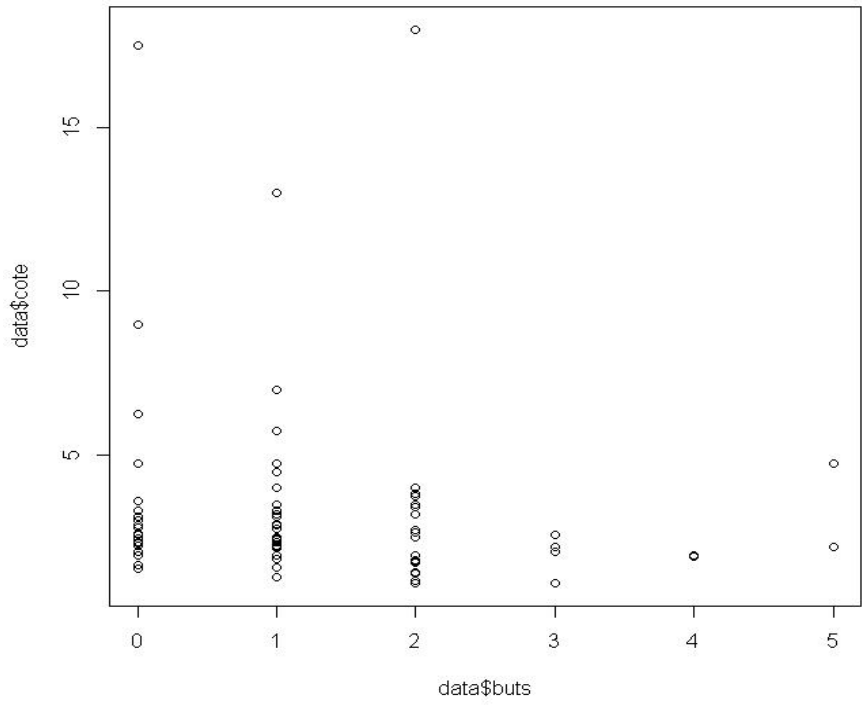
Anis TRABELSI
Romain RESPRIGET

Introduction

Le nombre de buts marqués par une équipe de football à domicile peut-il se modéliser grâce à une régression linéaire en fonction de la cote du match ou non ?

Introduction

Graphe du nombre de buts de matchs d'une journée de championnats européens en fonction de leur cote :



La régression linéaire semble mal adaptée

Plan

I. Présentation générale du modèle de Poisson
   1. Présentation théorique
   2. Présentation des données utilisées
II. Régression de Poisson – Présentation de l'estimation
   1. Théorie
   2. Estimation de notre modèle
III. Interprétation des résultats d'une régression de Poisson
   1. Application à nos données
   2. Limites de notre modèle : Avantages/Inconvénients

I. Présentation générale de la régression de Poisson

1. Présentation théorique
   a. Origine du modèle.

Loi de Poisson → 1838

→ Siméon Denis Poisson (1781-1840)

→ « *Recherches sur la probabilité des jugements en matière criminelle et en matière civile* »

N v.a.r dénombrent le nombre d'occurrences dans un laps de temps donné

I. Présentation générale de la régression de Poisson

1. Présentation théorique
   a. Origine du modèle
   b. Intérêt de la régression de poisson

Modèle de Poisson

→ Nombre de quelque chose ou évènement rare

Régression de Poisson

→Etude du nombre de quelque chose par période t (durée totale et variables explicatives connues)

I. Présentation générale de la régression de Poisson

1. Présentation théorique
   a. Origine du modèle
   b. Intérêt de la régression de poisson

   Exemples d'applications
   i. impact de jouer à domicile et de la cote d'un match sur le nombre de buts marqués
   ii. Nombre de naissances par césarienne public/privé
   iii. Nombre de PV de stationnement Paris/Rennes
   iv. Lien entre la cylindré d'un véhicule et son nombre de PV

I. Présentation générale de la régression de Poisson

1. Présentation théorique
   a. Origine du modèle
   b. Intérêt de la régression de poisson
   c. Présentation du modèle de Poisson

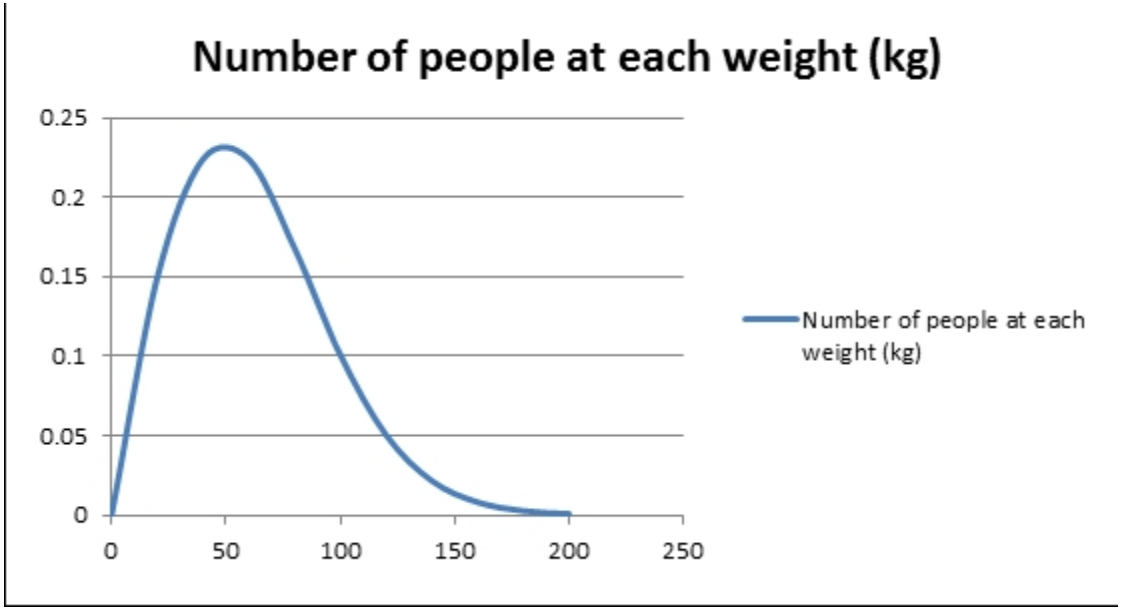
→ La loi de Poisson

Soit λ un réel et Y une variable aléatoire réelle,
$Y \sim P(\lambda)$ si et seulement si
quelque que soit l'entier naturel k, $P(Y = k) = e^{-\lambda} \frac{\lambda^k}{k!}$
En conséquence, on a $E(Y) = V(Y) = \lambda$

I. Présentation générale de la régression de Poisson

1. Présentation théorique
   a. Origine du modèle
   b. Intérêt de la régression de poisson
   c. Présentation du modèle de Poisson

→ Exemple de modélisation de la loi de Poisson



I. Présentation générale de la régression de Poisson

1. Présentation théorique
   a. Origine du modèle
   b. Intérêt de la régression de poisson
   c. Présentation du modèle de Poisson

→ Modèle de régression de Poisson

$$\ln(y) = \propto + \beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_k x_k$$

y réalisation de Y variable endogène suivant une loi de poisson

$\propto$ ordonnée à l'origine

β_i coefficient associé à la ième variable explicative x_i

Modèle « Loglinéaire »

I. Présentation générale de la régression de Poisson

1. Présentation théorique
2. Présentation des données utilisées
   a. Présentation des données

→ Données : Résultats de matchs de football et leurs cotes associées

→ Date : weekend du 1er Mars

→ Source : site des championnats espagnol, italien, français et allemand et le site Bwin

→ Variables : Y nombre de buts marqués par équipe

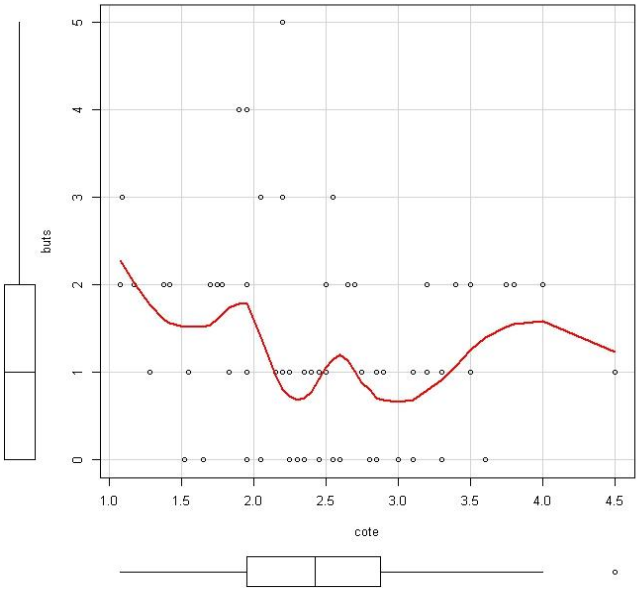
D indicatrice domicile ou non

X cote de l'équipe pour le match

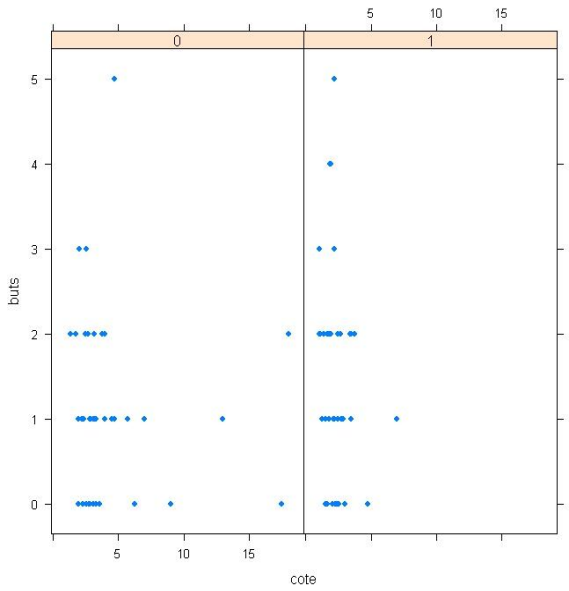
I. Présentation générale de la régression de Poisson

1. Présentation théorique
2. Présentation des données utilisées
   a. Présentation des données
   b. Statistiques descriptives

Nombre de buts en fonction de la cote



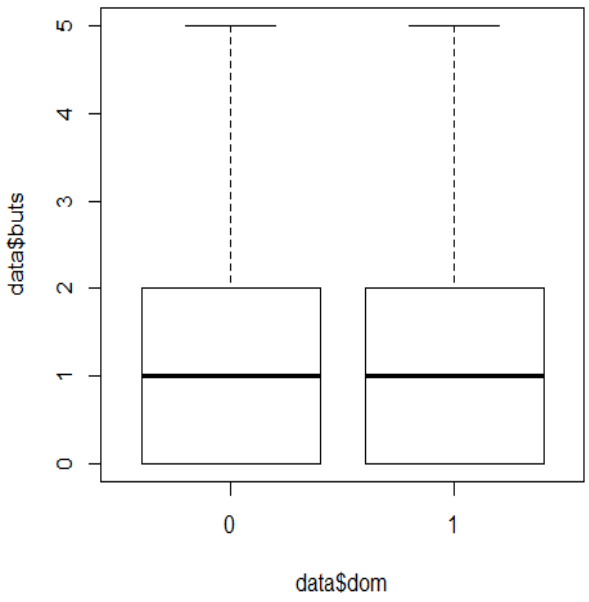
Nombre de buts en fonction de la cote avec la dummy domicile



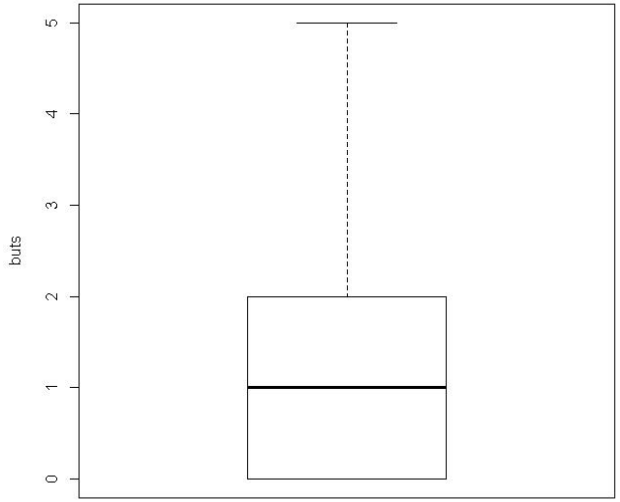
I. Présentation générale de la régression de Poisson

Statistiques descriptives :
Les Box-plots

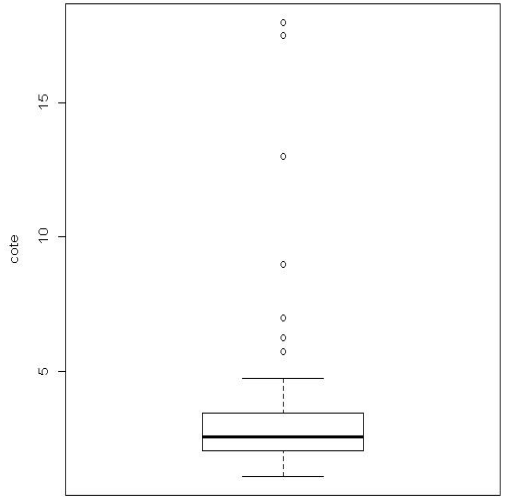
Boxplot du nombre de buts en fonction domicile/extérieur



Boxplot nombre de buts

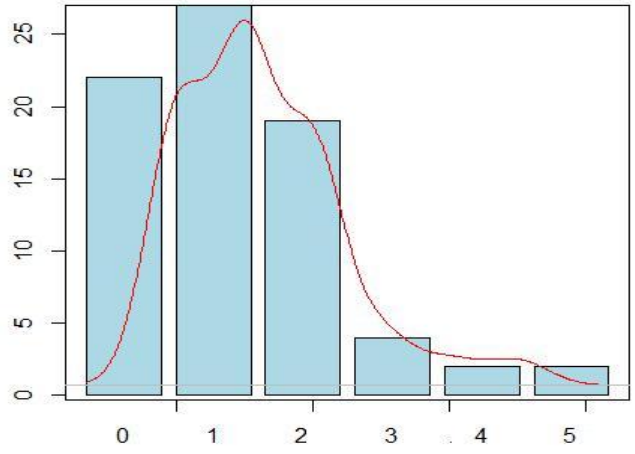


Boxplot cote d'une équipe



I. Présentation générale de la régression de Poisson

1. Présentation théorique
2. Présentation des données utilisées
   a. Présentation des données
   b. Statistiques descriptives
   c. Elaboration du modèle



Fréquence des buts par rapport au nombre de buts

→Courbe ressemblant fortement à celle du lissage d'une loi de Poisson

II. Régression de Poisson - Présentation de l'estimateur

1. Théorie
   a. Estimation du modèle

$$\ln[\mathrm{E}(\mathrm{Y})] = \ln(\lambda) = \alpha + \beta_1 x_1 + \cdots + \beta_j x_j$$

But : estimer $\propto$ et le vecteur β des coefficients β_i

Comment ? Méthode du maximum de vraisemblance

Vraisemblance :

$$L = \prod_{i=1}^{n} P(Y_i = k_i) = \prod_{i=1}^{n} e^{-\lambda_i} \frac{{\lambda_i}^{k_i}}{k_i!}$$

II. Régression de Poisson - Présentation de l'estimateur

1. Théorie
   a. Estimation du modèle

Vraisemblance : $$L = \prod_{i=1}^{n} P(Y_i = k_i) = \prod_{i=1}^{n} e^{-\lambda_i} \frac{{\lambda_i}^{k_i}}{k_i!}$$

avec n le nombre d'observations

$$\lambda_i = e^{\alpha + \beta x_i}$$

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ij} \end{pmatrix} \quad \text{et} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_j \end{pmatrix}$$

II. Régression de Poisson - Présentation de l'estimateur

1. Théorie
   a. Estimation du modèle

Logarithme de la Vraisemblance :

$$\ln(L) = \sum_{i=1}^{n} [\, k_i \ln(k_i) - \lambda_i] - constante$$

Maximisation grâce à la dérivée :

$$s(\alpha, \beta) = \begin{pmatrix} \frac{\partial \ln(L)}{\partial \alpha} \\ \frac{\partial \ln(L)}{\partial \beta} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} (y_i - \lambda_i) \\ \sum_{i=1}^{n} x_i (y_i - \lambda_i) \end{pmatrix}$$

$$= \sum_{i=1}^{n} (y_i - \lambda_i) \begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

II. Régression de Poisson - Présentation de l'estimateur

1. Théorie
   a. Estimation du modèle

Méthode de Newton-Raphson : Algorithme

$$\begin{pmatrix} \propto_{k+1} \\ \beta_{k+1} \end{pmatrix} = \begin{pmatrix} \propto_{k} \\ \beta_{k} \end{pmatrix} + I^{-1}(\propto_k, \beta_k) s(\propto_k, \beta_k)$$

où I^{-1} est la matrice de variance-covariance.

STOP quand on a k tel que $\begin{pmatrix} \propto_{k+1} \\ \beta_{k+1} \end{pmatrix} \approx \begin{pmatrix} \propto_{k} \\ \beta_{k} \end{pmatrix}$

II. Régression de Poisson - Présentation de l'estimateur

1. Théorie
   a. Estimation du modèle
   b. La déviance

Objectif : Une mesure de la qualité d'ajustement du modèle

Principe : comparer la varisemblance obtenue à celle du modèle de référence

Définition : $Déviance = 2[L_s - L_c]$

L_s valeur max de la log-vraisemblance modèle complet

L_c valeur max de la log-vraisemblance modèle estimé

II. Régression de Poisson - Présentation de l'estimateur

1. Théorie
   a. Estimation du modèle
   b. La déviance

Ainsi $$L_s = \sum_{i=1}^{n} [\, y_i \ln(y_i) - y_i] - constante$$

avec $\lambda_i = y_i$

et $$L_c = \sum_{i=1}^{n} [\, y_i \ln(\widehat{\lambda_i}) - \widehat{\lambda_i}] - constante$$

Avec $\widehat{\lambda_i} = \exp(\hat{\alpha} + \hat{\beta} x_i)$

Et après calcul : $$Déviance = 2 \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\widehat{\lambda_i}}\right) - (y_i - \widehat{\lambda_i})$$

II. Régression de Poisson - Présentation de l'estimateur

1. Théorie
2. Estimation de notre modèle

But : voir l'impact de la dummy domicile et de la côte

Modèle 1 : la seule variable exogène → côte

Modèle 2 : variables exogènes → côte et dummy domicile

Estimation du modèle Loglinéaire avec R :

→ « *glm(buts~cote, family=« poisson »,data=data)* »

→ « *glm(buts~dom+cote, family=« poisson »,data=data)* »

III. Interprétation des résultats d'une régression de Poisson

1. Application à nos données
   a. Modèle 1 : $Log(Y) = \propto + \beta_1 * X$

	Valeur estimée	Ecart-type	Statistique
Intercept	**0,36902**	0,16864	2,188
cote	**-0,04502**	0,04328	-1,04

$$Log(Nombre\ de\ buts) = 0{,}36902 - 0{,}04502 * cote$$

$$Nombre\ de\ buts = e^{0{,}36902} * e^{-0{,}04502} cote$$

En sommant avec nos données on a $Nombre\ de\ buts = 1.38$

III. Interprétation des résultats d'une régression de Poisson

1. Application à nos données
   a. Modèle 1 : $Log(Y) = \propto + \beta_1 * X$
   b. Modèle 2 : $Log(Y) = \propto + \beta_1 * X + \beta_2 * D$

	Valeur estimée	Ecart-type	Statistique
Intercept	**0,23971**	0,23702	1,011
dom1	**0,17212**	0,21939	0,785
cote	**-0,0332**	0,04485	-0,74

$$Log(Nombre\ de\ buts) = 0{,}23971 + 0.17212 * \mathrm{dom1} - 0{,}0332 * cote$$

$$Nombre\ de\ buts = e^{0{,}23971} * e^{0.17212} dom1 * e^{-0{,}0332} cote$$

Avec nos données :

$$Nombre\ de\ buts\ pour\ une\ équipe\ à\ domicile = 1.46$$
$$Nombre\ de\ buts\ pour\ une\ équipe\ à\ l'extérieur = 1.23$$

III. Interprétation des résultats d'une régression de Poisson

1. Application à nos données
2. Validation du modèle
   a. Analyse de la déviance avec une ANOVA

Analyse de la déviance				
Modèle 1 : buts~cote				
Modèle 2 : buts~dom+cote				
Resid	**Df**	**Déviance des résidus**	**Déviance**	
1	74	*89,755*		
2	89,135	*89,135*	**0,62037**	0,4309

→ Pas de différences entre les 2 modèles

→ On conserve le modèle sans la dummy

→ Déviances grandes → expliquées par données et variables

III. Interprétation des résultats d'une régression de Poisson

1. Application à nos données
2. Validation du modèle
   a. Analyse de la déviance avec une ANOVA
   b. Ajustement du modèle aux données

   → Statistique de Pearson

Test : $\begin{cases} H_0 : \textit{Le modèle décrit bien le lien entre les variables} \\ H_1 : \textit{Le modèle ne décrit pas bien le lien entre les variables} \end{cases}$

Statistique de Pearson : $$X^2 = \sum_{i=1}^{I} \frac{(n_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$$

III. Interprétation des résultats d'une régression de Poisson

1. Application à nos données
2. Validation du modèle
3. Limites du modèle : Avantages / Inconvénients
   a. Limites du modèle

→ Mauvaise qualité d'ajustement à nos données

→ Jeu de données pas assez grand

III. Interprétation des résultats d'une régression de Poisson

1. Application à nos données
2. Validation du modèle
3. Limites du modèle : Avantages / Inconvénients
   a. Limites du modèle
   b. Avantages

→ Spécialement adapté pour les variables discrètes

→ Assez simple d'utilisation

III. Interprétation des résultats d'une régression de Poisson

1. Application à nos données
2. Validation du modèle
3. Limites du modèle : Avantages / Inconvénients
   a. Limites du modèle
   b. Avantages
   c. Inconvénients

→ Problème de dispersion quasi-systématique
L'hypothèse forte de loi de Poisson $E(Y) = V(Y) = \lambda$ est assez difficile à vérifier dans la réalité

III. Interprétation des résultats d'une régression de Poisson

1. Application à nos données
2. Validation du modèle
3. Limites du modèle : Avantages / Inconvénients
   a. Limites du modèle
   b. Avantages
   c. Inconvénients

→ Nécessité d'avoir une variable dépendante avec une distribution poissonienne

(pas évident dans la pratique !)

Conclusion

Endogène = Variable de comptage

But de la régression : expliquer l'endogène par des explicatives (modèle log-linéaire !)

Notre exemple → nombre de buts moyen en fonction d'être à domicile ou non

Cependant qualité d'ajustement faible

(à cause de l'inconvénient principal de cette méthode et de la taille du jeu de données)