Projet modèles linéaires avancés:

Modèles linéaires mixtes généralisés

AMAHJOUR Walid

Novembre 2020

Table des matières

1. Introduction

2. Moindres carrés ordinaires

3. Modèle linéaire généralisé

4. Modèles linéaires mixtes généralisés

1. Introduction

Les modèles mixtes linéaires on une importante limitation, ils ne peuvent pas accepter des variables de réponse qui n'ont pas une distribution d'erreur normale. La plupart des données biologiques ne suivent pas l'hypothèse de normalité.
Nous allons voir comment on peut appliquer les modèles linéaires généralisés, qui sont des outils importants pour surmonter les hypothèses de distribution des modèles linéaires. Nous allons voire les principaaux distributions utilisées en fonction de la nature des variables de réponse, du concept de la fonction de lien et comment vérifier les hypothèses de ces modèles.
Nous allons utiliser plusieurs jeu de données biologique tel que mites, CO2, faramea.

2. Moindres carrés ordinaires

L'ensemble de données comprend 70 échantillons de mousses et d'acariens recueillis à la Station de biologie des Laurentides de l'Université de Montréal, Saint-Hippolyte, QC. Chaque échantillon comprend des mesures pour 5 variables environnementales et des données d'abondance pour 35 taxons d'acariens. Dans l'ensemble de données réduit que nous utiliserons tout au long de cet atelier, nous n'avons inclus que les 5 mesures environnementales et l'abondance d'un seul taxon d'acariens, «Galumna sp.» Notre objectif sera de modéliser l'abondance, l'occurrence (présence / absence) et la proportion de Galumna en fonction des 5 variables environnementales: par conséquent, nous avons également créé une variable présence / absence et une variable de proportion pour Galumna.

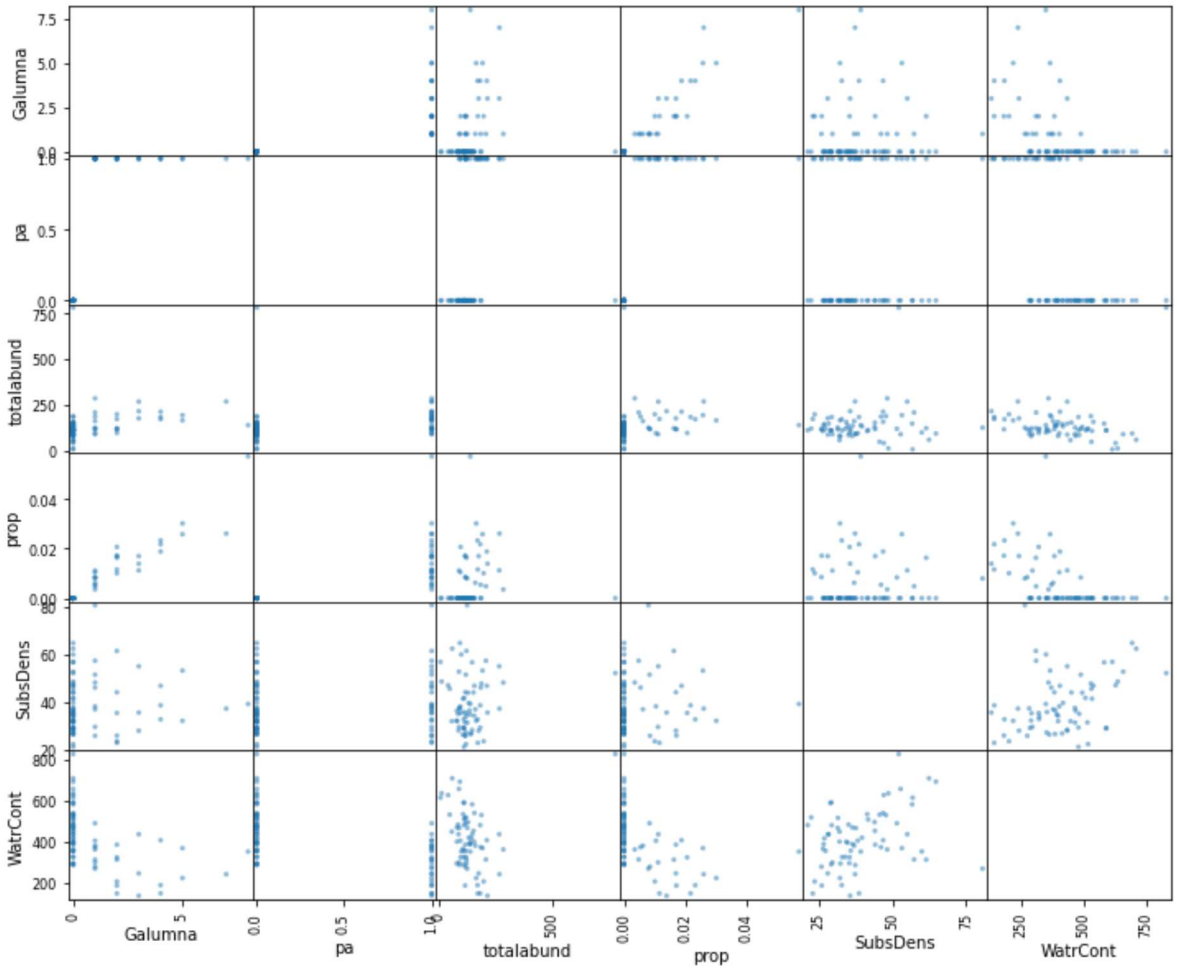
Out[3]:

	Galumna	pa	totalabund	pro
0	8	1	140	0.05714
1	3	1	268	0.01119
2	1	1	186	0.00537
3	1	1	286	0.00349
4	2	1	199	0.01005
...	...	...	...	
65	0	0	116	0.00000
66	0	0	781	0.00000
67	0	0	111	0.00000
68	0	0	184	0.00000
69	0	0	121	0.00000

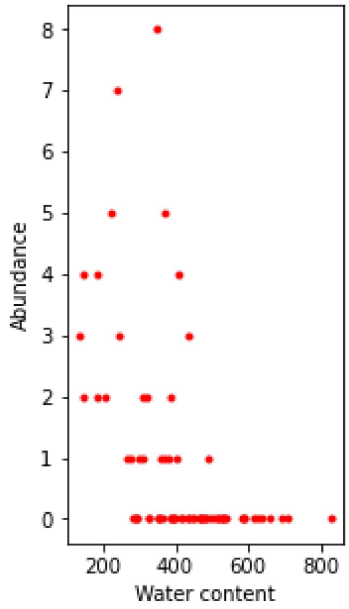
70 rows × 9 columns

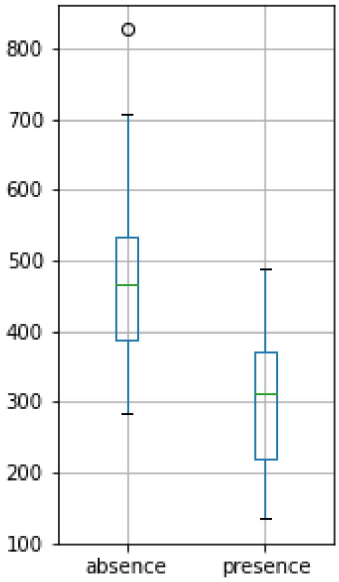
```
<class 'pandas.core.frame.Data
Frame'>
RangeIndex: 70 entries, 0 to 6
9
Data columns (total 9 column
s):
 #   Column      Non-Null Coun
t  Dtype
---  ------      -------------
-  -----
 0   Galumna     70 non-null
int64
 1   pa          70 non-null
int64
 2   totalabund  70 non-null
int64
 3   prop        70 non-null
float64
 4   SubsDens    70 non-null
float64
 5   WatrCont    70 non-null
float64
 6   Substrate   70 non-null
object
 7   Shrub       70 non-null
object
 8   Topo        70 non-null
object
dtypes: float64(3), int64(3),
object(3)
memory usage: 5.0+ KB
```

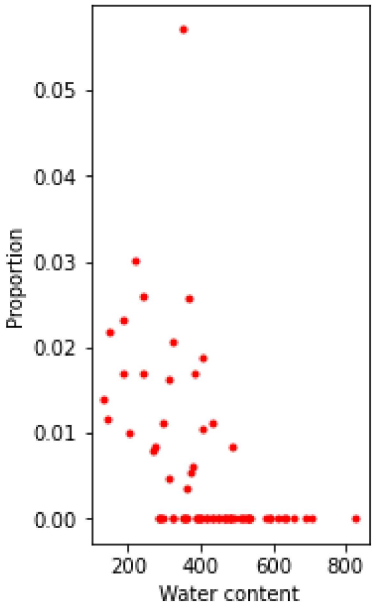
Pour voirsi il'y a des relations entre Galumna et les cinq variables environnementales, nous allons construire le plot suivant.



Nous constatons qu'il y aune relation negative entre la variable WatrCont et Galumna.Pour s'assuer nous allons construires les graphes suivants







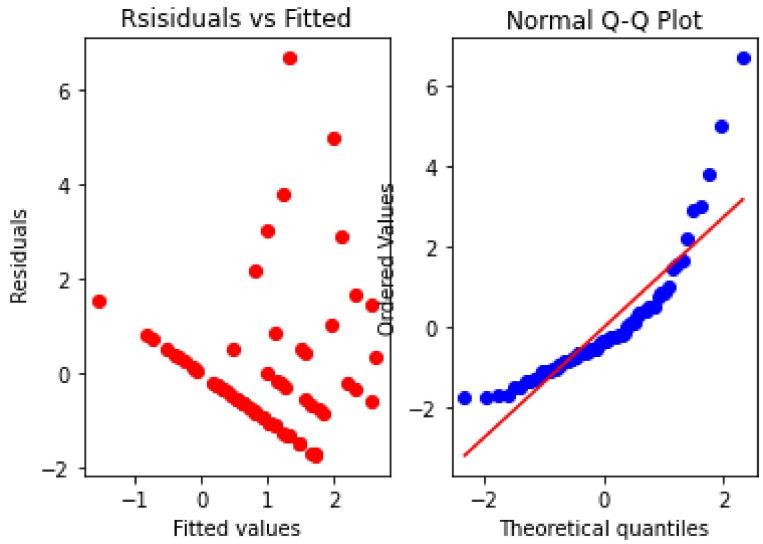
Nous allons faire des modèles linéaires afin de verifier les relation entre Galumna et WatrCont, Galumna et pa, Galumna et prop.

Interpretation des résultats des regressions: Mème si on a obtenue des résultats de regressions singnificatives on peut pas s'arreter ici, on doit donc verifier l'hypothèse de la normalité des risidus.

On s'interesse à la relation entre Galumna et WatrCont.Pour voir mieux la qualité de la régression on utulise souvent ces graphes.

Out[47]:

```
3.439348671737164
```



Interpretation des graphes: il est claire que l'hypothèse de la normalité des risidus est moins probable, comme le montres les deux histogrammes suivants.

Out[10]:

OLS Regression Results

Dep. Variable:	Galumna	**R-squared:**	
Model:	OLS	**Adj. R-squared:**	
Method:	Least Squares	**F-statistic:**	
Date:	Sun, 08 Nov 2020	**Prob (F-statistic):**	
Time:	13:05:03	**Log-Likelihood:**	
No. Observations:	70	**AIC:**	
Df Residuals:	68	**BIC:**	
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|
Intercept	3.4393	0.556	6.188	0.00
WatrCont	-0.0060	0.001	-4.723	0.00

Omnibus:	48.057	**Durbin-Watson:**	
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1
Skew:	2.184	**Prob(JB):**	8.
Kurtosis:	8.894	**Cond. No.**	1.3

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.33e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

sigma

modèle linéaire simple

$$y = \beta_0 + \beta_1 x_i + \varepsilon$$

y_i = valeur prédite d'une variable de réponse

β_0 = intercèpte

β_1 = pente

x_i = variables explicative

ε_i = modèle risiduel d'une distribution normale

Out[11]:

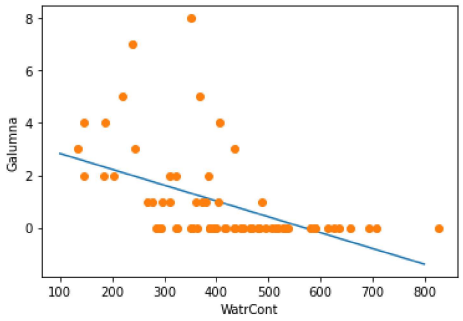
OLS Regression Results

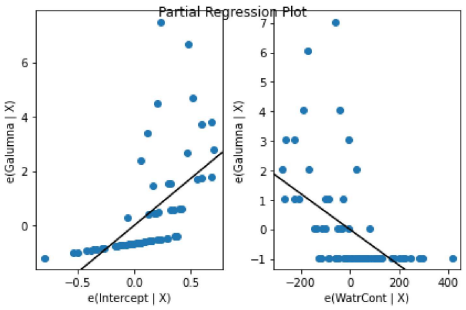
Dep. Variable:	pa	**R-squared:**	0.357
Model:	OLS	**Adj. R-squared:**	0.348
Method:	Least Squares	**F-statistic:**	37.80
Date:	Sun, 08 Nov 2020	**Prob (F-statistic):**	4.68e-08
Time:	13:05:03	**Log-Likelihood:**	-32.354
No. Observations:	70	**AIC:**	68.71
Df Residuals:	68	**BIC:**	73.20

Out[12]:

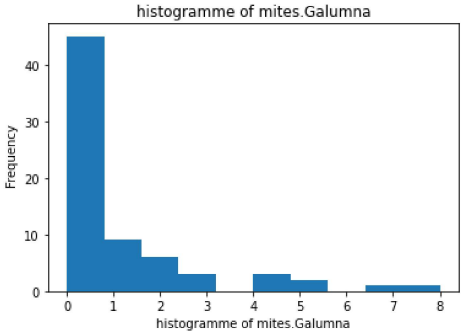
OLS Regression Results

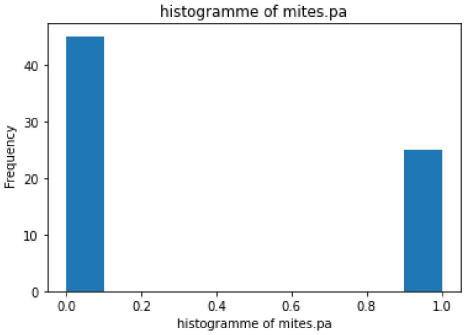
Dep. Variable:	prop	**R-squared:**	0.240
Model:	OLS	**Adj. R-squared:**	0.229
Method:	Least Squares	**F-statistic:**	21.49
Date:	Sun, 08 Nov 2020	**Prob (F-statistic):**	1.67e-05
Time:	13:05:03	**Log-Likelihood:**	231.65
No. Observations:	70	**AIC:**	-459.3
Df Residuals:	68	**BIC:**	-454.8






```
sigma= 1.51353094952848
```





$y_i \sim N(\mu = \beta_0 + \beta_1 X_i, \sigma^2)$ donc cette hypothese n'est pas vraie. Nous avons besoin d'une distribution avec une plage qui n'inclut que deux résultats possibles: zéro ou un. La distribution «Bernoulli» est une telle distribution.

3. Modèle linéaire généralisé

Pour contourner le probleme de la normalité des y_i on peut supposer que les ε_i suivent une loie de Poisson.
$y_i \sim Poisson(\lambda = \beta_0 + \beta_1 x_i)$

Avantages:

- Les valeurs prédites seront désormais des entiers au lieu de fractions
- Le modèle ne prédira jamais de valeurs négatives (Poisson est strictement positif)
- λ varie avec x (teneur en eau), ce qui signifie que la variance résiduelle variera également avec x . Cela signifie également que nous avons assoupli l'hypothèse d'homogénéité de la variance

Out[91]:

Generalized Linear Model Regression Result

Dep. Variable:	pa	**No. Observations:**	
Model:	GLM	**Df Residuals:**	
Model Family:	Poisson	**Df Model:**	
Link Function:	log	**Scale:**	
Method:	IRLS	**Log-Likelihood:**	
Date:	Sun, 08 Nov 2020	**Deviance:**	
Time:	20:11:28	**Pearson chi2:**	
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z
const	0.3596	0.766	0.469
mites.WatrCont	-0.0054	0.002	-2.937
Topo	0.8957	0.476	1.880

variables binaires

Une variable de réponse commune dans les ensembles de données écologiques est la variable binaire: nous observons un phénomène x ou son «absence».

- présence ou absence d'une éspece
- présence ou absence d'une maladie
- succés ou échec d'une experimentation

regression logit

$$g(p) = log\frac{p}{1-p}$$

Out[93]:

Generalized Linear Model Regression Result

Dep. Variable:	pa	**No. Observations:**	
Model:	GLM	**Df Residuals:**	
Model Family:	Binomial	**Df Model:**	
Link Function:	logit	**Scale:**	
Method:	IRLS	**Log-Likelihood:**	
Date:	Sun, 08 Nov 2020	**Deviance:**	
Time:	20:17:27	**Pearson chi2:**	
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z
const	4.4644	1.671	2.672
mites.WatrCont	-0.0158	0.005	-3.487
Topo	2.0908	0.735	2.843

```
exp(log(μ / (1 - μ)) = u / (1
- μ)
[0.9843118083493371, 8.0910340
00774105]
intervale de confiance
                    2.5%
97.5%
const            2.054154  367
3.645516
mites.WatrCont  0.974357
0.994368
Topo             1.556664    4
2.054571
```

```
<ipython-input-23-525fe09a8438
>:2: DeprecationWarning: Calli
ng Family(..) with a link clas
s as argument is deprecated.
Use an instance of a link clas
s instead.
  logit_reg =sm.GLM(mites.pa,s
m.add_constant(df.astype(floa
t)), data = mites, family =sm.
families.Binomial(link=sm.fami
lies.links.logit)).fit()
```

Out[23]:

Generalized Linear Model Regression Result

Dep. Variable:	pa	**No. Observations:**	
Model:	GLM	**Df Residuals:**	
Model Family:	Binomial	**Df Model:**	
Link Function:	logit	**Scale:**	
Method:	IRLS	**Log-Likelihood:**	
Date:	Sun, 08 Nov 2020	**Deviance:**	
Time:	13:05:05	**Pearson chi2:**	
No. Iterations:	6		
Covariance Type:	nonrobust		

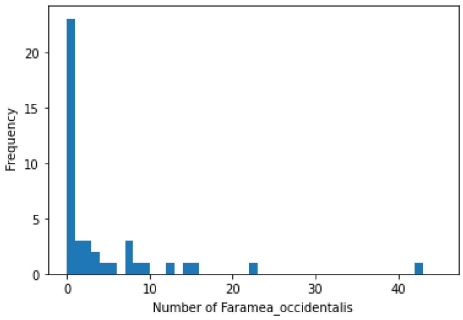
	coef	std err	z
const	4.4644	1.671	2.672
mites.WatrCont	-0.0158	0.005	-3.487
Topo	2.0908	0.735	2.843

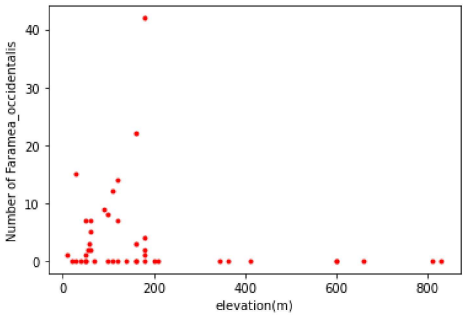
GLMs with count data

Out[26]:

	Unnamed: 0	UTM.EW	UTM.NS	Precipitation	Elevation	Age	Age.cat	Geology	Faramea_oc
0	B0	625754.0	1011569.0	2530.0	120	3	c3	Tb	
1	B49	626654.0	1011969.0	2530.0	120	3	c3	Tb	
2	p1	614856.9	1031786.4	2993.2	20	2	c2	Tc	
3	p2	613985.4	1030725.4	3072.0	100	3	c3	Tc	
4	p3	614674.3	1023801.5	3007.4	180	1	c1	Tc	
5	p4	615018.6	1023547.9	2999.8	180	1	c1	Tc	

plots





Poisson GLM

Negative binomial GLMs

Out[58]:

Generalized Linear Model Regression Result

Dep. Variable:	Faramea_occidentalis	**Obse**	
Model:	GLM	**Df R**	
Model Family:	Poisson		
Link Function:	log		
Method:	IRLS	**Li**	
Date:	Sun, 08 Nov 2020	**I**	
Time:	19:45:35	**Pear**	
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z	P>|
const	1.7687	0.110	16.092	0.00
Elevation	-0.0027	0.001	-4.253	0.00

Out[30]:

Generalized Linear Model Regression Results

Dep. Variable:	Faramea_occidentalis	**No. Observations:**	43
Model:	GLM	**Df Residuals:**	41
Model Family:	NegativeBinomial	**Df Model:**	1
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-100.37
Date:	Sun, 08 Nov 2020	**Deviance:**	91.089
Time:	13:05:08	**Pearson chi2:**	154.
No. Iterations:	12		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.1108	0.259	8.144	0.000	1.603	2.619
Elevation	-0.0052	0.001	-3.642	0.000	-0.008	-0.002

Generalized linear mixed models

Out[35]:

	Unnamed: 0	reg	popu	gen	rack	nutrient	amd	status	total_fruits
0	1	NL	3.NL	4	2	1	clipped	Transplant	0
1	2	NL	3.NL	4	1	1	clipped	Petri.Plate	0
2	3	NL	3.NL	4	1	1	clipped	Normal	0
3	4	NL	3.NL	4	2	1	clipped	Normal	0
4	5	NL	3.NL	4	2	8	clipped	Transplant	0
...	...	...	...	...	...	...	...	...	...
620	828	SW	1.SW	25	2	8	unclipped	Normal	5
621	831	SW	1.SW	25	1	1	clipped	Transplant	3
622	853	SW	1.SW	27	2	8	unclipped	Normal	5
623	854	SW	1.SW	27	2	8	clipped	Transplant	5
624	902	SW	7.SW	35	2	1	clipped	Normal	3

625 rows × 9 columns