

TD 1 — Modèle de Poisson (R)

Année universitaire 2025–2026 ■ Parcours Économie de la santé & Développement durable

Pierre Beaucoral

Contexte: On étudie d'anciennes données reliant **tabagisme** et **décès par cancer du poumon**. Variables : **age** (classes), **smoking status** (4 classes), **population** (centaines de milliers), **deaths** (décès annuels).

Préparation

Objectifs de ce TD

- Importer et préparer un tableau **comptages + exposition** (population à risque).
- Ajuster un **GLM Poisson** avec *offset* (log-exposition).
- Évaluer l'ajustement : **déviance** (vs modèle saturé) & **Pearson**.
- Comparer des modèles via **tests de rapport de vraisemblance (LR)**.
- Interpréter en **ratios de taux d'incidence (IRR)** et produire des **comptes attendus**.

Packages

Import et manipulation des données

Importer les données `smoking_dat.xlsx`

Dans l'énoncé, les données à importer sont `smoking_dat.xlsx` et les variables `age`, `smoking status`, `population`, `deaths`.

Note

Dictionnaire des variables :

- `age`: en classes (40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+).

- smoking status: 4 classes (ne fume pas / fume le cigare ou la pipe / fume la cigarette et le cigare ou la pipe ; fume seulement la cigarette)
- population: en centaine de milliers de personnes
- deaths: comptage des décès par cancer du poumon en un an.

Rows: 36

Columns: 4

```
$ age           <chr> "40-44", "45-59", "50-54", "55-59", "60-64", "65-69", "~
$ smoking_status <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "~
$ population     <dbl> 656, 359, 249, 632, 1067, 897, 668, 361, 274, 145, 104,~
$ deaths         <dbl> 18, 22, 19, 55, 117, 170, 179, 120, 120, 2, 4, 3, 38, 1~
```

Coder les deux variables enregistrées en texte avec des chiffres

En Stata on ferait `encode + i.variable`.

En R, il suffit de déclarer les variables comme `factor`.

```
[1] "40-44" "45-59" "50-54" "55-59" "60-64" "65-69" "70-74" "75-79" "80+"
```

```
[1] "cigarPipeOnly" "cigaretteOnly" "cigarettePlus" "no"
```

Note

Rappel : Les facteurs indiquent à R qu'il s'agit de variables qualitatives. Chaque modalité sera transformée en **variable indicatrice** (dummy) dans la régression.

Unité d'exposition

L'énoncé précise que `population` est en **centaines de milliers**. Pour une interprétation plus intuitive, on peut ramener l'exposition à l'unité **personne** (facultatif) :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9800000	36925000	85850000	155894444	230550000	605200000

Estimations

Appliquer un premier modèle de régression log-linéaire

Modèle Poisson **log-linéaire** avec effets de `smoking_status` et `age`, et *offset* `log(exposure)` : c'est l'équivalent de Stata `poisson deaths i.smokecod i.agecod, exposure(pop)`.

Tip

Pourquoi un modèle de Poisson ?

Les données sont des **comptages** (nombre de décès).

Le modèle de Poisson relie l'**espérance** de ces comptages à des variables explicatives par une fonction de lien log :

$$\log(\mathbb{E}[Y]) = X\beta$$

Cette structure garantit que la prédiction est **positive** et que la variance est proportionnelle à la moyenne (hypothèse de Poisson).

Nous voulons expliquer le nombre de décès par l'âge et le statut tabagique, en tenant compte de l'exposition.

Call:

```
glm(formula = deaths ~ smoking_status + age + offset(log(exposure)),  
     family = poisson(link = "log"), data = df)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-15.14514	0.06783	-223.293	< 2e-16 ***
smoking_statuscigarretteOnly	0.36915	0.03791	9.737	< 2e-16 ***
smoking_statuscigarrettePlus	0.17015	0.03643	4.671	3.00e-06 ***
smoking_statusno	-0.04781	0.04699	-1.017	0.309
age45-59	0.55388	0.07999	6.924	4.38e-12 ***
age50-54	0.98039	0.07682	12.762	< 2e-16 ***
age55-59	1.37946	0.06526	21.138	< 2e-16 ***
age60-64	1.65423	0.06257	26.439	< 2e-16 ***
age65-69	1.99817	0.06279	31.824	< 2e-16 ***
age70-74	2.27141	0.06435	35.296	< 2e-16 ***
age75-79	2.55858	0.06778	37.746	< 2e-16 ***
age80+	2.84692	0.07242	39.310	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 4055.984 on 35 degrees of freedom
Residual deviance: 21.487 on 24 degrees of freedom
AIC: 285.51

Number of Fisher Scoring iterations: 4

! Important

L'argument `offset(log(exposure))` ajoute le **log de l'exposition** avec un coefficient fixé à 1. Cela revient à modéliser un **taux** de décès (décès / population).

Calculer la déviance du modèle ajusté. DEV1

dev1	df	p_value
21.486738	24.000000	0.609872

Quel est l'effet de l'âge sur la probabilité de décès par cancer du poumon ?

Interprétez les coefficients associés aux **modalités d'âge** (comparées à la catégorie de référence) en termes d'IRR (voir section IRR) et/ou d'impact sur le **taux de décès** (à exposition fixée). (Discussion attendue.)

Interprétez les coefficients d'âge ($IRR = \exp(\text{coef})$) :

- $IRR > 1$: taux de décès plus élevé que la catégorie de référence.
- $IRR < 1$: taux plus faible.

Sauvegarde du modèle (utile pour LR tests)

Tester l'ajustement de ce modèle

Pour cela nous allons réaliser **deux tests** :

- **Deviance GOF** (modèle ajusté vs saturé).

test	statistic	df	p_value
Deviance GOF	21.48674	24	0.6098720
Pearson GOF	20.61936	24	0.6610658

- **Pearson GOF** (comparaison effectifs attendus vs observés). Les **degrés de liberté** correspondent ici au nombre de cellules moins le nombre de paramètres estimés (y compris l'interception). L'énoncé suggère **24 df** pour chacun de ces tests (voir justification plus bas).

Justifiez pourquoi ces tests sont effectués avec le même df.

Tip

Rappel : pour un GLM Poisson, $df = N - p$, où N est le nombre de cellules et p le nombre de paramètres estimés (y compris l'interception). Discuter les **conditions d'un**² (souvent $\hat{\mu} \geq 5$ dans la plupart des cellules).

Ajuster un modèle sans la variable smoke, et effectuer un test de rapport de vraisemblance entre ce nouveau modèle et celui précédemment sauvegardé

On ajuste un modèle **sans tabac** et on compare à mod1 par **LR test**. En Stata : `lrtest`. En R : `anova(mod0, mod1, test="Chisq")`.

Analysis of Deviance Table

```
Model 1: deaths ~ age + offset(log(exposure))
Model 2: deaths ~ smoking_status + age + offset(log(exposure))
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      27    191.723
2      24     21.487  3    170.24 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclure sur l'impact de l'usage du tabac sur la probabilité de décès par cancer du poumon.

Construire une nouvelle variable qui prend la valeur 1 si l'individu fume des cigarettes, 0 s'il n'en fume pas.

Créer une variable `cigarette_user` égale à 1 si l'individu fume des cigarettes, 0 sinon : l'énoncé demande de distinguer le type de produit et de concentrer l'attention sur la cigarette.

	cigarPipeOnly	cigaretteOnly	cigarettePlus	no
0	9	0	0	9
1	0	9	9	0

Ajuster un troisième modèle avec effet de l'âge et de cette variable d'usage de la cigarette.

Call:

```
glm(formula = deaths ~ cigarette_user + age + offset(log(exposure)),  
     family = poisson, data = df)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-15.15590	0.06378	-237.625	< 2e-16 ***
cigarette_user	0.26910	0.02757	9.762	< 2e-16 ***
age45-59	0.55342	0.07999	6.919	4.56e-12 ***
age50-54	0.98480	0.07682	12.820	< 2e-16 ***
age55-59	1.37640	0.06526	21.092	< 2e-16 ***
age60-64	1.64629	0.06256	26.317	< 2e-16 ***
age65-69	1.99023	0.06277	31.708	< 2e-16 ***
age70-74	2.26143	0.06432	35.161	< 2e-16 ***
age75-79	2.54560	0.06766	37.626	< 2e-16 ***
age80+	2.82907	0.07215	39.211	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 4055.984 on 35 degrees of freedom

Residual deviance: 92.237 on 26 degrees of freedom
AIC: 352.26

Number of Fisher Scoring iterations: 4

Analysis of Deviance Table

Model: poisson, link: log

Response: deaths

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			35	4056.0	
cigarette_user	1	58.3	34	3997.6	2.195e-14 ***
age	8	3905.4	26	92.2	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Comparer ce modèle avec le modèle initial par un test de rapport de vraisemblance

On compare le **modèle multiniveaux par type de tabac** (mod1) avec le **modèle binaire cigarette vs non** (mod2).

Analysis of Deviance Table

Model 1: deaths ~ cigarette_user + age + offset(log(exposure))
Model 2: deaths ~ smoking_status + age + offset(log(exposure))

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	26	92.237			
2	24	21.487	2	70.75	4.334e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

IRR (modèle 1) — exp(coef) avec IC 95%

term	estimate	conf.low	conf.high	p.value
age80+	1.723470e+01	1.496968e+01	1.988626e+01	0.000000e+00
age75-79	1.291740e+01	1.132626e+01	1.477500e+01	0.000000e+00
age70-74	9.693017e+00	8.558760e+00	1.101560e+01	6.852385e-273
age65-69	7.375556e+00	6.533324e+00	8.357251e+00	2.978913e-222
age60-64	5.229045e+00	4.633991e+00	5.922596e+00	4.943059e-154
age55-59	3.972749e+00	3.501347e+00	4.522490e+00	3.581870e-99
age50-54	2.665487e+00	2.294116e+00	3.100656e+00	2.658489e-37
age45-59	1.739985e+00	1.487796e+00	2.036040e+00	4.376858e-12
smoking_statuscigaretteOnly	1.446509e+00	1.343418e+00	1.558691e+00	2.099342e-22
smoking_statuscigarettePlus	1.185481e+00	1.104252e+00	1.273769e+00	3.001586e-06
smoking_statusno	9.533182e-01	8.692799e-01	1.045138e+00	3.090007e-01
(Intercept)	2.645745e-07	2.312255e-07	3.016715e-07	0.000000e+00

Le type de produit fumé semble-t-il influencer la probabilité de décès par cancer du poumon ?

En Poisson log-linéaire : $IRR = \exp(\text{coef})$. On reporte aussi des **IC 95%** exponentiés.

Extensions

Comptes attendus & tableau Observé vs Attendu

Vérification d'éventuelle sur-dispersion

Le GLM Poisson suppose $\text{Var}(Y) = E[Y]$. Si $\text{Var}(Y) \gg E[Y]$, la **sur-dispersion** peut invalider les tests usuels (SE sous-estimés).

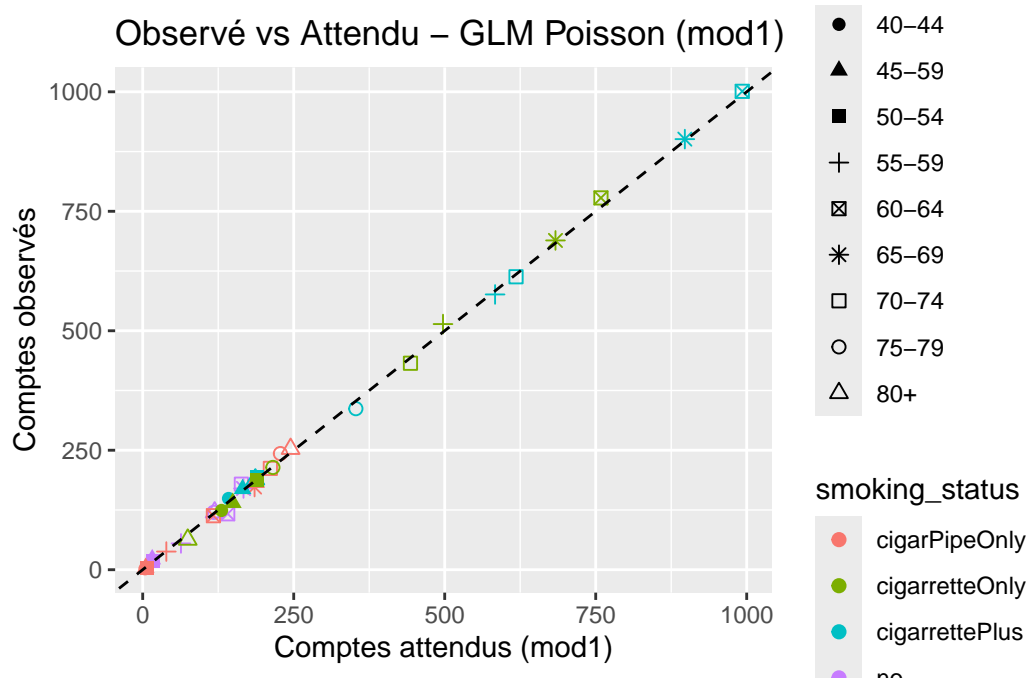
```
# Overdispersion test
```

```
dispersion ratio = 0.859
Pearson's Chi-Squared = 20.619
p-value = 0.661
```

En cas de sur-dispersion marquée, envisagez **Quasi-Poisson** (`family = quasipoisson`) ou **Négative Binomiale** (`MASS::glm.nb`) et comparez l'ajustement.

age	smoking_status	exposure	y_obs	mu_hat
40-44	cigarPipeOnly	14,500,000.00	2.00	3.84
40-44	cigarretteOnly	341,000,000.00	124.00	130.50
40-44	cigarrettePlus	453,100,000.00	149.00	142.11
40-44	no	65,600,000.00	18.00	16.55
45-59	cigarPipeOnly	10,400,000.00	4.00	4.79
45-59	cigarretteOnly	223,900,000.00	140.00	149.10
45-59	cigarrettePlus	303,000,000.00	169.00	165.36
45-59	no	35,900,000.00	22.00	15.76
50-54	cigarPipeOnly	9,800,000.00	3.00	6.91
50-54	cigarretteOnly	185,100,000.00	187.00	188.82
50-54	cigarrettePlus	226,700,000.00	193.00	189.53
50-54	no	24,900,000.00	19.00	16.74
55-59	cigarPipeOnly	37,200,000.00	38.00	39.10
55-59	cigarretteOnly	327,000,000.00	514.00	497.17
55-59	cigarrettePlus	468,200,000.00	576.00	583.40
55-59	no	63,200,000.00	55.00	63.33
60-64	cigarPipeOnly	84,600,000.00	113.00	117.04
60-64	cigarretteOnly	379,100,000.00	778.00	758.66
60-64	cigarrettePlus	605,200,000.00	1,001.00	992.58
60-64	no	106,700,000.00	117.00	140.73
65-69	cigarPipeOnly	94,900,000.00	173.00	185.19
65-69	cigarretteOnly	242,100,000.00	689.00	683.37
65-69	cigarrettePlus	388,000,000.00	901.00	897.57
65-69	no	89,700,000.00	170.00	166.87
70-74	cigarPipeOnly	82,400,000.00	212.00	211.32
70-74	cigarretteOnly	119,500,000.00	432.00	443.30
70-74	cigarrettePlus	203,300,000.00	613.00	618.07
70-74	no	66,800,000.00	179.00	163.31
75-79	cigarPipeOnly	66,700,000.00	243.00	227.95
75-79	cigarretteOnly	43,600,000.00	214.00	215.54
75-79	cigarrettePlus	87,100,000.00	337.00	352.89
75-79	no	36,100,000.00	120.00	117.62
80+	cigarPipeOnly	53,700,000.00	253.00	244.86
80+	cigarretteOnly	11,300,000.00	63.00	74.53
80+	cigarrettePlus	34,500,000.00	189.00	186.49
80+	no	27,400,000.00	120.00	119.11

Graphiques (facultatifs)



Bilan

- Le modèle de Poisson permet d'estimer des **taux de décès** en fonction du tabagisme et de l'âge.
- Les **tests d'ajustement** (déviante, Pearson) valident le modèle si p-value élevée.
- Le **tabagisme** a un impact significatif sur la mortalité par cancer du poumon.
- Les **IRR** offrent une interprétation intuitive : *par rapport à la catégorie de référence, combien de fois le taux de décès est-il multiplié.*

Conseil pratique : en recherche appliquée, vérifiez toujours la sur-dispersion et documentez les hypothèses de variance (Poisson vs quasi-Poisson vs binomiale négative).