Un exemple de régression linéaire

Exercice 1 (*Calculs pour une régression linéaire*)

Soient $p \in]0;1[$, et $X,Y \hookrightarrow \mathcal{G}(p)$ indépendantes.
On s'interesse au couples de variables aléatoires :
- $S = \max(X,Y)$
- $I = \min(X,Y)$

1. Justifier les relations *a priori* :
- $I+S = X+Y$.
- $I \cdot S = X \cdot Y$.
- $I^2+S^2 = X^2+Y^2$.

2. Rappels sur la loi géométrique On a choisi $X \hookrightarrow \mathcal{G}(p)$.

a) Rappeler $X(\Omega)$, et pour $n \in X(\Omega)$, la probabilité $\mathbb{P}(X=n)$.

Réponse : $X(\Omega) = \mathbb{N}^*$, et pour $n \geqslant 1$, $\mathbb{P}(X=n) = p \cdot q^{n-1}$.)

b) Rappeler l'expression de $\mathbb{E}[X]$ et de $\mathrm{Var}(X)$.
Par la formule de Kœnig-Huygens, en déduire : $\mathbb{E}\left[X^2\right] = \frac{1+q}{p^2}$.

Réponse : $\mathbb{E}[X] = \frac{1}{p}$, $\mathrm{Var}(X) = \frac{q}{p^2}$, $\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2$.)

c) Rappeler, pour $n \in \mathbb{N}$, la probabilité : $\mathbb{P}(X > n)$. *(fonction d'**antirépartition** de X.)*

Réponse : $\mathbb{P}(X < n+1) = \sum_{k=1}^{n} p \cdot q^{k-1}$, $\mathbb{P}(X \geqslant n) = q^{n-1}$.)

3. Lois marginales de I,S

a) Montrer l'égalité d'événements :
- $[S \leqslant k] = [X \leqslant k] \cap [Y \leqslant k]$,
- $[I > k] = [X > k] \cap [Y > k]$.

b) En déduire :
- $\mathbb{P}(S \leqslant n) = 1 - 2q^n + \left(q^2\right)^n$,
- $\mathbb{P}(I > n) = \left(q^2\right)^n$.

c) Déduire enfin :
- $\mathbb{P}(S=n) = 2 \cdot p \cdot q^{n-1} - (1-q^2) \cdot \left(q^2\right)^{n-1}$, pour $n \geqslant 1$,
- $I \hookrightarrow \mathcal{G}\left(1-q^2\right)$.

4. Calcul de la covariance

a) Justifier que l'on a : $\mathbb{E}[I] = \frac{1}{1-q^2}$.

b) Montrer : $\mathbb{E}[S] = \mathbb{E}[X] + \mathbb{E}[Y] - \mathbb{E}[I]$. *(On utilisera l'une des relations de **1.**.)*
En déduire : $\mathbb{E}[S] = \frac{2p-1}{p^2}$.

c) Montrer : $\mathbb{E}[I \cdot S] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$. *(On utilisera **1.**, et l'indépendance de X,Y.)*
En déduire : $\mathbb{E}[I \cdot S] = \frac{1}{p^2}$.

d) Obtenir enfin : $\mathrm{Cov}(I,S) = \frac{1}{p^2} - \frac{2p-1}{p^3 \cdot (2-p)} = \frac{1}{p^3 \cdot (2-p)}\left[p \cdot (2-p) - (2p-1)\right] = \frac{1-p^2}{p^3 \cdot (2-p)}$.

Exercice 2 (*Calculs pour la régression linéaire*)

Soient X et Y deux variables aléatoires admettant un moment d'ordre 2.
Toutes les espérances, variances, covariances apparaissant convergent donc bien.
On fait la régression linéaire de Y par X, en approximant Y par $\widehat{Y} = a\cdot X + b$.
On cherche pour quelles valeurs de $a,b \in \mathbb{R}$, l'erreur quadratique $r_{a,b} = \mathbb{E}\big[(Y-aX-b)^2\big]$.

1. **Détermination des coefficients a,b optimaux**
   - **a)** Par la formule de Kœnig-Huygens, montrer : $r_{a,b} = \mathrm{Var}(Y-aX) + \big(\mathbb{E}[Y]-a\cdot\mathbb{E}[X]-b\big)^2$.
   - **b)** Montrer que : $\mathrm{Var}(Y-aX) = \left[a\cdot\sigma_X - \frac{\mathrm{Cov}(X,Y)}{\sigma_X}\right]^2 + \mathrm{Var}(Y) - \frac{1}{\mathrm{Var}(X)}\cdot\big[\mathrm{Cov}(X,Y)\big]^2$
     (On écrira : $\mathrm{Var}(Y-aX) = \mathrm{Var}(Y) - 2a\cdot\mathrm{Cov}(X,Y) + a^2\cdot\sigma_X^2$*)*
   - **c)** En déduire que $r_{a,b}$ est minimisée pour :
     - $a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$,
     - $b = \mathbb{E}[Y] - a\cdot\mathbb{E}[X]$.
2. On choisit ces valeurs de a,b. Le résidu est noté : $\epsilon = Y - a\cdot X - b$.
   - **a)** Montrer que $\mathbb{E}[\epsilon] = 0$. En déduire : $\mathbb{E}[a\cdot X + b]$.
   - **b)** Montrer que : $\mathrm{Var}(\epsilon) = \mathrm{Var}(Y) - \frac{1}{\mathrm{Var}(X)}\cdot\big[\mathrm{Cov}(X,Y)\big]^2$.
   - **c)** Montrer que : $\mathrm{Cov}(X,\epsilon) = 0$.
     En déduire que : $\mathrm{Var}(a\cdot X+b) + \mathrm{Var}(\epsilon) = \mathrm{Var}(Y)$
   - **d)** Montrer :
     - $\mathrm{Var}(a\cdot X+b) = \rho^2(X,Y)\cdot\mathrm{Var}(Y)$,
     - $\mathrm{Var}(\epsilon) = \big[1-\rho^2(X,Y)\big]\cdot\mathrm{Var}(Y)$.

Détermination pratique en Scilab

La commande `reglin` permet de faire la régression linéaire d'un échantillon `X` par `Y`.

- **Syntaxe**

```
// x,y sont deux échantillons de même longueur
[a,b] = reglin(x,y) // la régression linéaire s'écrit alors :    yr = a*x + b
```

- **Écart-type du résidu**
  La syntaxe `[a,b,sig] = reglin(x,y)` retourne de plus `sig` l'écart-type du résidu.
- **Part expliquée, part inexpliquée de la variance**
  - On obtient la part expliquée $\mathrm{Var}(a\cdot X+b) = \rho^2(X,Y)\cdot\mathrm{Var}(Y)$ comme suit :

```
partExpliquee = variance(a*x+b)
```

  - On obtient la part inexpliquée $\mathrm{Var}(\epsilon) = \big[1-\rho^2(X,Y)\big]\cdot\mathrm{Var}(Y)$ comme suit :

```
partInexpliquee = sig^2
coefDetermination = 1 - sig^2 / variance(Y)
// coefDetermination proche de 1 <=> régression linéaire de qualité
// coefDetermination proche de 0 <=> piètre régression linéaire
```