Statistique mathématique

Examen final — 6 mai 2020

Durée : 3 heures. Fournissez des réponses complètes aux questions : justifiez vos réponses, expliquez votre raisonnement, détaillez vos calculs. Concernant la présentation, écrivez proprement et encadrez vos résultats. Les questions marquées d'un astérisque (⋆) sont un peu plus difficiles, vous êtes libres de les ignorer dans un premier temps et d'y revenir plus tard. Bonne chance !

Estimation et croissance exponentielle

Dans ce sujet nous nous intéressons à un sujet d'actualité : la propagation d'un virus dans une population. Expérimentalement, on constate souvent que lors de la première phase d'une épidémie, le nombre de personnes infectées augmente de manière exponentielle. Il est alors crucial d'estimer la *vitesse* de cette croissance exponentielle, par exemple pour pouvoir estimer le nombre de personnes infectées à un horizon de temps fixé. Une quantité associée à cette vitesse est le R_0, le nombre moyen de personnes qu'une personne malade infecte. Dans tout le sujet nous ferons l'hypothèse simplificatrice que nous avons accès au nombre exact de personnes infectées (toutes les personnes infectées sont détectées).

Partie I : le modèle SIR

Le modèle compartimental le plus simple en épidémiologie est le modèle SIR (Susceptible - Infected - Recovered, Kermack–McKendrick *c.a.* 1930). Dans ce modèle, la population se compose d'un nombre constant d'individus N, qui se séparent à chaque instant t en trois groupes : les personnes susceptibles d'être infectées ($S(t)$), les personnes infectées ($I(t)$), et les personnes guéries ($R(t)$). La dynamique de ces trois groupes est régie par un système d'équations différentielles couplées :

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\beta I S}{N}\,, \quad \frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\beta I S}{N} - \gamma I\,, \quad \text{et} \quad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I\,,$$

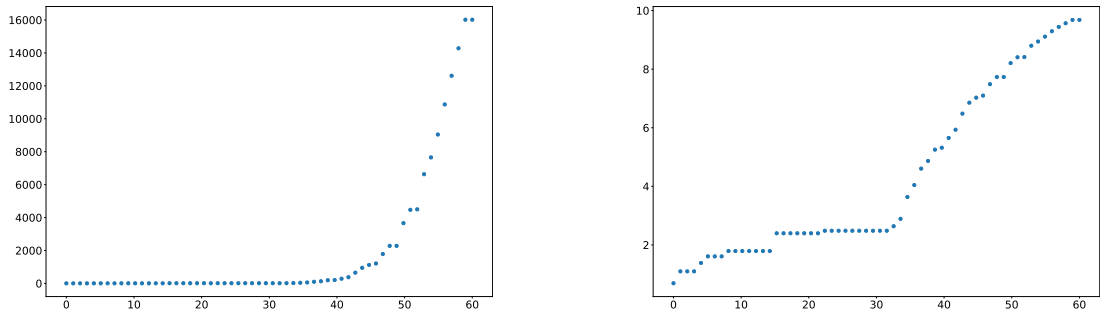
où β et γ sont des constantes positives. Ce qu'on appelle R_0 est en fait la constante égale à β/γ.

1. Prouvez que $S(t) + I(t) + R(t) = N$ pour tout $t > 0$.
2. Prouvez que, pour tout $t > 0$,
$$\frac{\mathrm{d}I}{\mathrm{d}t} = \left(R_0 \frac{S}{N} - 1\right) \gamma I\,.$$
3. Supposons que l'on soit dans les premiers jours de l'épidémie. Dans ce régime, $S \approx S(0)$. On fait l'hypothèse que $S(0)$ est très proche de N. Montrez que dans ce cas I satisfait une équation différentielle ordinaire à coefficients constants.
4. Résolvez l'équation différentielle obtenue dans la question 3 et montrez que
$$I(t) = I(0) \exp(\lambda t)\,, \tag{1}$$
avec λ une constante positive que vous écrirez en fonction des données du problème.
5. Expliquez ce qui se passe quand $R_0 < 1$, quand $R_0 = 1$, et quand $R_0 > 1$. Faites des dessins.

Partie II : régression linéaire

Nous nous tournons maintenant vers l'analyse de données épidémiques réelles. Plus précisément, nous allons estimer λ dans le cas de l'épidémie de COVID-19 sur le territoire national. Dans le graphique de gauche, nous reproduisons le nombre de cas détectés cumulé en France métropolitaine du 22 janvier au 23 mars 2020 d'après la base de données de l'Université John Hopkins. [1] A droite, les mêmes données en échelle logarithmique pour le nombre d'infections.

1. `https://github.com/CSSEGISandData/COVID-19`



En d'autres termes, à gauche nous avons tracé $t \mapsto I(t)$ et à droite $t \mapsto \log I(t)$.

1. L'hypothèse de croissance exponentielle vous paraît-elle justifiée ?

Nous nous limitons maintenant aux données observées à partir du 2 mars, soit $n = 22$ points. On note $p_i = I(t_i)$ le nombre de personnes infectées à la date i ainsi que $p_0 = I(0)$. De plus, on notera $y_i = \log p_i$. Dans cette partie, nous faisons l'hypothèse que (1) est approximativement vérifiée. On en déduit que

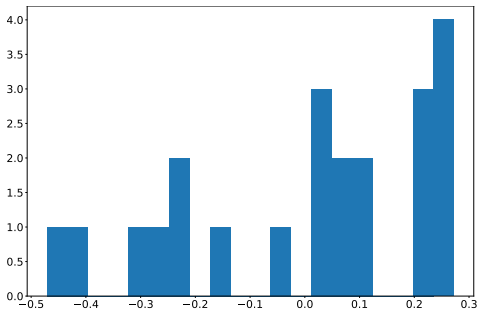
$$y_i = \log p_i = x_0 + \lambda t_i + \varepsilon_i \,,$$

avec ε_i un terme d'erreur. Ainsi nous sommes ramenés à un problème d'estimation dans le modèle linéaire.

2. Expliquez à quoi correspondent les ε_i dans cette situation. Donnez un exemple concret.
3. On calcule les valeurs numériques suivantes :

$$\overline{x} = 11.00\,, \quad \overline{y} = 7.82\,, \quad \mathrm{Cov}\,(x,y) = 9.35\,, \quad \mathrm{Var}\,(x) = 44.17\,, \quad \text{et} \quad \mathrm{Var}\,(y) = 2.03\,.$$

   Donnez les valeurs de $\hat{x}_0$ et $\hat{\lambda}$ obtenues par régression linéaire simple.
4. On admet que les malades restent en moyenne $D = 10$ jours contagieux, et que $\gamma = 1/D$. Grâce à l'estimée $\hat{\lambda}$ de λ et à la formule obtenue dans la question 4 de la partie I, donnez une estimation de R_0 pour le COVID-19.
5. Estimez le nombre de personnes infectées au 30 mars ($t = 29$).
6. On suppose maintenant que les ε_i sont i.i.d. $\mathcal{N}\left(0, \sigma^2\right)$. Donnez un estimateur $\hat{\sigma}_n^2$ de σ^2.
7. Donnez un intervalle de confiance approché à 95% pour $\hat{y}_{29}$. Déduisez-en en intervalle de confiance approché à 95% pour $I(29)$. On ne calculera pas les valeurs numériques.
8. Ci-dessous nous avons tracé un histogramme des résidus $\hat{y}_i - y_i$. Au vu de ce graphique, l'hypothèse gaussienne vous paraît-elle justifiée ? Comment pourrait-on s'en assurer rigoureusement ?



Partie III : régression de Poisson

Dans cette partie, nous allons considérer un autre modèle pour les p_i. Nous allons faire l'hypothèse que chaque p_i suit une *distribution de Poisson* et qu'ils sont indépendants. Rappelons qu'une variable aléatoire X suit la loi de Poisson de paramètre ν si

$$\forall k \in \mathbb{N}, \qquad \mathbb{P}\left(X = k\right) = \frac{\nu^k}{k!}\mathrm{e}^{-\nu}\,,$$

où $k!$ est la *factorielle* de k ($0! = 1$, $(k+1)! = (k+1)\cdot k!$). Nous nous proposons d'estimer λ par la méthode du maximum de vraisemblance.

1. Montrer que si $X \sim \text{Poisson}(\nu)$, $\mathbb{E}[X] = \nu$. Ainsi nous pouvons modéliser les p_i par des Poisson de paramètres $p_0 \exp(\lambda t_i)$ indépendantes les unes des autres.
2. Sous cette hypothèse, écrire la vraisemblance $\mathcal{L}(p_1, \ldots, p_n | \lambda, p_0)$.
3. En déduire que l'estimateur du maximum de vraisemblance pour (p_0, λ) est solution de

$$\max_{\lambda, p_0} \sum_{i=1}^{n} \left\{ p_i \log p_0 + \lambda t_i p_i - p_0 \mathrm{e}^{\lambda t_i} \right\}$$

4. Supposons maintenant p_0 connu. Montrez que trouver $\hat{\lambda}$ revient à maximiser une fonction de λ qu'on notera f.
5. Montrez que f est concave, et donc qu'il suffit de résoudre une équation en λ que vous écrirez. Pensez-vous que cela soit possible en formule fermée ?
6. Prenons $p_0 = 191$ (la valeur expérimentale, cela revient à considérer que $\varepsilon_1 = 0$). Une méthode de résolution approchée de $f'(\lambda) = 0$ nous donne $\hat{\lambda} = 0.22$. Comparez cette valeur à celle obtenue dans la partie II. Calculez le R_0 obtenu et commentez votre résultat.