Exercice supplémentaire

On cherche à analyser la différence de revenu selon le sexe au sein d'une entreprise. Pour cela, on modélise le revenu R_i du i-ième individu d'un échantillon de n personnes par :

$$R_i = \mu + \beta s_i + \varepsilon_i, \quad 1 \leqslant i \leqslant n,$$

où

- on observe les variables aléatoires $R_1, \ldots, R_n$;
- le vecteur $(s_1, \ldots, s_n)$ est **déterministe** avec $s_i = 1$ si le i-ième individu est une femme, et $s_i = 0$ sinon ;
- les variables aléatoires $\varepsilon_1, \ldots, \varepsilon_n$ sont indépendantes et identiquement distribuées selon la loi normale $\mathcal{N}(0, \sigma^2)$ avec $\sigma^2 > 0$ **connue** ;
- le paramètre $\mu \in \mathbb{R}_+$ est supposé **connu**.
- le paramètre $\beta \in \mathbb{R}$ est **inconnu**.

On cherche à estimer β à partir du n-échantillon $(R_1, \ldots, R_n)$.

1. Comment s'interprète les coefficients μ et β dans cette analyse de différence de salaire ?
2. Quelle est la loi de R_1 ? Les variables aléatoires $(R_1, \ldots, R_n)$ sont-elles indépendantes ? Ont-elles la même loi ?
3. Calculer la log-vraisemblance de $(R_1, \ldots, R_n)$.
4. À quelle condition sur $(s_1, \ldots, s_n)$ l'estimateur du maximum de vraisemblance $\widehat{\beta_n}$ de β est-il bien défini ? Comment s'interprète cette condition dans cette analyse de différence de salaire ? En supposant que cette condition est vérifiée, calculer l'estimateur du maximum de vraisemblance $\widehat{\beta_n}$.
5. Quelle est la loi de $\widehat{\beta_n}$?
6. En déduire le biais et l'erreur quadratique moyenne de $\widehat{\beta_n}$.
7. Soit $\alpha \in]0, 1[$. Construire un intervalle de confiance au niveau $(1 - \alpha)$ pour β.
8. (a) Construire un test de niveau α pour l'hypothèse (H_0) :"le salaire ne dépend pas du sexe" contre (H_1) :"le salaire dépend du sexe".
   (b) Calculer l'erreur de seconde espèce du test construit précédemment. Comment se comporte l'erreur de seconde espèce en fonction de α ?
9. Construire un test de niveau α pour l'hypothèse (H_0) :"le salaire des hommes est plus élevé que celui des femmes" contre (H_1) :"le salaire des hommes est moins élevé que celui des femmes".
10. *Pour aller plus loin*
    On suppose maintenant que le paramètre $\theta = (\mu, \beta) \in \mathbb{R}^2$ est **inconnu**. On cherche à estimer $\theta = (\mu, \beta)$ à partir du n-échantillon $(R_1, \ldots, R_n)$. À quelle condition sur $(s_1, \ldots, s_n)$ l'estimateur du maximum de vraisemblance $\widehat{\theta_n} = (\widehat{\mu}, \widehat{\beta})$ de θ est-il bien défini ? En supposant que cette condition est vérifiée, calculer l'estimateur du maximum de vraisemblance $\widehat{\theta_n}$.