TP2: Régression linéaire - Travaux pratiques

Visseho Adjiwanou, PhD.

31 October 2021

Exercice 1

Supposons que $y = \alpha + \beta * x + \epsilon$, où les ϵ sont indépendants et identiquement distribuées (iid) avec la moyenne 0 et la variance σ^2. Supposons que les données sont divisées uniformément en deux groupes désignés par les indices a et b, et β est estimé par $\beta^* = (y_a - y_b)/(x_a - x_b)$ où y_a est la moyenne de toutes les observations y du groupe a, etc. 1. Définir la forme algébrique de y_a, y_b, x_a, x_b 2. Montrer que β^* est un estimateur sans biais de β 3. Trouver la variance de β^*, $Var\beta^*$ 4. Démontrer que $Var\beta^* = (2\sigma)^2/T(x_a - x_b)^2$ dans le cas spécifique où le nombre d'observations dans chaque groupe est T/2, T étant le nombre total d'observations 5. Comment répartiriez-vous les observations entre les deux groupes? Pourquoi?

Exercice 2

upposons que nous ayons les échantillons x_1, x_2 et x_3, choisis au hasard dans une distribution de moyenne 4 et variance 9. Vous avez deux estimations de la moyenne: $\mu^* = (x_1 + x_2 + x_3)/3$ et $\mu^{**} = (x_1 + x_2 + x_3)/4$

1. Calculez l'espérance des deux estimateurs ($E\mu^*$ et $E\mu^{**}$)
2. Calculer leurs variances
3. Quel estimateur choisirez-vous? Pourquoi?

Exercice 3

L'erreur quadratique moyenne (EQM ou MSE en anglais) de l'estimateur $\hat{\theta}$ est défini par $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$, $(\hat{\theta} - \theta)$ est appelé l'erreur d'échantillonnage, et $E(\hat{\theta} - \theta)$ est le biais. 1. Définissez dans votre propre mot votre compréhension de MSE 2. Démontrez que $MSE(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2$ = Variance + biais carré

Exercice 4

Démontrer que la valeur attendue d'une fonction de perte constituée du carré de la différence entre β et son estimation (c'est-à-dire le carré de l'erreur d'échantillonnage) est identique à la somme de la variance et du biais au carré.

Exercice 5

La vraie relation entre X et Y dans la population est donnée par: $Y_i = 2 + 3X_i + \epsilon_i$. Supposons que les valeurs de X dans l'échantillon de 10 observations sont 1, 2,... .., 10. Les valeurs des termes d'erreurs sont tirées au hasard parmi une population normale de moyenne nulle et de variance 1.

- $\epsilon_1 = 0.464$

- ϵ_2 = 0,137
- ϵ_3 = 2,455
- ϵ_4 = -0,323
- ϵ_5 = -0,068
- ϵ_6 = 0.296
- ϵ_7 = -0,288
- ϵ_8 = 1,298
- ϵ_9 = 0,241
- $\epsilon_1 0$ = -0,957

1. Déterminez les 10 valeurs observées de X et Y
2. Placez les informations sur un graphique en utilisant (ggplot)
3. Utilisez les formules des moindres carrés pour estimer (manuellement) les coefficients de régression et leurs erreurs standard et comparer les résultats avec les valeurs vraies
4. Tracez la droite estimée (ou prédite) en utilisant R dans le même graphique qu'en 2.
5. Estimer les paramètres en utilisant cette fois-ci R.

Exercice 6

Examinez les données ci-dessous sur les prix et les quantités d'oranges vendues dans un supermarché sur douze jours consécutifs.

Prix (.01$/lb)	Quantité (lb)
100	55
90	70
80	90
70	100
70	90
70	105
70	80
65	110
60	125
60	115
55	130
50	130

Soit X_i le prix demandé et Y_i la quantité vendue le jour de la vente. Supposons en outre que la fonction de demande est de la forme $Y_i = \alpha + \beta * X_i + \epsilon_i$ et que les hypothèses de base du modèle de régression normale classique sont satisfaites.

1. Rappeler ces hypothèses.
2. Estimer les paramètres du modèle (manuellement)
3. Estimer la droite de régression de l'échantillon et représenter-la dans un graphique en même temps que les nuages de points (avec ggplot)
4. Calculer les estimations biaisées et non biaisées de la variance du terme d'erreur
5. Calculez la variance estimée de α^* et de β^*.
6. Calculer le R^2 et Commenter.
7. Comparer vos résultats avec ce que vous obtenez avec R. Relier chaque estimation de R à ce que vous estimez manuellement.

8. Qu'avez-vous du tableau de régression R que vous n'avez pas encore calculé manuellement? Avez-vous une idée de ce qu'il faut faire?