PARIS 7 SD1
2010-2011

TP machines n° 4 (groupes MASS 3 et Math Info 5)

Régression linéaire

1. Introduction Dans cette section, $\hat{\beta}_0$ et $\hat{\beta}_1$ sont l'intercept et la pente de la droite de régression dérivés de l'échantillon bivarié $(x_i, y_i)_{1 \leqslant i \leqslant n}$.

- Vérifier que les échantillons sont centrés :

$$\sum_{i}^{n} \left(y_i - \hat{\beta}_1 x_i - \hat{\beta}_0 \right) = 0$$

- Vérifier que les résidus $\hat{e}_i = y_i - \hat{\beta}_1 x_i - \hat{\beta}_0$ ne sont pas corrélés avec les x_i, i.e. :

$$\sum_{i}^{n} \hat{e}_i x_i = 0$$

On pourra utiliser que $\min_{a \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \bar{y}_n - a(x_i - \bar{x}_n))^2 = \sum_{i=1}^{n} \left(y_i - \bar{y}_n - \hat{\beta}_1 (x_i - \bar{x}_n) \right)^2$

On cherche à expliquer la variable Y à partir de la variable X par une relation de la forme $\mathrm{Y} = \gamma \mathrm{X}^{\delta}$, où $\gamma > 0$ et $\delta \in \mathbb{R}$. Les variables X et Y sont supposées à valeurs positives.

- Montrer que l'on peut transformer X et Y pour obtenir deux variables X′ et Y′ liées par une fonction affine.
- Si on part d'un échantillon bivarié $(x_i, y_i)_{1 \leqslant i \leqslant n}$, quelles sont les valeurs de γ et δ minimisant $\sum_{i=1}^{n} \left(log \left(log \frac{y_i}{\gamma x_i^{\delta}} \right) \right)^2$

Si l'on impose que l'intercept $\hat{\beta}_0$ soit égal à 0, calculer le coefficient de régression $\hat{\beta}_1$ minimisant $\sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$ en fonction de $\bar{x}_n$, $\bar{y}_n$, σ_x, σ_y et le facteur de corrélation ρ.

2. Mise en pratique L'échantillon `cars` contient les vitesse et distances de freinage sur plusieurs tests. On cherche à étudier le comportement de la distance `dist` en fonction de la vitesse `speed`. Commenter le `plot (cars)`. On cherche à savoir si une dépendance linéaire est pertinente : `lm(dist  speed, cars` . Interpréter. On rapelle que `Multiple R-squared` renvoie le carré du coefficient de corrélation ρ^2. Comme

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 x_i - \hat{\beta}_0 \right) 2 = n \left(1 - \rho^2 \right) \sigma_y^2,$$

plus `Multiple R-squared` sera prohe de 1 et meilleure sera le régression !
Est-il plus pertinent d'essayer d'expliquer `dist` par `speed*speed` ?