M2 recherche DM 3: régression linéaire

Emeline Schmisser, `emeline.schmisser@math.univ-lille1.fr`, bureau 314 (bâtiment M3).

Exercice 1 (4 points)
On dispose d'observations $x_1, \dots, x_n$ et $y_1, \dots, y_n$. On suppose que $Y = aX$. Pour simplifier les notations, on pourra noter $\mathbb{E}(X) = \sum x_i/n$ et $\mathbb{E}(X^2) = \sum x_i^2/n$ (et ainsi de suite).

1. (1 point) Donner la formule permettant de trouver $\hat{a}$.
2. (1 point) Que vaut la somme des résidus ?
3. (0.5 points) Quelle est la covariance empirique entre x et les résidus ?
4. (0.5 points) Quelle est la covariance empirique entre y et les résidus ?
5. (1 point) Si $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, quelle est la loi de a ?

Exercice 2 (4 points)
On sait que les variables Y suivent une loi $Y = b\exp(aX)$. On voudrait tout de même faire une régression linéaire.

1. (2 points) Avec quelles variables va-t-on faire une régression linéaire ?
2. (1 point) Donner les formules permettant de calculer a et b.
3. (1 point) On observe les variables y_i avec un bruit. Comment doit être ce bruit pour qu'avec les variables avec lesquelles on fait la régression linéaire, on ait un bruit additif ($u_i = c_1 v_i + c_2 + \varepsilon$) ?

Exercice 3 (2 points)
Cette fois-ci, on ne veut pas faire une régression de moindres carrés. On veut minimiser

$$|y_i - b|$$

1. (2 points) Que vaut b ?

Exercice 4 (10 points)
Télécharger le fichier "logements.csv". On veut expliquer le prix d'un logement en fonction de sa surface. Le but est bien évidemment de trouver le meilleur modèle possible. On pourra utiliser pour cet exercice soit les fonctions qu'on a tabulée lors du TP, soit les fonctions déjà programmées dans R.

On commence par étudier le modèle linéaire.

1. (0.5 points) Tracer l'évolution du prix en fonction de la surface.
2. (0.5 points) Faire une régression linéaire du prix en fonction de la surface. Ajouter la droite de régression sur le graphique.
3. (0.5 points) Quelle est l'expression des coefficients dans ce cas ?
4. (1 point) Expliquer quelle loi suivent les estimateurs des coefficients si

$$prix = a * surface + b + bruit$$

avec *bruit* des variables indépendantes identiquement distribuées gaussiennes centrées et de variance σ^2.

5. (2 points) Tester si les coefficients que l'on a trouvé sont nuls. Le cas échéant, refaire une régression linéaire avec le modèle que vous pensez le plus approprié. Tester si les coefficients sont nuls. Ajouter la nouvelle droite de régression sur le modèle. Comparer si besoin est les modèles avec le critère qui vous semble le plus pertinent.

   On sait que le loyer ne varie en général pas linéairement avec la surface, les petites surfaces reviennent souvent plus cher au mètre carré que les grandes.
   On suppose maintenant que

$$prix = a * surface + b\sqrt{surface} + c$$

6. (0.5 points) Faire une régression dans ce cas. Ajouter la courbe de régression au graphique. Commenter.
7. (1 point) Faire le test de nullité des coefficients. Conclure.
   On suppose maintenant que

$$prix = a * \sqrt{surface} + b$$

8. (1 point) Faire la régression dans ce cas. Tracer la courbe de régression. Commenter.
9. (1 point) Faire le test de nullité des coefficients. Si un des coefficients semble petit, refaire la régression avec le modèle qui vous semble le mieux adapté.
10. (2 points) Utiliser le critère de sélection et le critère de sélection ajustée pour choisir le meilleur modèle. Est-ce que cela vous semble cohérent avec le graphique ?