TISD – FICHE 5

Régression linéaire simple (R et SAS)

Adrien Hardy, adrien.hardy@math.univ-lille1.fr

1 Données immobilières (avec R)

Le fichier "donneesImmobilieres_euros.txt", disponible sur Moodle, contient les variables `surface` (en m^2) et `prix` (en euros, par mois) d'un échantillon de 34 appartements en location dans un même quartier de Paris. On veut décrire précisément l'interdépendance entre ces deux variables.

1. Faire un graphique de la variable `prix` en fonction de la variable `surface`. Commenter.

On considère le modèle de régression linéaire $Y = aX + b + \varepsilon$, où X et Y représentent respectivement les variables `surface` et `prix`.

2. Quelle est la variable expliquée, et quelle est la variable explicative ?

3. Calculer les estimateurs $\hat{a}$ et $\hat{b}$ des coefficients a et b.

4. Superposer au nuage de points obtenu à la question 1 la droite d'équation $y = \hat{a}x + \hat{b}$.

Pour obtenir plus rapidement ces coefficients avec **R**, on utilise commande `lm(Y~X)` pour expliquer Y par X.

5. Retrouver les coefficients $\hat{a}$ et $\hat{b}$ obtenu ci-dessus à l'aide de cette commande ; qui est `intercept` ?

6. Calculer le coefficient de détermination R^2. Commenter.

Si l'on nomme cette regression linéaire simple, par exemple `RLS <- lm(Y~X)`, on peut obtenir plus d'informations en tapant `summary.lm(RLS)`.

7. Retrouver R^2 dans ces informations ; quel est son nom ?

8. Effectuer un test de Fisher (test de significativité) : Quelle est l'hypothèse $\mathcal{H}_0$, quelle est la statistique d'intérêt, ainsi que son comportement sous $\mathcal{H}_0$, puis comparer au quantile d'une loi de Fisher de paramètres appropriés, et conclure. Retrouver les résultats de ce test dans `summary.lm`.

9. Etudier les résidus de la régression : Moyenne, écart-type, histogramme, dessiner-les sur un graphe. Est-ce que l'hypothèse de résidus gaussiens est validée ?

10. Calculer l'estimateur sans biais $\hat{\sigma}^2$ de σ^2, puis donner un intervalle de confiance à 95% des paramètres a et b.

11. Prédire le prix d'un appartement de $140m^2$ avec un intervalle de confiance à 95%.

12. Si vous louez un $35m^2$ à Paris pour 800 euros par mois, avez-vous fait une bonne affaire ?

2 Explication du pic d'ozone (avec SAS)

Le fichier SAS "ozone2.sas7bdat", disponible sur Moodle, contient les variables suivantes, pour une série de journées :

- l'identifiant de la journée (`date`),
- le maximum d'ozone (`maxO3`)
- l'heure à laquelle le maximum d'ozone a été obtenu (`heure`),
- les températures à 6h, 9h, 12h, 15h, 18h (resp. `T6` à `T18`)
- la nébulosité à 6h, 9h, 12h, 15h, 18h (resp. `Ne6` à `Ne18`)
- la projection du vent sur l'axe est-ouest à 12h (`Vx`),
- le maximum d'ozone de la veille (`maxO3v`).

Remarque : Ce jeu de données n'est pas identique à celui utilisé dans le TP4.

Le but de cet exercice est de comprendre comment la valeur des pics d'ozone est liée à d'autres grandeurs physiques facilement mesurables (température, heure, nébulosité, vent), afin d'avoir des approximations de la qualité de l'air faciles et rapides à obtenir.

1. Observer la corrélation entre les différentes variables à l'aide de `proc corr`. Quelle est la variable la moins corrélée linéairement avec `maxO3` ? Et la variable (différente de `maxO3`) la plus corrélée ?

Maintenant, on s'intéresse au lien entre la valeur du pic d'ozone `maxO3` et la température à midi `T12`.

2. Analyser les variables `maxO3` et `T12` indépendamment : moyenne, boîte à moustache, histogramme, ...

3. Effectuer la régression de `maxO3` en fonction de la variable `T12` à l'aide de `proc reg`. Donner les coefficients $\hat{a}$, $\hat{b}$, et R^2. Est-ce que le test de significativité de Fisher est concluant ? Et que peut-on dire de la qualité de la regression ?

4. Que représentent la zone en bleu et celle délimitée par des pointillets dans le dernier graphique ? A l'oeil nu, donner des bornes approximatives sur la valeur de y à 95% si $x = 20$.

5. Demander de plus à la `proc reg` de tracer le graphe de `maxO3` en fonction de `T12`.

Pour stocker les résultats dans fichier de la bibliothèque locale malib, rajouter dans `proc reg` :
```
output out=malib.resultat_reg p=predictions r=residus;
```

6. Faire une `proc univariate` sur les résidus ε_j et tester s'ils peuvent être gaussien.

7. S'il reste du temps, refaire cette analyse pour `maxO3` et d'autres variables de votre choix.