Statistique descriptive – 1re année [1]

Feuille de T.P. 8.

Objectifs. Ajustement linéaire entre deux variables quantitatives.

Instructions. Ce T.P. et votre compte-rendu sont à finir d'ici **Vendredi 28 décembre** minuit. Le compte-rendu devra être déposé sur *moodle*. Il devra être nommé :

`TP8_Prenom_Nom.pdf`

Si le compte-rendu a été fait en binôme, les deux étudiants doivent le déposer sur *moodle* et spécifier **explicitement** dans le texte les deux noms.

Nous reprenons dans ce T.P. les données du jeu de données *Recensement85* du T.P. précédent, où l'on avait créé une variable contenant le logarithme des salaires. Sur Moodle, enregistrer le jeu de données correspondant *Recensement85Bis.sta* (DonneesTP8) dans vos documents. Créer un nouveau classeur dans *Statistica*, et ajouter cette feuille de données au classeur.

Nous allons étudier l'ajustement linéaire du nombre d'années d'expérience en fonction de l'âge et celui du salaire en fonction de l'âge .

1 Etude du nombre d'années d'expérience en fonction de l'âge

1.1 Ajustement linéaire du nombre d'années d'expérience en fonction de l'âge

1. Proposer un graphique permettant d'illustrer un lien éventuel entre le nombre d'années d'expérience et l'âge de chacun des 534 individus.
2. Calculer le coefficient de corrélation linéaire entre le nombre d'années d'expérience et l'âge, comme dans le T.P.7. Vous semble-t-il pertinent d'envisager un ajustement linéaire du nombre d'années d'expérience en fonction de l'âge ?
3. Essayons ! Calculer les coefficients $\hat{a}$ et $\hat{b}$ de la droite des moindres carrés du nombre d'années d'expérience en fonction de l'âge. On pourra utiliser le menu *Statistiques* > *Modèles Avancés* > *Modèles de régression*. Choisir *Régression simple*. Sélectionner les variables, cliquer sur *OK*. Dans l'onglet *Base* de la nouvelle boite de dialogue, choisir *Coefficients*. **Conserver cette boîte de dialogue ouverte, elle sera utile dans la suite du T.P.**.
4. Représenter la droite des moindres carrés sur le nuage de points. On pourra cocher la case correspondant au type d'ajustement *linéaire* dans la boîte de dialogue du nuage de points correspondant, et dans l'onglet *Avancé* de cette même boîte de dialogue, cocher les cases pertinentes pour vérifier les calculs des deux questions précédentes.
5. Conclure quant à la qualité de l'ajustement linéaire obtenu.

1. Enseignant responsable des TP : G.Chagny (gaelle.chagny@parisdescartes.fr) et C.Laclau (charlotte.laclau@parisdescartes.fr)

1.2 Analyse des résidus et de la variance

Notons X la variable « Age » et Y la variable « Experience ».

1. Créer une feuille de données contenant les valeurs ajustées $\hat{Y}_k = \hat{a}X_k + \hat{b}$ et les résidus $\hat{e}_k = Y_k - \hat{Y}_k$ correspondants, pour $k = 1, \ldots, 534$. On pourra retourner dans la boîte de dialogue de la régression simple (question 3 de la Section 1.1). Dans l'onglet *Résidus*, choisir *Val prévues & résidus*. **Définir cette nouvelle feuille de données comme feuille active, pour toute la suite de la section**. Renommer convenablement les différentes variables.
2. On cherche à commenter les caractéristiques des valeurs ajustées et résidus :
   - Comparer la moyenne des valeurs ajustées et des valeurs observées.
   - Représenter sur un même graphique les valeurs ajustées et les valeurs observées (on pourra utiliser un nuage de points).
   - Calculer la moyenne et l'écart-type des résidus. Commenter.
3. Représenter sur un graphique tous les points $(X_k, \hat{e}_k)_{1 \leq k \leq 534}$. Pour cela,
   - ajouter une colonne (une variable), à la feuille de données contenant les résidus. Copier, sur la feuille de données de départ, les valeurs de la variable *Age*, et les coller dans la nouvelle colonne de la feuille des résidus.
   - utiliser le type de graphique *Nuage de points* pour représenter les points demandés.

   Pensez-vous qu'il existe encore une information dans les résidus qui n'est pas prise en compte par l'ajustement linéaire de Y par rapport à X ?
4. Représenter l'histogramme en fréquence des résidus. Commenter.
5. La formule de décomposition de la variance est

$$\mathrm{Var}(Y) = \mathrm{Var}(\hat{Y}) + \mathrm{Var}(\hat{e})\,,$$

   où $\mathrm{Var}(\hat{Y})$ est la variance expliquée par le modèle linéaire et $\mathrm{Var}(\hat{e})$ est la variance dite résiduelle (non expliquée par le modèle). Donner les valeurs de ces variances. Interpréter.
6. Le coefficient de détermination est donné sous le nom de *R square*. Il est défini comme le rapport :

$$R^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)}\,.$$

   Donner la valeur du coefficient de détermination. Retrouver cette valeur à partir du coefficient de corrélation linéaire. Quelle est la proportion de la variance de la variable « Experience » expliquée par la régression linéaire ?
7. Peut-on en conclure que la relation linéaire entre « Experience » et « Age » suffit à expliquer la variabilité de « Experience » ?

2 Ajustement linéaire du logarithme du salaire en fonction de l'âge

Reprendre l'ensemble des questions de la partie précédente en les appliquant cette fois à une régression linéaire de la variable « Log_Wage » en fonction de la variable « Age ».