Corrigé de l'exercice 1

Jean-François Bickel

Statistique II – SP08

- L’objectif est d’apprendre à lire et interpréter les résultats produits par SPSS
- Les références théoriques se trouvent dans les documents « Introduction à l’analyse de régression » et « Introduction à l’analyse de régression (2) »
- Le premier des documents cités contient aussi la manière de procéder pour demander une analyse de régression via l’interface graphique de SPSS

Syntaxe SPSS

```
regression
 /missing listwise
 /statistics defaults ci
 /noorigin
 /dependent i05wy
 /method=enter age05.
```

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.285[a]	.081	.081	55147.345

a. Valeurs prédites : (constantes), Age durant l'année de l'interview

- r est une mesure de corrélation entre x et y (ici x=âge et y=revenu)
- Il s'agit d'une mesure de la force de l'association
- La corrélation entre âge et revenu de .29
- Son signe positif indique une relation positive entre les deux variables
- Dans le cas de la régression bi-variée, la valeur de r est identique à celle du coefficient standardisé de b (voir plus bas)

- r^2, dit aussi coefficient de détermination, est une mesure de réduction proportionnelle de l'erreur
- Il s'agit également d'une mesure d'association entre x et y
- r^2 peut être calculé directement à partir de r
- En régression multivariée, on l'utilise de manière générale pour mesure l'adéquation globale (fit) du modèle aux données

- Le r^2 ajusté est une mesure qui corrige r^2 pour tenir compte du nombre de paramètres dans le modèle et la perte de degré de liberté résultant de l'accroissement de ce nombre
- Dans une régression multivariée, quand interviennent de nombreux paramètres, on préfère souvent utiliser le r^2 ajusté plutôt que simplement la mesure de r^2

- L’erreur standard de l’estimation (standard error of the estimate) est l’expression qu’utilise SPSS pour désigner l’écart type conditionnel de y
- Rappelons que sa formule est

$$s = \sqrt{SSE / (n - 2)}$$

- Ici s = 55147.345

ANOVA[b]

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	1.13E+012	1	1.127E+012	370.490	.000[a]
	Résidu	1.27E+013	4184	3041229647		
	Total	1.39E+013	4185			

a. Valeurs prédites : (constantes), Age durant l'année de l'interview

b. Variable dépendante : Revenu annuel du travail (sans considération brut ou net)

- Dans le tableau « ANOVA » figurent les résultats des calculs des « sommes des carrés des erreurs » de prédiction
- La ligne « Total » réfère à la somme des carrés des erreurs en l'absence de prédicteur, soit
  $TSS = \sum (y - \bar{y})^2$
- La ligne « Résidu » réfère à la somme des erreurs avec l'équation de régression, soit
  $SSE = \sum (y - \hat{y})^2$

- La ligne « Régression » est la différence entre TSS et SSE, soit le numérateur de la mesure de r^2
- Elle indique la quantité de la variation totale de y qui est expliquée par x en utilisant la droite des moindre carrés
- Si on divise sa valeur par la somme totale des erreurs (TSS), on obtient r^2

- Le tableau « ANOVA » rapporte également les degrés de liberté (dl) associés aux différentes sommes des carrés
- Total: dl = n – 1 = 4185
- Résidu: dl = n – 2 = 4184
- Régression: dl = 1
  (=nombre de variables indépendantes dans le modèle de régression)

- Si on divise la valeur du résidu par son degré de liberté, on obtient la mesure s^2 rapportée dans la colonne « Carré moyen » (*mean square*)
- Ici s^2 vaut 3041229647
- Si on prend la racine carré de s^2, on obtient s
  - qui n'est autre que l'écart type conditionnel de y indiqué sous « Erreur standard de l'estimation » du tableau « Récapitulatif »

- SPSS fournit également les résultats d'un test de F selon lequel l'ensemble des coefficients des x sont égales à zéro
- I.e. un test de l'absence de toute relation entre x et y
- Dans un tel cas, le modèle ne serait tout simplement pas « meilleur » que celui où ne figure aucun x

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Intervalle de confiance à 95% de B	
		B	Erreur standard	Bêta	t	Signification	Borne inférieure	Borne supérieure
1	(constante)	8056.675	2836.483		2.840	.005	2495.662	13617.689
	Age durant l'année de l'interview	1284.654	66.742	.285	19.248	.000	1153.804	1415.503

a. Variable dépendante : Revenu annuel du travail (sans considération brut ou net)

- L'équation de prédiction est (en arrondissant):

  revenu = 8057 + 1285(âge)

- Le revenu augmente donc en moyenne de b=1285 francs pour chaque année d'âge supplémentaire

- Ou encore, les personnes d'un âge donné ont un revenu en moyenne plus élevé de 12850 francs que leurs cadets âgés de 10 ans de moins
- Par exemple, les personnes âgées de 40 ans ont un revenu moyen (prédit) de:

  8057 + 1285(40) = 59457 francs

- Dans SPSS, « Beta » dénote le coefficient standardisé
  (et non β qui est inconnu!)
- Dans une régression bi-variée, sa valeur est égale à la corrélation r entre x et y
- Elle est positive et égale à .29
- Et se lit: « Quand l'âge augmente d'un écart type, le revenu augmente de 0.29 écart-type »

- Le tableau « Coefficients » rapporte également l'erreur standard (se) pour le coefficient b, soit 66.74
- Cette valeur est une estimation de la variabilité du coefficient b dans le cas où seraient sélectionnés de manière répétée des échantillons aléatoires de même taille dans la population de référence, et que dans chacun de ces échantillons serait calculée la même équation de prédiction

- Pour tester l'hypothèse d'indépendance $\beta = 0$,
  on effectue un test de t, la valeur de t étant
  $t = b/se = 19.25$ (cf. colonne « t »)
- Ce test s'effectue avec n – 2 = 4184 degré de liberté
- La valeur de p est indiquée sous signification:
  elle est très petite (<.001), donc l'hypothèse d'indépendance peut être rejetée

- L'intervalle de confiance indique les bornes entre lesquelles on peut être confiant (à 95%) que se situe β
- Ici, β se situe (en arrondissant) entre 1154 et 1416
- Donc, pour chaque année d'âge supplémentaire, le revenu moyen dans la population augmente dans une « fourchette » de 1154 à 1416 francs

- Dans la population, le revenu moyen des personnes de 40 ans se situe entre:

  8057 + 1154(40) et 8057 + 1416(40)

  donc entre

  54217 et 64697 francs