**EPICB325 - Statistiques pour Sciences humaines II**

**TP8 (Partie pratique d'un examen)**

*Les réponses proposées dans ce correctif sont très complètes. On ne s'attend pas que vous soyez aussi détaillés lors de l'examen, mais vous devez toujours expliquer et interpréter vos résultats.*

***Conseils préliminaires***

* Lisez attentivement les questions et répondez dans ce fichier (directement après chaque question).
* **Sauvez très régulièrement votre fichier Word!**
* Quand vous avez répondu à toutes les questions, collez les codes R que vous avez utilisés à la fin de ce document, dans la section ***Appendice: code R***.
* Dans les exercices, il faut parfois ajouter une figure. Vous pouvez exporter un graphique à partir de Rstudio en cliquant sur *"Export…"* au-dessus du graphique, puis *"Copy to clipboard…".* Vous pouvez changer la largeur et/ou hauteur du graphique avant de cliquer sur *"Copy Plot"*. Ensuite vous le collez directement dans le fichier Word.
* A la fin, sauvegardez ce fichier comme pdf (*Fichier -> Enregistrer sous…*, puis dans *Format du fichier*, choisissez pdf). Pour le nom, suivez la convention *filière-nom-prénom*, ou *filière* peut être *pol* ou *commu*. Par exemple, *commu-kiriliouk-anna.pdf.*
* Chargez le fichier pdf ainsi obtenu dans l'espace dédié sur webcampus (*Examen janvier 2022: remise).*

***Description des données***

Les données concernent les salaires annuels des professeurs universitaires aux États-Unis en 2008-2009. Ces données ont été collectés dans le but de discerner une potentielle différence entre les salaires des hommes et des femmes.

Le fichier salaires contient n = 397 observations des variables suivantes:

* rank : indique le grade d'un professeur; assistant (premier grade), associate (deuxième grade), full (troisième grade).
* discipline : indique la discipline d'un professeur; A (département théorique) ou B (département appliqué).
* yrs.since.phd: nombre d'années depuis l'obtention du doctorat.
* yrs.service: nombre d'années de travail en tant que professeur.
* sex: le sexe d'un professeur, female (femme) ou male (homme).
* salary: le salaire d'un professeur, en milliers de dollars par an.

Récupérez le fichier salaires.RDS sur webcampus, chargez-le, et vérifiez que tout est en ordre avec str avant de procéder:

* data <- readRDS("salaires.RDS")
* str(data)

'data.frame': 397 obs. of 6 variables:

$ rank : Factor w/ 3 levels "assistant","associate",..: 3 3 1 3 3 2 3 3 3 3 ...

$ discipline : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...

$ yrs.since.phd: int 19 20 4 45 40 6 30 45 21 18 ...

$ yrs.service : int 18 16 3 39 41 6 23 45 20 18 ...

$ sex : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 1 ...

$ salary : num 139.8 173.2 79.8 115 141.5 ...

***Questions***

Veuillez répondre directement en dessous de chaque question (ajoutez de l'espace si nécessaire).

1. Quel est le salaire moyen des professeurs? Quel est le salaire moyen des professeurs féminins? Quelle est la proportion des professeurs ayant le grade associate ?

*Le salaire moyen d'un professeur est de 113 707 $ par an. Le salaire moyen d'un professeur féminin est de 101 002 $ par an. La proportion des professeurs ayant le grade "associate" est de 16.12%.*

1. Est-ce que les hommes sont plus représentés dans les départements théoriques ou appliqués, comparé aux femmes? Si vous voulez, vous pouvez montrer un graphique qui illustre votre réponse.

*Le graphe montre que les proportions des hommes et femmes dans les deux types de départements sont quasiment identiques. On peut calculer que la proportion des hommes dans les départements appliquées (B) est de 90.28% tandis que leur proportion dans les départements théoriques (A) est de 90.06 %. Ils sont donc (très !) légèrement plus présents dans les départements appliquées, tandis que les femmes sont très légèrement plus présentes dans les départements théoriques (9.95 % contre 9.72 %).*

**

1. Ajustez un modèle de régression linéaire (appelez-le m1) pour prédire le salaire d'un professeur en fonction du nombre d'années depuis l'obtention du doctorat. Interprétez le coefficient de détermination de ce modèle.

*Le coefficient de détermination r2 est de 17.58 %. Le nombre d'années depuis l'obtention du doctorat explique donc 17.58 % des variations du salaire d'un professeur.*

1. Écrivez l'équation de la droite de régression estimée à la question 3). À quel salaire un professeur pourrait s'attendre 10 ans après l'obtention de son doctorat?

*La droite estimée est*

*salary = 91.7187 + 0.9853 yrs.since.phd*

*Dix ans après son doctorat, un professeur peut s'attendre à un salaire de 91.7187 + 9.853 = 101.572, c’est-à-dire, 101 572 $ dollars par an.*

1. Est-ce que l'hypothèse d'une variance constante des résidus est raisonnable pour ce modèle?

*Non, la graphe ci-dessous montre que la variabilité des résidus augmente quand le salaire augmente. On pourrait y remédier en prenant le logarithme des salaires, mais cela complique l’interprétation du modèle.*



1. Ajustez un modèle de régression linéaire (appelez-le m2) pour prédire le salaire d'un professeur en fonction du nombre d'années depuis l'obtention du doctorat et en fonction du sexe. Faites un nuage de points de yrs.since.phd contre salary et ajoutez les deux droites de régression estimées, une pour les hommes, et une autre pour les femmes. N'oubliez pas d'ajouter une légende. Que voyez-vous?

*Le graphique ci-dessous montre que pour un même nombre d'années depuis l'obtention du doctorat, les hommes gagnent un peu plus que les femmes.*

**

1. Interprétez **soigneusement** les coefficients du modèle m2.

* *beta0 = 85.1818; le salaire d'un professeur féminin qui vient d'obtenir son doctorat est de 85 182 $ par an.*
* *beta1 = 0.9581; avec chaque année passée depuis l'obtention du doctorat, un professeur (homme ou femme) gagnera 958 $ par an supplémentaires.*
* *beta2 = 7.9236; pour un même nombre d'années écoulés depuis l'obtention du doctorat, un professeur masculin gagnera 7924 dollars par an de plus qu'un professeur féminin.*

1. Donnez un intervalle de confiance pour le coefficient de régression de la variable yrs.since.phd et interprétez-le. Utilisez un niveau .

*L'intervalle de confiance pour beta1 de niveau 0.01 est donné par (0.678, 1.238). C'est-à-dire, nous sommes confiants à 99 % qu'avec chaque année passée depuis l'obtention du doctorat, un professeur (homme ou femme) gagnera entre 678 et 1238 $ par an supplémentaires.*

1. Considérons le test d'hypothèses contre , où est le coefficient de la variable sex. Comment peut-on interpréter cette hypothèse alternative? Donnez la p-valeur associée à ce test d'hypothèse et expliquez si la variable sexest significative au niveau .

*L'hypothèse alternative signifie que, pour un homme et une femme professeurs ayant obtenu leur doctorat au même année, l'homme gagnera milliers de dollars par an de plus que la femme (car est présumé positif).*

*La p-valeur de ce test unilatéral est égale à 0.0915/2 = 0.04575. Nous pouvons donc rejeter l'hypothèse nulle est conclure que cette variable est significative au niveau . Nous sommes donc confiants à 95.425% que le sexe est une variable pertinente pour prédire le salaire d'un professeur, et que le salaire des professeurs masculins aux Etats-Unis est, en moyenne et ceteris paribus, plus élevé que celui les femmes.*

1. Ajustez un modèle de régression linéaire (appelez-le m3) pour prédire le salaire d'un professeur en fonction de yrs.since.phd, rank, discipline, et sex. À quel salaire un professeur féminin de grade associate d'un département appliqué pourrait s'attendre 10 ans après l'obtention de son doctorat?

*10 ans après l'obtention de son doctorat, un professeur féminin de grade « associate » d'un département appliquée peut s'attendre à gagner 95 536 $ par an.*

1. Dans le modèle m3, arrive-t-on à la même conclusion concernant la significativité de la variable sexqu’avec le modèle m2 ? Pourquoi (pas)? Même question pour la variable yrs.since.phd? Avez-vous une idée pourquoi c'est le cas?

*Pour la variable "sex", la p-valeur est plus élevée qu'avant; on trouve 0.262/2 = 0.131 et on ne peut plus rejeter l'hypothèse nulle au niveau . On n'a donc pas assez de  « preuves » pour conclure que les salaires des professeurs masculins sont, en moyenne, plus élevés que ceux des professeurs féminins. Pareil pour la variable yrs.since.phd – elle était significative dans le modèle m2, mais pas dans le modèle m3. Une possible explication serait la forte association positive entre yrs.since.phd et rank, et entre sex et rank. En effet, les différences salariales sont majoritairement expliquées par le rang d'un professeur (et un haut rang est bien sûr lié au nombre d’années passés depuis l’obtention du doctorat). En plus, les hommes sont proportionnellement plus présents parmi les hauts rangs que les femmes.*

1. Supposons maintenant que l'on veuille utiliser la variable salary en tant que variable explicative. Pourrait-on prédire le rang d'un professeur en fonction de son salaire avec une régression logistique? Pourquoi (pas)?

*Non car la variable qualitative « rang » a trois catégories. Une régression logistique sert à prédire une variable à deux catégories / valeurs.*

1. Ajustez un modèle de régression logistique pour prédire le sexe d'un professeur en fonction de son salaire. Interprétez le coefficient associé . Que se passe-t-il quand on ajoute rank comme deuxième variable explicative?

*Quand le salaire augmente d'une unité (= 1000 $ par an), la cote d'être un professeur masculin est multipliée par = 1.0186. Étant donné que > 1, la cote (et donc la probabilité) d'être un professeur masculin augmente quand le salaire augmente.*

*Quand on ajoute le rang comme deuxième variable explicative, le salaire n’est plus significative (la p-valeur d’un test bilatéral augmente de 0.00659 à 0.167). Cette observation rejoint celui faite à la question 11), c’est-à-dire, le lien entre le sexe et le salaire s’explique majoritairement par le rang d’un professeur.*