Université Paris Cité

Université Paris-Cité

Projet final de logiciel de Statistiques

Salaires Universitaires

Auteurs:
Cellier Roxane
Peraud Marie
Ritter Braha
Promo 2021-2022

Superviseurs:
Mr. Jérome Collet
Mr. Renaud Mozet

28 avril 2022

Sommaire

1 Introduction:

Nous avons à notre disposition un jeu de données à étudier. Il s'agit d'observations concernant les employés d'une faculté des États-Unis. L'objectif de ces observations est d'étudier les différences salariales parmi les employés, pour en estimer les causes et fournir un modèle fiable expliquant le salaire.

Nous allons premièrement étudier sommairement le jeu de données, pour en détailler ses variables ainsi que leurs critères de positions et dispersions. Nous chercherons également des similitudes avec des lois existantes, et formulerons des hypothèses de corrélation. Ensuite, nous allons tenter d'affirmer ou d'infirmer ces hypothèses en étudiant différents modèles de régressions, pour déduire des conclusions plus représentatives de la réalité. Pour finir, nous allons tenter de simplifier la visualisation de notre jeu de données, pour classer sous différents groupes les employés respectant des critères similaires, et étudier si des groupements sont prévisibles ou non.

2 Statistique descriptive:

2.1 Variables:

Notre jeu de données est composé de 397 observations effectuées sur des employés d'une faculté des États-Unis. Les observations sont détaillées par 6 variables caractérisant les employés. Nous avons trois variables quantitatives et 3 variables qualitatives.

Une variable quantitative est une variable représentée par une valeur numérique, généralement avec une unité de mesure de référence. Nos variables quantitatives sont ici:

- ***Yrs since PhD***, qui représente le nombre d'années depuis la validation de la thèse;
- ***Yrs service***, qui représente le nombre d'années de service , ou ancienneté;
- ***Salary***, qui représente le salaire de l'employé en dollars, calculé sur neuf mois.

Une variable qualitative est une variable représentant un état ou un statut. Nous différencions les variables qualitatives ordinales [1] des variables qualitatives nominales [2]. Nos variables qualitatives sont ici:

- ***Rank***, qui représente le poste occupé par la personne; nous pouvons choisir de classer ses valeurs:
  - *Prof (Professeur)*
  - *AssocProf (Professeur Associé)*
  - *AsstProf (Assistant Professeur)*
- ***Sex***, qui représente le sexe de l'employé; il peut prendre les valeurs Male (homme) et Female (femme);
- ***Discipline***, qui représente le domaine d'étude de l'employé; il peut prendre les valeurs A (théorique) et B (appliquée).

Au vu de notre objectif, nous prendrons comme principale variable expliquée le salaire.

[1] variables pouvant être classées
[2] variables ne pouvant pas être ordonnées

2.2 Analyses univariées:

2.2.1 Répartition/ Absurdité:

La première chose que nous pouvons remarquer est la mauvaise répartition de certaines variables qualitatives. En effet, pour effectuer une analyse cohérente, il faut que les variables explicatives soient également distribuées. Ainsi, cela nous permet d'évaluer pour chaque groupe un échantillon sensiblement de même taille. Avoir un échantillon trop petit risque de fausser les calculs de moyennes et de variances, et ainsi de donner une mauvaise estimation du modèle de régression. Il sera donc important de prendre en compte ces répartitions dans nos conclusions.

La variable la plus disparate est le sexe des employés. Sur 397 observations, il n'y a que 39 femmes – soit 9,8%. Cette forte inégalité rendra difficile l'expression de modèles selon le sexe, et les conclusions tirées de ces modèles ne feront pas preuve d'une bonne fiabilité.

Le poste des employés montre également un grand déséquilibre. En effet, nous pouvons compter 266 professeurs, 64 professeurs associés et 67 assistants professeurs, ce qui veut dire que les deux tiers des observations ne représentent que les professeurs. Les modèles estimés selon les postes risquent donc de présenter un coefficient plus fiable pour les professeurs.

Malgré un nombre légèrement plus important d'employés dans le domaine appliqué, la répartition entre les deux disciplines reste plutôt équilibrée. Nous considérons que cet écart n'aura donc pas grande conséquence dans notre analyse.

2.2.2 Positions et dispersions:

Une autre partie de l'analyse consiste en l'étude des critères de position et de dispersion des données. Nous allons donc visualiser la moyenne, l'écart-type, ainsi que la valeur maximum et la valeur minimum de chacune des variables quantitatives.

Variable	N	Moyenne	Ec-type	Somme	Minimum	Maximum
YsincePhD	397	22.31486	12.88700	8859	1.00000	56.00000
YService	397	17.61461	13.00602	6993	0	60.00000
salary	397	113706	30289	45141464	57800	231545

Figure 1: Statistiques simples des variables quantitatives

2.2.3 Corrélation:

Notre prochain objectif est d'avoir une idée des rapports que peuvent avoir chaque variables les unes avec les autres. Les tests de corrélation vont nous permettre d'établir des rapports de dépendance entre les différentes variables pour, par la suite, déduire les modèles de régression les plus intéressants. Une première partie de ces rapports va nous être donnée par le tableau de corrélation des variables quantitatives. Nous allons donc visualiser s'il existe des relations entre le salaire, le nombre d'années depuis l'obtention du doctorat, et le nombre d'années de service.

	YsincePhD	YService	salary
YsincePhD	1.00000	0.90965 <.0001	0.41923 <.0001
YService	0.90965 <.0001	1.00000	0.33474 <.0001
salary	0.41923 <.0001	0.33474 <.0001	1.00000

Figure 2: Tableau des coefficients de corrélation des variables quantitatives

Ce tableau ne nous permet cependant pas de visualiser les rapports de dépendance avec les variables qualitatives. Pour cela, nous allons procéder à une autre méthode, appelée le test du Chi-2. Étant donné que nous cherchons à expliquer le salaire, nous allons nous concentrer sur les résultats des tests, testant l'indépendance du salaire avec les variables qualitatives. Plus précisément, nous allons regarder les résultats du test de Mantel-Haenszel. L'hypothèse nulle de ce test est l'indépendance des deux variables concernées. En choisissant le seuil 0.1%, nous allons considérer que les deux variables ne sont pas indépendantes si la p-value de ce test est inférieure à ce seuil. Par SAS, nous obtenons les valeurs suivantes:

Indépendance avec le salaire	*p-value du test*
sex	0,0058
rank	< 0,0001
discipline	< 0,0001

Grâce à nos deux tests, nous pouvons déduire certaines liaisons entre plusieurs variables. Le premier test des corrélations nous montre qu'il n'y a pas l'air d'avoir une réelle dépendance entre le salaire et nos deux autres variables quantitatives. Nous voyons par contre un fort rapport entre l'ancienneté et le nombre d'années depuis la validation de la thèse, ce qui paraît logique en suivant le fonctionnement d'un doctorat. Le salaire semble malgré tout sensiblement lié au nombre d'années depuis l'obtention du doctorat.

Le test du Chi-2 nous permet, quant à lui, de rejeter l'hypothèse d'indépendance du salaire avec le poste et la discipline. La p-value du test étant supérieure au seuil définit pour le test du sexe, nous pouvons estimer que l'hypothèse de base n'est pas rejetée. Ainsi, nous pouvons considérer qu'il existe un rapport significatif de corrélation du salaire avec le poste, la discipline, et le nombre d'années depuis la validation de la thèse.

2.2.4 Similarités avec des lois théoriques:

Pour tenter de prévoir l'allure de la répartition des variables quantitatives, nous allons étudier la similarité de chacune de ces variables avec des lois théoriques. Les principales lois que nous allons observer sont la loi normale et la loi de Student. Pour chacun de ces tests, nous allons observer les p-values, qui représentent la probabilité que la variable considérée suive une loi normale ou de Student:

	p-value (loi normale)	*p-value (loi de Student)*
salary	< 0,0001 < 0,0100 < 0,0050 < 0,0050	< 0,0001
YsincePhD	< 0,0001 < 0,0100 < 0,0050 < 0,0050	< 0,0001
YService	< 0,0001 < 0,0100 < 0,0050 < 0,0050	< 0,0001

Les p-values des différents tests sont globalement très petites, nous allons donc choisir de rejeter l'hypothèse de similarité de loi. Ainsi, nous pouvons assez rapidement conclure que nos variables quantitatives ne suivent ni une loi normale, ni une loi de Student. Nous ne pourrons donc pas les étudier en tant que telles.

2.3 Analyses Multivariées:

Pour avoir une première idée des informations présentées dans le jeu de données, nous allons effectuer plusieurs analyses multivariées. L'analyse multivariée est l'étude des relations d'une variable par rapport aux autres. Nous allons ici principalement nous concentrer sur le salaire, pour le visualiser de façon sommaire en fonction de nos autres variables et des données présentées. Nous observerons tout de même certains croisements de variables ne prenant pas en compte le salaire.

Comme vu précédemment, les valeurs de certaines variables sont réparties de façon très inégale. La plus flagrante est le nombre de femmes interrogées comparé au nombre d'hommes. Cela se remarque très bien en croisant cette comparaison avec la répartition entre les différents départements. Nous avons estimé que l'écart général entre les deux départements n'était pas suffisant pour impacter de manière significative notre analyse. Mais nous remarquons une forte différence entre la représentation des sexes qui, elle, risque d'être fortement impactante dans nos résultats.

Nous pouvons également nous intéresser aux postes occupés par les employés selon leur sexe. Ainsi, nous constatons toujours une grande différence dans l'effectif entre le nombre de femmes et d'hommes. Nous visualisons aussi la disparité des effectifs selon les postes, comme mentionné plus tôt, avec une majorité de professeurs.

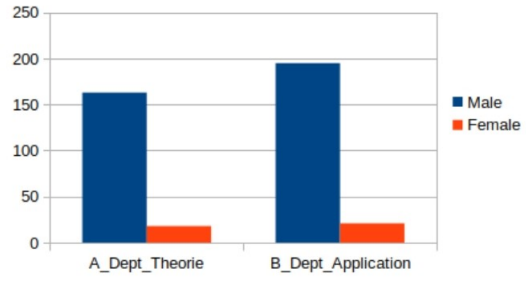

Figure 3: Répartition homme/femme par département d'étude

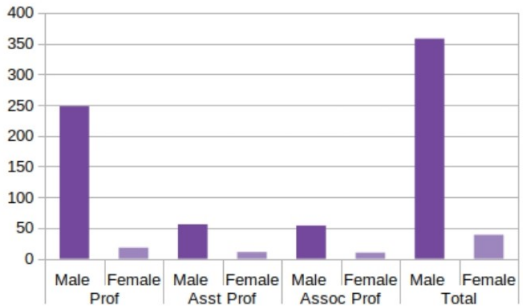

Figure 4: Répartition de l'effectif

Concentrons-nous maintenant sur la valeur du salaire par rapport à d'autres variables, en commençant par le département d'étude. Nous constatons que le salaire moyen des employés du département des applications est supérieur au salaire moyen des employés du département théorique – environ $10 000.

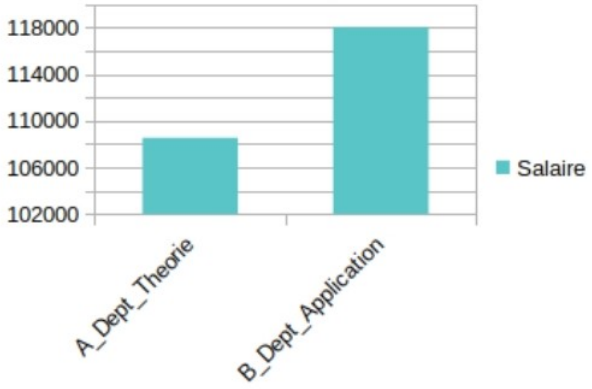

Figure 5: Salaire moyen par département d'étude

Nous allons à présent regarder les salaires moyens selon le poste et le sexe des employés. La première chose à remarquer est que le salaire des femmes est en moyenne plus faible que celui des hommes, peu importe le poste occupé. Nous pouvons aussi voir que les professeurs gagnent en moyenne plus que les professeurs associés, et que ces derniers gagnent en moyenne plus que les assistants professeurs. Cette différence est cependant logique au vu de la hiérarchie des postes proposée plus tôt.

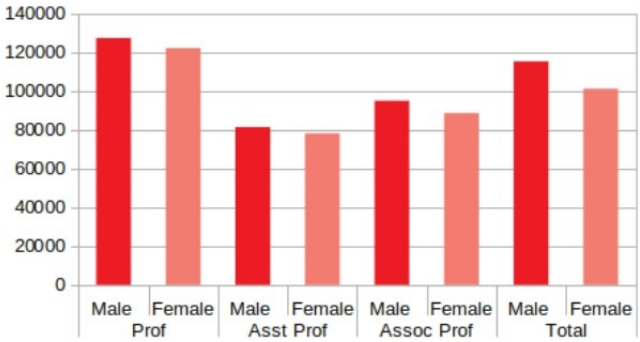

Figure 6: Répartition des salaires homme/femme

2.4 Premières conclusions et hypothèses:

Nous pouvons tirer de ces premières analyses plusieurs hypothèses, et justifier certaines de nos tentatives de modèles par nos résultats des tests de corrélation et du Chi-2, examinant la dépendance de nos variables les unes aux autres.

L'une de ces hypothèses serait d'établir un lien entre le salaire de l'employé et son sexe. Nous avons en effet remarqué à plusieurs reprises que les salaires moyens étaient plus faibles pour les femmes. En contradiction avec certaines de ces observations, le test du Chi-2 semble cependant montrer qu'il n'y a pas de corrélation significative entre le salaire et le sexe. Il serait donc intéressant

d'étudier un modèle dépendant de ces deux variables pour en estimer sa fiabilité. Nous nous devons cependant de rappeler que les femmes sont faiblement représentées dans ce jeu de données, et donc que l'étude de cette hypothèse peut s'avérer peu concluante.

Un modèle dépendant du poste occupé semble quant à lui pertinent. Il est plutôt naturel de penser qu'un poste mieux gradé rapportera plus que les autres. De plus, nos tests du Chi-2 semblent en effet démontrer qu'il y a un rapport existant entre la valeur du salaire et le poste. Il serait intéressant de vérifier cette hypothèse par nos modèles de régressions. Au vu de la matrice de corrélation des variables quantitatives, une troisième idée intéressante serait d'expliquer le salaire par le nombre d'années depuis l'obtention du doctorat. De façon plutôt contre-intuitive, les tests semblent indiquer que l'obtention du doctorat influe plus fortement sur le salaire que l'ancienneté. Nous allons donc, dans la partie suivante, tenter de croiser ces données pour chercher les meilleurs modèles de régression linéaire possibles.

3 Régressions:

3.1 Définitions et objectifs:

La régression est une méthode d'analyse permettant d'estimer une variable, appelée variable expliquée, par une ou plusieurs autres variables qui lui sont corrélées, appelées variables explicatives. Nous allons ici nous concentrer sur des modèles de régression linéaire, qui vont ainsi estimer notre variable expliquée par une relation linéaire de nos variables explicatives. Nous obtiendrons donc, pour chacune de nos valeurs explicatives, un coefficient représentant son poids dans l'explication de la variable estimée, ainsi qu'une valeur représentant sa pertinence. Cette donnée est la ***p-value*** de T. Elle représente la probabilité de ne pas rejeter l'hypothèse que le coefficient en question soit nul [3]. Elle fournit une appréciation de l'apport du coefficient et de la significativité de la variable dans le modèle. Nous allons, dans la suite de notre analyse, considérer que le coefficient est intéressant dans le modèle si sa p-value est inférieure à une marge valant 0.1%.

Nous allons commencer par présenter des modèles de régression linéaire simple, c'est-à-dire dépendant d'une seule variable explicative, puis nous allons chercher par la suite des modèles de régression multiple, impliquant plusieurs variables explicatives. Parmi toutes les régressions que nous aurons modélisées, nous choisirons selon certains critères la ou lesquelles seront les plus représentatives de la réalité.

Notre objectif principal étant d'étudier les facteurs justifiants le salaire d'un employé, cette dernière sera notre variable expliquée, et nous considérerons toutes les autres comme des variables explicatives.

3.2 Modèles simples:

Au vu des différents tests de corrélations et de dépendance, ainsi que de nos hypothèses, un modèle de régression simple intéressant semble être la représentation du salaire selon le poste occupé par l'employé. Via le logiciel SAS, nous obtenons le modèle de régression suivant:

[3] elle est réalisée sous un test de Student

Paramètre	Estimation		Erreur type	Valeur du test t	Pr > |t|
Constante	126772.1090	B	1449.073400	87.48	<.0001
rank AssocPro	-32895.6715	B	3290.465693	-10.00	<.0001
rank AsstProf	-45996.1239	B	3230.539828	-14.24	<.0001
rank Prof	0.0000	B	.	.	.

Figure 7: Modèle de régression linéaire reliant le salaire et le poste

La colonne comprenant la lettre B indique que plusieurs modèles sont possibles. Par défaut lors d'estimation par des variables qualitatives, SAS définit la moyenne la plus haute comme référence. Le logiciel a choisi dans notre cas le poste de professeur. Ainsi, la constante choisie par le modèle de régression représente la moyenne de salaire pour le poste de professeur. Les coefficients associés aux autres postes représentent l'estimation de la moyenne de la différence entre leurs salaires et ceux des professeurs. Comme observé précédemment, nous voyons que les assistants professeurs semblent gagner moins que les professeurs associés, qui eux même gagnent moins que les professeurs.

Une autre variable semblant significativement corrélée avec le salaire est le nombre d'années depuis la validation de la thèse. La procédure GLM nous propose donc d'estimer le salaire via le modèle de régression suivant:

Paramètre	Estimation	Erreur type	Valeur du test t	Pr > |t|
Constante	91718.68545	2765.792261	33.16	<.0001
YsincePhD	985.34212	107.365126	9.18	<.0001

Figure 8: Modèle de régression linéaire reliant le salaire et le nombre d'années depuis la validation de thèse

La constante ici représente le salaire moyen d'un employé venant de valider son doctorat, c'est-à-dire qui a validé sa thèse depuis 0 année. Le coefficient associé à notre variable représente la pente de notre droite de régression. Le coefficient étant positif, nous comprenons bien que le salaire augmente au fil des années, mais cet élément n'est pas suffisant pour définir le modèle comme fiable.

Les p-values de chacun des tests de nullité des coefficients sont sur ces deux modèles inférieures à notre seuil. Nous pouvons donc choisir de rejeter cette hypothèse, et en conclure que ces coefficients semblent expliquer de manière plutôt fiable le salaire. Cette conclusion ne permet cependant pas de déterminer si le modèle en lui-même est intéressant ou non.

Nous n'allons pas expliciter l'ensemble des modèles de régression simple que nous avons visualisés, nous nous contentons ici de détailler les modèles qui semblent les plus pertinents au vu de nos premiers tests. Notre objectif étant d'expliquer le salaire d'un employé en fonction des autres variables, nous avons soumis à SAS tous les modèles simples possibles. Nous avons ainsi obtenu cinq modèles, dont les détails des résultats seront explicités plus tard, lors d'une analyse plus globale. C'est cette analyse plus détaillée qui nous permettra de décider si nos hypothèses de dépendance sont plausibles ou non.

3.3 Modèles multiples:

Un modèle linéaire simple n'est cependant peut-être pas suffisant pour bien expliquer le salaire d'un employé. Nous allons donc également nous intéresser à différents modèles de régression multiple, principalement des modèles linéaires et des modèles croisant plusieurs variables entre elles. Nous ne présenterons ici que certains modèles impliquant deux variables, mais nos tests se porteront également sur des modèles plus complexes.

Nous pouvons commencer par un modèle regroupant les variables d'apparence les plus corrélées avec le salaire, à savoir ici le poste de l'employé et le nombre d'années depuis l'obtention du doctorat. Le modèle proposé est donc le suivant:

Paramètre	Estimation		Erreur type	Valeur du test t	Pr > |t|
Constante	129046.6512	B	3940.518926	32.75	<.0001
rank AssocPro	-33928.2402	B	3689.269671	-9.20	<.0001
rank AsstProf	-47860.4174	B	4412.634934	-10.85	<.0001
rank Prof	0.0000	B	.	.	.
YsincePhD	-80.3704		129.465073	-0.62	0.5351

Figure 9: Modèle de régression linéaire expliquant salaire par le poste et le nombre d'années depuis la validation de thèse

De la même façon que pour les modèles simples, la lettre B indique que plusieurs coefficients sont possibles pour nos variables qualitatives. SAS a ici fait le choix de considérer comme référence la moyenne du salaire des professeurs pour la plus petite valeur possible du nombre d'années depuis la validation de la thèse (ici 11 ans pour les professeurs). Les coefficients correspondent donc aux différences en moyenne par rapport à la constante choisie. Pour chaque rang, il représente la moyenne de la différence entre la constante et le salaire moyen d'un employé de ce rang ayant validé son doctorat depuis 11 ans. Pour la variable quantitative, le coefficient représente la pente de nos courbes de régression. Étonnamment, ce coefficient est négatif, ce qui implique que le salaire moyen baisse si le nombre d'années depuis la thèse augmente.

Nous pouvons cependant constater que sa p-value est bien supérieure au seuil que nous avons prédéfini, et donc que ce coefficient n'est pas vraiment pertinent dans notre modèle.

Nous pouvons également nous intéresser à un modèle tentant d'expliquer le salaire selon le poste et le sexe d'un employé. La procédure SAS nous fournit le modèle suivant:

Paramètre	Estimation		Erreur type	Valeur du test t	Pr > |t|
Constante	127106.5944	B	1473.546638	86.26	<.0001
sex Female	-4942.9511	B	4026.126994	-1.23	0.2203
sex Male	0.0000	B	.	.	.
rank AssocPro	-32457.8208	B	3307.632440	-9.81	<.0001
rank AsstProf	-45519.0801	B	3251.760657	-14.00	<.0001
rank Prof	0.0000	B	.	.	.

Figure 10: Modèle de régression linéaire expliquant le salaire par le poste et le sexe

Encore une fois, SAS nous dit que plusieurs modèles sont possibles, et en choisit un par défaut. La constante représente la moyenne de salaire d'un professeur homme. Le coefficient rattaché au sexe "femme" représente la moyenne de différence entre le salaire moyen d'un homme professeur et d'une femme professeur. Les coefficients rattachés aux postes représentent la moyenne de la différence de salaire entre les hommes professeurs, assistants professeurs et professeurs associés.

Nous remarquons encore que la p-value du coefficient du sexe "femme" est largement supérieure à notre seuil. Ainsi, pour ce modèle, nous pouvons estimer que le sexe n'est pas une variables significative dans notre modèle.

3.4 Analyses et choix des modèles:

Pour analyser nos modèles de régression, nous devons étudier certaines données présentées par SAS suite à l'utilisation de la procédure GML, que nous n'avons pas détaillées précédemment. Les principales valeurs que nous allons regarder seront représentées sous forme de tableau, nous permettant ainsi de comparer chaque modèle proposé en fonction de ces données, pour ensuite choisir celui que nous estimerons le meilleur.

La première valeur que nous regardons est la ***p-value*** de F. Cette valeur représente la probabilité de ne pas rejeter l'hypothèse que tous les coefficients du modèle soient nuls – réalisée sous un test de Fisher – c'est-à-dire qu'on rejette cette hypothèse si au moins un coefficient est significativement différent de 0. Elle fournit ainsi une appréciation du modèle dans son ensemble. Nous allons ici considérer que le modèle est intéressant si cette p-value est inférieure à une marge valant 0.1%.

La deuxième valeur à observer est le ***R-carré***, qui mesure la proportion de variation de la vraie valeur de la variable expliquée autour de la moyenne donnée par le modèle proposé. Tout comme la p-value de F, cette valeur permet de donner une idée de la cohérence du modèle. Nous cherchons à obtenir une valeur proche de 1, cependant cela ne suffira pas pour départager nos modèles.

La troisième valeur que nous allons regarder est la ***Racine MSE***. Cette valeur représente l'erreur quadratique moyenne du modèle proposé aux valeurs du jeu de données. Nous préférerons généralement une Racine MSE faible.

Enfin, la dernière valeur importante pour notre analyse est le ***Coefficient de variation***. C'est un indicateur de comparaison de l'écart-moyen résiduel du modèle à la moyenne de la variable expliquée. Nous l'utilisons pour comparer directement les modèles entre eux. Nous préférons un Coefficient de Variation bas, indiquant que les données ont tendance à bien suivre le modèle proposé.

3.4.1 Modèles simples:

Nous allons tout d'abord comparer nos modèles de régression simples. Le tableau suivant résume les valeurs importantes de chaque modèle pour nous permettre de les comparer plus simplement:

Modèles considérés	*p-value de F*	*R-carré*	*Racine MSE*	*Coeff. de Var.*
rank	< 0,000 1	0,394 251	23 633,67	20,784 81
sex	0,005 7	0,019 213	30 034,61	26,414 16
YsincePhD	< 0,000 1	0,175 755	27 533,59	24,214 62
YService	< 0,000 1	0,112 054	28 577,74	25,132 90
discipline	0,001 8	0,024 362	29 955,66	26,344 73

En regardant les p-value des tests de chacun des modèles, nous remarquons que deux modèles dépassent le seuil fixé précédemment. Il y a donc une possibilité non négligeable que tous les coefficients de ces modèles soient nuls. Nous pouvons donc considérer que le sexe et la discipline ne permettent pas d'exprimer convenablement le salaire par un modèle de régression linéaire. Pour les trois modèles restants, nous allons comparer les autres valeurs pour établir notre préférence.

Nos valeurs pour le R-carré sont plutôt éloignées de 1; cependant nous pouvons constater que les modèles ayant un R-carré le plus éloigné de 1 sont les modèles considérés comme peu intéressants grâce à la p-value. Ainsi nous pouvons tenter un premier classement de nos modèles de régression : le modèle dépendant du rang de l'employé semble être le plus correct, suivi de celui dépendant du nombre d'années depuis l'obtention du doctorat, et en dernier le modèle dépendant du nombre d'années de service.

Nos Racines MSE appuient cette idée. Les valeurs les plus faibles sont en effet sur les modèles classés précédemment.

Les valeurs des coefficients de variation de nos modèles semblent soutenir également notre premier classement. En effet, le modèle ayant le coefficient le plus faible est le modèle dépendant du rang de l'employé. De la même manière, les modèles dépendant du sexe et de la discipline ont un coefficient plus élevé que les autres, ce qui conforte notre décision de ne pas les considérer comme des variables explicatives intéressantes.

Au vu de notre analyse, il semble que le meilleur moyen de prédire le salaire de manière linéaire et avec une seule variable explicative est d'utiliser un modèle dépendant du rang de l'employé. Le modèle de régression proposé par SAS est celui que suivent les données le plus fidèlement, selon les critères que nous avons considérés. Notre étude nous montre également que, en opposition avec l'une de nos hypothèses, le sexe n'a qu'une influence minime et apparement non significative sur le salaire d'un employé. C'est en effet le modèle dépendant du sexe qui possède le R-carré et le Coefficient de Variation le plus élevé de tous les modèles, faisant de lui le moins représentatif de la réalité.

3.4.2 Modèles multiples:

Nous allons maintenant étudier les valeurs pour nos modèles de régression multiple. Nous allons résumer les valeurs des régressions multiples que nous avons modélisées pour certains modèles, pour pouvoir les analyser de la même manière que pour les régressions simples. Nous allons considérer les modèles où les variables sont prises séparément, et les modèles où les données sont croisées, nous proposant ainsi plus de coefficients.

Modèles considérés (séparés et croisés)	*p-value de F*	*R-carré*	*Racine MSE*	*Coeff. de Var.*
rank discipline rank*discipline	< 0,000 1 < 0,000 1	0,444 980 0,446 250	22 651,19 22 683,06	19,920 76 19,948 79
sex rank sex*rank	< 0,000 1 < 0,000 1	0,396 566 0,396 686	23 618,47 23 676,45	20,771 44 20,822 43
sex YsincePhD sex*YsincePhD	< 0,000 1 < 0,000 1	0,181 698 0,176 673	27 468,93 27 553,15	24,157 76 24,231 82
rank YsincePhD rank*YsincePhD	< 0,000 1 < 0,000 1	0,394 845 0,343 666	23 652,13 24 631,98	20,801 04 21,662 78
rank YService rank*YService	< 0,000 1 < 0,000 1	0,397 154 0,283 029	23 606,95 25 744,69	20,761 31 22,641 36
YsincePhD discipline YsincePhD*discipline	< 0,000 1 < 0,000 1	0,240 077 0,242 221	26 470,97 26 433,59	23,280 09 23,247 22

Nous nous rendons vite compte que la p-value de F ne nous permettra pas d'estimer directement qu'un modèle est plus cohérent qu'un autre. Il va nous falloir étudier le reste des valeurs pour choisir un modèle. Nous pouvons également remarquer que la plupart des résultats sont proches entre les modèles multiples linéaires et les modèles croisés. Une observation plus poussée nous permettra de définir si un modèle croisé est plus intéressant qu'un modèle linéaire.

L'observation des R-carré nous montre de manière assez flagrante que les modèles dépendants du poste de l'employé semblent les plus intéressants. En effet, les modèles linéaires rapprochant le poste et la discipline, et le poste et le nombre d'années de service ont un meilleur R-carré. Concernant les modèles croisés, nous obtenons les meilleurs R-carré en croisant le poste avec la discipline, et le poste avec le sexe.

De la même manière que pour les modèles simples, les valeurs obtenues pour les Racines MSE portent notre choix sur les mêmes modèles que l'analyse des résultats des R-carré, que ce soit pour les modèles linéaires ou les modèles croisés.

Nous allons enfin comparer les coefficients de variation des modèles proposés. Nous voyons que les modèles avec un meilleur coefficient sont en effet ceux prenant en compte le nombre d'années depuis la validation de la thèse, comme observé grâce aux précédentes valeurs.

Il semble ainsi que le meilleur moyen d'estimer le salaire par deux variables est d'utiliser le poste de l'employé et sa discipline de travail, ce qui ne faisait pas partie de nos premières hypothèses. Le deuxième meilleur modèle relie le poste et le nombre d'années de service, ce qui semble logique intuitivement. De plus, malgré une certaine corrélation entre la salaire et le nombre d'années depuis la validation de la thèse, les résultats obtenus pour les modèles prenant en compte cette variable nous montrent que de tels modèles de régression ne seraient pas de si bons estimateurs de la réalité. En croisant les variables entre elles, le meilleur modèle reste inchangé. Cependant le deuxième meilleur modèle devient alors le modèle croisant le sexe et le poste de l'employé. Ces résultats n'étaient pas forcément les plus évidents, puisque nos meilleurs modèles de régressions parmi ceux proposés impliquent chaque fois une variable peu corrélée au salaire.

Pour nous permettre de traiter plus de possibilités, nous avons fourni à une procédure SAS un grand nombre de modèles. Cela nous permet d'étudier un large panel de modèles différents plus rapidement, puisque le logiciel effectue directement les comparaisons et nous renvoie ce qu'il estime être le meilleur modèle. Nous avons ainsi testé une vingtaine de modèle pour obtenir le résultat suivant:

Paramètre	DDL	Estimation	Erreur type	Valeur du test t
Intercept	1	133394	1952.246828	68.33
discipline*rank A AssocPro	1	-50333	4858.038921	-10.36
discipline*rank A AsstProf	1	-59458	5024.902821	-11.83
discipline*rank A Prof	1	-13445	2781.889635	-4.83
discipline*rank B AssocPro	1	-32117	4165.488877	-7.71
discipline*rank B AsstProf	1	-48800	3972.011518	-12.29
discipline*rank B Prof	0	0	.	.

Figure 11: Meilleur modèle de régression multiple selon SAS

Le meilleur modèle choisi par le logiciel est celui que nous avons classé deuxième d'après les différentes valeurs analysées. La meilleure manière d'expliquer le salaire des employés est donc en utilisant leur poste, et la discipline dans laquelle ils travaillent.

4 Classification:

4.1 Analyse en composantes principales:

4.1.1 Définitions et Principes:

L'analyse en composantes principales est une méthode d'analyse et de modélisation des données. Nous allons chercher à simplifier le jeu de données, en transformant les variables originales corrélées en nouvelles variables décorrélées. Ainsi, ces nouvelles variables, appelées "composantes principales", forment un nouveau repère de visualisation. L'un des objectifs d'une ACP est donc de réduire la dimension du problème en diminuant le nombre de variables, tout en maximisant l'information conservée. Pour simplifier la représentation graphique du jeu de données, nous allons tenter de passer de 3 dimensions à 2 dimensions.

L'ACP va également nous permettre de visualiser d'une autre manière les rapports de corrélation entre les variables. Pour celà, nous allons tracer le cercle de corrélation, qui va nous permettre d'analyser par la suite la projections de nos observations sur nos composantes – ou axes – principales. Nous pourrons ainsi notamment étudier s'il y a ou non regroupement des individus suivants certaines variables, et si certains individus se démarquent des autres, pour éventuellement en tirer une idée de classification.

4.1.2 Applications à nos données:

Nos variables n'ayant pas les mêmes mesures et unités, nous allons donc effectuer une ACP standard. Nous nous retrouvons avec 3 composantes principales. L'objectif étant de diminuer la dimension de notre représentation, nous allons définir la qualité de représentation des variables selon les différents axes. En partant du premier axe principal, cette qualité de représentation va nous permettre d'estimer si l'axe considéré représente bien les variables. Pour pouvoir effectuer une représentation en dimension supérieure, on somme les qualités de représentations des groupements d'axes considérés. Étant donné le classement par ordre croissant des valeurs propres au début de l'ACP – assurant que le premier axe est le "meilleur", nous allons principalement étudier les qualités du premier axe principal (1D) et du premier plan principal (2D).

Eigenvalues

	Dim.1	Dim.2	Dim.3
Variance	2.156	0.758	0.086
% of var.	71.864	25.281	2.855
Cumulative % of var.	71.864	97.145	100.000

Figure 12: Tableau des qualités de représentation des axes principaux

Nous allons définir pour notre analyse que les variables sont bien représentées par un axe, plan ou hyperplan, si la qualité de représentation du groupement est supérieure à 75%. D'après le tableau des qualités de représentation, nous dépassons le seuil défini à partir de la deuxième dimension. Nous allons donc faire notre ACP sur le premier plan principal. Maintenant que nous avons choisi nos dimensions, nous pouvons tracer notre cercle des corrélations. Grâce au logiciel RStudio, nous obtenons le cercle suivant:

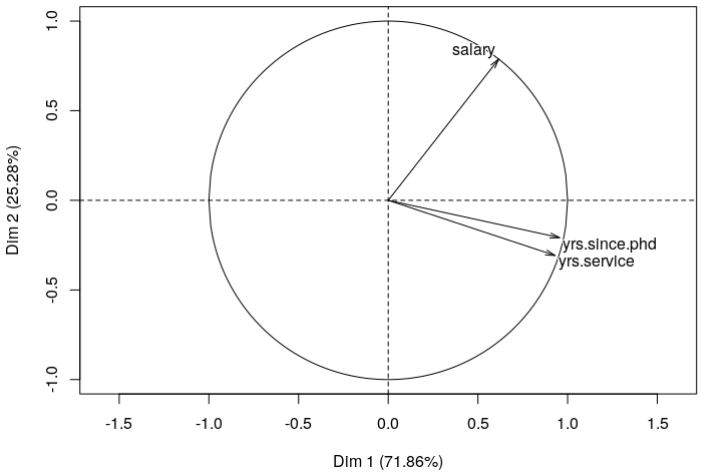

Figure 13: Cercle de corrélation

Le cercle nous permet de déduire des rapports de corrélations entre nos variables. Nous pouvons ainsi constater que le nombre d'années de service et le nombre d'années depuis la validation de la thèse sont fortement corrélés, comme nous le montrait déjà notre tableau de corrélation calculé par SAS. Le salaire cependant ne semble pas significativement corrélé aux autres variables.

En analysant nos flèches, nous pouvons interpréter notre plan et ainsi affirmer que notre premier axe (Dim 1) est très corrélé au nombre d'années de service et au nombre d'années depuis la validation de la thèse. Cela signifie que lors de la projection de nos individus sur ce plan, leur position selon cet axe sera très représentative de la valeur de ces deux variables. L'orientation de la flèche du salaire signifie que cette variable est bien représentée par le plan, mais n'est pas pas corrélée à un axe de manière évidente. Cependant elle semble plus corrélée avec le deuxième axe principal (Dim 2). Ainsi les individus placés dans le quart supérieur droit du plan auront un salaire plus élevé.

Les individus trop proches de l'origine du plan sont considérés comme étant mal représentés par la projection. Nous allons donc nous intéresser principalement aux individus éloignés du centre. Pour visualiser nos variables qualitatives sur cette projection, nous allons colorer nos individus selon leurs valeurs. Cela nous permet ainsi d'étudier les phénomènes de corrélation entre les variables qualitatives et quantitatives. Nous allons également grâce à ça déterminer s'il est possible de classer nos individus. Regardons tout d'abord une coloration selon le sexe:

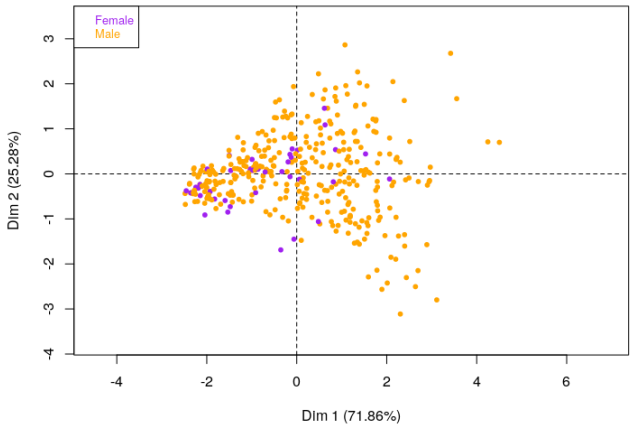

Figure 14: Projection classée selon le sexe des individus

Ce que nous voyons ici complémente nos résultats précédents selon lesquels le sexe n'a pas de réelle influence dans notre jeu de données. Nous pouvons en effet remarquer qu'il n'y a pas de délimitation précise entre les hommes et les femmes. La faible représentation des femmes peut fausser notre analyse, mais la répartition semble ici être aléatoire. Nous allons maintenant observer les colorations selon la discipline et le poste des employés.

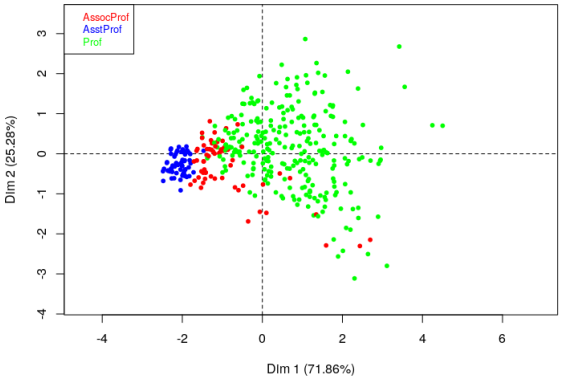

Figure 15: Projection classée selon le poste des individus

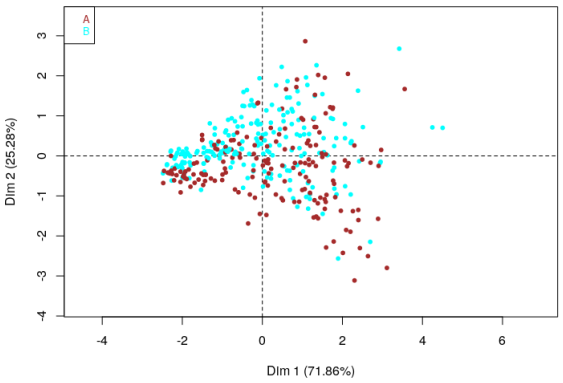

Figure 16: Projection classée selon la discipline

La répartition selon la discipline semble également être aléatoire, ce qui est en accord avec nos hypothèses et nos régressions selon lesquelles la discipline n'influence pas le salaire. Nous remarquons cependant immédiatement une démarcation très claire entre les différents postes. Il nous semble alors naturel de chercher à classer les employés selon leur poste au sein de la faculté.

4.2 Classification Ascendante Hiérarchique:

4.2.1 Définitions et principes:

La classification ascendante hiérarchique est une méthode de classification itérative. Le but de cette classification est de rassembler des individus selon une variable prédéfinie. Ainsi, nous cherchons à créer des classes dans lesquelles les individus se rapprochent fortement, mais telles que les classes se démarquent les unes des autres. La méthode que nous utiliserons ici est la méthode de Ward. L'idée de cette méthode est de considérer au départ chaque individu comme une classe à part entière. Nous allons ainsi fusionner les classes maximisant l'augmentation de la variance intra [4]. La classification étant itérative, nous répétons cette étape jusqu'à n'obtenir plus qu'une classe globale contenant l'ensemble des individus. En reliant les nouvelles classes entre elles à chaque étape, nous obtenons un arbre de regroupement. Nous pourrons ainsi, selon le nombre de classes voulues, séparer notre jeu de données en choisissant où couper les branches de notre arbre.

[4] moyenne des variances des sous-nuages

4.2.2 Application à nos données:

Pour tracer le dendrogramme de notre jeu de données, nous utilisons les librairies du logiciels RStudio. Nous obtenons ainsi l'arbre de classification suivant:

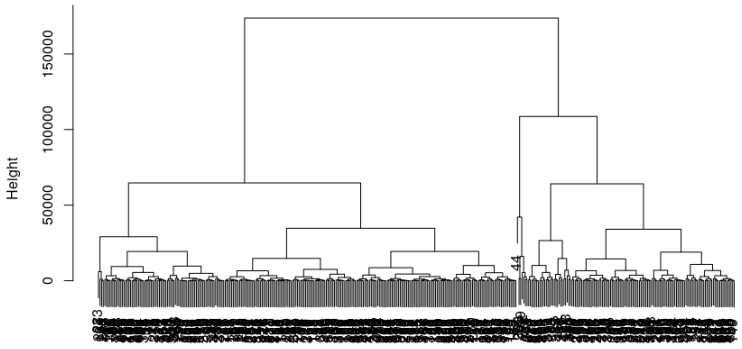

Figure 17: Dendrogramme résultat de la CAH

Étant donné le nombre d'individus, le bas de notre arbre n'est pas très lisible. De ce fait, les résultats et classes proposées par cette méthode seront difficilement interprétables.

Index des images