Rapport de méthodes

Enquête suisse sur la structure des salaires 2006

Aspects méthodologiques du modèle des salaires «Salarium»

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la statistique OFS

Neuchâtel, 2008

Statistik der Schweiz

Die vom Bundesamt für Statistik (BFS) herausgegebene Reihe «Statistik der Schweiz» gliedert sich in folgende Fachbereiche:

0	Statistische Grundlagen und Übersichten
1	Bevölkerung
2	Raum und Umwelt
3	Arbeit und Erwerb
4	Volkswirtschaft
5	Preise
6	Industrie und Dienstleistungen
7	Land- und Forstwirtschaft
8	Energie
9	Bau- und Wohnungswesen
10	Tourismus
11	Mobilität und Verkehr
12	Geld, Banken, Versicherungen
13	Soziale Sicherheit
14	Gesundheit
15	Bildung und Wissenschaft
16	Medien, Informationsgesellschaft, Sport
17	Politik
18	Öffentliche Verwaltung und Finanzen
19	Kriminalität und Strafrecht
20	Wirtschaftliche und soziale Situation der Bevölkerung
21	Nachhaltige Entwicklung und Disparitäten auf regionaler und internationaler Ebene

Statistique de la Suisse

La série «Statistique de la Suisse» publiée par l'Office fédéral de la statistique (OFS) couvre les domaines suivants:

0	Bases statistiques et produits généraux
1	Population
2	Espace et environnement
3	Vie active et rémunération du travail
4	Economie nationale
5	Prix
6	Industrie et services
7	Agriculture et sylviculture
8	Energie
9	Construction et logement
10	Tourisme
11	Mobilité et transports
12	Monnaie, banques, assurances
13	Protection sociale
14	Santé
15	Education et science
16	Médias, société de l'information, sport
17	Politique
18	Administration et finances publiques
19	Criminalité et droit pénal
20	Situation économique et sociale de la population
21	Développement durable et disparités régionales et internationales

Statistik der Schweiz

Methodenbericht

Statistique de la Suisse

Rapport de méthodes

Enquête suisse sur la structure des salaires 2006

Aspects méthodologiques du modèle des salaires «Salarium»

Auteurs **Beatriz Andrade, Monique Graf**
Office fédéral de la statistique

Editeur **Office fédéral de la statistique**

Neuchâtel, 2008

Préambule

La section LOHN de l'Office fédéral de la statistique a demandé au «Service de méthodes statistiques» (METH) de développer un modèle statistique qui permet de prédire des salaires à l'aide de l'ESS 2006. Ce rapport décrit en détails la méthode utilisée. Il décrit également la validation du modèle. Ce travail a été réalisé par Beatriz Andrade et Monique Graf, Meth.

Un grand merci à la section LOHN, en particulier à Jacques Méry et Judith Häfliger pour les fructueuses discussions tout au long de ce travail. Nous remercions aussi Markus Eichenberger de DiSo Solution AG, qui a fait l'exploitation des données et a contribué à l'écriture des programmes SAS.

Résumé

Ce rapport comporte des explications mathématiques et des présentations graphiques du modèle des salaires développé. La méthode utilisée pour faire des prédictions de salaires est présentée. Des exemples d'application pour une branche économique spécifique sont donnés à différents niveaux. Puis, un chapitre explique la qualité obtenue avec ces modèles. Pour terminer, des comparaisons entre les fonctions de répartition théoriques et empiriques sont présentées pour valider les modèles.

Mots-clé

Rapport de méthodes; LSE; procédure TRANSREG; régression; transformations; fonctions de répartition; validation

Complément d'information: Beatriz Andrade, tél. 032 713 68 19
Beatriz.Andrade@bfs.admin.ch
Monique Graf, tél. 032 713 66 15
Monique.Graf@bfs.admin.ch

Réalisation: Service de méthodes statistiques, OFS

Diffusion: Office fédéral de la statistique
CH-2010 Neuchâtel
Tél. 032 713 60 60 / Fax 032 713 60 61
Order@bfs.admin.ch

Internet: http://www.statistik.admin.ch

Numéro de commande: 338-0053

Prix: Gratuit

Série: Statistique de la Suisse

Domaine: 0 Bases statistiques et produits généraux

Langue du texte original: Français

Graphisme/Layout: OFS

ISBN: 978-3-303-00410-4

Table des matières

1 Introduction

L'objectif de ce rapport est de faire une description du modèle qui a été développé pour créer le calculateur de salaire "Salarium" sur le site de l'OFS.

Le but est de proposer un outil qui permette aux internautes de situer leur salaire par rapport au marché. L'enquête suisse sur la structure des salaires (ESS) 2006 a permis de construire un modèle prédictif du salaire pour le secteur privé qui compte environ 1.2 million de salaires. Les données du secteur public ne sont pas prises en compte. Ce modèle tient compte non seulement de la branche économique et de la taille de l'entreprise, mais aussi des caractéristiques individuelles des salariés et des postes de travail, telles que la formation, la position professionnelle, les années de service, le niveau des qualifications requises pour le poste de travail et le type d'activité exercée dans l'entreprise.

2 Contexte

2.1 Source des données

L'enquête suisse sur la structure des salaires (ESS) est réalisée depuis 1994 tous les deux ans. Les entreprises suisses ont été réparties en strates selon la branche d'activité (classes NOGA 2), la taille (en fonction du nombre d'employés : de 3 à 19, de 20 à 49 et plus de 50) et la grande région (certaines grandes régions ont été subdivisées pour tenir compte de besoins cantonaux). Dans ces strates, un tirage aléatoire simple sans remise a été effectué, puis, dans chaque entreprise, des salariés ont été sélectionnés, à nouveau selon un tirage simple sans remise. Chaque salaire est pondéré par un poids d'extrapolation dépendant des taux de tirage aux deux niveaux (strate et entreprise), du taux de réponse et du taux d'occupation.

2.2 Introduction au modèle des salaires

Le modèle développé permet de prédire le logarithme du salaire brut standardisé à l'aide des variables explicatives. Notons que l'on n'a retenu que les réponses complètes pour les variables du modèle. 15 attributs ont été considérés, parmi eux figurent la branche et la taille de l'entreprise ainsi que les attributs individuels des employés et du poste de travail. Le logarithme du salaire (variable à expliquer) est approché par une méthode de régression avec des transformations des variables explicatives. Ces transformations seront décrites plus en détail dans le chapitre 4. Les transformations mentionnées optimisent la relation entre le logarithme du salaire et chacune des variables explicatives. Elles sont estimées en même temps que les paramètres du modèle. De chaque variable transformée est déduit un facteur d'influence sur le salaire.

2.2.1 La variable à expliquer

Salaire mensuel brut, en francs, standardisé

Les montants relevés sont convertis en salaires mensuels standardisés, c'est-à-dire qu'ils sont recalculés sur la base d'un équivalent plein temps de 4 semaines 1/3 à 40 heures de travail.

Les composantes du salaire : le salaire brut du mois d'octobre (y c. les cotisations sociales à la charge du salarié pour les assurances sociales, les prestations en nature, les versements réguliers de primes, de participations au chiffre d'affaires et de commissions), ainsi que les allocations pour le travail en équipe et le travail le dimanche ou de nuit, un douzième du 13^{e}

salaire et un douzième des paiements spéciaux annuels. Ne sont pas prises en compte les allocations familiales et les allocations pour enfants.

Cette variable est notée $mbls$ (monatlicher Bruttolohn, standardisiert).

2.2.2 Les variables explicatives

Les variables explicatives sont des facteurs non ordonnés à plusieurs niveaux qui caractérisent la personne. Les 14 variables explicatives, quantitatives et qualitatives utilisées dans notre modèle sont les suivantes :

Niveau des qualifications requises pour le poste de travail (anforni) : variable qualitative

- 1 = Poste comportant les travaux les plus exigeants et les tâches les plus difficiles
- 2 = Poste requérant un travail indépendant et qualifié
- 3 = Poste requérant des connaissances professionnelles spécialisées
- 4 = Poste comportant des activités simples et répétitives

Age du salarié (alter) : variable quantitative

Formation du salarié (ausbild) : variable qualitative

- 1 = Haute école universitaire (UNI, EPF)
- 2 = Haute école spécialisée (HES), haute école pédagogique (HEP)
- 3 = Formation professionnelle supérieure, écoles supérieures
- 4 = Brevet d'enseignement
- 5 = Maturité
- 6 = Apprentissage complet
- 7 = Scolarité obligatoire, formation professionnelle acquise exclusivement en entreprise
- 8 = Scolarité obligatoire, sans formation professionnelle complète
- 9 = Autres formations complètes

Domaine d'activité (taetigk) : variable qualitative

- Activités proches de la production
  - 10 = Fabrication et transformation de produits
  - 11 = Activités de la construction
  - 12 = Mise en service, réglage et maintenance
  - 13 = Restauration, arts manuels
- Services
  - 20 = Définition des buts et de la stratégie de l'entreprise
  - 21 = Comptabilité, gestion du personnel
  - 22 = Secrétariat, travaux de chancellerie
  - 23 = Autres activités commerciales et administratives
  - 24 = Logistique, tâches d'état-major
  - 25 = Expertises, conseils, vente
  - 26 = Achat et vente de produits de base et d'équipement
  - 27 = Vente au détail de biens de consommation et de services
  - 28 = Recherche et développement
  - 29 = Analyse, programmation, "operating"
  - 30 = Planifier, construire, réaliser, dessiner
  - 31 = Transp. de personnes et de marchandises, communications
  - 32 = Services de sécurité, de surveillance
  - 33 = Activités médicales, sociales et dans le domaine des soins
  - 34 = Soins corporels, nettoyage des vêtements

- 35 = Nettoyage et hygiène publique
- 36 = Activités pédagogiques
- 37 = Activités de l'hôtellerie-restauration, économie domestique
- 38 = Culture, information, sport, loisirs et divertissements
- 40 = Autres activités

Sexe du salarié (geschle) : variable binaire
- 0 = Hommes
- 1 = Femmes

Taille de l'entreprise (ta3) : variable qualitative
- 1 = Nombre de salariés < 20
- 2 = 20 <= Nombre de salariés < 50
- 3 = Nombre de salariés >= 50

Grande région (gr)
- 1 = VD, VS, GE
- 2 = BE, FR, SO, NE, JU
- 3 = BS, BL, AG
- 4 = ZH
- 5 = GL, SH, AR, AI, SG, GR, TG
- 6 = LU, UR, SZ, OW, NW, ZG
- 7 = TI

Statut de séjour (natkat) : variable qualitative
- 1 = Suisses

  Etrangers
- 2 = Courte durée
- 3 = Séjours
- 4 = Etablis
- 5 = Frontaliers
- 6 = Autres

Position professionnelle (Codification ESS) (berufst) : variable qualitative
- 1 = Cadre supérieur
- 2 = Cadre inférieur
- 3 = Sans fonction de cadre

Années de service (dienstja) : variable quantitative

Taux d'occupation standardisé (ibgrs) : variable réelle

Variables sous forme binaire : présence-absence

- Paiements spéciaux pour toute l'année 2006 (sonderza)
- Salarié payé à l'heure ou par mois (bezstd / iwaz)
- 13ième salaire pour toute l'année 2006 (xiiimloh)

2.2.3 Cas niveau de qualification 2, 3, 4

Dans le cas du niveau de qualification 2, 3 et 4 le logarithme du salaire est modélisé pour chaque branche (donnée par NOGA2, avec regroupements usuellement définis dans l'ESS) séparément et indépendamment. Plus précisément, la branche est définie par la variable UNOGA du REE, qui est la branche de la majorité des employés de l'entreprise. La variable définissant la subdivision en branches est notée nog_2. Comme il y a un modèle de régression par

branche, les 14 variables expliquées sous le chapitre 2.2.2 sont alors les variables explicatives du modèle.

2.2.4 Cas niveau de qualification 1

Le niveau de qualification 1 (anforni=1) des postes comportant les travaux les plus exigeants et les tâches les plus difficiles (environ 5% des salaires de l'ESS) a été modélisé à part. En effet, les fonctions de prédiction valables pour les autres niveaux ne convenaient pas pour le niveau 1 (voir chapitre 7). L'approche par rapport au modèle pour les niveaux de qualification 2, 3 et 4 a changé, puisqu'on n'a pas assez de données par classe d'activité. On a donc élaboré des modèles basés sur des regroupements de nogas. Les regroupements ont été choisis de la manière suivante :

TAB. 1 Regroupements de nogas

Regroupement	Branches économiques (Noga)	
Groupe 1	10-40	Secteur 2 Production (sans la construction)
Groupe 2	45	Construction
Groupe 3	50-52	Commerce, réparation
Groupe 4	55	Hôtellerie et restauration
Groupe 5	60-64	Transports et communication
Groupe 6	65-67	Act. financières ; assurances
Groupe 7	70-74	Informatique ; R-D ; services fournis aux entreprises
Groupe 8	80	Enseignement
Groupe 9	85	Santé et activités sociales
Groupe 10	90-93	Autres services collectifs et personnels

La variable explicative du niveau de qualification a été enlevée puisqu'elle est maintenant fixe et la variable explicative nog_2 a été introduite dans le modèle, de façon à obtenir le même niveau de détail pour anforni=1. Le développement mathématique est le même pour les deux cas.

2.2.5 Les limites du modèle

Les limites du modèle sont les suivantes :

– Dans le modèle développé on n'a considéré que les réponses complètes pour les variables du modèle. C'est-à-dire qu'on a enlevé les cas où il y avait des valeurs manquantes.
– On n'a pas considéré d'interactions entre les variables explicatives dans le modèle.

3 Quelques rappels théoriques sur le thème de la régression et des fonctions de répartition

3.1 L'analyse de régression

L'analyse de régression est une technique qui permet de décrire la relation entre une variable dépendante Y et une ou plusieurs variables indépendantes $X = (x_1, x_2, ..., x_p)$. Dans le cas où on cherche une relation entre Y et un seul X, on parle de régression simple, lorsqu'il y a plusieurs X on parle de régression multiple.

3.1.1 La régression linéaire simple

Le cas spécial de modèles de régression concerne les modèles linéaires. On parle alors de régression linéaire simple lorsque les données sont de la forme $(y_i, x_i), i = 1, ..., n$ où n=nombre d'observations. Comme modèle on choisit

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

On suppose alors qu'il existe une correspondance linéaire entre x_i et Y_i. Les données y_i correspondent aux réalisations des variables aléatoires Y_i, les x_i ne sont pas aléatoires mais correspondent aux valeurs mesurées. On suppose que $\mathbf{E}(\epsilon) = 0$ et $\mathbf{Var}(\epsilon) = cste$. Le but de l'analyse de régression consiste à déterminer les paramètres inconnus β_0 et β_1. L'équation d'estimation devient alors :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{2}$$

Les résidus sont définis par

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i). \tag{3}$$

Dans le nuage de points de y versus x, ceci correspond à la distance verticale entre les valeurs observées y_i et la valeur prédite $\hat{y}_i$.

Pour trouver les coefficients $\hat{\beta}_0$ et $\hat{\beta}_1$, on utilise le principe de **moindres carrés**. On choisit les coefficients $\hat{\beta}_0$ et $\hat{\beta}_1$ de telle façon à ce que l'écart entre les valeurs observées (y_i) et les valeurs prédites par la droite de régression $(\hat{y})$ soit le plus faible possible.

3.1.2 La régression linéaire multiple

On considère plusieurs variables indépendantes $(X_1, X_2, ..., X_p)$ pour estimer la variable dépendante Y.

Les objectifs sont :
– obtenir une meilleure prédiction de Y que lors d'une régression linéaire simple
– créer un modèle qui permet d'expliquer la variable Y en utilisant plusieurs variables indépendantes qui la décrivent

Dans ce cas la fonction générale est :

$$Y_i = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \beta_3 x_{3_i} + ...\beta_p x_{p_i} + \epsilon_i \tag{4}$$

où $(i = 1, 2, ..., n)$.

3.2 Fonctions de répartition

Par une fonction de répartition on entend les fréquences relatives cumulées des données ordonnées. Il s'agit donc d'une fonction monotone croissante avec des sauts aux différentes valeurs d'observations. Ici nous considérons les fonctions de répartition pondérées par les poids d'extrapolation : les fréquences relatives sont les sommes cumulées des poids des données ordonnées, divisées par la somme totale des poids. La même méthode est utilisée pour estimer la médiane des salaires. On a donc

$$F(y) = P(Y \leq y).$$

4 La procédure TRANSREG [3]

Avec la procédure TRANSREG (transformation regression) de SAS, on peut estimer des modèles linéaires avec des transformations optionnelles non linéaires sur les variables telles que des splines par exemple.

Les données peuvent contenir des variables mesurées sur des échelles nominales, ordinales, intervalles et ratio. Tout mélange de ces types de variables est permis pour les variables dépendantes et indépendantes.

La procédure TRANSREG peut transformer :
- des variables nominales en transformant la valeur des modalités (scoring) afin de minimiser une erreur quadratique, soit en remplaçant chaque modalité par une variable indicatrice (binaire)
- des variables ordinales en faisant un “scoring“ monotone des catégories ordonnées afin que l’ordre soit faiblement préservé (des catégories adjacentes peuvent être fusionnées) et que l’erreur quadratique soit minimisée. Les variables ordinales peuvent aussi être transformées en rangs.
- des variables quantitatives par intervalles ou par ratio en utilisant des splines ou d’autres transformations telles que logarithme, box-cox, logit et arcsinus.

PROC TRANSREG élargit le modèle linéaire général en fournissant des transformations optimales de variables qui sont déduites itérativement de la méthode des moindres carrés (alternating least squares).

PROC TRANSREG itère jusqu’à ce qu’il y ait convergence en alternant
- la recherche par moindres carrés des paramètres du modèle étant donné les scores ou transformations de variables
- la recherche des transformations étant donné les paramètres du modèle

Les transformations utilisées dans le cadre du modèle des salaires (monotone, opscore et spline) sont des cas particuliers de la quantification optimale (optimal scaling). La méthode de la quantification optimale peut être définie comme un problème de régression par moindres carrés avec éventuellement des contraintes. La quantification optimale peut être accomplie en introduisant une matrice d’incidence qui est basée sur la quantification d’origine de la variable et du type de transformation défini pour cette variable. La quantification optimale de la variable est une combinaison linéaire des colonnes de cette matrice. Ces transformations sont décrites en détail dans la référence [4].

Rappelons qu’on appelle modalité une valeur possible d’une variable catégorielle.

Spline : L’instruction spline permet de trouver une fonction B-spline, qui est une fonction deux fois dérivable définie par morceaux par des polynômes. Dans notre cas on a utilisé des polynômes d’ordre cubique.

Monotone : L’instruction monotone permet de trouver une transformation monotone avec la restriction que l’ordre des modalités soit faiblement préservé, c’est-à-dire que deux catégories ordonnées adjacentes peuvent prendre la même valeur.

Opscore : L’instruction opscore attribue un score à chaque modalité d’une variable catégorielle. La méthode de Fisher (optimal scoring) [4] est utilisée.

Les transformations ci-dessus sont appliquées itérativement. Pour les nouvelles variables créées, l’ajustement du modèle est au moins aussi bon que pour les variables originales.

5 Modélisation du salaire brut standardisé

5.1 Modélisation du salaire brut standardisé

Le modèle développé est basé sur les salaires bruts standardisés $mbls$, c'est-à-dire qui ont été convertis en un temps de travail standard de 4 semaines 1/3 à 40 heures de travail.

Nous prenons en compte les valeurs $\log(mbls)$, où $\log$ est le logarithme naturel, comme expliqué au paragraphe 5.1.2.

5.1.1 Formulation mathématique

Un modèle additif est défini de la façon suivante :

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon \tag{5}$$

où les erreurs ϵ sont indépendantes des X_j, $\mathbf{E}(\epsilon) = 0$ et $\mathbf{Var}(\epsilon) = \sigma^2$. Les f_j sont des fonctions arbitraires à une variable pour chaque prédicteur.

5.1.2 La transformation appliquée au salaire

Comme les données du salaire sont très dispersées, on a cherché une transformation adéquate pour ce type de données. On a étudié les transformations boxcox qui sont définies de la manière suivante :

$$\varphi_0(x) = \log(x) \tag{6}$$

$$\varphi_\lambda(x) = \frac{x^\lambda - 1}{\lambda} \quad \text{si} \quad \lambda \neq 0 \tag{7}$$

Pour décider quelle transformation est la plus appropriée pour le modèle des salaires on a analysé les fonctions de répartition des résidus du salaire en utilisant d'une part un $\lambda \neq 0$ et d'autre part un $\lambda = 0$. Les fonctions de répartition se ressemblent dans les deux cas.
Si on regarde la qualité du modèle on peut constater que le $\mathbf{R^2}$ est supérieur dans le cas de la transformation avec $\lambda = 0$ (logarithmique) et le λ optimal pour une transformation boxcox est négatif.

Vu ces résultats et vu qu'une transformation logarithmique est plus simple à utiliser on a choisi celle-ci.

5.1.3 Les transformations appliquées aux variables explicatives

Les variables explicatives ont été transformées en utilisant la procédure TRANSREG, voir chapitre 4. Parmi les transformations offertes par TRANSREG on a utilisé les suivantes :

TAB. 2 Type de transformations pour le modèle anforni 2, 3 et 4

Variable	Transformation
mbls	log
anforni	monotone
alter	spline
ausbild	opscore
taetigk	opscore
geschle	opscore
ta3	opscore
gr	opscore
natkat	opscore
berufst	monotone
dienstja	monotone
ibgrs	monotone
sonderza	opscore
bezstd_iwaz	opscore
xiiimloh	opscore

Notons que pour le modèle anforni=1, la variable explicative anforni a été remplacée par la variable explicative nog_2. On a appliqué une transformation “opscore“ sur cette variable.

5.1.4 Présentation du modèle

Dans notre cas le modèle des salaires pour anforni 2, 3 et 4 qu'on a ajusté avec la procédure TRANSREG devient :

$$\log(mbls) \quad = \quad \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon \tag{8}$$

où

$$\begin{aligned}\sum_{j=1}^{p} f_j(X_j) \quad = \quad & \beta_{anforni} * tanforni + \beta_{ausbild} * tausbild + \beta_{taetigk} * ttaetigk + \beta_{geschle} * tgeschle \\ & + \beta_{gr} * tgr + \beta_{natkat} * tnatkat + \beta_{berufst} * tberufst + \beta_{dienstja} * tdienstja \\ & + \beta_{ibgrs} * tibgrs + \beta_{sonderza} * tsonderza + \beta_{xiiimloh} * xiiimloh + \beta_{ta3} * tta3 \\ & + \beta_{bezstd\_iwaz} * tbezstd\_iwaz + \beta_{alter} * talter \end{aligned} \tag{9}$$

Les erreurs ϵ sont supposées indépendantes. Ayant spécifié un poids dans TRANSREG, par défaut la variance ϵ est proportionnelle à l'inverse du poids, ce qui n'est pas très réaliste, mais n'a pas d'influence sur les estimations de variance d'échantillonnage (voir plus loin l'application de SURVEYREG).

Ce modèle permet de prédire les salaires bruts standardisés pour chaque branche économique et

- un niveau de qualification
- un âge
- une formation
- un domaine d'activité
- un sexe
- une taille de l'entreprise
- une grande région
- une nationalité
- une position professionnelle
- des années de service
- un taux d'occupation standardisé
- des paiements spéciaux
- un salarié payé à l'heure ou au mois
- un 13ième salaire fixés.

Pour le modèle anforni=1, le modèle des salaires est défini pour chaque regroupement de noga par l'équation suivante (la variable explicative du niveau de qualification (anforni) est remplacée par la variable de la branche économique (nog_2)) :

$$\log(mbls) \quad = \quad \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon \tag{10}$$

où

$$\begin{aligned}\sum_{j=1}^{p} f_j(X_j) \quad = \quad & \beta_{nog\_2} * tnog\_2 + \beta_{ausbild} * tausbild + \beta_{taetigk} * ttaetigk + \beta_{geschle} * tgeschle \\ & + \beta_{gr} * tgr + \beta_{natkat} * tnatkat + \beta_{berufst} * tberufst + \beta_{dienstja} * tdienstja \\ & + \beta_{ibgrs} * tibgrs + \beta_{sonderza} * tsonderza + \beta_{xiiimloh} * xiiimloh + \beta_{tta3} * ttta3 \\ & + \beta_{bezstd\_iwaz} * tbezstd\_iwaz + \beta_{alter} * talter \end{aligned} \tag{11}$$

5.1.5 Les poids dans le modèle

On a utilisé la pondération $\frac{gewibgrs}{\frac{1}{n}\sum gewibgrs}$ où n=nombre d'observations dans la procédure TRANSREG. Dans ce cas une somme pondérée des carrés de résidus est minimisée. L'introduction des poids n'a pas d'influence sur les degrés de liberté et le nombre d'observations, mais affecte les autres calculs.

5.2 La qualité du modèle

Le coefficient de détermination multiple $\mathbf{R^2}$ est défini de la manière suivante :

$$\mathbf{R^2} \quad = \quad \frac{SC_{reg}}{SC_{total}}, \tag{12}$$

où somme des carrés total (SC_{total}) = somme des carrés liés à l'erreur (SC_{erreur}) + somme des carrés liés à la régression (SC_{reg}).

Le $\mathbf{R^2}$ varie entre 0 et 1. Ce coefficient de détermination mesure la partie expliquée de la variable dépendante en présence de la variable indépendante.

Le coefficient de variation permet d'évaluer la qualité de l'ajustement. En ajoutant des variables explicatives au modèle le $\mathbf{R^2}$ augmente, même si les nouvelles variables explicatives sont très liées à la variable dépendante. Le calcul du coefficient de détermination ajusté permet d'une part de tenir compte de l'augmentation du nombre de variables explicatives et d'autre part de réduire SC_{erreur} par rapport à SC_{total}.

Le coefficient de détermination est défini par :

$$\mathbf{R^2_{adj}} = 1 - \frac{\frac{SC_{erreur}}{DL_{residus}}}{\frac{SC_{total}}{DL_{total}}} \tag{13}$$

où

- $DL_{residus}$ = nombre de degrés de liberté des résidus : $(DL_{residus} = n - (p + 1))$ où n = nombre d'individus, p = nombre de paramètres à estimer
- $DL_{regression}$= nombre de degrés de liberté de la régression : $DL_{regression} = p$ (nombre de variables)
- DL_{total} = nombre total de degrés de liberté : $DL_{total} = DL_{residus} + DL_{regression} = n - 1 = p + n - (p + 1))$

Les statistiques globales des deux modèles sont :

TAB. 3 Statistiques globales

	Modèle anforni 2, 3 et 4		Modèle anforni 1	
Statistique	$\mathbf{R^2}$	$\mathbf{R^2}$ ajusté	$\mathbf{R^2}$	$\mathbf{R^2}$ ajusté
Moyenne	0.6107	0.5983	0.4812	0.4542
Min	0.4383	0.4140	0.4096	0.3606
Max	0.8934	0.8858	0.5609	0.5328

En comparaison le $\mathbf{R^2}$ ajusté obtenu dans l'étude "Frauenlöhne, Männerlöhne" de l'office cantonal de la statistique de Zurich monte à 0.566 (au total) et à 0.576 (pour les hommes) et 0.443 (pour les femmes).

6 Les résultats

6.1 Les tableaux créés par le programme pour calculer le salaire prédit et les graphiques correspondants

Pour chaque noga les tableaux suivants ont été calculés :

- Les tableaux de transformation
- Les coefficients du modèle
- La distribution des résidus, par le moyen d'une liste de tous les percentiles
- Les tableaux avec la valeur la plus fréquente pour chaque variable
- Les tableaux avec l'influence sur le salaire par certaines variables clés

Ces tableaux sont expliqués en détail dans les paragraphes suivants.

Des présentations graphiques ont été faites pour les exemples suivants :

– Exemple de la transformation âge, niveau de qualification et formation
– Exemple de l'influence sur le salaire de la variable anforni

6.1.1 Les tableaux de transformation

Les tableaux de transformation sont construits pour chaque variable.

Exemples "anforni" et "ausbild" de la "noga 45" pour le modèle anforni 2, 3 et 4 et le modèle anforni=1 :

TAB. 4 Exemple de transformations de la variable anforni du modèle anforni 2, 3 et 4

anforni	Transformations anforni
2	2.0868
3	2.8986
4	4.1211

Dans le modèle anforni=1 la variable du niveau de qualification est constante donc il n'y pas de transformation pour cette variable.

TAB. 5 Exemple de transformations de la variable ausbild

ausbild	Transformations ausbild	
	modèle anforni 2, 3 et 4	modèle anforni=1
1	-1.0346	-1.2676
2	-1.0524	0.9237
3	2.6642	3.9646
4	6.6099	-0.9156
5	5.0146	3.0159
6	6.2456	5.8099
7	7.4466	6.0580
8	7.5874	6.4043
9	7.6219	3.4117

6.1.2 Graphiques des transformations âge, niveau de qualification et formation

L'exemple représenté dans la figure 1 nous montre que le salaire augmente en fonction de l'âge jusqu'à atteindre un maximum à 48 ans et diminue légèrement après cet âge.

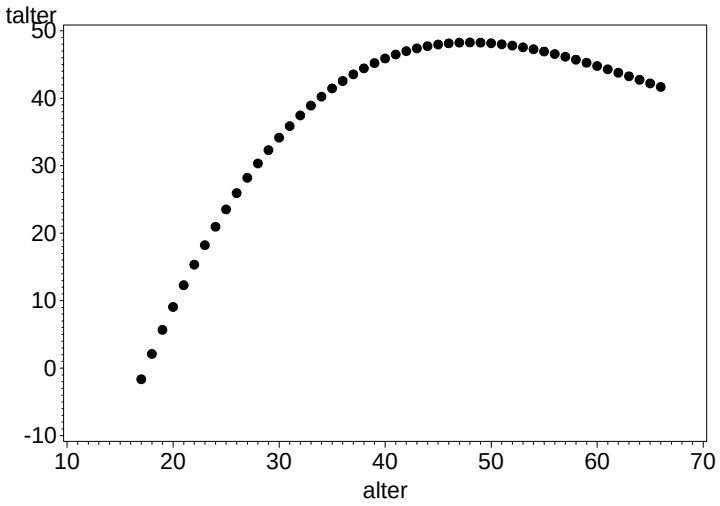

FIG. 1 Exemple de la transformation de l'âge de la noga 45 et pour le modèle des salaires anforni 2, 3 et 4

L'exemple du niveau de qualification représenté dans la figure 2 montre qu'avec le niveau de qualification 2 on gagne plus qu'avec un niveau de qualification 3 ou 4. Ceci est dû au fait que le coefficient $\beta_{anforni}$ est négatif. Donc après multiplication on obtient la figure 5, voir chapitre 6.1.8.

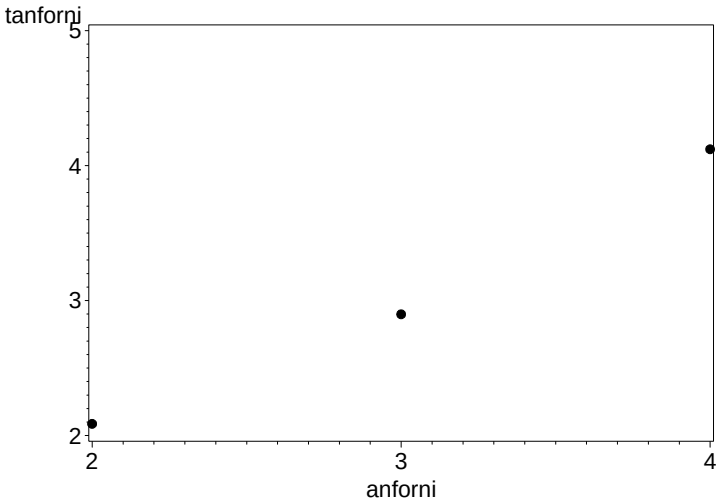

FIG. 2 Exemple de la transformation du niveau de qualification de la noga 45 et pour le modèle des salaires anforni 2, 3 et 4

L'exemple de la formation représenté dans la figure 3 montre qu'on gagne plus avec un niveau de formation correspondant à 1 (UNI, EPF) et 2 (HES, HEP) qu'avec un niveau de formation de 7, 8 et 9 (scolarité obligatoire et autres formations complètes). Ceci est dû au fait que le coefficient $\beta_{ausbild}$ est négatif. Donc après multiplication on obtient la figure 6.

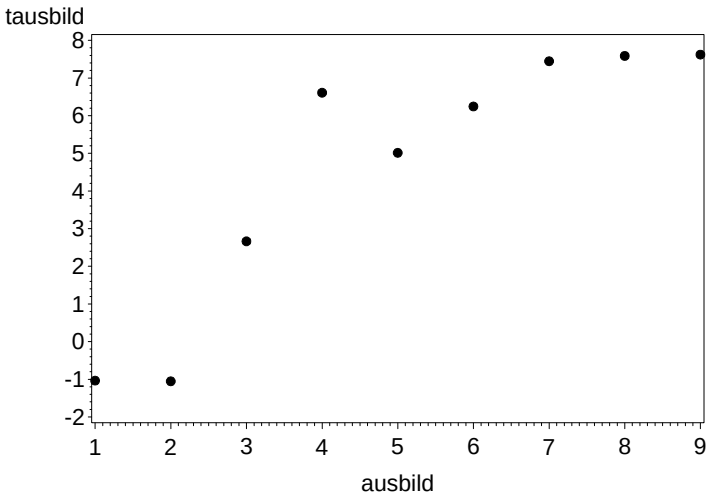

FIG. 3 Exemple de la transformation de la formation de la noga 45 et pour le modèle des salaires anforni 2, 3 et 4

6.1.3 Les coefficients du modèle et le tableau ANOVA

Le programme calcule les coefficients du modèle pour chaque variable. Une sortie pour chaque noga se présente sous la forme suivante :

Exemple “noga 45“ pour le modèle anforni 2, 3 et 4 :

TAB. 6 Tableau de coefficients

Variable	DL	Coefficient	StdErr	tVALUE	SC_{reg}	CM_{reg}	F valeur	p valeur
Intercept	1	8.8595	0.0102	870.56	20717.7	20717.7	757867	$<$.0001
Opscore(ausbild)	8	-0.0273	.	.	44.1	5.5	201.42	$<$.0001
Opscore(taetigk)	21	0.0045	.	.	22.0	1.0	38.35	$<$.0001
Opscore(geschle)	1	-0.1414	0.0031	-45.18	28.1	28.1	1026.78	$<$.0001
Opscore(ta3)	2	0.0385	.	.	47.1	23.5	861.01	$<$.0001
Opscore(gr)	6	-0.0134	.	.	26.5	4.4	161.76	$<$.0001
Opscore(natkat)	5	0.0036	.	.	1.3	0.3	9.37	$<$.0001
Opscore(sonderza)	1	0.087	0.0025	34.8	32.4	32.4	1186.75	$<$.0001
Opscore(bezst_iwaz)	1	0.035	0.0020	17.54	5.6	5.6	205.08	$<$.0001
Opscore(xiiimloh)	1	0.0777	0.0030	25.65	17.3	17.3	632.46	$<$.0001
Spline(alter)	3	0.0064	.	.	185.9	62.0	2266.77	$<$.0001
Monotone(dienstja)	54	0.0025	.	.	15.7	0.3	10.63	$<$.0001
Monotone(anforni)	2	-0.0438	.	.	27.5	13.8	503.66	$<$.0001
Monotone(berufst)	2	-0.0621	.	.	44.6	22.3	815.42	$<$.0001
Monotone(ibgrs)	162	-0.2153	.	.	27.7	0.2	6.25	$<$.0001

Remarquons que les écart-types (StdErr) ne sont calculés que si les degrés de liberté DL=1 (variable binaire ou quantitative). Dans le cas d’une variable catégorielle ou d’un spline, la signification du paramètre est donnée par un test F.

Le tableau ANOVA est donné par :

TAB. 7 Tableau ANOVA

Source	DL	SC	CM	F valeur	p valeur
Modèle	269	1081.324	4.01979	147.05	<.0001
Erreur	47621	1301.806	0.02734	.	.
Total corrigé	47890	2383.129	.	.	.

Pour le modèle anforni=1 on obtient pour la noga 45 les résultats suivants :

TAB. 8 Tableau de coefficients

Variable	DL	Coefficient	StdErr	tVALUE	SC_{reg}	CM_{reg}	F valeur	p valeur
Intercept	1	9.1486	0.0565	161.97	1924.80	1924.80	26235.40	<.0001
Opscore(nog_2)	0	0	0	.	.	.	.	.
Opscore(ausbild)	8	-0.0362	.	.	11.60	1.45	19.76	<.0001
Opscore(taetigk)	16	0.0139	.	.	17.58	1.10	14.98	<.0001
Opscore(geschle)	1	-0.1945	0.0212	-9.19	4.72	4.72	64.36	<.0001
Opscore(ta3)	2	0.0954	.	.	15.70	7.85	107.01	<.0001
Opscore(gr)	6	-0.0230	.	.	5.38	0.90	12.21	<.0001
Opscore(natkat)	5	-0.0115	.	.	0.50	0.10	1.37	0.2326
Opscore(sonderza)	1	0.2415	0.0125	19.35	25.94	25.94	353.60	<.0001
Opscore(bezstd_iwaz)	1	0.0407	0.0235	1.73	0.18	0.18	2.46	0.1169
Opscore(xiiimloh)	1	0.0692	0.0148	4.67	1.53	1.53	20.91	<.0001
Spline(alter)	3	0.0074	.	.	14.02	4.67	63.69	<.0001
Monotone(dienstja)	52	0.0017	.	.	0.78	0.01	0.20	1
Monotone(berufst)	2	-0.1039	.	.	6.35	3.17	43.25	<.0001
Monotone(ibgrs)	90	-0.6161	.	.	16.09	0.18	2.44	<.0001

Dans le cas de la noga 45 la variable nog_2 est fixée à 45 donc vaut zéro dans le modèle.

Le tableau ANOVA est donné par :

TAB. 9 Tableau ANOVA

Source	DL	SC	CM	F valeur	p valeur
Modèle	188	207.0448	1.101302	15.01	<.0001
Erreur	3120	228.9033	0.073366	.	.
Total corrigé	3308	435.9481	.	.	.

Description du tableau de coefficients :

Variable : indique le nom de la variable et la transformation appliquée

Degrés de liberté (DL) : indique le nombre de degrés de liberté de la variable

Coefficient : indique le coefficient de régression de cette variable

(StdErr) : est l'écart-type

t valeur (tVALUE) : est égale au coefficient estimé divisé par l'écart-type

Somme des carrés (SC_{reg}) : est définie au paragraphe 5.2

Carré moyen : c'est la somme des carrés divisée par le nombre de degrés de liberté

F valeur : indique que la part de la variance de la variable dépendante expliquée par le modèle est autant de fois plus importante que la variance de la variable dépendante qui reste inexpliquée.

p valeur : indique la p-valeur, c'est-à-dire la signifiance de la variable.

Dans le tableau ANOVA le carré moyen de l'erreur CM_{erreur} est calculé. Notons que $CM_{erreur} = {std_e}^2$ (voir chapitre 7.2.1). A partir de ce carré moyen on peut calculer la valeur F du tableau de coefficients qui est défini par : $\frac{CM_{reg}}{CM_{erreur}}$.

6.1.4 Comparaison avec la procédure SURVEYREG

Avec la procédure TRANSREG on n'a considéré l'information du plan d'échantillonnage que via les poids d'extrapolation et en introduisant les variables définissant les strates comme variables explicatives (la strate est définie par les catégories croisées de nog_2, région et classe de taille).

La procédure SURVEYREG permet de tenir compte du plan d'échantillonnage. On a appliqué cette procédure sur les variables transformées par TRANSREG et on a comparé les résultats obtenus avec ceux de la procédure de TRANSREG. Les analyses ont montré que les estimations des coefficients ($\hat{\beta}$) sont identiques et le tableau ANOVA reste le même, c'est-à-dire que les carrés moyens, la somme des carrés et la valeur de F ne changent pas. Par contre il y a un changement dans les écart-types des coefficients de régression puisque la formule mathématique n'est pas la même dans les deux cas. Pour plus de détails sur le plan d'échantillonnage, voir Graf (2004).

Dans le cas de TRANSREG la formule pour la matrice de covariance de $\hat{\beta}$ est donné par :

$$\hat{V} = (X'WX)^{-}s^2 \tag{14}$$

où W est la matrice diagonale avec les poids (voir définition au chapitre 5.1.5) sur la diagonale, s^2 est égal au carré moyen de l'erreur et X est la matrice des variables explicatives transformées.

La procédure SURVEYREG utilise la linéarisation de Taylor pour estimer la matrice de covariance-variance des coefficients de régression estimés. Soit h = strate, i = grappe (entreprise), r_{hik} = résidus du modèle 8 et $w_{hik} = gewibgrs$ et soit

$$r = y - X\hat{\beta} \tag{15}$$

où le (h, i, k) ième élément est r_{hik}. (h, i, k) numérote les observations. On calcul les vecteurs lignes $1 \times p$ (p = nombre variables explicatives) :

$$e_{hik} = w_{hik} r_{hik} (x_{hik,1}, x_{hik,2}, ..., x_{hik,p}) \tag{16}$$

$$e_{hi.} = \sum_{k=1}^{m_{hi}} e_{hik} \tag{17}$$

$$\bar{e}_{h..} = \frac{1}{n_h} \sum_{i=1}^{n_h} e_{hi.} \tag{18}$$

et la matrice $p \times p$:

$$G = \frac{n-1}{n-p} \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi.} - \bar{e}_{h..})'(e_{hi.} - \bar{e}_{h..}) \tag{19}$$

La matrice de covariance de $\hat{\beta}$ calculé par la procédure SURVEYREG est alors donné par :

$$\hat{V} = (X'WX)^{-} G (X'WX)^{-} \tag{20}$$

Elle ne tient donc compte que de la stratification et non du tirage dans l'entreprise.

La sortie des tableaux des coefficients de la procédure SURVEYREG pour l'exemple "noga 45" et modèle anforni 2, 3 et 4 est la suivante :

TAB. 10 Tableau de coefficients

Paramètre	Coefficient	StdErr	DenDF	t valeur	Probt
Intercept	8.8595089	0.02327586	2399	380.63	<.0001
Tausbild	-0.0272733	0.00155676	2399	-17.52	<.0001
Ttaetigk	0.0044653	0.00037871	2399	11.79	<.0001
Tgeschle	-0.1413917	0.00647384	2399	-21.84	<.0001
Tta3	0.0384717	0.00194602	2399	19.77	<.0001
Tgr	-0.0133860	0.00109297	2399	-12.25	<.0001
Tnatkat	0.0036038	0.00081462	2399	4.42	<.0001
Tsonderza	0.0870455	0.00685539	2399	12.70	<.0001
Tbezstd_iwaz	0.0349693	0.00487204	2399	7.18	<.0001
Txiiimloh	0.0776855	0.01222572	2399	6.35	<.0001
Talter	0.0063840	0.00012994	2399	49.13	<.0001
Tdienstja	0.0025068	0.00017154	2399	14.61	<.0001
Tanforni	-0.0437781	0.00260432	2399	-16.81	<.0001
Tberufst	-0.0620622	0.00344925	2399	-17.99	<.0001
Tibgrs	-0.2152584	0.01449250	2399	-14.85	<.0001

DenDF indique les degrés de liberté du dénominateur pour les tests F et les tests t. Par défaut DenDF est égal au nombre de grappes moins le nombre des strates.

Les coefficients de régression (Tableaux 6 et 10) sont les mêmes, mais en comparant les écart-types obtenus par les procédures TRANSREG et SURVEYREG on peut constater que la méthode TRANSREG est trop optimiste en comparaison avec la procédure SURVEYREG. Cependant dans ce cas, les résultats des tests de signification concordent.

Discussion

La matrice de covariance des paramètres de régression issue de TRANSREG, équation (14), est basée sur l'hypothèse que la variance des erreurs ϵ_i est proportionnelle à l'inverse du poids, alors que l'hypothèse sous-jacente à SURVEYREG est que la variance du résidu dépend du plan d'échantillonnage. Dans l'enquête ESS où les taux de sondage sont souvent très élevés, le plan a un effet important. L'équation (20) tient compte de la stratification et de la correction de population finie au niveau entreprise seulement, si bien que les résultats sont approximatifs. Les tables ANOVA sont les mêmes dans les deux cas, car SURVEYREG fait les calculs sans tenir compte du plan d'échantillonnage.

En résumé, TRANSREG est utilisé pour l'estimation des transformations des variables et les tables ANOVA. Nous pouvons utiliser TRANSREG pour estimer les paramètres du modèle,

car ils sont toujours identiques aux résultats de SURVEYREG (Tableaux 6 et 10), mais pour obtenir les variances approximatives des coefficients basées sur le plan d'échantillonnage, nous devons passer par SURVEYREG.

6.1.5 La distribution des résidus, par le moyen d'une liste de tous les percentiles

Un extrait de la sortie des quartiles 25 %, 50 % et 75 % pour la noga 45 pour le modèle anforni 2, 3 et 4 et le modèle anforni=1 est représenté ci-dessous :

TAB. 11 Tableau des résidus et percentiles

	Résidus	
Percentiles	modèle anforni 2, 3 et 4	modèle anforni=1
25	-0.0852	-0.1564
50	-0.0045	-0.0204
75	0.0827	0.1250

6.1.6 Les tableaux avec la valeur la plus fréquente pour chaque variable

On calcule aussi la valeur la plus fréquente de chaque variable explicative par noga pour le modèle anforni 2, 3 et 4 et par regroupement de noga pour le modèle anforni=1. On note ceci par exemple pour la formation :

$\widehat{ausbild}_{ref}$ et notant la valeur de la transformation de cette variable $\widehat{tausbild}_{ref}$.

Ceci nous permet de mettre cette valeur par défaut, si la personne qui entre son profil dans le modèle des salaires sur internet, ne répond pas à cette question. Ce calcul est fait pour les attributs qui ne sont pas obligatoires dans l'outil sur internet. Les attributs obligatoires et non obligatoires sont définis dans le tableau ci-dessous.

TAB. 12 Variables obligatoires et non obligatoires

Variables oblig.	Variables non oblig.
nog_2	ta3
anforni	ausbild
gr	alter
taetigk	dienstja
berufst	geschle
heures par semaine	natkat
	sonderza
	bezstd / iwaz
	xiiimloh

6.1.7 Les tableaux avec l'influence sur le salaire par certaines variables clés

En prenant l'exponentiel de l'équation 8 on obtient :

$$
\begin{aligned}
mbls \;=\; & \exp(\beta_0 + \beta_{anforni} * tanforni + \beta_{ausbild} * tausbild + \beta_{taetigk} * ttaetigk \\
& + \beta_{geschle} * tgeschle + \beta_{gr} * tgr + \beta_{natkat} * tnatkat + \beta_{berufst} * tberufst \\
& + \beta_{dienstja} * tdienstja + \beta_{ibgrs} * tibgrs + \beta_{sonderza} * tsonderza + \beta_{xiiimloh} * xiiimloh \\
& + \beta_{ta3} * tta3 + \beta_{bezstd\_iwaz} * tbezstd\_iwaz + \beta_{alter} * talter + \epsilon) \\
& = \exp(\beta_0) * \underbrace{\exp(\beta_{anforni} * tanforni)}_{\alpha_{anforni}} * \underbrace{\exp(\beta_{ausbild} * tausbild)}_{\alpha_{ausbild}} * \underbrace{\exp(\beta_{taetigk} * ttaetigk)}_{\alpha_{taetigk}} \\
& * \underbrace{\exp(\beta_{geschle} * tgeschle)}_{\alpha_{geschle}} * \underbrace{\exp(\beta_{gr} * tgr)}_{\alpha_{gr}} * \underbrace{\exp(\beta_{natkat} * tnatkat)}_{\alpha_{natkat}} \\
& * \underbrace{\exp(\beta_{berufst} * tberufst)}_{\alpha_{berufst}} * \underbrace{\exp(\beta_{dienstja} * tdienstja)}_{\alpha_{dienstja}} * \underbrace{\exp(\beta_{ibgrs} * tibgrs)}_{\alpha_{ibgrs}} \\
& * \underbrace{\exp(\beta_{sonderza} * tsonderza)}_{\alpha_{sonderza}} * \underbrace{\exp(\beta_{xiiimloh} * xiiimloh)}_{\alpha_{xiiimloh}} * \underbrace{\exp(\beta_{ta3} * tta3)}_{\alpha_{ta3}} \\
& * \underbrace{\exp(\beta_{bezstd\_iwaz} * tbezstd\_iwaz)}_{\alpha_{bezstd\_iwaz}} * \underbrace{\exp(\beta_{alter} * talter)}_{\alpha_{alter}} * \tilde{\epsilon}) \qquad (21)
\end{aligned}
$$

où $\tilde{\epsilon} = \exp(\epsilon)$

L'étude de l'influence des différentes variables sur le salaire peut se faire puisqu'il s'agit de facteurs multiplicatifs indépendants. [1] On peut donc l'étudier en regardant le rapport de salaires prédits dont on ne change qu'une seule variable explicative en maintenant toutes les autres variables explicatives égales. Par exemple pour le niveau de qualification ceci nous donne en développant :

$$
\frac{mbls_{anforni_m}}{mbls_{anforni_n}} = \frac{\exp(\beta_{anforni} tanforni_m)}{\exp(\beta_{anforni} tanforni_n)} \quad \text{où } m, n \in \{1, ..., 4\} \qquad (22)
$$

6.1.8 Exemple de l'influence des variables âge et niveau de qualification sur le salaire

En remplaçant dans les expressions du type 22 le dénominateur par exemple par le terme $\exp(\beta_{alter}\widehat{talter_{ref}})$ ou $\exp(\beta_{anforni}\widehat{tanforni_{ref}})$ ou encore $\exp(\beta_{ausbild}\widehat{tausbild_{ref}})$, ceci nous permet de fixer la valeur la plus fréquente à 1 dans les graphiques.

Sur l'axe des x on met la variable explicative et sur l'axe des y on met la variable explicative transformée multipliée par le coefficient correspondant et en prenant l'exponentiel en standardisant le tout par rapport à la valeur la plus fréquente. Sur l'axe des y on a par exemple pour le niveau de qualification :

$$
\frac{\exp(\beta_{anforni} tanforni_m)}{\exp(\beta_{anforni}\widehat{tanforni_{ref}})} \qquad (23)
$$

[1] Ceci est dû au fait que le modèle présenté ici ne tient pas compte d'interactions.

Dans la noga 45 et le modèle avec anforni 2, 3 et 4, la valeur la plus fréquente de l'âge monte à 42 ans, pour le niveau de qualification à 3 et pour la formation à 6.

La figure 4 montre que le salaire monte jusqu'à un âge de 48 ans et redescend ensuite.

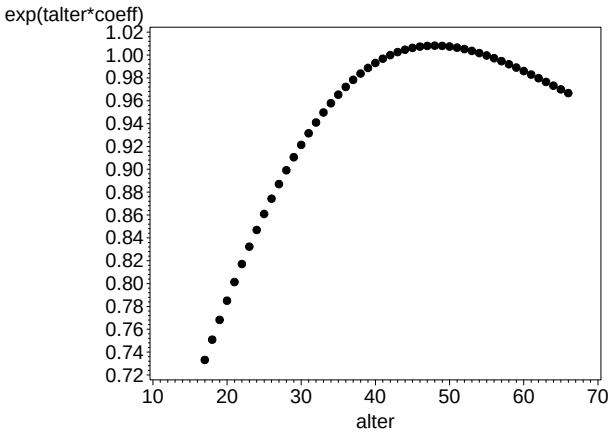

FIG. 4 Exemple effet multiplicatif de l'âge sur le salaire pour la noga 45 et pour le modèle des salaires anforni 2, 3 et 4. Age de référence : 42 ans

Dans la figure 5 on voit que le salaire descend si le niveau de qualification passe de 2 à 4. Ce n'est pas surprenant, car le niveau 2 correspond à la qualification la deuxième plus haute et le niveau 4 correspond au niveau de qualification le plus bas. Pour la variable formation on peut

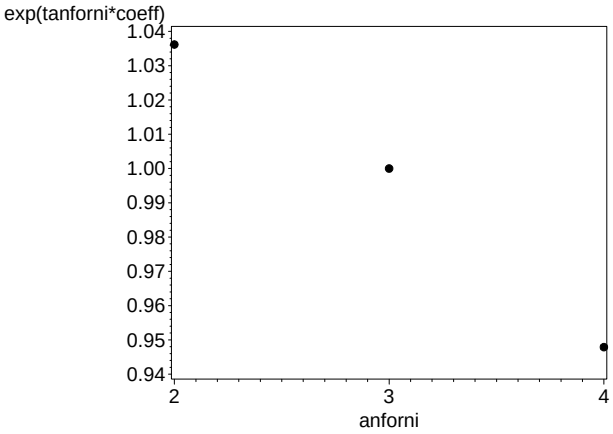

FIG. 5 Exemple effet multiplicatif du niveau de qualification sur le salaire pour la noga 45 et pour le modèle des salaires anforni 2, 3 et 4. Niveau de qualification de référence : 3

observer que le salaire est plus haut pour les niveaux 1 et 2 que pour les niveaux 7, 8 et 9 (voir figure 6).

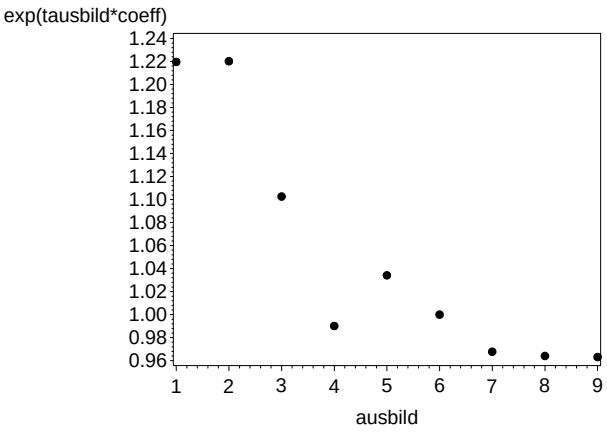

FIG. 6 Exemple effet multiplicatif de la formation sur le salaire pour la noga 45 et pour le modèle des salaires anforni 2, 3 et 4. Formation de référence : 6

Exemple du niveau de qualification

Le tableau 13 est tiré de la noga 45 et du modèle des salaires avec anforni 2, 3 et 4 et montre les changements du salaire si quelqu'un change son niveau de qualification. On voit par exemple si une personne change du niveau de qualification 2 au niveau de qualification 3 son salaire diminue de 3.5%.

TAB. 13 Tableau avec les changements d'un niveau de qualification à un autre pour le modèle anforni 2, 3 et 4 et la noga 45 (les chiffres ont été arrondis)

	"anforni" à la fin		
"anforni" au début	2	3	4
2	1	0.965	0.915
3	1.036	1	0.948
4	1.093	1.055	1

Notons que ce tableau ne dépend plus de la valeur de référence. On peut donc également l'établir pour les variables où on ne défini pas de référence.

Sur la page internet ces tableaux sont disponibles pour les attributs suivants : niveau de qualification, région, position professionnelle et le sexe. Ils représentent directement les changements du salaire en $\%$ et en franc. Pour le modèle anforni=1 le niveau de qualification est fixe et les changements de salaire sont disponibles que pour la région, la position professionnelle et le sexe.

6.1.9 Présentation d'une fonction de répartition pour un profil donné

Pour chaque variable explicative, on détermine la valeur la plus fréquente dans la classe d'activité. Le profil type (pour la classe d'activité) est défini par la combinaison de ces valeurs.

Le profil-type pour la noga 45 et pour le modèle des salaires anforni 2, 3 et 4 est donné par :

– anforni=3

– ausbild=6
– taetigk=11
– geschle=0
– ta3=3
– gr=2
– natkat=1
– alter=42
– dienstja=0
– berufst=3
– ibgrs=1
– sonderza=non
– bezst=non
– xiiimloh=oui

Quartiles pour un profil donné

Les quartiles 25%, 50% et 75% pour une combinaison choisie de variables explicatives - un profil donné - sont obtenus par la méthode suivante :
Pour un profil donné, un quartile est calculé en ajoutant au logarithme du salaire prédit par le modèle le quartile correspondant de la distribution des écarts au modèle. Le résultat est un quartile du logarithme du salaire. Il est retransformé pour aboutir à un quartile du salaire.

Pour ce faire on part de l'équation 21. On obtient pour chaque percentile $p = 1, 2, ...99$ l'équation suivante :

$$mbls_p = \exp(\hat{s} + \epsilon_p) \tag{24}$$

où

$$\begin{aligned} \hat{s} = {} & \beta_0 + \beta_{anforni} * tanforni + \beta_{ausbild} * tausbild + \beta_{taetigk} * ttaetigk \\ & + \beta_{geschle} * tgeschle + \beta_{gr} * tgr + \beta_{natkat} * tnatkat + \beta_{berufst} * tberufst \\ & + \beta_{dienstja} * tdienstja + \beta_{ibgrs} * tibgrs + \beta_{sonderza} * tsonderza \\ & + \beta_{xiiimloh} * xiiimloh + \beta_{ta3} * tta3 + \beta_{bezstd\_iwaz} * tbezstd\_iwaz \\ & + \beta_{alter} * talter \end{aligned} \tag{25}$$

Donc les équations pour les 3 quartiles du salaire sont données par :

$$mbls_{25} = \exp(\hat{s} + \epsilon_{25}) \tag{26}$$

$$mbls_{50} = \exp(\hat{s} + \epsilon_{50}) \tag{27}$$

$$mbls_{75} = \exp(\hat{s} + \epsilon_{75}) \tag{28}$$

On peut alors obtenir les trois quartiles pour ce profil-type de la noga 45 :

TAB. 14 Quartiles du profil-type de la noga 45

noga	25%	50%	75%
45	5292	5737	6259

La fonction de répartition pour ce profil-type est donnée par :

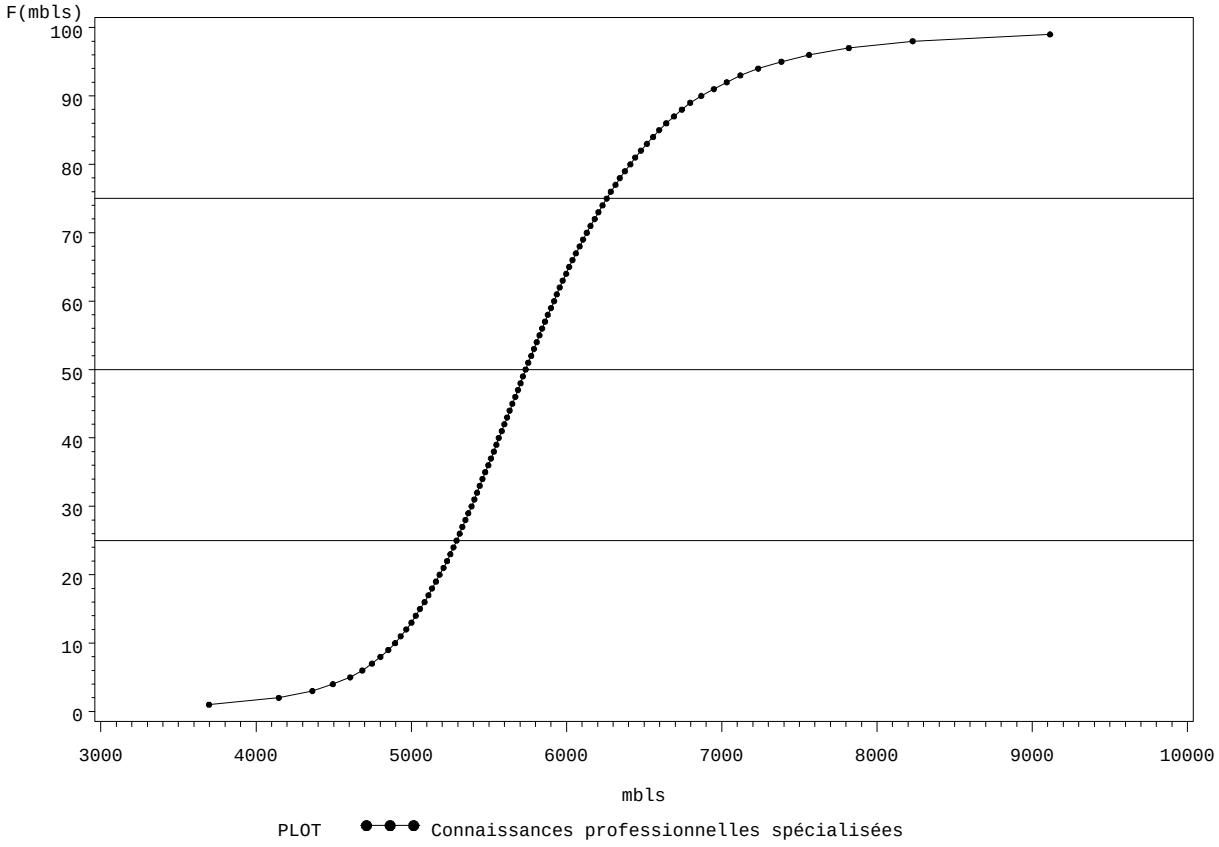

FIG. 7 Exemple de la fonction de répartition pour le profil-type de la noga 45 avec le modèle des salaires anforni 2, 3 et 4

Si on fixe toutes les variables explicatives aux valeurs du profil-type sauf le niveau de qualification on obtient trois fonctions de répartition qui correspondent aux niveaux de qualification 2, 3 et 4.

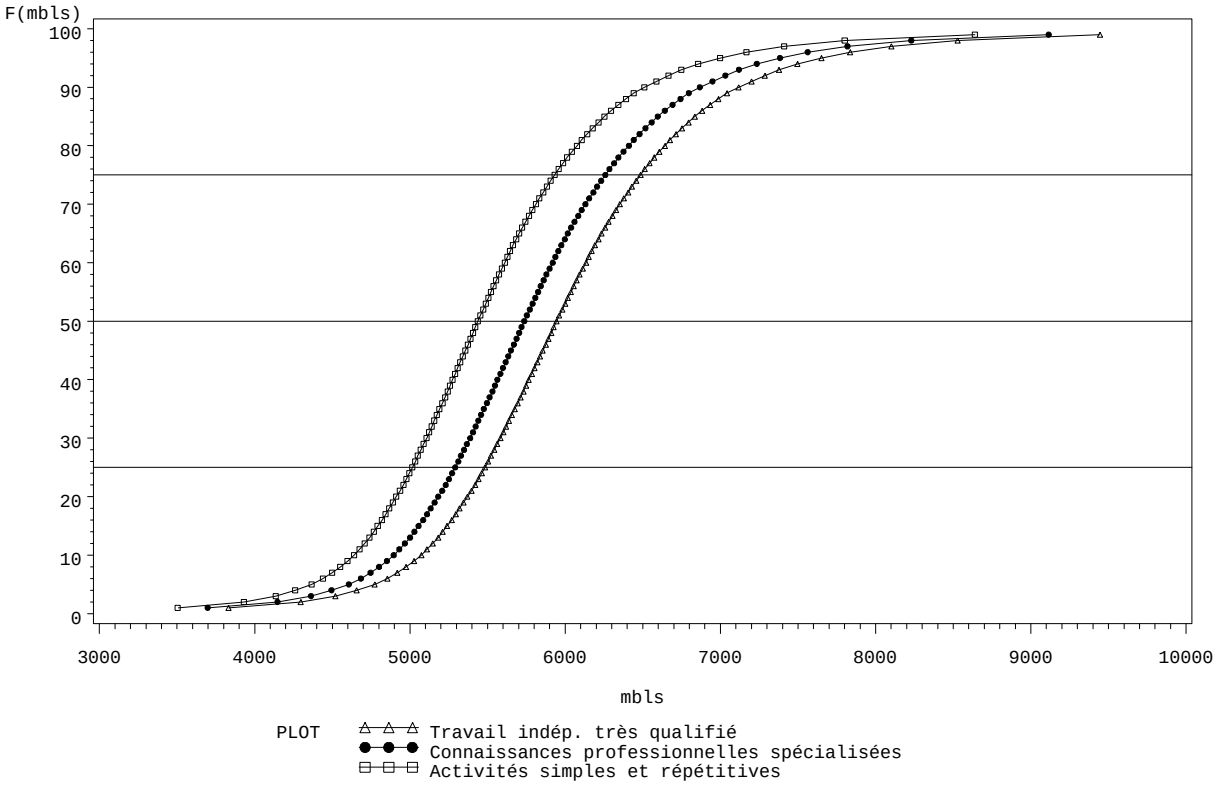

FIG. 8 Exemple de la fonction de répartition pour le profil-type de la noga 45 en variant la variable niveau de qualification avec le modèle des salaires anforni 2, 3 et 4

7 Validation du modèle par des fonctions de répartition

La validation d'un modèle statistique peut se faire de différentes façons. L'analyse des résidus permet par exemple de tester la validité d'un modèle et de détecter d'éventuels défauts. Les méthodes d'analyse de résidus sont principalement des méthodes d'analyse graphique. La validation peut aussi se faire au niveau d'agrégats pour lesquelles on a des valeurs empiriques.

Dans notre cas on a utilisé les deux méthodes mentionnées ci-dessus, c'est-à-dire :

- L'analyse des résidus
- La comparaison des fonctions de répartition pour différents niveaux d'agrégation
  - Calculée directement sur les données
  - Déduite du modèle

7.1 Analyse des résidus

On définit les résidus comme étant les différences entre les valeurs observées et les valeurs estimées par un modèle de régression. La partie non expliquée par le modèle de régression est représentée par ces résidus. Pour standardiser les résidus, on a utilisé l'écart-type calculé par la procédure TRANSREG.

On a analysé le graphique des résidus standardisés en fonction des valeurs du salaire prédit. Dans la figure 9 on peut voir que la dispersion des résidus est homogène sur tout le spectre. Pour cette raison on prend la distribution marginale comme distribution des erreurs. C'est-à-dire qu'on approche les erreurs ϵ de l'équation (8) par $\hat{\epsilon}$.

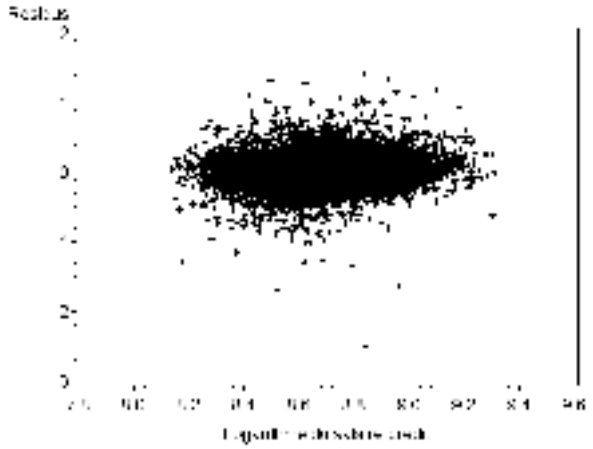

FIG. 9 Résidus en fonction du salaire prédit pour la noga 45 et le modèle anforni 2, 3 et 4

7.2 Comparaison des résultats du modèle empirique par un modèle théorique

7.2.1 Construction de la fonction de répartition au niveau individuel

Soit k un individu et $\hat{s}_k$ le salaire prédit par le modèle pour son profil. A cause de la variabilité des salaires, un salaire $y \neq \hat{s}_k$ est possible pour le même profil. L'écart à la valeur prédite est alors (à l'échelle logarithmique)

$$e_k(y) = \log(y) - \log(\hat{s}_k) \tag{29}$$

En calculant le quantile correspondant à $e_k(y)$ relativement à la fonction de répartition des résidus du modèle, on peut alors trouver la probabilité $F_k(y)$ d'observer un salaire $\leq y$ pour le profil donné. Nous posons donc :

$$F_k(y) = F_e(e_k(y)) \tag{30}$$

où F_e est la fonction de répartition des résidus du modèle pour tous les profils.
Techniquement, le calcul de $F_k(y)$ demande une interpolation de la fonction de répartition des résidus, car en général $e_k(y)$ ne correspondra pas exactement à un résidu observé. On observe que dans tous les modèles estimés la fonction de répartition des résidus standardisés est pratiquement symétrique et peut être approchée par une loi de Student (standardisée). Seul le paramètre de la loi de Student (le nombre de degrés de libertés) change d'un modèle à l'autre, voir figures 10 à 13, graphique en haut à gauche.

Approximation de la fonction de répartition des résidus à l'aide de la loi Student L'interpolation est effectuée de la manière suivante. Soit
F_e = Fonction de répartition des résidus et
F_{T_ν} = Fonction de répartition de la loi de Student à ν degrés de liberté.
On sait que la variance d'une variable aléatoire T_ν suivant une loi de Student est

$$\mathbf{Var}(T_\nu) = \frac{\nu}{\nu - 2}$$

Donc, si $X = \sqrt{\frac{\nu-2}{\nu}} T_\nu$, alors $\mathbf{Var}(X) = 1$ et

$$\mathbf{P}(X < x) = \mathbf{P}(\sqrt{\frac{\nu-2}{\nu}} T_\nu < x) = \mathbf{P}\left(T_\nu < \sqrt{\frac{\nu}{\nu-2}}\, x\right) = F_{T_\nu}\left(\sqrt{\frac{\nu}{\nu-2}}\, x\right) \tag{31}$$

En remplaçant dans l'équation (31) x par $e_k(y)/std_e$ où std_e est l'écart-type estimé des résidus, on obtient une approximation de la fonction de répartition des résidus :

$$F_e(e_k(y)) \cong F_{T_\nu}\left(\sqrt{\frac{\nu}{\nu-2}} \frac{e_k(y)}{std_e}\right) \tag{32}$$

Estimation de ν Le ν optimal a été déterminé par une optimisation non linéaire qui minimise la distance quadratique entre les fonctions de répartition. Soit $e_p = F_e^{-1}(p/100)$ le p-ième centile de la distribution des résidus. La fonction objectif à minimiser est donnée par :

$$SC_{erreur} = \sum_{p=1}^{99} \left(\frac{e_p}{std_e} - \sqrt{\frac{\nu-2}{\nu}} F_{T_\nu}^{-1}(p/100) \right)^2 \tag{33}$$

Les valeurs optimales de ν pour toutes les nogas sont données dans le tableau dans l'Annexe A.

Finalement, à partir d'une liste de salaires y prédéfinie (par exemple de 2000 à 26000), on peut calculer les probabilités $F_k(y)$ correspondantes en évaluant $F_e(e_k(y))$ dans l'équation (30) par l'approximation de Student de l'équation (32).

7.2.2 Construction de la fonction de répartition des salaires pour des domaines

Un domaine est l'ensemble de tous les profils observés dans l'échantillon possédant une certaine caractéristique (tous les hommes, tous les emplois d'une grande région, ...). Notons D ce domaine.

La fonction de répartition d'un domaine évaluée en y est donnée par la moyenne pondérée des $F_k, k \in D$ (équation 34) :

$$F_D(y) = \frac{\sum_{k \in D} w_k F_k(y)}{\sum_{k \in D} w_k} \tag{34}$$

ceci pour chaque salaire y donné et avec $w_k = gewibgrs$.

Cette moyenne pondérée considérée comme fonction de y, nous définit la fonction de répartition cumulée du domaine D.

TAB. 15 Tableau avec les probabilités $F_{D_i}(y)$

	D_1 =anforni2	D_2 =anforni3	D_3 =anforni4	salaire y
$F_D(2000)$	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	2000
$F_D(2500)$	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	2500
...	...	...	...	...

Exemple des domaines définis par les niveaux de qualification Considérons dans le modèle pour anforni 2, 3 et 4 les trois domaines définis D_1, D_2, D_3 par anforni. L'équation (34) peut alors être mise sous forme d'un tableau de dimension (nombre de salaires y fixés) x 3 : Dans le tableau 15, $f_{1,1} = \sum_{k \in D_1} w_k F_k(2000)/(\sum_{k \in D_1} w_k)$, etc. Les trois colonnes des anforni 2, 3 et 4 nous donnent pour chaque niveau de qualification (domaine) une fonction de répartition cumulée. Les fonctions de répartition cumulées sur toutes les données, par sexe ou par grande région sont obtenues par la même méthode.

7.2.3 Exemples de l'approximation de la fonction de répartition empirique par des fonctions de répartition théoriques

La validation du modèle proposée est la comparaison de ces fonctions de répartition estimées à l'aide du modèle avec les fonctions de répartition empiriques calculées directement. Un exemple des résultats pour les nogas 45 et 85 se trouvent aux figures 10 à 13.

On a analysé les domaines suivants :

- Comparaison de la fonction de répartition empirique et la fonction de répartition cumulée sur toute la noga
- Comparaison de la fonction de répartition empirique et la fonction de répartition cumulée par niveau de qualification
- Comparaison de la fonction de répartition empirique et la fonction de répartition cumulée par région
- Comparaison de la fonction de répartition empirique et la fonction de répartition cumulée par sexe

Les figures 10, 11, 12 et 13 permettent de visualiser les domaines qu'on a construit par sexe, niveau de qualification et région pour le modèle anforni 2, 3 et 4 et le modèle anforni=1. On présente ici les résultats pour les nogas 45 et 85.

Les points des fonctions de répartition théoriques correspondent aux valeurs y du salaire pour lesquelles 34 a été calculée.

Par exemple, la première image de la figure 10 indique que l'ajustement de la fonction de répartition empirique par la loi de Student avec un degré de liberté de 3.51 est très bon.

La deuxième image montre l'ajustement de la fonction de répartition empirique par la fonction de répartition cumulée sur toutes les données de la noga.

Les images suivantes montrent l'ajustement de la fonction de répartition empirique par les fonctions de répartition cumulées par niveau de qualification, région et par sexe.

En général l'ajustement des fonctions de répartition empiriques par des fonctions de répartition théoriques est très bon. L'analyse sur l'ensemble des nogas a montré que l'ajustement est en général meilleur dans le modèle anforni 2, 3 et 4 que dans le modèle anforni=1.

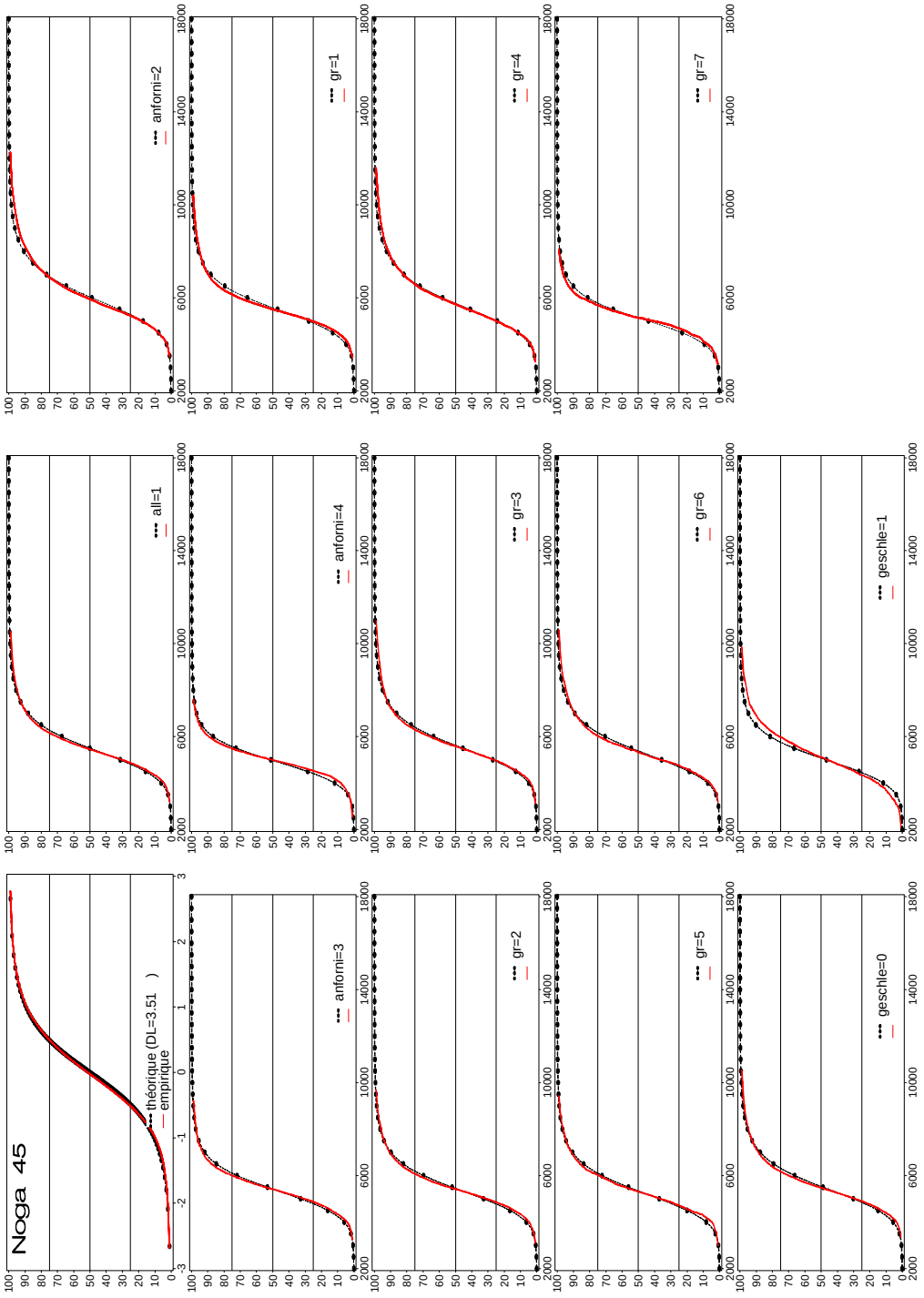

FIG. 10 Comparaison des fonctions de répartition empiriques et des fonctions de répartition cumulées de la noga 45 et le modèle avec anforni 2, 3 et 4. Trait simple : fonction de répartition empirique. Points : valeurs y où les fonctions de répartition théoriques sont calculées. Ces points sont interpolés par spline pour donner la fonction de répartition théorique.

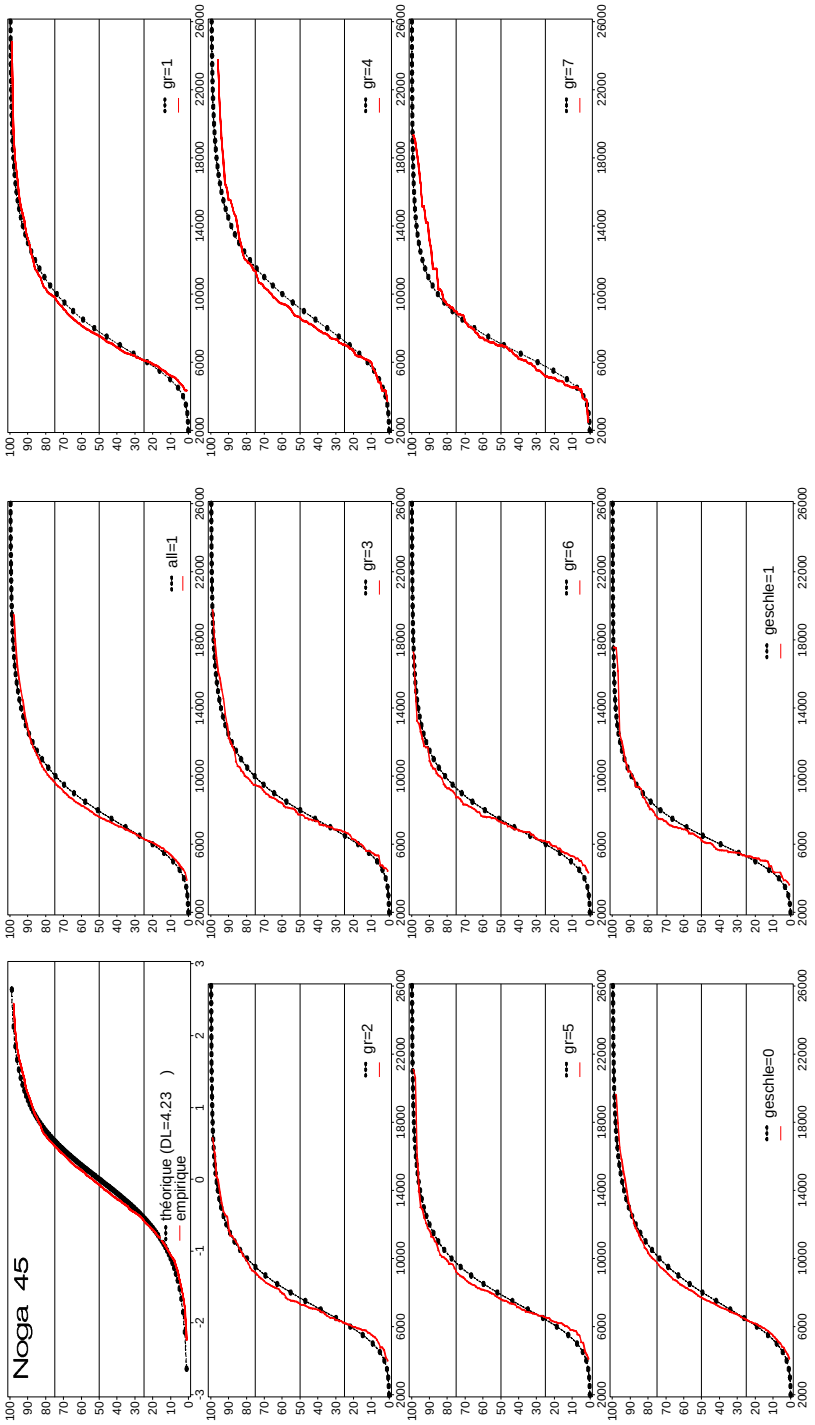

FIG. 11 Comparaison des fonctions de répartition empiriques et des fonctions de répartition cumulées de la noga 45 et le modèle avec anforni=1. Trait simple : fonction de répartition empirique. Points : valeurs y où les fonctions de répartition théoriques sont calculées. Ces points sont interpolés par spline pour donner la fonction de répartition théorique.

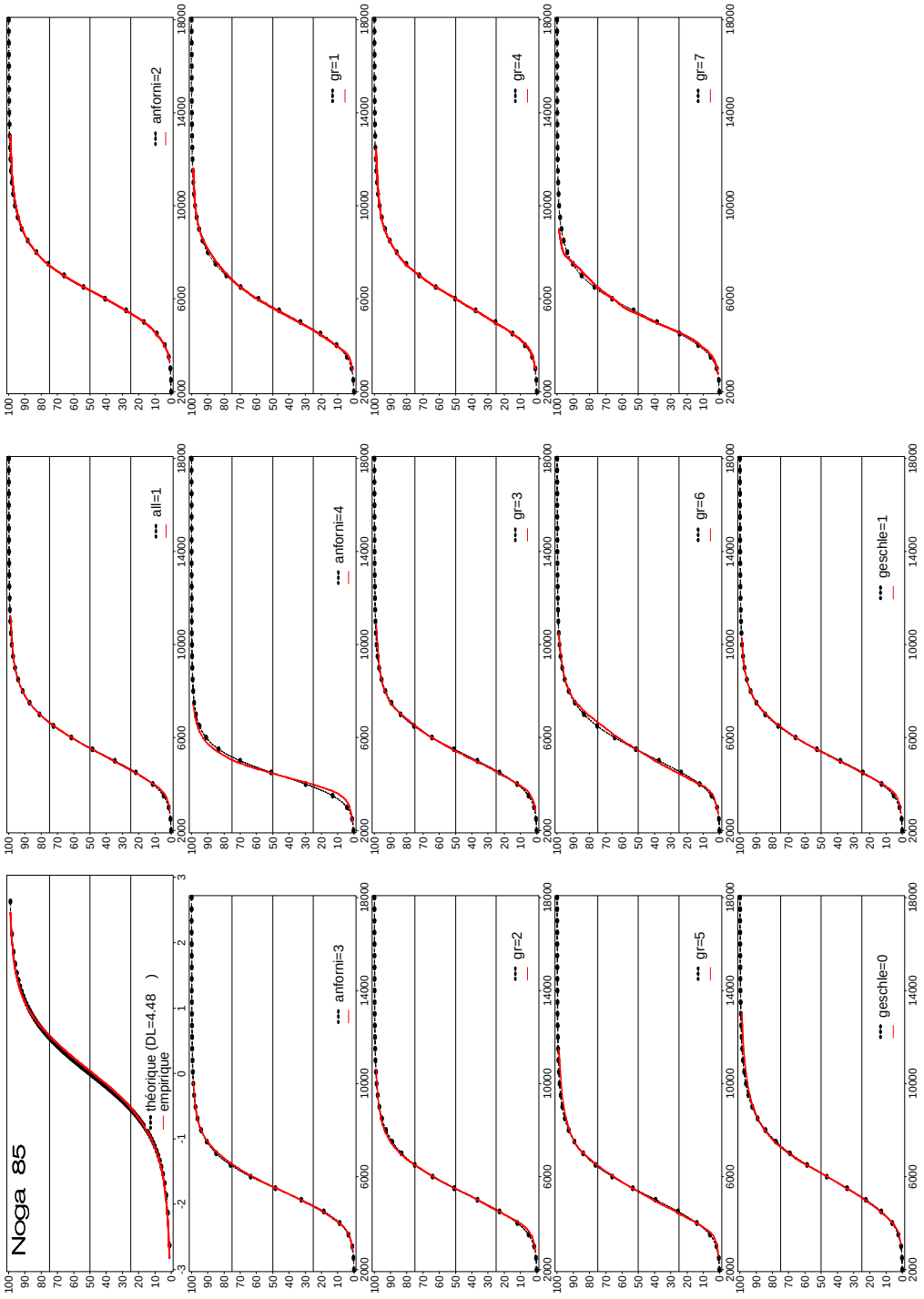

FIG. 12 Comparaison des fonctions de répartition empiriques et des fonctions de répartition cumulées de la noga 85 et le modèle avec anforni 2, 3 et 4. Trait simple : fonction de répartition empirique. Points : valeurs y où les fonctions de répartition théoriques sont calculées. Ces points sont interpolés par spline pour donner la fonction de répartition théorique.

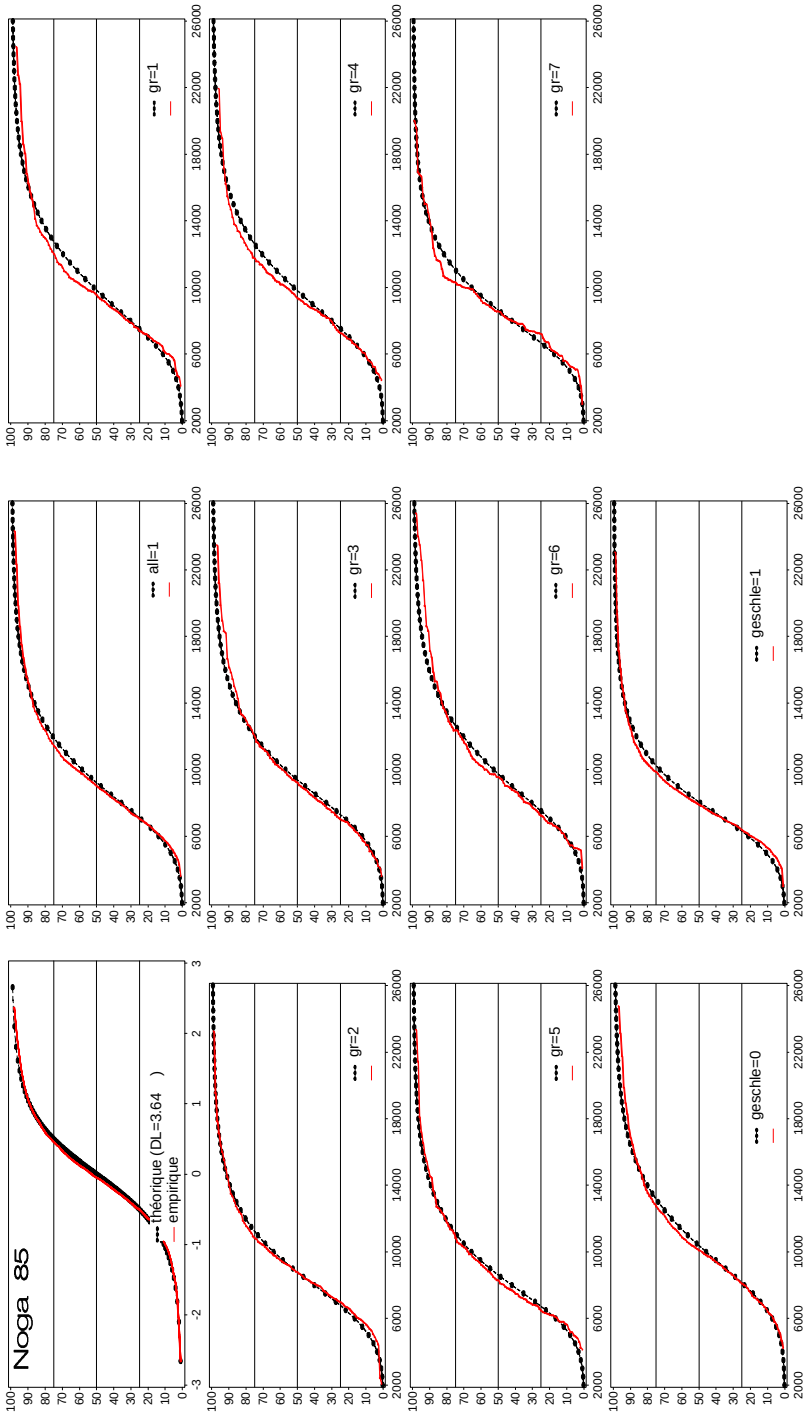

FIG. 13 Comparaison des fonctions de répartition empiriques et des fonctions de répartition cumulées de la noga 85 et le modèle avec anforni=1. Trait simple : fonction de répartition empirique. Points : valeurs y où les fonctions de répartition théoriques sont calculées. Ces points sont interpolés par spline pour donner la fonction de répartition théorique.

8 Conclusions

Le but de ce travail était de développer un outil qui permette aux personnes intéressées de situer leur salaire par rapport au marché. Pour ce faire on s'est basé sur les données de l'ESS 2006 du secteur privé. La méthode utilisée est une méthode de régression avec des transformations de variables. Elle permet de transformer les variables indépendantes de telle sorte que la corrélation entre la variable dépendante et les variables indépendantes soit optimale.

Nos analyses des modèles des salaires développés ont montré qu'il est possible de trouver une bonne prédiction des salaires avec ces modèles. Les modèles ont pu être validés par des fonctions de répartition théoriques pour les principaux domaines. Le niveau de qualification des postes comportant les travaux les plus exigeants et les tâches les plus difficiles (environ 5 % des salaires de l'ESS) a été modélisé à part. En effet, les modèles prédictifs valables pour les autres niveaux ne convenaient pas pour ce niveau.

Ce rapport contient la partie théorique qui a été utilisée pour développer le calculateur de salaire "Salarium" sur le site de l'OFS (URL http://www.bfs.admin.ch/bfs/portal/fr/index/themen/03/04/blank/key/lohnstruktur/salarium.html). Cette application interactive permet d'obtenir des informations salariales pour un poste de travail spécifique (branche, économique, région, etc.) et pour des caractéristiques individuelles à choix (âge, formation, etc.). Les variables prédictives ont été définies au chapitre 6.1.6. Les informations salariales sont les suivantes :

- le salaire mensuel brut (valeur centrale ou médiane)
- la dispersion des salaires (intervalle interquartile)
- certains facteurs influençant le salaire (tableau des variations du salaire par région, niveau de qualification, position professionnelle et sexe)
- la comparaison avec les données salariales personnelles (montant salarial défini sur la base des mêmes composantes que celles du salaire médian).

Références

[1] T.J. Hastie and R.J. Tibshirani (1990). *Generalized Additive Models* . Chapman & Hall.

[2] R. L. Chambers, C. J. Skinner (2003). *Analysis of Survey Data*. Wiley series in survey methodology.

[3] Warren F. Kuhfeld. SAS OnlineDoc 9.1.3. The TRANSREG Procedure. URL http://support.sas.com/onlinedoc/913/docMainpage.jsp.

[4] Warren F. Kuhfeld (1990). SAS Technical Report R-108 : Algorithms for PRINQUAL and TRANSREG Procedures. URL http://support.sas.com/kb/23/addl/fusion_23806_1_r108_59040.pdf.

[5] Bentz D., Tschannen A. (2007). Frauenlöhne, Männerlöhne. Eine Bestandesaufnahme in der Zürcher Privatwirtschaft. Statistisches Amt Zürich. URL http://www.statistik.zh.ch/themenportal/themen/down.php?id=488&fn=2007_14.pdf.

[6] (2008). L'enquête suisse sur la structure des salaires - Panorama salarial 2006. *Statistique de la Suisse*, OFS. URL http://www.bfs.admin.ch/bfs/portal/fr/index/themen/03/22/publ.Document.108492.pdf.

[7] Graf, M. (2004). Enquête suisse sur la structure des salaires 2002. Plan d'échantillonnage et extrapolation pour le secteur privé. *Rapport de méthode 338-0025*, Office fédéral de la statistique, Neuchâtel. URL http://www.bfs.admin.ch/bfs/portal/de/index/infothek/erhebungen__quellen/methodenberichte.Document.38557.pdf.

[8] Graf, R. (2006). Löhne. Ortsübliche Branchenlöhne in 7 Schweizer Regionen. Neue erweiterte Ausgabe 2006. Schweizerischer Gewerkschaftsbund.

A Annexes

A.1 Tableau avec les paramètres ν optimal et la qualité de l'approximation pour le modèle anforni 2, 3 et 4

noga	ν optimal	SC_{erreur}
10	3.04358	0.17361
15	4.64773	0.05081
16	3.60694	0.28142
17	5.17056	0.17655
18	3.78447	0.59586
19	2.96546	0.68858
20	3.22940	0.14890
21	6.71202	0.27545
22	4.68865	0.08358
23	5.50446	0.06420
25	3.72911	0.38164
26	3.83611	0.12858
27	3.90714	0.09095
29	4.03627	0.07263
30	4.57133	0.04384
33	4.91158	0.33535
36	3.44974	0.07615
40	4.43130	0.50685
45	3.50988	0.13526
50	3.66693	0.04702
51	4.36701	0.07265
52	3.75271	0.27152
55	3.63125	0.12011
60	3.48210	0.28981
61	3.14097	0.11387
62	4.51944	0.84963
63	3.83563	0.14846
64	4.38413	0.08791
65	5.50755	0.40481
66	4.18620	0.12513
67	4.05876	0.24850
70	3.87948	0.08558
72	4.65028	0.02631
73	4.58034	0.12622
80	4.38881	0.26014
85	4.48127	0.24868
90	3.99484	0.17683
91	4.32307	0.66082
92	3.46565	0.09832
93	2.89463	0.17286

SC_{erreur} est la somme des carrés de l'erreur standardisée qui est définie par 33. En analysant le tableau A.1 on voit que les sommes des carrés de l'erreur standardisée atteignent un maximum dans les nogas 19, 62 et 91. Les nogas 19 et 62 sont deux nogas avec peu de données se qui explique une qualité d'approximation moins bonne. Les graphiques de comparaison des

fonctions de répartition théoriques et empiriques montrent le même comportement.

A.2 Tableau avec les paramètres ν optimal et la qualité de l'approximation pour le modèle anforni 1

noga	ν optimal	SC_{erreur}
1040	4.13883	0.89763
45	4.23365	1.18366
5052	4.84345	0.45859
55	4.53405	0.85458
6064	4.09823	0.48425
6567	4.49855	2.15709
7074	4.54637	0.70959
80	3.72163	0.37493
85	3.64429	0.61021
9093	5.02158	0.25052

En analysant le tableau A.2 on voit que la somme des carrés de l'erreur standardisée pour le regroupement 6567 est le plus grand. Ceci se reflète aussi dans les graphiques des comparaisons des fonction de répartition théoriques et empiriques. Pour les regroupements 80 et 9093 la qualité de l'approximation est la meilleure de ce modèle. Les graphiques correspondants confirment ceci.

A.3 Quelques explications sur les programmes SAS

Cinq programmes SAS permettent de faire les calculs des salaires prédits et des graphiques correspondants. La structure des programmes est pour chaque modèle (anforni 2, 3 et 4, ainsi que anforni 1) la suivante :

- Programme principal
- Programme avec l'approximation du modèle empirique par un modèle théorique

Les programmes principaux permettent de calculer les tableaux décrits sous le chapitre 6.1 pour chaque modèle.

Les programmes avec l'approximation du modèle empirique par un modèle théorique permettent de calculer le degré de liberté optimal ν_{opt} de la loi de Student qui permet d'approcher la fonction de répartition des résidus (voir chapitre 7.2.1). Ce même programme permet en plus de faire des fonctions de répartition des salaires pour les domaines comme expliqué sous le chapitre 7.2.2.

Pour le modèle anforni 1 on a une deuxième version du programme graphique qui permet de faire aussi les fonctions de répartition par noga au lieu de les faire par regroupement de nogas.

A.3.1 Les programmes du modèle anforni 2, 3 et 4

Le programme principal pour ce modèle contient une macro noga qui permet de calculer les tableaux décrites sous le chapitre 6.1 pour chaque noga.

En plus de ces macros les fonctionnalités suivantes sont prévues dans ce programme :

- Une macro qui permet de faire des interpolations des variables ibgrs, alter et dienstja
- Création d'un fichier où les résidus sont standardisés avec std_e voir chapitre 7.2.1
- Calcul de la fonction de répartition empirique des résidus voir chapitre 7.2.1

– Création des listes concernant la convergence de la procédure TRANSREG ainsi que du tableau ANOVA sous forme de pdf

Le programme avec l'approximation du modèle empirique par un modèle théorique pour ce modèle contient la même macro noga comme dans le programme principal et permet d'une part de calculer le ν_{opt} en utilisant une macro Student. Comme input le fichier avec les résidus standardisés est utilisé. Avec la SAS procédure nlin et la méthode Marquardt ce degré de liberté optimal est déterminé, voir chapitre 7.2.1. A partir de ce degré optimal les percentiles de la loi de Student sont déterminés et la fonction de répartition des résidus avec la loi de Student et la fonction empirique sont construites, voir chapitre 7.2.2.

Pour des salaires fixés les fonctions de répartition des salaires par domaines sont alors calculées en utilisant comme fichiers d'entrée les fichiers suivants :

– liste avec les ν_{opt}
– le fichier avec les résidus standardisés du programme principal.

A.3.2 Les programmes du modèle anforni 1

Le programme principal pour le modèle anforni 1 contient la définition des regroupements de noga (voir chapitre 2.2.4). On rajoute dans le modèle la variable explicative nog_2 et en enlève la variable explicative anforni.

A part ces changements la structure du fichier reste la même.

Les programmes avec l'approximation du modèle empirique par un modèle théorique ont la même structure comme dans le cas du modèle anforni 2, 3 et 4. Les deux programmes permettent de calculer la fonction de répartition des résidus avec la loi de Student et la fonction empirique. A partir de là les programmes calculent les fonctions de répartition des salaires par domaine. Ils se distinguent de la manière suivante :

– Le premier programme fait ces fonctions par regroupement de noga
– Le deuxième programme fait ces fonctions par noga

Un grand merci à Markus Eichenberger de DiSo Solution AG, qui a documenté ces programmes SAS et contribué à leur écriture.

Methodenberichte des Dienstes Statistische Methoden des BFS
Rapports de méthodes du Service de méthodes statistiques de l'OFS
Methodology reports published by the SFSO's Statistical Methods Unit

Andrade, B., Graf, M. (2008). Enquête suisse sur la structure des salaires 2006. Aspects méthodologiques du modèle des salaires "Salarium". Numéro de commande : 338-0053

Renaud, A. (2008). Statistique de l'emploi. Révision 2007 : cadre de sondage et échantillonage. Numéro de commande : 338-0052

Graf, E. (2008). Pondérations du SILC pilote. SILC_I vague 2, SILC_II vague 1, SILC_I et SILC_II combinés. Numéro de commande : 338-0051

Kilchmann, D. (2008). Statistik der sozialmedizinischen Institutionen 1999-2004 und Krankenhausstatistik 1999-2002. Einsetzungen für fehlende Daten. Bestellnummer : 338-0050

Renaud, A. (2008). Technologies de l'information et de la communication. Estimations sur la base de la statistique de la valeur ajoutée. Numéro de commande : 338-0049

Assoulin, D. (2007). Wertschöpfungsstatistik. Einsetzungsversuche für fehlende Antworten grosser Unternehmen. Bestellnummer : 338-0048

Kilchmann, D. (2007). Beherbergungsstatistik Campingplätze. Stichprobenrahmen und Schätzverfahren 2005/06. Bestellnummer : 338-0047

Gabler, S., Häder, S. (2007). Haushalts- und Personenerhebungen. Machbarkeit von Random Digit Dialing in der Schweiz. Bestellnummer : 338-0046

Ferrez, J., Graf, M. (2007). Enquête suisse sur la structure des salaires. Programmes R pour l'intervalle de confiance de la médiane. Numéro de commande : 338-0045

Renaud, A. (2007). Harmonisation de la scolarité obligatoire en Suisse (HarmoS). Design général de l'enquête et échantillon des écoles. Numéro de commande : 338-0044

Potterat, J. (2007). Betriebszählung 2005. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer : 338-0043

Hulliger, B. (2006). Umweltschutzausgaben der Unternehmen 2003, Stichprobenplan, Datenaufbereitung und Schätzverfahren. Bestellnummer : 338-0042

Renfer, J.-P. (2006). Enquête sur les chiffres d'affaires du commerce de détail. Plan d'échantillonnage et méthodes d'estimation. Numéro de commande : 338-0041

Salamin, P.-A. (2006). Statistique de l'aide sociale dans le domaine de l'asile. Plan de sondage et extrapolations pour l'enquête pilote 2005. Numéro de commande : 338-0040

Renaud, A. (2006). Statistique suisse des bénéficiaires de l'aide sociale. Pondération des communes 2004. Numéro de commande : 338-0039

Graf, M. (2006). Swiss Earnings Structure Survey 2002-2004. Compositional data in a stratified two-stage sample : Analysis and precision assessment of wage components. Order number : 338-0038

Potterat, J. (2006). Pensionskassenstatistik 2004. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer : 338-0037

Potterat, J. (2006). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2004. Stichprobenplan, Gewichtung und Schätzverfahren. Bestellnummer : 338-0036

Kilchmann, D. (2006). Vierteljährliche Wohnbaustatistik. Stichprobenplan, statistische Datenaufarbeitung und Schätzverfahren 2005. Bestellnummer : 338-0035

Kilchmann, D. (2006). Erhebung über Forschung und Entwicklung in der schweizerischen Privatwirtschaft 2004. Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Bestellnummer : 338-0034

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 2. Bestellnummer : 338-0033

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 1. Bestellnummer : 338-0032

Graf, M., Matei, A. (2005). Enquête suisse sur la structure des salaires 2002. La précision du salaire brut standardisé médian. Numéro de commande : 338-0031

Graf, E., Renfer, J.-P. (2005). Enquête suisse sur la santé 2002. Plan d'échantillonnage, pondération et estimation de la précision. Numéro de commande : 338-0030

Potterat, J. (2005). Mietpreis-Strukturerhebung 2003. Gewichtung und Schätzverfahren. Bestellnummer : 338-0029

Potterat, J. (2005). Landwirtschaftliche Betriebszählung 2003. Schätzverfahren für die Zusatzerhebung. Bestellnummer : 338-0028

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Order number : 338-0027

Kilchmann, D. (2004). Revision des Schweizerischen Lohnindex. Schätzmethoden der Lohnindices und deren Varianzschätzer. Bestellnummer : 338-0026

Graf, M. (2004). Enquête suisse sur la structure des salaires 2002. Plan d'échantillonnage et extrapolation pour le secteur privé. Numéro de commande : 338-0025

Renaud, A. (2004). Analyse de données d'enquêtes. Quelques méthodes et illustration avec des données de l'OFS. Numéro de commande 338-0024

Renaud, A., Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Numéro de commande : 338-0023

Graf, M. (2004). Fusion de données. Etude de faisabilité. Numéro de commande : 338-0022

Potterat, J. (2003). Mietpreis-Strukturerhebung 2003. Entwicklung des Stichprobenplans und Ziehung der Stichprobe. Bestellnummer : 338-0021

Potterat, J. (2003). Landwirtschaftliche Betriebszählung 2003. Stichprobenplan der Zusatzerhebung. Bestellnummer : 338-0020.

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sur-couverture (E-sample). Numéro de commande : 338-0019

Hulliger, B. (2003). Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Erhebung über F+E in der schweizerischen Privatwirtschaft 2000. Bestellnummer : 338-0018

Renfer, J.-P. (2003). Enquête 2000 sur la recherche et le développement dans l'économie privée en Suisse. Plan d'échantillonnage. Numéro de commande : 338-0017

Potterat, J. (2003). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe. Schätzverfahren. Bestellnummer : 338-0016

Graf, M., Matei, A. (2003). Stratégie de choix des modèles de désaisonnalisation. Application aux séries de l'emploi total. Numéro de commande : 338-0015

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Methoden für die Datenbereinigung. Bestellnummer : 338-0014

Renaud, A. (2002). Programme international pour le suivi des acquis des é lèves (PISA). Plans d'échantillonnage pour PISA 2000 en Suisse. Numéro de commande : 338-0013

Renfer, J.-P. (2002). Enquête 2001 sur les coûts et l'utilité de la formation des apprentis du point de vue des établissements. Plan d' échantillonnage. Numéro de commande : 338-0012

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Stichprobenplan und Schätzverfahren für die provisorischen Ergebnisse. Bestellnummer : 338-0011

Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Numéro de commande : 338-0010

Renaud, A., Eichenberger P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Numéro de commande : 338-0009

Kilchmann, D., Hulliger, B. (2002). Stichprobenplan für die Obstbaumzählung 2001. Bestellnummer : 338-0008

Graf, M. (2002). Passage du concept établissement au concept entreprise. Numéro de commande : 338-0007

Salamin, P.A. (2001). La technique de la double enquête pour la statistique du transport routier de marchandise. Numéro de commande : 338-0006

Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro de commande : 338-0005

Potterat, J., Hulliger, B. (2001). Schätzung der Sägereiproduktion mit der Sägerei-Erhebung PAUL. Bestellnummer : 338-0004

Graf, M. (2001). Désaisonnalisation. Aspects méthodologiques et application à la statistique de l'emploi. Numéro de commande : 338-0003

Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer : 338-0002

Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande : 338-0001

Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen : Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer : 338-0000

de Rossi, F.-X. (1998). Méthodes statistiques pour le compte routier suisse.

Hulliger, B., Kassab, M. (1998). Evaluation of Estimation Methods for the Survey on Environment Protection Expenditures of Swiss Communes.

Salamin, P.A. (1998). Etablissement d'une clef de passage pondérée entre l'ancienne (NGAE 85) et la nouvelle nomenclature (NOGA 95) générale des activités économiques.

Peters, R. (1998). Extrapolation des données de l'enquête de structure sur les loyers.

Bender, A., Hulliger, B. (1997). Enquête suisse sur la population active : rapport de pondération pour 1996.

Salamin, P.A. (1997). Evaluation de la Statistique de l'emploi.

Peters, R. (1997). Etablissement du plan d'échantillonnage pour l'enquête 1996 sur la recherche et le développement dans l'économie privée en Suisse.

Peters, R. (1997). Enquête 1996 sur la structure des salaires en Suisse : établissement du plan d'échantillonnage.

Peters, R. (1996). Pondération des données de l'enquête sur la famille en Suisse.

Comment, T., Hulliger, B., Ries, A. (1996). Gewichtungsverfahren für die Schweizerische Arbeitskräfteerhebung (1991-1995).

Hulliger, B. (1996). Haushalterhebung Familie 1994 : Stichprobenplan, Stichprobenziehung und Reservestichproben.

Peters, R., Hulliger, B. (1996). Schätzverfahren für die Lohnstruktur-Erhebung 1994 / Procédure d'estimation pour l'enquête de 1994 sur la structure des salaires.

Peters, R. (1996). Schéma de pondération des indices PAUL.

Hulliger, B., Peters, R. (1996). Enquête sur le comportement de la population suisse en matière de transport en 1994 : plan d'échantillonnage et pondération.

Hulliger, B. (1996). Gütertransportstatistik 1993 : Schätzverfahren mit Kompensation der Antwortausfälle.

Salamin, P.A. (1995). Estimation des flux pour le module II des comptes globaux du marché de travail.

Peters, R. (1995). Enquête de structure sur les loyers : établissement d'un plan d'échantillonnage stratifié.

Hulliger, B. (1995). Konjunkturelle Mietpreiserhebung : Stichprobenplan und Schätzverfahren.

Schwendener, P. (1995). Verbrauchserhebung 1990 - Vertrauensintervalle.

Peters, R., Hulliger, B. (1994). La technique de pondération des données : application à l'enquête suisse sur la santé.

Hulliger, B., Peters, R. (1994). Enquête sur la structure des salaires en Suisse : stratégie d'échantillonnage pour le secteur privé.

Publikationsprogramm BFS

Das Bundesamt für Statistik (BFS) hat – als zentrale Statistikstelle des Bundes – die Aufgabe, statistische Informationen breiten Benutzerkreisen zur Verfügung zu stellen.

Die Verbreitung der statistischen Information geschieht gegliedert nach Fachbereichen (vgl. Umschlagseite 2) und mit verschiedenen Mitteln

Programme des publications de l'OFS

En sa qualité de service central de statistique de la Confédération, l'Office fédéral de la statistique (OFS) a pour tâche de rendre les informations statistiques accessibles à un large public.

L'information statistique est diffusée par domaine (cf. verso de la première page de couverture); elle emprunte diverses voies:

Diffusionsmittel	*Kontakt* *N° à composer*	*Moyen de diffusion*
Individuelle Auskünfte	032 713 60 11 info@bfs.admin.ch	Service de renseignements individuels
Das BFS im Internet	www.statistik.admin.ch	L'OFS sur Internet
Medienmitteilungen zur raschen Information der Öffentlichkeit über die neusten Ergebnisse	www.news-stat.admin.ch	Communiqués de presse: information rapide concernant les résultats les plus récents
Publikationen zur vertieften Information (zum Teil auch als Diskette/CD-Rom)	032 713 60 60 order@bfs.admin.ch	Publications: information approfondie (certaines sont disponibles sur disquette/CD-Rom)
Online-Datenbank	032 713 60 86 www.statweb.admin.ch	Banque de données (accessible en ligne)

Nähere Angaben zu den verschiedenen Diffusionsmitteln liefert das laufend nachgeführte Publikationsverzeichnis im Internet unter der Adresse www.statistik.admin.ch→Aktuell→Publikationen.

La liste des publications, mise à jour régulièrement, donne davantage de détails sur les divers moyens de diffusion. Elle se trouve sur Internet à l'adresse www.statistique.admin.ch →Actualités→Publications.

Methodenberichte des Dienstes Statistische Methoden
Rapports de méthodes du Service de méthodes statistiques
Methodology Reports by the Statistical Methods Unit

Die Methodenberichte beschreiben die mathematischen und statistischen Methoden, die den Resultaten und Analysen der öffentlichen Statistik zu Grunde liegen. Sie enthalten ausserdem die Evaluation und Entwicklung von neuen Methoden im Hinblick auf eine zukünftige Anwendung. Diese Publikationen sollen einerseits die verwendeten Methoden dokumentieren, um Transparenz und Wissenschaftlichkeit sicher zu stellen, und sie sollen andererseits die Zusammenarbeit mit den Hochschulen und der Wissenschaft fördern.

Zur Illustration der beschriebenen mathematischen Konzepte, werden im Bericht numerische Resultate aufgeführt. Diese sind allerdings nicht als offizielle Resultate der betreffenden Erhebungen zu verstehen. Ebenfalls können die tatsächlich angewendeten Methoden leicht von den hier beschriebenen abweichen.

Die Methodenberichte sind auf der Internetseite des BFS in elektronischer Form verfügbar.

Les rapports de méthodes décrivent les méthodes mathématiques et statistiques à la base des résultats et des analyses de la statistique publique. Ils présentent également l'évaluation et le développement de nouvelles méthodes en vue d'une application future. Ces publications visent d'une part à documenter les méthodes utilisées ou envisagées dans un souci de transparence et de rigueur scientifique, et d'autre part à favoriser la collaboration avec le monde scientifique et universitaire.

Les résultats numériques présentés dans les rapports de méthodes illustrent les concepts mathématiques décrits, mais ne sont pas des résultats officiels des enquêtes concernées. De même, les méthodes réellement appliquées peuvent différer légèrement de celles décrites dans ces rapports.

Les rapports de méthodes sont disponibles sous forme électronique sur le site internet de l'OFS.

Ce rapport comporte des explications mathématiques et des présentations graphiques du modèle des salaires développé. La méthode utilisée pour faire des prédictions de salaires est présentée. Des exemples d'application pour une branche économique spécifique sont donnés à différents niveaux. Puis, un chapitre explique la qualité obtenue avec ces modèles. Pour terminer, des comparaisons entre les fonctions de répartition théoriques et empiriques sont présentées pour valider les modèles.

N° de commande: 338-0053
Commandes: 032 713 60 60 Fax: 032 713 60 61 E-Mail: order@bfs.admin.ch
Prix: gratuit
ISBN 978-3-303-00410-4