Rapport - Projet SAS

Sarah CHIKHAOUI

Données Benefits

Master 2 IMB
2023/2024

Partie 1. Analyse descriptive du jeu de données

Description du jeu de données :

Ce jeu de données contient sur une période de 1982 à 1991, **4877 observation** de "*Blue collar workers*" (cols bleus en français) c'est-à-dire les personnes effectuant un travail manuel (ouvrier d'usine, secteur du bâtiment…) par opposition aux cols blancs (travail de bureau).
Ce jeu comporte des informations personnelles comme l'âge, le sexe et la situation professionnelle (ancienneté, date de perte d'emploi…).
Le but ici est d'essayer d'établir un lien entre bénéficier de l'assurance chômage et temps de retour à l'emploi.

Description des variables :

Le jeu de données contient 18 variables dont 5 quantitatives (en bleu) et 14 qualitatives (en rouge).

- **stateur** : *state unemployment rate* soit le taux de chômage de l'état dans lequel habite l'individu (en pourcentage)
- **statemb** : *state maximum benefit level* soit le taux maximum d'assurance chômage auquel on peut bénéficier dans cet état
- **state** : code correspondant à l'état de résidence de l'individu
- **age** : âge de l'individu
- **tenure** : ancienneté (en années) dans l'emploi perdu
- **joblost** : un facteur à 4 niveaux (chômage partiel, poste supprimé, saisonnier_emploi_terminé, autre) expliquant la raison de la perte d'emploi
- **nwhite** : l'individu est-il non blanc ?
- **school12** : l'individu a-t-il effectué plus de 12 années d'école ?
- **sex** : sexe de l'individu .
- **bluecol** : l'individu est-il un ouvrier ? Ici, tous les individus du jeu de données le sont
- **smsa** : l'individu vit-il dans une zone statistique métropolitaine (MSA), c'est à dire une zone métropolitaine centrée sur une seule grande ville qui exerce une influence substantielle sur la région.
- **married** : l'individu est-il marié ?
- **dkids** : l'individu a-t-il au moins un enfant ?
- **dykids** : l'individu a-t-il au moins un enfant en bas âge (de 0 à 5 ans) ?
- **yrdispl** : un facteur de 10 niveaux indiquant l'année de perte de l'emploi au moment de la demande de chômage (de 1982 = 1 à 1991=10).
- **rr** : taux de remplacement: le rapport entre le montant de l'assurance chômage et le salaire du dernier emploi du demandeur, sur une période d'une semaine
- **head** : dans le cas où l'individu habite en couple et subvient seul aux besoins matériels/substantiels de la famille
- **ui** : indique si l'individu a obtenu des prestations chômage suite à sa demande. C'est notre variable d'intérêt

Valeurs manquantes ?

Comme le montrent la ligne de code ci-dessous et son résultat, le jeu de données ne contient pas de valeur manquante

```
proc means data = benefits NMISS N;
run;
```

Variable	Nbre manquant	N
stateur	0	4877
statemb	0	4877
state	0	4877
age	0	4877
tenure	0	4877
yrdispl	0	4877
rr	0	4877

2. analyse des variables du jeu de données : traitements univariés (répartition, adéquation à une loi théorique...)

Notre jeu de données traite du chômage des ouvriers aux USA (depuis l'année 1972).

Ce jeu de données contient 4877 observation correspondant aux ouvriers et 18 variables.

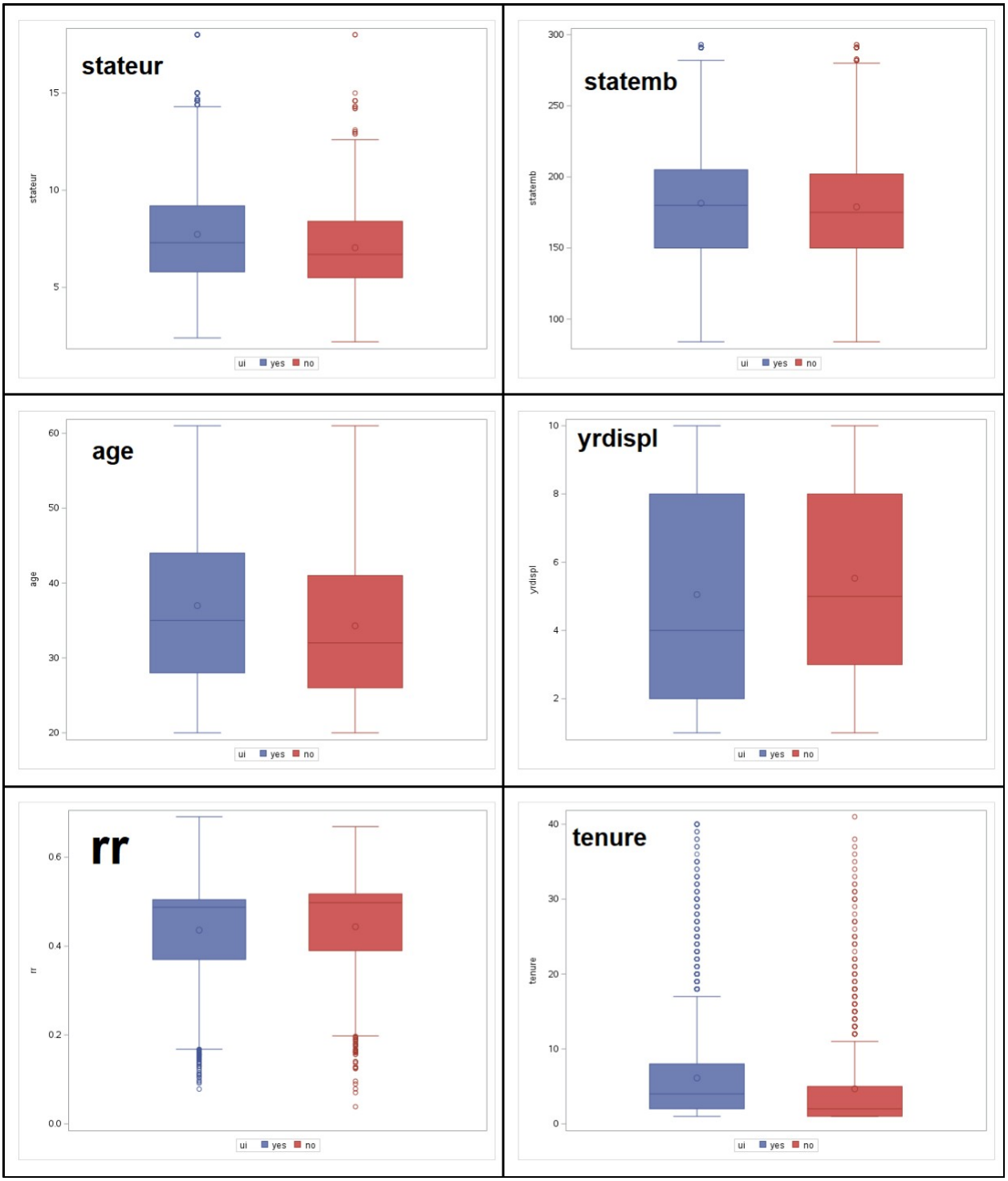
Le but de nôtre étude est de savoir si bénéficier de l'assurance chômage aux USA prolonge ou diminue le temps de retour à l'emploi. Or le statut d'obtention de cette assurance correspond à 2 modalités (obtenue ou non). Alors je propose d'évaluer le lien entre le statut d'obtention de cette assurance et les autres variables par une régression logistique binaire.

Je remarque que la variable bluecol est en fait constate donc inutile à l'explication d'une variable, on ne l'intègre alors pas au modèle que l'on considère relatif à une population ouvrière.

Afin de proposer un modèle le moins biaisé possible, il est recommandé d'entreprendre au préalable une analyse des observations. Ainsi on cherche à comprendre la répartition des observations et à identifier de

potentielles valeurs extrêmes. Pour ce faire, on a recours à une méthode de visualisation offerte par le box-plot.

Box-plots et interprétation des résultats :



	Stateur	Statemb	Age	Yrdispl	RR	Tenure
min	2.2	84.0	20.0	1.0	0.04	1.0
max	18.0	293.0	61.0	10.0	0.69	41.0
moyenne	7.5	180.7	36.1	5.2	0.44	5.7
médiane	7.2	177.0	34.0	5.0	0.49	3.0
Q1	5.7	150.0	28.0	2.0	0.38	2.0
Q3	9.0	205.0	43.0	8.0	0.51	7.0

Stateur et Statemb : ces 2 variables sont assez dispersées. La quasi-totalité des valeurs de Stateur se situent entre 4 et 14 (92.5% pour être exact). Quant à Statemb, 96% des données se situent entre 100 et 290.
Ni le boxplot de Stateur, ni celui de Statemb ne présentent d'asymétrie flagrante, c'est-à-dire que dans les deux cas, on a à peu près autant de valeurs au-dessus qu'en dessous de la boîte (qui pour rappel contient 50% des observations). Les 2 variables contiennent également peu de valeurs aberrantes.
Pour statemb, on ne distingue pas de différence notable entre les deux classe (ui = yes et ui = no). En revanche, pour stateur, les valeurs sont en moyenne plus grandes pour ui = yes que pour ui = non. Cela signifie que les personnes qui ont demandé et obtenu des aides au chômage l'ont fait (en moyenne) dans des états où le taux de chômage était plus élevé, ce qui semble parfaitement logique : s'il y a plus de chômage dans un état que dans un autre, il y aura également plus de demande d'aides au chômage.

Age : pour cette variable on constate une asymétrie. En-dehors de la boîte (50% des données), les données semblent plus abondantes au-dessus de celle-ci qu'en dessous. On a donc plus d'individus âgés de plus de 40 ans que de moins de 25 ans. On remarque que l'âge est en moyenne plus élevée pour les personnes ayant demandé des aides que pour ceux n'en ayant pas demandé.

Yrdispl : la variable Yrdispl semble uniformément répartie entre 1 et 10, sans asymétrie particulière ou une quantité trop importante de valeurs aberrantes. En comparant les 2 boxplots, on peut également noter que les personnes ayant demandé des aides étaient des personnes ayant perdu leur emploi (en moyenne) plus tôt que ceux n'ayant pas fait de demande.

RR : la variable RR comporte un nombre non négligeable de valeurs aberrantes. Il s'agit des valeurs inférieures à 0,2. La différence selon la variable ui n'est pas notable, les 2 boxplots étant quasiment identiques.

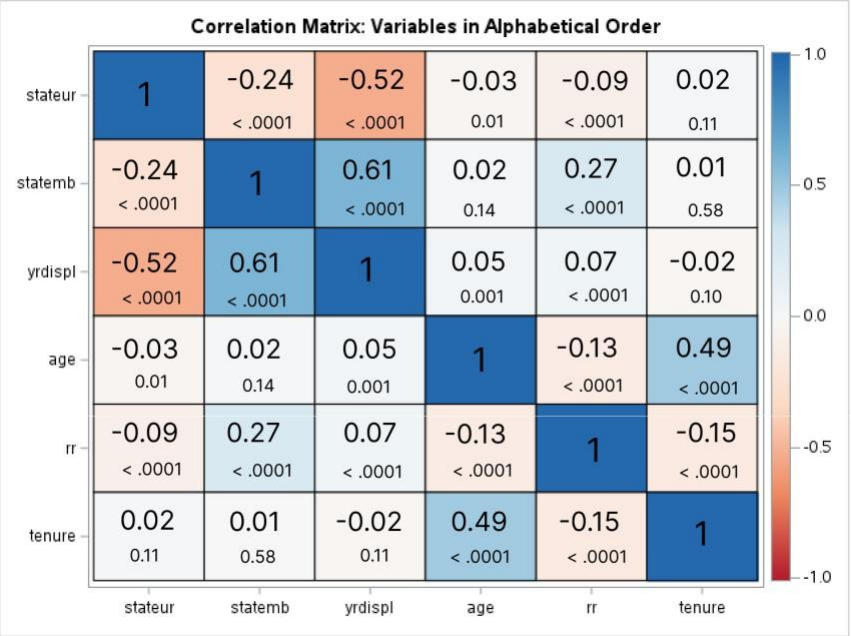
Tenure : Ici, 75% des données sont inférieures à 7. Pourtant on trouve un nombre important de valeurs aberrantes (près de 8% au-dessus de 15). Il s'agit des personnes ayant perdu leur emploi malgré une grande ancienneté. En comparant les 2 boxplots, on s'aperçoit que l'ancienneté moyenne des personnes ayant demandé des aides est légèrement plus élevé que ceux ne l'ayant pas fait.

Parmi les variables explicatives, certaines sont notamment quantitatives continues. La suite de l'analyse a pour but de détecter:

-lesquelles présentent des liaisons linéaires entre elles deux à deux

- à quel degré. En effet, certaines de ces variables peuvent mesurer le même phénomène, ce qui peut provoquer une augmentation de la variance des coefficients du modèle de régression, les rendre instables et difficiles à interpréter. Ce caractère problématique pour toute régression est appelé la multicolinéarité. Ainsi pour détecter ce phénomène je vais devoir identifier quelles variables sont susceptibles d'être corrélées entre elles. Pour ce faire, dans un 1er temps, je trace les graphes de tous les binômes de variables qualitatives continues possibles selon le statut de la réponse ui (par la couleur). Ensuite, je calcule et interprète leurs coefficients de corrélation.

Matrice des coefficients de corrélation de Pearson obtenus avec la proc corr et l'option Pearson :



Cette matrice présente donc les coefficients de corrélation de Pearson calculés sur les variables quantitatives 2 à 2

Chaque coefficient de corrélation est calculé selon la formule suivante : $r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$. La procédure corr de SAS renvoie également les p-valeurs relatives au test de significativité de chaque coefficient. Si cette valeur est inférieure à 0,05 on peut conclure que le coefficient est significativement différent de zéro au seuil de 5%. Il s'agit du test de sphéricité de Bartlett.

Le coefficient de corrélation est compris entre -1 et 1. Plus il se rapproche de ces deux extrêmes, plus la corrélation linéaire entre les 2 variables est forte. Elle vaut 1 (respectivement -1) si une des variables est une fonction affine croissante (respectivement décroissante) de l'autre et elle vaut 0 si les variables ne sont pas linéairement corrélées (elles peuvent cependant être corrélées non-linéairement).

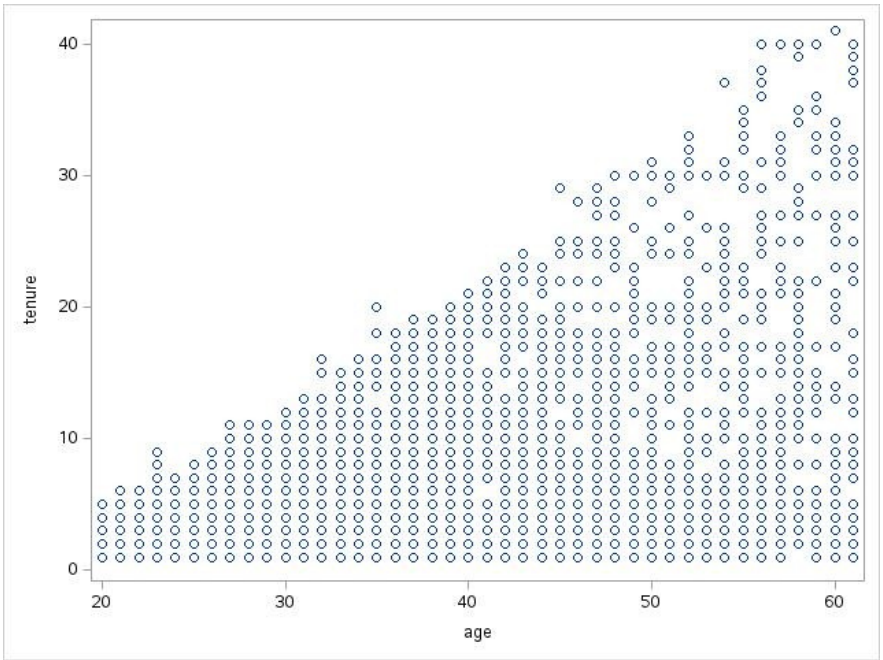
En plus d'une matrice de corrélation, le graphique constitue une heatmap permettant d'identifier plus aisément les coefficients notables. Plus on tend vers le rouge, plus coefficient se rapproche de -1 et plus on tend vers le bleu et plus le coefficient est proche de 1.

Ici, on peut noter quelques coefficients intéressant :

- Statemb & Yrdispl : 0.61
Ici, j'ai une corrélation positive assez forte. Elle signifierait qu'à mesure que l'on se rapproche de 1991, la valeur de Statemb (soit le taux maximum d'assurance chômage auquel on peut bénéficier dans l'état où vit l'individu) augmente également.
Comme pour la corrélation entre Stateur et Yrdispl, celle-ci peut être expliquée en regardant l'évolution du taux d'assurance chômage aux Etats-Unis entre 1982 et 1991. Ce taux a connu une croissance durant cette période, ce qui explique que pour une valeur plus élevée de Yrdispl, on a également une valeur plus élevée de Statemb.

- Stateur & Yrdispl : -0.52
Il s'agit d'une corrélation négative relativement forte. Cela voudrait dire qu'à mesure que l'année de perte d'emploi augmente, le taux de chômage dans l'état de l'individu diminue. En regardant les courbes du taux de chômage des États-Unis entre 1982 et 1991 (période sur laquelle s'étend notre jeu de données), on s'aperçoit qu'il atteignait près de 10% en 1982 pour finalement valoir autour de 7% en 1991 avec même un pic à 5% en 1989.
Il semble donc logique, que pour des valeurs plus élevées de Yrdispl soit une perte d'emploi plus proche de 1991 que de 1982, la valeur de Stateur soit plus faible quand on connaît la tendance de la courbe du taux de chômage américain sur cette période.

- Tenure & Age : 0.49
Cette corrélation importante entre l'Age et le nombre d'années d'ancienneté avec la perte d'emploi peut être expliqué assez logiquement. Il semble raisonnable de penser que les personnes avec une plus grande ancienneté soient des personnes plus âgées. A l'inverse, les individus avec une faible ancienneté sont plus susceptibles d'être des personnes relativement jeunes.
Il est donc normal d'observer cette corrélation positive entre Tenure et Age
Certaines valeurs de Tenure peuvent également être exclues en fonction de l'âge. Par exemple, un individu de 20 ans ne pourra pas avoir une ancienneté de 10 ans ou plus. Idem pour une personne âgée de 40 ans et une ancienneté de 25 ans ou plus. Ceci explique l'allure du nuage de points ci-dessous de Tenure en fonction de Age (proc sgplot).



3. construction de modèles de régressions simples

Dans un premier temps, je vais effectuer des régressions simples, c'est-à-dire que je vais tenter d'expliquer la variable d'intérêt à l'aide d'une seule variable explicative.

Le tableau ci-dessous récapitule les différentes régressions simples et donne les valeurs de l'intercept et du coefficient correspondant à la variable considérée.

Variable	Intercept	Coefficient	Coefficient 2 (var qualitatives)	Coefficient 3 (var qualitatives)
stateur	0.091	-0.117		
statemb	-0.523	-0.001		
age	0.12	-0.025		
tenure	-0.547	0.042		
yrdispl	-1.034	0.050		
rr	-1.075	0.691		
joblost	-0.687	0.148 (other)	0.168 (position_abolished)	0.044 (seasonal_job_ended)
nwhite	-0.790	0.026 (no)		
school12	-0.742	-0.048 (no)		
sex	-0.805	-0.062 (female)		
smsa	-0.792	-0.066 (no)		
dkids	-0.772	0.006 (no)		
dykids	-0.774	0.005 (no)		
head	-0.777	-0.014 (no)		

Les modèles linéaires simples avec les varibales stateur, rr et joblost sont ceux avec le coefficient le plus élevé et sont donc les variables explicatives qui expliquent le mieux notre variable d'intérêt **ui**

4. Analyse de la liaison entre la variable que l'on veut expliquer et les autres variables du jeu de données

A présent, je peux raisonnablement proposer un modèle de régression logistique de nos données en modélisant la variable réponse UI par une loi de Bernoulli (prenant les valeurs 0 ou 1) de paramètre p, dont la fonction de lien naturelle est la fonction logit. Aussi puisque le but de cette étude concerne le fait bénéficier de l'assurance chômage, il faut alors prendre la modalité "yes" comme valeur de référence pour la variable réponse ui.
Néanmoins une méthode plus rapide d'ajustement des modèles de régression est la procédure stepwise avec comme valeur "yes" pour le paramètre *event* .
Le choix d'intégration au modèle, des variables explicatives candidates, repose sur un critère statistique, relatif au modèle considéré sans la variable en question. Celui-ci est soumis au test global de nullité des coefficients qui confronte H0 : « les coefficients sont tous nuls » vs H1 : « il existe au moins un de ces coefficients qui soit soit non nul ». Les valeurs des p-values de ce test indiquerons si ensembles, les variables candidates rende ou non, le modèle de régression significatif. A chaque étape de sélection d'un modèle, toutes les variables introduites dans le modèle précédent sont ré-examinées. En effet, une variable considérée comme la plus significative à une étape de l'algorithme peut, à une étape ultérieure, devenir non significative.
Dans le cas de la procédure stepwise implémentée dans sas, cette méthode utilise ici 3 critères que sont: l'AIC , le BIC aussi appelé SBC et la déviance, du modèle considéré. Dans un 1er temps, la méthode considère le modèle trivial restreint à la constate uniquement (sans variables explicatives). Les 3 critères récemment évoqués sont alors calculés puis comparés à ceux de chaque modèle où une seule variable explicative est ajoutée en plus de la constate. Ensuite, parmi tous les modèles candidats à la sélection, celui de plus faible AIC sera proposé puis soumis à 3 tests de sélection de modèles que sont : le test du score, wald et du rapport des vraisemblances. j'en reparlerai lors de l'interprétation de la table *Statistique d'ajustement du modèle* en sortie de la procédure logistique.

Le tableau ci-dessous récapitule le modèle final retourné par les 3 régressions logistiques que je vais effectué, à savoir : méthode stepwise, méthode backward et méthode forward. Les lignes correspondent aux variables explicatives retenues ou non, tandis que les colonnes correspondent aux 3 méthodes de régression.
Les cases remplies par un X correspondent aux variables n'ayant pas été retenu.
Les cases dont le fond est vert indiquent des coefficients positifs alors que les fonds rouges indiquent les coefficients négatifs.

variable / methode	stepwise	backward	forward
intercept	1.356	1.356	1.263
stateur	-0.094	-0.094	-0.095
statemb	-0.006	-0.006	-0.006
age	-0.025	-0.025	-0.025
tenure	-0.032	-0.032	-0.032
yrdispl	0.062	0.062	0.062

rr	0.725	0.725	0.851
joblost = other	0.208	0.208	0.214
joblost = position_abolished	0.241	0.241	0.249
joblost = job_ended	-0.049	-0.049	-0.057
nwhite	X	X	X
school12	X	X	X
sex = female	X	X	-0.079
smsa = no	-0.084	-0.084	-0.085
dkids	X	X	X
dykids = no	0.112	0.112	0.107
head = no	-0.126	-0.126	-0.096

On remarque que les méthodes stepwise et backward renvoient exactement le même modèle : les mêmes variables ainsi que les mêmes coefficients.
Pour la méthode forward en revanche, une variable explicative est ajoutée au modèle. Il s'agit de la variable **sex** dont le coefficient vaut -0.079 (avec comme modalité de référence sex = female).
On peut également noter quelques différences de coefficients entre la méthode forward et les 2 autres méthodes. Pour la méthode forward, les coefficients des variables stateur, joblost = job_ended, smas = no, dykids = no ainsi que l'intercept sont légèrement inférieurs à ceux des 2 autres méthodes tandis que ceux des variables rr, joblost = other, joblost = position_abolished et head = no sont légèrement supérieurs. Les variable nwhite et school12 ne sont en revanche sélectionnées dans aucun des 3 modèles.

Interprétation de la proc logistic avec l'option stepwise:

On remarque que les premières tables de sortie de la procédure stepwise traitent du choix des valeurs de références pour toutes variables et en particulier celles qui seront gardées dans le modèle final. La tables "Profil de réponse" indique bien que la variable réponse ui prend comme modalité de référence 'no'. La table "Information sur les niveaux de classe" indiquent que pour les variables dichotomiques, la variable ordinale yrdispl, les variables joblost et sexe, sas a choisi respectivement comme valeurs de référence: 'yes', 'yrdispl= 1', 'joblost= slack_work' ou chômage partiel, et "sex=man" . Ces choix de sas se vérifient aussi dans la table "Analyse des valeurs estimées du maximum de vraisemblance" que je interpréterais plus loin.

Les tables "Récapitulatif sur la sélection Stepwise" et "Analyse des effets Type 3" présentent respectivement les variables prédictives sélectionnées pour le modèle final avec les statistiques de test et p-values<0.05 associées aux tests du Score et de Wald respectivement pour chaque table. La table "Analyse des valeurs estimées du maximum de vraisemblance" présente (de gauche à droite pour chaque colonne de la table):

- le nombre de degrés de liberté de chaque variable exogène sélectionnée dans le modèle final
- les estimations des coefficients de ces variables, soit, les log(OR) de chacune d'entre elles
- les écarts-types de l'estimation de ces coefficients
- à nouveau les statistiques de test et p-values<0.05 du test de Wald.

Note déjà que l'avantage de l'OR est qu'il reste interprétable même si la prévalence de l'événement d'intérêt est différente de celle de la population générale. En effet, les données ne représentent qu'un échantillon de la population générale, il n'est donc pas représentatif de celle-ci. Dans cette table, on observe que toutes les variables ont chacune une estimation de son ORs<1, et leur p-values du test de Wald révèle que ces estimations sont significatives. Selon le type de la variable l'interprétation diffère. Observe alors les résultats des estimations des ORs en colonne 1 de la table 'Estimation du rapport de cotes' ainsi que leur intervalles de confiance respectifs.

Estimation du rapport de cotes			
Effet	Estimation du point	Intervalle de confiance de Wald à95%	
stateur	1.087	1.022	1.155
statemb	1.006	1.001	1.011
state 11 vs 95	3.908	1.152	13.258
state 12 vs 95	1.349	0.446	4.080
state 13 vs 95	2.168	0.680	6.909
state 14 vs 95	1.947	0.717	5.281
state 15 vs 95	1.914	0.562	6.513
state 16 vs 95	1.957	0.649	5.900
state 21 vs 95	1.687	0.629	4.529
state 22 vs 95	1.310	0.484	3.547
state 23 vs 95	1.386	0.516	3.728
state 31 vs 95	1.514	0.535	4.285
state 32 vs 95	1.343	0.392	4.605
state 33 vs 95	1.482	0.535	4.105
state 34 vs 95	1.121	0.405	3.103
state 35 vs 95	1.980	0.660	5.942
state 41 vs 95	1.016	0.356	2.894
state 42 vs 95	1.566	0.493	4.980
state 43 vs 95	1.232	0.393	3.860
state 44 vs 95	1.309	0.435	3.938
state 45 vs 95	0.599	0.187	1.917
state 46 vs 95	1.154	0.360	3.698
state 47 vs 95	1.572	0.530	4.660
state 51 vs 95	0.765	0.241	2.433
state 52 vs 95	1.679	0.497	5.667
state 53 vs 95	0.466	0.117	1.866
state 54 vs 95	0.998	0.338	2.949
state 55 vs 95	1.742	0.543	5.584
state 56 vs 95	1.189	0.448	3.158
state 57 vs 95	1.810	0.573	5.718
state 58 vs 95	1.666	0.532	5.224
state 59 vs 95	0.676	0.250	1.824
state 61 vs 95	1.943	0.615	6.138
state 62 vs 95	1.550	0.488	4.920
state 63 vs 95	1.235	0.372	4.098
state 64 vs 95	1.017	0.313	3.307
state 71 vs 95	1.125	0.373	3.388
state 72 vs 95	0.901	0.301	2.695
state 73 vs 95	0.866	0.308	2.431
state 74 vs 95	0.728	0.277	1.918
state 81 vs 95	1.224	0.415	3.612
state 82 vs 95	1.026	0.340	3.099
state 83 vs 95	0.885	0.300	2.612
state 84 vs 95	0.634	0.223	1.806
state 85 vs 95	0.660	0.215	2.029
state 86 vs 95	0.756	0.230	2.483
state 87 vs 95	1.310	0.443	3.877

state 46 vs 95	1.154	0.360	3.698
state 47 vs 95	1.572	0.530	4.660
state 51 vs 95	0.765	0.241	2.433
state 52 vs 95	1.679	0.497	5.667
state 53 vs 95	0.466	0.117	1.866
state 54 vs 95	0.998	0.338	2.949
state 55 vs 95	1.742	0.543	5.584
state 56 vs 95	1.189	0.448	3.158
state 57 vs 95	1.810	0.573	5.718
state 58 vs 95	1.666	0.532	5.224
state 59 vs 95	0.676	0.250	1.824
state 61 vs 95	1.943	0.615	6.138
state 62 vs 95	1.550	0.488	4.920
state 63 vs 95	1.235	0.372	4.098
state 64 vs 95	1.017	0.313	3.307
state 71 vs 95	1.125	0.373	3.388
state 72 vs 95	0.901	0.301	2.695
state 73 vs 95	0.866	0.308	2.431
state 74 vs 95	0.728	0.277	1.918
state 81 vs 95	1.224	0.415	3.612
state 82 vs 95	1.026	0.340	3.099
state 83 vs 95	0.885	0.300	2.612
state 84 vs 95	0.634	0.223	1.806
state 85 vs 95	0.660	0.215	2.029
state 86 vs 95	0.756	0.230	2.483
state 87 vs 95	1.310	0.443	3.877
state 88 vs 95	0.831	0.261	2.641
state 91 vs 95	0.671	0.220	2.051
state 92 vs 95	1.476	0.479	4.547
state 93 vs 95	0.963	0.360	2.573
state 94 vs 95	0.856	0.282	2.598
age	1.020	1.012	1.027
tenure	1.028	1.014	1.041
joblost other vs slack_wo	0.542	0.471	0.625
joblost position vs slack_wo	0.518	0.408	0.656
joblost seasonal vs slack_wo	0.748	0.532	1.052
sex female vs male	1.238	1.040	1.473
married no vs yes	0.778	0.678	0.894
rr	0.381	0.185	0.786
head no vs yes	1.174	1.003	1.375
yrdispl 1 vs 10	2.118	1.265	3.546
yrdispl 2 vs 10	1.319	0.831	2.094
yrdispl 3 vs 10	1.117	0.741	1.683
yrdispl 4 vs 10	1.071	0.739	1.552
yrdispl 5 vs 10	1.283	0.885	1.861
yrdispl 6 vs 10	0.920	0.649	1.303
yrdispl 7 vs 10	0.933	0.656	1.326
yrdispl 8 vs 10	0.772	0.569	1.049
yrdispl 9 vs 10	0.966	0.723	1.290

En effet, pour les variables quantitatives continues, cette inégalité signifie que, toutes choses égales par ailleurs, le risque que ui=0 est moins élevé lorsque chacune d'elle (indépendamment des autres) augmente d'une unité, par rapport à la modalité de référence ui='yes'. Pour ces variables (state, statemb, age et tenure) remarque déjà qu'elles ont toutes un OR proche de 1 par valeur supérieur, néanmoins leur intervalles de confiance indiquent tous que ces variables n'influent pas de manière significative la probabilité du risque d'obtenir l'assurance chômage. En effet, la valeur 1 n'appartient à aucun des intervalles de confiance de ces variables.

Pour la variable qualitative joblost à 3 modalités (sans la modalité de référence), on observe pour chacune d'elles, un OR <1 mais seule la modalité 'seasonal' présente un intervalle de confiance qui contient la valeur 1. Ainsi on peut dire que toutes choses égales par ailleurs, un individu ayant terminé son emploi saisonnier non renouvelé, aura un risque 0.748 fois moins élevé d'obtenir l'assurance chômage qu'un individu ayant perdu son emploi pour cause de chômage partiel.

Pour la variable quantitative state les états ayant le code 45, 51, 53, 54, 59, 72, 73, 74, 83 à 86, 88, 91, 93, et 94 ont chacun un OR<1 dont l'intervalle de confiance contient la valeur 1, ce qui rend cet OR significatif. Ainsi un individu résident dans l'un de ces états, aura un risque moins élevé d'obtenir l'assurance chômage qu'un individu résidant dans l' état correspondant au code 95.

Pour la variable quantitative yrdispl, l'année 1991 est prise comme année de référence. Ainsi on observe que l'effet de l'année 1982 n'est pas significatif puisque l'intervalle de confiance de l'OR associé ne contient pas valeur 1. On remarque que de 1983 à 1986, chaque OR associé est supérieur à 1, cela signifie qu'un individu ayant perdu son emploi entre 1983 à 1986 a un risque plus élevé d'obtenir l'assurance chômage qu'un individu l'ayant perdu en 1991. Néanmoins, entre 1987 et 1990, la situation s'inverse. En effet, puisque tous les OR associés sont inférieurs à 1 et significatifs, cela signifie qu'un individu ayant perdu son emploi durant cette période aura un risque moins élevé d'obtenir une assurance chômage qu'un individu l'ayant perdu en 1991.

Pour les variables dichotomique (sex, married, head), aucun d'entre elles ne possède d'OR dont l'intervalle de confiance contient la valeur 1. Ainsi on peut dire que ces variables n'ont pas d'effet significatif sur le risque d'obtenir l'assurance chômage.

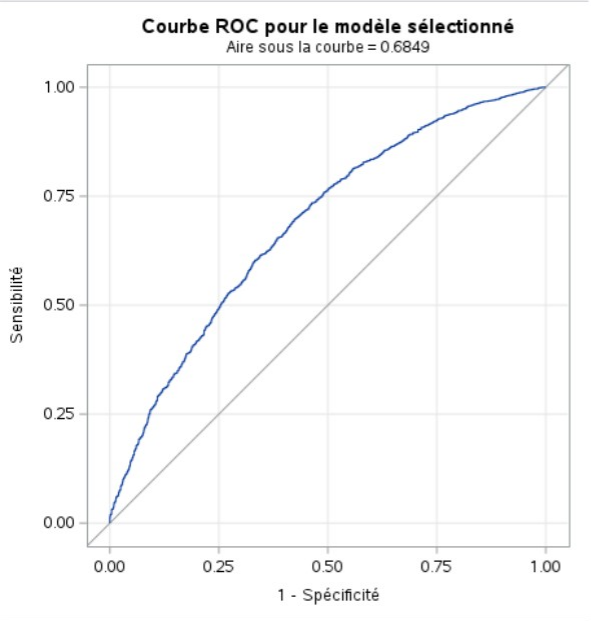
Matrice de confusion obtenue avec l'option CTABLE

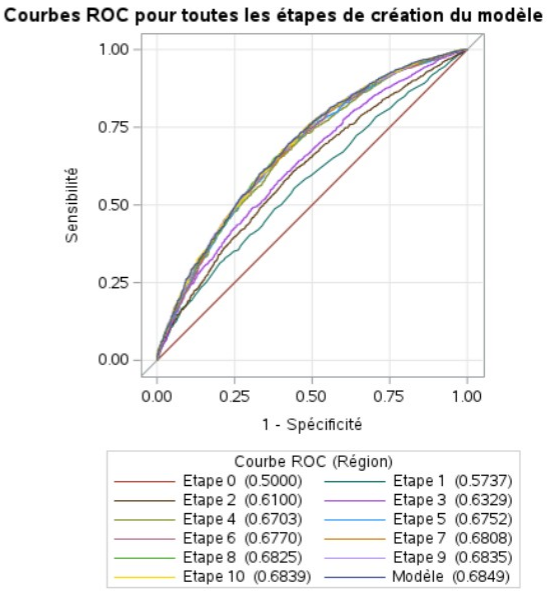
Matrice de confusion								
Correct		Incorrect		Pourcentages				
Evénement	Non-événement	Evénement	Non-événement	Correct	Sensi-bilité	Spéci-ficité	Préd pos	Préd nég
212	3183	152	1330	69.6	13.7	95.4	58.2	70.5

Cette matrice a pour but de mesurer la qualité d'un système de classification. Ici, cela mesure donc le taux d'attribution de labels (ui) corrects et incorrectes aux données en suivant le modèle. Ici, le modèle prédit 69.6% des labels correctement.

Interprétons maintenant la courbe de ROC du modèle sélectionné en fonction des étapes de sélection.

Je sais que la courbe ROC représente l'évolution de la sensibilité (taux de vrais positifs) en fonction de 1 – spécificité (taux de faux positifs) pour toutes les valeurs seuils possibles du modèle étudié. Ici la sensibilité est la capacité du modèle à bien détecter les individus ayant perçu l'assurance chômage et la spécificité est la capacité du test à bien détecter ceux qui n'en bénéficient pas. Autrement dit, la courbe de ROC cetitue un autre point de vue sur la qualité de la modélisation en fonction du nombre de variables explicatives. L'aire sous la courbe ROC (ou AUC) est un indicateur de la qualité de la prédiction. Pour notre modèle, l'AUC=0.68 signifie que, dans 68% des cas, un individu tiré au sort de la population des individus bénéficiant de l'assurance a une valeur du test diagnostique supérieure à celle d'un individu tiré au sort dans la population des individus qui n'en bénéficient pas.





Au vu de nôtre analyse je ne peux pas conclure. Cependant, une analyse plus poussée notamment avec des données concernant le temps de retour à l'emploi je aurait permis de répondre à la question initiale.