Meyssa BEDDAR
Tom BLACHON
Matthieu SIMOES

Master MéDAS
02/04/2022

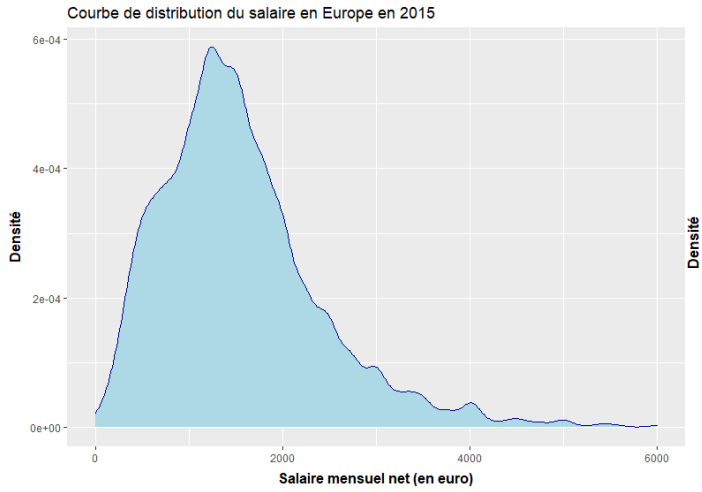
Projet de statistique :
Etude du salaire au Portugal

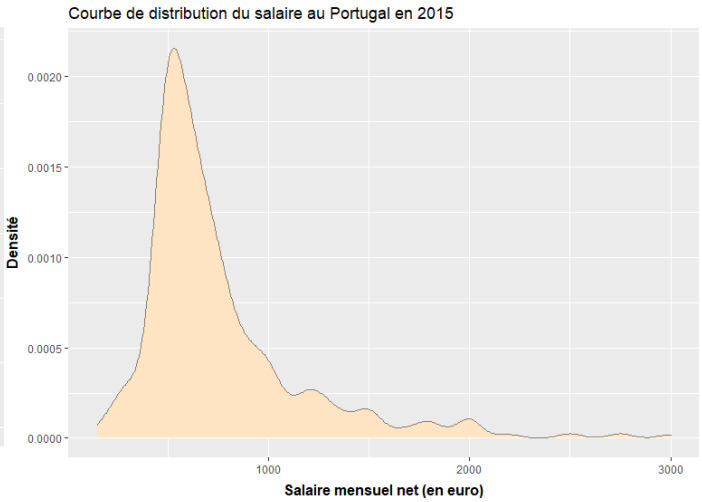
Contexte de l'étude :

Notre étude se base sur les données collectées lors d'enquêtes européennes réalisés par *Eurofound* auprès de travailleurs de différents pays d'Europe en 2015. En nous servant de cette base de données, nous analyserons la situation des travailleurs au Portugal. Nous étudierons la rémunération mensuelle nette de ces derniers, que nous comparerons à celle des autres travailleurs européens. Dans un second temps, nous tenterons de comprendre quels sont les facteurs qui, au Portugal, impactent le salaire d'un travailleur, de façon positive ou négative.

Question 1 :

Avant d'observer la distribution du salaire au Portugal, une rapide observation du salaire en Europe nous permet d'identifier une répartition très inégale. En effet, en Europe, le salaire médian est de 1400€ mensuel net, mais le salaire maximal est de 271140€ mensuel net. Comme nous pouvions nous y attendre, une très faible part des individus observés touche un salaire largement supérieur aux autres travailleurs. Aussi, pour améliorer la lisibilité de nos observations graphiques, nous avons décidé de tronquer notre base en retirant ces individus au salaire aberrant en fixant le salaire mensuel maximal à 6000€. Cela représente 98 individus sur 22514, soit 0,4% de la base. Les analyses se baseront toutefois uniquement sur le jeu de données complet.

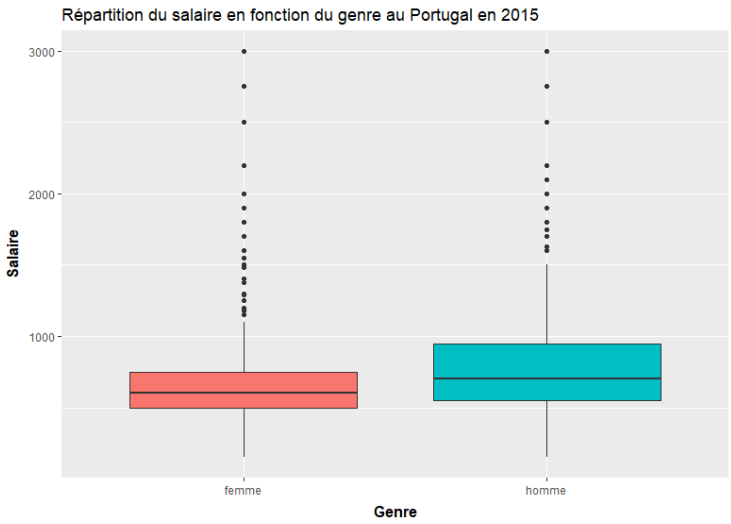


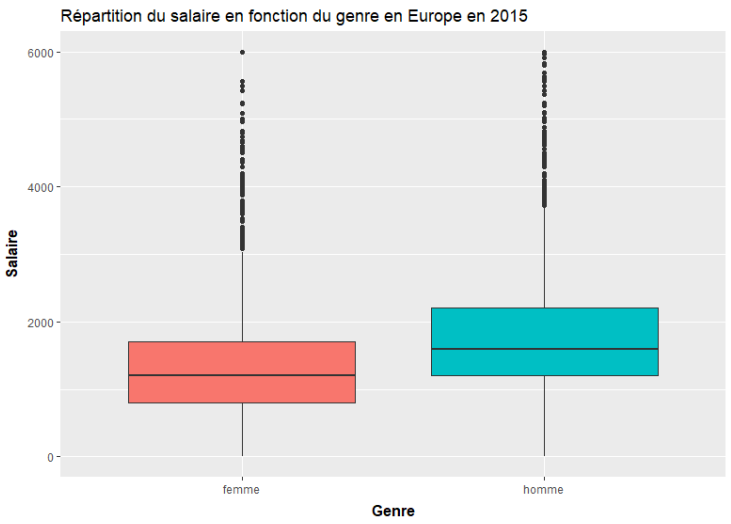


L'observation des courbes de distribution du salaire à l'échelle européenne et portugaise nous permet de constater rapidement que la rémunération au Portugal est bien inférieure à la moyenne en Europe.

L'étude des quartiles nous confirme cette tendance : au Portugal, le salaire au 1er quartile est de 500€, le salaire médian est de 630€ et le salaire au 3ème quartile est de 850€ mensuel net. Cela signifie que 50% des travailleurs gagne entre 500€ et 850€ mensuel, tandis que seul 25% des travailleurs touche entre 850€ et 3000€ mensuel, qui est le salaire maximum. A titre comparatif, rappelons que le salaire médian en Europe est de 1400€ net mensuel, donc 50% des travailleurs en Europe touche plus que cette somme.

Observons maintenant les écarts de salaires entre les hommes et les femmes :





Les boites à moustaches sont révélatrices des inégalités de salaires entre les hommes et les femmes. Au Portugal et en Europe, les femmes demeurent moins bien rémunérées que les hommes.

A première vue au Portugal, le salaire des femmes se situe majoritairement entre 500€ et 750€, alors qu'il se situe entre 550€ et 950€ chez les hommes. Notons qu'en Europe, le salaire au 1er quartile est bien plus bas chez les femmes que chez les hommes, accroissant encore les disparités (bien qu'il reste plus élevé qu'au Portugal).

Enfin, le calcul du salaire médian nous révèle qu'il est de 600€ mensuel pour les femmes et de 700€ pour les hommes au Portugal. Les femmes y sont donc payées 14% moins bien que les hommes. En Europe, le salaire médian d'une femme est de 1200€ contre 1600€ pour un homme. Les femmes y sont donc payées 25% moins bien que les hommes.

Question 2 :

Après avoir constaté les écarts de salaire entre les hommes et les femmes, à l'échelle du Portugal et du reste de l'Europe, intéressons-nous aux différentes caractéristiques des salariés portugais en fonction de leur genre. Pour cela, nous avons réalisé le tableau ci-dessous :

	Hommes	Femmes
taux d'individus	45%	55%
âge moyen	43 ans	43 ans
salaire moyen	828€	704€
salaire médian	700€	600€
ancienneté moyenne	11 ans	11 ans
taux d'emploi en CDI	75%	69%
taux d'emploi en CDD	12%	17%
taux d'emploi en interim	2.2%	1.8%
taux d'emploi en alternance/stage	11%	12%
taux de secteur privé	82%	77%
taux de secteur public	18%	23%
taux de travailleurs étrangers	8.2%	8.5%
taux de satisfaction au travail	85%	84%

Table des caractéristiques des travailleurs portugais par genre

Outre le salaire, les différences notables de notre échantillon, composé à 45% d'hommes et à 55% de femmes, concerne la nature du contrat de travail. En effet, si 75% des salariés hommes sont en CDI, ce n'est le cas que pour 69% des femmes. Ces dernières sont alors plus employées que les hommes en CDD. Les contrats à durée déterminée sont généralement plus précaires et moins bien payés, ce qui pourrait ainsi expliquer les différences de salaire.

De plus, nous pouvons également constater que les salariées portugaises femmes sont plus présentes dans le secteur public (à 23%) que les hommes (à 18%). Il est également possible que ce facteur puisse expliquer les écarts de salaires en fonction du genre. Toutefois, nous ne pouvons l'affirmer sans avoir étudié l'impact de ces variables sur le salaire.

Notons enfin que l'âge moyen des travailleurs portugais, de 43 ans, est identique pour les deux genres. Il en va de même pour l'ancienneté moyenne, de 11 ans, et pour le taux de travailleurs étrangers, avoisinant 8% pour les hommes comme pour les femmes.

Question 3.a. :

Intéressons-nous maintenant aux variables suivantes : l'âge, le genre, la nationalité, la nature du contrat de travail et le secteur d'activité. Afin d'identifier véritablement l'impact de chacune de ces variables sur le salaire des travailleurs portugais ainsi que leur significativité, nous devons réaliser un modèle linéaire. Dans celui-ci, nous indiquons que notre variable à estimer est le salaire et nous observons les coefficients ainsi que l'intervalle de confiance de chaque variable :

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	853.2751729	73.31972	11.6377317	0.0000000	709.413166	997.137180
age	0.3820194	1.05718	0.3613569	0.7179019	-1.692294	2.456333
gender1	-126.4583740	23.03405	-5.4900618	0.0000000	-171.653927	-81.262821
nationality1	-8.3463859	41.90519	-0.1991731	0.8421641	-90.569348	73.876576
CDI1	224.1518460	36.63302	6.1188476	0.0000000	152.273507	296.030185
CDD1	59.0923013	45.74764	1.2917015	0.1967314	-30.670019	148.854622
interim1	-27.8500403	88.21071	-0.3157218	0.7522735	-200.929944	145.229863
Private1	-254.5727644	28.43095	-8.9540715	0.0000000	-310.357673	-198.787856

Table des coefficients et intervalles de confiance du modèle linéaire n°1

Ainsi, nous pouvons déclarer que, toutes choses étant égales par ailleurs :

- Si l'âge augmente de 1 an, le salaire augmente de 0,38€. Mais ce coefficient n'est pas significativement différent de 0 au seuil de 5% : l'intervalle de confiance comprend 0 donc l'estimation est imprécise. On ne sait pas si l'âge a un impact positif ou négatif sur le salaire.
- Pour les femmes, le salaire moyen estimé serait 126€ de moins que celui des hommes. Le coefficient est significatif au seuil de 5%.
- Pour les travailleurs de nationalité portugaise, on estime leur salaire moyen à 8€ de moins que pour les travailleurs étrangers. Ce coefficient n'est pas significativement différent de zéro au seuil de 5%, l'estimation sera donc imprécise.
- Le salaire moyen des travailleurs en CDI est estimé à 224€ de plus que pour les travailleurs en alternance. Le coefficient est significativement différent de zéro au seuil de 5%. Cependant, l'intervalle de confiance, compris entre 152 et 296, n'est pas resserré autour de l'estimation qui sera donc imprécise.
- Le salaire moyen des travailleurs en CDD est estimé à 59€ de plus que pour les travailleurs étant en alternance ou en stage. Mais ce coefficient n'est pas significativement différent de zéro au seuil de 5% et son intervalle de confiance est extrêmement large : l'estimation est donc très imprécise. De plus, remarquons que même la borne inférieure de l'intervalle du coefficient du CDI est supérieure au coefficient du CDD. Un contrat en CDI aura donc plus d'impact qu'un contrat en CDD sur le salaire.
- De même, le salaire moyen des travailleurs en intérim est estimé à 28€ de moins que celui des travailleurs en alternance. Ce coefficient demeure moins impactant que celui des contrats en CDI et en CDD. De plus il n'est pas significativement différent de zéro au seuil de 5% et son intervalle de confiance n'est pas resserré autour de l'estimation. L'estimation de ce coefficient est donc très imprécise.
- Enfin, pour les travailleurs du secteur privé, le salaire moyen serait 255€ de moins que celui des travailleurs du secteur public. Le coefficient est par ailleurs significatif au seuil de 5%.

Globalement, les variables les plus impactantes sur le salaire d'un travailleur portugais sont le genre, les contrats en CDI et le travail ou non dans le secteur privé. Ces trois variables sont

également les seules dont le coefficient est significativement différent de 0 au seuil de 5%. Ce sont donc les seules variables pour lesquelles nous pouvons affirmer avec certitude qu'elles ont un impact positif ou négatif sur le salaire mensuel net.

Question 3.b. :

Ainsi, comme mentionné plus tôt, le genre aurait un impact significatif sur le salaire des travailleurs portugais. Plus précisément, toutes choses égales par ailleurs, on observe une différence moyenne de salaire de 126 euros en défaveur des femmes, significative au seuil de 5%.

On peut alors se demander si ces écarts diminueraient avec l'âge, toutefois, l'interaction entre l'âge et le sexe n'apparaît pas significative au seuil de 5% au vu de la significativité statistique du coefficient du terme d'interaction s'élevant à 0.08.

Le tableau des coefficients suivant permet d'apercevoir la non-interaction de ces deux variables :

	ESTIMATE	**STD. ERROR**	**T VALUE**	**PR(>|T|)**
INTERCEPT	776.343	85.441	9.086	< 2e-16
AGE	2.265	1.508	1.502	0.1334
GENDER	23.623	88.835	0.266	0.7904
NATIONALITY	-9.557	41.872	-0.228	0.8195
PRIVATE	-253.712	28.409	-8.931	< 2e-16
CDI	221.254	36.636	6.039	2.12e-09
CDD	54.547	45.779	1.192	0.2337
INTERIM	-32.137	88.162	-0.365	0.7155
AGE:GENDER	**-3.523**	**2.014**	**-1.749**	**0.0805**

Question 3.c. :

L'équation d'analyse de la variance est la suivante :

$$\sum_{i=1}^{N}(y_i-\bar{y})^2 = \sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{N}\hat{\varepsilon}_i^{\,2}$$

$$SCT = SCE + SCR$$

Avec SCT étant la somme des carrés totale, SCE la somme des carrés expliquée et SCR la somme des carrés des résidus.

Plus la variance expliquée est proche de la variance totale (c'est-à-dire la variance résiduelle est faible), meilleure est la régression.

Après calcul, nous obtenons une variance résiduelle s'élevant à 158649988, une variance expliquée de 26772196 et une variance totale s'élevant à 185422184. La part de variance résiduelle est donc de 85.5%. On peut alors s'interroger sur la qualité de notre régression.

Question 3.d. :

Pour analyser la qualité globale du modèle, nous avons décidé de nous concentrer sur deux indicateurs :

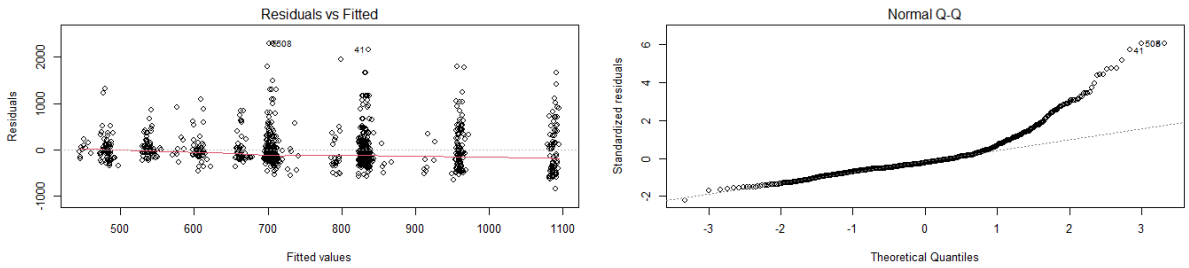
- Le R^2, mesurant le rapport entre la variance expliquée par le modèle et la variance totale,
- Et la p-value associée à la statistique de Fisher (F-test).

Notre modèle explique 14,4% de la variance des salaires, avec un R^2 ajusté à 13,9%. Il serait donc intéressant de le maximiser après avoir effectué l'analyse des résidus.

Aussi, la p-value associée à la statistique de Fisher étant inférieure à 1, nous pouvons rejeter l'hypothèse d'absence de significativité globale des variables. Ainsi, au moins une variable n'est pas significativement différente de 0, et le modèle est globalement significatif.

Question 3.e. :

Pour réaliser un diagnostic plus précis de notre régression, nous pouvons d'abord visualiser la différence entre les salaires calculés et les salaires réels, et voir si les résidus suivent une loi normale :



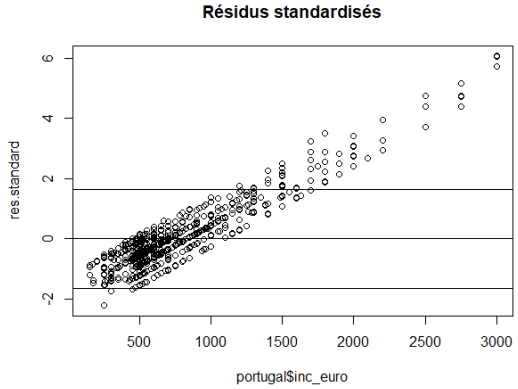
On remarque d'abord que certaines observations sont très mal reconstituées par le modèle, notamment les individus 508 et 41, avec des écarts entre les valeurs estimées et les valeurs réelles très importants (près de 2000euros de différence). Aussi, on observe que les résidus ne suivent pas une loi normale, avec une courbe s'éloignant de plus en plus de la diagonale à l'extrémité droite. Cela indique qu'il existerait des points atypiques dans notre jeu de données. On retrouve par ailleurs encore une fois sur cette deuxième visualisation les individus 508 et 41. Ces derniers semblent se distinguer par un salaire plus élevé, s'élevant pour les deux à 3000 euros mensuels. On peut donc imaginer qu'un salaire bien plus élevé que la moyenne pourrait avoir impacté le modèle.

Pour compléter ces informations, nous pouvons nous baser sur trois indicateurs supplémentaires :

- Les résidus standardisés, normalisés par l'écart-type, qui vont permettre d'identifier les points mal reconstitués par le modèle qui reflètent les observations atypiques;
- Les résidus studentisés qui vont quant à eux permettre d'identifier les points influençant fortement la régression ;
- Et les mesures leviers qui permet également d'identifier les indicateurs d'influence.

Le graphique des résidus standardisés suivant permet d'appréhender que certaines observations sortent des tuyaux, notamment lorsque les salaires excèdent 1500 euros mensuels. Aussi, plus les salaires sont élevés, plus l'éloignement est important. Ces dernières concernent près de 7% du jeu de données des travailleurs portugais.

Effectivement, le tableau de répartition des salaires ci-dessous comparant les salaires des observations sortant des tuyaux avec ceux de l'ensemble de la base permet de confirmer la différence conséquente de salaire :



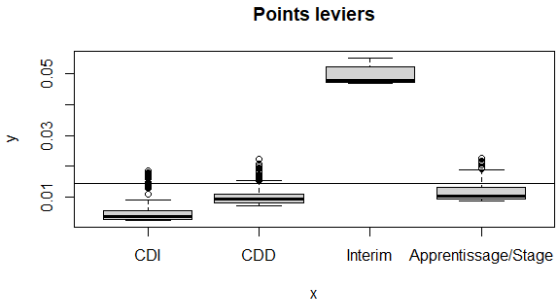
BASE	MIN	Q1	MEDIANE	MOYENNE	Q3	MAX
OBSERVATIONS ATYPIQUES – RESIDUS STANDARDISES	250	1500	1800	1847	2000	3000
ENSEMBLE DE LA BASE (PORTUGAL)	150	500	630	759	850	3000

On peut donc supposer que le modèle peine à reconstituer le salaire des travailleurs portugais les plus aisés.

L'analyse des résidus studentisés indique également que les observations impactant fortement le modèle concernent les travailleurs portugais les plus aisés. Effectivement, les observations retenues comme influençant la régression sont exactement identiques à celles retenues lors de l'analyse des résidus standardisés.

La mesure des points leviers apporte de nouvelles informations quant aux individus impactant fortement la régression. Cette mesure indique que les points influents ne concernent pas particulièrement les travailleurs ayant les salaires les plus élevés, mais plutôt les travailleurs n'étant pas en CDI, et même plus précisément ceux qui sont en intérim. Effectivement, si dans le jeu de données initial des travailleurs portugais ces derniers représentent 2% des travailleurs, ils représentent 24% des points leviers identifiés.

Les graphique et tableau suivant permettent d'appréhender ce phénomène :



BASE	CDI	CDD	INTERIM	APPRENTISSAGE / STAGE
POINTS LEVIERS	13%	32%	24%	30%
ENSEMBLE DE LA BASE (PORTUGAL)	72%	15%	2%	11%

Question 3.f. :

L'analyse des résidus ayant identifié 162 travailleurs portugais comme étant des points influents ou comme ayant des valeurs atypiques (notamment les travailleurs étant contrat d'intérim ou ayant un salaire bien plus élevé que la moyenne des travailleurs portugais), nous avons décidé de les exclure pour effectuer une nouvelle régression et donc améliorer la qualité du modèle. Aussi, nous avons décidé d'inclure la variable relative à l'expérience, qui pour l'instant avait été écartée pour la construction du modèle, étant donné qu'elle pourrait impacter fortement le revenu.

Nous obtenons alors les indicateurs suivant pour les deux modèles :

Modèle	*R^2*	*R^2 adj.*	*Sigma*	*Stat.*	*P.value*	*logLik*	*AIC*	*BIC*	*Deviance*
Initial	0.144	0.139	379	26.6	8.52^{e-34}	-8163	16344	16389	158649988
Final	0.22	0.21	227	37.9	5.69^{e-47}	-6484	12987	13031	4848600

Ainsi, les indicateurs à maximiser (R^2, R^2 ajusté, Statistique de Fisher) augmentent avec le nouveau modèle, et les indicateurs à minimiser (Sigma, vraisemblance, AIC, BIC, déviance) diminuent.

Le modèle final apparaît donc comme étant meilleur que le modèle initial.

Pour améliorer encore la qualité du modèle, on pourrait également ajouter des informations concernant le niveau d'étude, le secteur d'activité ou encore le lieu de l'emploi.

Question 4. :

Pour un même modèle, une seule variable devient significative. En effet, on a le genre, le fait d'avoir un CDI et le secteur d'appartenance qui est significatif pour notre modèle appliqué sur les données du Portugal. De plus, comme nous l'avons déjà dit, nous avons la variable Age,

nationalité, CDD et alternance qui possède la valeur 0 compris dans leur intervalle de confiance à 95%.

Lorsque l'on passe sur l'ensemble des pays, il y a plusieurs variables qui deviennent significatives en plus des autres déjà significatives sur les données du Portugal. Il y a la variable Age, si l'âge augmente de 1 an, le salaire augmente de 10€. La variable nationalité est presque significative avec une p-value de 9%. La fait d'être en intérim ou en CDD devient aussi significatives quant à l'explication du salaire sur l'ensemble du pays.

En effet, malgré la même significativité pour les variables CDI, genre et le secteur d'appartenance, les coefficients ne sont pas similaires. Pour la variable genre, le coefficient est presque 4 fois plus grand que pour les données du Portugal : -126 → -505. Sur l'ensemble des pays, le fait d'être une femme multiplie par 4 l'écart de salaire moyen entre les hommes et les femmes par rapport aux données du Portugal. Pour la variable binaire CDI, le coefficient est multiplié par 3 : 224 → 774. Sur l'ensemble des pays le fait d'être CDI a donc un plus grand impact sur le salaire par rapport au Portugal. Pour le secteur d'appartenance, il y a un écart de 100 €. Sur les données du Portugal, le coefficient est de -254 et on passe à -328 sur l'ensemble des pays. Le fait de travailler dans un secteur privé baisserai de 100 euros le salaire moyen sur l'ensemble des pays rapport au Portugal.

Lorsque l'on regarde la variance expliquée, il y a aussi un écart non négligeables. La variance expliquée du modèle sur le Portugal est de 14% alors que la variance expliquée du modèle sur l'ensemble des pays est de 3%. Le modèle sur le Portugal arrive à avoir 11% de plus de part de variance expliqué par rapport à l'ensemble des pays. On a donc le modèle sur le Portugal qui arrive à mieux expliquer le salaire par rapport au modèle sur le Portugal. On peut expliquer cette différence par la différence de ligne entre les données sur le Portugal et les donnée sur l'ensemble des pays. En effet, les donnée du Portugal possède 1 110 lignes alors que les donnée sur l'ensemble des pays possède 22 514 lignes.

On peut donc dire que les résultats ne sont pas convergés car les variables significatives ne sont pas les mêmes entre les deux modèles. Lorsque les variables sont significatives sur les deux modèles, l'écart entre les coefficients est non négligeable. La différence sur l'impact sur le salaire sera donc plus visible.

Question 5. :

Comme nous l'avons déjà vu, on a un écart de 11% sur le R^2 entre les deux modèles. L'AIC permet de valider aussi cet écart car l'AIC pour l'ensemble des pays est de 312 340 alors que l'AIC sur le Portugal est de 12 228, soit 25 fois plus petit. La déviance est aussi très révélatrice sur la différence de qualité de régression entre les deux modèles. La déviance pour l'ensemble des pays est 10^3 plus grande que la déviance pour le Portugal. De plus la vraisemblance entre les deux modèles est aussi très différente. On voit que la vraisemblance pour le Portugal est 26 fois plus petite que la vraisemblance pour l'ensemble des pays.

On peut donc dire que la qualité du modèle sur l'ensemble des pays est moins bonne que la qualité du modèle sur le Portugal car les indicateurs de qualité sont nettement meilleurs pour le modèle sur le Portugal. L'explication de cette différence est probablement encore la différence de nombre de lignes. En effet, le fait d'avoir plus de lignes pour les données sur l'ensemble devient un avantage. Le modèle possède plus de ligne pour s'entrainer et réaliser une meilleur prédiction.