

Décès du cancer

Dans les pays de l'OCDE

Introduction

La Hongrie est le pays membre de l'OCDE où les décès dus au cancer sont les plus fréquents. L'OMS estime que ces décès sont liés à une consommation dangereuse de tabac, d'alcool, ainsi que des infrastructures de santé insuffisantes.

Il s'agit de vérifier l'existence d'un lien entre ces facteurs et les décès dus au cancer. Pour cela, nous construisons une équation de régression linéaire en utilisant les données de santé de l'OCDE.

I. Les données

Nous cherchons à expliquer les décès par cause de cancer par un ensemble de variables de santé. Pour cela, nous constituons une base de données à partir des données de l'OCDE que nous complétons par les données du PIB de la Banque Mondiale, ainsi que celles sur la consommation de tabac de l'*Institute for Health Metrics and Evaluation*.

Le tableau suivant donne la liste des variables ainsi que leur unité :

Décès dus au cancer	Espérance de vie (0)	Fumeurs quotidiens	Consommation d'éthanol
Pour 100 000 habitants	Années	%	Litres nets par habitant
Lits d'hôpitaux	Médecins	Dépenses de santé	PIB par habitant
Pour 1 000 habitants	Pour 1 000 habitants	USD par habitant	USD PPA constants

Nous allons d'abord chercher un lien linéaire entre les morts du cancer et les variables explicatives. Nous utilisons le test de corrélation de Spearman. Le tableau suivant donne l'indice de corrélation de Spearman ainsi que la p-value de ce test entre les morts du cancer et les autres variables :

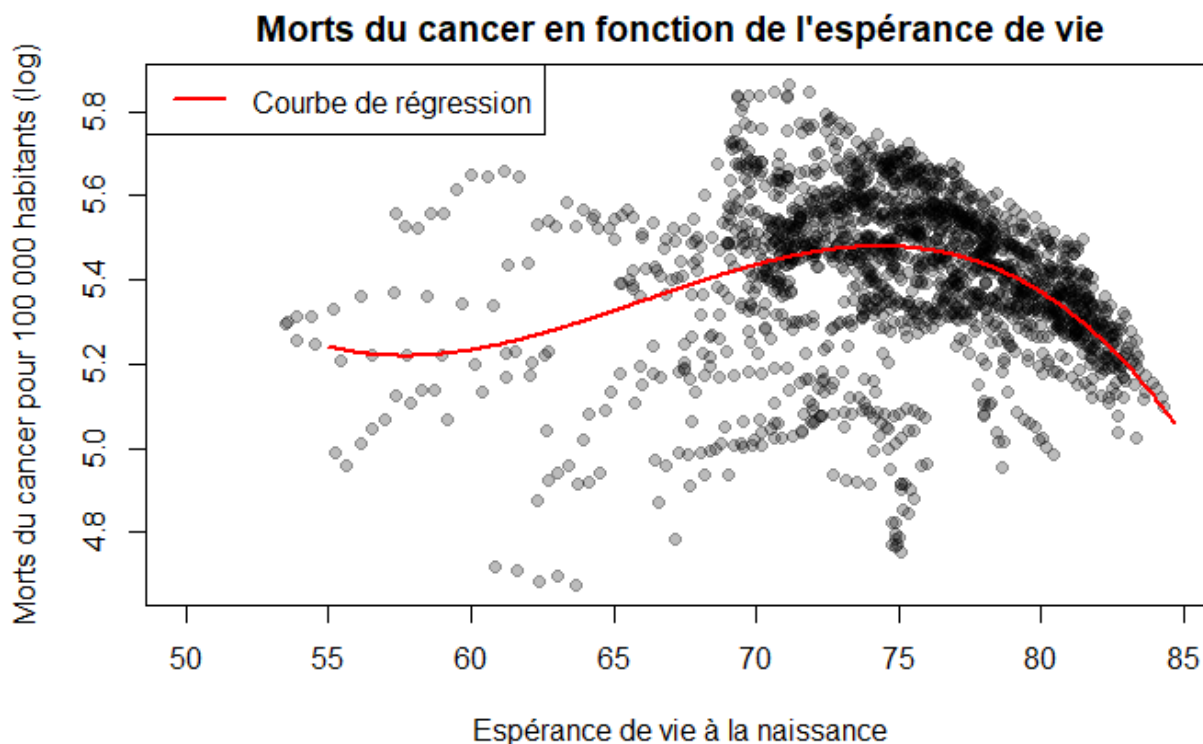
Variable explicative	Corrélation avec les morts du cancer	p-value
Alcool	0,62	0,00
Fumeurs	0,46	0,00
Lits d'hôpitaux	0,39	0,01
PIB	0,17	0,29
Médecins	0,18	0,30
Dépenses de santé	0,15	0,34
Espérance de vie	-0,11	0,51
CT-Scanner	-0,01	0,96

Nous choisissons un niveau α de 5 %. Nous voyons qu'il y a, comme attendu, une corrélation positive fortement significative entre morts du cancer et consommations d'alcool et de tabac. et entre morts du cancer et les lits d'hôpitaux. Il y a de plus une corrélation positive significative avec le nombre de lits d'hôpitaux.

En revanche, il n'y a pas de corrélation significative avec le PIB, le nombre de médecins, les dépenses de santé, l'espérance de vie à la naissance, et les CT-Scanners. Nous nous

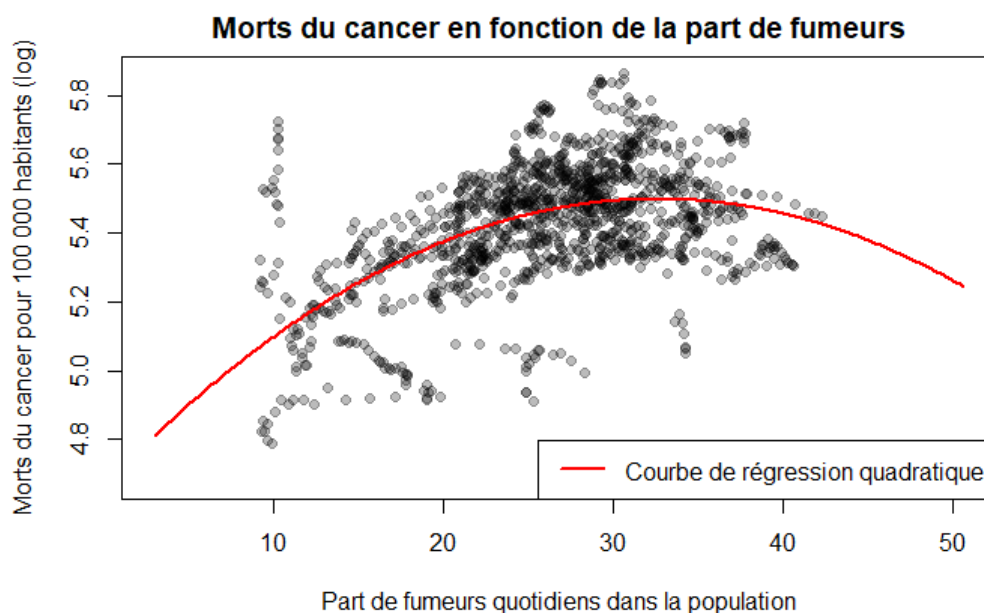
pourtant attendons à l'existence d'un lien entre espérance de vie et morts du cancer, le cancer étant une maladie qui touche surtout les personnes les plus âgées.

Pour tenter de trouver un lien entre morts du cancer et espérance de vie, nous représentons le nuage de points correspondant :



La courbe rouge correspond à la courbe de régression polynomiale de degré 3 des morts du cancer en fonction de l'espérance de vie. On voit qu'il semble bien exister un lien entre les deux variables, mais qu'il n'est pas linéaire, et donc pas détectable par une simple corrélation de Spearman. Le lien semble positif quand l'espérance de vie est inférieure à 75 environ, mais négatif quand elle surpasse ce niveau.

De manière similaire, il semble exister une relation en U inversé entre les morts du cancer et la part de fumeurs quotidiens dans la population :



Nous intégrerons donc le taux de fumeurs au carré dans notre équation de régression.

Nous allons désormais étudier les corrélations entre les variables. Nous utilisons le même test de corrélation que précédemment. Le résultat est cette matrice de corrélation :

Corrélation	Alcool	CT-Scanners	Fumeurs	PIB	Dépenses	Lits	Espérance
CT-Scanners	0,12						
Fumeurs	0,51	0,01					
PIB	0,35	0,47	-0,05				
Dépenses	0,34	0,41	-0,19	0,97			
Lits	0,70	0,33	0,49	0,31	0,28		
Espérance	0,06	0,30	-0,15	0,75	0,75	0,14	
Médecins	0,49	0,31	0,11	0,47	0,47	0,41	0,37

Comme attendu, les dépenses de santé par habitant sont très fortement corrélées au PIB par habitant. De plus, PIB et dépenses de santé sont corrélées très fortement à l'espérance de vie. Nous devons donc choisir l'une de ces trois variables pour représenter le niveau de vie des habitants. Nous choisissons de garder le PIB car les dépenses ne sont pas en USD PPA constants.

Nous allons diviser les données entre les pays d'Europe de l'est et les autres. En effet, les pays d'Europe de l'est ont des chiffres de décès du cancer plus élevés :

Médiane autres	Médiane Europe de l'est
226.70	242,25

Nous allons donc d'abord comparer les médianes des variables explicatives entre les pays d'Europe de l'est et les autres.

Pour cela, nous créons la variable Europe de l'est, qui vaut 1 si le pays vient de cette région, et 0 sinon. Les pays d'Europe de l'est sont les pays du continent Europe se situant à l'est de l'Allemagne, l'Autriche, l'Italie et de la Finlande. Cela inclut les pays des Balkans et la Russie.

Nous divisons nos données selon cette variable et nous appliquons le test de Wilcoxon-Mann-Whitney (fonction *wilcox.test* sur R) entre les deux jeux de données ainsi créés. Le résultat du test pour chaque variable explicative est le tableau suivant :

Row.names	Médiane autre	Médiane est	Rapport	p-value
Morts du cancer	226,70	242,25	1,07	0,00
Fumeurs	16,50	28,20	1,71	0,00
Lits d'hôpitaux	4,12	7,16	1,74	0,00
Alcool	8,60	11,00	1,28	0,00
Médecins	2,29	3,03	1,32	0,00
Espérance de vie	74,95	71,50	0,95	0,00
PIB	19 572,68	15 858,12	0,81	0,00
CT-Scanner	16,60	13,82	0,83	0,00
Dépenses de santé	1 400,48	1 399,26	1,00	0,44

La p-value correspond à la p-value du test. Nous trouvons une médiane significativement différente pour toutes les variables sauf pour les dépenses de santé.

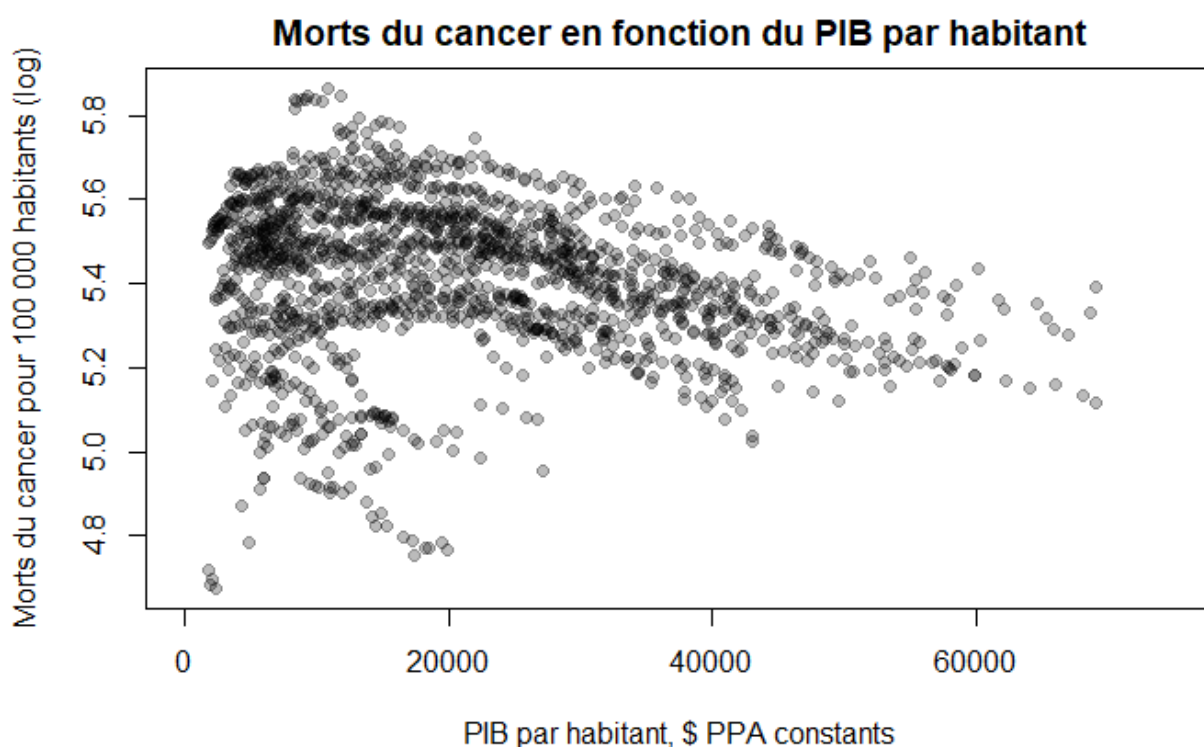
En particulier, le test confirme la différence de médiane pour les morts du cancer, Les consommations de tabac et d'alcool sont également quasiment deux fois plus élevées chez les pays de l'est que chez les autres. L'espérance de vie, le PIB, et les équipements en CT-scanners y sont significativement inférieurs. Toutefois, les pays de l'est ont beaucoup plus de lits d'hôpitaux et de médecins pour 1 000 habitants.

II. Régression linéaire et tests

Nous devons choisir les variables à intégrer à notre régression linéaire. Les résultats des comparaisons de médianes entre les pays de l'est et les autres nous mène à introduire la variable Europe de l'est à l'équation. Nous choisissons d'étudier l'année 2012, car c'est l'année pour laquelle nous avons le plus de données pour le plus de pays.

Pour mesurer le niveau de vie, on a vu que nous retenons le PIB par habitant. Nous représentons l'importance des infrastructures médicales par les lits d'hôpitaux, car nous n'avons sur l'année 2012 pas assez de données sur les médecins pour 1 000 habitants ou sur les CT-Scanners.

De manière similaire à l'espérance de vie, le nuage de points entre morts du cancer et PIB par habitant prend la forme d'un U inversé. Nous intégrons donc à notre équation de régression le PIB au carré.



Pour détecter des différences de structures entre les pays de l'est et les autres, nous intégrons des variables d'interaction entre la variable Europe de l'est et les autres variables explicatives.

Finalement, notre première équation de régression est la suivante :

$$\ln(Morts_i) = \beta_1 + \gamma_1 Est_i + \beta_2 \ln(Alcool_i) + \beta_3 Fumeurs_i + \beta_4 Fumeurs_i^2 + \beta_5 \ln(Lits_i) + \beta_6 PIB_i + \beta_7 PIB_i^2 + \sum_{k=2}^7 \gamma_k Est_i \cdot Variable_{ki} + \epsilon_i$$

Nous disposons de 39 observations complètes sur l'année 2012 avec ces variables, dont 10 en Europe de l'est. Nous trouvons, pour cette équation, les résultats suivants :

term	estimate	std.error	statistic	p.value
(Intercept)	3,98828204145	0,2263	17,6266	0,00***
est	2,41771410148	4,2470	0,5693	0,57
log(Alcool)	0,11430363807	0,0395	2,8919	0,01**
est*log(Alcool)	-0,41782879676	0,2669	-1,5653	0,13
Fumeurs	0,08638407598	0,0264	3,2770	0,00***
Fumeurs ²	-0,00186323019	0,0006	-2,9113	0,01**
est*Fumeurs	-0,05783087082	0,1859	-0,3110	0,76
est*Fumeurs ²	0,00087608997	0,0032	0,2728	0,79
log(Lits)	-0,10606642987	0,0380	-2,7880	0,01**
est*log(Lits)	0,22202950246	0,2042	1,0873	0,29
PIB	0,00000953118	0,0000	2,5170	0,02*
PIB ²	-0,00000000008	0,0000	-2,2667	0,03*
est*PIB	-0,00002490678	0,0002	-0,1101	0,91
est*PIB ²	-0,00000000003	0,0000	-0,0062	1,00

Nous trouvons bien un lien entre les morts du cancer et les facteurs que nous avons énoncés :

- L'alcool : le coefficient de $\log(\text{Alcool})$ est significatif et positif.
- Le tabac : les coefficients Fumeurs et Fumeurs^2 sont significatifs et dessinent bien une relation en U inversé car Fumeurs^2 est négatif.
- Les dépenses dans la santé : les coefficients représentant les dépenses dans la santé sont les coefficients du PIB et du PIB^2 (cf. corrélation entre PIB et dépenses dans la santé). Là encore on retrouve bien une relation en U inversé significative.

De plus, les équipements de santé, c'est-à-dire les lits d'hôpitaux réduisent les morts du cancer significativement.

Cependant, les coefficients relatifs à l'Europe de l'est sont tous non-significatifs. Nous effectuons un test de Chow pour vérifier l'existence d'une différence de structure de l'équation selon l'appartenance à l'Europe de l'est.

Pour cela, nous créons un modèle sans les variables d'interaction et nous effectuons un test des modèles emboîtés entre ce modèle et le modèle déjà estimé. Nous testons :

$$H_0 : \gamma_2 = \gamma_3 = \dots = \gamma_7 = 0$$

vs.

$$H_1 : \exists i / \gamma_i \neq 0$$

La fonction *anova* de R nous donne une p-value de $0,2526 > \alpha = 0,05$. Nous rejetons donc H_0 , il n'y a pas de différence significative entre les deux modèles. Nous utiliserons donc maintenant le modèle sans variable d'interaction :

$$\ln(\text{Morts}_i) = \beta_1 + \gamma_1 \text{Est}_i + \beta_2 \ln(\text{Alcool}_i) + \beta_3 \text{Fumeurs}_i + \beta_4 \text{Fumeurs}_i^2 + \beta_5 \ln(\text{Lits}_i) + \beta_6 \text{PIB}_i + \beta_7 \text{PIB}_i^2 + \epsilon_i$$

Cette équation nous donne les coefficients suivants :

term	estimate	std.error	statistic	p.value	
(Intercept)	4,0732	0,1595	25,5434	0,00	***
est	0,2485	0,0454	5,4665	0,00	***
$\log(\text{Alcool})$	0,1007	0,0394	2,5531	0,02	*
Fumeurs	0,0787	0,0165	4,7833	0,00	***
Fumeurs^2	-0,0017	0,0004	-4,5471	0,00	***
$\log(\text{Lits})$	-0,0990	0,0379	-2,6131	0,01	*
PIB	9,42E-06	0,0000	2,6659	0,01	*
PIB^2	-7,37E-11	0,0000	-2,3186	0,03	*

Les coefficients sont tous significatifs à un niveau $\alpha = 0,05$. On retrouve les mêmes relations significatives que dans le modèle initial. Le R^2 ajusté vaut 0,7615, ce qui est fort. Cette fois, cependant, la variable Europe de l'est est fortement positive et significative. Les pays d'Europe de l'est ont donc bien un taux de décès du cancer disproportionné par rapport aux autres pays de l'OCDE.

Nous allons maintenant considérer le modèle constitué uniquement des infrastructures et dépenses de santé. Les coefficients sont les suivants :

term	estimate	std.error	statistic	p.value	
(Intercept)	5,02220485451	0,1095	45,8671	0,00	***
log(Lits)	0,14394685324	0,0428	3,3660	0,00	**
PIB	0,00000368568	0,0000	0,7221	0,48	
PIB ²	-0,000000000004	0,0000	-0,7113	0,48	

Nous réalisons un test des modèles emboîtés entre ce modèle et le modèle sans interactions. Le principe est similaire au test que nous avons réalisé pour rejeter le modèle avec interactions. Cette fois, nous testons la nullité des coefficients correspondant aux lits et au PIB. La p-value du test est de l'ordre du 100-millionième, donc on rejette H_1 et on conserve les coefficients testés dans l'équation de régression.

Nous vérifions la troisième hypothèse de la régression linéaire avec un test de Shapiro-Wilk :

H_0 : les résidus suivent une loi normale

vs.

H_1 : les résidus ne suivent pas une loi normale

La fonction *shapiro.test* de *r* appliquée aux erreurs observées de notre équation sans interactions nous donne une p-value de $0,1846 > 0,05 = \alpha$. Nous ne rejetons donc pas H_0 , et considérons que les résidus sont normaux.

Nous avons donc une équation sans interactions très significative qui nous donne une relation en U inversé entre les morts du cancer et les dépenses de santé, de même pour le nombre de fumeurs quotidiens. Cette dernière relation est surprenante, la cigarette étant un très important facteur de développement de cancers. Les coefficients suggèrent que la part de fumeur augmente le nombre de décès du cancer jusqu'à une part de 23 % de la population adulte, et que le taux baisse à partir d'un PIB de 64 000 USD PPA constants par habitant.

L'alcool a lui aussi une influence positive sur les morts du cancer. On trouve aussi que l'équipement hospitalier permet de réduire les morts du cancer. Enfin, les pays d'Europe de l'est ont des morts significativement supérieures.

Notre équation est plutôt robuste, avec un R^2 ajusté de plus de 0,75 et des résidus normaux.

Conclusion

Nous cherchions à confirmer l'existence d'un lien entre le taux de décès du cancer et des facteurs de risques liés aux mauvaises habitudes, et à trouver l'importance des infrastructures pour diminuer les risques.

Nous avons trouvé que l'alcool augmente le taux de décès du cancer, et que le tabac l'augmente jusqu'à une part critique de 23 % de la population adulte fumant quotidiennement. L'équipement de santé permet de réduire les risques, et les pays les plus

riches sont ceux qui ont le moins de morts du cancer, avec une relation en U inversé. Cette dernière relation pourrait venir d'un certain nombre de pays ayant une bonne espérance de vie à la naissance mais des infrastructures de santé insuffisantes par manque de richesse.

Les pays d'Europe de l'est ont un taux de mort largement plus élevé, mais notre modèle ne présente pas de changement de structure entre les pays. Si notre modèle est robuste et respecte les hypothèses de la régression linéaire, nous pourrions étudier un modèle avec plus d'observations et/ou plus de variables explicatives pour trouver la raison de cette différence géographique.

Pour conclure, nos résultats sont cohérents avec les facteurs cités par l'OMS pour expliquer le risque élevé lié au cancer en Hongrie.

Bibliographie

Document

OECD/European Observatory on Health Systems and Policies (2017), Hungary: Country Health Profile 2017, State of Health in the EU, OECD Publishing, Paris/European Observatory on Health Systems and Policies, Brussels. <http://dx.doi.org/10.1787/9789264283411-en>

Données

- Données sur la santé :

OECD data. <https://data.oecd.org/health.htm>, consulté le 05/04/2022.

- Données sur le niveau de vie :

Banque mondiale. <https://donnees.banquemondiale.org/indicateur/NY.GDP.PCAP.PP.CD>, consulté le 05/04/2022.

Institute for Health Metrics and Evaluation (via Our World in Data). <https://ourworldindata.org/smoking>, consulté le 05/04/2022.