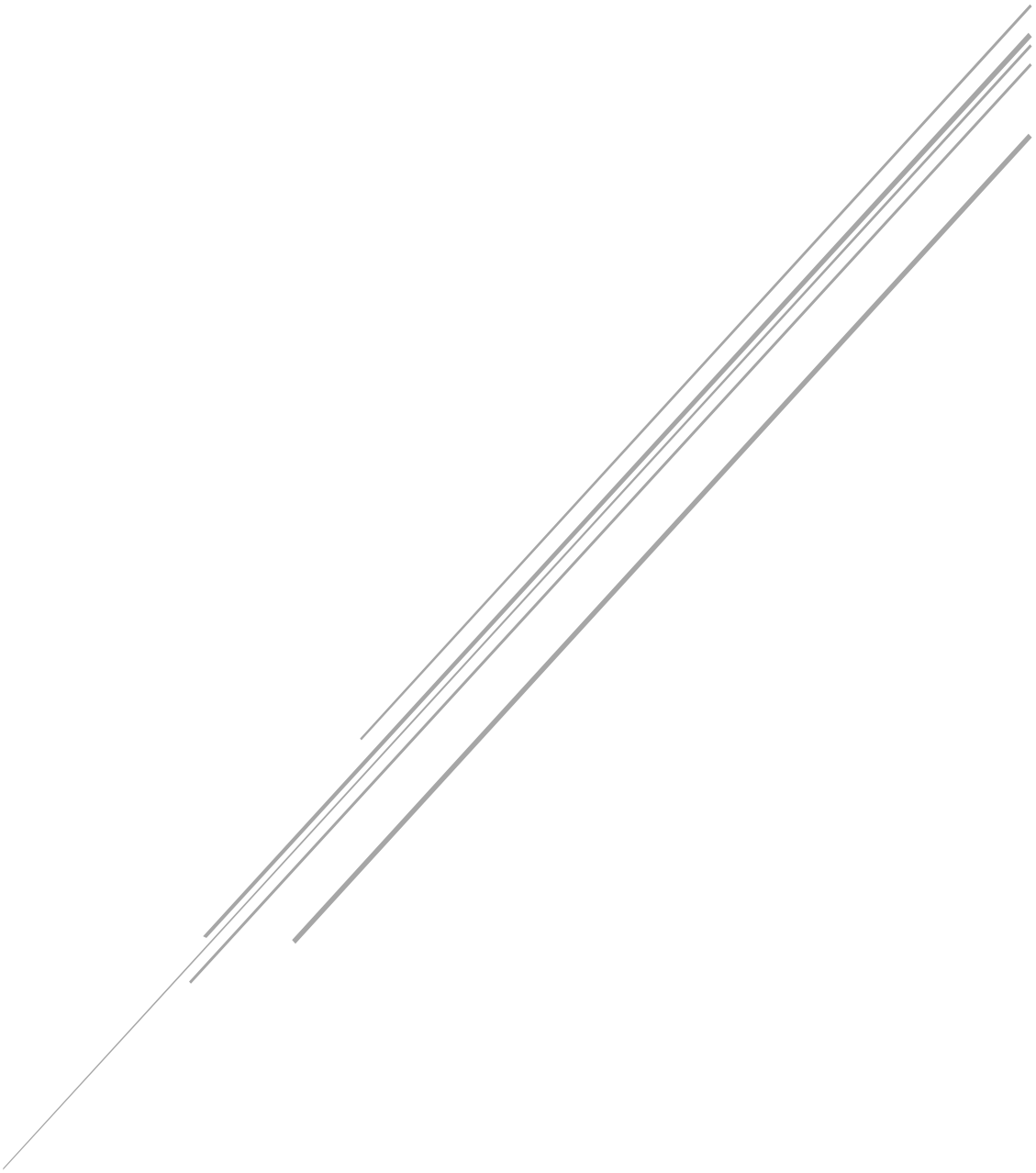
PROJET D'ECONOMETRIE

Quel est l'impact des facteurs de risque sur le nombre de décès suite à un cancer en France ?



FAUJOUR Clara
GUIBERT Marie
PAILLARD Loevane

Aujourd'hui, le cancer est l'une des causes de mortalité les plus importantes au monde. En 2018, selon les estimations publiées par le Centre international de recherche sur le cancer, les nouveaux cas de cancers seraient au nombre de 18.1 millions dans le monde. En France, sur cette même période, 182 000 nouveaux cas ont été détectés et 157 400 décès ont été recensés. Cette maladie est une prolifération de cellules malignes qui détruisent le système immunitaire. Dans la majeure partie des cas, les organes touchés sont les poumons, les seins ou encore le pancréas.
Une tumeur peut apparaître suite à de fortes expositions à des facteurs de risque. Ainsi, nous analyserons l'impact de certains facteurs sur les décès, en France, liés à cette pathologie. Nous avons décidé d'étudier le volume des dépenses de santé, la consommation d'alcool par habitant, le taux de personnes en situation d'obésité et enfin le taux de fumeurs. Nous avons modélisé cette étude de la façon suivante :

$$Y_{décès} = X1_{santé} + X2_{alcool} + X3_{obésité} + X4_{fumeurs}$$

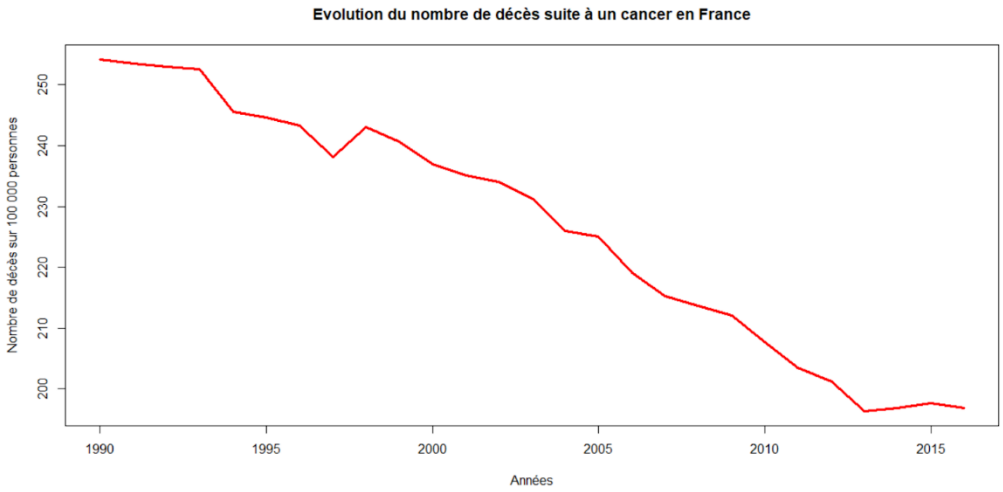
Dans un premier temps, nous décrirons nos données et effectuerons des statistiques descriptives afin de nous orienter au mieux dans le choix de notre modèle. Dans un second temps, nous présenterons notre démarche et mettrons en place différents modèles afin d'en déterminer le plus approprié. Enfin, après analyse de l'ensemble de nos résultats, nous réaliserons des tests pour approfondir notre étude.

Tout d'abord, nous avons analysé l'ensemble de nos variables quantitatives indépendamment les unes des autres. En prenant nos données sur le site de l'OCDE, nous avons pu étudier le nombre de cancers pour 100 000 personnes entre 1990 et 2016, en France. Afin d'expliquer ces données nous avons choisi quatre variables explicatives. L'étude de notre variable endogène est passée par l'analyse des dépenses de santé (en dollars US courants par habitant), de la consommation d'alcool (en litres par habitant), de la part de la population en situation de surpoids ou d'obésité et enfin du taux de fumeurs en France.
Pour pouvoir réaliser notre projet efficacement, nous avons choisi d'effectuer des modifications sur notre base de données. Le passage en dollars constants des dépenses de santé nous a semblé primordial afin de conclure de façon cohérente sur l'évolution de nos données sans prendre en compte l'inflation. Ainsi, nous avons utilisé le déflateur pour passer les observations en dollars US constants. Aussi, certaines de nos variables explicatives comportaient des sauts sur quelques années, nous avons alors réalisé des interpolations linéaires pour combler ces écarts et avoir des données sur l'ensemble de la période étudiée.

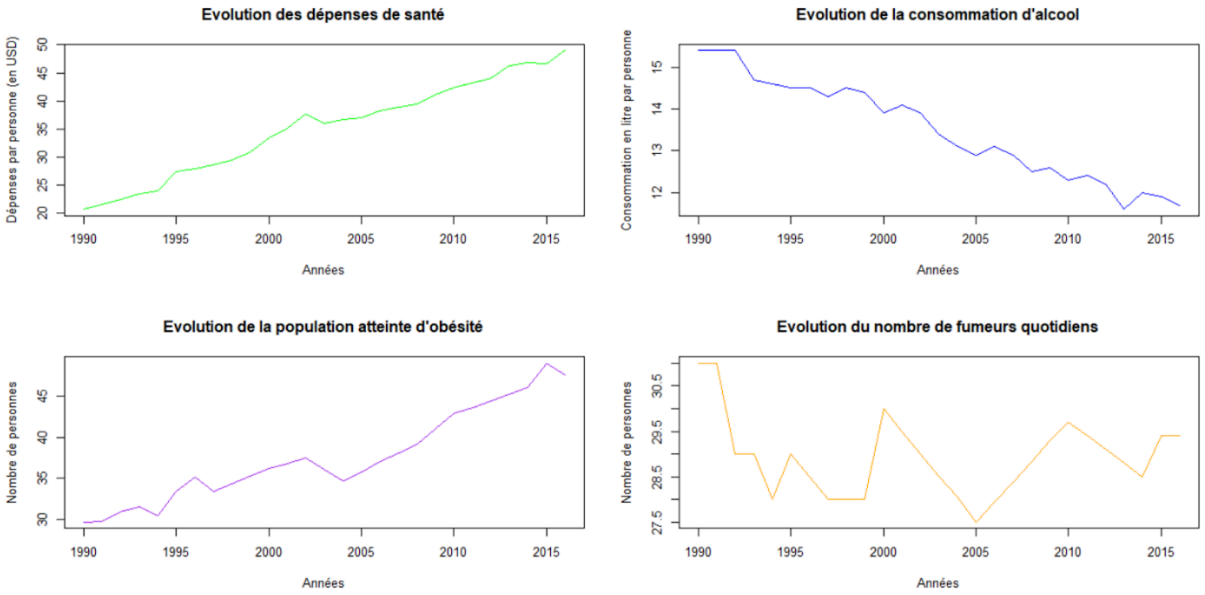
Notre base de données est constituée de six composantes avec des unités diverses impliquant donc des interprétations différentes. Par exemple, en 1990, l'OCDE comptabilise environ 254 décès liés au cancer pour 100 000 habitants tandis qu'en 1995, environ 245 sont recensés. Concernant le taux d'obésité, entre 1990 et 1995, nous observons une hausse de 3.7 points de pourcentage. Par ailleurs, le taux de fumeurs atteint les 31% en 1990. Cette hétérogénéité de mesures est à prendre en compte pour la suite de notre étude. Voici les premières lignes de notre base de données :

	Date	Trend	Y_deces	X1_sante	X2_alcool	X3_obesite	X4_fumeurs
1	1990	1	254.2	20.70461	15.4	29.7	31
2	1991	2	253.5	21.55808	15.4	29.8	31
3	1992	3	253.0	22.40609	15.4	30.9	29
4	1993	4	252.6	23.41486	14.7	31.5	29
5	1994	5	245.5	24.04168	14.6	30.5	28
6	1995	6	244.6	27.48770	14.5	33.4	29

Puis, nous avons représenté graphiquement nos variables afin de visualiser au mieux leur évolution dans le temps. La courbe ci-dessous nous montre la tendance du nombre de décès. Effectivement, celle-ci diminue globalement depuis 1990. Aussi, nous remarquons quelques fluctuations en 1998 et en 2015. Par exemple, en 1997, nous observons 238 décès alors qu'en 1998, sur 100 000 personnes, 240 sont décédées d'un cancer.



Comme constaté sur les premières lignes de notre base de données, la consommation d'alcool tend à la baisse alors que les autres variables présentent une augmentation au fil du temps. Seul le taux de fumeurs n'exprime pas de tendance claire. Ces affirmations sont confirmées par les graphiques suivants :

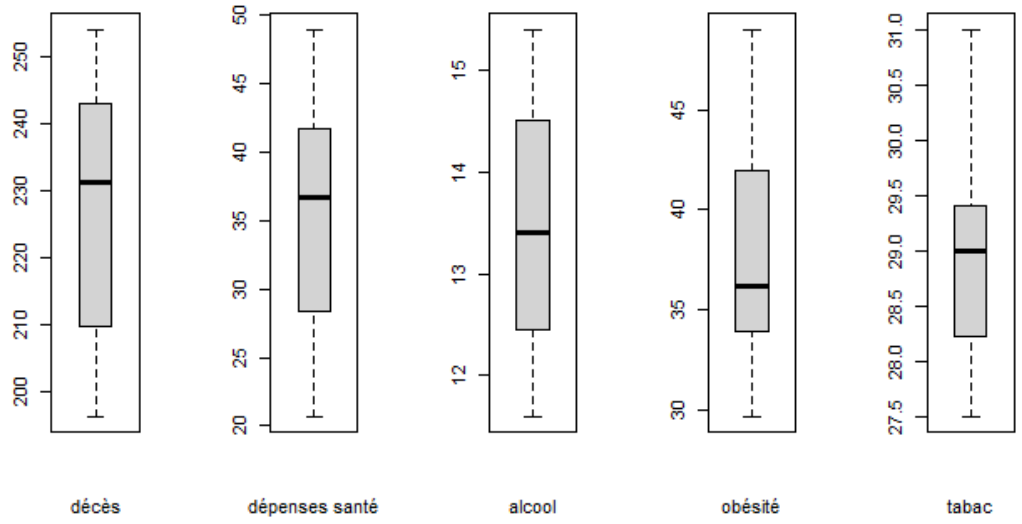


L'évolution de nos données présente des profils radicalement différents. En effet, les dépenses de santé et le taux de la population en situation de surpoids ou d'obésité augmentent sur la période étudiée. Entre 1990 et 2015, la part de la population atteinte d'obésité (ou de surpoids) est en hausse de 19.3 points de pourcentage. Par ailleurs, la consommation d'alcool diminue de façon constante sur l'ensemble de la période étudiée. Quant à l'évolution du taux de fumeurs, il est clair qu'elle est instable et ne nous permet pas d'établir de conclusion à ce sujet.

Afin de poursuivre notre analyse et pour visualiser au mieux nos données, nous pouvons observer le résumé suivant ainsi que les boxplots associés à nos variables.

```
      Date          Y_deces         X1_sante       X2_alcool       X3_obesite      X4_fumeurs   
 Min.   :1990   Min.   :196.3   Min.   :20.70   Min.   :11.60   Min.   :29.70   Min.   :27.50  
 1st Qu.:1996   1st Qu.:209.8   1st Qu.:28.33   1st Qu.:12.45   1st Qu.:33.90   1st Qu.:28.22  
 Median :2003   Median :231.3   Median :36.70   Median :13.40   Median :36.20   Median :29.00  
 Mean   :2003   Mean   :226.6   Mean   :35.14   Mean   :13.49   Mean   :37.61   Mean   :28.92  
 3rd Qu.:2010   3rd Qu.:243.2   3rd Qu.:41.75   3rd Qu.:14.50   3rd Qu.:41.98   3rd Qu.:29.40  
 Max.   :2016   Max.   :254.2   Max.   :48.97   Max.   :15.40   Max.   :49.00   Max.   :31.00  
```

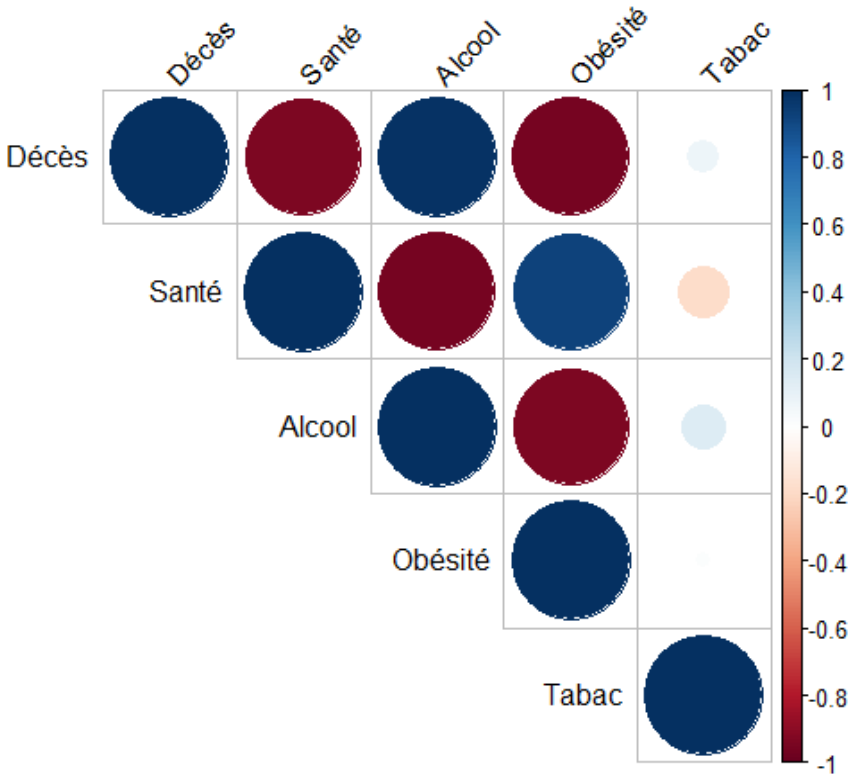
Ce résumé nous renvoie les principales statistiques sur nos observations. Premièrement, entre 1990 et 2016, la moyenne du nombre de décès suite à un cancer est d'environ 227 personnes (pour 100 000), en France. Ensuite, concernant les dépenses de santé, nous pouvons observer une assez forte dispersion puisque le minimum est de 20.7 dollars US constants face à quasiment 50 dollars sur cette même période. En 26 ans, l'évolution de cette variable est considérable puisque les dépenses ont plus que doublé. Par ailleurs, la consommation d'alcool moyenne est de 13.5 litres par habitant par an, représentant 1.125 litres par mois. La part de la population en situation de surpoids ou d'obésité présente une tendance claire comme l'a montré le graphique auparavant, c'est pourquoi les valeurs extrêmes sont éloignées. En effet, on note une différence de presque 20 points de pourcentage sur cette période. Enfin, le taux de fumeurs ne révèle pas de statistique notable puisque les fluctuations de cette variable sont limitées.



A l'aide de ces boxplots, nous pouvons voir l'étendue des distributions de nos variables. Dans l'ensemble, ces dernières ont des tendances marquées, de croissance ou de décroissance, et se traduisent donc par une variance importante entre la première et la dernière valeur. Les quatre premiers boxplots illustrent cette réalité. A l'inverse, le taux de fumeurs quotidien présente de faibles fluctuations avec des valeurs majoritairement comprises entre 28 et 29.5%, ainsi la distribution de ses observations est davantage concentrée autour de sa moyenne.

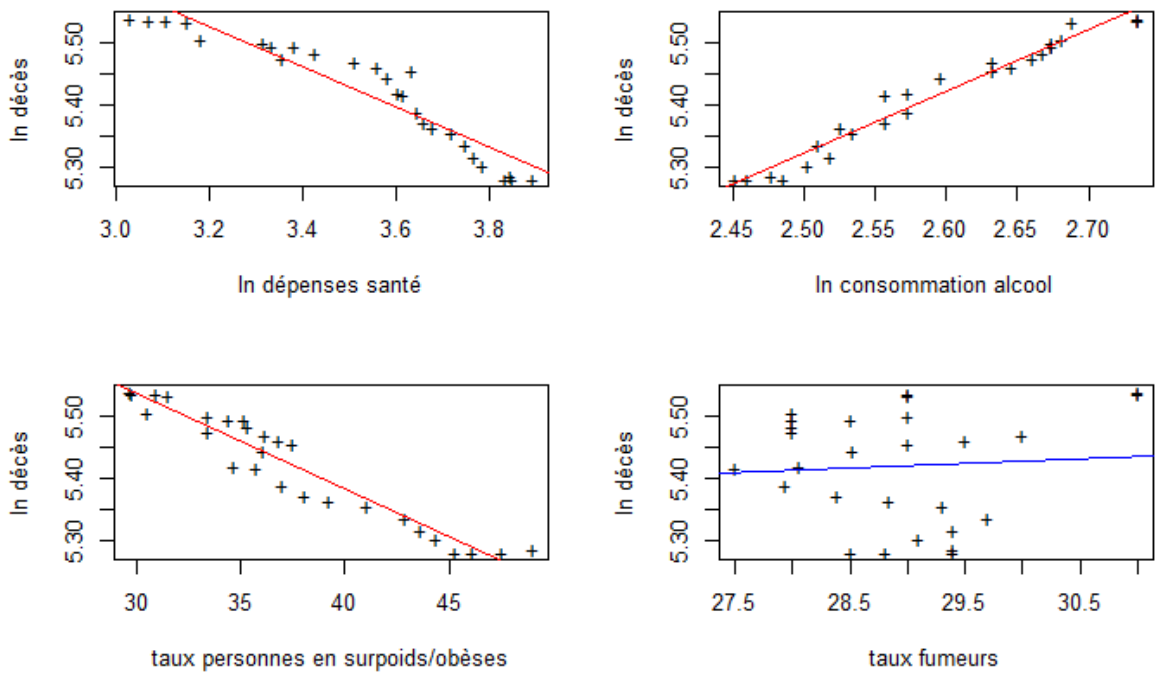
Notre variable endogène, ayant une tendance décroissante, se voit donc corrélée positivement à la variable consommation alcool et négativement aux dépenses de santé ainsi qu'au taux d'obésité. Ces corrélations doivent cependant être interprétées avec attention. Elles n'impliquent pas, telles quelles, de causalité directe. Leurs tendances sont similaires, et l'impact n'est peut-être pas totalement nul, mais nous ne pouvons pas établir de conclusion à partir de cette matrice. Notre modèle ayant pour but d'expliquer la mortalité suite à un cancer, les liens entre nos variables sont donc importants. La matrice de corrélation ci-dessus présente cette situation. En effet, notre variable endogène, ayant une tendance à la baisse, est directement en lien avec nos variables explicatives. Par exemple, lorsque le nombre de décès augmente, nous observons que la consommation d'alcool s'accroît également, la corrélation est donc positive et l'interprétation cohérente. Cependant, les dépenses de santé et le taux d'obésité présentent des tendances différentes de celle du nombre de décès et

l'interprétation n'est pas la même. Dans ce cas, les corrélations doivent être interprétées avec attention puisqu'elles n'impliquent pas forcément de causalité directe. Enfin, cette figure expose le fait que la variable tabac et décès ne sont pas en lien, pouvant paraître étonnant.



L'ensemble de cette analyse met en avant la singularité de la variable concernant le taux de fumeurs en France. Que ce soit sur la tendance, la distribution ou encore la corrélation, elle présente des différences notoires qui nous mène donc à se questionner sur sa place dans notre base de données.

Suite à cette étude statistique, nous avons pris connaissance d'un problème d'unités au sein de notre base de données. En effet, les décès dus à un cancer, les dépenses de santé et la consommation d'alcool sont exprimées dans diverses unités alors que les autres variables explicatives sont simplement exprimées en pourcentages. Ces différentes mesures peuvent altérer nos résultats, nous avons donc décidé d'utiliser le passage au logarithme pour les trois variables non exprimées en pourcentage. Avant cette transformation, nous nous sommes assurées que notre démarche était cohérente.



Nous avons observé une tendance linéaire concernant les trois premières variables explicatives. Par contre, le taux de fumeurs ne montre pas de tendance linéaire, comme présenté auparavant. Ces graphiques ont donc confirmé la nécessité du passage en log pour trois de nos variables. Après transformation nous avons obtenu un nouveau résumé :

```
      Date          Trend          Y_deces        X1_sante       X2_alcool       X3_obesite
 Min.   :1990   Min.   : 1.0   Min.   :196.3   Min.   :20.70   Min.   :11.60   Min.   :29.70
 1st Qu.:1996   1st Qu.: 7.5   1st Qu.:209.8   1st Qu.:28.33   1st Qu.:12.45   1st Qu.:33.90
 Median :2003   Median :14.0   Median :231.3   Median :36.70   Median :13.40   Median :36.20
 Mean   :2003   Mean   :14.0   Mean   :226.6   Mean   :35.14   Mean   :13.49   Mean   :37.61
 3rd Qu.:2010   3rd Qu.:20.5   3rd Qu.:243.2   3rd Qu.:41.75   3rd Qu.:14.50   3rd Qu.:41.98
 Max.   :2016   Max.   :27.0   Max.   :254.2   Max.   :48.97   Max.   :15.40   Max.   :49.00
   X4_fumeurs      deflateur         lndeces         lnsante         lnalcool
 Min.   :27.50   Min.   : 70.37   Min.   :5.280   Min.   :3.030   Min.   :2.451
 1st Qu.:28.22   1st Qu.: 77.69   1st Qu.:5.346   1st Qu.:3.344   1st Qu.:2.522
 Median :29.00   Median : 85.01   Median :5.444   Median :3.603   Median :2.595
 Mean   :28.92   Mean   : 85.89   Mean   :5.419   Mean   :3.528   Mean   :2.598
 3rd Qu.:29.40   3rd Qu.: 95.02   3rd Qu.:5.494   3rd Qu.:3.732   3rd Qu.:2.674
 Max.   :31.00   Max.   :100.52   Max.   :5.538   Max.   :3.891   Max.   :2.734
```

Les données en logarithme ne sont pas interprétables telles quelles. Pour la variable à expliquer, sa moyenne est de 5.419 mais cela n'a rien de concret et sa signification est difficile à comprendre. C'est au niveau des résultats du modèle que l'interprétation des valeurs sera facilitée, en effet l'ensemble des variables seront lues en pourcentage.
Grâce à l'ensemble de notre travail et transformations sur nos observations, nous avons abouti à une base de données optimale pour poursuivre notre étude. Nous allons désormais établir différents modèles.

Dans un premier temps, nous avons voulu tester un modèle sur nos données initiales sans transformation en log. Nous avons obtenu un coefficient de détermination (R^2) de 0.9807, très proche de 1, traduisant une bonne qualité du modèle. De plus, les variables concernant la consommation d'alcool et le taux d'obésité se sont révélées significatives à 5%. En effet les valeurs de la p-value, se trouvant dans la dernière colonne, sont, pour ces deux variables,

inférieures à 0.05. En revanche, les dépenses de santé et le taux de fumeurs ne passent pas le test à 5% ni à 10%. D'après ce modèle, ces variables ne semblent pas avoir de réel impact sur le nombre de décès suite à un cancer. Le problème qu'expose cette représentation est surtout, comme précisé précédemment, dans l'interprétation des résultats. Il faut apporter une vigilance quant à l'unité de chaque variable. Par exemple, lorsque la consommation d'alcool augmente d'un litre par personne par an, les décès suite à un cancer sont en hausse de 12.26 pour 100 000 habitants.

```
Call:
lm(formula = Y_deces ~ X1_sante + X2_alcool + X3_obesite + X4_fumeurs,
    data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0785 -2.7825  0.4149  2.0779  5.5913

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.15361   41.23675   2.744  0.01185 *
X1_sante      0.08194    0.37233   0.220  0.82785
X2_alcool    12.25919    2.14600   5.713 9.57e-06 ***
X3_obesite   -1.08196    0.38305  -2.825  0.00987 **
X4_fumeurs   -0.48937    0.77566  -0.631  0.53460
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.009 on 22 degrees of freedom
Multiple R-squared:  0.9807,    Adjusted R-squared:  0.9772
F-statistic: 279.2 on 4 and 22 DF,  p-value: < 2.2e-16
```

Dans un second temps, nous nous sommes penchées sur un second modèle basé sur les données transformées en logarithme. Ce dernier est un modèle semi-log, puisque toutes les variables n'ont pas subi un passage au logarithme. Les résultats présentés ci-dessous sont meilleurs que ceux obtenus par le premier modèle. En effet, avec un R^2 de 0.9817, le modèle dévoile une meilleure qualité d'ajustement. De plus, la variable obésité a gagné en significativité et la variable santé s'est grandement améliorée. En revanche, le taux de fumeurs semble avoir moins d'impact sur les décès suite à un cancer que dans le modèle précédent.

```
Call:
lm(formula = lndeces ~ lnsante + lnalcool + X3_obesite + X4_fumeurs,
    data = donnees)

Residuals:
      Min        1Q    Median        3Q       Max
-0.018989 -0.011258  0.001102  0.008733  0.024848

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.493571   0.395431   8.835 1.09e-08 ***
lnsante      0.040520   0.037425   1.083  0.29067
lnalcool     0.777151   0.106952   7.266 2.80e-07 ***
X3_obesite  -0.005445   0.001526  -3.569  0.00172 **
X4_fumeurs  -0.001095   0.003501  -0.313  0.75749
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01309 on 22 degrees of freedom
Multiple R-squared:  0.9817,    Adjusted R-squared:  0.9784
F-statistic: 295.8 on 4 and 22 DF,  p-value: < 2.2e-16
```

Dans un troisième temps, nous avons établi un autre modèle reposant sur les conclusions du précédent. Nous écartons la variable tabac, pour gagner en significativité. Ces nouveaux résultats sont préférables, en effet, le R^2 est identique à son prédécesseur et on ajoute à cela une augmentation de la significativité de l'intégralité des coefficients.

```
Call:
lm(formula = lndeces ~ lnsante + lnalcool + X3_obesite, data = donnees)

Residuals:
      Min        1Q    Median        3Q       Max
-0.018691 -0.010481  0.002232  0.008802  0.025229

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.464585   0.376795   9.195 3.64e-09 ***
lnsante      0.044800   0.034142   1.312 0.202414
lnalcool     0.773734   0.104284   7.419 1.52e-07 ***
X3_obesite  -0.005681   0.001299  -4.374 0.000222 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01283 on 23 degrees of freedom
Multiple R-squared:  0.9817,    Adjusted R-squared:  0.9793
F-statistic: 410.4 on 3 and 23 DF,  p-value: < 2.2e-16
```

Enfin, pour davantage prendre en compte la dimension temporelle de nos données, nous avons établi un dernier modèle. Premièrement, nous avons passé notre base de données en série temporelle pour pouvoir analyser clairement la tendance du nombre de décès. Pour examiner si la tendance est plutôt linéaire ou polynomiale, nous avons réalisé deux régressions.

```
Call:
lm(formula = lndeces ~ Trend + Trend2, data = donnees)

Residuals:
      Min        1Q    Median        3Q       Max
-0.021808 -0.007614 -0.001339  0.009183  0.026470

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -43.7188    10.3418  -4.227 0.000296 ***
Trend        17.1389     3.8174   4.490 0.000152 ***
Trend2       -1.4891     0.3522  -4.228 0.000296 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01212 on 24 degrees of freedom
Multiple R-squared:  0.9829,    Adjusted R-squared:  0.9815
F-statistic:   691 on 2 and 24 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = lndeces ~ Trend, data = donnees_t)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0289690 -0.0137406  0.0002055  0.0136217  0.0250776

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.5739886  0.0062073  897.97   <2e-16 ***
Trend       -0.0110573  0.0003875  -28.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01568 on 25 degrees of freedom
Multiple R-squared:  0.9702,    Adjusted R-squared:  0.969
F-statistic: 814.4 on 1 and 25 DF,  p-value: < 2.2e-16
```

Naturellement, c'est la première option qui apparaît comme la plus adaptée. Ainsi, avec ces résultats, nous avons choisi d’intégrer la variable Trend (tendance linéaire) dans notre modèle afin de le rendre plus précis. Cette régression multiple renvoie une qualité d’ajustement supérieure. Les significativités de nos variables changent de façon notoire. L’ensemble d’entre elles correspondent à un risque de première espèce de 10%. Ce choix de variables représente ainsi la combinaison la plus précise pour poursuivre notre analyse.

```
Call:
lm(formula = lndeces ~ lnsante + lnalcool + X3_obesite + Trend,
    data = donnees)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0139479 -0.0083263 -0.0008414  0.0064935  0.0195261

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.866535   1.350976  -0.641  0.52788
lnsante      0.148165   0.042410   3.494  0.00206 **
lnalcool     0.363269   0.152020   2.390  0.02587 *
X3_obesite  -0.002575   0.001438  -1.790  0.08717 .
Trend        0.907144   0.275151   3.297  0.00329 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01073 on 22 degrees of freedom
Multiple R-squared:  0.9877,    Adjusted R-squared:  0.9855
F-statistic: 442.6 on 4 and 22 DF,  p-value: < 2.2e-16
```

Le nouveau modèle correspond donc à l'équation suivante :

$$Y_{lndécès} = X1_{lnsanté} + X2_{lnalcool} + X3_{obésité} + X4_{trend}$$

A l'issue de ces démarches, nous avons analysé plus précisément notre modèle. La réalisation de tests liés aux contraintes ou liés à la stabilité du modèle nous ont permis de conclure quant à l'hypothèse de nullité des coefficients ou encore d'observer d'éventuels changements structurels.

Les écarts entre les valeurs observées et les valeurs prévues par notre modèle de régression sont concentrés autour de 0. Les résidus semblent être distribués selon une loi normale centrée en ce point. En réalisant un test de Shapiro, nous observons une p-value valant 0.1427, supérieure à 0.10, l'hypothèse de normalité des résidus est donc conservée.

```
        Shapiro-Wilk normality test

data:  residuals(reg_trend)
W = 0.94279, p-value = 0.1427
```

De plus, la différence entre les valeurs extrêmes vaut 0.04392 traduisant une dispersion faible. La valeur de ces résidus accentue la qualité du modèle.

```
Call:
lm(formula = lndeces ~ lnsante + lnalcool + X3_obesite + Trend,
    data = donnees)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0139479 -0.0083263 -0.0008414  0.0064935  0.0195261

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.866535   1.350976  -0.641  0.52788
lnsante      0.148165   0.042410   3.494  0.00206 **
lnalcool     0.363269   0.152020   2.390  0.02587 *
X3_obesite  -0.002575   0.001438  -1.790  0.08717 .
Trend        0.907144   0.275151   3.297  0.00329 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01073 on 22 degrees of freedom
Multiple R-squared:  0.9877,    Adjusted R-squared:  0.9855
F-statistic: 442.6 on 4 and 22 DF,  p-value: < 2.2e-16
```

Afin de pouvoir procéder à l'interprétation de nos variables, la réalisation des tests de significativité des coefficients et de significativité globale du modèle est essentielle. Nous avons d'abord évalué cette dernière grâce au test de Fisher. Dans notre situation, la p-value est très largement inférieure à un risque de première espèce de 10%. Ainsi, on rejette très fortement l'hypothèse correspondant à l'absence de significativité globale de notre modèle. Puis, l'hypothèse de normalité des résidus étant vérifiée, nous avons pu analyser les résultats du test de Student. Celui-ci nous permet de conserver ou non l'hypothèse selon laquelle le coefficient est significatif au sein du modèle. Afin de rejeter cette hypothèse pour l'ensemble de nos variables, nous supposons donc un risque de première espèce de 0.10. En effet, l'ensemble des coefficients sont supérieurs (en valeur absolue) à 1.717, valeur théorique de la loi de Student à 22 degrés de libertés. Une autre façon de démontrer que nos coefficients

sont bien significatifs dans ce modèle est d'observer la p-value associée à chacun d'entre eux. Si la valeur est inférieure à notre seuil alors l'hypothèse de nullité du coefficient est rejetée. Nous pouvons remarquer que toutes les valeurs sont strictement inférieures à 0.10.

A présent, l'interprétation de chaque variable est donc possible. Dans un cadre purement statistique, toutes choses égales par ailleurs, lorsque les dépenses de santé augmentent de 1%, les décès dus aux cancers augmentent de 0.1482%. Ce résultat peut paraître aberrant puisque la corrélation entre ces deux variables devrait être négative. Dans un cadre plus réaliste, leurs tendances étant contradictoires, l'augmentation des dépenses de santé impliquerait une diminution des décès dus aux cancers. De son côté, si le taux d'obésité augmente de 1%, les victimes du cancer diminuent de 0.0026 points de pourcentage. La relation linéaire entre ces deux variables est donc décroissante. Par ailleurs, la hausse de 1% de la consommation d'alcool entraîne un accroissement de 0.3633% du nombre de décès. L'ensemble de notre modèle est donc à nuancer puisque les valeurs obtenues sont exclusivement de nature mathématique.

Pour finir, le coefficient de détermination dévoile une très bonne qualité d'ajustement : 0.9877. Toutefois, une valeur si élevée peut cacher un problème d'endogénéité impliquant des résultats inexacts.

A ce stade de notre projet, nous disposons d'un modèle statistique globalement satisfaisant. En complément, nous allons établir différents tests pour préciser la pertinence et la stabilité du modèle.

D'abord, nous avons comparé deux modèles emboîtés : le premier sans contrainte, composé de l'ensemble de nos variables, et le second avec contraintes, comprenant exclusivement la variable dépense de santé. Ce dernier suppose la nullité des coefficients de la variable alcool, obésité et de la tendance. Le deuxième modèle étant un cas particulier du premier, la comparaison de leur variance par la méthode de l'anova nous a permis d'observer leur pouvoir explicatif. Le modèle réduit présente une variance plus élevée que celui de référence. De plus, l'écart entre ces valeurs est considérable puisqu'il est d'environ 0.02199. Ainsi, la complexité de notre modèle influe beaucoup sur son caractère explicatif et aura donc un pouvoir de prédiction plus précis.

```
Analysis of Variance Table

Model 1: lndeces ~ lnsante + lnalcool + X3_obesite + Trend
Model 2: lndeces ~ lnsante
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1     22 0.0025333
2     25 0.0245292 -3 -0.021996 63.672 5.248e-11 ***
```

Ensuite, nous avons mis en place un autre test permettant de tester la stabilité temporelle du modèle. Grâce aux graphiques réalisés auparavant, nous avons pu observer une date de rupture dans la variable obésité. En 2004, la population en situation de surpoids ou d'obésité chute de 1.4 point de pourcentage alors que la tendance était à la hausse depuis 1990. Nous avons donc observé si l'impact de nos variables explicatives différait avant et après cette date. La création d'une variable indicatrice nous a permis de mettre en place ce test et d'identifier si les coefficients de l'équation sont identiques ou non entre ces deux périodes. Nous avons réalisé une régression multiple sur notre modèle complet correspondant au modèle sans contrainte. Pour examiner cette possible date de rupture, nous avons effectué une autre régression sur deux sous périodes : avant et après 2004. Ces deux tableaux ci-dessous constituent donc notre modèle avec contraintes. Notre observation va alors se porter sur la variation du coefficient lié à la variable obésité. Entre ces deux périodes, les coefficients ne sont pas constants et présentent une différence notable. Aussi, la significativité des autres

variables explicatives change, ce qui montre un changement structurel. Ces deux sous-modèles sont donc plus adaptés pour expliquer le nombre de décès suite à un cancer. En conclusion, notre modèle n'est pas stable dans le temps puisque la variable obésité présente une date de rupture. Il faudrait donc considérer deux sous-modèles pour analyser les fluctuations du nombre de décès de 1990 à 2016.

```
Call:
lm(formula = lndeces ~ lnsante + lnalcool + X3_obesite:obesite_av2004 +
    Trend, data = donnees)

Residuals:
      Min        1Q    Median        3Q       Max
-0.016921 -0.007668 -0.001756  0.008650  0.020956

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -2.551e+00  1.164e+00  -2.192  0.03925 *
lnsante                     1.724e-01  4.934e-02   3.495  0.00205 **
lnalcool                    2.683e-01  1.670e-01   1.607  0.12230
Trend                       1.230e+00  2.280e-01   5.393 2.05e-05 ***
X3_obesite:obesite_av2004   4.253e-07  2.984e-04   0.001  0.99888
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01149 on 22 degrees of freedom
Multiple R-squared:  0.9859,    Adjusted R-squared:  0.9834
F-statistic: 385.7 on 4 and 22 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = lndeces ~ lnsante + lnalcool + X3_obesite:obesite_ap2004 +
    Trend, data = donnees)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0178108 -0.0070146  0.0000536  0.0071743  0.0253059

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.4561653  1.4251013  -1.022 0.317982
lnsante                     0.1326387  0.0555921   2.386 0.026072 *
lnalcool                    0.2400008  0.1508281   1.591 0.125828
Trend                       1.0684662  0.2624833   4.071 0.000508 ***
X3_obesite:obesite_ap2004 -0.0003387  0.0003107  -1.090 0.287464
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01119 on 22 degrees of freedom
Multiple R-squared:  0.9867,    Adjusted R-squared:  0.9842
F-statistic: 406.8 on 4 and 22 DF,  p-value: < 2.2e-16
```

Au terme de notre analyse, nous avons sélectionné un modèle composé de quatre variables explicatives. Le choix de ce croisement de variables nous a permis d'aboutir à des résultats très satisfaisants. Nos données sont significatives au seuil de 10% et le coefficient de détermination est très proche de 1. Globalement, notre modèle semble donc complet et qualitatif. De plus, grâce à différents tests réalisés, nous constatons que l'ensemble de nos variables ont une importance au sein du modèle.

Cependant, notre représentation étant simplifiée, celle-ci révèle certaines limites. En effet, un R^2 aussi important peut provoquer une estimation biaisée. En outre, nous avons été confrontées à des interprétations anormales quant à la relation entre nos variables. De plus, un changement structurel a été dévoilé par un de nos tests, remettant en cause la stabilité de notre modèle. D'un point de vue scientifique, nous pouvons observer que la variable tabac, omise dans notre analyse, impacte la proportion de décès. Notre étude n'étant pas exhaustive, nos conclusions relèvent d'un cadre purement économétrique.

Nous avons pu, à notre échelle, tenter de déterminer l'impact de certains facteurs entraînant un décès suite à un cancer. Au sein de notre modélisation, le financement des services de santé joue un rôle non négligeable dans la lutte contre cette maladie. Les habitudes de vie peuvent, à l'inverse, impacter négativement l'état de santé des patients. Le cancer est l'une des premières causes de décès en France, il paraît donc évident que les éléments mis en avant ici, ne sont pas les seuls en cause. L'apparition de cette pathologie ne répond pas à une science parfaite, nous pouvons donc nous demander dans quelle mesure sa prédiction et son étude peuvent être améliorées. Et par la suite, si ces démarches économétriques peuvent aider à la recherche médicale.