Rapport: Projet Modèle Linéaire

Ayoub EL HOUDRI, Nawfel BACHA, Marine RASOLOFO, Manoah ABENZOAR

CY Tech - ING2 – MI 1

05 Décembre 2021

1. Introduction

Ce projet consiste à la mise en place d'un modèle linéaire pour expliquer la variable **Score (indicateur de bonheur dans chaque pays)** en fonction de 6 autres variables expliquatives: **GDP_per_capita, Social_support, Healthy_life_expectancy, Freedom_to_make_life_coices, Generosity** and **Perception_of_corruption**. Au fur et à mesure, nous essaierons d'ajuster le modèle en améliorant la qualité de prédiction tout en proposant un calcul d'erreur bien détaillé à chaque étape.

Dans la section suivante nous présenterons plus en détails les données utilisés pour l'entrainement et l'évaluation du modèle.

2. Data: World Happiness Report (2019)

Le World Happiness Report est une enquête historique sur l'état du bonheur dans le monde. Le premier rapport a été publié en 2012, le deuxième en 2013, le troisième en 2015 et le quatrième dans la mise à jour 2016. Le World Happiness 2017, qui classe 155 pays selon leur niveau de bonheur, a été publié aux Nations Unies lors d'un événement célébrant la Journée internationale du bonheur le 20 mars. . Des experts de premier plan dans tous les domaines - économie, psychologie, analyse d'enquêtes, statistiques nationales, santé, politiques publiques, etc. - décrivent comment les mesures du bien-être peuvent être utilisées efficacement pour évaluer les progrès des nations. Les rapports examinent l'état du bonheur dans le monde aujourd'hui et montrent comment la nouvelle science du bonheur explique les variations personnelles et nationales du bonheur.

Les données sont disponibles ici [World Hapiness Report 2019] sous format csv, et se présentent ainsi:

	A	B	C	D	E	F	G	H	I
1		Country or region	Score	GDP_per_capita	Social_support	Healthy_life_expectancy	Freedom_to_make_life_choices	Generosity	Perceptions_of_corruption
2	1	Finland	7.769	1.34	1.587	0.986	0.596	0.153	0.393
3	2	Denmark	7.6	1.383	1.573	0.996	0.592	0.252	0.41
4	3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341
5	4	Iceland	7.494	1.38	1.624	1.026	0.591	0.354	0.118
6	5	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298
7	6	Switzerland	7.48	1.452	1.526	1.052	0.572	0.263	0.343
8	7	Sweden	7.343	1.387	1.487	1.009	0.574	0.267	0.373
9	8	New Zealand	7.307	1.303	1.557	1.026	0.585	0.33	0.38
10	9	Canada	7.278	1.365	1.505	1.039	0.584	0.285	0.308
11	10	Austria	7.246	1.376	1.475	1.016	0.532	0.244	0.226
12	11	Australia	7.228	1.372	1.548	1.036	0.557	0.332	0.29
13	12	Costa Rica	7.167	1.034	1.441	0.963	0.558	0.144	0.093
14	13	Israel	7.139	1.276	1.455	1.029	0.371	0.261	0.082
15	14	Luxembourg	7.09	1.609	1.479	1.012	0.526	0.194	0.316
16	15	United Kingdom	7.054	1.333	1.538	0.996	0.45	0.348	0.278
17	16	Ireland	7.021	1.499	1.553	0.999	0.516	0.298	0.31
18	17	Germany	6.985	1.373	1.454	0.987	0.495	0.261	0.265
19	18	Belgium	6.923	1.356	1.504	0.986	0.473	0.16	0.21
20	19	United States	6.892	1.433	1.457	0.874	0.454	0.28	0.128
21	20	Czech Republic	6.852	1.269	1.487	0.92	0.457	0.046	0.036
22	21	United Arab Emirates	6.825	1.503	1.31	0.825	0.598	0.262	0.182
23	22	Malta	6.726	1.3	1.52	0.999	0.564	0.375	0.151
24	23	Mexico	6.595	1.07	1.323	0.861	0.433	0.074	0.073
25	24	France	6.592	1.324	1.472	1.045	0.436	0.111	0.183
26	25	Taiwan	6.446	1.368	1.43	0.914	0.351	0.242	0.097

Figure 1: Echantillon des données utilisées

3. Régression Linéaire

3.1. Modèle linéaire

On commence par une visualisation des données pour voir la répartition des données s'il y a des informations évidentes à en tirer.

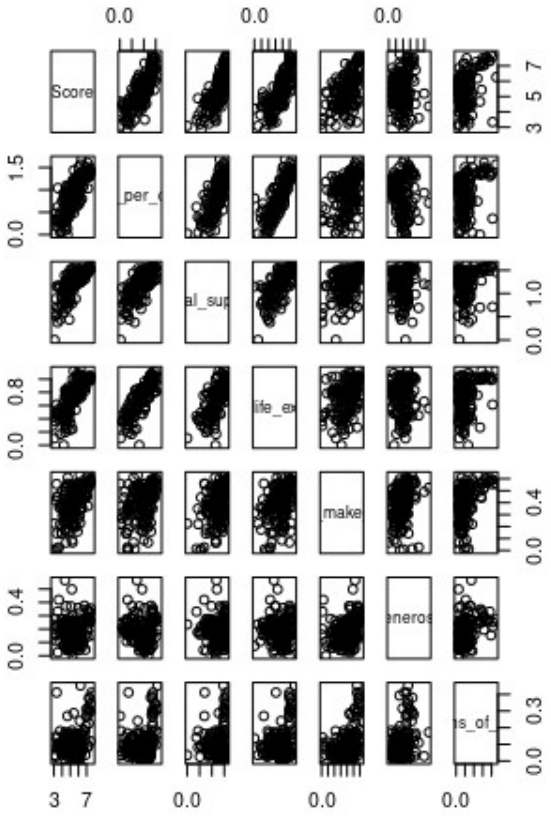

Figure 2: Visualisation des données

Par la suite on construire le modèle linéaire utilisant les 6 variables comme variables explicatives et la variable **Score** comme variable à expliquer, le résultats est donné sous forme de 7 coefficients (coefficients and intercept).

```
Call:
lm(formula = Score ~ GDP_per_capita + Social_support + Healthy_life_expectancy +
     Freedom_to_make_life_choices + Generosity + Perceptions_of_corruption,
     data = df)

Coefficients:
                   (Intercept)                GDP_per_capita                Social_support
Healthy_life_expectancy
                        1.7952                        0.7754                        1.1242
1.0781
Freedom_to_make_life_choices                    Generosity     Perceptions_of_corruption
                        1.4548                        0.4898                        0.9723
```

Figure 3: Les coefficients estimés du modèle (RStudio)

Ce qui suit sont les résultats tu tableau excel pour le calcul des coefficients estimés su modèle à l'aide de la méthode des moindres carrés ordinaires.

Coefficients estimés =(X'X)^{-1}X'Y	
â0=	1.7952
â1=	0.7754
â2=	1.1242
â3=	1.0781
â4=	1.4548
â5=	0.4898
â6=	0.9723

Figure 4: Les coefficients estimés du modèle (Excel)

On remarque que les coefficients sont les mêmes, ce qui signifie que la fonction `lm` sous R utilise la méthode des moindres carrés ordianires pour estimer les coefficients du modèle.

Afin d'avoir plus d'informations sur le modèles et ses variables, on a appliquer la fonction `summary` à notre modèle et on a le résulat suivant:

```
Call:
lm(formula = Score ~ GDP_per_capita + Social_support + Healthy_life_expectancy +
     Freedom_to_make_life_choices + Generosity + Perceptions_of_corruption,
     data = df)

Residuals:
      Min       1Q   Median       3Q      Max
-1.75304 -0.35306  0.05703  0.36695  1.19059

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1.7952     0.2111   8.505 1.77e-14 ***
GDP_per_capita                    0.7754     0.2182   3.553 0.000510 ***
Social_support                    1.1242     0.2369   4.745 4.83e-06 ***
Healthy_life_expectancy           1.0781     0.3345   3.223 0.001560 **
Freedom_to_make_life_choices      1.4548     0.3753   3.876 0.000159 ***
Generosity                        0.4898     0.4977   0.984 0.326709
Perceptions_of_corruption         0.9723     0.5424   1.793 0.075053 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5335 on 149 degrees of freedom
Multiple R-squared:  0.7792,    Adjusted R-squared:  0.7703
F-statistic: 87.62 on 6 and 149 DF,  p-value: < 2.2e-16
```

Figure 5: Informations sur le modèle

Les calculs sur Excel ont donnés le même résultat su du coefficient de corrélation $R^2 = 0.7792$.

Somme $(Y-\hat{Y})^2$ (SCR) =	**42.412**
Somme $(Y-Y\_bar)^2$ (SCT) =	**192.051**
R^2 (Coefficient de corrélation du model) =	**0.7792**

*Figure 6: Coefficient de corrélation du modèle calculé à partir de **SCT** et **SCR***

Ceci signifie que la variable **Score** est expliquée avec un pourcentage de **78%** par les variables **GDP_per_capita, Social_support, Healthy_life_expectancy, Freedom_to_make_life_coices, Generosity** and **Perception_of_corruption**.

3.2. Vérifications des Hypothèses sur les résidus

3.2.1. Hypothèse 1: Espérance nulle des résidus

L'hypothèse est vérifiée en calculant la moyenne des résidus à l'aide de la fonction `mean` sous R, ou la fonction AVERAGE sous Excel. Sur R on a le résultats suivant: `mean(e) → -4.896841e-17` (presque égale à 0), Sur excel on a le résultat suivant:

Moyenne des Résidus =	**0.000**

*Figure 7: Moyenne des résidus calculés à l'aide de la fonction **AVERAGE** sur Excel*

3.2.2. Hypothèse 2: Homoscédasticité des résidus

On a effectué un test de Breush-Pagan sur le model construit pour évaluer si la variance est constante: `ncvTest(initial_model)`, et qui renvoi ce qui suit:

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.77625, Df = 1, p = 0.18261
```

*Figure 8: Résultats de la fonction **ncvTest***

On remarque qu'on a une p-value = 0.18261 > 5% ce qui signifie que L'hypothèse d'homoscédasticité des résidus est vérifiée et la variance reste constante.
Sur Excel, l'hypothèse est vérifiée à l'aide d'un graphe, qui montre que les résidus sont bien répartis en fonction des **Scores** prédits et n'ont pas une distribution spéciale.

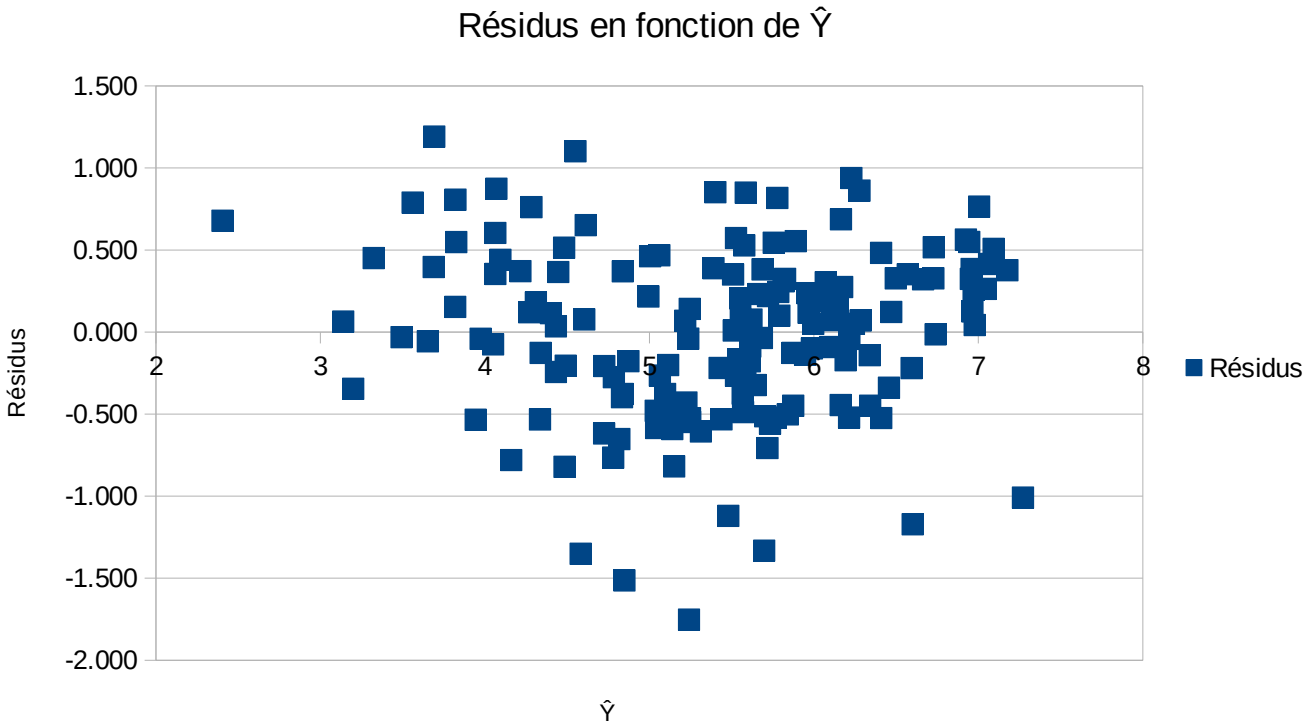

Figure 9: Répartitions des résidus en fonctions des valeurs prédites de la variable ***Score***

3.2.3. Hypothèse 3: Non corrélation des résidus

Le test de Durbin-Watson a montré qu'il est susceptible d'y avoir une corrélation entre les résidus puisque la fonction `dwtest(initial_model)` renvoi le résuilat suivant:

```
    Durbin-Watson test

data:  initial_model
DW = 1.6484, p-value = 0.01097
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 10: Le résultat du test de Durbin-Watson pour la corrélation

Le résultat montre que La statistique de test DW = 1.6484 est proche de 2, mais la p-value = 0.01097 < 5% ce qui montre qu'il y'a une possibilité de corrélation entre les résidus.
Sauf, que le graphe qui représente la répartition des résidus en fonction des valeurs prédites de la variable **Score**, confime l'hypothèse de non corrélation des résidus, vue que la répartition n'a pas de forme particulière, comme ce qui apparaît ci-dessous sur R *(Figure 11)* ainsi que le résultat sur Excel *(Figure 10).*

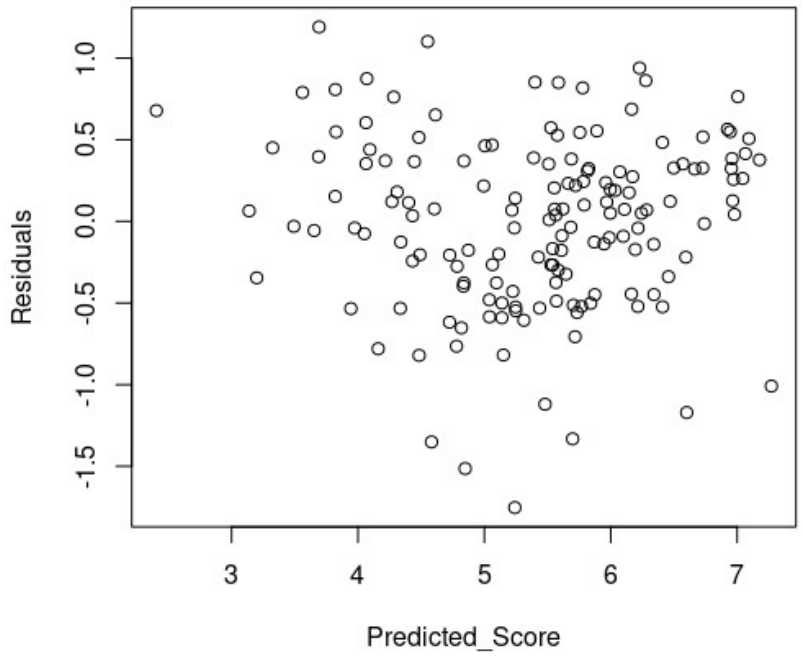

Figure 11: Répartitions des résidus en fonctions des valeurs prédites de la variable ***Score***

3.2.4. Hypothèse 4: Normalité des résidus

L'hypothèse de normalité des résidus est prouvée à l'aide du **QQ-Plot** qui représente les résidus standardisés en fonction des quantiles de la loi normale. La représentation a été faite sous R *(Figure 12)* ainsi que Excel *(Figure 13).*

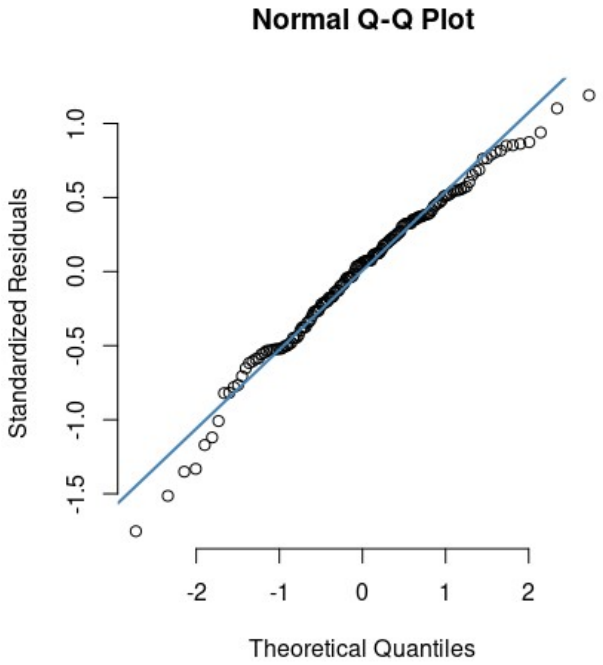

Figure 12: QQ-plot des résidus (Rstudio)

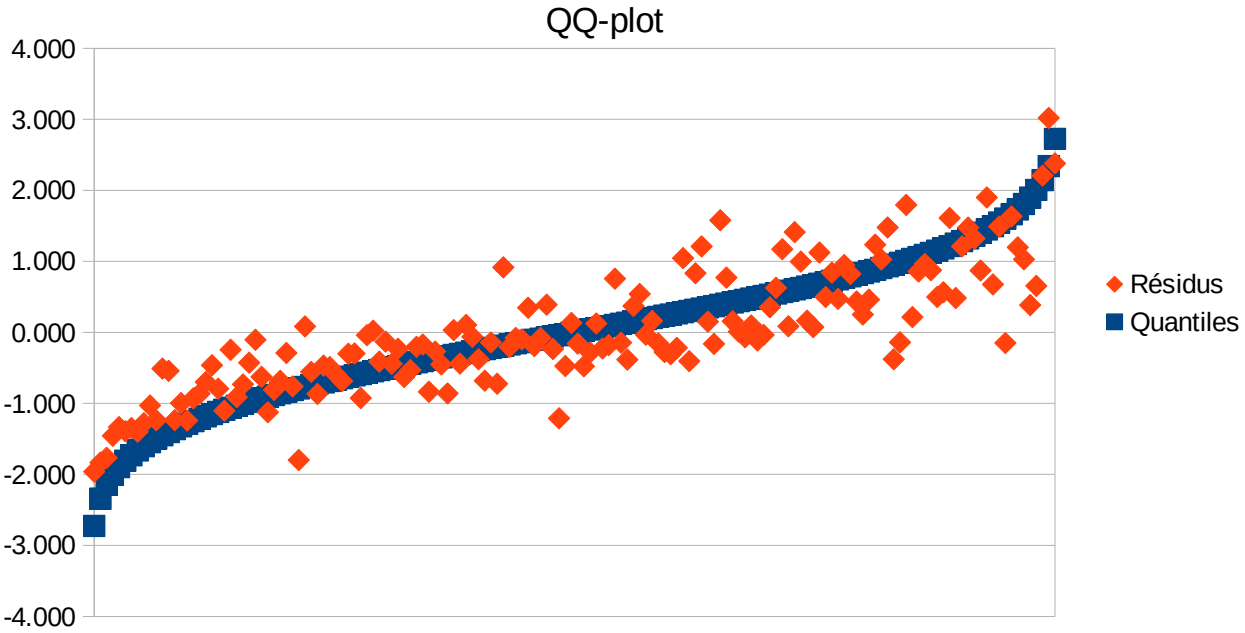

Figure 13: QQ-plot des résidus (Excel)

On remarque que les résidus sont répartis suivant la ligne droite en bleu. Ce qui montre que l'hypothèse est vérifiée.

3.3. Détection des valeurs atypiques

3.3.1. Détection des valeures abérrantes

On éffectuant la statistique de Student pour détecter les valeurs abérrantes, on a le résultat suivant:

	Score	GDP_per_capita	Social_support	Healthy_life_expectancy	Freedom_to_make_life_choices	Generosity	Perceptions_of_corruption
12	7.167	1.034	1.441	0.963	0.558	0.144	0.093
34	6.262	1.572	1.463	1.141	0.556	0.271	0.453
67	5.653	0.677	0.886	0.535	0.313	0.220	0.098
76	5.430	1.438	1.277	1.122	0.440	0.258	0.287
99	4.944	0.569	0.808	0.232	0.352	0.154	0.090
102	4.883	0.393	0.437	0.397	0.349	0.175	0.082
130	4.366	0.949	1.265	0.831	0.470	0.244	0.047
131	4.360	0.710	1.181	0.555	0.525	0.566	0.172
148	3.488	1.041	1.145	0.538	0.455	0.025	0.100
152	3.334	0.359	0.711	0.614	0.555	0.217	0.411
153	3.231	0.476	0.885	0.499	0.417	0.276	0.147

Figure 14: Les valeurs abérrantes déctées par la statistique de Student

Une représentation graphique des données pour bien visualiser les points abbérents est affichée *(Figure 15).*

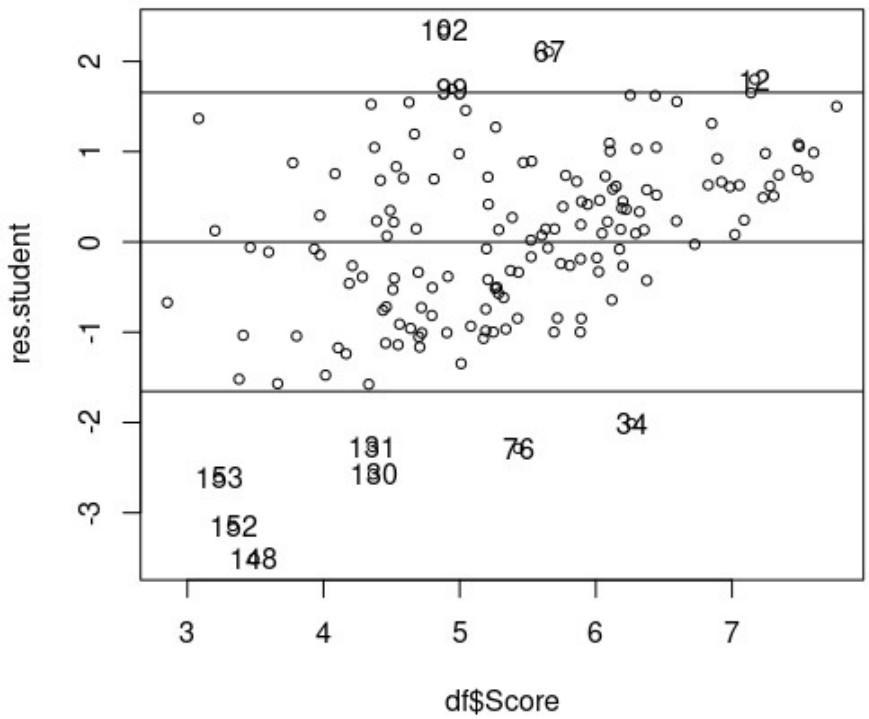

Figure 15: Représentation des valeurs abérrantes ainsi que les seuils de Student

On a fait une détection des veleurs abérrantes à l'aide du critère du Levier.

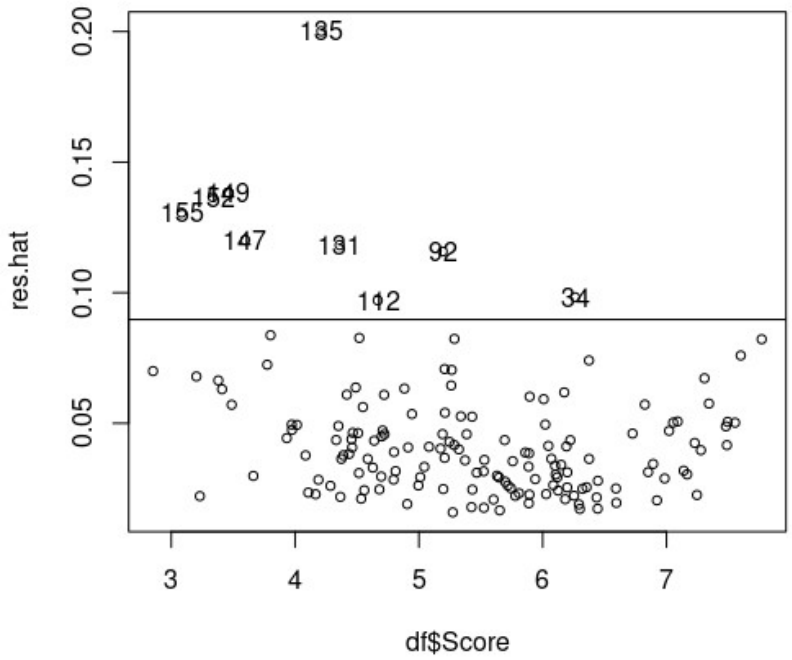

Figure 16: Représentation des valeurs abérrantes ainsi que le seuil du levier

On remarque que les deux méthode des détection des points abérrants, donnent des résultats différents: Le critère de Student détecte plus de valeurs abérrantes que le critère du Levier, ce qui signifie que le test de Student est plus précis.

3.3.1. Détection des valeures influentes

Pour détecter les valeurs atypiques on a utilisé la distance de Cook à l'aide de la fonction `ols_plot_cooksd_chart(initial_model)` du package `olsrr`, on a représenter la distance de Cook de chaque point des données ainsi que le seuil (Threshold).

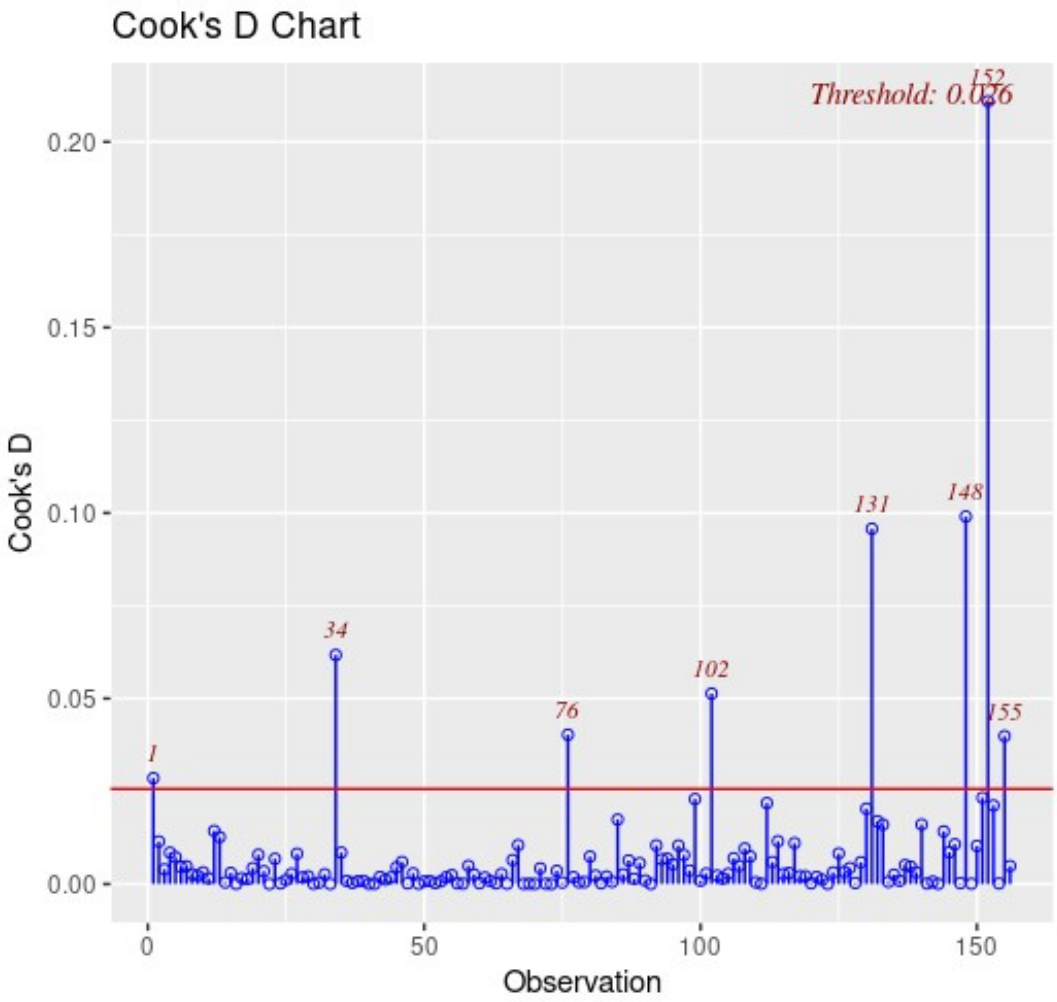

Figure 17: Représentation des valeurs atypiques (qui sont au dessus du seuil=0.026)

3.4. Evaluation du modèle et ses variables

```
Call:
lm(formula = Score ~ GDP_per_capita + Social_support + Healthy_life_expectancy +
    Freedom_to_make_life_choices + Generosity + Perceptions_of_corruption,
    data = df)

Residuals:
      Min       1Q   Median       3Q      Max
-1.75304 -0.35306  0.05703  0.36695  1.19059

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1.7952     0.2111   8.505 1.77e-14 ***
GDP_per_capita                    0.7754     0.2182   3.553 0.000510 ***
Social_support                    1.1242     0.2369   4.745 4.83e-06 ***
Healthy_life_expectancy           1.0781     0.3345   3.223 0.001560 **
Freedom_to_make_life_choices      1.4548     0.3753   3.876 0.000159 ***
Generosity                        0.4898     0.4977   0.984 0.326709
Perceptions_of_corruption         0.9723     0.5424   1.793 0.075053 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5335 on 149 degrees of freedom
Multiple R-squared:  0.7792,    Adjusted R-squared:  0.7703
F-statistic: 87.62 on 6 and 149 DF,  p-value: < 2.2e-16
```

Figure 5: Informations sur le modèle

3.4.1. Significativité de chaque variable

On remarque que toute les variables sont significatives (suivies de ‘***’ et ‘**’) en ayant une p-valeur très petite , à l’exception de la variable **Generosity** (la moins significative du modèle ‘ ‘) et **Perceptions_of_corruption** (’.’) qui ont des p-valeurs plus grandes, ce qui signifie qu’elles peuvent êtres supprimés du modèle.

3.4.2. Significativité du modèle

On teste la significativité du modèle à l’aide du test de Fisher. On a une statistique de Fisher F–statistic = 87.62 élevée et une p-value du modèle très petite, ce qui prouve l’existence d’au moins une variable explicative significative d’où le modèle peut être considéré significatif.

3.2. Vérification de la collinéarité et sélection des variables

3.2.1. Calcul de VIF et matrice de corrélation

On a calculé le VIF (Variance Inflation Factor) pour chaque variable explicative à l'aide de la fonction `vif` du package `car`, et on les a représentés sous forme de bar plot pour chaque variable, pour voir si elle dépasse le seuil qui égale à 4.

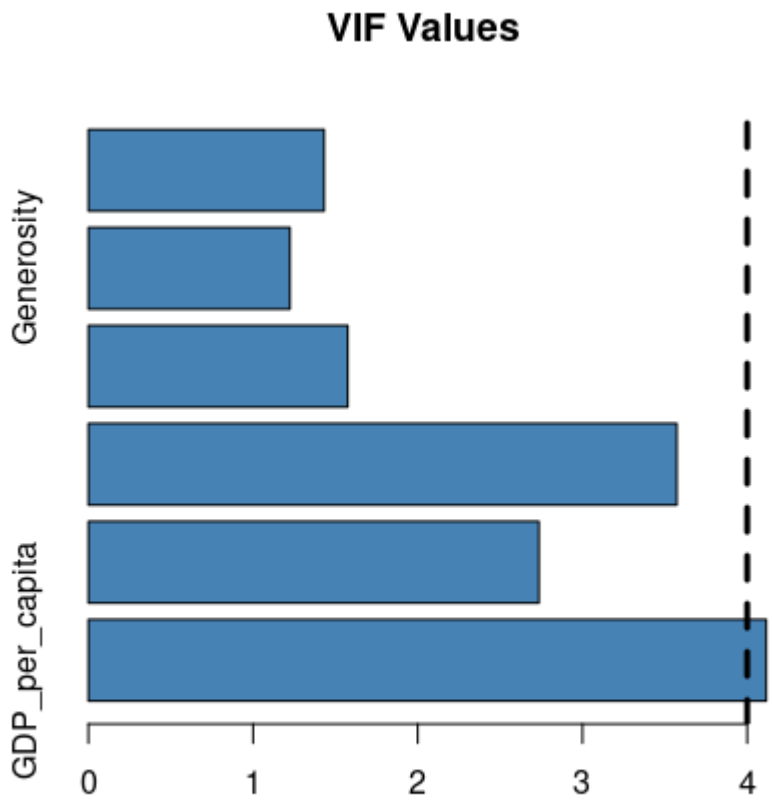

Figure 18: Représentation des valeurs atypiques (qui sont au dessus du seuil=0.026)

On remarque que la valeur du VIF de la variable **GDP_per_capita** est susceptible d'être à l'origine d'une collinéarité entre les variables explicatives.

On a représenté aussi la matrice de corrélation entre les variables explicatives, comme ci-dessus :

```
                          GDP_per_capita Social_support Healthy_life_expectancy Freedom_to_make_life_choices
Generosity
GDP_per_capita                 1.00000000     0.75490573              0.83546212
0.3790791 -0.07966231
Social_support                 0.75490573     1.00000000              0.71900946
0.4473332 -0.04812645
Healthy_life_expectancy        0.83546212     0.71900946              1.00000000
0.3903948 -0.02951086
Freedom_to_make_life_choices   0.37907907     0.44733316              0.39039478
1.0000000  0.26974181
Generosity                    -0.07966231    -0.04812645             -0.02951086
0.2697418  1.00000000
Perceptions_of_corruption      0.29891985     0.18189946              0.29528281
0.4388433  0.32653754
                                  Perceptions_of_corruption
GDP_per_capita                                    0.2989198
Social_support                                    0.1818995
Healthy_life_expectancy                           0.2952828
Freedom_to_make_life_choices                      0.4388433
Generosity                                        0.3265375
Perceptions_of_corruption                         1.0000000
```

Figure 19: Matrice de corrélation entre les variables explicatives

Le résultat montre que les variables **GDP_per_capita** a une corrélation non négligeable avec les 2 variable **Healthy_life_expectancy** et **Social_support**

3.2.2. Sélection des variables (critère du AIC)

Pour sélectionner les variables du nouveau modèle dans le but d'améliorer le modèle initial, on a éliminé la variable la moins significative en appliquant l'algorithme Backward du AIC (élimination d'une variavble par étape jusuqu'à l'atteinte de la plus petite valeur possible du AIC), ceci en appliquant la fonction `ols_step_best_subset` du package `olsrr` ce qui donne le résultat suivant:

```
                                                    Best Subsets Regression
-------------------------------------------------------------------------------------------------------------------------------------------------------
Model Index    Predictors
-------------------------------------------------------------------------------------------------------------------------------------------------------
      1        GDP_per_capita
      2        GDP_per_capita Freedom_to_make_life_choices
      3        GDP_per_capita Social_support Freedom_to_make_life_choices
      4        GDP_per_capita Social_support Healthy_life_expectancy Freedom_to_make_life_choices
      5        GDP_per_capita Social_support Healthy_life_expectancy Freedom_to_make_life_choices Perceptions_of_corruption
      6        GDP_per_capita Social_support Healthy_life_expectancy Freedom_to_make_life_choices Generosity Perceptions_of_corruption
-------------------------------------------------------------------------------------------------------------------------------------------------------
```

```
                                                    Subsets Regression Summary
-------------------------------------------------------------------------------------------------------------------------------------------
                          Adj.        Pred
Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC         SBC        MSEP       FPE       HSP       APC
-------------------------------------------------------------------------------------------------------------------------------------------
  1        0.6303      0.6278      0.6209    97.4734    325.9328    -118.5555    335.0824    71.9330    0.4670    0.0030    0.3794
  2        0.7128      0.7090      0.7024    43.8071    288.5446    -155.4108    300.7440    56.2499    0.3675    0.0024    0.2985
  3        0.7536      0.7487      0.7413    18.2747    266.6419    -176.5689    281.8912    48.5786    0.3194    0.0021    0.2594
  4        0.7709      0.7649      0.7552     8.5538    257.2385    -185.3769    275.5377    45.4552    0.3007    0.0019    0.2442
  5        0.7777      0.7703      0.7567     5.9683    254.5400    -187.6894    275.8890    44.4026    0.2955    0.0019    0.2401
  6        0.7792      0.7703       0.754     7.0000    255.5295    -186.5260    279.9284    44.4140    0.2974    0.0019    0.2416
-------------------------------------------------------------------------------------------------------------------------------------------
 AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria
```

Figure 20: Résultat de la fonction **ols_step_best_subset**

On remarque que le modèle avec le plus petit AIC (254.54) est le 5ème modèle:

```
5         GDP_per_capita Social_support Healthy_life_expectancy Freedom_to_make_life_choices Perceptions_of_corruption
```

c'est-à-dire le modèle initial sans la variable **Generosity** (la variable explicative la moins significative, ce qui est cohérent avec ce qui est présenté dans la partie **3.4.1. Significativité de chaque variable**).

```
Call:
lm(formula = Score ~ GDP_per_capita + Social_support + Healthy_life_expectancy +
    Freedom_to_make_life_choices + Perceptions_of_corruption,
    data = df)

Residuals:
      Min       1Q   Median       3Q      Max
-1.82997 -0.35344  0.05803  0.35977  1.17522

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1.8689     0.1973   9.471  < 2e-16 ***
GDP_per_capita                    0.7455     0.2161   3.450 0.000728 ***
Social_support                    1.1180     0.2368   4.722 5.33e-06 ***
Healthy_life_expectancy           1.0840     0.3344   3.241 0.001467 **
Freedom_to_make_life_choices      1.5340     0.3666   4.185 4.84e-05 ***
Perceptions_of_corruption         1.1176     0.5218   2.142 0.033839 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5335 on 150 degrees of freedom
Multiple R-squared:  0.7777,    Adjusted R-squared:  0.7703
F-statistic:   105 on 5 and 150 DF,  p-value: < 2.2e-16
```

Figure 21: Informations sur le nouveau modèle

Après le choix du modèle ci-dessus comme nouveau modèle on remarque qu'on a amélioré Pred. R-square de 0.7540 à 0.7567 la statistique de Fisher (F-statistique) qui est passée de 87.62 à 105 ce qui signifie qu'on a amélioré la significativité globale du modèle.

3.2.3. Validation Croisée

3.2.3.1. K-Fold

On a effectué une validation 10-Fold 10 fois (pour une meilleure estimation de l'erreur) sur le modèle initial ainsi que le nouveau modèle pour comparer l'erreur de prédiction (RMSE). On aboutit aux résultats suivants:

```
Linear Regression

156 samples
  6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 140, 140, 140, 140, 140, 140, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5442248  0.7800542  0.4345855
```

Figure 22: Résultats de l'algorithme 10-Fold sur le modèle initial

```
Linear Regression

156 samples
  5 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 141, 140, 140, 140, 140, 140, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5403849  0.7816133  0.4310819
```

Figure 23: Résultats de l'algorithme 10-Fold sur le nouveau modèle

On remarque que l'erreur de prediction (RMSE) est passée de 0.5442248 à 0.5403849, ce qui signifie que le modèle est amélioré.

3.2.3.1. LOOCV

On a effectué une validation LOOCV, ce qui donne les résultats suivants:

Linear Regression

```
156 samples
  6 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 155, 155, 155, 155, 155, 155, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5503624  0.7542225  0.4346722
```

Figure 24: Résultats de la validattion croisée LOOCV sur le modèle initial

Linear Regression

```
156 samples
  5 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 155, 155, 155, 155, 155, 155, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.5472434  0.7569402  0.4308614
```

Figure 25: Résultats de la validattion croisée LOOCV sur le nouveau modèle

Comme dans la validation croisée K-Fold présentée ci-dessus, on remarque que l'erreur de prédiction a baissé de 0.5503624 pour le modèle initial à 0.5472434 pour le nouveau modèle.

3. Régularisation de Ridge et de LASSO

Dans tout ce qui vient toute comparaison établie entre modèles, concerne le nouveau modèle (sans la variable **Generosity**) et chacun des modèles de Ridge et de LASSO.

3.1. Régularisation de Ridge

Pour améliorer le modèle encore plus, on a mis en place une régularisation de Ridge, qui a pour but de minimiser l'erreur de prédiction (RMSE). Le modèle est sans constante, entrainé sur nos donées centrées réduites.

On a commencé par déterminer la valeur de lambda de Ridge a plus petite qui est égale à 0.004322487. Le graphe ci-dessous explique bien le choix de cette valeur pour lambda qui tient compte des erreurs quadratiques moyennes.

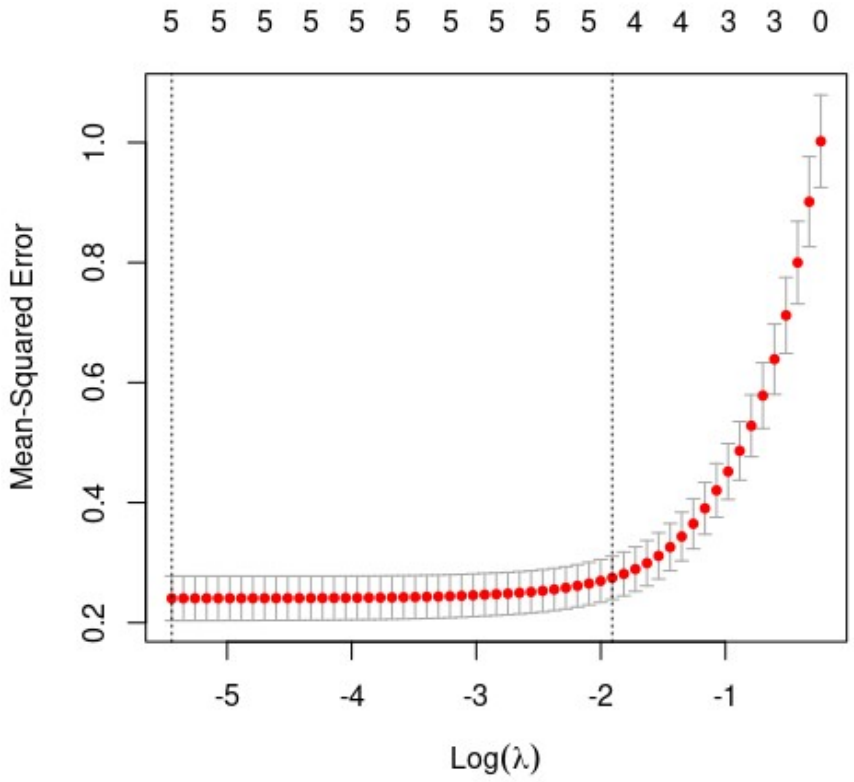

Figure 26: L'erreur quadratique moyenne en fonction de lambda Ridge en échelle logarithmique

Le calcul des coefficients de Ridge sur R et sur Excel, donne les mêmes résultats comme indiqué ci-dessous:

â _Ridge(λ)
0.26680
0.30051
0.23579
0.19747
0.09491

Erreur moyenne quadratique (RMSE) =	**0.4699**

Figure 27: Coefficients de Ridge estimés et RMSE (Excel)

```
                                s0
(Intercept)                 2.467173e-17
GDP_per_capita              2.656972e-01
Social_support              2.988175e-01
Healthy_life_expectancy     2.350779e-01
Freedom_to_make_life_choices 1.957819e-01
Perceptions_of_corruption   9.216769e-02
```

Figure 28: Coefficients de Ridge estimés (R)

Le modèle de Ridge a un coefficient de corrélation $R^2 = 0.77769$ (presque égal au R^2 du nouveau modèle) mais une erreur de prédiction RMSE = 0.4699795 plus petite (en comparaison avec le nouveau modèle) `RMSE_Ridge <- (mean((y_predicted_Ridge - y)**2))**0.5` et `RMSE_Ridge` renvoi `0.4699795.`

3.1. Régularisation de LASSO

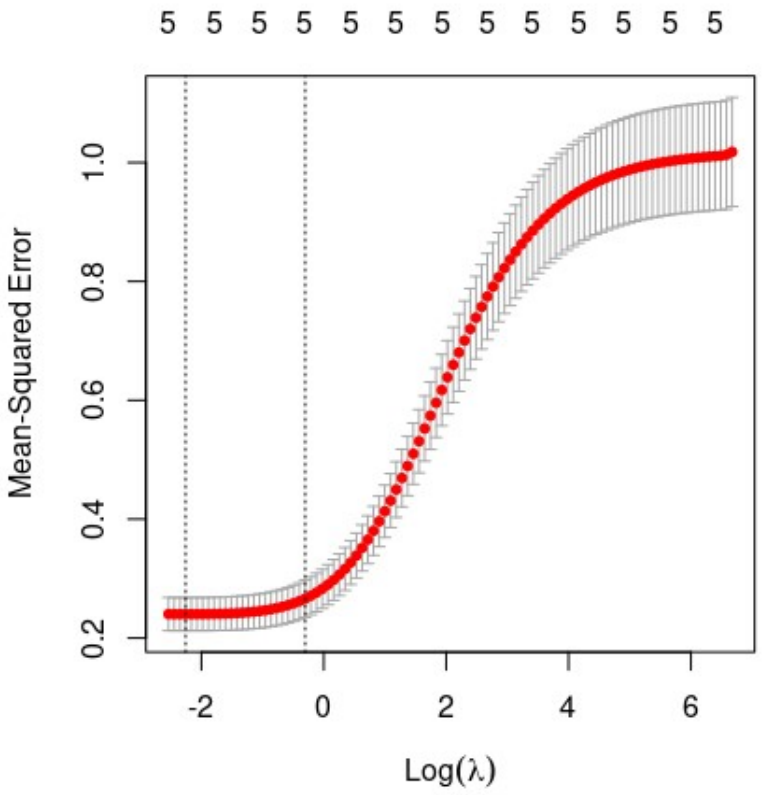

Figure 28: L'erreur quadratique moyenne en fonction de lambda LASSO en échelle logarithmique

On a effectué une régularisation de LASSO sur nos données centrées réduites.

On a commencé par déterminer la valeur de lambda de LASSO a plus petite qui est égale à 0.004709764. Le graphe ci-dessous explique bien le choix de cette valeur pour lambda qui tient compte des erreurs quadratiques moyennes.

Le calcul des coefficients de LASSO sur R, donne le résultat suivant:

```
                                    s0
(Intercept)                 2.361649e-17
GDP_per_capita              2.566052e-01
Social_support              2.818961e-01
Healthy_life_expectancy     2.359969e-01
Freedom_to_make_life_choices 1.897404e-01
Perceptions_of_corruption   9.473742e-02
```

Figure 29: Coefficients de LASSO estimés (R)

Le modèle de Ridge a un coefficient de corrélation $R^2 = 0.777675$ (presque égal au R^2 du nouveau modèle) mais une erreur de prédiction RMSE = 0.4709764 plus petite (en comparaison avec le nouveau modèle) `RMSE_LASSO <- (mean((y_predicted_LASSO - y)**2))**0.5` et `RMSE_LASSO` renvoi `0.4699795.`

5. Conclusion

On remarque qu'à travers les différents étapes du modèle initial aux régularisations de Ridge et de LASSO en passant par le nouveau modèle, on a essayé d'améliorer la qualité de prédiction du modèle en minimisant l'erreur et les résultats montrent clairement que l'erreur a diminué et la prédiction est devenue plus précise.