Health and Social-Economics Factors Influencing Life Expectancy, a report using linear regression

Yinzhou Liu

12/17/2022

Introduction

From ancient mythologies where Gilgamesh chases for immortality to Qin Shi Huang' attempting to achieve eternity , mankind has done everything imaginable to not die. While it is uncertain if mankind can ever achieve such remarkable feat, what we have achieved is a very impressive increase in life expectancy. Why is life expectancy important? Life expectancy is an important indicator that tells the overall physical health of a population, any shift in this statistics can be used to describe changes in mortality, most notably, it forms the human development index along with GDP according to the UN(Murillo,2021).

Scholars have analyzed this statistics in the past.Like a paper that focuses only on environmental effects in 2009 by Mariani, Perez-Barahona and Raffin, where they explained how some countries fall into a vicious cycle of low life expectancy correlating with low environmental level through mathematical modeling and other papers where scholars analyzed life expectancy only in certain regions, like Freeman, et al in 2020 where they combined both quantitative and opinion-based analysis for life expectancy only in Ethiopia, Brazil and USA.This report is different in the sense that I'm performing analysis for life expectancy in all countries using regressions. I'm attempting to cover more than one variable, from social economics to health, and estimating their effects on our interested outcome variable. We can interpret important information from coefficients as to which variable positively or negative influences life expectancy, this provides preliminary guidelines for countries who aim to increase life expectancy .

Method

Our data set is compiled from GHO(global health observatory) and UNESCO websites by user "mmattson", it features life expectancy data from 2000 to 2016 for all countries. We will be splitting the data into training and testing sets for model validation purposes. An explanation of the variables used can be found in the appendix.

Our research aims to investigate various health and social economic factors effect on life expectancy by finding the "best fitted" linear regression model.We first fit a model with every reasonable variable to achieve a full model. Then we check the two additional conditions. Using scatter plot of predicted and actual value to check condition 1, and pairwise scatter-plots between predictors to check condition 2. Condition 1 would hold if scattering of points on the scatter plot is random and condition 2 would hold if we observe linear relations between predictors in our pairwise plots. If both condition holds, we proceed to graph out residual plots of predicted value, predictor variables and QQ plots to check for assumptions. Linearity assumption holds if no apparent pattern exists in residual plots, uncorrelated error assumption holds if no large clusters of residuals exists separated from the rest, constant variance exists if no fanning patterns exists in residual plots and normality assumption holds if our QQ plot shows not a lot of deviation from the diagonal line. If above assumptions and/or condition doesn't hold, apply Boxcox transform on variables, if the assumptions and conditions hold or we have tried transforming variables, continue with the model and acknowledge the limitations.

After the previous step, we should have a full model, we aim to reduce the model by first checking for multicollinearity, and remove any variable with VIF larger than 5. We proceed to assess each variables P-values from our full model, remove any insignificant variables, this leads us to a reduced model. But using partial F-test via ANOVA, we test whether the removed variables are significant, if the P-value is large enough, we conclude that removing the variable is a good choice. We repeat this step several times until all of our variables are significant according to the p-value. If we have achieve different models with same number of variables, we compare them each by AIC and adjusted-R squared, those with lower AIC and higher adjusted-R squared is deemed optimal. We should arrive at an ideal model by now, recheck the conditions and assumptions and perform transformation if necessary.

With our ideal model, we check for outliers, leverages and influential points. We will check influential points via DFFITS, we then consider the points, evaluate whether they are removable based on the context. At last, we validate our model by performing the same variable transformation on our testing data separated earlier and using the ideal model from training data, then we compare the coefficients of the variables and the R-squared for our training and testing data, the model is deemed valid if the R-squared is similar and the coefficients are within two standard errors. If deemed invalid, interpret as such, acknowledge the limitations and try to figure out why.

Results

Summary tables consisting of our response and predictor variables of both training and testing data can be seen below, we do see an overall similar means in most variables between the two data sets, but differences in quartile data. This can be expected due to the limited number of data entries, with only 37 data entries for testing data and 146 for training data, this suggests potential problems during validation steps. More discussions on this topic can be seen in the next section.

Table 1: Summary stats of training dataset

	Life Expect.	Mortalityage 1-4	Alcohol Consump/L	BMI	Thin%*	Obese%*	%Access basic water	Health Exp**	Avg Vax***
	Min. :53.04	Min. :0.000100	Min. : 0.00091	Min. :20.60	Min. : 0.100	Min. : 1.000	Min. : 38.85	Min. : 2.312	Min. :39.00
	1st Qu.:66.26	1st Qu.:0.000270	1st Qu.: 1.54618	1st Qu.:23.93	1st Qu.: 1.500	1st Qu.: 3.750	1st Qu.: 79.70	1st Qu.: 4.572	1st Qu.:85.08
	Median :73.33	Median :0.000580	Median : 4.33735	Median :26.20	Median : 3.300	Median : 7.950	Median : 95.17	Median : 6.410	Median :92.50
	Mean :72.10	Mean :0.001943	Mean : 4.99667	Mean :25.64	Mean : 4.445	Mean : 8.236	Mean : 86.75	Mean : 6.646	Mean :88.35
	3rd Qu.:77.25	3rd Qu.:0.002839	3rd Qu.: 7.53801	3rd Qu.:26.90	3rd Qu.: 6.600	3rd Qu.:11.200	3rd Qu.: 99.38	3rd Qu.: 8.433	3rd Qu.:96.67
	Max. :83.08	Max. :0.014615	Max. :20.18246	Max. :32.20	Max. :18.000	Max. :26.700	Max. :100.00	Max. :17.197	Max. :99.00

[a] * = for age 5 to 19

[b] Health expenditure as percentage of GDP

[c] *** = Average Vax rate for measles,polio and diphtheria

Table 2: Summary stats of testing dataset

	Life Expect.	Mortality age1-4	Alcohol Consump/L	BMI	5Thin%*	Obese%*	%Acces s basic water	Health Exp**	Avg Vax***
	Min. :52.94	Min. :0.0001100	Min. : 0.04576	Min. :21.30	Min. : 0.200	Min. : 1.800	Min. : 53.18	Min. : 2.713	Min. :52.33
	1st Qu.:66.42	1st Qu.:0.0002887	1st Qu.: 1.34969	1st Qu.:23.27	1st Qu.: 1.875	1st Qu.: 3.050	1st Qu.: 78.76	1st Qu.: 4.304	1st Qu.:78.92
	Median :74.92	Median :0.0006525	Median : 3.09702	Median :26.10	Median : 5.050	Median : 9.200	Median : 96.17	Median : 5.926	Median :92.33
	Mean :71.72	Mean :0.0022162	Mean : 4.57239	Mean :25.64	Mean : 5.781	Mean : 8.269	Mean : 88.60	Mean : 6.174	Mean :87.77
	3rd Qu.:76.09	3rd Qu.:0.0033112	3rd Qu.: 7.50454	3rd Qu.:27.38	3rd Qu.: 7.175	3rd Qu.:11.750	3rd Qu.: 99.01	3rd Qu.: 7.775	3rd Qu.:95.83
	Max. :84.17	Max. :0.0118050	Max. :11.47825	Max. :32.20	Max. :26.900	Max. :21.700	Max. :100.00	Max. :12.221	Max. :99.00

[a] * = for age 5 to 19

[b] Health expenditure as percentage of GDP

[c] *** = Average Vax rate for measles,polio and diphtheria

According to pairwise scatter plots, our full model doesn't satisfy condition 2, after Boxcox transforming, we use new variables, we use square root of alcohol, $percentage\ of\ those\ who\ are\ thin\ aged\ 5-19^{0.33}$, $percentage\ access\ to\ basic\ water^{4.20}$, $avg\ vax\ rate\ across\ measles, polio, diphtheria^{5.95}$.

For the transformed full model now, normality assumption holds according to QQ plot, linearity assumptions holds as our residual plots improve and we can say we don't observe any patterns, constant variance assumption might not hold as we can see higher residuals for countries/data points with lower life expectancy, uncorrelated error assumption holds as we can only observe one large cluster of residuals with outliers. We will acknowledge this limitation and proceeds. All our variables have VIF <5, so no variable is removed. We reduce our model by observing P-values, and remove variables alcohol and average vaccination rate due to insignificance due to high P-values. We test using partial F-test via ANOVA and both full and reduced model 1, and our resulted P-value suggested we should remove the variables. We attempt to further reduce our model by removing two more variables with bigger p-value but slightly significant, we repeat our partial F-test and this time our P-value suggests we should keep the variables. By comparing the AIC and adjusted R-squared, we can see that reduced model 1 with 6 predictors has the lowest AIC with similar adjusted R-squared despite having less variables, this is our ideal model.

A total of 11 leverages, 6 outliers and 10 influential points exist. Given the context, only countries where major natural disasters and conflicts occurred can be removed(like Haiti Earthquake in 2010)as such change in life expectancy can't be explained by our variables. Upon inspection, no data is removed, none fit the criteria and every data is valuable when researching on a global level.

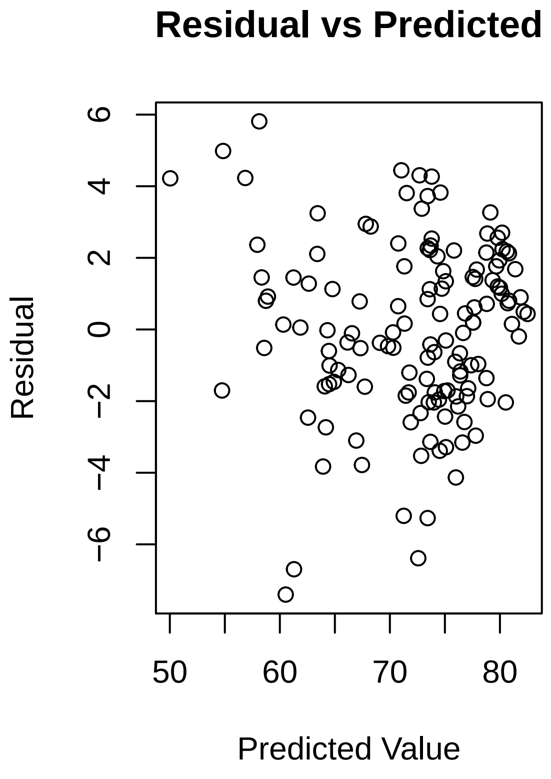
According to Figure 1 in the appendix, condition 1 holds for our final model, we conclude that our linear model is sufficient, condition 2 might not hold due to only observing linear relations between few predictors. The four assumptions all hold except for constant variance, experiencing similar larger variances for lower life expectancy countries as observed below and figure 2 in appendix.

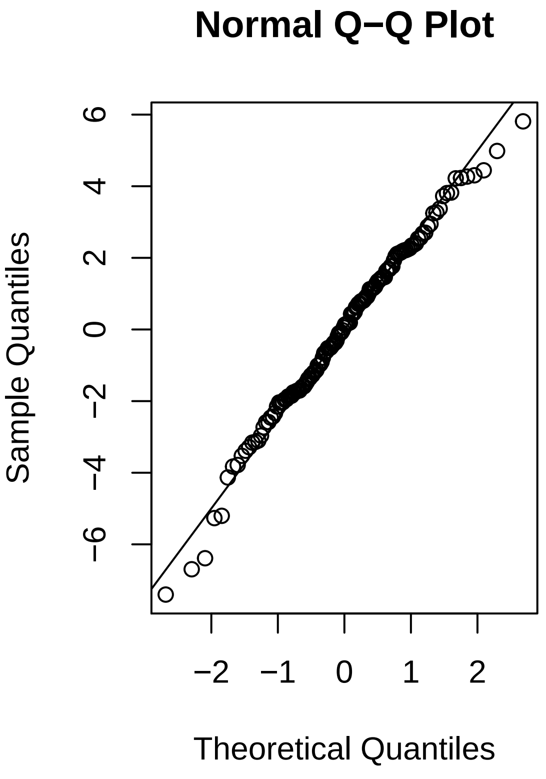
Validation is done by running the same model on our testing data set and applying the same transformation,from table below we observe similar coefficients for some variables between training and testing but other coefficients are different beyond two standard error, we observe similar adjusted R-squared. The two conditions doesn't hold for this model and the same can be said about the four assumptions, our model is deemed invalid by using our testing data.

Discussion

From our table below, for our final training model, on an average for a country,we do see a very steep negative slope on child mortality rate, a 1252 decrease in life expectancy per 1% point increase in mortality rate for those age 1-4, we also expect a 1.265 decrease in life expectancy per 1 point increase in BMI, a 0.189 increase in life expectancy per 1 point increase in percentage of obesity among age 5-19 and a 0.188 increase in life expectancy per 1% point increase in current health expenditure as percentage of GDP. Interpretation for thinness prevalent in age 5-19 and access to basic water are harder due to transformations, but we can see a 4.19 drop in life expectancy per 1% point increase in $percentage\ of\ prevalent\ of\ thinness\ in\ age\ 5-19^{0.33}$ and a 0.00000004 increase in life expectancy per 1% point in $percentage\ access\ to\ basic\ water^{4.2}$. These coefficients explains these variables influence on life expectancy on a global level, and can be used as guidelines when a government aims to increase life expectancy.

	Training model	Testing model
(Intercept)	103.078	130.538
	(5.673)	(11.449)
age1_4mort	−1252.499	−1436.294
	(136.649)	(273.577)
bmi	−1.265	−1.652
	(0.214)	(0.350)
age5_19thintrans	−4.190	−9.762
	(0.674)	(1.953)
age5_19ob	0.189	−0.090
	(0.072)	(0.189)
basic_watertrans	4×10^{-8}	6×10^{-8}
	(5×10^{-9})	(1×10^{-8})
che_gdp	0.188	−1.160
	(0.088)	(0.384)
Num.Obs.	138	36
R2	0.892	0.915
R2 Adj.	0.887	0.897
AIC	651.8	181.4
BIC	675.3	194.0
Log.Lik.	−317.919	−82.681
RMSE	2.42	2.41





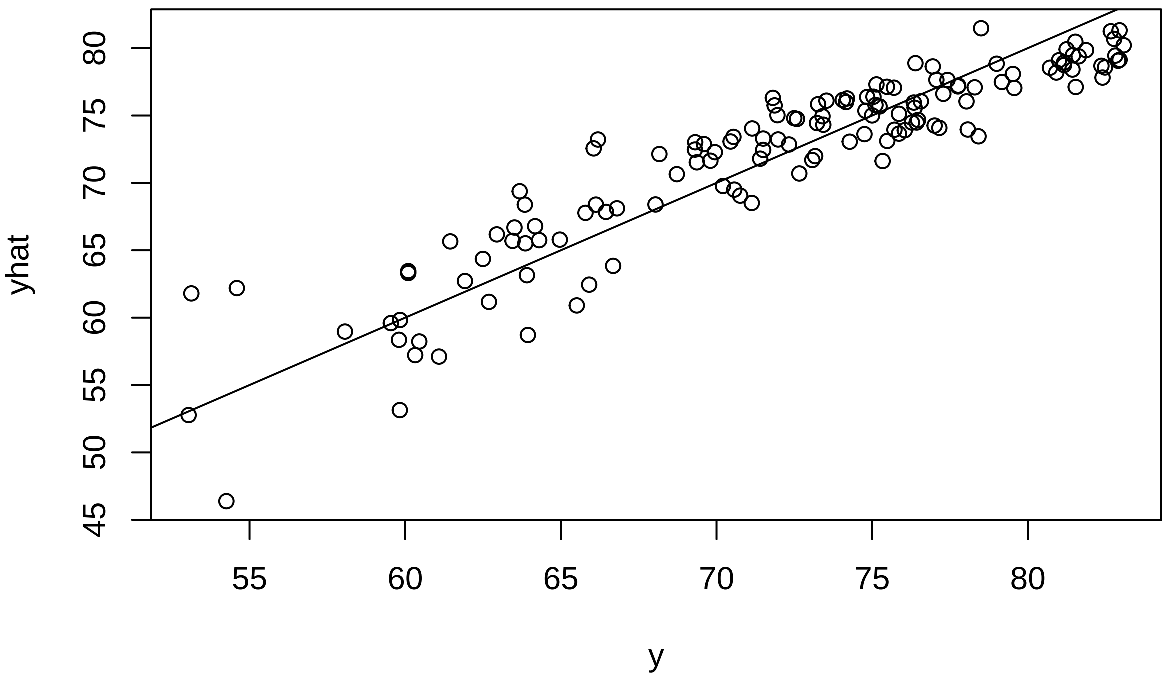
Limitations:

The model doesn't satisfy the constant variance assumption even after transformation as observed above, this leads to low prediction accuracy for countries with very low life expectancy values, it can't be corrected due to the nature of the data, the model is also not validated due to limited data in testing set, good representation of the variables cannot be guaranteed,thus achieving very different coefficients. Given the nature of our research topic, our data entries are limited. One reason that our model isn't validated is because by splitting into training and testing set, we are limiting only about 35 data points during the validation step. The model is deemed invalid, so we should interpret with care but it can still serve as guidelines for policy makers, and interpret coefficients as trends rather than precise estimates.

Appendix

An explanation of variables used, all numerical.
life_expect: Life expectancy in years.
age1_rmot: Mortality rate in age 1-4.
alcohol: Per capita consumption per litre of pure alcohol.
bmi: BMI, body mass index.
age5_19thin: Prevalence of thinness in age5-19, in percentage.
age5_19ob: Prevalence of obesity in age5-19, in percentage.
basic_water: Access to basic water, in percentage.
che_gdp : Current health expenditure as percentage of GDP, in percentage.
avgvac: Average vaccination rate across measles, polio, diphtheria, in percentage.

Figure 1: Plots checking for additional conditions for final model



```
## integer(0)
```

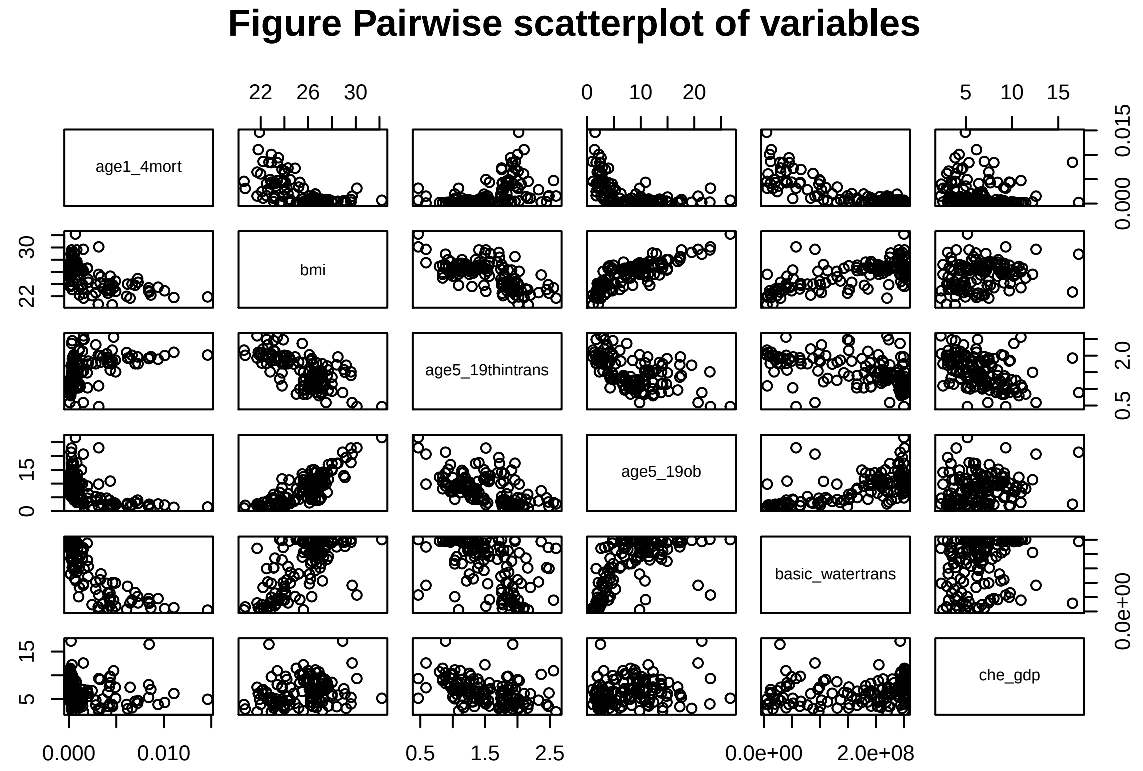
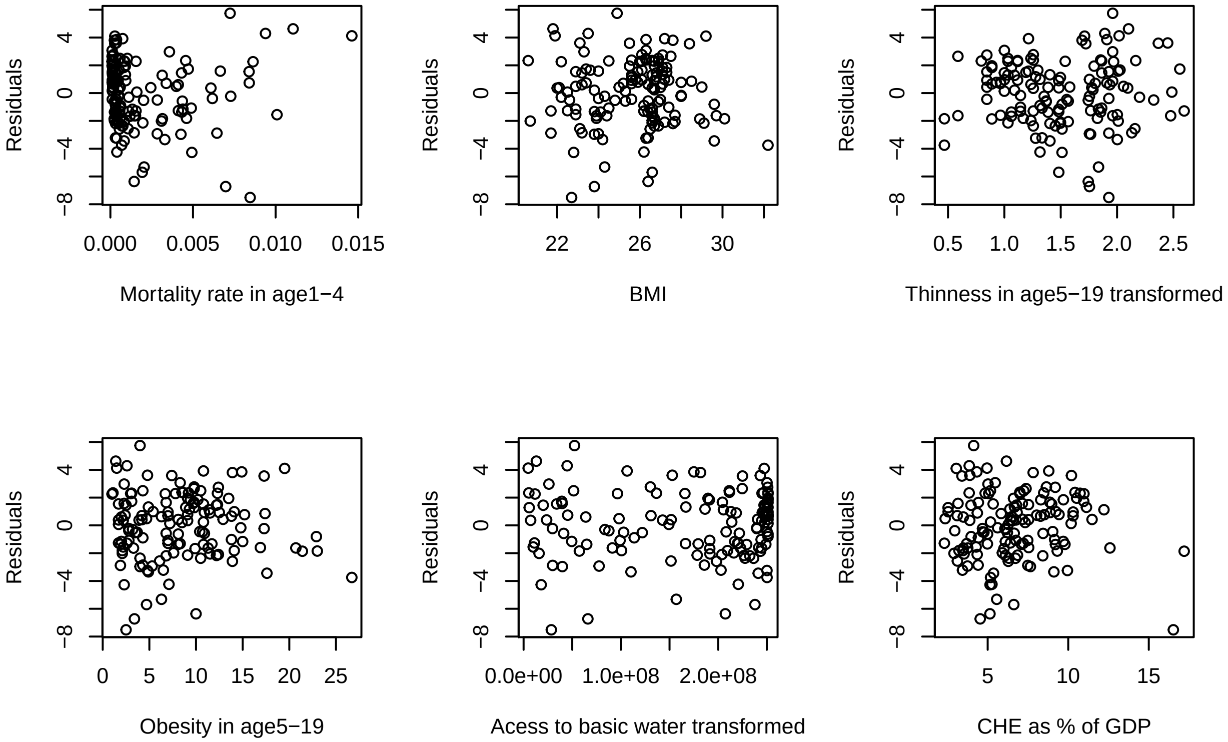


Figure 2: Remaining residual plots for final model



Citation

Murillo, P. I. L. (2021, July 14). The life expectancy: What is it and why does it matter. CENIE. Retrieved October 19, 2022, from https://cenie.eu/en/blogs/age-society/life-expectancy-what-it-and-why-does-it-matter

Mariani, F., Pérez-Barahona, A., & Raffin, N. (2009, December 2). Life expectancy and the environment. Journal of Economic Dynamics and Control. Retrieved October 19, 2022, from https://www.sciencedirect.com/science/article/pii/S0165188909002164?casa_token=Oy2OecoPqwIAAAA A%3AKH2azMuYdj4QvqE2V2NQ2VLc98iuyQyiJtK0RE2ipxpBNoy6zLkQNZDVioiswi6q7ybTXemKjt-x

Freeman, T., Gesesew, H. A., Bambra, C., Giugliani, E. R. J., Popay, J., Sanders, D., Macinko, J., Musolino, C., & Baum, F. (2020, November 10). Why do some countries do better or worse in life expectancy relative to income? an analysis of Brazil, Ethiopia, and the United States of America - International Journal for equity in health. BioMed Central. Retrieved October 20, 2022, from https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-020-01315-z

Kabir, M. (2008). Determinants of Life Expectancy in Developing Countries. Retrieved October 20, 2022, from https://www.jstor.org/stable/40376184?searchText=life%20expectancy%20developing&searchUri=%2Faction%2FdoBasicSearch%3FQuery%3Dlife%2Bexpectancy%2Bdeveloping&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&refreqid=fastly-default%3A0b04f948b1b29dfc6eba08c0325479cf

modelsummary: Beautiful, customizable, publication-ready model summaries in R. Retrieved December 19, 2022, from https://www.rdocumentation.org/packages/modelsummary/versions/0.2.0

Vincent A. modelsummary: regression tables with side-by-side models. Retrieved December 19, 2022, from https://vincentarelbundock.github.io/modelsummary/articles/modelsummary.html