Rotterdam School of Management
Erasmus University

Assignment 1

Linear Regressions with Cross-Sectional and Panel Data

Group 16

Khoi Gia Pham 523755kg

David Lensen 667268dl

Catalina Betivu 532560cb

Victoria Noodt 690974vn

Mashal Aliasi 543873ma

Course: BIM Research Method

MSc Business Information Management

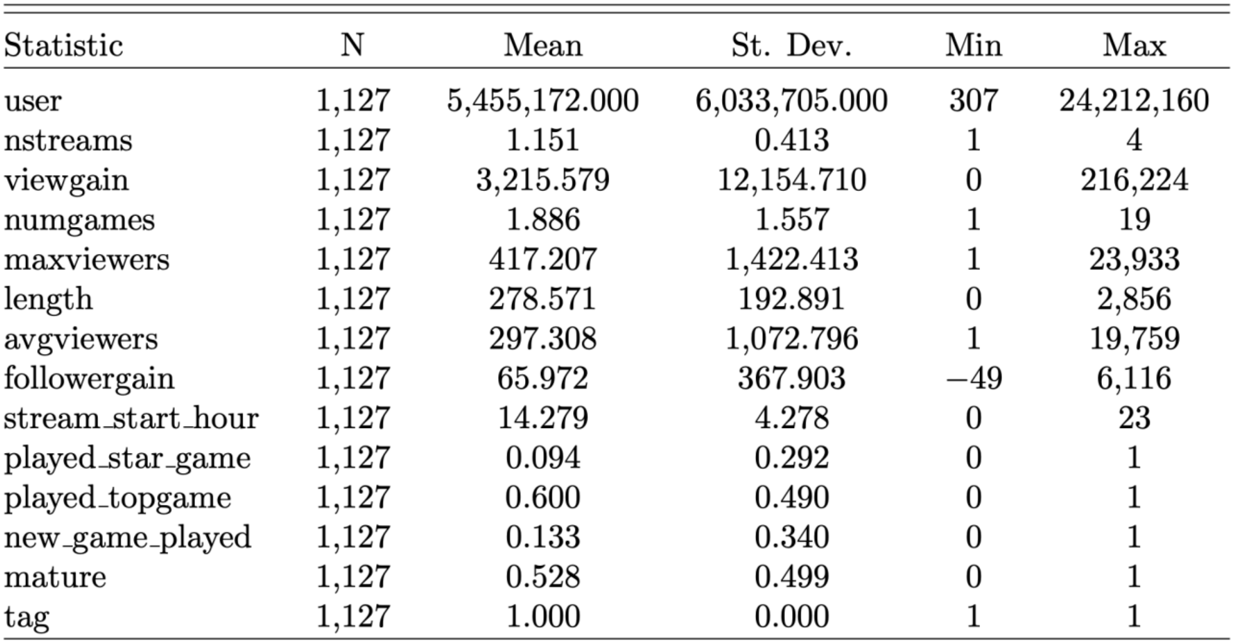
Rotterdam School of Management, Erasmus University

RSM - a force for positive change

Exercise 1. Linear regression analysis with cross-sectional data

1a. Descriptive statistics of the dataset

This section briefly discusses the mean values of the variables in table 1. The variable *length* has a mean of 278.571 minutes, which is the average number of minutes participated streamers streamed on the observation day. Next, the variable *nstreams* presents the average number of streams the streamers started, which is 1.151 streams (*nstreams*). The mean viewers gained was 3215.579 viewers (*viewgain*), and the average number of games played is 1.886 (*numgames*). Furthermore, the maximum number of viewers viewing the stream at any given point on the observation day was on average 417.207 viewers (*maxviewers*), the mean number of viewers viewing the stream was 297.308 viewers (*avgviewers*) and the streamer gained on average 65.972 followers on observation day (*followergain*). Moreover, on average, 9,4% of streamers played Fortnite (*played_star_game*), 60% played their top game (*played_topgame*), and 52,8% of the content was rated as mature (*mature*). Finally, the mean of the hour in which the streams started is 4.270 (*stream_start_hour*)



Statistic	N	Mean	St. Dev.	Min	Max
user	1,127	5,455,172.000	6,033,705.000	307	24,212,160
nstreams	1,127	1.151	0.413	1	4
viewgain	1,127	3,215.579	12,154.710	0	216,224
numgames	1,127	1.886	1.557	1	19
maxviewers	1,127	417.207	1,422.413	1	23,933
length	1,127	278.571	192.891	0	2,856
avgviewers	1,127	297.308	1,072.796	1	19,759
followergain	1,127	65.972	367.903	−49	6,116
stream_start_hour	1,127	14.279	4.278	0	23
played_star_game	1,127	0.094	0.292	0	1
played_topgame	1,127	0.600	0.490	0	1
new_game_played	1,127	0.133	0.340	0	1
mature	1,127	0.528	0.499	0	1
tag	1,127	1.000	0.000	1	1

Table 1. Descriptive statistics on twitch_data

1b. Relationship between dependent variable *length* and three independent variables n*streams, viewgain and numgames*

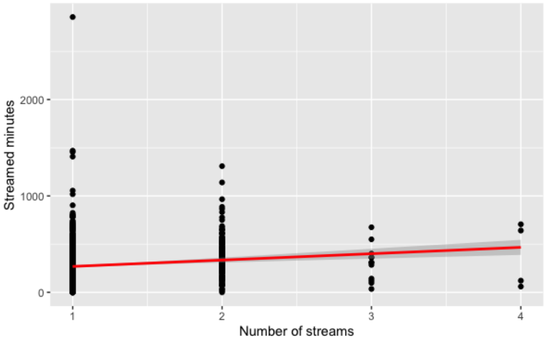


Figure 1. length vs. *nstreams*

Firstly, the upward slope regression line (red line) from the first scatter plot depicts a slight positive correlation between the dependent variable *length* and independent variable *nstreams*. Meaning that when the number of streams increases, the streaming time also increases. There is also an abnormal observation located above 2000 streamed minutes when the number of streams is 1

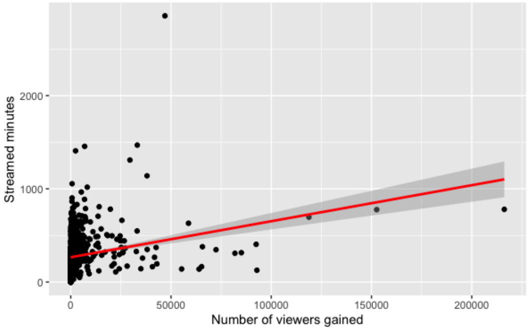


Figure 2 length vs. viewgain

Secondly, figure 2 shows the relationship between *length* and *viewgain*. The upward-sloping regression line indicates a stronger positive correlation, meaning that when number of viewers increases, the number of minutes streamed also increases. There are outliers near 3000 streamed minutes when the number of viewers gained is 5000.

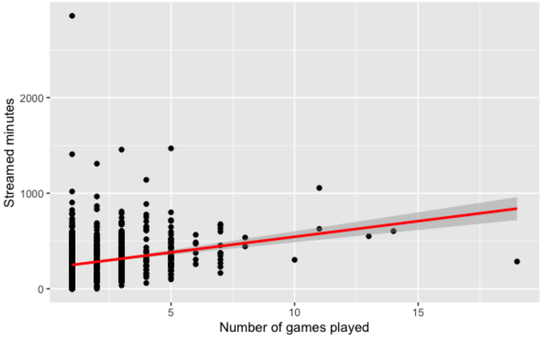


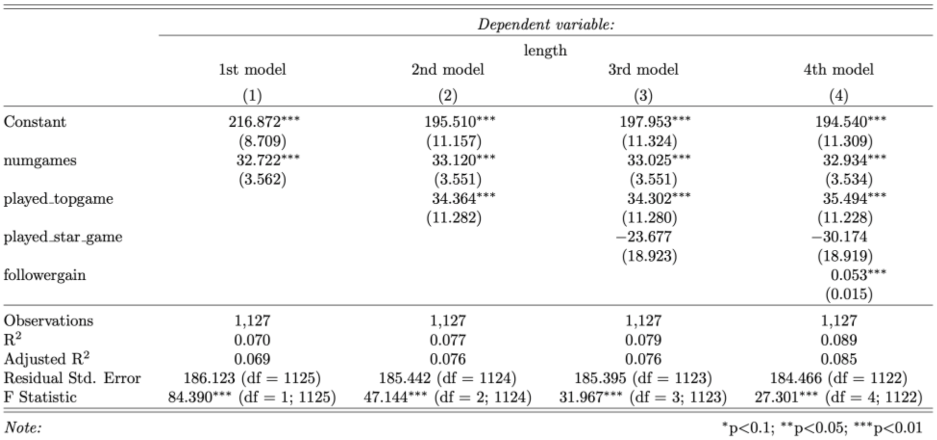
Figure 3 length vs. *numgames*

Thirdly, a positive correlation is also recorded between the *length* and *numgames*. The upward slope regression line indicates that when the number of games played by a streamer on observation day increases, the number of minutes also streamed increases. There are outliers above the 2000 streamed minutes when the number of games played is around 1 and around 400 streamed minutes when the number of games played is 18.

1c. Four different linear regression models

We selected *Numgames, Played_topgame, Played_star_game, and Followergain* as four independent variables for the following reasons. First, as for *Numgames,* we are intuitively convinced that the more games are being played, the more minutes a player will stream. Second, *Played_topgame* is because the player tends to stream longer while enjoying playing his or her favourite games. Third, regarding *Played_star_game,* since Fortnite has been one of the most popular games over the past years, we want to examine its addictiveness on the stream length. Finally, we include *Followergain* because intuitively, increases in followers gained encourage the player to stream longer.

Table 2 below presents the regression results of the four different models



	Dependent variable:			
	length			
	1st model	2nd model	3rd model	4th model
	(1)	(2)	(3)	(4)
Constant	216.872***	195.510***	197.953***	194.540***
	(8.709)	(11.157)	(11.324)	(11.309)
numgames	32.722***	33.120***	33.025***	32.934***
	(3.562)	(3.551)	(3.551)	(3.534)
played_topgame		34.364***	34.302***	35.494***
		(11.282)	(11.280)	(11.228)
played_star_game			−23.677	−30.174
			(18.923)	(18.919)
followergain				0.053***
				(0.015)
Observations	1,127	1,127	1,127	1,127
R^2	0.070	0.077	0.079	0.089
Adjusted R^2	0.069	0.076	0.076	0.085
Residual Std. Error	186.123 (df = 1125)	185.442 (df = 1124)	185.395 (df = 1123)	184.466 (df = 1122)
F Statistic	84.390*** (df = 1; 1125)	47.144*** (df = 2; 1124)	31.967*** (df = 3; 1123)	27.301*** (df = 4; 1122)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2. Linear regression results table

The regression equations and of the four models are presented below:

1. *length = 32.722 * numgames + 216.872.*
2. length = *33.120 * numgames + 34.364 * played_topgame + 195.510*
3. length = *33.025 * numgames + 34.302 * played_topgame - 23.677 * played_stargame + 197.953*
4. length = *32.934 * numgames + 35.494 * played_topgame - 30.174 * played_stargame + 0.053 * followergain + 194.540*

We focus on explaining coefficients from the final model since it includes all IVs and has the highest $\mathbf{R^2}$. First, the estimated coefficient of *numgames* is positive (32.934) and significant (p-value < 0.05). Second, the estimated coefficient of *played_topgame* is positive (*35.494*) and significant (p-value < 0.05). Third, the estimated coefficient of *played_stargame* is negative (-*30.174*) and not sig significant (p-value > 0.05). Fourth, the estimated coefficient of *played_topgame* is positive (*35.494*) and significant (p-value < 0.05).

We observed that $\mathbf{R^2}$ increases every time an additional variable is added to the model. In theory, R^2 indicates how much the variance of DV *length* is captured in our model. And R^2 always increase if the number of independent variables increases. The R^2 from the (4th model) is 0.089, meaning that 8.9% of the variance for DV: *length* is explained by four IVs: *nstreams*, *viewgain* and *numgames*

1d. How to make the streamer stream six hours a day

Only the independent variables (IVs) with statistically significant estimated coefficients (p-value <0.05) are relevant to make the streamer stream six hours (360 minutes) a day. And we solve the regression equations from the 4th model to calculate the required IVs changes (identified as *x*).

1. *Numgames:* $360 = 194.540 + 32.934x \Rightarrow x = 5.024$
2. *Player_topgame:* $360 = 194.540 + 35.494x \Rightarrow x = 4.662$
3. *Followergain:* $360 = 194.540 + 0.053x \Rightarrow x = 3121.887$

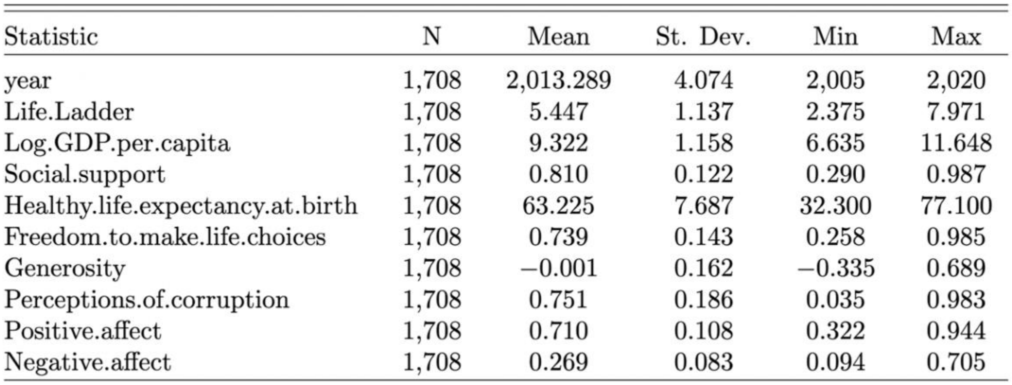
To make the streamer stream 6 hours long, the number of games played by the streamer should be 5.024, and the number of follers gained should be 3,121.887 followers on the observation day. However, *player_topgame* can only have binary values. This means that when the streamer plays their top game, the *length* only increases by 35.494 minutes.

1e. Theory on the causality of effects

The regression coefficients cannot be interpreted as a causal effect in this case, as not all the three conditions for causality are met. Based on the observation fay, we can only verify the condition that length and other variables vary together. However, the two other conditions are not fulfilled. First, other possible causal factors are not eliminated, which may cause the relationship between the dependent variable *length* and the independent variables to be a coincidence. This relationship is known as a spurious correlation. Second, time order is ambiguous, and we cannot verify that *length* happens before other independent variables.

Exercise 2. WHR Data

2a. Descriptive statistics of the dataset



Statistic	N	Mean	St. Dev.	Min	Max
year	1,708	2,013.289	4.074	2,005	2,020
Life.Ladder	1,708	5.447	1.137	2.375	7.971
Log.GDP.per.capita	1,708	9.322	1.158	6.635	11.648
Social.support	1,708	0.810	0.122	0.290	0.987
Healthy.life.expectancy.at.birth	1,708	63.225	7.687	32.300	77.100
Freedom.to.make.life.choices	1,708	0.739	0.143	0.258	0.985
Generosity	1,708	−0.001	0.162	−0.335	0.689
Perceptions.of.corruption	1,708	0.751	0.186	0.035	0.983
Positive.affect	1,708	0.710	0.108	0.322	0.944
Negative.affect	1,708	0.269	0.083	0.094	0.705

Table 3. Descriptive statistics on whr_data

Table 3 summarizes the descriptive statistics of the panel data collected to measure a country's aggregated happiness from 2005 to 2020. Overall, the total number of observations from our collected data is 1,708, and the number of countries surveyed (*Country.name*) is 155 since we removed the missing values in the original dataset.

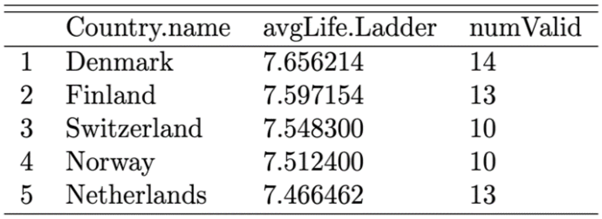
Most of the observations were made in 2013, as 2,013.289 is the mean value of the variable *year*. Notably, the happiness score, *Life.Ladder*, has a mean of 5.447 (on a scale from 1 to 10), entailing the national average responses to the question of life evaluations.

As for two economic-related variables, *Log.GDP.per.capita* and *Perceptions.of.corruption* variables average at 9.322 and 0.751, respectively. The former variable reflects approximation to GDP growth in Purchasing Power Parity (PPP), while the latter illustrates a relatively high perception of governmental and business corruption.

Concerning the societal and health-related variables, the binary variable *Social.support*, which refers to the availability of support in case of trouble, has a relatively high mean value of 0.810. Next, the *Freedom.to.make.life.choices* variable, which refers to the national average of satisfaction in freedom, also has a high mean of 0.739. Meanwhile, the *Generosity* variable, which represents the residual of regressing the national average of donations decisions in the past month, has a negative mean of -0.001. Finally, the healthy life expectancy variable, *Healthy.life.expectancy.at.birth*, averages approximately 63 years.

Moreover, the *Positive.affect* variable, equating happiness, love, and enjoyment effects, has a mean of 0.710. This figure is considerably higher than that of *Negative.affect,* which is 0.269. *Negative.affect* is the average of worry, sadness and anger affects.

2b. Five countries with the highest average happiness score & Normal distribution test



	Country.name	avgLife.Ladder	numValid
1	Denmark	7.656214	14
2	Finland	7.597154	13
3	Switzerland	7.548300	10
4	Norway	7.512400	10
5	Netherlands	7.466462	13

Table 4. Top five countries based on average happiness score

The five countries with the highest happiness on average are Denmark, Finland, Switzerland, Norway, and the Netherlands. Table 4 presents the countries' average happiness score (avgLife.Ladder) and their number of complete observations (numValid) during the period surveyed.

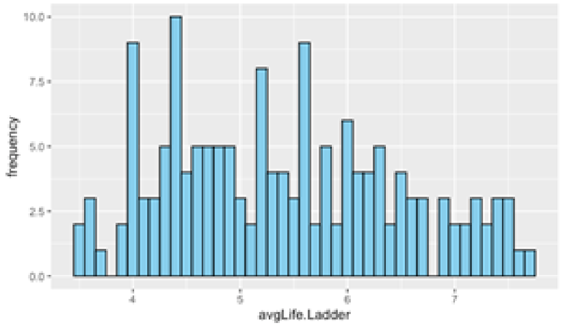


Figure 2 Average happiness score by frequency

As shown in Figure 3, the histogram visualizing the average happiness score is right-skewed. Thus, we can graphically conclude that the happiness variable does not follow a normal distribution. Our Shapiro-Wilk test results further support this claim. The test p-value = 0.0009622 < 0.005 suggest that we can reject the null hypothesis: the variable is normally-distributed.

2c. Regression model with happiness

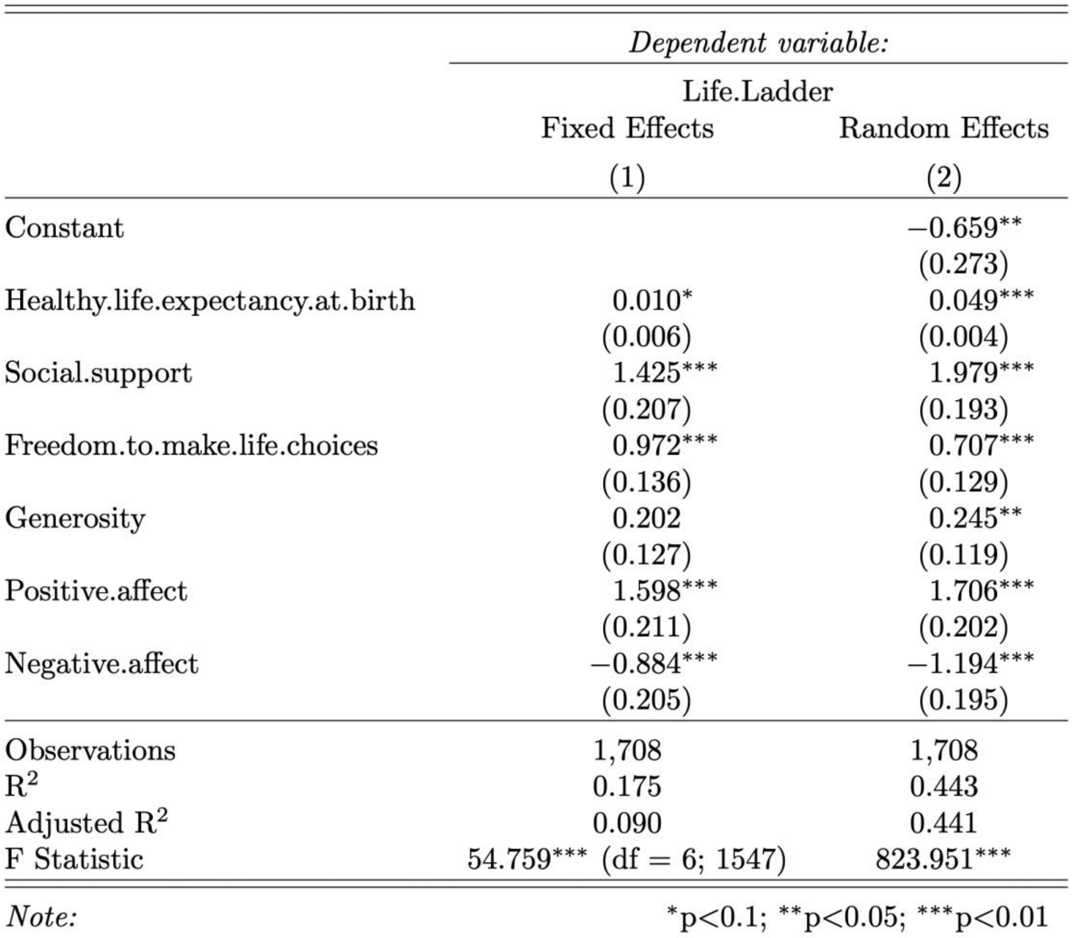
Our regression model includes the dependent variable *Life.Ladder* and six independent variables: *Healthy Life Expectancy (HLE), Social support, Freedom to make life choice, Generosity, Positive affect, and Negative affect*.

1. Regression equations

Model	**Regression equations**
Fixed effects	$\text{Life.Ladder} = 0.010*\textit{Healthy.life.expectancy.at.birth}_{it} + 1.425*\textit{Social.Support}_{it} +$ $0.972*\textit{Freedom.to.make.life.choices}_{it} + 0.202*\textit{Generosity}_{it} + 1.598*\textit{Positive.affect}_{it}$ $+ (-0.884)*\text{Negative.affect}_{it}+\gamma_{it}+\varepsilon_{it}$ $(i =1,\ldots,155 ; t=1,\ldots,16)$
Random effects	$\text{Life.Ladder} = -0.659 + 0.049*\textit{Healthy.life.expectancy.at.birth}_{it} +$ $1.979*\textit{Social.Support}_{it} +$ $0.707*\textit{Freedom.to.make.life.choices}_{it} + 0.245*\textit{Generosity}_{it} + 1.706*\textit{Positive.affect}_{it}$ $+ (-1.194)*\text{Negative.affect}_{it}+\varepsilon_{it}$ $(i =1,\ldots,155 ; t=1,\ldots,16)$

Table 5. Fixed effects & Random effects models regression equations

2. Fixed effects and random effects model results



	Dependent variable:	
	Life.Ladder	
	Fixed Effects	Random Effects
	(1)	(2)
Constant		-0.659^{**}
		(0.273)
Healthy.life.expectancy.at.birth	0.010^{*}	0.049^{***}
	(0.006)	(0.004)
Social.support	1.425^{***}	1.979^{***}
	(0.207)	(0.193)
Freedom.to.make.life.choices	0.972^{***}	0.707^{***}
	(0.136)	(0.129)
Generosity	0.202	0.245^{**}
	(0.127)	(0.119)
Positive.affect	1.598^{***}	1.706^{***}
	(0.211)	(0.202)
Negative.affect	-0.884^{***}	-1.194^{***}
	(0.205)	(0.195)
Observations	1,708	1,708
R^2	0.175	0.443
Adjusted R^2	0.090	0.441
F Statistic	54.759^{***} (df = 6; 1547)	823.951^{***}
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 6. Fixed effects and Random effects models results

3. Coefficients interpretations

In the panel data *whr_data*, entity index is *Country.name* and time index is *year*. In other words, *Country.name* is the fixed effect level, and *year* is the timestamp. Both the Fixed effects (FE) and Random effects model (RE) that we used present specification estimates fixed *Country.name* effect (One-way individual effect). And the models' resulting coefficients are discussed as follows.

In the FE model, the first IV, *Healthy.life.expectancy.at.birth* has a positive coefficient of 0.010 but is not significant (p-value>0.05). Secondly, the coefficient of *Social.support* equals 1.425 and is significant (p-value<0.01), suggesting that the happiness score increases by 1.425 when there is *Social.Support*. The third coefficient implies a 0.972 increase in happiness score if *Freedom.to.make.life.choices* increase by 1, and it is significant (p-value<0.01). Next, the coefficient of *Generosity* is positive (0.202) but is insignificant. Meanwhile, the coefficients of

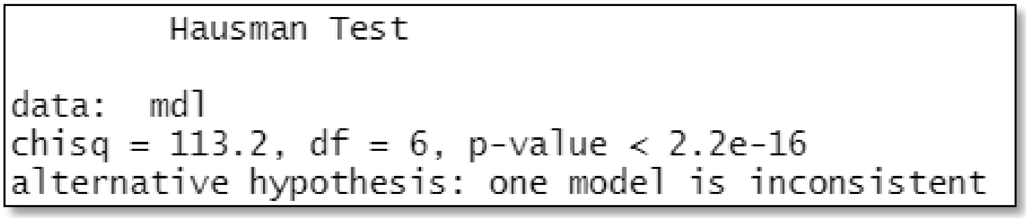
Positive.affect and *Negative.affect* are 1.598 and -0.884, respectively. And both coefficients are significant (p-value<0.01). One unit increase in the former suggests an increase of 1.598 in happiness score, while one unit increase in the latter suggests a 0.884 decrease in happiness score.

In the RE model, the constant is -0.659 and is significant (p-value<0.05), meaning the average happiness score (*Life.Ladder)* is -0.659 when all IVs are zero. All the IVs' coefficients in the model are significant at 0.01 level, except for that of *Generosity*, which is significant at 0.05 level. The first variable, *Healthy.life.expectancy.at.birth*, has a positive coefficient of 0.049, suggesting a 0.049 increase in happiness score when this variable increases by 1 unit. Next, the coefficient of *Social.support* equals 1.979, suggesting that the happiness score increases by 1.979 when there is *Social.support.* The third coefficient implies a 0.707 increase in happiness score if *Freedom.to.make.life.choices* increase by 1. Also, the fourth coefficient implies a 0.245 increase in happiness score if G*enerosity* increases by 1. Meanwhile, the coefficients of *Positive.affect* and *Negative.affect* are 1.706 and -1.194, respectively. One unit increase in the former suggests an increase of 1.706 in happiness score, while one unit increase in the latter suggests a 1.194 decrease in happiness score.

When comparing both models, a few things are interesting. For all independent variables, the strength of the coefficient is higher for the random effects model, except for the *Freedom.to.make.life.choices* variable. Secondly, the R^2 of the random effects model is much higher (0.443) as opposed to the fixed effects model (0.175).

2d. Hausman Test

To decide whether the Fixed effects or Random effects model should be preferred for our regression specification, we used **Hausman Test: phtest**



```
        Hausman Test

data:  mdl
chisq = 113.2, df = 6, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

Figure 4. Hausman test result

Results of **phtest**: p-value < 2.2e-16 < 0.05 indicates the test is significant. Thus, the null hypothesis that the preferred model is the Random effects model can be rejected. We should choose the Fixed Effects model rather than the Random Effects model. Two additional theoretical reasons to support the preference for the Fixed effects model are:

- Fixed effects models leverage *within* unit variation by considering the fact that one entity (Country.name) has multiple observations. Meanwhile, Random effects models consider a mix of *between* and *within* variation estimates, and assume the error term is not correlated with the IVs
- Fixed effects models also control for time-constant unobserved factors (or unobserved heterogeneity/ characteristics) that may influence the regression results.