Statistics & Methodology Group Project

Group 01 Members - ANR
Sema Karan - 924823
Cigdem Dalbudak -371926
Levi Pols - 320513
Vratislav Frits Kosovsky - 341057

Submission Date: 12 October 2018

Questions & Answers

2.1 Multiple linear regression

1. Which countries are represented in these data?

The countries represented in the data as table below;

Code (V2)	156	276	356	643	840
Country	China	Germany	India	Russia	United States

2. What are the sample sizes for each country represented in these data?

The sample sizes per country in the data as table below;

Code (V2)	156	276	356	643	840
Country	China	Germany	India	Russia	United States
Sample Size	2300	2046	5659	2500	2232

3. Overall, is there a significant effect of country on feelings of happiness?

Yes, there is a significant effect of country on feelings of happiness *(DF=4, MSE=60.74, F-value=132.8, p-value: < 2.2e-16).*

4. Which country has the highest level of feelings of happiness?

US has the highest feeling of happiness. A linear regression model with unweighted coding shows the following statistics *(slope = -0.1742, t-value = -13.8363, p = 2.869637e-43).*

5. Which country has the lowest level of feelings of happiness?

Russia has the lowest feeling of happiness. A linear regression model with unweighted coding shows the following statistics (*slope = 0.19883, t-value = 16.4881, p = 1.555794e-60)*

6. How do the country-specific levels of feelings of happiness change after controlling for subjective state of health?

After controlling for subjective state of health, the country-specific levels of feelings of happiness change, as shown in the following table:

	intercept	U.S	China	Germany	India	Russia
Without SSOH	1.9103	-0.1743	0.0835	0.0007	-0.1088	0.1988
With SSOH	1.2506	-0.0986	0.0923	0.0173	-0.0747	0.0637

A change in pattern did also occur. Instead of Russia, now China is the country with the lowest level of Feelings of Happiness (*slope = 0.0923, t-value = 7.994, p = 1.402361e-15).*

2.2 Continuous variable moderation

1. After controlling for country, does the importance people afforded to democracy (DemImp) significantly predict the extent to which they think their country is being run democratically (DemRun)?

Yes, the importance people afforded to democracy significantly predicts the extent to which they think their country is being run democratically *(R^2=0.0862, F=1581.1 , $p < 0.001$).*

2. After controlling for country, does the DemImp → DemRun effect vary as a function of peoples' satisfaction with their lives (SWL)?

Yes, the DemImp → DemRun effect varies as a function of SWL *(R^2=0.0002, F=4.0839, p = 0.0433)*

3. Within what range of SWL is the DemImp → DemRun simple slope from Question 2 statistically significant?

We probed the interaction using the Johnson-Neyman technique, to find the region of significance wherein the conditional effect of DemImp on DemRun is statistically significant. The two roots the Johnson-Neyman technique produced are:

low	*high*
29.8366	*1581.3792*

Neither of the roots fall within the observed range of SWL. This means that we are in one of two situations:

1. *The focal effect is significant across the entire range of SWL.*
2. *The focal effect is not significant anywhere within the range of SWL.*

We tested the simple slope of DemImp → DemRun at the mean of SWL, which was significant *(β= 0.270985, SE = 0.0083, t = 32.655, p < 0.001).*

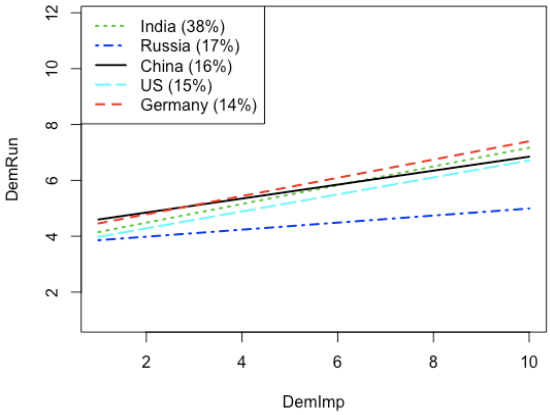
This means that we are in the first situation; the DemImp → DemRun effect is significant across the entire range of SWL.

2.3 Categorical variable moderation

1. After controlling for SWL, does the DemImp → DemRun effect vary significantly by country?

Yes, after controlling for SWL the effect does vary significantly by country. *(F Statistic:28.645, p-value: 2.2e-16).*

2. Visualize the results from Question 1 in a suitable way.



3. For which country is the effect of DemImp on DemRun strongest, after controlling for SWL?

The effect is the strongest for India. See the Simple slopes table below *(β=0.33561, SE = 0.01163, t = 28.8556, p = 3.75e-178)*

Simple slopes table

Countries	Simple.Slopes	Standard.Errors	t.values	p.vlaues
India	0.33561	0.01163069	28.855544	3.75E-178
Germany	0.3266343	0.02709018	12.057298	2.55E-33
US	0.3039117	0.02127616	14.284139	5.57E-46
China	0.2500002	0.02596144	9.629674	6.95E-22
Russia	0.1260681	0.01663399	7.578945	3.69E-14

4. For which country is the effect of DemImp on DemRun weakest, after controlling for SWL?

The effect is the weakest for Russia. See the Simple slopes table *(β=0.1260681, SE = 0.00167, t = 7.578945, p = 3.69e-14)*

5. Are the simple slopes referenced in Questions 3 and 4 statistically significant?

The simple slopes mentioned in Questions 3 and 4 are all statistically significant, as the Simple slopes table shows *(p = 3.75e-178 < 0.05, p = 3.75e-178 <0.05)*

2.4 Predictive modeling

In this section, you will be building linear regression models to predict people's reported satisfaction with the financial situation of their household (FinSat).

1. Select and list three (theoretically justified) sets of predictors (or functions thereof, e.g., interactions or polynomials) to use in predicting FinSat.

Our selected sets of variables are demonstrated below;

	Predictors				
Set 1	Country Code	Marital Status	Age	Sex	CountChild[1]
Set 2	Age	HighestEd[2]			
Set 3	IncomeScale[3]	RichImportant[4]			

[1] How many children do you have?

[2] Education Level

[3] Scale of incomes

[4] It is important to this person to be rich; to have a lot of money and expensive things

2. Briefly explain why you expect the three sets of predictors you chose in Question 1 to perform well. That is, explain your rationale for defining these three sets.

For the first set of predictors, we intended to group the indicators which are mainly close to addressing the financial position of households and therefore can be used in predicting the satisfaction of the financial situation of households. After controlling for country, marital status and number of children are related to finance in the house affairs because being married and having children means relatively more costs and therefore more financial burdens on households. Age and gender of the householders are also strong predictors affecting the income and expenses in the household.

Our second set of predictors is associated with the expectation of the individuals, which in turn can give proper results in predicting the satisfaction of the financial position of households. The starting point here is that with the increase in age and increase in education level, individuals become more sceptical about their lives including financial issues and satisfaction in their life.

For the last set of predictors, we intended to group real financial position of household and his/ her attitude about giving importance to be rich. Current situation and expectations on the other hand are strong in the sense that they identify satisfaction about his /her current situation.

3. Use 10-fold cross-validation to compare the predictive performance of the three models define in Question 1.

The 10-fold cross-validation errors of the models are below:

Models	**CVE**
SatFinHou~CountryCode + MaritalStatus + Age + Sex + CountChild	5.539912
SatFinHou~Age + HighestEd	5.844054
SatFinHou~IncomeScale + RichImportant	5.003561

4. Which of the three models compared in Question 3 performed best?

Our third model in which we used income scale and giving importance to be a rich person as predictors is the best among our three model *(min CVE= 5.003561.)*

5. What is the estimated prediction error of the best model?

4.9154 is the estimated prediction error of our best model.

6. Based on the selection you made in Question 4, what can you say about the attributes that are are important for predicting financial satisfaction?

According to the performances of our models, income scale of households and giving importance to be a rich person are important in predicting financial satisfaction of households. For the model, the p- value associated with the F-statistics is nearly zero, therefore we have strong evidence that at least one of our predictors in our best model is associated with financial satisfaction of households.

Summary of the regression model is below:

```
Residuals:
   Min      1Q  Median      3Q     Max
-7.6537 -1.5066  0.1075  1.5670  6.2547

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.13119    0.07120   43.98   <2e-16 ***
IncomeScale    0.45951    0.01010   45.48   <2e-16 ***
RichImportant  0.15457    0.01365   11.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.237 on 11787 degrees of freedom
Multiple R-squared:  0.157,     Adjusted R-squared:  0.1569
F-statistic:  1098 on 2 and 11787 DF,  p-value: < 2.2e-16
```