Which countries and categories have higher suicide rates and why?

The purposes of the analysis is to find out which factors are attributed to higher suicide rates. The analysis is country based, considers such factors as country GDP, Human development index, age, gender, unemployment rate and average household size.

Key conclusions

- The suicide risk increases with ages. People who commit suicide most often are not teenagers, but rather elder people above 55+
- Men commit suicide much more often than women. Besides, the gap in suicide rates increases for men and women as they age
- Countries with high GDP and HDI have higher suicide rates, contrary to expectations. One of the explanations can be the higher number of one-person household and childless family, i.e. lonely people. The correlation between smaller households and higher suicide rates is confirmed in the analysis.
- Overall, it turns out that: the highest suicide risk category is 55+ year men living in a rich country; the lowest suicide risk category is young (<25 years old) woman in a poor country

Data Sources

Suicide Rates Overview 1985 to 2016: compares socio-economic info with suicide rates by year and country https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

Households mean size: https://www.prb.org/international/indicator/hh-size-av/map/country

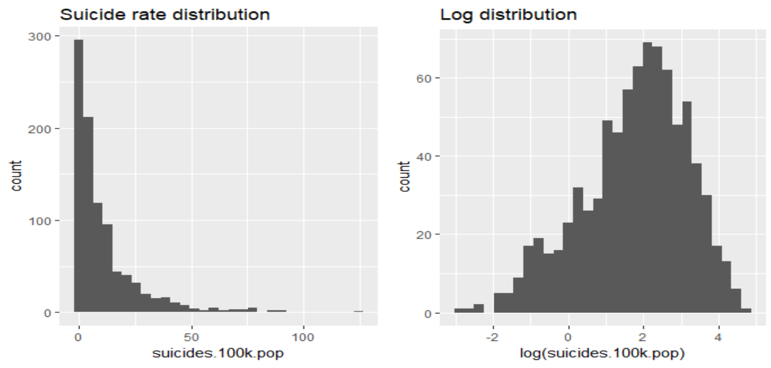
Unemployment rates: https://data.oecd.org/unemp/unemployment-rate.htm

EDA

Since my goal was the analysis of influencing factors, rather than dynamic analysis I used the data only for one year. The latest year with complete data for most countries is 2014.

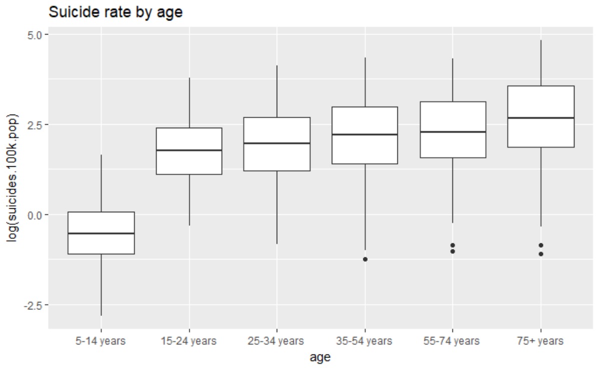
Suicide rate: dependent variable

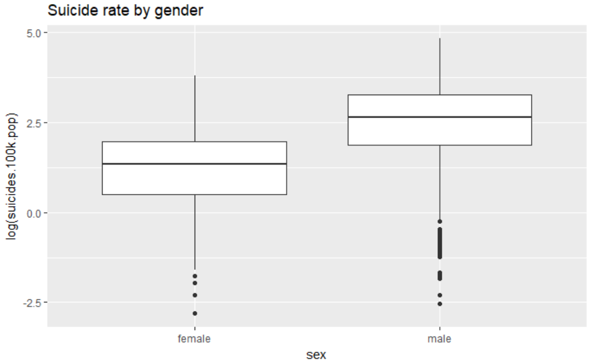
Suicide rate is the number of suicides per 100K of population in a country. In the histogram we see that it's skewed to the left (most countries have low suicide rates) which is challenging for prediction task. For further analysis I will use log(suicide rate), so that the distribution of dependent variable is close to normal, with mean of 2.0 and sd 1.0 (I will use these for the prior values in Bayesian models).



Explanatory variables: age group, gender, HDI

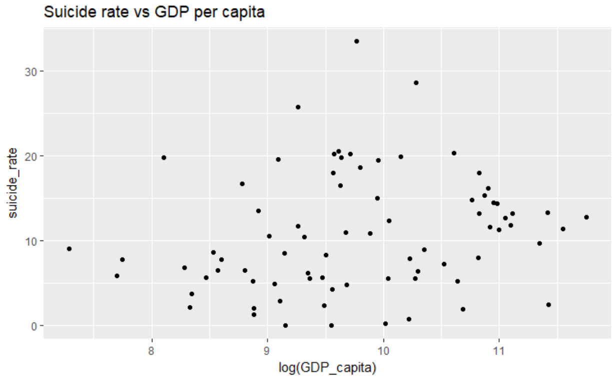
On the charts we see clearly that the suicide rate is different by age group: the higher the age, the higher the suicide rate. Also, by gender: men have much higher suicide rate. Therefore, formed a hypothesis that the suicide rate has a correlation with gender is age and included the variables into the model.

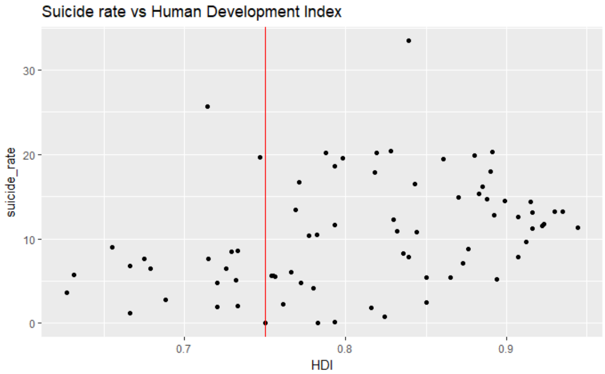




I also want to check whether the suicide rate somehow depends on country's wellbeing measured by GDP/capita and HDI. The below chart shows no correlation between GDP/capita (I use log value because GDP/capita distribution is skewed) and suicide rates.

However, if we look at the Human Development Index, although there's no clear correlation, we can see that high suicide rates appear in the countries with high HDI (>0.75). Therefore, I will use HDI as a variable in the model, but I transform it into the binary variable (below and above 0.75 threshold) instead of numeric since there's no direct correlation.





Analysis approach

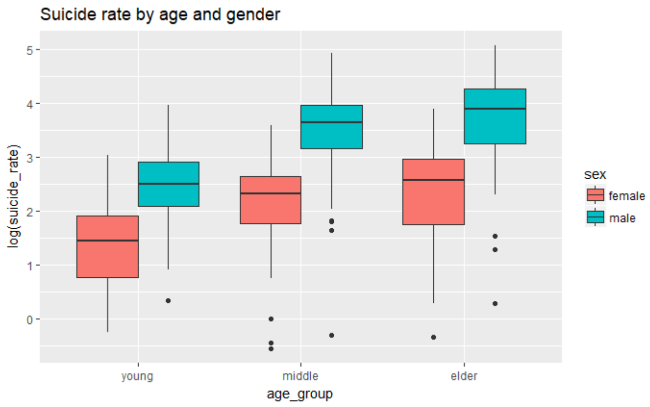
The key variables from the initial dataset which appeared to have effect on the suicide rates from EDA were: age group; gender; Human Development Index (HDI) of the country. I use them for the initial model.

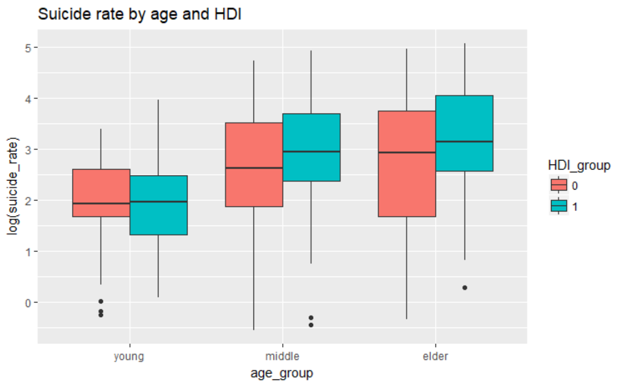
Since all of these were categorical variables, I will use ANOVA modes of three types: 1) Simple additive model, with no interaction between the variables at all, 2) Full Cell model with full interaction between the variables 3)) Mixed model, with partial interaction between the variables. I apply both Frequentist and Bayesian approach to each model and compare the results.

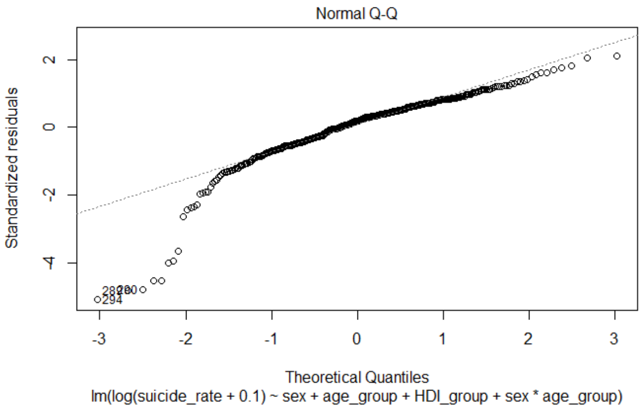
	OLS regression model	**Bayesian regression model**
Simple Additive	Log(suicide)~ Gender+Age_group+HDI	mu (log(suicide))= int + alpha*sexmale+ beta[1]*age_groupmiddle + beta[2]*age_groupelder+theta*HDI_group1

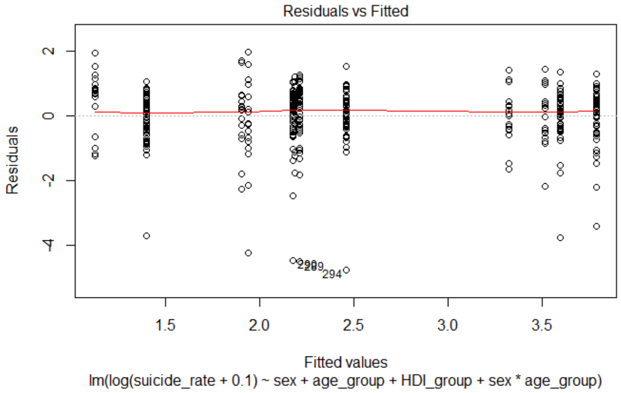
Full Cell	Log(suicide)~ Gender+Age_group+HDI+ Gender*Age_group+ Gender*HDI+ Age_group*HDI+ Gender*Age_group*HDI	mu (log(suicide))= mu[sex, age_group, HDI_group]
Mixed Cell	Log(suicide)~ Gender+Age_group+HDI+ Gender*Age_group	mu (log(suicide))= mu[sex, age_group]+HDI_group
Model priors	n/a	Mean: normal distribution with prior (2.0, 1.0) based on log(suicide rate) distribution SD: Inverse Gamma distribution with non-informative priors

For the Mixed model design I chose HDI as non-interactive variable, since the HDI effect on suicide rate, as was seen during EDA, is lower than the effect of age and gender. Therefore, for model simplicity we may not calculate separate HDI group effect for each of 6 groups.





<table>
<tr><th colspan="3">OLS regression models</th></tr>
<tr><th>Additive model</th><th>Full cell model</th><th>Mixed model</th></tr>
<tr><th colspan="3">Regression formula</th></tr>
<tr><td>Log(suicide)~ Gender+Age_group+HDI</td><td>Log(suicide)~ Gender+Age_group+HDI+ Gender*Age_group+ Gender*HDI+ Age_group*HDI+ Gender*Age_group*HDI</td><td>Log(suicide)~ Gender+Age_group+HDI+ Gender*Age_group</td></tr>
<tr><th colspan="3">Key residuals criteria (zero conditional mean, constant standard deviation, normality)</th></tr>
<tr><td>All criteria met (see Residual chart)</td><td>All criteria met</td><td>All criteria met</td></tr>
<tr><td colspan="3">




</td></tr>
</table>

These residuals charts are for the third model, but they look similar for all three models. The outliers both on the qqplot chart (left part) and Residuals vs Fitted are attributed to observations with 0 suicide rates (from countries with small population and low-risk categories). In the model they are given very small value 0.01 when applying logged suicide rate.		
Key model statistics		
RSE: 1.14 (397 DF) Adjusted R2: 0.3614 F-statistics: 57.75 (397 DF) AIC: 1253.26 BIC: 1277.24	RSE: 1.124 (390 DF) Adjusted R2: 0.3793 F-statistics: 23.28 (390 DF) AIC: 1248.72 BIC: 1300.68	RSE: 0.94 (395 DF) Adjusted R2: 0.4397 F-statistics: 53.45(395 DF) AIC: 1104.13 BIC: 1136.11
Statistically significant coefficients (mean, sd)		
Int = young, women, low HDI **int 0.88 0.15** **male 1.41 0.11** **age_middle 0.98 0.14** **age_elder 1.06 0.14** **high HDI 0.29 0.13**	*Int = young, women, low HDI* **int 1.57 0.27** **male*old age 1.73 0.55** **male:old_age*high HDI -1.5 0.63** **female*old age*high HDI 1.61 0.45**	*Int = young, women, low HDI* **int 1.13 0.14** **male 1.06 0.16** **middle age 0.78 0.16** **old age 0.81 0.16** **+ male:old_age 0.51 0.23** **HDI_group1 0.27 0.11**
Model results interpretation		
- young women living in low-HDI countries are expected to have the lowest suicide rate = 0.9 log values, or e^0.9=2.4 cases per 100K - being male increases suicide rate by 1.4 log value, i.e. e^1.4=4 times compared to women, or to 9.6 cases per 100K population - middle age and elder age group have statistically significant higher suicide rates, of approximately 1.0 logged value, or 2.7 times higher than young people - living in a country with high HDI group increases the risk of suicide by 0.3 log value, or 1.35 times	- young woman in low-HDI country have the lowest suicide rate, but it's exp(0.7)=2 times higher estimation than obtained in the first model - the difference in suicide rate between young women and men is not statistically significant, unlike the first model - the difference between young and middle-age people suicide rates is not statistically significant, for neither of genders - The old age leads to statistically significant higher suicide rate, but only for men living in low-HDI countries. Their expected suicide rate is exp(1.7)=5.5 times higher compared to people of middle&young age - For old men living in high-HDI countries there is also statistically significant difference, but much smaller, only exp(1.73-1.5)=1.25 times higher - living in high-HDI country has a significant negative impact on old women, i.e. the suicide rates increases by 1.6 logged values, or 5.0 times compared to young&middle age women. This, however, can be a biased estimate since the coefficients for simpler parameters: old age women, and women in high-HDI countries, turned non-statistically significant. And this leads to overestimation for the combined category.	- the suicide rate for young women in low HDI countries is 1.1 logged value, which is between the first model (0.9) and the second model (1.6) results - being male increases suicide rate by 1.06 log value, or 2.9 times - women of middle and old age have 0.8 logged value (2.2 times) higher suicide rate. This is close to the first model result, expect in the first model it was for both men and women. - For men of older age the risk factor increases by additional 0.5 logged value (1.65) - high HDI country adds 0.3 to logged suicide rate, which corresponds with the first model result

OLS regression models summary:

- Mixed model turns the most effective based on the criteria (AIC, BIC, RSE) combining higher explanatory power and simplicity
- The results of the Full cell model is very different from results of the Simple additive and Mixed model. The problem with OLS regression model with too many interaction variables, is that if one interaction coefficient is not statistically significant, the interaction coefficient of higher order can be overestimated. For example, if the women old_age=-0.44 and HDI_group=-0.41 coefficients were statistically significant, we would not have the old women in high_HDI as the highest risk group, their risk rate would be similar to what we got in the other models. In this Full Cell model a large portion of coefficients turn not statistically significant and this could bias the result. When I the complex model returns too many not-significant coefficients, I try simpler model as the next step instead of relying on complex model.
- Simple additive and Mixed model provide similar results, but the Mixed model catches the higher impace of aging on men than on women, which is undistinguished in the Simple model.
- In all models, the lowest risk group are young women. The highest rish group in first and third models are old men in high_HDI countries.

OLS regression Mixed model: expected mean suicide rates (per 100K people) for different categories

	Young	Middle	Old
Women – low HDI	3.1	6.9	6.9
Women – high HDI	4.2	9.3	9.3
Men – low HDI	8.9	19.9	32.8
Men – high HDI	12.1	26.8	44.3

Bayesian regression models		
Additive model	**Full cell model**	**Mixed model**
Convergence and effective size (out of 30 000)		
Scale reduction factor: 1.0-1.01 alpha beta[1] beta[2] int sig theta 10K 8.2K 8.1K 3.1K 29K 4.4K	Scale reduction factor: 1.0 Effective size is close to 30K for all parameters due to low autocorrelation	Scale reduction factor: 1.0-1.03 int mu (6 groups) sig theta 0.4 K ~0.6K 28K 1K (because of high autocorrelation I had to increase the simulation size from 1e4 to 1e5)
Residuals criteria met (zero conditional mean, constant standard deviation, normality)		
All criteria met	All criteria met	All criteria met
DIC (Mean deviance + penalty)		
1305+6=1311	1299+12=1311	1303+8=1311
Statistically significant coefficients (mean, sd)		
Int = young, women, low HDI **int 1.0 0.2** *+ male gender* **alpha 1.4 0.15** *+ middle age* **beta[1] 0.9 0.2**	**mu[female,young, low HDI] 1.6 0.34** **mu[male, young, low HDI] 2.2 0.34** **mu[female, middle, low HDI] 1.9 0.34** **mu[male, middle, low HDI] 3.2 0.34** **mu [female, old, low HDI] 1.2 0.34** **mu[male, old, low HDI] 3.3 0.34**	**int** 2.0 0.45 **mu[female,young]** -1.19 0.42 mu[male,young] -0.14 0.42 mu[female,middle age] -0.42 0.42 **mu[male,middle age] 1.0 0.41** mu[female, old] -0.4 0.41

+ elder age **beta[2] 1.0 0.2** **sig 1.52 0.05** *+ high HDI (not statistically significant)* theta 0.26 0.17	**mu[male, young, high HDI] 1.2 0.21** **mu[male, young, high HDI] 2.3 0.21** **mu[female, middle, high HDI] 2.1 0.21** **mu[male, middle, high HDI] 3.5 0.21** **mu[female, old, high HDI] 2.3 0.21** **mu[male, old, high HDI] 3.7 0.21** **sig 1.5 0.05**	**mu[male,old] 1.16 0.41** **sig 1.4 0.05** theta(HDI) 0.27 0.15
Model results interpretation		
Difference from OLS regression		
HDI coefficient is statistically significant in the OLS regression, but not in Bayesian because of larger standard deviations	While all coefficients are statistically significant in Full Cell Bayesian model, in the OLS regression the only statistically significant coefficients are: old men (mu[1,3,1] , mu[1,3,2]), old women in high HDI countries (mu[1,3,2]).	Similar to model 1, HDI coefficient is statistically significant in the OLS regression, but not in Bayesian because of larger standard deviations

Models results:

- Unlike OLS regression models, the Bayesian models show the same efficiency with the same DIC of 1311. In this case I would choose the Fuel Cell model as the preferred one. The Mixed model returns a large portion of non-statistically significant interaction variables which may bias the result; while the Simple model doesn't catch different effect of the variables on different groups (e.g. age on genders)
- HDI factor turned not statistically significant in Bayesian Simple and Mixed models, although it was significant in OLS regression with high HDI countries showing higher suicide rates. The Full Cell model, however, shows the difference: all categories, except for young women, have higher suicide rate in high-HDI countries, especially old people.
- The lowest risk categories are young women with 0.9 log mean (or 2.5 cases per 100K). Two out of three models show statistically significant increased rate for men
- All models show increased suicide rate with higher age: middle age > young age; old age>middle age. But for men the effect with age is significantly higher

Bayesian regression Full Cell model: expected mean suicide rates (per 100K people) for different categories

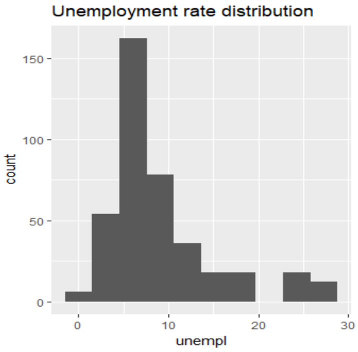
	Young	Middle	Old
Women – low HDI	5.1	6.6	3.4
Women – high HDI	3.3	8.3	10.2
Men – low HDI	8.9	24.5	27.7
Men – high HDI	10.8	33.4	40.9

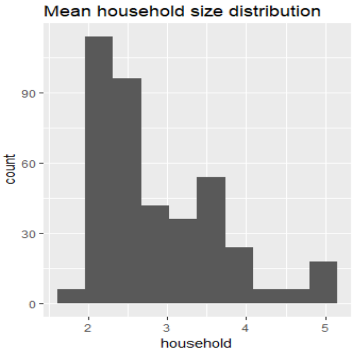
Model with additional parameters

In an attempt to explain why countries with higher GDP and HDI, contrary to expectation, have higher suicide rate, I've decided to try to use additional explanatory variables. It is believed that unemployment and loneliness increase the risk of suicide. Therefore, I have added unemployment rates in countries to the original dataset, and the mean household size by countries. Below is the quick summary of the model.

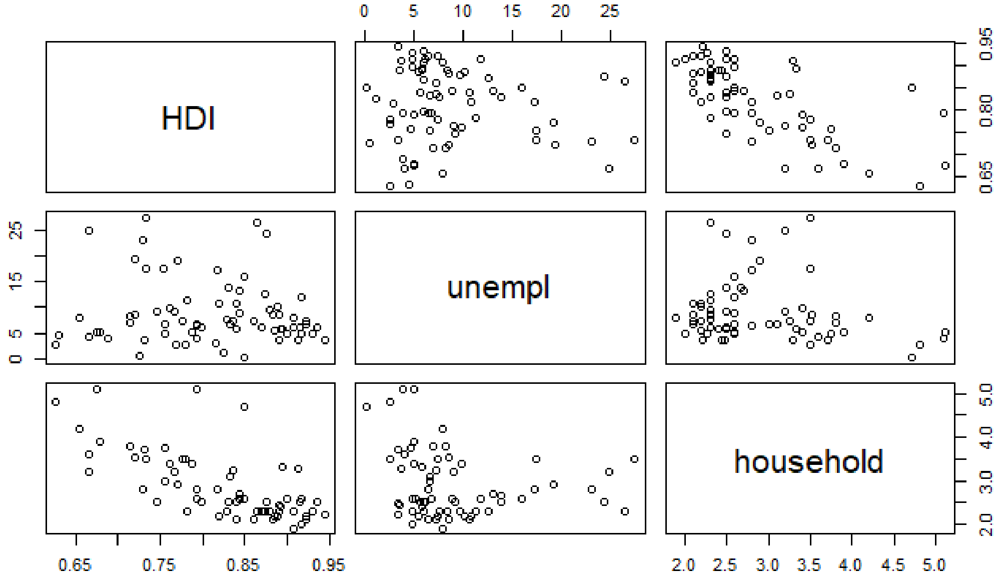
EDA

Distribution: The distribution of mean household size is skewed to the left, therefore, I will use log of the variable for the model, so that the distribution is closer to normal.





Variables correlation: HDI and mean household size have a strong negative correlation, therefore, we can get skewed results if we use both variables in the model. Since in this case I'm interested in the effect of household structure on suicide rate specifically, I will use the mean household size and drop HDI.



Model design

The model was designed based on the mixed ANOVA model, with two numeric variables added and HDI excluded.

Algorithm: mean (log(suicide))= mean [sex, age_group]+unemployment+hoseuhold_size

Prior distribution of variables:

mean[sex, age_group] ~ dnorm(2.0, 1.0); based on logged suicide rate distribution

standard deviation ~ inverse Gamma with non-informative priors;

unemployment ~ dnorm (9.0, 6.0), based on unemployment rate distrubution

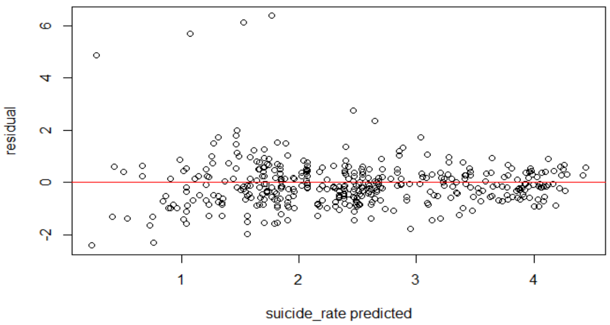
hoseuhold_size ~ dnorm (1.0, 0,2), based on household size distribution

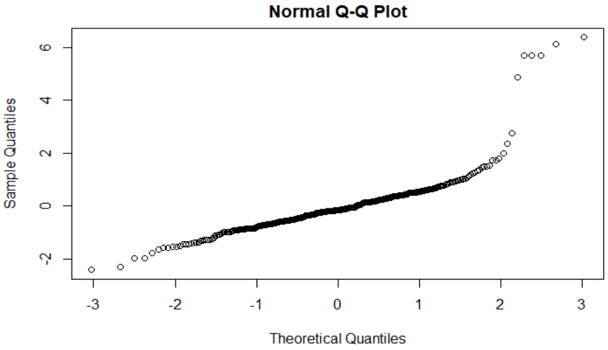
Model quality

- DIC=1238 is lower than DIC=1311 of the three ANOVA models, therefore the model is more efficient

- all variables are convergent with scale factor=1

- residuals meet the key criteria (conditional mean = 0 ; constant standard deviation; normal distribution). Outliers attributed to zero suicide rates from small population groups (small country + low risk group), which were assigned 0.01 value for log purpose, log(0.01)=-6. The standard deviation is slightly smaller at higher values, but the difference is not critical





Model results

Statistically significant coefficients	Mean	SD
int	**4.30026**	**0.47261**
mu[female, young]	**-0.80676**	**0.41055**
mu[2,1]	0.24207	0.41064
mu[1,2]	-0.03807	0.41061
mu[male, middle]	**1.38179**	**0.41089**
mu[1,3]	-0.0691	0.41102
mu[male, old]	**1.56743**	**0.41084**
unemployment	**-0.02373**	**0.01205**
mean household size	**-1.92344**	**0.28759**
sig	**1.42587**	**0.05111**

- coefficients for both added variables (unemployment and average household size) are statistically significant. Again, this model also shows the impact of old age is higher for men than women, for women the age impact turned not statistically significant.

- average household size is correlated with suicide rate: 1% larger household size decreases the suicide risk by 2%. This is a very important finding in understanding why richer countries have higher suicide rates: they have a much higher share of one-person households, old people living separate from children, a higher share of childless families.

- there's a small, but statistically significant, coefficient of unemployment rate. This model shows that higher unemployment rate is correlated with slightly lower suicide rates, which is contrary to expectation. My assumption is that there is another explanatory (omitted) variable correlated both with the unemployment rate and the suicide rate (e.g. non-registered employment in low-income countries), which creates this effect. However, the effect is very small, therefore, I do not go deeper within this work