A GLM-based analysis of Gender Equality in the EU

Hendrik Walter

August 22, 2019

1 Executive Summary

In this report the impact of different societal factors on Gender Equality in the EU-28 Countries is analysed. Data from education, professional and personal life as well as healthcare is taken into consideration. These overarching parameters were derived from performance indicators (e.g. mean earnings). A Beta distribution is used to predict the Index based in the data. It is of interest if the parameters with which the above mentioned factors could be zero, which implies no relationship, or non zero, which implies a relationship. To derive this conclusion several models are proposed. The best of these initial models is selected using an Information Criteria, a method commonly used to compare the fit of different models to a particular data set. The best model turned out to not contain any temporal effects. The parameters in the final model were: Work, Money, Time, Power, Health and Knowledge. All their parameters in the model differed statistically significant from zero, meaning that is unlikely to obtain these data if they were zero. The error in the estimation of these parameters are small, which gives extra confidence. However, collinearity is high, which indicates that many of the parameters affect each other, e.g. if one rises, the other one does too. Residuals were used to understand the fit of the model. The conclusion of this work is, that women in powerful positions have the single most significant effect on Gender Equality in a country, while health related scores influence equality less. All factors have a positive relationship with the Index. To take this work further, it is possible to consider the performance indicators used to derive the overarching parameters, to pinpoint the most significant effect.

2 Introduction

This report aims to investigate the impact of certain societal factors on Gender Equality between members of the EU-28. Receiving more and more media [2] and research attention [4] in recent years and with political campaigns on the way [1], Gender Equality is present topic in most countries. As the EU monitors Gender Equality throughout the different countries over time, while collecting additional data on the country, it is of special interest how these external factors influence Gender Equality. It is not only of interest what impacts the Gender Equality in itself, but it can be helpful to governments to focus resources on certain areas to improve overall Gender Equality in the country. Using a model can help to quantify how improvements in certain areas might affect the Gender Equality and help to predict Gender Equality for potential future scenarios.

3 Data

Data on Gender Equality in the EU-28 is collected by an EU institution named "European Institute for Gender Equality" on a regular basis and published every year in a report. The report highlights the development of the Gender Equality Index for every European member state and compiles a final index between 0 and 100, which 100 meaning perfect equality between males and females. The index is derived from 6 core categories which each core category itself being a score between 0 and 100 compiled from indicators covering different aspects of society. The categories are:

- **Work**: Includes how equal the genders are in terms of taking part in the job market, as well as if they take part how equal they are in specific sectors as well as the career prospects. A high score implies equal prospects and part-taking in work.
- **Money**: Includes the financial situation of each citizen as income and earnings and the economic situation for the population as a whole (e.g. income distribution, risk of poverty). A higher money index implies more financial resources per person and good economic situation.
- **Time**: The way time is spent by women on taking care of housework and caring for friends and family as well as if participation in social activities and charities happen.
- **Power**: Divided into political, economic and social powers, measured by board positions filled with women in the influential institutions and organizations in the respective sectors.
- **Health**: Combines life expectancy and risk factors (e.g. alcohol consumption) as well as accessibility to public healthcare.
- **Knowledge**: The share of females experiencing education as well as how equal they are, if taking part. This section focuses on tertiary education.

In general a higher score implies a more equal situation.

The data used in this report were collected over multiple years, namely 2005, 2010, 2012, 2015, which adds another possible component to the data, allowing to observe Gender Equality over time. The report from 2015 includes the data of the years 2005, 2010 and 2012 and was used for this study [3].

With data available for each member of the EU-28 and four years available, the total number of observations available is 112. In the plot below, each dot represents a country. The variation within each group on the x-axis doesn't imply an earlier or later collection day, but was introduced to improve visibility of overlapping data points.

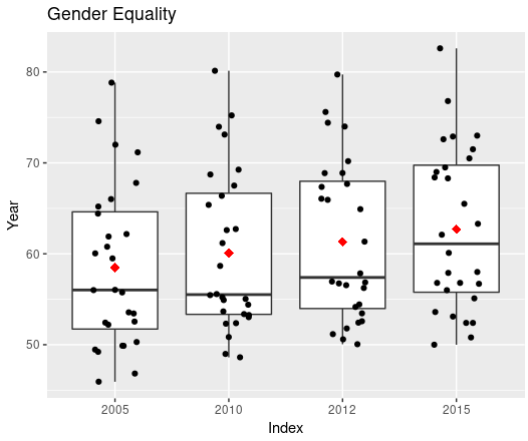

Figure 1: Boxplot Gender Equality

It seems that over the years the mean and the median were both increasing, while the range and the variation about the mean seems to be fairly consistent. There seems to be some positive relationship between the years and the gender equality in a country as the mean increases over time.

Now taking the data of one year, say 2015, it is possible to observe a few linear relationships between some of the covariates but also with Index. The plot below visualizes this:

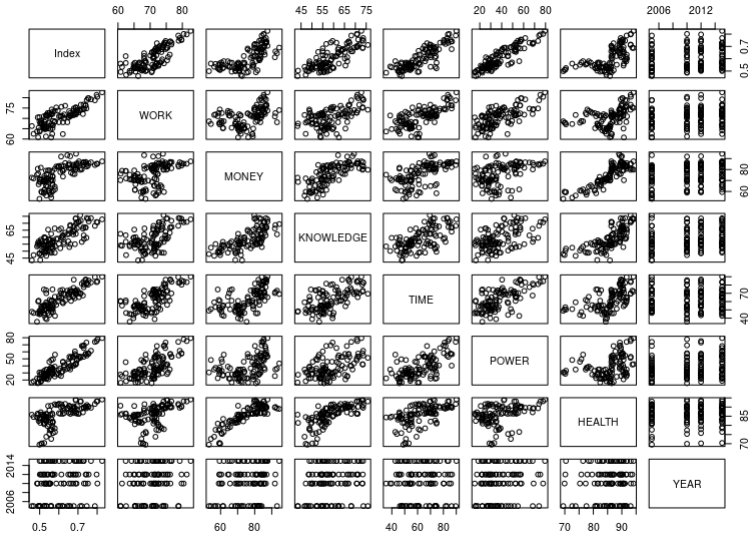

Figure 2: 2015 Plot

Between Index and Time, Power, Work and Knowledge there seems to be a fairly strong linear relationship, Money could potentially have a polynomial impact on the Index while the impact on Health could be interpreted as linear, but it is less strong than the other ones. There are almost certainly some non linear relationship between health and the Gender Equality Index. In terms of data, it seems that high Gender Equality Indices are also possible with low Health Scores. This plot also indicated that there might be collinearity between some covariates as there seem to be linear relationships. Taking Knowledge and Money as an example, it seems that they impact each other fairly strongly, which also counts for many other relationships.

4 Model formulation

The data at hand is bounded between 0 and 100 and is continuous. Considering that the lower and upper bound is difficult to enforce by link functions, it seems sensible to also select the overarching function based on these constraints. This leads to a Beta GLM. Beta-distributions are bounded between 0 and 1, which makes the mapping between the two different scales straightforward. The index is divided by 100 for the purpose of fitting the model. The parameters α and β allow the Beta(α,β) distribution to take various shapes, making it also a flexible model. These two parameters will also be estimated from the data.

The PDF of the Beta function is defined as followed:

$$f(x) = \frac{x^{\alpha-1} \cdot (1-x)^{\beta-1}}{B(\alpha, \beta} \tag{1}$$

where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{2}$$

The first Model includes all listed parameters in 3 as well as the year as a factor variable. This model includes no higher powers. A variation on this model is to consider the year covariate as a continuous variable instead. This could have particular relevance for predictions. Finally, it seemed that Money

could have a higher power relationship with Index. This is the third model considered. The initial models prior variable selection are:

Model 1:

$$logit(\mathbb{E}(Index))) = \beta_0 + \beta_1 WORK + \beta_2 MONEY + \beta_3 KNOWLEDGE + \beta_4 TIME + \beta_5 POWER + \beta_6 HEALTH + \beta_7 2010 + \beta_8 2012 + \beta_9 2015 \quad (3)$$

Model 2:

$$logit(\mathbb{E}(Index))) = \beta_0 + \beta_1 WORK + \beta_2 MONEY + \beta_3 KNOWLEDGE + \beta_4 TIME + \beta_5 POWER + \beta_6 HEALTH + \beta_7 YEAR \quad (4)$$

Model 3:

$$logit(\mathbb{E}(Index)) = \beta_0 + \beta_1 WORK + \beta_2 MONEY + \beta_3 KNOWLEDGE + \beta_4 TIME + \beta_5 POWER + \beta_6 HEALTH + \beta_7 YEAR + \beta 10 MONEY^2 \quad (5)$$

5 Results

In this section, the three models from above will be fitted. The main criteria in this section for variable selection will be p-values Wald tests. While there are known error rates with p-values, they tend to be more conservative than alternatives such as BICs, which tend to prefer the best predictive model. In this report, it is of most interest what effected the scores most, rather than being able to give the best predictions. Running this process with ICs instead might yield a very different result [5].

Note that Model2 and Model3 are in fact nested. Applying a log likelihood test is a simple way to distinguish between the two. The test statistics is calculated as followed:

$$2 * log(LR_{M_3,M_2}) = 2 * (log(\mathcal{L}(\theta_{M_3}) - (\mathcal{L}\theta_{M_2}))) \quad (6)$$

$$2 * (435.6915 - 434.9724) = 1.434 \quad (7)$$

The difference in parameters is one, hence a comparison against a $\mathcal{X}_1^2$-distribution gives the p-value. In this case this is 0.23. It is possible to conclude, that the simpler model Model 2 sufficiently captures the trend in the data.

This test is not applicable for the comparison between Model 1 and Model 2. In this case, Information Criteria are the method of choice as the two models are not nested. Here, BIC is used, which tends to restrict more than AICs.

$$BIC = -2 * (log(\mathcal{L}(\theta_{M_i}) + log(n) * p_i \quad (8)$$

The BIC for Model2 is -827.48, while the BIC for Model 1 is -818.15. As the BIC for Model 2 is considerably smaller, Model 2 will be the model of choice for further variable selection.

5.1 Variable Selection

In Figure 2 it seemed that there could a high correlation between some of the variables. And indeed, calculating the VIF scores like:

$$VIF_k = \frac{1}{1 - R^2} \tag{9}$$

This give the following results for the different covariates:

MONEY	WORK	KNOWLEDGE	TIME	POWER	HEALTH	YEAR
4.79	2.41	2.79	2.95	2.19	4.66	1.21

Table 1: Different VIF scores

The VIF scores seem slightly inflated, but there is no major reason for concern. Both Health and Money are close to 5, often considered the cut off point. If the opportunity arises, one could consider removing either Health or Money, as they seem to be highly correlated. However, as both are still below 5, this is something to keep in mind rather than actively pursue. With this extra knowledge, it is possible to proceed into the variable selection.

As mentioned before, the aim of this analysis is not perfect prediction, but rather the description of the effects of the covariates on Gender Equality. Considering this, a Wald tests is used to identify the most significant factors.

Running a Wald test on the full Model 2 gives:

	z value	Pr(>|z|)
Intercept	-1.31	0.19
WORK	10.63	<2e-16
MONEY	11.99	<2e-16
KNOWLEDGE	23.64	<2e-16
TIME	22.141	<2e-16
POWER	63.85	<2e-16
HEALTH	2.36	0.0185
YEAR	-0.488	0.6252

Work, Money, Knowledge, Time and Power are highly significant, even on the 0.01% level. Health is also significant on the 5% level, while Year does not seem to be significant. Based on these p-levels, Year can be removed from the model.

Running another Wald test on the remaining parameters, does not yield any change in significance, which leaves all remaining covariates in the model. Revisiting the VIF scores at this stage shows that some of the VIF scores went down slightly. These differences are calculated $VIF_{new} - VIF_{old}$:

MONEY	WORK	KNOWLEDGE	TIME	POWER	HEALTH
-0.27	-0.08	-0.039	-0.29	-0.02	0.13

Table 2: Difference in VIF scores

5.2 Model fit

After selecting a model and covariates for said model, the final question remaining is the question if the final model is appropriate for the data. It might be the best models of all models considered, but still not appropriately describe the data.

A first step is looking at the residuals. Here, raw, Pearson and Deviance residuals are considered in the plot below. Note that the Gender Equality Index is mapped from the 0-100 scale to 0-1.

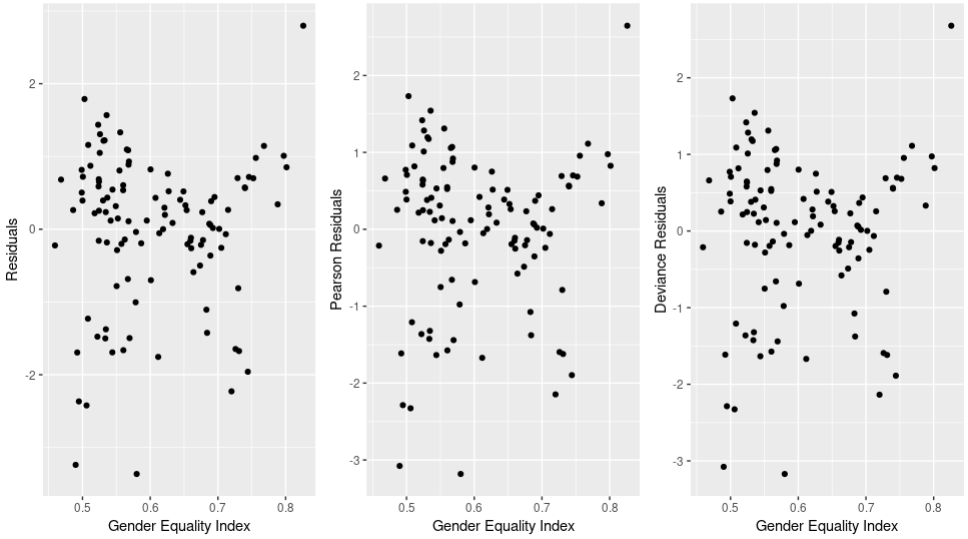

Figure 3: (a) raw residuals (b) Pearson Residuals (c) Deviance Residuals

While the variance of the Pearson residuals are 1.009, which is perfectly close to the expected 1. This doesn't come as a surprise, as there is a dispersion parameter included in the model. The dispersion parameter which was estimated using an identity link, is highly significant with a p-value of $7.23e-14$ and ensures a standardized variance across the whole range. If this parameter was not included, the data would have been overdispersed. Within the model on the 0-1 scale the parameter Φ, to normalize the overdispersion was taken to be 9239. However, there is a slight quadratic shape in all three of the residual plots, which is reason for concern as some polynomial effect might have been missed out. Throughout this report a quadratic model has been considered, but failed to convince statistically. None of the other factors seem to have a quadratic relationship with the Index and also the effect plots below show satisfying results, which leads to the conclusion that the final model is the best model at hand.

As mentioned above, good results are yielded from the different effects plots. Thee different covariates are displayed below in against the Index. These plots are on a link scale and should be linear. There doesn't seem to be any major outliers in any of the covariates.

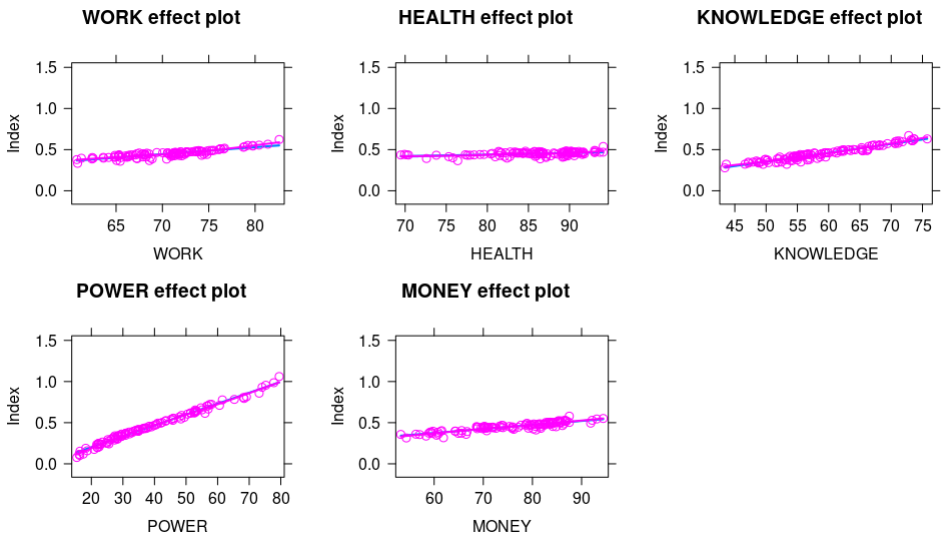

Figure 4: Effects of the different covariates

None of the plots above indicate that there is a problem with independence, which is also confirmed by a runs test. The p-value of the runs test is 0.13.

Overall, the model can be considered to fit the data presented.

5.3 Parameter estimations

The parameters and the associated standard errors were estimated for 0-100 scale to be the following:

	Estimate (0-100)	Std Error (0-100)	Upper CI	Lower CI
Intercept	-222.4	5.4	-211.82	-232.99
WORK	0.49	0.05	0.588	0.392
MONEY	0.84	0.07	0.978	0.703
KNOWLEDGE	1.01	0.04	1.088	0.932
TIME	0.6	0.03	0.659	0.541
POWER	1.32	0.02	1.359	1.281
HEALTH	0.2	0.07	0.337	0.063

Table 3: Parameter estimations

The standard errors seem small compared to the estimations, which was also recognized in the p-values during the variable selection process. The negative intercept is likely to be due to the fact that data only ranges between 45 and 82, far away from the intercept.

6 Discussion

The results section has shown that the Beta model without any temporal component yields best results in explaining the data. There were strong collinearity between some of the covariates, but they weren't strong enough to form an argument to remove one of them from the model. Another small caveat to this model are the residuals which seem to lean towards a quadratic shape. However, no quadratic relationship could be identified beyond the one considered, which failed to convince. Considering the good fit overall, this has little impact on the final conclusion.

The remaining covariates all have different factors in their relationship to the Gender Equality Index. Between Power and Gender Equality the parameter estimate is particularly high. With a 95% confidence interval of (1.281,1.359) it seems that the share of women in powerful positions in politics, the economy and NGOs is the single most determining factor how equal a society is. The standard errors are comparably small. This gives high confidence that there is a relationship close to the parameters given above. As already seen throughout the analysis, the healthcare system has very correlation with the Gender Equality Index, even though it is still significant. Knowledge seems to translate 1:1 into the Gender Equality Index, which is a good sign for all policy makers, that investments in education systems is likely to pay off with better Gender Equality. The collinerarity between all these factors also quantifies that none of these areas can be truly changed on its own without impacting the others. Taking the education system as an advantage: More female graduates will result in more Money and Work for these women, which will eventually lead to more females in leadership positions. Giving Power and Money to people will also enable them to spend the time based on their choices.

In the previous sections it was possible to see that a Beta distribution is suitable for a GLM-based analysis of Gender Equality in the EU and was able to produce statistically significant results with a model fitting well to the data.

References

[1] Eu comission policies. `https://ec.europa.eu/info/policies/justice-and-fundamental-rights/gender-equality_en`. Accessed: 2018-11-21.

[2] The power-sharing dream: Where women rule in the world. `https://www.bbc.co.uk/news/world-44454914`. Accessed: 2018-11-21.

[3] European Institute for Gender Equality. Gender equality index. 2015.

[4] Jill Rubery. Gender mainstreaming and gender equality in the eu: the impact of the eu employment strategy. *Industrial Relations Journal*, 33(5):500–522.

[5] Galit Shmueli. To explain or to predict? *Statist. Sci.*, 25(3):289–310, 08 2010.

7 Appendix

```
    # https://data.europa.eu/euodp/data/dataset/gender-equality-index
library(betareg)
library(car)
library(effects)
library(ggplot2)
library(lawstat)
library(gridExtra)

## Read data
data <- read.csv("data/data.csv", header=TRUE)
data.2005 <- read.csv("data/gender-equality-index-2005.csv",
    header=TRUE)[2:29,]
data.2005$YEAR <- rep(2005,nrow(data.2005))
data.2010 <- read.csv("data/gender-equality-index-2010.csv",
    header=TRUE)[2:29,]
data.2010$YEAR <- rep(2010,nrow(data.2010))
data.2012 <- read.csv("data/gender-equality-index-2012.csv",
    header=TRUE)[2:29,]
data.2012$YEAR <- rep(2012,nrow(data.2012))
data.2015 <- read.csv("data/gender-equality-index-2015.csv",
    header=TRUE)[2:29,]
data.2015$YEAR <- rep(2015,nrow(data.2015))
# combine dataframes
data <- rbind(data.2005,data.2010)
data <- rbind(data,data.2012)
data <- rbind(data,data.2015)

# these are the columns to investigate
cols <- c("Country","Gender.Equality.Index","WORK","MONEY",
    "KNOWLEDGE","TIME","POWER","HEALTH","YEAR")
data <- data[,cols]
# rename Index
names(data) <- c("Country","Index","WORK","MONEY","KNOWLEDGE",
    "TIME","POWER","HEALTH","YEAR")
head(data)

## plots for introduction
## used
ggplot(data, aes(x=as.factor(YEAR), y=Index)) + geom_boxplot()+
    geom_jitter(position=position_jitter(0.2))+
    stat_summary(fun.y=mean, geom="point", shape=18,size=3, color="red")+
    labs(title="Gender Equality",x ="Index", y = "Year")
## USED
pairs(data[,2:9])

##### Beta distribution approach ###
# set values min =0 and max =1
# convert scale from 0-100 to 0-1
data[,2] <- data[2]/100

# Model 1
# all covariates and year as factor variable
Model1 <- betareg(Index~MONEY+WORK+KNOWLEDGE+TIME+POWER+HEALTH+
```

```
    as.factor(YEAR),data=data)

# Model2
# year as continuous variable
Model2 <- update(Model1,.~.-as.factor(YEAR)+YEAR)

## Model 3
# health as quadratic term
Model3 <- update(Model2,.~.+I(MONEY^2))

##### Model Selection ####
ll.model1 <- Model1$loglik
ll.model2 <- Model2$loglik
ll.model3 <- Model3$loglik

# Model 3 and Model 2 are nested
D2.3 <- 2*(ll.model3-ll.model2)
# difference in parameter is one
# model 2 is better
1-pchisq(D2.3,1)

# Take model 2 as BIC is lower
BIC(Model1)
-2*ll.model1+log(nrow(data))*11
BIC(Model2)
-2*ll.model2+log(nrow(data))*9

## Select Variables
# get vif scores
vif1 <- vif(Model2)
# none are above 5, but close to it
print(vif)

# see wald test
summary(Model2)

# get rid of year - not significant
Model2 <- update(Model2,.~.-YEAR)

# all covariates now significant
summary(Model2)

# get second vif scores, see if something changed
vif2 <- vif(Model2)
# are lower now
print(vif2)
# difference in vif scores
vif2-vif1[1:6]

## Residuals
# standard residuals
res <- residuals(Model2)
var(res)

# pearson residuals
res.pearson <- residuals(Model2, type="pearson")
```

```
# around 1 - but there is also a dispersion parameter
var(res.pearson)

# deviance residuals
res.dev <- residuals(Model2, type="deviance")
var(res.dev)

# plot the residuals
plot.res.pearson <- ggplot(data,aes(x=Index,y=res.pearson))+ geom_point()+
    ylab("Pearson Residuals")+xlab("Gender Equality Index")
plot.res <- ggplot(data,aes(x=Index,y=res))+ geom_point()+
    ylab("Residuals")+xlab("Gender Equality Index")
plot.res.dev <- ggplot(data,aes(x=Index,y=res.dev))+ geom_point() +
    ylab("Deviance Residuals")+xlab("Gender Equality Index")
grid.arrange(plot.res,plot.res.pearson,plot.res.dev,ncol=3)

## Effects Plot
# Plot the effects on link scale, should be linear
e_work <- plot(effect("WORK",Model2,residuals=TRUE),
    partial.residuals=TRUE, type="link")
e_health <- plot(effect("HEALTH",Model2,residuals=TRUE),
    partial.residuals=TRUE, type="link")
e_knowledge <- plot(effect("KNOWLEDGE",Model2,residuals=TRUE),
    partial.residuals=TRUE, type="link")
e_power <- plot(effect("POWER",Model2,residuals=TRUE),
    partial.residuals=TRUE, type="link")
e_money <- plot(effect("MONEY",Model2,residuals=TRUE),
    partial.residuals=TRUE, type="link")
# Arrange the plots - they seem to be linear, all good
grid.arrange(e_work,e_health,e_knowledge,e_power,e_money,nrow=2)

# runs test - not significant
runs.test(residuals(Model2,type="pearson"))

# obtain parameters for presentation
summary(Model2)
```