Capstone CYO HUMAN FREEDOM INDEX

Stefan Reinhard

5/30/2020

Introduction and overview

The Human Freedom Index (HFI) presents a broad measure of human freedom, understood as the absence of coercive constraint. The index is calculated by country on an a scale of 0 to 10, where 10 represents more freedom. Indicators are distinct between personal and economic freedom.

The Human Freedom Index dataset is provided by kaggle https://www.kaggle.com/gsutters/the-human-freedom-index and CATO https://www.cato.org. For this project we use the the direct source from CATO, but the dataset are identical.

The goal of this project is to determine the importance of variables in the dataset with a machine learning algorithms. The index uses 76 distinct indicators of personal and economic freedom, the indicator are grouped and weighted to in total 11 variables.

Personal Freedom (50%)

1. Rule of Law (12.5%)
2. Security and Safety (12.5%)
3. Movement (5%)
4. Religion (5%)
5. Association, Assembly, and Civil Society D. Expression and Information (5%)
6. Identity and Relationships (5%)

Economic Freedom (50%)

7. Size of Government (10%)
8. Legal System and Property Rights (10%)
9. Sound Money (10%)
10. Freedom to Trade Internationally (10%)
11. Regulation (10%)

To simplify, in this project we take the 11 indicators in account including their average by group (personal and economic freedom) and the overall score of the HFI (hf_score). The aim of this project is to compare the hf score with the predicted importance of the indicators. Therefore 4 linear models are calculated and finally compared to the actual outcome of original score. With the new model winner and losers are shown and their shift visualized. Root mean squared error (RMSE) is defined as a loess function. RMSE is read as a standard deviation, a result of 1 leads to a error of 1 on the scale of 0 to 10. Preferably it shows the change of the model (RMSE) during the building process.

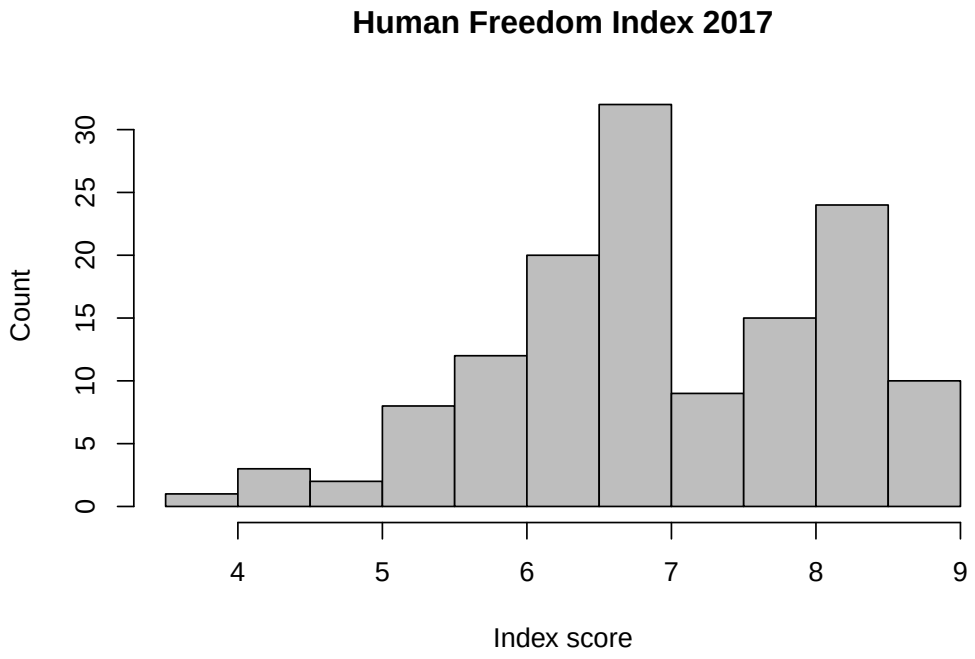
$$RMSE = \sqrt{\frac{1}{N}\sum_{u,i}(\hat{y}_{u,i} - y_{u,i})^2}$$

First we do an exploratory data analysis to get to know the data. We focus on the outcome for 2017 and the grouped indicators. There are 136 listed countries with associated information (ISO, region), 11 indicators as mentioned, and 3 scores (hf_score, pf_score, ef_score).

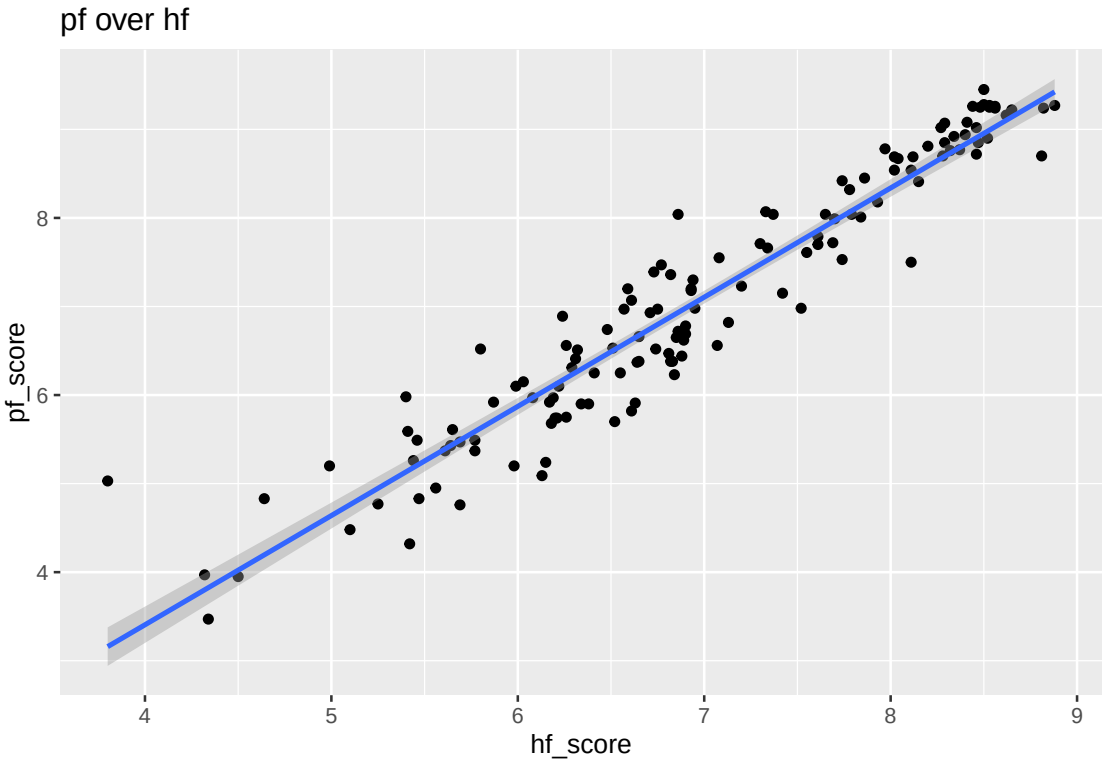
```
glimpse(data2017)
```

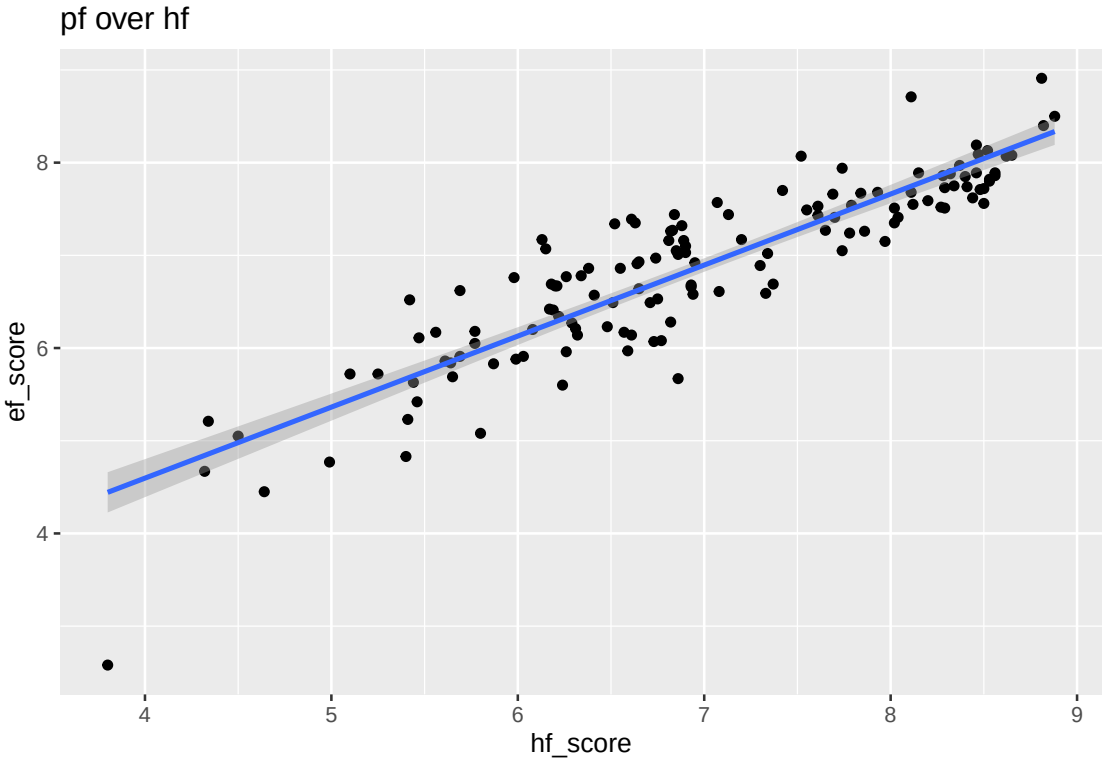
```
## Rows: 136
## Columns: 19
## $ year            <int> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 201...
## $ ISO_code        <chr> "ALB", "DZA", "AGO", "ARG", "ARM", "AUS", "AUT", "A...
## $ countries       <chr> "Albania", "Algeria", "Angola", "Argentina", "Armen...
## $ region          <chr> "Eastern Europe", "Middle East & North Africa", "Su...
## $ hf_score        <dbl> 7.84, 4.99, 5.40, 6.86, 7.42, 8.62, 8.48, 6.22, 6.6...
## $ pf_rol          <dbl> 5.3, 3.8, 3.4, 5.7, 4.9, 7.8, 8.2, 4.3, 5.9, 3.1, 5...
## $ pf_ss           <dbl> 9.3, 7.8, 8.1, 8.8, 9.1, 9.9, 9.9, 9.5, 7.4, 7.0, 9...
## $ pf_movement     <dbl> 10.0, 5.8, 6.7, 10.0, 8.3, 10.0, 10.0, 8.3, 4.2, 6....
## $ pf_religion     <dbl> 9.2, 4.9, 8.8, 7.8, 6.2, 9.1, 7.9, 5.5, 7.5, 5.9, 5...
## $ pf_association  <dbl> 10.0, 5.0, 4.0, 7.5, 6.5, 10.0, 10.0, 3.0, 5.0, 7.0...
## $ pf_expression   <dbl> 8.6, 7.3, 6.5, 8.7, 7.2, 9.4, 9.3, 4.0, 4.1, 7.0, 5...
## $ pf_identity     <dbl> 5.8, 0.0, 5.0, 10.0, 8.2, 9.0, 10.0, 5.8, 5.0, 1.7,...
## $ pf_score        <dbl> 8.01, 5.20, 5.98, 8.04, 7.15, 9.16, 9.25, 6.10, 5.9...
## $ ef_government   <dbl> 7.5, 3.6, 6.8, 5.7, 7.4, 7.0, 5.7, 4.8, 7.0, 8.2, 6...
## $ ef_legal_gender <dbl> 1.0, 0.8, 0.8, 0.8, 1.0, 1.0, 1.0, 0.7, 0.5, 0.8, 0...
## $ ef_money        <dbl> 9.6, 7.3, 5.6, 6.5, 9.5, 9.5, 9.4, 6.8, 9.4, 7.0, 7...
## $ ef_trade        <dbl> 8.3, 2.8, 3.2, 6.5, 8.2, 7.6, 8.1, 7.3, 7.4, 6.0, 6...
## $ ef_regulation   <dbl> 7.8, 5.4, 5.7, 5.6, 7.5, 8.5, 7.5, 7.2, 7.8, 6.7, 7...
## $ ef_score        <dbl> 7.67, 4.77, 4.83, 5.67, 7.70, 8.07, 7.71, 6.34, 7.3...
```

With focus on the distribution of the HF index, we see 2 peaks, one between 6.5 and 7 and another between 8 and 8.3.



Now we want to see the correlation between the human freedom score and the personal respectively the economic freedom score. We expect a close to linear correlation.





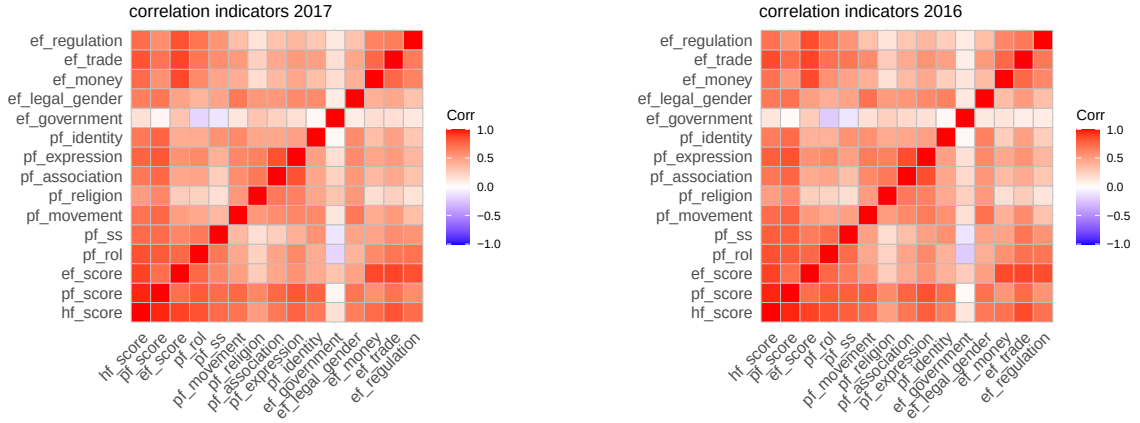
As we have seen the distribution of the score we check if the visual expectation is right and calculate the correlation.

```
#correlation ef, pf, hf
cor(data2017[, c(5,13,19)])
```

```
##           hf_score  pf_score  ef_score
## hf_score 1.0000000 0.9550136 0.8946544
## pf_score 0.9550136 1.0000000 0.7219297
## ef_score 0.8946544 0.7219297 1.0000000
```

The correlation between hf and pf is 0.955, slightly higher than between hf and ef 0.895, but we expect it linear.

Now we focus on the indicators, we see that indicators may lead to different correlation to each other, the indicator ef_government, which correlates more weakly than all other variables, is noticeable here. In general we don't see a difference between 2016 and 2017.



That the variables are mostly not independent and start with the analysis of the prediction of the individual indicators.

Methods/analysis

We start with the first simple model to predicted the score according the average of all variables regardless of their weighting. The equation is as follows.

$$pred\_score_i = \frac{pf\_rol_i + pf\_ss_i + pf\_movement_i + ... + ef\_regulation_i}{11}$$

The RMSE with all indicators based on the average is:

```
rsme_model1 <- RMSE(model1$hf_score, model1$pred_score)
rsme_model1
```

```
## [1] 0.4737959
```

To improve the first model we predict the score according the CATO as seen on the overview of the first page. The equation is as follows.

$$pred\_score_i = pf\_rol_i * 0.125 + pf\_ss_i * 0.125 + pf\_movement_i * 0.05 + ... + ef\_regulation_i * 0.01$$

The RMSE with all indicators based on the weighted model 2 is:

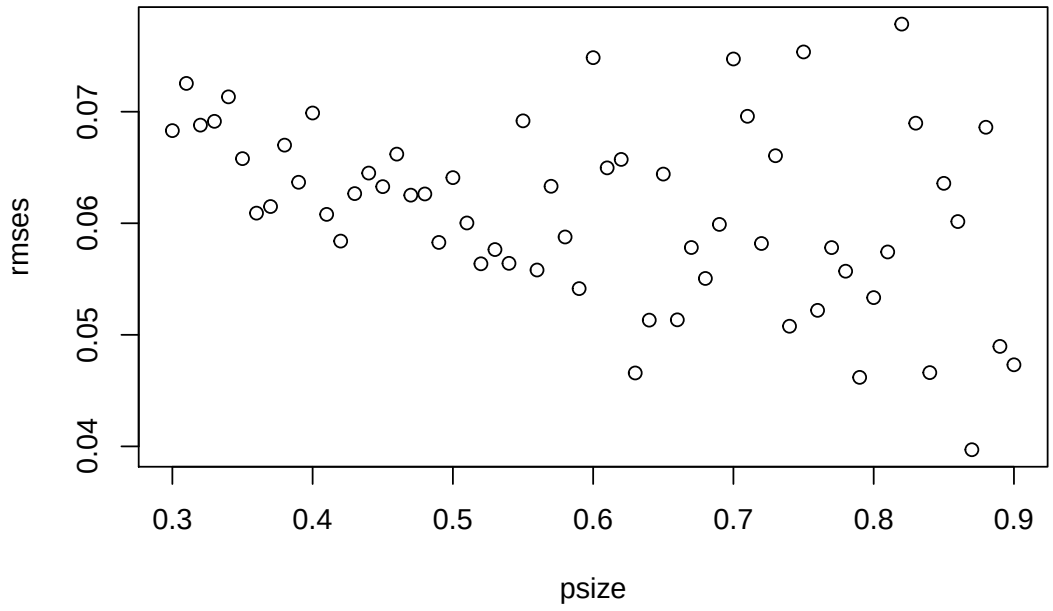
```
rsme_model2 <- RMSE(model2$hf_score, model2$pred_score)
rsme_model2
```

```
## [1] 0.4669578
```

Model 2 has a lower RSME, what not surprising since we know the score is built this way. We compare now the hf_score and pred_score of top rated countries and it looks like the scores don't exactly match.

```
##    hf_score   countries pred_score
## 1      8.88 New Zealand     8.1150
## 2      8.82 Switzerland     8.0800
## 3      8.81   Hong Kong     8.1025
## 4      8.65      Canada     7.9600
## 5      8.62   Australia     7.9475
## 6      8.56     Denmark     7.8575
## 7      8.56  Luxembourg     7.8400
## 8      8.53     Finland     7.7500
## 9      8.53     Germany     7.9000
## 10     8.52     Ireland     7.8750
## 11     8.50 Netherlands     7.7900
## 12     8.50      Sweden     7.8175
```

For the next models, we calculate a fit model to optain coefficients closer to the prediction based on the data of 2017. Therefore, in a first step the best matching data partition size is determined.



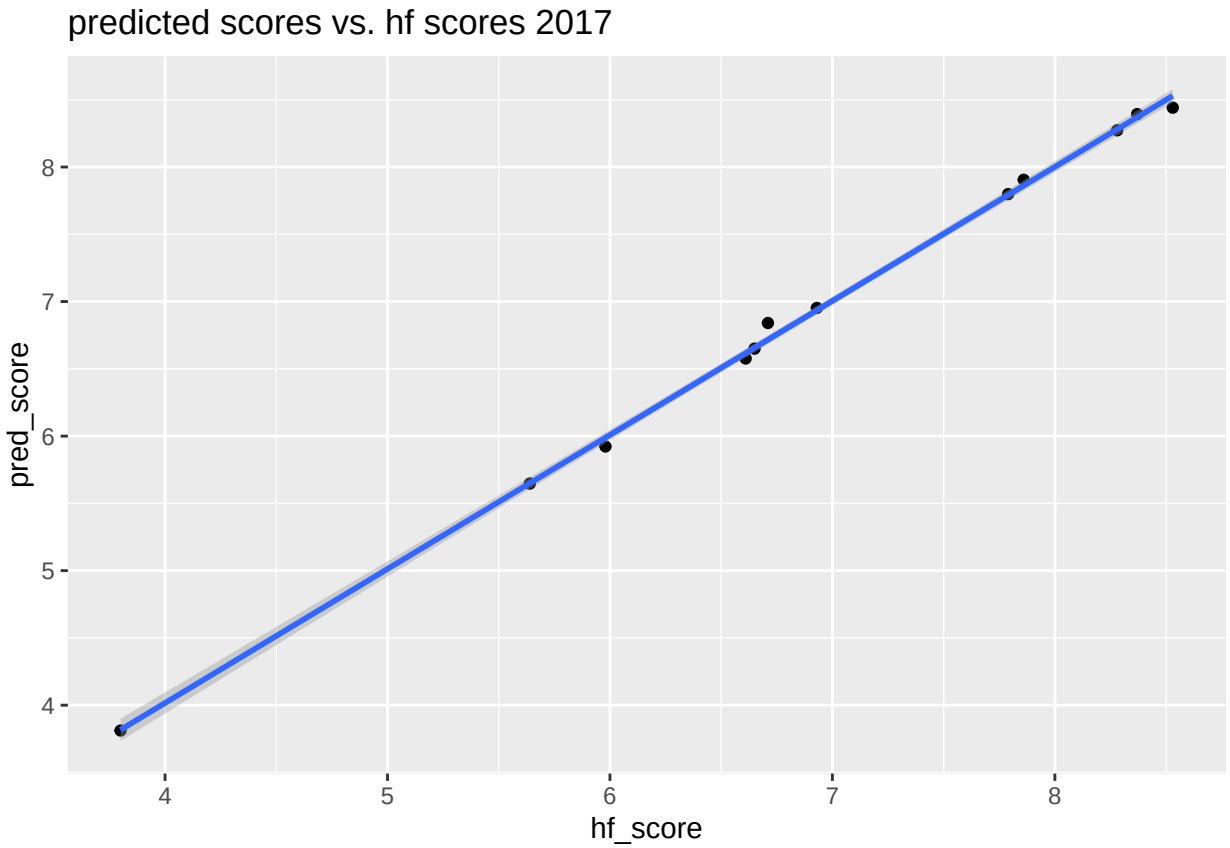
```
min(rmses)
```

```
## [1] 0.03970213
```

```
psize[which.min(rmses)]
```

```
## [1] 0.87
```

The data partition is split to a train set 0.9 and the test set of 0.1. Now we calculate the fit model to obtain coefficients based on the data of 2017.



The coefficients based on the fitted model are slightly different to the original.

```
##     (Intercept)          pf_rol           pf_ss     pf_movement     pf_religion
##     -0.15775105      0.18738475      0.13343871      0.04844705      0.04319749
##  pf_association   pf_expression     pf_identity   ef_government ef_legal_gender
##      0.04317739      0.05421190      0.04943618      0.09632355      0.19614115
##        ef_money        ef_trade   ef_regulation
##      0.10420887      0.09717430      0.12985263
```

A third model with the coefficients of the fitted model is calculated according following equation.

$$pred\_score_i = pf\_rol_i * coefficient_{pf\_rol} + ... + ef\_regulation_i * coefficient_{ef\_regulation}$$

The RMSE with all indicators and fitted coefficents based on the model 3 is:

```
rsme_model3 <- RMSE(model3$hf_score, model3$pred_score)
rsme_model3
```

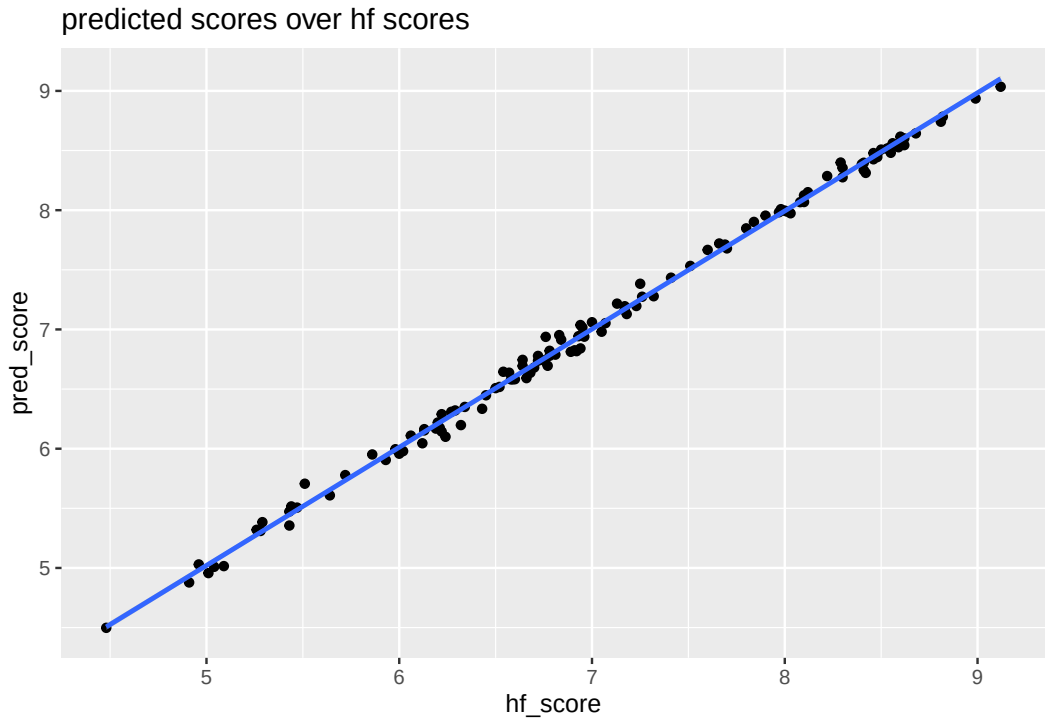
```
## [1] 0.05264264
```

We see a lower RSME for the model 3 compared tho the former models and the scores are closer the the original hf_score as expected. The top countries with predicted scores from model 3 are:

Table 1: hf_score compared to predicted score for the top countries

countries	hf_score	pred_score
New Zealand	8.88	8.778
Switzerland	8.82	8.762
Hong Kong	8.81	8.736
Canada	8.65	8.601
Australia	8.62	8.591
Denmark	8.56	8.571
Luxembourg	8.56	8.484
Finland	8.53	8.441
Germany	8.53	8.571
Ireland	8.52	8.511
Netherlands	8.5	8.475
Sweden	8.5	8.498

For the final model we are going to repeat the steps on the full dataset. Now we calculate the fit model to obtain coefficients based on the data of 2017.



The coefficents based on the fitted model are slightly different to the original.

```
##     (Intercept)          pf_rol           pf_ss     pf_movement     pf_religion
##     -0.18341432      0.19724894      0.13204052      0.04958719      0.04285139
##  pf_association   pf_expression     pf_identity   ef_government ef_legal_gender
##      0.04449375      0.04636675      0.04833533      0.09236958      0.25917652
##        ef_money        ef_trade   ef_regulation
##      0.10826029      0.09446472      0.12882842
```

With the coefficents of the fitted model of all data, a forth model is calculate according following equation.

$$pred\_score_i = pf\_rol_i * coefficient\_full_{pf\_rol} + ... + ef\_regulation_i * coefficient\_full_{ef\_regulation}$$

The RMSE with all indicators and fitted coefficents based on the model 4 is:

```
rsme_model4 <- RMSE(model4$hf_score, model4$pred_score)
rsme_model4
```

```
## [1] 0.05458624
```

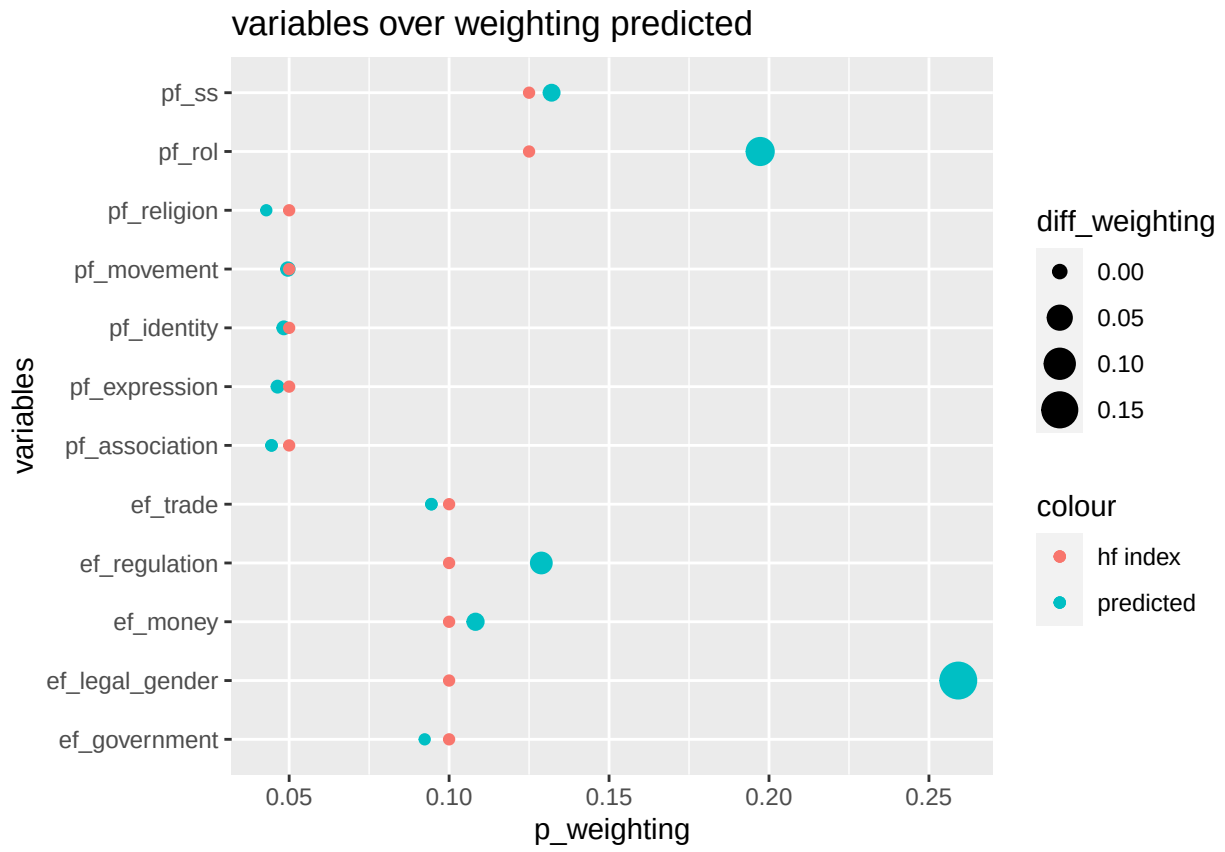
The RSME is slighlty higher compared to the third model and correlation between the predicted coefficients ant the actual score becomes lower.

```
cor(p_weighting,hf_weighting, method="pearson")
```

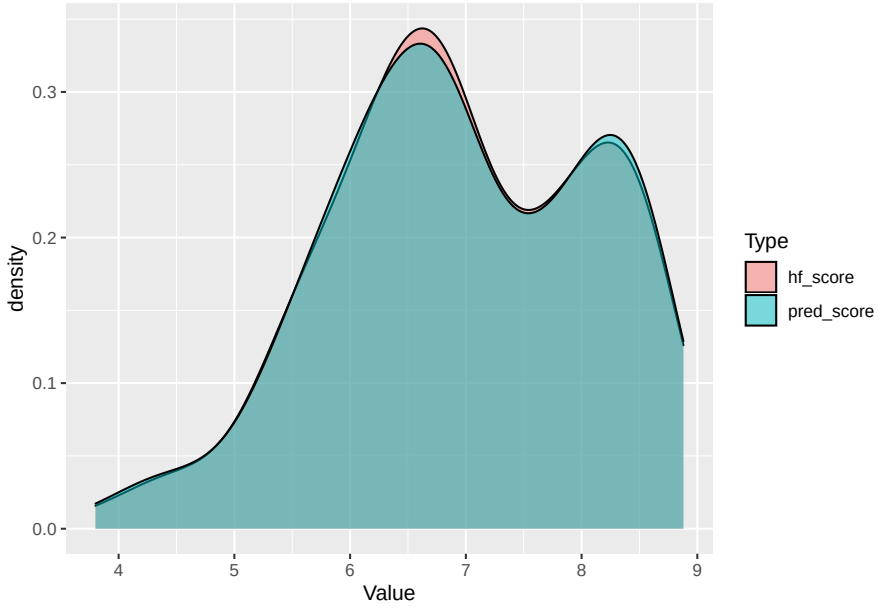
```
## [1] 0.7536293
```

Resuls

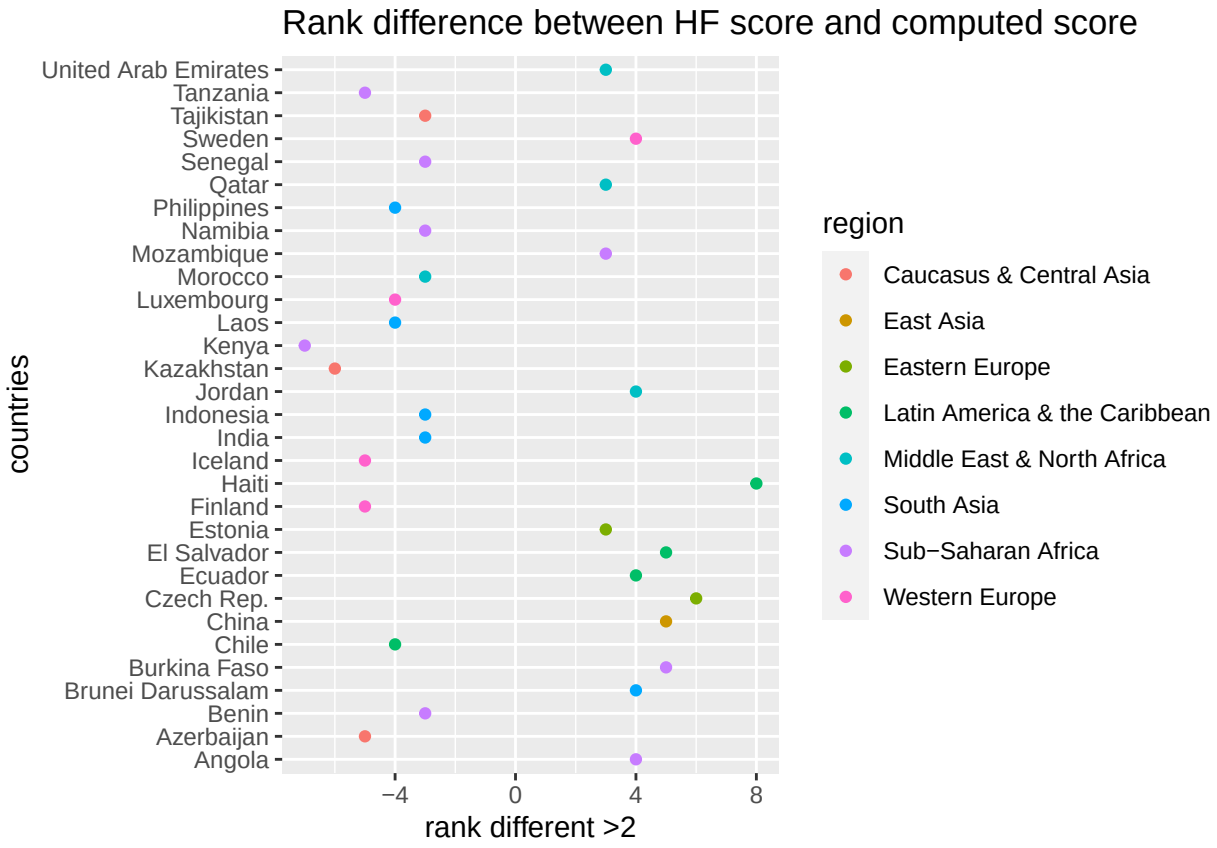
We have a model, that predicts the score based on coefficients predicted form the dataset. Following plot gives an overview of the changes to the weighting of the variables. We see the biggest different on pf_rol and ef_legal_gender, both are much higher ranked. Are we underestimate these variables?



The density of the scores is displayed in following plot, we see a small shift at the peaks.



The countries with the biggest changes and their direction from the original score to the data optimized are the following.



We see winner and loser countries with new model, the top winner with a increase of two or more ranks are the following.

Table 2: Increase of number of countries with a difference of ranks by 2 or more between HFI coefficients and the fitted model

countries	region	rank_diff
Haiti	Latin America & the Caribbean	8
Czech Rep.	Eastern Europe	6
El Salvador	Latin America & the Caribbean	5
Burkina Faso	Sub-Saharan Africa	5
China	East Asia	5
Sweden	Western Europe	4

These are the countries which lost freedom by two or more ranks acording the new model.

Table 3: Decrease of number of countries with a difference of ranks by 2 or more between HFI coefficients and the fitted model

countries	region	rank_diff
Kenya	Sub-Saharan Africa	-7
Kazakhstan	Caucasus & Central Asia	-6
Finland	Western Europe	-5
Iceland	Western Europe	-5
Tanzania	Sub-Saharan Africa	-5
Azerbaijan	Caucasus & Central Asia	-5

The biggest winner regions with countries increased by 2 or more ranks are in the top table, contries with 2 or more ranss declined are at the bottom.

Table 4: Increase of number of countries with a difference of ranks by 2 or more by region

region	count
Latin America & the Caribbean	3
Middle East & North Africa	3
Sub-Saharan Africa	3
Eastern Europe	2
East Asia	1
South Asia	1
Western Europe	1

Table 5: Decrease of number of countries with a difference of ranks by 2 or more between HFI by region

region	count
Latin America & the Caribbean	1
Middle East & North Africa	1
Caucasus & Central Asia	3
Western Europe	3
South Asia	4
Sub-Saharan Africa	5

For completeness all rmse for all models are as following.

Table 6: Model and RSME

method	RMSE
average all	0.4738
weighted HFI	0.467
mutated coefficients 2017	0.05264
mutated coefficients all years	0.05459

Conclusion

It is surprising that the hf_score cannot be derived from the indicators.With a general linear model we could predict more suitable coefficients from the data, the weighting does not correlate as strong as expected with the coefficients of the HFI.

Why this shift occurs cannot be assessed, probably the importance of variables is not due to the data but to subjective evaluation, which may be closer to reality.

Appendix

Environment

```
##                _
## platform       x86_64-apple-darwin15.6.0
## arch           x86_64
## os             darwin15.6.0
## system         x86_64, darwin15.6.0
## status
## major          3
## minor          6.3
## year           2020
## month          02
## day            29
## svn rev        77875
## language       R
## version.string R version 3.6.3 (2020-02-29)
## nickname       Holding the Windsock
```

References

Ian Vásquez and Tanja Porčnik, The Human Freedom Index 2018: A Global Measurement of Personal, Civil, and Economic Freedom (Washington: Cato Institute, Fraser Institute, and the Friedrich Naumann Foundation for Freedom, 2018).