

**Бонусное домашнее задание**  
**Практикум по разведывательному ФА**  
**Рубанов Владислав БПТ 201**

**Описание данных и теоретические гипотезы**

Передо мной стоит задача выделения разных **пониманий-трактовок демократии**, свойственных для России, которые будут рассмотрены как латентные факторы.

В качестве **эмпирической базы** будет представлен массив данных из 6-й волны исследования World Values Survey (WVS). Для анализа будут взяты блок вопросов по трактовке демократии (переменные V131 – V139). Все они представляют собой некоторую возможную характеристику демократии, применимость которой к концепту „демократии“ нужно оценить по шкале от 1 до 10, где 1 — „совсем не является «существенной характеристикой демократии»“; а 10 — „определенно является «существенной характеристикой демократии»“.

Кратко опишем имеющиеся вопросы (переменные), чтобы мы могли предположить, что может стоять за их совместной изменчивостью:

1. V131: «Governments tax the rich and subsidize the poor».  
V131: «Правительство облагает налогом богатых и субсидирует бедных».
2. V132: «Religious authorities ultimately interpret the laws».  
V132: «Религиозные власти в конечном счете интерпретируют законы».
3. V133: «People choose their leaders in free elections».  
V133: «Люди выбирают своих лидеров на свободных выборах».
4. V134: «People receive state aid for unemployment».  
V134: «Люди получают государственную помощь по безработице».
5. V135: «The army takes over when government is incompetent».  
V135: «Армия берет власть в свои руки, когда правительство некомпетентно».
6. V136: «Civil rights protect people from state oppression».  
V136: «Гражданские права защищают людей от государственного угнетения».
7. V137: «The state makes people's incomes equal».  
V137: «Государство делает доходы людей равными».
8. V138: «People obey their rulers».  
V138: «Люди подчиняются своим правителям».
9. V139: «Women have the same rights as men».  
V139: «Женщины имеют те же права, что и мужчины».

Очевидно, что ответы на некоторые из этих вопросов будут связаны между собой — это заметно невооруженным взглядом, поскольку они „вписываются“ в схожее понимание демократии. Следовательно, вероятно, мы будем наблюдать некоторую совместную изменчивость в наших данных.

Для предположения о том, какие латентные факторы могут объяснять совместную изменчивость данных, хотелось бы прежде всего обратиться к статье *К. Вельцеля и Х. Кирша*, посвященной авторитарным представлениям демократии и объяснению так называемого „**парадокса демократии**“:

сосуществование массового стремления к демократии с отсутствием последней<sup>1</sup>. В данной статье авторы, также обращаясь к WVS, доказывают, что, **чем больше людей неправильно понимают демократию** (авторитарным образом, а не либеральным), **тем в большей степени фактический уровень демократии в стране отстает от того уровня, который декларируют сами люди** (поскольку таким образом происходит легитимация авторитарного режима). Кроме того, в их исследовании Россия заняла высокое место в рейтинге стран по распространенности авторитарного понимания демократии (между Китаем и Эквадором).

Итак, К. Вельцель и Х. Кирш в исследовании явно выделяют два фактора, объясняющих совместную изменчивость в данных: это **авторитарное понимание демократии** (authoritarian notions of democracy — ANDs) относительно вопросов V132, V135 и V138, а также **либеральное понимание демократии** (liberal notions of democracy — LNDs) относительно вопросов V133, V136 и V139, доказывая значимость этих факторов на уровне всех государств и наблюдений, включенных в WVS, поэтому было бы глупо игнорировать данные наработки. Однако в исследовании они явно не учитывают оставшиеся три вопроса относительно трактовок демократии: V131, V134 и V137, поскольку они не попадают под предмет их исследования. Можно предположить, что совместная изменчивость в ответах на данные три вопроса может быть связана с **пониманием людьми демократии как социальной справедливости или перераспределения благ**, что действительно сложно отнести к предыдущим двум латентным факторам.

Таким образом, поскольку данный блок вопросов довольно очевидно делится на 3 подгруппы, в качестве первого теоретического предположения можно ожидать увидеть три латентных фактора, свойственных для России, которые будут описывать совместную изменчивость переменных, а именно:

1. *авторитарные взгляды* (трактовки) для V132, V135, v138;
2. *либеральные взгляды* (трактовки) для V133, V136, V139;
3. *установка на социальную справедливость* для V131, V134, V137.

Кроме того, **можно предположить**, что между 3 фактором будет присутствовать корреляция с 1 фактором (в большей степени) и со 2 фактором (в меньшей степени), но между 1 и 2 факторами корреляции быть не должно. Таким образом, наличие взаимосвязи между 3 фактором и другими двумя может стать **проблемной точкой разведывательного этапа**.

После того, как мы выдвинули теоретическое предположение относительно латентных факторов, настало время посмотреть на сами данные и описать их.

**Загрузим** набор данных (ответы на представленные выше вопросы респондентов из России), **выберем** интересующие переменные и удалим строки, в которых есть пропущенные ответы.

```
library(psych)
library(haven)

data <- read_sav("WVS6_RU.sav")

# отберем интересующие переменные
data <- subset(data, select = c(V131:V139))
data <- na.omit(data)
head(data)

## # A tibble: 6 x 9
##       V131      V132      V133 V134 V135 V136 V137 V138 V139
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
```

<sup>1</sup>Welzel, C. H. Kirsch (2017). „Democracy Misunderstood: Authoritarian Notions of Democracy around the Globe.“ *World Values Research* 9 (1): 1–29.

```
## 1      8 [8]      3 [3] 10 [An essential char~ 9 [9] 6 [6] 9 [9] 8 [8] 4 [4] 4 [4]
## 2      8 [8]      3 [3] 10 [An essential char~ 9 [9] 6 [6] 9 [9] 8 [8] 4 [4] 4 [4]
## 3      8 [8]      3 [3] 10 [An essential char~ 9 [9] 6 [6] 9 [9] 8 [8] 4 [4] 4 [4]
## 4      8 [8]      3 [3] 10 [An essential char~ 9 [9] 6 [6] 9 [9] 8 [8] 4 [4] 4 [4]
## 5      8 [8]      3 [3] 10 [An essential char~ 9 [9] 6 [6] 9 [9] 8 [8] 4 [4] 4 [4]
## 6      8 [8]      3 [3] 10 [An essential char~ 9 [9] 6 [6] 9 [9] 8 [8] 4 [4] 4 [4]
```

Отлично, теперь **опишем** наши данные: посмотрим на описательные статистики, корреляционную матрицу и их распределение.

```
describe(data)
```

##	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
## V131	1	1849	7.18	2.69	8	7.52	2.97	1	10	9	-0.75	-0.44	0.06
## V132	2	1849	3.82	2.56	3	3.54	2.97	1	10	9	0.61	-0.55	0.06
## V133	3	1849	8.13	2.37	9	8.54	1.48	1	10	9	-1.29	0.86	0.06
## V134	4	1849	7.91	2.42	9	8.30	1.48	1	10	9	-1.12	0.40	0.06
## V135	5	1849	4.77	2.85	5	4.61	2.97	1	10	9	0.23	-1.01	0.07
## V136	6	1849	8.03	2.28	9	8.40	1.48	1	10	9	-1.19	0.73	0.05
## V137	7	1849	7.30	2.57	8	7.62	2.97	1	10	9	-0.75	-0.34	0.06
## V138	8	1849	6.37	2.71	6	6.54	2.97	1	10	9	-0.32	-0.85	0.06
## V139	9	1849	8.22	2.31	9	8.60	1.48	1	10	9	-1.24	0.72	0.05

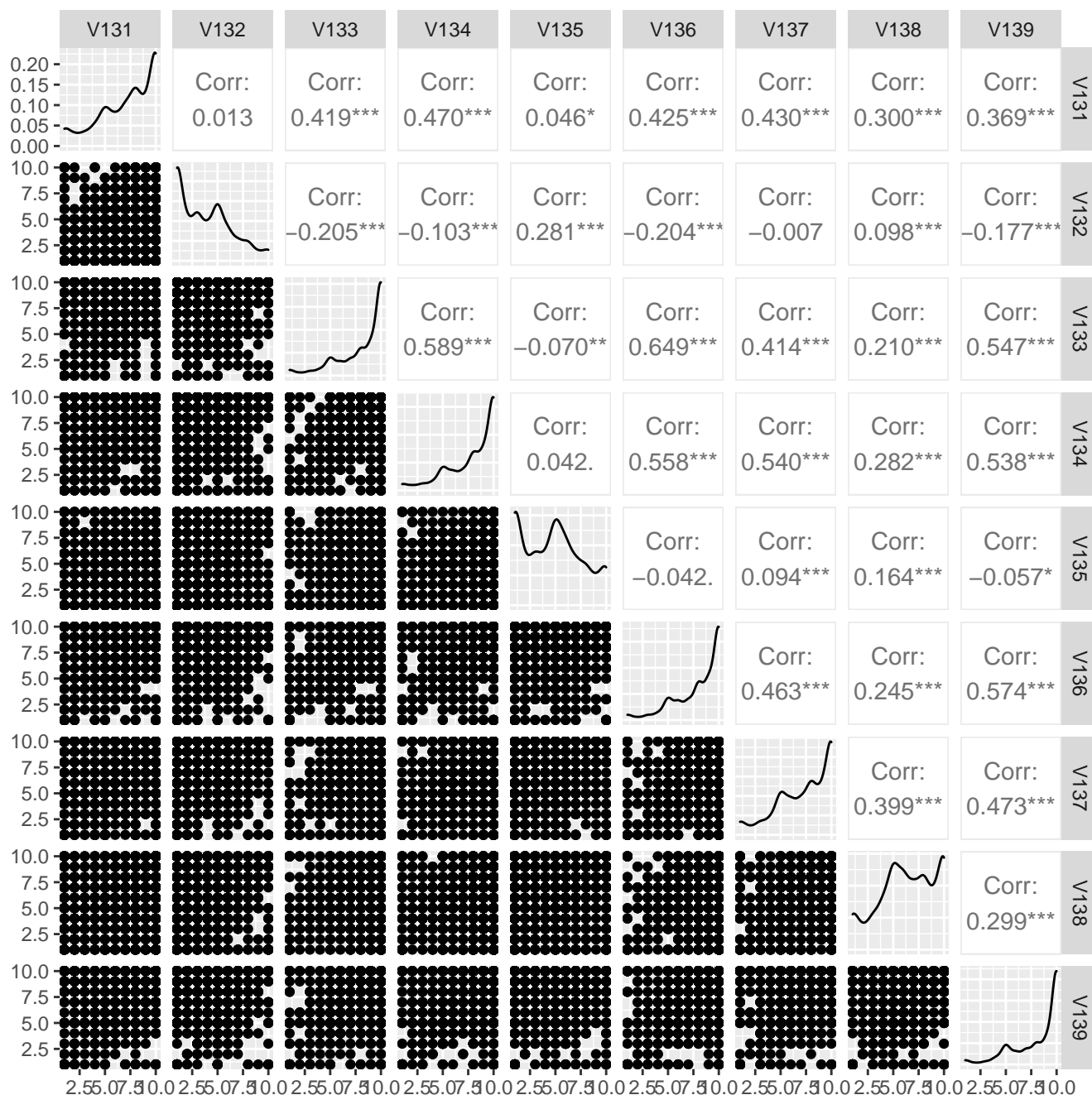
```
corr <- cor(data, method = "spearman")
corr
```

##	V131	V132	V133	V134	V135	V136
## V131	1.00000000	-0.07239203	0.4470183	0.5202907	0.01405544	0.4451097
## V132	-0.07239203	1.00000000	-0.3099204	-0.1893274	0.29102405	-0.3141552
## V133	0.44701825	-0.30992043	1.00000000	0.6158371	-0.11770327	0.6807199
## V134	0.52029070	-0.18932740	0.6158371	1.00000000	-0.01411250	0.5785297
## V135	0.01405544	0.29102405	-0.1177033	-0.0141125	1.00000000	-0.1075092
## V136	0.44510974	-0.31415515	0.6807199	0.5785297	-0.10750920	1.00000000
## V137	0.46410966	-0.07962482	0.4407782	0.5631498	0.05792072	0.4849326
## V138	0.32511683	0.05014757	0.2123351	0.3013520	0.15728553	0.2539548
## V139	0.41159266	-0.26939238	0.5860151	0.5651308	-0.11951070	0.5947433

##	V137	V138	V139
## V131	0.46410966	0.32511683	0.4115927
## V132	-0.07962482	0.05014757	-0.2693924
## V133	0.44077824	0.21233514	0.5860151
## V134	0.56314976	0.30135200	0.5651308
## V135	0.05792072	0.15728553	-0.1195107
## V136	0.48493264	0.25395477	0.5947433
## V137	1.00000000	0.40877002	0.4970162
## V138	0.40877002	1.00000000	0.3071577
## V139	0.49701619	0.30715768	1.00000000

```
library(GGally)
ggpairs(data)
```



**Во-первых**, можно заметить, что выделенные группы вопросов по большей части имеют похожие средние и медианные значения: у либеральных показателей они выше всех, у социальных — выше среднего, а у авторитарных — низкие (*кроме вопроса V138 про подчинение правителям, и это интересно*).

**Во-вторых**, мы действительно можем наблюдать наличие совместной изменчивости наблюдаемых показателей. Теоретическое предположение о том, что связь между факторами может существовать (если выбирать 3 фактора), а также о том, что наибольшая взаимосвязь между факторами прослеживается между либеральным представлением о демократии и социальной справедливостью, а между авторитарным представлением о демократии и другими факторами — нет — подтвердилось (**кроме вопроса V138** про подчинение правителям, что, опять же, примечательно). Также, безусловно, присутствует сильная связь внутри выделенных факторов (пожалуй, кроме вопроса V132 про религиозную составляющую).

Таким образом, выбор числа факторов будет лежать **между двумя и тремя факторами**. Вероятно, его можно описать как выбор между большей очевидностью выводов и меньшим числом объясненной информации (2 фактора) и между большей объясненной информацией (3 фактора), но скоррелированностью факторов и, следовательно, затрудненной интерпретацией. Также нужно держать в голове вопрос V138, поскольку он положительно коррелирует со всеми остальными показателями.

## Формальные тесты

Однако зададимся вопросом: **оправданно ли реализовывать факторный анализ** на этих данных, учитывая наличие некоторой фрагментарности в датасете, как было замечено из корреляционной матрицы? Ведь это может затруднить поиск общих концептов.

Для ответа на него применим **критерий (статистику) Кайзера-Майера-Олкина**, которая сравнит частные корреляции и обычные.

```
KMO(data)

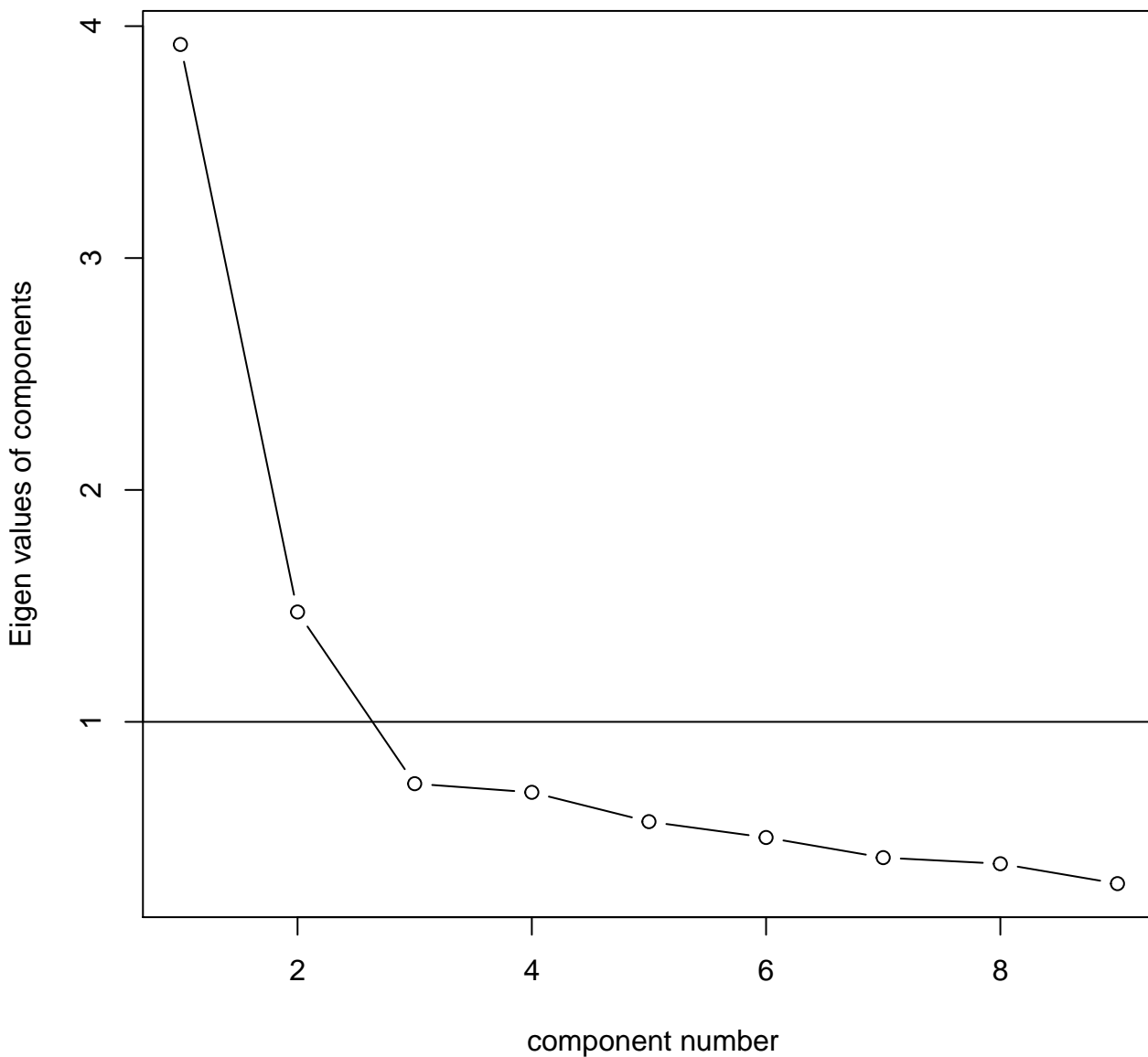
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data)
## Overall MSA = 0.87
## MSA for each item =
## V131 V132 V133 V134 V135 V136 V137 V138 V139
## 0.91 0.67 0.86 0.88 0.59 0.87 0.88 0.84 0.90
```

Итак, можно заметить, что общее **значение статистики КМО очень высоко: 0.87**, что говорит о низких значениях частных корреляций, а значит есть общие концепты, которые могли бы объяснить данные. Значения по каждому из наблюдаемых показателей также очень высоки: большинство больше 0.8, а два — 0.9 и больше. В этом плане выбиваются только два показателя V132 и V135 (про религию и армию), имеющие очевидную „авторитарную окраску“, это логично, что они имеют достаточно сильную частную корреляцию.

Теперь посмотрим на **график**, с помощью которого мы сможем определить количество латентных факторов.

```
VSS.scree(corr)
```

scree plot



Итак, мы видим, что, **во-первых**, переломной точкой на графике является число 3 (значит нужно выбрать или это число, или предыдущее — до переломного). **Во-вторых**, собственные значения больше 1 у двух факторов. Значит, согласно обоим методам, нам **стоит выбрать именно два латентных фактора**.

Вероятно, стоит признать, что концепты либерального понимания демократии и демократии как социальной справедливости в контексте России во многом *описывают одно и то же* (вспоминаем их сильную корреляцию между собой) и **объединить их в один фактор**, противопоставив его более очевидному фактору авторитарного понимания демократии. Таким образом, на данный момент логичнее всего выделить два следующих латентных фактора:

- *авторитарное понимание демократии;*
- *либеральное понимание демократии и социальная справедливость.*

Хотя необходимо отметить, что, теоретически, *либерализм слабо связан с социальной справедливостью, равенством результатов и перераспределением*, т.е. можно предположить, что россияне связывают либерализм с элементами социализма. *Это может заслуживать отдельного исследования.*

Таким образом, модель ФА в данном случае будет выглядеть следующим образом:

$$X_j = \alpha_{1j} \times F_1 + \alpha_{2j} \times F_2 + \varepsilon_j$$

## Оценка модели разведывательного ФА

Далее оценим модель разведывательного ФА, взяв два латентных фактора, используя метод главных факторов (осей) и вращение „varimax“.

```
efa1 <- fa(r = corr, nfactors = 2, fm = "pa", rotate = "varimax")
efa1

## Factor Analysis using method = pa
## Call: fa(r = corr, nfactors = 2, rotate = "varimax", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##          PA1   PA2   h2    u2 com
## V131  0.64  0.03 0.41 0.59 1.0
## V132 -0.16  0.56 0.35 0.65 1.2
## V133  0.71 -0.38 0.64 0.36 1.5
## V134  0.77 -0.12 0.61 0.39 1.1
## V135  0.04  0.47 0.23 0.77 1.0
## V136  0.72 -0.35 0.64 0.36 1.4
## V137  0.73  0.10 0.54 0.46 1.0
## V138  0.49  0.27 0.31 0.69 1.6
## V139  0.70 -0.27 0.56 0.44 1.3
##
##                               PA1  PA2
## SS loadings                   3.31 0.98
## Proportion Var                 0.37 0.11
## Cumulative Var                 0.37 0.48
## Proportion Explained           0.77 0.23
## Cumulative Proportion          0.77 1.00
##
## Mean item complexity = 1.2
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are 36 and the objective function was 3.19
## The degrees of freedom for the model are 19 and the objective function was 0.07
##
## The root mean square of the residuals (RMSR) is 0.02
## The df corrected root mean square of the residuals is 0.03
##
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##
##                               PA1  PA2
## Correlation of (regression) scores with factors 0.93 0.77
## Multiple R square of scores with factors         0.87 0.59
## Minimum correlation of possible factor scores    0.73 0.18

print(efa1$loadings, cutoff=0.3)
```

```
##
## Loadings:
##      PA1      PA2
## V131  0.639
## V132      0.565
## V133  0.707 -0.380
## V134  0.774
## V135      0.475
## V136  0.721 -0.350
## V137  0.729
## V138  0.487
## V139  0.698
##
##              PA1      PA2
## SS loadings  3.309 0.980
## Proportion Var 0.368 0.109
## Cumulative Var 0.368 0.477
```

Прежде всего, мы видим перед собой **таблицу стандартизированных факторных нагрузок**, основанных на корреляционной матрице. То есть это собственные векторы, которые показывают, насколько наши факторы объясняют наблюдаемые показатели (вопросы про демократию). Ниже можно увидеть эту табличку со значениями факторных нагрузок, которые превышают 0,3.

Что можно сказать о закономерностях? Мы помним, что факторная нагрузка в квадрате — это доля объясненной вариации посредством этого фактора. В целом можно сказать, что **доля общности** (объясненной фактором информации) **лишь в 3-х случаях превышает 50%**: для V133, V135, V138. Большие специфичности (необъясненная информация) говорят о том, что в последствии придется изменять спецификацию модели, чтобы избежать больших ошибок — это еще одна **проблемная точка** работы.

Тем не менее, уже можно **наметить структуру данных**, предположив, как в последствии могут выглядеть закономерности. По коэффициентам и долям объясненной информации можно утверждать, что **PA1** отвечает за либерально-перераспределительные ценности, а **PA2** — за авторитарное понимание демократии.

Обращаясь к показателю *complexity*, можно заметить, что *большая часть показателей в среднем так или иначе связана с одним фактором*. Так, вопросы V131, 134, 136, 137, 139 скорее можно отнести к первому фактору, а вопросы V132, 135 — ко второму. При этом вопросы V133, V138 нельзя явно отнести к какому-то одному фактору ( $\cos \geq 1.5$ ).

Обращаясь ко второй таблице, показывающей **информативность решения**, можно заметить, что два выделенных латентных фактора объяснили лишь 48% информации, что, безусловно, мало. Причем на долю первого фактора (предположительно, либерально-перераспределительных ценностей) пришлось сразу 77% объясненной информации, а на второй фактор — 23%. Таким образом, можно заметить **еще одну проблемную точку: неинформативность нашего решения**.

Однако, при этом, значение **RMSR** (корень квадрата остатков) маленькое: всего 0.02 (эмпирическим порогом является 0.08). Также можно обратиться к значению  $R^2$  для обоих факторов: для первого фактора (либерально-перераспределительные ценности) он высок: 87%, для второго фактора (авторитарное понимание демократии) — ниже: 59%

## Дополнительная модель

По итогам данного анализа может возникнуть резонное желание **выделить 3 латентных фактора** — попробуем сделать это. В таком случае модель будет выглядеть следующим образом:

$$X_j = \alpha_{1j} \times F_1 + \alpha_{2j} \times F_2 + \alpha_{3j} \times F_3 + \varepsilon_j$$



Протестируем модель:

```
efa2 <- fa(r = corr, nfactors = 3, fm = "pa", rotate = "varimax")
efa2

## Factor Analysis using method = pa
## Call: fa(r = corr, nfactors = 3, rotate = "varimax", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      PA1   PA3   PA2   h2   u2 com
## V131  0.41  0.48  0.03 0.40 0.60 2.0
## V132 -0.21 -0.03  0.57 0.37 0.63 1.3
## V133  0.79  0.28 -0.24 0.75 0.25 1.4
## V134  0.58  0.52 -0.09 0.61 0.39 2.0
## V135 -0.03  0.09  0.50 0.26 0.74 1.1
## V136  0.63  0.41 -0.28 0.64 0.36 2.2
## V137  0.34  0.68  0.02 0.58 0.42 1.5
## V138  0.13  0.54  0.18 0.34 0.66 1.3
## V139  0.47  0.52 -0.29 0.57 0.43 2.6
##
##
##      PA1   PA3   PA2
## SS loadings      1.91 1.78 0.84
## Proportion Var    0.21 0.20 0.09
## Cumulative Var    0.21 0.41 0.50
## Proportion Explained 0.42 0.39 0.19
## Cumulative Proportion 0.42 0.81 1.00
##
## Mean item complexity = 1.7
## Test of the hypothesis that 3 factors are sufficient.
##
## The degrees of freedom for the null model are 36 and the objective function was 3.19
## The degrees of freedom for the model are 12 and the objective function was 0.03
##
## The root mean square of the residuals (RMSR) is 0.01
## The df corrected root mean square of the residuals is 0.02
##
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##
##      PA1   PA3   PA2
## Correlation of (regression) scores with factors 0.81 0.79 0.71
## Multiple R square of scores with factors        0.66 0.63 0.51
## Minimum correlation of possible factor scores    0.33 0.25 0.01

print(efa2$loadings, cutoff=0.3)

##
## Loadings:
##      PA1   PA3   PA2
## V131  0.411  0.485
## V132           0.570
## V133  0.787
## V134  0.578  0.519
## V135           0.501
```

```
## V136 0.626 0.409
## V137 0.335 0.681
## V138 0.536
## V139 0.467 0.518
##
##          PA1    PA3    PA2
## SS loadings 1.907 1.777 0.838
## Proportion Var 0.212 0.197 0.093
## Cumulative Var 0.212 0.409 0.502
```

Однако следует заметить, что **ситуация не изменилась в лучшую сторону**. *Во-первых*, общая доля общности увеличилась лишь на 0.02, достигнув ровно 50%. *Во-вторых*, доли объясненной информации для каждого отдельного показателя всего 1 раз увеличились больше, чем на 5%. *В-третьих*, теперь большая часть показателей в среднем связана с 2 факторами: как раз из опасения подобной ситуации на предыдущем шаге было выбрано 2 фактора, а не 3. Теперь можно заметить, что **либеральное понимание демократии и запрос на перераспределение действительно очень сильно связаны между собой**, что затрудняет их интерпретацию при раздельном оценивании.

## Выводы по итогам разведывательного ФА

Следует отметить, что **полученные результаты разведывательного ФА расходятся с моими исходными предположениями**. Несмотря на то, что, казалось бы, представленные из WVS вопросы должны довольно очевидно объясняться выделенными двумя (или тремя) факторами. Однако следует признать, что, тем не менее, данные латентные факторы имеют право на существования, хоть они и объясняют только половину общности.

В качестве единственного **полезного (эффективного) вывода** можно выделить тот факт, что для россиян ответы на вопросы, связанные с либеральным пониманием демократии и на вопросы, связанные с перераспределением благ очень связаны между собой (иногда — неотделимы друг от друга), что является теоретически неверным представлением о демократии (а именно как о справедливом порядке, где государство перераспределяет ресурсы в пользу бедных). Вторая модель скорее похожа на так называемый „скандинавский социализм“, свойственный Дании, Исландии, Норвегии, Швеции и Финляндии.