НИУ ВШЭ, ОП «Политология», 2022
Курс «Регрессионный анализ: панельные данные и каузальность»

Бонусное домашнее задание 1
Рубанов Владислав, БПТ 201

Для удобства буду подкреплять свои рассуждения фрагментами кода из R.

Задание 1.

Мы задались следующим вопросом: *"Можно ли получить $\hat{\beta}_1$ посредством оценивания разных моделей отдельно на подгруппах, заданных пространственными единицами?„*

И ответ на него: **"Да, можно„**.

Эту процедуру действительно можно назвать **взвешиванием**, и вот почему:

- В начале мы *строим N моделей* (по числу пространственных единиц, которые мы имеем (всего i моделей, где $i \in \{1, 2, ..., N\}$).
  Тогда у нас получится N моделей вида:

  $$\hat{y}_{t_{\{i\}}} = \hat{a}_{0_{\{i\}}} + \hat{a}_{1_{\{i\}}} \cdot \hat{x}_{it}$$

  Оценка константы $\hat{a}_{0_{\{i\}}}$ нас сейчас не интересует, поэтому мы **забираем** все получившиеся $\hat{a}_{1_{\{i\}}}$ — *оценки коэффициентов при предикторе.*

  Задумаемся, *как модель должна их учитывать?* Очевидно, что не просто как среднее арифметическое всех оценок, ведь это было бы странно — мы бы теряли много информации в таком случае. Логично, что какие-то страны вносят больший вклад в модель, а какие-то — меньший.

- Отличным критерием информативности оценок той или иной пространственной единицы, с которой мы работаем может служить **условная дисперсия**, а точнее ***доля*** от той общей дисперсии зависимой переменной, которую принимает на себя некоторая пространственная единица (или которая как бы *"приходится„* на эту пространственную единицу). Тогда доля условной дисперсии при i-й пространственной единице должна хорошо объяснять для нас разницу в информации, которая приходится на ту или иную единицу.

  Формально это можно записать так:

  $$weight_i = \frac{Var(X|country = i)}{\sum_{i=1}^{N} Var(X|country = i)}$$

- Теперь кажется логичным, что именно **доля условной дисперсии** может выступать в качестве *"веса„* полученных оценок $\hat{a}_{1_{\{i\}}}$. Так, мы будем ориентироваться на **вклад** каждой пространственной единицы в общую дисперсию (информацию) независимой переменной. В том числе логично, что те пространственные единицы, у которых нет изменчивости, *не будут вносить никакой вклад* в общую оценку: это логично, т.к. их условная дисперсия равна **0**: эти наблюдения во временной перспективе совсем не изменяются во времени, они являются лишь точкой на графике, которая не изменяется. Кроме того, это исключит небольшие (около-)случайные колебания.

- Таким образом, взвешенная сумма оценок отдельных регрессионных моделей приведет нас к общей оценке $\hat{\beta}_1$:

  $$\hat{\beta}_1 = \sum_{i=1}^{N} \hat{a}_{1i} \cdot \frac{Var(X|country = i)}{\sum_{i=1}^{N} Var(X|country = i)}$$

Итак, наша искомая оценка $\hat{\beta}_1$ является **взвешенной суммой оценок** соответствующих коэффициентов $\hat{a}_{1_{\{i\}}}$.

Понятно, что все эти предпосылки работают, когда мы имеем дело со **сбалансированной панелью**. Но что делать, если мы работаем с несбалансированными данными?

Нужно обязательно **учесть количество наблюдений**. Например, в качестве веса можно будет использовать не просто условную вариацию, а условную вариацию, домноженную на количество наблюдений по каждой подгруппе по аналогии с F-статистикой в рамках модели ANOVA. Такая нормировка выровняет показатели.

Проделаем все эти шаги в R:

```
library(haven)
library(plm)
library(dplyr)
library(psych)

panel<-read_dta("RAPDC_lab1.dta")

# получим значения условной вариации предиктора state_capacity
var <- summarize(group_by(panel, country), var(state_capacity))
var
```

```
## # A tibble: 27 x 2
##    country        `var(state_capacity)`
##    <chr>                          <dbl>
##  1 Albania                       1.21
##  2 Armenia                       0.745
##  3 Azerbaijan                    0.0748
##  4 Belarus                       0.0542
##  5 Bulgaria                      0.276
##  6 Croatia                       1.83
##  7 Czech Republic                0.0317
##  8 Estonia                       0.549
##  9 Georgia                       1.53
## 10 Hungary                       0.0561
## # ... with 17 more rows
```

```
# оценим набор моделей, чтобы получить оценки коэф-тов для каждой страны
a <- group_by(panel, country) %>%
  do(data.frame(beta = coef(lm(fh_polity ~ state_capacity, data = .))[2]))
a
```

```
## # A tibble: 27 x 2
## # Groups:   country [27]
##    country           beta
##    <chr>            <dbl>
##  1 Albania         1.32
##  2 Armenia         0.468
##  3 Azerbaijan     -1.28
##  4 Belarus        -8.00
##  5 Bulgaria        0.0766
##  6 Croatia         1.77
```

```
##  7 Czech Republic -2.64
##  8 Estonia         1.09
##  9 Georgia         0.519
## 10 Hungary         2.20
## # ... with 17 more rows
```

```
# запишем значения чистых оценок и условную вариацию
m <- as.data.frame(merge(a, var, by ="country"))
m
```

```
##            country        beta var(state_capacity)
## 1          Albania  1.32327974          1.20776729
## 2          Armenia  0.46759613          0.74544856
## 3       Azerbaijan -1.28345476          0.07479485
## 4          Belarus -7.99785371          0.05415636
## 5         Bulgaria  0.07660255          0.27623487
## 6          Croatia  1.76947840          1.82857786
## 7   Czech Republic -2.63679748          0.03171778
## 8          Estonia  1.08637328          0.54881883
## 9          Georgia  0.51867369          1.52757049
## 10         Hungary  2.20258823          0.05609440
## 11      Kazakhstan -0.48613953          0.70864877
## 12 Kyrgyz Republic -2.68101161          0.06227268
## 13          Latvia -0.23238248          0.24980317
## 14       Lithuania  1.02531974          0.25107826
## 15       Macedonia -0.32781987          0.59351373
## 16         Moldova  3.49571683          0.01927932
## 17        Mongolia -1.16029865          0.06448780
## 18          Poland  1.05161728          0.21313029
## 19         Romania  2.14798550          0.31834659
## 20          Russia -0.42715599          0.08617639
## 21          Serbia  1.43752581          3.42483921
## 22 Slovak Republic  3.63235219          0.09000245
## 23        Slovenia  1.39162188          0.05640271
## 24      Tajikistan  0.22775110          3.36315158
## 25    Turkmenistan -0.24134946          0.02389123
## 26         Ukraine  0.26484051          0.11511840
## 27      Uzbekistan  0.58017913          0.15969442
```

```
# посчитаем взвешенные коэффициенты
m$coef <- m$beta*(m$`var(state_capacity)`/sum(m$`var(state_capacity)`))
sum(m$coef)
```

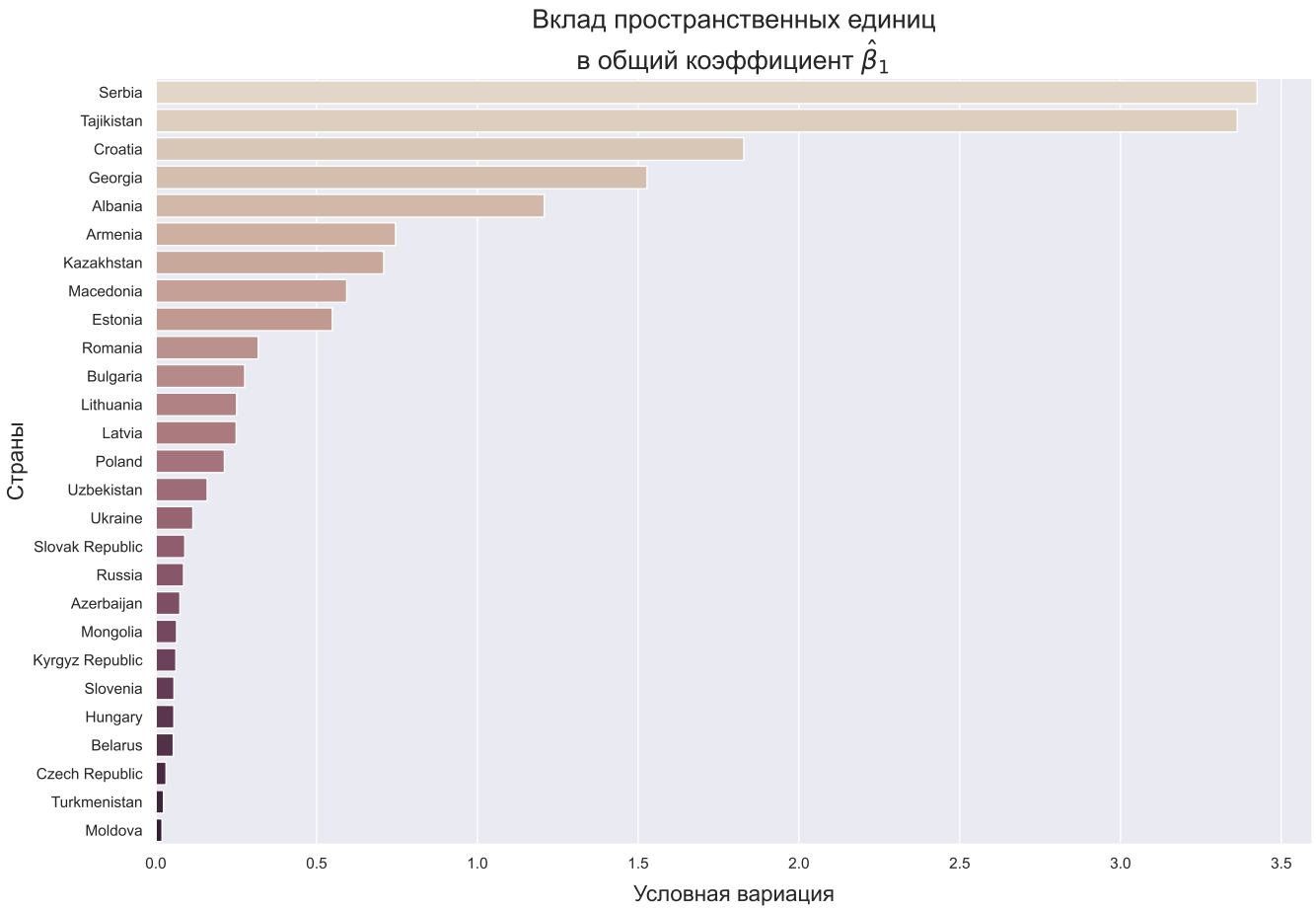
```
## [1] 0.7845941
```

Задание 2.
Покажем, какие страны получили наибольший вес в формировании оценки коэффициента при предикторе “государственная состоятельность„, а какие — наименьший.

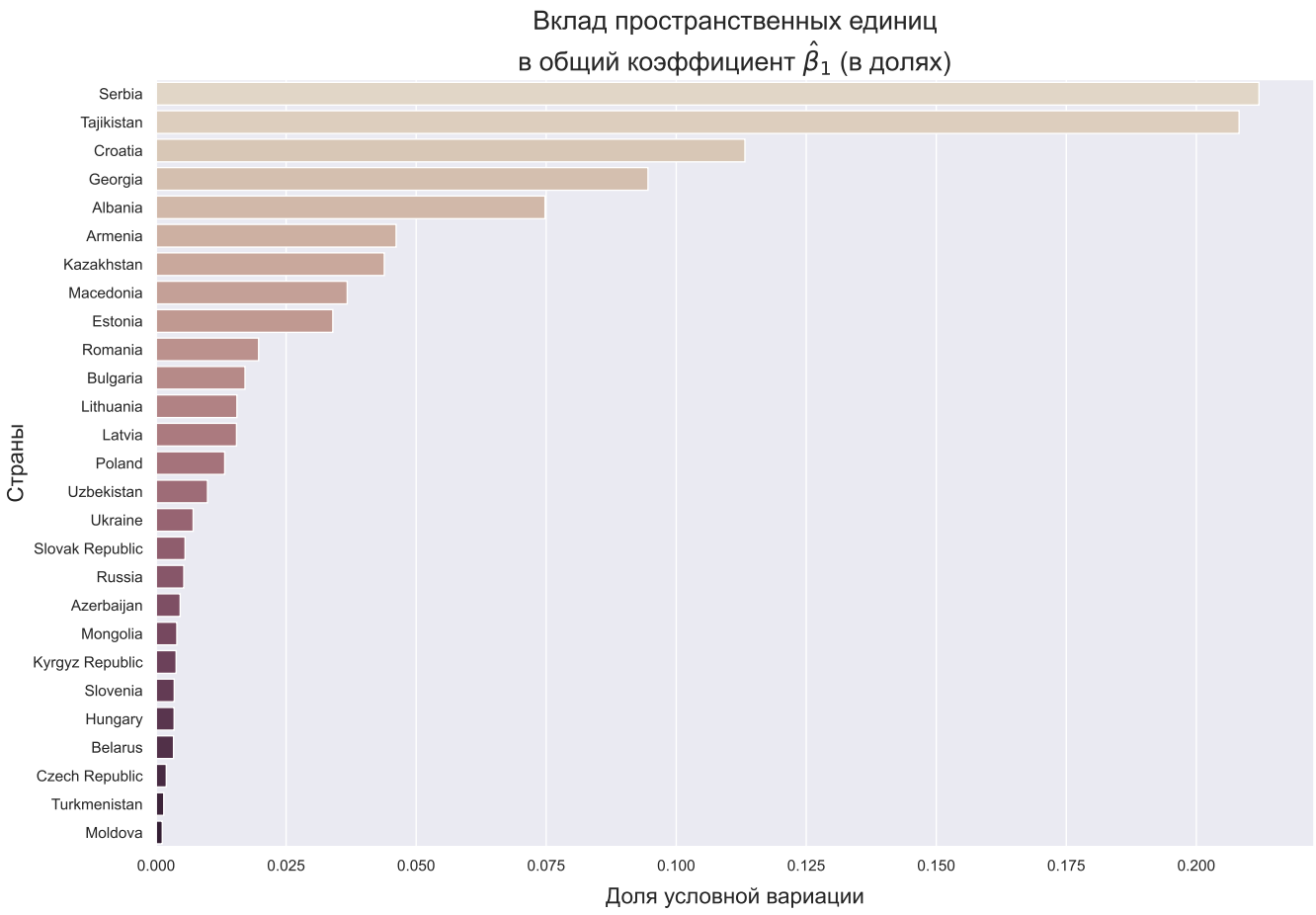
P.S. Построение графиков будет производиться в Python через seaborn.

Чтобы явно показать (и, может быть, удивиться), какие страны в наибольшей степени оказывают влияние на полученный результат, **проиллюстрируем** весь процесс взвешивания, описанный в предыдущем пункте.

Итак, в начале *выведем условную вариацию* предиктора “государственная состоятельность„, которая в дальнейшем будет выступать “весом„ для оценки коэффициента на каждую страну:



Для удобства можно также вывести те же значения, но **в терминах долей**, т.е. уже непосредственно весов в явном виде (разделив все значения на $\sum_{i=1}^{N} Var(X|country = i)$):



Таким образом, мы видим, что **наибольший вклад** в оценку $\hat{\beta}_1$ с большим отрывом будут вносить такие страны, как Сербия и Таджикистан. За ними также идут Хорватия, Грузия, Албания. Вес других стран составляет менее 5%.

Что стоит за этими цифрами? Как было сказано ранее, это значит, что для обозначенных выше стран наблюдается наибольшая **условная вариация** предиктора "государственная состоятельность,,, т.е. внутри этих пространственных единиц есть реальное изменение этого показателя во временной перспективе, на которое мы можем опираться — в этом заложено много информации.

Выведем это напрямую:

```
# выведем несколько примеров
panel[panel$country=='Serbia', ]

## # A tibble: 5 x 4
##   country period fh_polity state_capacity
##   <chr>    <dbl>     <dbl>          <dbl>
## 1 Serbia       1      2.65           1.88
## 2 Serbia       2      1.72           3.40
## 3 Serbia       3      6.55           3.31
## 4 Serbia       4      8.25           5.72
## 5 Serbia       5      8.67           6.35

panel[panel$country=='Tajikistan', ]

## # A tibble: 5 x 4
##   country    period fh_polity state_capacity
```

```
##   <chr>       <dbl>     <dbl>          <dbl>
## 1 Tajikistan      1      2.5           0
## 2 Tajikistan      2      1.63          0.859
## 3 Tajikistan      3      3.15          2.57
## 4 Tajikistan      4      3             3.22
## 5 Tajikistan      5      3             4.58
```

```
panel[panel$country=='Croatia', ]
```

```
## # A tibble: 5 x 4
##   country period fh_polity state_capacity
##   <chr>    <dbl>     <dbl>          <dbl>
## 1 Croatia      1      4.17           4.86
## 2 Croatia      2      3.85           6.19
## 3 Croatia      3      7.9            7.30
## 4 Croatia      4      8.87           8.15
## 5 Croatia      5      9.33           7.87
```

```
panel[panel$country=='Georgia', ]
```

```
## # A tibble: 5 x 4
##   country period fh_polity state_capacity
##   <chr>    <dbl>     <dbl>          <dbl>
## 1 Georgia      1      5.17           1.28
## 2 Georgia      2      6.12           1.75
## 3 Georgia      3      6.33           3.16
## 4 Georgia      4      7.07           3.72
## 5 Georgia      5      6.71           4.11
```

Действительно, мы видим, что для этих стран мы можем наблюдать довольно большую изменчивость для предиктора *state_capacity*. На теоретическом уровне это значит, что государственная состоятельность этих государств сильно изменилась за рассматриваемый период. И, как видно, это вызвало за собой смену политического режима (повлекло изменения в зависимой переменной *fh_polity*).

Можно также подтвердить тот факт, что изменений в предикторе *state_capacity* действительно существенны, посмотрев на описательные статистики по этой переменной:

```
# описательные статистики
summary(panel)
```

```
##    country              period    fh_polity      state_capacity
##  Length:135         Min.   :1   Min.   : 0.250   Min.   :0.000
##  Class :character   1st Qu.:2   1st Qu.: 4.083   1st Qu.:3.756
##  Mode  :character   Median :3   Median : 7.083   Median :4.944
##                     Mean   :3   Mean   : 6.367   Mean   :5.034
##                     3rd Qu.:4   3rd Qu.: 8.917   3rd Qu.:6.293
##                     Max.   :5   Max.   :10.000   Max.   :9.090
```

```
describe(panel)
```

```
##               vars   n  mean   sd median trimmed   mad  min   max range  skew
## country*         1 135 14.00 7.82  14.00   14.00 10.38 1.00 27.00 26.00  0.00
## period           2 135  3.00 1.42   3.00    3.00  1.48 1.00  5.00  4.00  0.00
```

```
## fh_polity         3 135  6.37 2.99   7.08    6.62  3.15 0.25 10.00  9.75 -0.60
## state_capacity    4 135  5.03 1.72   4.94    5.02  1.85 0.00  9.09  9.09  0.03
##                kurtosis   se
## country*          -1.23 0.67
## period            -1.33 0.12
## fh_polity         -0.91 0.26
## state_capacity    -0.22 0.15
```

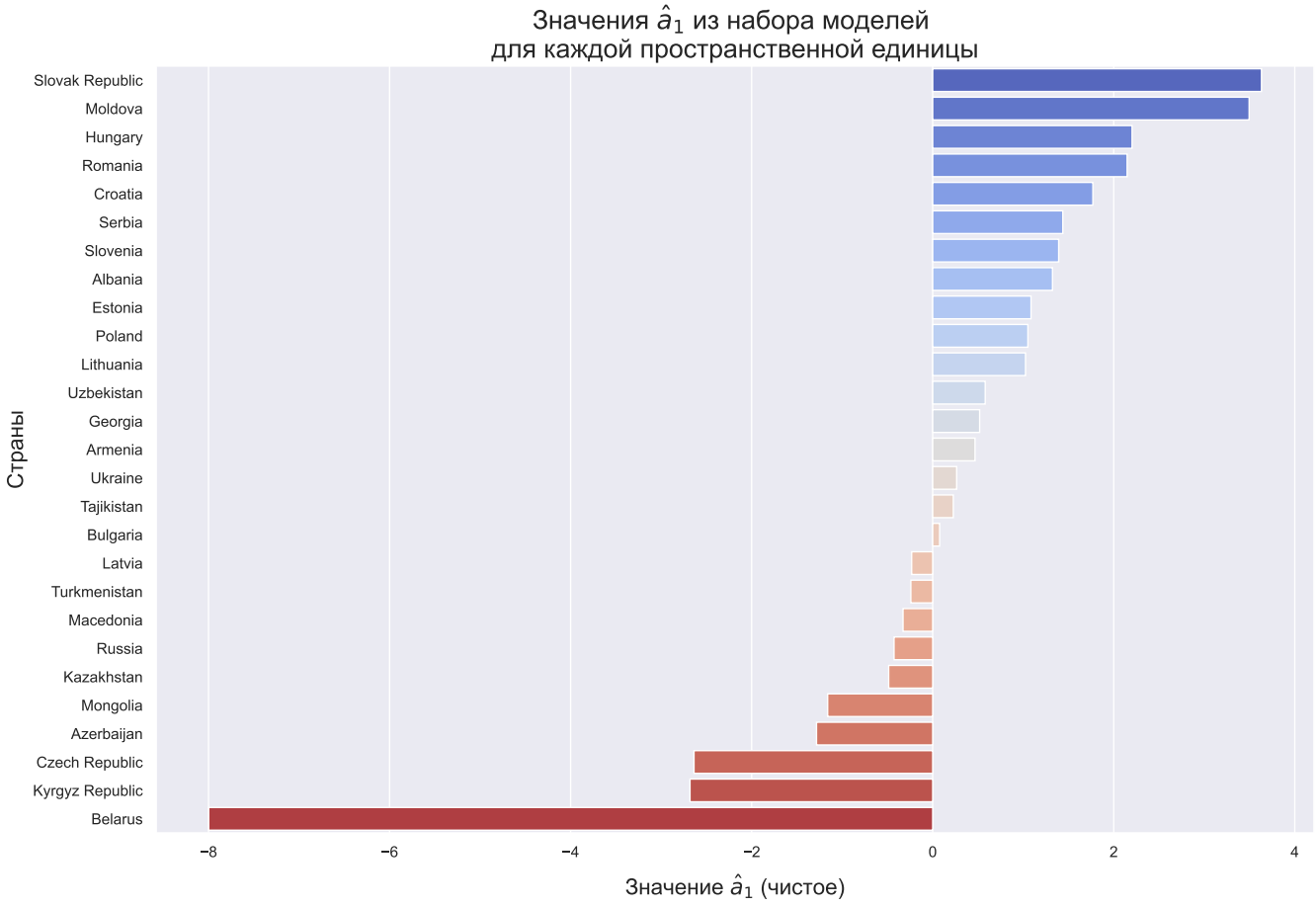
Видим, что **изменения в этих странах действительно существенны**: часто это был переход от типично низких значений из нижней четверти к типично высоким из высокой четверти. Вероятно, это можно связать с процессами демократизации и распадом "советского блока,,, а также интеграционными процессами в Европе.

Следующие же страны, наоборот, внесут наименьший вклад в итоговую модель. Посмотрим на них:

```
# описательные статистики
panel[panel$country=='Czech Republic', ]

## # A tibble: 5 x 4
##   country        period fh_polity state_capacity
##   <chr>           <dbl>     <dbl>          <dbl>
## 1 Czech Republic      1      7.48           7.45
## 2 Czech Republic      2      9.58           7.21
## 3 Czech Republic      3      9.58           7.51
## 4 Czech Republic      4      9.7            7.08
## 5 Czech Republic      5      9.5            7.27

panel[panel$country=='Turkmenistan', ]

## # A tibble: 5 x 4
##   country      period fh_polity state_capacity
##   <chr>         <dbl>     <dbl>          <dbl>
## 1 Turkmenistan      1      1.33           3.64
## 2 Turkmenistan      2      0.25           3.40
## 3 Turkmenistan      3      0.25           3.76
## 4 Turkmenistan      4      0.25           3.72
## 5 Turkmenistan      5      0.25           3.78

panel[panel$country=='Moldova', ]

## # A tibble: 5 x 4
##   country period fh_polity state_capacity
##   <chr>    <dbl>     <dbl>          <dbl>
## 1 Moldova      1      5.67           3.86
## 2 Moldova      2      7.08           3.93
## 3 Moldova      3      7.57           3.98
## 4 Moldova      4      7.33           4.22
## 5 Moldova      5      7.62           4.04
```

Действительно, можно заметить, что изменение переменной *state_capacity* у них минимально (в то время как *fh_polity иногда меняется*, причем довольно сильно — в дальнейшем нам пригодится это наблюдение).

Далее было бы интересно посмотреть на **“чистые„ значения** $\hat{a}_{1_{\{i\}}}$, которые получаются при оценивании отдельных моделей на *каждую пространственную единицу* i:



Здесь значения $\hat{a}_{1_{\{i\}}}$ отсортированы по убыванию. Можно предположить, что они соответствуют связи предиктора “государственная состоятельность„ и отклика — индекса демократии политического режима в отдельных пространственных единицах.

По всей видимости, наибольшая связь между этими двумя показателями прослеживается в таких странах, как *Словакия, Молдавия, Венгрия, Румыния* и др. Однако мы **не видели** их в топе предыдущего графика, посвященного условной вариации, а, значит, *мы не можем экстраполировать эффект в данных пространственных единицах на все страны в нашей LSDV-модели*. Точно так же, как и эффект стран с сильной отрицательной оценкой коэффициента: *Белоруссии, Киргизии, Чехии*. Возможно, условная вариация для них была столь мала, что даже небольшие колебания в переменных создали такие большие оценки по абсолютному значению. Было бы ошибочно брать их с одинаковым весом или тем более давать больший вес большим значениям. Как раз немного ранее мы уже смотрели на данные по *Чехии* и убедились в этом.

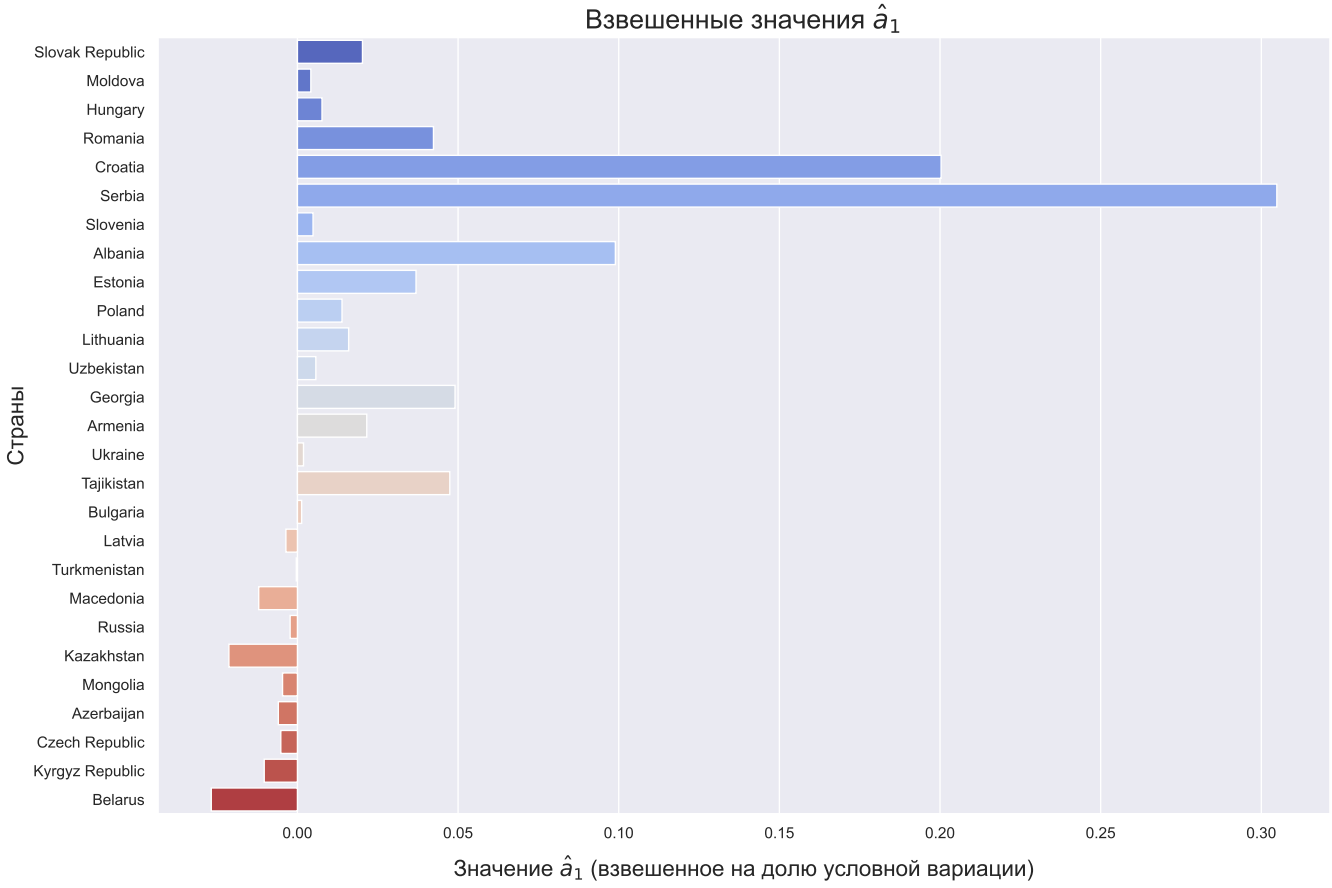
Можем еще раз проверить это, посмотрев на данные:

```
# выведем несколько примеров
panel[panel$country=='Slovak Republic', ]

## # A tibble: 5 x 4
##   country         period fh_polity state_capacity
##   <chr>            <dbl>     <dbl>          <dbl>
## 1 Slovak Republic      1       7.5           6.50
```

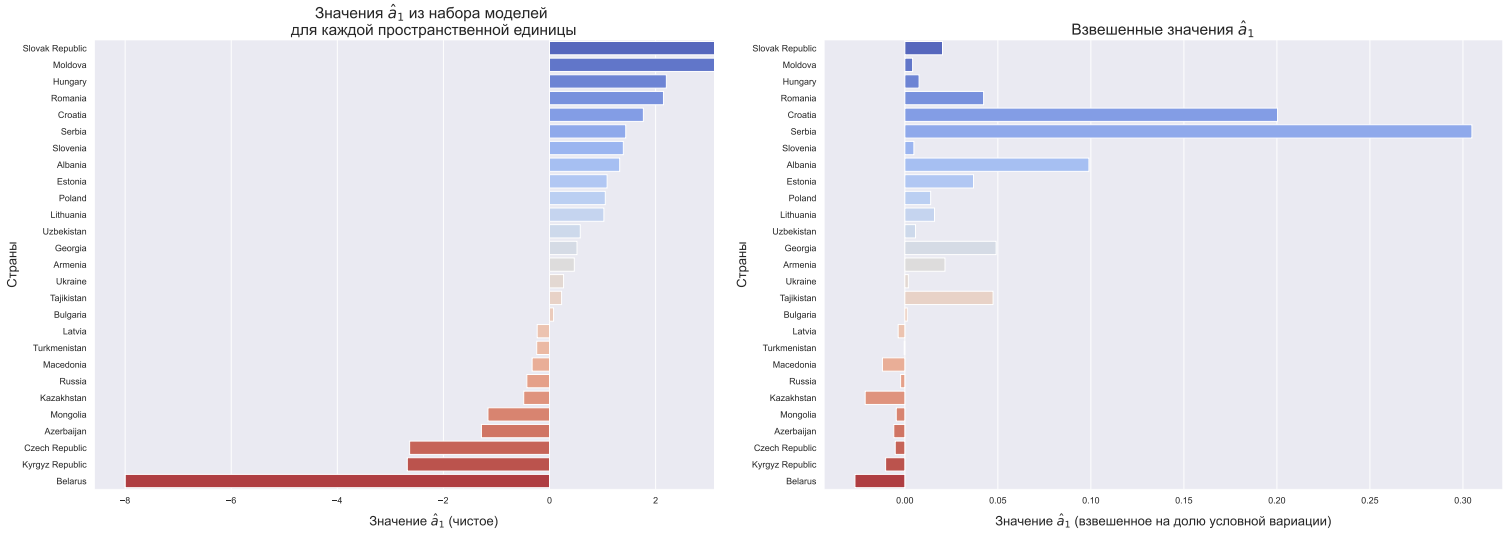
```
## 2 Slovak Republic      2      8.02           6.71
## 3 Slovak Republic      3      9.33           6.91
## 4 Slovak Republic      4      9.9            7.06
## 5 Slovak Republic      5     10              7.27

panel[panel$country=='Moldova', ]

## # A tibble: 5 x 4
##   country period fh_polity state_capacity
##   <chr>    <dbl>     <dbl>          <dbl>
## 1 Moldova      1      5.67           3.86
## 2 Moldova      2      7.08           3.93
## 3 Moldova      3      7.57           3.98
## 4 Moldova      4      7.33           4.22
## 5 Moldova      5      7.62           4.04

panel[panel$country=='Belarus', ]

## # A tibble: 5 x 4
##   country period fh_polity state_capacity
##   <chr>    <dbl>     <dbl>          <dbl>
## 1 Belarus      1      6.42           4.33
## 2 Belarus      2      3.13           4.52
## 3 Belarus      3      1.58           4.56
## 4 Belarus      4      1.17           4.74
## 5 Belarus      5      1.17           4.94
```

Действительно, мы можем увидеть, что в этих странах, **во-первых**, изменение fh_polity превосходит изменение $state\_capacity$ (т.е. уровень демократии растет быстрее, чем уровень государственной состоятельности). **Во-вторых**, например, в Белоруссии, уровень государственной состоятельности практически не изменяется, в то время как уровень демократии снижается от выше среднего до практически нулевого значения. Значит мы действительно **не можем использовать оценки** коэффициента при предикторе для этих стран в чистом виде.

Далее выведем **взвешенные значения** $\hat{a}_{1_{\{i\}}}$, сохранив тот же порядок стран:



Теперь для удобства выведем **два графика рядом** для сравнения:



(a) Чистые значения

(b) Взвешенные значения

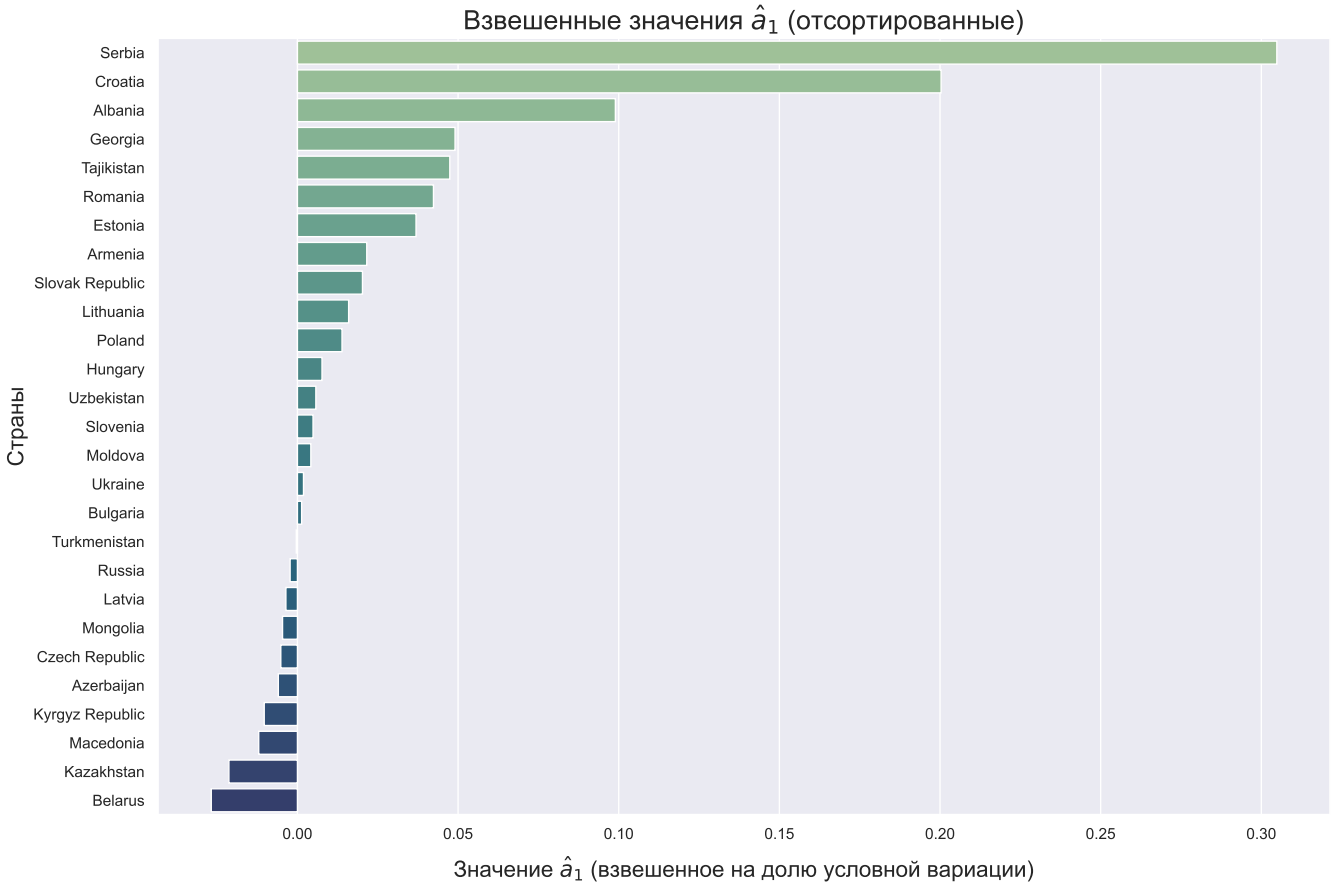
Рис. 1: Сравнение чистых и взвешенных оценок коэффициентов при предикторе

Итак, что мы видим: **вклад стран, в которых не наблюдалось сильного изменения предиктора, сильно уменьшился**: например, *Словакия и Белоруссия* как полярные примеры.

Кроме того, вклад стран, в которых практически не наблюдалась условная изменчивость показателя уровня государственной состоятельности, **минимален**.

Можно заметить, как вклад (казалось бы, не такой большой без взвешивания) стран с большой условной вариацией (*Сербия, Хорватия* и др.) ***“раздулся„*** по сравнению с изначальной чистой оценкой.

Таким образом, можно представить **“рейтинг„** стран по значению взвешенных оценок, т.е. вкладу в итоговую оценку $\hat{\beta}_1$:



На самом деле, по относительным значениям он **очень похож на самый первый график**, посвященный вкладу пространственных единиц в общую оценку. Поэтому просто доли условной вариации уже во многом определяют дальнейший вклад единиц по конкретным значениям оценок.

Задание 3.
Ранее уже были даны некоторые предположения относительно того, с чем содержательно могут быть связаны такие результаты. Так, часто изначальные (чистые) значения оценок могут **ввести нас в некоторое заблуждение**, если мы будем использовать их напрямую. Ведь, когда мы оцениваем модель, она видит значения только по одной пространственной единице. И может возникнуть такая ситуация, что *ключевой предиктор изменяется совсем немного, а отклик растет очень быстро*. Логично, что в такой ситуации (как например было со *Словакией*) оценка коэффициента при предикторе будет очень большой, но такой результат **нельзя проецировать** на все наблюдения в равной силе, т.к. в данной ситуации условная вариация предиктора очень низкая, по сравнению с остальными странами. Возможно, такое сильное изменение отклика было вызвано другими факторами — тогда мы бы столкнулись с **проблемой эндогенности**, если бы взяли оценку в чистом виде.

Следовательно, мы можем попытаться выявить, в каких странах изменение отклика действительно связано с интересующим нас предиктором, а в каких — с другими факторами. Поэтому весьма логично, что мы берем результаты стран, где изменение отклика можно связать с изменением предиктора, с большим весом, т.к. **мы можем доверять** этим значениям и распространить их на всю совокупность территориальных единиц.

Таким образом, мы можем утверждать, что знание о полученных весах на практике дает нам информацию о том, в каких единицах у нас прослеживается большая условная вариация интересующих нас предикторов. *Это знание будет очень полезным при интерпретации результатов.*

Кроме того, на самом деле, мы можем увидеть **направление и силу связи** в каждой из подгрупп, которые мы выделяем. Можно заметить, где наши теоретические предпосылки и гипотезы работают (и насколько хорошо), а где — не очень. Так, мы можем заметить, что в нашем примере в Белоруссии, даже с учетом взвешивания, получилась относительно большая оценка коэффициента при предикторе. Это свидетельствует о том, что там наша гипотеза не работает, несмотря на общий положительный результат для большинства стран ($\hat{\beta}_1$ положительная). Развивая предложенный дизайн исследования, можно было бы разделить все имеющиеся страны на несколько *подвыборок* в зависимости от уровня связи ключевого предиктора и отклика и оценить модель, предполагающую **разную взаимосвязь** между уровнем государственной состоятельности и уровнем демократии.

Кстати, об общей оценке $\hat{\beta}_1$. Ведь, суммировав полученные взвешенные значения $\hat{a}_{1_{\{i\}}}$, мы можем получить интересующую нас общую оценку, которая будет идентичной оценке из LSDV-модели:

```
# оценим LSDV-модель и сравним коэф-ты
LSDV_country <- lm(fh_polity~state_capacity+country, data = panel)
summary(LSDV_country)

##
## Call:
## lm(formula = fh_polity ~ state_capacity + country, data = panel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2753 -0.4968  0.1919  0.5423  3.9501
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.7102295  0.8322841   3.256 0.001512 **
## state_capacity         0.7845941  0.1343982   5.838 5.74e-08 ***
## countryArmenia         0.0003584  0.6999010   0.001 0.999592
## countryAzerbaijan     -2.8211369  0.7265017  -3.883 0.000179 ***
## countryBelarus        -3.6416852  0.6855667  -5.312 5.95e-07 ***
## countryBulgaria        1.4102365  0.6856351   2.057 0.042135 *
```

```
## countryCroatia         -1.2813502  0.7262585  -1.764 0.080532 .
## countryCzech Republic   0.7278715  0.7478433   0.973 0.332602
## countryEstonia          0.9959629  0.7095131   1.404 0.163296
## countryGeorgia          1.3685292  0.7465533   1.833 0.069563 .
## countryHungary          1.3291870  0.7347050   1.809 0.073237 .
## countryKazakhstan      -3.4400550  0.6898298  -4.987 2.38e-06 ***
## countryKyrgyz Republic -1.2676060  0.7121942  -1.780 0.077938 .
## countryLatvia           1.7385071  0.6844088   2.540 0.012516 *
## countryLithuania        2.1021607  0.6912702   3.041 0.002965 **
## countryMacedonia       -0.0471028  0.6954089  -0.068 0.946124
## countryMoldova          1.2013274  0.6972578   1.723 0.087790 .
## countryMongolia         1.8325965  0.6835142   2.681 0.008500 **
## countryPoland           1.4646948  0.7143880   2.050 0.042782 *
## countryRomania          0.6633450  0.6886991   0.963 0.337626
## countryRussia          -0.4983025  0.6928297  -0.719 0.473568
## countrySerbia          -0.3827508  0.6941581  -0.551 0.582517
## countrySlovak Republic  0.8342225  0.7269053   1.148 0.253678
## countrySlovenia        -0.0353723  0.8449212  -0.042 0.966685
## countryTajikistan      -1.8143519  0.7798914  -2.326 0.021880 *
## countryTurkmenistan    -5.1161202  0.7079871  -7.226 7.69e-11 ***
## countryUkraine          1.1699991  0.6940497   1.686 0.094757 .
## countryUzbekistan      -4.3890763  0.7400814  -5.931 3.76e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.08 on 107 degrees of freedom
## Multiple R-squared:  0.8961,Adjusted R-squared:  0.8699
## F-statistic: 34.18 on 27 and 107 DF,  p-value: < 2.2e-16
```

```
sum(m$coef)
```

```
## [1] 0.7845941
```

Действительно, они схожи.

Кроме того, мы можем **увидеть, на какие страны модель прежде всего ориентируется при построении результата**, т.е. какие единицы "оттягивают,, итоговый результат. Получив подобную информацию, мы можем либо продолжить работу, либо, если мы увидим, что некоторая единица "сдавливает,, оценки других пространственных единиц. В таком случае нам стоит задуматься: возможно, необходимо добавить новых данных или пересмотреть имеющиеся на предмет ошибок. Возможно, такая ситуация является проявлением своего рода проблемы **"выбросов,,**.