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

Домашнее задание 1
Рубанов Владислав, БПТ 201

Задание 1

Перед началом непосредственно регрессионного анализа, хотелось бы **ознакомиться с данными**, с которыми мы будем работать. Прежде всего, нужно держать в голове, что мы работаем с логарифмами (натуральными), т.е. мы будем предсказывать степень числа e, особенно это актуально для интерпретации.

Кроме того, посмотрим на **описательные статистики и распределение** данных:

```
    library(haven)
    library(plm)
    library(ggplot2)
    library(dplyr)
    library(lmtest)
    library(sandwich)
    library(psych)
    library(GGally)

    panel<-read_dta("RAPDC_hw1.dta") # загрузим данные

    # описательные статистики
    summary(panel)
```

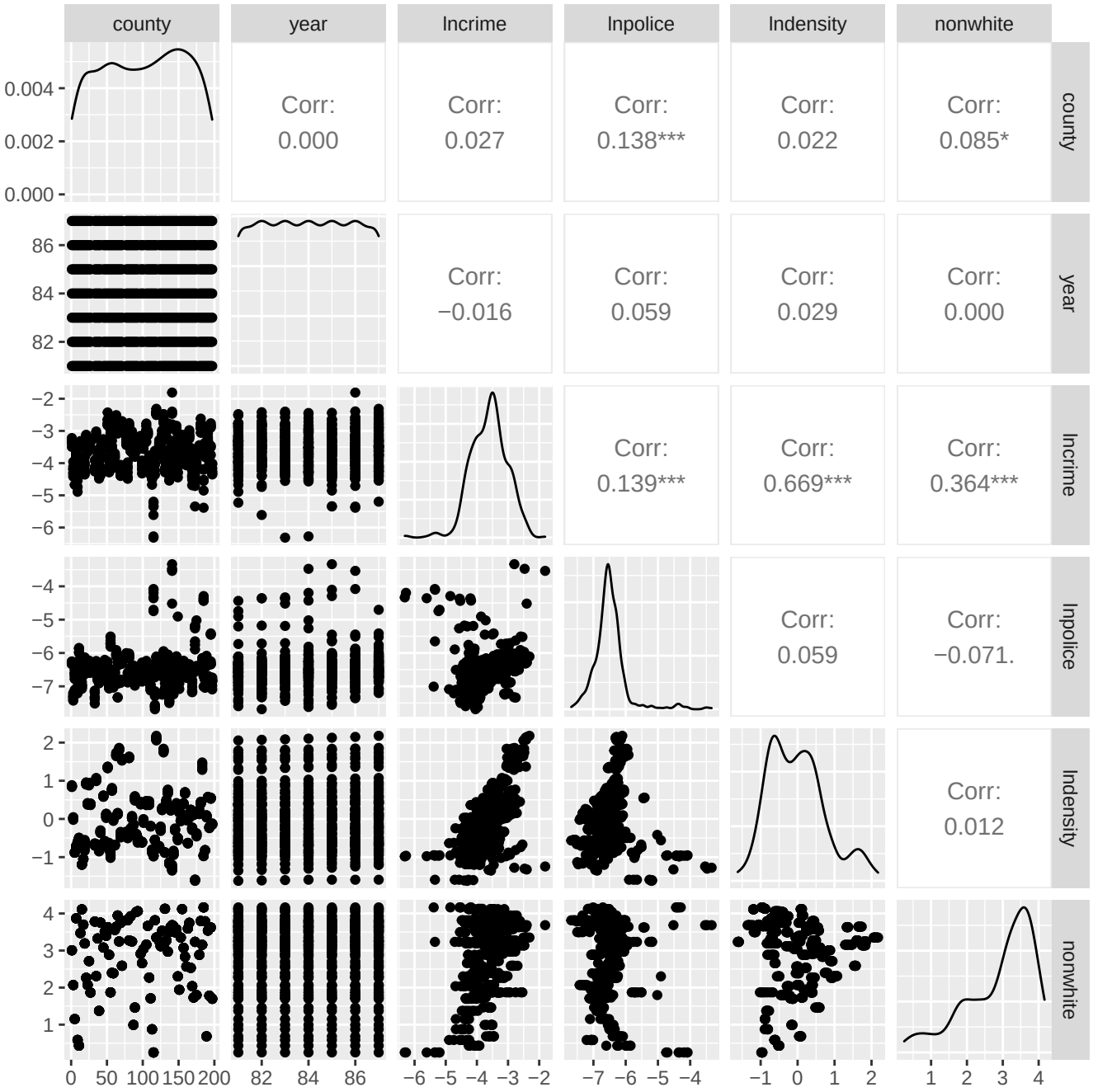
```
##      county           year       lncrime          lnpolice
##  Min.   :  1.0   Min.   :81   Min.   :-6.314   Min.   :-7.688
##  1st Qu.: 51.0   1st Qu.:82   1st Qu.:-3.998   1st Qu.:-6.733
##  Median :103.0   Median :84   Median :-3.560   Median :-6.536
##  Mean   :100.6   Mean   :84   Mean   :-3.609   Mean   :-6.491
##  3rd Qu.:151.0   3rd Qu.:86   3rd Qu.:-3.260   3rd Qu.:-6.318
##  Max.   :197.0   Max.   :87   Max.   :-1.809   Max.   :-3.336
##    lndensity           nonwhite
##  Min.   :-1.62091   Min.   :0.2497
##  1st Qu.:-0.62934   1st Qu.:2.3030
##  Median :-0.04857   Median :3.2127
##  Mean   :-0.01593   Mean   :2.9134
##  3rd Qu.: 0.41066   3rd Qu.:3.6434
##  Max.   : 2.17789   Max.   :4.1643
```

```
    describe(panel)
```

```
##           vars   n   mean    sd median trimmed   mad   min    max  range  skew
## county       1 630 100.60 58.04 103.00  101.00 74.13  1.00 197.00 196.00 -0.06
## year         2 630  84.00  2.00  84.00   84.00  2.97 81.00  87.00   6.00  0.00
## lncrime      3 630  -3.61  0.57  -3.56   -3.60  0.57 -6.31  -1.81   4.50 -0.43
## lnpolice     4 630  -6.49  0.53  -6.54   -6.54  0.31 -7.69  -3.34   4.35  2.22
## lndensity    5 630  -0.02  0.77  -0.05   -0.08  0.80 -1.62   2.18   3.80  0.62
```

```
## nonwhite     6 630   2.91  0.95   3.21    3.03  0.81  0.25   4.16   3.91 -0.96
##           kurtosis   se
## county       -1.23 2.31
## year         -1.26 0.08
## lncrime       1.28 0.02
## lnpolice      9.15 0.02
## lndensity     0.03 0.03
## nonwhite      0.11 0.04
```

```
    # распределение и корреляции
    ggpairs(panel)
```



Можно заметить, что логарифм плотности населения очень **сильно связан** с откликом — логарифмом числа преступлений на человека (**корреляция** равна 0.694). Другие два предиктора имеют более умеренную корреляцию: 0.185 и 0.169 для логарифма числа полицейских на д.н. и логарифма небелого населения соответственно.

Что касается *описательных статистик*: логично, что само по себе число полицейских на душу населения будет очень маленьким: в среднем около 0.0015 на 1 д.н., также как и значение числа преступлений на 1 человека — примерно 0.027.

“Плотность населения„ измеряется в числе людей на квадратную милю — среднее равно 0.98 и **“небелое население„** измеряется в доле небелого населения — 18.36 в среднем.

Также важно отметить, что распределение логарифма числа полицейских на д.н. **скошено влево**, а лог. числа небелого населения — **вправо**.

Оценим регрессионную модель без учета панельной структуры данных (**pooled model**).

В ней ***откликом*** будет выступать логарифм числа преступлений на человека, а ***предикторами*** — логарифм числа полицейских на душу населения, логарифм плотности населения и логарифм показателя небелого населения.

Кроме того, проведем **центрирование** всех предикторов, чтобы отойти от неправдоподобных значений, таких как “при одном полицейском на душу населения„ (т.к. мы работаем с логарифмами, а $e^0 = 1$):

```
    pooled <- plm(lncrime~I(lnpolice-mean(lnpolice)) +
                  I(lndensity-mean(lndensity)) +
                  I(nonwhite-mean(nonwhite)), data=panel, model="pooling")
    summary(pooled)
```

```
## Pooling Model
##
## Call:
## plm(formula = lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity -
##     mean(lndensity)) + I(nonwhite - mean(nonwhite)), data = panel,
##     model = "pooling")
##
## Balanced Panel: n = 90, T = 7, N = 630
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.9480749 -0.1873083  0.0071238  0.1910687  1.8208435
##
## Coefficients:
##                                Estimate Std. Error   t-value  Pr(>|t|)
## (Intercept)                   -3.609225   0.014641 -246.5096 < 2.2e-16 ***
## I(lnpolice - mean(lnpolice))   0.137782   0.027944    4.9306 1.051e-06 ***
## I(lndensity - mean(lndensity)) 0.485777   0.018949   25.6359 < 2.2e-16 ***
## I(nonwhite - mean(nonwhite))   0.219439   0.015391   14.2579 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    206.38
## Residual Sum of Squares: 84.542
## R-Squared:      0.59036
## Adj. R-Squared: 0.58839
## F-statistic: 300.719 on 3 and 626 DF, p-value: < 2.22e-16
```

Задание 1.1

Проинтерпретируем полученные результаты, в начале в классических терминах:

- $\hat{\beta}_0 = -3.61$ — в среднем значение натурального логарифма числа преступлений на человека равно -3.61 (или $e^{-3.61} = 0.027$) при условии равенства натуральных логарифмов числа полицейских, плотности населения и небелого населения своим средним значениям (которые обозначены выше);

- $\hat{\beta}_1 = 0.14$ — с увеличением порядка натурального логарифма числа полицейских на д.н. на 1 значение (во временной перспективе), порядок натурального логарифма числа преступлений на человека увеличивается в среднем на 0.14 при прочих равных условиях;

- $\hat{\beta}_2 = 0.49$ — с увеличением порядка натурального логарифма значения плотности населения на 1 значение (во временной перспективе), порядок натурального логарифма числа преступлений на человека увеличивается в среднем на 0.49 при прочих равных условиях;

- $\hat{\beta}_3 = 0.22$ — с увеличением порядка натурального логарифма значения небелого населения на 1 значение (во временной перспективе), порядок натурального логарифма числа преступлений на человека увеличивается в среднем на 0.22 при прочих равных условиях.

Итак, можно отметить, что *интерпретация несколько затруднена*, поскольку мы вынуждены работать не с чистыми значениями, а с их логарифмами — нелинейной функцией.

Однако мы можем сравнивать полученные значения между собой по величине, поскольку они все находятся в одной шкале — натуральных логарифмах.

Далее можно воспользоваться интерпретацией, предложенной **Д. Гуджарати**[1]:

- $\hat{\beta}_1 = 0.14$ — с увеличением значения числа полицейских на д.н. на 1 **процент** (во временной перспективе), значение числа преступлений на человека увеличивается в среднем на 0.14 процентов при прочих равных условиях;

- $\hat{\beta}_2 = 0.49$ — с увеличением значения плотности населения на 1 **процент** (во временной перспективе), значение числа преступлений на человека увеличивается в среднем на 0.49 процента прочих равных условиях;

- $\hat{\beta}_3 = 0.22$ — с увеличением значения небелого населения на 1 **процент** (во временной перспективе), значение числа преступлений на человека увеличивается в среднем на 0.22 процента при прочих равных условиях.

Итак, можно заметить, что у всех предикторов наблюдается **положительная и значимая связь** с откликом — натуральным логарифмом числа преступлений на д.н. Самым "влиятельным,, из предикторов является натуральный логарифм значения **плотности населения**. Это весьма логично, и это было заметно еще на разведывательном этапе — у данного предиктора наблюдается большая взаимосвязь с откликом. К тому же, это объяснимо на содержательном уровне.

Также все оценки **статистически значимы** на любом уровне значимости. Но, возможно, их значимость *завышена*, поскольку мы оценили pooled-модель, и у нас очень много (на самом деле не независимых) наблюдений: 360.

Задание 1.2

Таким образом, оценивание модели объединенной регрессии (pooled model) на массиве панельных данных **может привести к двум большим проблемам**:

1. **Aggregation bias**, поскольку наши наблюдения уже не являются независимыми. На самом деле, это разные подвыборки, в которых могут наблюдаться разные взаимосвязи как внутри групп, так и между группами, в то время как мы даем им всем одинаковые веса и усредняем эффект по всей выборке, создавая смещение и сталкиваясь с эндогенностью.

[1] Gujarathi, D. M. (2004). Gujarati: Basic Econometrics. McGraw-hill, p. 175.

2. **Значимые (ложно) оценки** из-за маленьких значений стандартных ошибок оценок. Причиной этого являются завышенные t-статистики, т.к., когда мы объединяем панель в одну большую выборку, у нас очень много наблюдений, что приводит к уменьшению дисперсии оценок. На самом деле, некоторые из предикторов могут оказаться статистически незначимыми при рассмотрении их на страновом или временном уровнях после процедуры "взвешивания„.

Задание 2

Оценим модель посредством МНК с фиктивными переменными (**LSDV-модель** с набором дамми-переменных на константы).

```
    LSDV <- lm(lncrime~I(lnpolice-mean(lnpolice)) +
             I(lndensity-mean(lndensity)) +
             I(nonwhite-mean(nonwhite)) +
             factor(county), data=panel)
    summary(LSDV)

##
## Call:
## lm(formula = lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity -
##     mean(lndensity)) + I(nonwhite - mean(nonwhite)) + factor(county),
##     data = panel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77068 -0.08539 -0.00100  0.08589  0.74125
##
## Coefficients: (1 not defined because of singularities)
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -3.384844   0.196779 -17.201  < 2e-16 ***
## I(lnpolice - mean(lnpolice))    0.213866   0.027961   7.649 9.42e-14 ***
## I(lndensity - mean(lndensity)) -0.055861   0.233589  -0.239 0.811088
## I(nonwhite - mean(nonwhite))    0.612488   0.190111   3.222 0.001351 **
## factor(county)3                -0.132301   0.089408  -1.480 0.139527
## factor(county)5                 0.077613   0.141270   0.549 0.582963
## factor(county)7                -1.010351   0.535465  -1.887 0.059716 .
## factor(county)9                 0.401178   0.171441   2.340 0.019647 *
## factor(county)11                0.725701   0.210432   3.449 0.000608 ***
## factor(county)13               -0.393039   0.451991  -0.870 0.384921
## factor(county)15               -1.357166   0.688585  -1.971 0.049242 *
## factor(county)17               -1.006209   0.580308  -1.734 0.083504 .
## factor(county)19               -0.951455   0.406282  -2.342 0.019552 *
## factor(county)21                0.561621   0.168008   3.343 0.000887 ***
## factor(county)23                0.379112   0.100936   3.756 0.000192 ***
## factor(county)25                0.100227   0.106732   0.939 0.348124
## factor(county)27                0.696758   0.126430   5.511 5.54e-08 ***
## factor(county)33               -1.158712   0.507404  -2.284 0.022783 *
## factor(county)35                0.624292   0.181985   3.430 0.000649 ***
## factor(county)37               -0.705604   0.432614  -1.631 0.103471
## factor(county)39                0.108436   0.147450   0.735 0.462412
## factor(county)41               -0.954157   0.429576  -2.221 0.026755 *
```

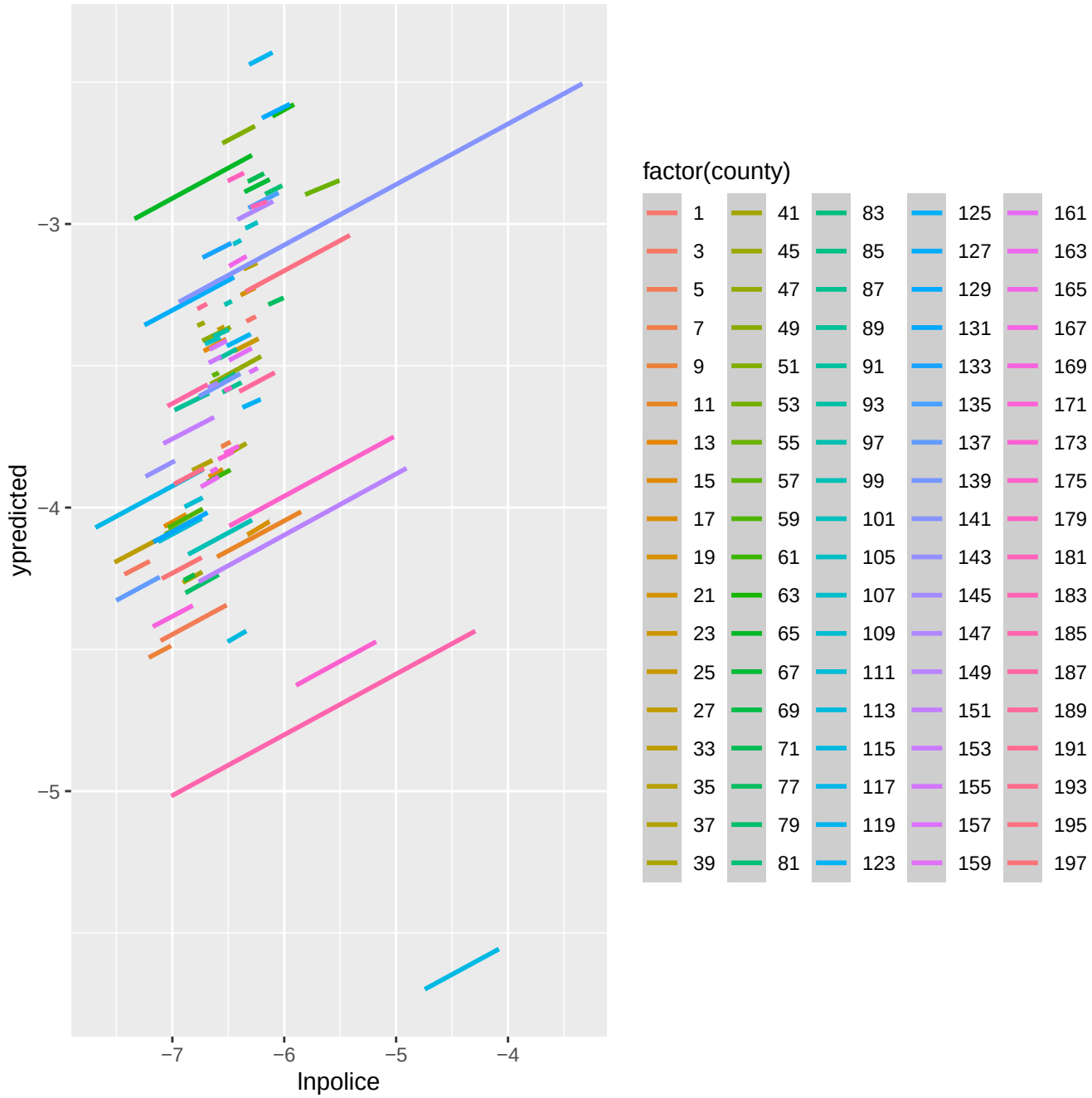
```
## factor(county)45                0.018587   0.120601   0.154 0.877575
## factor(county)47               -0.539785   0.441721  -1.222 0.222241
## factor(county)49               -0.270841   0.271077  -0.999 0.318182
## factor(county)51                0.318586   0.121983   2.612 0.009260 **
## factor(county)53               -0.609346   0.354237  -1.720 0.085978 .
## factor(county)55                0.926725   0.224050   4.136 4.10e-05 ***
## factor(county)57                0.240500   0.118371   2.032 0.042671 *
## factor(county)59               -0.245169   0.120681  -2.032 0.042691 *
## factor(county)61               -0.914416   0.470735  -1.943 0.052595 .
## factor(county)63                0.330095   0.145679   2.266 0.023854 *
## factor(county)65               -0.039508   0.361045  -0.109 0.912906
## factor(county)67                0.366484   0.210440   1.742 0.082165 .
## factor(county)69               -1.317292   0.447097  -2.946 0.003355 **
## factor(county)71                0.784126   0.250347   3.132 0.001830 **
## factor(county)77               -0.551026   0.452859  -1.217 0.224224
## factor(county)79               -1.411509   0.473853  -2.979 0.003024 **
## factor(county)81                0.290808   0.164830   1.764 0.078249 .
## factor(county)83               -0.786706   0.443688  -1.773 0.076777 .
## factor(county)85                0.005922   0.253943   0.023 0.981404
## factor(county)87                0.972324   0.173774   5.595 3.51e-08 ***
## factor(county)89                0.749816   0.224954   3.333 0.000918 ***
## factor(county)91               -0.781405   0.497592  -1.570 0.116916
## factor(county)93               -0.717122   0.536820  -1.336 0.182157
## factor(county)97                0.132811   0.125322   1.060 0.289729
## factor(county)99               -0.579721   0.292790  -1.980 0.048214 *
## factor(county)101              -0.075818   0.238095  -0.318 0.750276
## factor(county)105               0.211745   0.162410   1.304 0.192869
## factor(county)107              -0.130900   0.252917  -0.518 0.604979
## factor(county)109              -0.111057   0.084821  -1.309 0.190988
## factor(county)111               0.127023   0.093552   1.358 0.175105
## factor(county)113               0.118625   0.135213   0.877 0.380702
## factor(county)115              -1.108905   0.198255  -5.593 3.55e-08 ***
## factor(county)117              -1.013097   0.488328  -2.075 0.038496 *
## factor(county)119               0.758978   0.255527   2.970 0.003108 **
## factor(county)123              -0.293005   0.428479  -0.684 0.494380
## factor(county)125              -0.429033   0.296826  -1.445 0.148926
## factor(county)127              -0.175599   0.259366  -0.677 0.498675
## factor(county)129               0.657269   0.217333   3.024 0.002612 **
## factor(county)131              -1.382542   0.620295  -2.229 0.026237 *
## factor(county)133               0.121696   0.174813   0.696 0.486634
## factor(county)135               0.406375   0.096256   4.222 2.85e-05 ***
## factor(county)137              -1.148504   0.566537  -2.027 0.043131 *
## factor(county)139              -0.594295   0.281222  -2.113 0.035039 *
## factor(county)141              -0.338539   0.637610  -0.531 0.595673
## factor(county)143              -0.849199   0.536445  -1.583 0.114006
## factor(county)145              -0.524233   0.368844  -1.421 0.155812
## factor(county)147              -0.000202   0.244101  -0.001 0.999340
## factor(county)149              -0.471743   0.212663  -2.218 0.026953 *
## factor(county)151               0.340649   0.092295   3.691 0.000246 ***
## factor(county)153              -0.336033   0.290965  -1.155 0.248647
## factor(county)155              -0.752557   0.400422  -1.879 0.060729 .
```

```
## factor(county)157              -0.251427   0.151571  -1.659 0.097738 .
## factor(county)159              -0.007493   0.087930  -0.085 0.932121
## factor(county)161              -0.274935   0.149007  -1.845 0.065572 .
## factor(county)163              -0.904759   0.466346  -1.940 0.052890 .
## factor(county)165              -0.269919   0.343684  -0.785 0.432582
## factor(county)167              -0.168460   0.103309  -1.631 0.103553
## factor(county)169              -0.366547   0.121440  -3.018 0.002662 **
## factor(county)171               0.307553   0.106031   2.901 0.003877 **
## factor(county)173              -1.654159   0.625361  -2.645 0.008405 **
## factor(county)175              -0.018718   0.124048  -0.151 0.880116
## factor(county)179              -0.165416   0.161860  -1.022 0.307257
## factor(county)181               0.020497   0.273988   0.075 0.940393
## factor(county)183               0.327500   0.137444   2.383 0.017529 *
## factor(county)185              -2.339980   0.650021  -3.600 0.000348 ***
## factor(county)187              -0.720971   0.544581  -1.324 0.186098
## factor(county)189               1.144283   0.268570   4.261 2.41e-05 ***
## factor(county)191              -0.204224   0.198163  -1.031 0.303198
## factor(county)193               0.255073   0.096439   2.645 0.008410 **
## factor(county)195              -0.288172   0.219184  -1.315 0.189154
## factor(county)197                    NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1736 on 538 degrees of freedom
## Multiple R-squared:  0.9214,Adjusted R-squared:  0.9081
## F-statistic: 69.32 on 91 and 538 DF,  p-value: < 2.2e-16
```

Можно также **визуализировать** результат на примере одного из предикторов:

```
# визуализация
panel$ypredicted <- LSDV$fitted

ggplot(panel, aes(x = lnpolice, y = ypredicted, color = factor(county)))+geom_smooth(metho
```



Видно, что график получился *крайне неинтерпретируемым*, однако на нем заметно **предположение об одинаковой связи** каждого из предикторов (конкретно здесь — логарифма числа полицейских на д.н.) и отклика во всех пространственных единицах. Кроме того, выдача также получилась **громоздкой** из-за 90 пространственных единиц.

Сравним полученные результаты с результатами модели, построенной без учета панельной структуры данных.

- Прежде всего, заметно, что один из коэффициентов при предикторе — натуральный лог. плотности населения **стал очень сильно незначимым**. Этому можно найти объяснение: ведь во многом именно он олицетворял ту изменчивость между разными пространственными единицами, которую учла (“съела„) LSDV-модель через большое число дамми. Ведь логично, что плотность населения — одна из характеристик, которая по большей части *отличается от округа к округу* (если они только не нарезаны “по линейке„, как в Африке или в Австралии, т.е. если между ними есть географические и демографические различия) и *не изменяется во времени*. Поэтому спецификация с дамми-переменными на штаты учла эту изменчивость, сделав коэффициент незначимым и, по сути, бесполезным. Кроме того, нельзя забывать про то, что теперь мы рабо-

таем с большим числом подвыборок. Вероятно, это также могло быть проявлением aggregation bias.

- Кроме того, упала значимость коэффициента при предикторе — натуральный лог. небелого населения, хотя оценка и осталась значимой на уровне значимости 0.01. Это также логично, ведь уровень небелого населения, вероятно, тоже зависит от того или иного округа внутри и почти не изменяется во времени, однако не настолько, как плотность населения. Кроме того, именно эта переменная могла подвергнуться завышению значимости при оценки объединенной модели.
  При этом само значение оценки коэф-та при предикторе натуральный лог. небелого населения увеличилось в три раза (до 0.61), что свидетельствует о смещенных результатах pooled-модели. Теперь именно этот предиктор вносит **наибольший вклад** в значение отклика.

- Коэф-т при предикторе — натуральный лог. числа полицейских на д.н. несколько увеличился (до 0.21), но не сильно.

- Также важно, что все значимые коэф-ты при предикторах остались **положительными**, как и в объединенной модели.

Для примера **проинтерпретируем** оценки коэффициентов при нескольких статистически-значимых дамми-переменных:

- $\hat{\gamma}_{21} = 0.56$ — среднее значение натурального логарифма числа преступлений на человека в 21 округе штата Северная Каролина примерно на 0.56 больше, чем среднее значение натурального логарифма числа преступлений на человека в 1 округе штата Северная Каролина (базовой категории) при условии равенства всех предикторов среднему значению, или в целом составляет $e^{-3.38+0.56} = 0.060$.

- $\hat{\gamma}_{115} = -1.11$ — среднее значение натурального логарифма числа преступлений на человека в 115 округе штата Северная Каролина примерно на 1.11 меньше, чем среднее значение натурального логарифма числа преступлений на человека в 1 округе штата Северная Каролина (базовой категории) при условии равенстве всех предикторов среднему значению, или в целом составляет $e^{-3.38-1.11} = 0.011$.

Наконец, подумаем над тем, каковы **недостатки указанной спецификации** модели.

1. Модель явно **перегружена**. Большая часть дамми-переменных незначимы, поскольку, вероятно, деление на округа как на пространственные единицы, слишком мало, чтобы получить значимые различия. Это было бы более актуально для штатов. Кроме того, Мы в принципе имеем очень большую и слабо интерпретируемую выдачу.

2. В спецификации модели мы имплицитно задаем **одинаковый характер взаимосвязи** для всех пространственных единиц. Возможно, это не так, особенно для такого большого числа единиц (90 округов).

3. Как одно из следствий громоздкой спецификации — мы можем иметь дело со **“вздутием„ коэффициента детерминации** R^2, который у нас составляет очень большое значение: примерно 0.92.

4. Модели не удалось получить оценки для округа 197. Это произошло из-за совпадения данных с каким-то из других округов. Вероятно, при фиксировании временных эффектов, такого бы не случилось.

Задание 3

Теперь оценим модель, используя **внутригрупповое преобразование** (within-group transformation), чтобы избавиться от столь громоздкой выдачи:

```
    fe <- plm(lncrime~I(lnpolice-mean(lnpolice)) +
                I(lndensity-mean(lndensity)) +
                I(nonwhite-mean(nonwhite)), data = panel, index=c("county", "year"), effect =
    summary(fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity -
##     mean(lndensity)) + I(nonwhite - mean(nonwhite)), data = panel,
##     effect = "individual", model = "within", index = c("county",
##         "year"))
##
## Balanced Panel: n = 90, T = 7, N = 630
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -0.77068152 -0.08538992 -0.00099913  0.08588997  0.74125457
##
## Coefficients:
##                                 Estimate Std. Error t-value Pr(>|t|)
## I(lnpolice - mean(lnpolice))    0.213866   0.027961  7.6487 9.42e-14 ***
## I(lndensity - mean(lndensity)) -0.055860   0.233589 -0.2391   0.8111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    17.991
## Residual Sum of Squares: 16.218
## R-Squared:      0.098544
## Adj. R-Squared: -0.053933
## F-statistic: 29.406 on 2 and 538 DF, p-value: 7.5878e-13
```

Однако теперь мы видим лишь две оценки коэффицикента при предикторе из трех. **Коэффициент при предикторе "натуральный лог. небелого населения,, не позволяет полностью оценить модель с внутригрупповым преобразованием.**

Это логично, поскольку, данный показатель не изменяется во времени ни для одной пространственной единицы. Вероятно, он был взят из некоторой статистической базы, в которой в исследуемый период не проводилось переписи или другой процедуры для обновления демографических характеристик. Убедимся в этом:

```
    # небольшая выборка
    panel[panel$county==1, 6]
```

```
## # A tibble: 7 x 1
##   nonwhite
##      <dbl>
## 1     3.01
```

```
## 2     3.01
## 3     3.01
## 4     3.01
## 5     3.01
## 6     3.01
## 7     3.01
```

```
    panel[panel$county==17, 6]
```

```
## # A tibble: 7 x 1
##   nonwhite
##      <dbl>
## 1     3.70
## 2     3.70
## 3     3.70
## 4     3.70
## 5     3.70
## 6     3.70
## 7     3.70
```

```
    panel[panel$county==131, 6]
```

```
## # A tibble: 7 x 1
##   nonwhite
##      <dbl>
## 1     4.13
## 2     4.13
## 3     4.13
## 4     4.13
## 5     4.13
## 6     4.13
## 7     4.13
```

Действительно, можно увидеть, что **все значения являются константными внутри пространственных единиц**. А если мы вспомним, как считаются оценки коэффициентов при предикторах в модели с внутригрупповым преобразованием:

1. 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:

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

2. Далее мы забираем интересующие нас оценки коэффициентов при предикторе $\hat{a}_{1_{\{i\}}}$ и берем их со следующим весом:

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

   , то есть в качестве веса выступает доля условной вариации у той или иной пространственной единицы.

3. В результате искомая оценка $\hat{\beta}_1$ получается следющим образом:

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

Получается, что мы не смогли получить оценку коэф-та при предикторе "натуральный лог. небелого населения„, поскольку, с одной стороны, все оценки от каждой пространственной единицы должны были идти туда *с нулевым весом*, но, вообще говоря, *они просто не могли быть получены*, т.к. в каждой из моделей на подвыборке значения небелого населения представляли собой просто константные значения — в таком случае оценка для коэф-та при этом предикторе даже не может быть рассчитана, она как будто "живет своей жизнью„, по сравнению с откликом.

И это вовсе не недостаток модели с фиксированными эффектами, и даже не недостаток регрессионной модели в целом. Ведь весьма логично, что мы хотим брать с большим весом значения оценок на основе таких подвыборок, в которых прослеживается максимальное разнообразие данных по интересующей нас независимой переменной. И как раз хорошим критерием информативности оценок является **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). В некоторой степени мы смотрим на то, больше или меньше можно "доверять„ тем или иным пространственным (или временным) единицам. Так мы ориентируемся на **вклад** каждой пространственной единицы в общую дисперсию (информацию) независимой переменной. В том числе логично, что те пространственные единицы, у которых нет изменчивости, *не вносят 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 исключает небольшие (около-)случайные колебания.

Соответственно, нет ничего удивительного в том, что мы не смогли получить оценки для этого предиктора, ведь у нас просто **отсутствует изменение во временной перспективе** по нему при фиксировании территориальных единиц. Скорее это проблема самих данных и их устройства.

Задание 4

Тем не менее, мы видели, что у многих дамми-переменных на пространственные единицы были незначимые оценки. **Проверим гипотезу о том, что все индивидуальные эффекты равны нулю**, т.е. нужна ли нам вообще такая модель, которая учитывает панельную структуру, или можно оценить обычную регрессионную модель без поправок на неоднородность?

Для этого нам понадобится **F-test Фишера**. Вспомним его нулевую и альтернативную гипотезу для данного случая:

$$H_0 : \gamma_1 = ... = \gamma_{N-1} = 0$$
$$H_1 : \gamma_i \neq \gamma_j$$

, где γ — коэф-т при дамми-переменной для пространственных единиц.

Реализуем тест:

```
    #F-test
    pFtest(fe, pooled)

##
##  F test for individual effects
##
## data:  lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity - mean(lndensity)) +  ...
## F = 25.756, df1 = 88, df2 = 538, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Итак, мы видим, что значение $p - value$ очень маленькое. Это значит, что у нас **есть основания отвергнуть нулевую гипотезу** в пользу альтернативы на любом уровне значимости.

Таким образом, мы можем с большой уверенностью утверждать, что для наших данных **больше подходит модель с фиксированными эффектами**. Это логично, ведь, все-таки, мы могли видеть значимые коэф-ты при дамми-переменных для многих округов.

Задание 5

Оценим **модель со случайными эффектами**:

```
    re <- plm(lncrime~I(lnpolice-mean(lnpolice)) +
            I(lndensity-mean(lndensity)) +
            I(nonwhite-mean(nonwhite)), data=panel, index=c("county", "year"), model="random")
    summary(re)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity -
##     mean(lndensity)) + I(nonwhite - mean(nonwhite)), data = panel,
##     model = "random", index = c("county", "year"))
##
## Balanced Panel: n = 90, T = 7, N = 630
##
## Effects:
##                   var std.dev share
## idiosyncratic 0.03014 0.17362 0.217
## individual    0.10852 0.32942 0.783
## theta: 0.8046
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -9.6731e-01 -8.9767e-02 -2.6845e-05  9.4573e-02  9.1317e-01
##
## Coefficients:
##                                 Estimate Std. Error   z-value  Pr(>|z|)
## (Intercept)                    -3.609225   0.035529 -101.5848 < 2.2e-16 ***
## I(lnpolice - mean(lnpolice))    0.197241   0.026254    7.5128 5.787e-14 ***
## I(lndensity - mean(lndensity))  0.464491   0.045105   10.2980 < 2.2e-16 ***
## I(nonwhite - mean(nonwhite))    0.221979   0.037265    5.9567 2.573e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    25.181
## Residual Sum of Squares: 19.002
## R-Squared:      0.2454
## Adj. R-Squared: 0.24178
## Chisq: 203.575 on 3 DF, p-value: < 2.22e-16
```

Задание 5.1

Чтобы сделать выбор в пользу модели со случайными эффектами или модели без учета панельной

структуры данных, проверим гипотезу о том, что **дисперсия случайного индивидуального эффекта равна 0**:

$$H_0 : Var(\alpha_i) = 0$$
$$H_1 : Var(\alpha_i) > 0$$

Для этого реализуем **тест Бреуша-Пагана** на гомоскедастичность:

```
    #Тест Бреуша-Пагана
    plmtest(pooled, type=c("bp"))
```

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan)
##
## data:  lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity - mean(lndensity)) +  ...
## chisq = 1124.6, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Можно заметить очень маленькое значение $p - value$. Это позволяет нам **отвергнуть нулевую гипотезу** в пользу альтернативы на любом уровне значимости.

Таким образом, нам однозначно **лучше делать выбор в пользу модели со случайными эффектами**, по сравнению с pooled-моделью. Опять же, это весьма логично, учитывая структуру наших данных и предыдущие тезисы.

Задание 5.2

Далее **сравним результаты** модели со случайными эффектами с результатами модели с фиксированными эффектами.

- Оценка константы основным образом не отличается.
- Оценка коэф-та при предикторе логарифм числа полицейских на д.н. также не отличается.
- Оценка коэф-та при предикторе логарифм плотности населения вновь стала значимой и большой (логично, ведь больше нет набора дамми на простр. ед.).
- Оценка коэф-та при предикторе логарифм небелого населения уменьшилась в три раза и стала статистически более значимой .

Задумаемся о допущениях, предпосылках и ограничениях для использования модели со случайными эффектами:

1. прежде всего, для работы с моделью со случайными эффектами, наша выборка должна иметь **(квази-)экспериментальный дизайн** — как раз для учета межпространственной изменчивости (α_i), иначе мы столкнемся с **эндогенностью**. Выборка желательно должна избираться **случайно**.

   В нашем же случае мы работаем с довольно *специфической* выборкой — только по одному штату, хотя и с большим числом округов. Это не похоже на случайный отбор, явно прослеживается определенная специфика, сложно говорить о случайности.

2. Должно соблюдаться следующее **условие**: $Cor(\alpha_i, \varepsilon_{it}) = 0$.

   В нашем случае это скорее тоже не так. Ведь, как было подмечено ранее, многие различия можно было объяснить пространственными различиями, которые не изменяются во времени (географические, демографические факторы). Как только мы добавили дамми на такие фиксированные эффекты, некоторые предикторы сразу же потеряли в значимости. Скорее всего, в нашем случае межпространственная и внутрипространственная ошибки связаны, что является нарушением одного из основных допущений модели со случайными эффектами.

3. Из предыдущего тезиса можно выдвинуть предположение, что условие $Cor(\alpha_i, xit) = 0$ также может **нарушаться** из-за особенностей данных.

Попробуем посмотреть на данные, чтобы **проверить** эти предположения эмпирически.

Отсортируем выборку по одному из предикторов (плотности населения) и возьмем 10 округов с самым высоким значением данной переменной и 10 — с самым низким. Далее проделаем такую же сортировку, но по зависимой переменной — логарифму числа преступлений на человека и сравним, много ли среди них будет пересечений:

```
    # небольшая выборка
    # Отсортируем массивы
    dfsort1 <- panel[order(panel$lndensity), ]
    dfsort2 <- panel[order(panel$lndensity, decreasing = TRUE), ]

    # Выведем номера округов
    A = unique(head(dfsort1, 57)['county'])
    B = unique(head(dfsort2, 65)['county'])

    # Отсортируем массивы по y
    dfsort3 <- panel[order(panel$lncrime), ]
    dfsort4 <- panel[order(panel$lncrime, decreasing = TRUE), ]

    # Выведем номера округов
    C = unique(head(dfsort3, 30)['county'])
    D = unique(head(dfsort4, 35)['county'])

    intersect(A, C)
```

```
## # A tibble: 5 x 1
##   county
##    <dbl>
## 1    173
## 2    137
## 3    115
## 4    185
## 5    113
```

```
    intersect(B, D)
```

```
## # A tibble: 6 x 1
##   county
##    <dbl>
## 1    119
## 2    129
## 3     63
## 4     81
## 5     71
## 6     51
```

Мы вывели **пересечения множеств** и видим, что в первом случае (для минимальных значений) у нас **5 пересечений из 10**, т.е. 50%, а во втором (для максимальных) — **6 из 10**, т.е. 60%. Посмотрим на эти значения:

```
    # min
    A
```

```
## # A tibble: 10 x 1
##    county
##     <dbl>
##  1    173
##  2    141
##  3     15
##  4    137
##  5     17
##  6     55
##  7    115
##  8    185
##  9    143
## 10    113
```

```
    C
```

```
## # A tibble: 10 x 1
##    county
##     <dbl>
##  1    115
##  2    185
##  3    173
##  4      9
##  5    169
##  6      5
##  7    113
##  8    137
##  9     79
## 10     39
```

```
    # max
    B
```

```
## # A tibble: 10 x 1
##    county
##     <dbl>
##  1    119
##  2     67
##  3    129
##  4     63
##  5     81
##  6     71
##  7    183
##  8     51
##  9     35
## 10     21
```

```
    D
```

```
## # A tibble: 10 x 1
##    county
##     <dbl>
##  1    141
##  2    119
##  3     51
##  4    129
##  5     63
##  6     55
##  7    181
##  8     71
##  9     65
## 10     81
```

Из выдачи можно увидеть определенные **"кластеры,,** округов. Они определенно связаны, ведь они определенно расположены территориально рядом с друг другом. Таким образом, ***ни о какой случайности нельзя говорить***.
Вероятно, лучше выбрать **модель с фиксированными эффектами**, чтобы учесть эту неизменяющуюся во времени, но меняющуюся от одного округа к другому вариацию.

Задание 5.3

Далее протестируем посредством **теста Хаусмана** отсутствие корреляции между индивидуальными эффектами и предикторами.

Вспомним гипотезу, которую мы будем проверять:

$$H_0 : Cor(\alpha_i, x_{it}) = 0$$
$$H_1 : Cor(\alpha_i, x_{it}) \neq 0$$

Реализуем тест:

```
    phtest(fe, re)

##
##  Hausman Test
##
## data:  lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity - mean(lndensity)) +  ...
## chisq = 6.4278, df = 2, p-value = 0.0402
## alternative hypothesis: one model is inconsistent
```

Итак, мы видим довольно пограничное значение $p - value$. *Формально* у нас **есть основания отвергнуть нулевую гипотезу** ($0.04 < 0.05$) в пользу альтернативы, которая свидетельствует в пользу модели с фиксированными эффектами.

Однако стоит отметить, что тест Хаусмана имеет несколько **ограничений**, связанных с *корректной спецификацией модели* (которая неидеальна в нашем случае) и *большой выборкой* (это условие, пожалуй, выполняется). Поэтому, учитывая ограничение на спецификацию, мы можем дополнить аргументацию обозначенными содержательными предпосылками и точно сделать выбор в пользу FE-модели.

Таким образом, как и предполагалось, у нас **наблюдаются нарушение** одного из ключевых предположений для модели со случайными эффектами о равенстве корреляции между индивидуальными эффектами и предикторами 0.

Также можно подумать над тем, почему **значение статистики** получилось не таким большим. Это также говорит о том, что разница между оценками моделей (смещение) не такая большая, а также

что дисперсия оценок не столь мала. Это может быть последствием эндогенности: возможно, предположение об одинаковом эффекте предикторов на отклик не выполняется для большинства простр-х единиц.

Задание 6

Протестируем альтернативные спецификации модели. Начнем с включения модели с **фиксированными эффектами на временные периоды**.

Оценим ее *в полном виде*:

```
    LSDV_time <- lm(lncrime~I(lnpolice-mean(lnpolice)) +
                   I(lndensity-mean(lndensity)) +
                   I(nonwhite-mean(nonwhite)) +
                   factor(year), data = panel)
    summary(LSDV_time)
```

```
##
## Call:
## lm(formula = lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity -
##     mean(lndensity)) + I(nonwhite - mean(nonwhite)) + factor(year),
##     data = panel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94317 -0.18825  0.00802  0.20157  1.82448
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -3.531276   0.038396 -91.969  < 2e-16 ***
## I(lnpolice - mean(lnpolice))    0.144228   0.027752   5.197 2.75e-07 ***
## I(lndensity - mean(lndensity))  0.486468   0.018776  25.909  < 2e-16 ***
## I(nonwhite - mean(nonwhite))    0.219686   0.015246  14.410  < 2e-16 ***
## factor(year)82                 -0.006373   0.054270  -0.117  0.90656
## factor(year)83                 -0.095444   0.054283  -1.758  0.07919 .
## factor(year)84                 -0.154422   0.054306  -2.844  0.00461 **
## factor(year)85                 -0.158556   0.054323  -2.919  0.00364 **
## factor(year)86                 -0.099972   0.054298  -1.841  0.06607 .
## factor(year)87                 -0.030878   0.054320  -0.568  0.56994
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.364 on 620 degrees of freedom
## Multiple R-squared:  0.6019,Adjusted R-squared:  0.5961
## F-statistic: 104.2 on 9 and 620 DF,  p-value: < 2.2e-16
```

Или *в сокращенном* (с внутригрупповым преобразованием):

```
    fe_time <- plm(lncrime~I(lnpolice-mean(lnpolice)) +
                  I(lndensity-mean(lndensity)) +
                  I(nonwhite-mean(nonwhite)) +
                  factor(year), data = panel, index=c("county", "year"), effect = "time", model
    summary(fe_time)
```

```
## Oneway (time) effect Within Model
##
## Call:
## plm(formula = lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity -
##     mean(lndensity)) + I(nonwhite - mean(nonwhite)) + factor(year),
##     data = panel, effect = "time", model = "within", index = c("county",
##         "year"))
##
## Balanced Panel: n = 90, T = 7, N = 630
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.9431668 -0.1882550  0.0080151  0.2015693  1.8244787
##
## Coefficients:
##                                 Estimate Std. Error t-value  Pr(>|t|)
## I(lnpolice - mean(lnpolice))   0.144228   0.027752   5.197 2.753e-07 ***
## I(lndensity - mean(lndensity)) 0.486468   0.018776  25.909 < 2.2e-16 ***
## I(nonwhite - mean(nonwhite))   0.219686   0.015246  14.410 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    204.51
## Residual Sum of Squares: 82.16
## R-Squared:      0.59826
## Adj. R-Squared: 0.59243
## F-statistic: 307.766 on 3 and 620 DF, p-value: < 2.22e-16
```

Какие особенности можно заметить?

1. Во-первых, в модели на временные периоды **оценки всех предикторов статистически значимы**. Теперь они обозначают то, насколько процентов в среднем изменится значение числа преступлений на человека с увеличением значения того или иного предиктора на 1 **процент** (в пространственной перспективе) при прочих равных условиях. *Данная модель кажется довольно устойчивой.*

2. Во-вторых, бОльшая часть оценок коэф-в при дамми на временные периоды (годы 1981–1987) статистически **значимы** на уровне значимости 0.1; и многие — на уровне значимости 0.01. Учитывая тот факт, что в качестве базовой категории был взят 1981 г. и то, что оценки при всех дамми отрицательные (хотя некоторые незначимы), это **открывает более интересную интерпретацию** о том, что в 1981 г. наблюдалось самое высокое значение логарифма числа преступлений на человека при условии равенства всех предикторов среднему значению, а затем по каким-то причинам снижалось (причем не монотонно, но больше уровня 1981 г. больше не поднималось).

3. В-третьих, в целом, можно было бы оттолкнуться от данной спецификации и взять два года с самым высоким и низким средним значения зависимой переменной и проанализировать их.

Далее оценим весьма спорную спецификацию: **twoway model** (с включением и пространственных, и временных эффектов).

```
    fe_twoways <- plm(lncrime~I(lnpolice-mean(lnpolice)) +
                      I(lndensity-mean(lndensity)) +
                      I(nonwhite-mean(nonwhite)) +
                      factor(county), data = panel, index=c("county", "year"), effect = "twoways
    summary(fe_twoways)

## Twoways effects Within Model
##
## Call:
## plm(formula = lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity -
##     mean(lndensity)) + I(nonwhite - mean(nonwhite)) + factor(county),
##     data = panel, effect = "twoways", model = "within", index = c("county",
##         "year"))
##
## Balanced Panel: n = 90, T = 7, N = 630
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.6840621 -0.0726079  0.0021417  0.0750564  0.6941206
##
## Coefficients:
##                                 Estimate Std. Error t-value  Pr(>|t|)
## I(lnpolice - mean(lnpolice))    0.237944   0.026092  9.1194 < 2.2e-16 ***
## I(lndensity - mean(lndensity)) 0.835914   0.323056  2.5875  0.009931 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    16.123
## Residual Sum of Squares: 13.778
## R-Squared:      0.1455
## Adj. R-Squared: -0.010307
## F-statistic: 45.2915 on 2 and 532 DF, p-value: < 2.22e-16
```

В такой модели **кардинально меняется интерпретация**, поскольку она равносильна тому, что мы, пытаясь учесть и пространственную, и временную изменчивость, в начале делаем центрирование по стране, а потом — по временному периоду.

Мы получили всего два коэффициента, **столкнувшись с такой же проблемой**, как ранее — отсутствием изменчивости показателя логарифма небелого населения внутри пространственных групп.

Несмотря на то, что обе оценки получили значимыми, **им не стоит доверять**: недаром такая модель называется *"tricky„*. Ведь теперь, после двух преобразований, оценки коэф-ов показывают то, как в среднем при увеличении того или иного предиктора на 1 процент как в пространственной, так и во временной перспективе, изменяется значение отклика при прочих равных (на сколько процентов).

Важно отметить, что **аналитического решения данной модели не существует** в явном виде, статистические пакеты прибегают к специальным преобразованиям для нахождения результатов анализа.

Более того, для применения данной модели необходим (квази-)эксперементальный дизайн, например, Dif-in-Dif: число единиц N должно быть равно 2-м. **Данное условие невыполнимо в нашем случае**.

Таким образом, нам точно **стоит отказаться** от такой спецификации.

Говоря о выборе между FE моделями на пространственные единицы и временные периоды, я бы предложил сделать выбор в пользу **второй** по следующим причинам:

1. Она менее громоздка и более интерпретируема содержательно: гораздо удобнее и логичнее сравнивать между собой 7 временных периодов, чем 90 округов.

2. Она лучше статистически: все ключевые предикторы статистически значимы на большом уровне доверия, большая часть дамми-переменных на годы также значимы.

3. Она кажется более устойчивой, поскольку в каждом из временных периодов находится большое число пространственных единиц. Это более надежно, чем наблюдения за 7 лет по 90 округам, т.к. во втором случае выбросы могут иметь большое влияние на итоговую оценку.

4. При фиксировании временных периодов мы можем избавиться от проблемы отсутствия внутригрупповой изменчивости предиктора логарифм небелого населения на уровне округов: мы можем посчитать оценки и больше доверять им, ведь теперь мы говорим об изменчивости в пространственной перспективе.

Теперь, когда мы определились с моделью на содержательном уровне, проведем несколько статистических тестов на устойчивость (и не только).

- Во-первых, проведем **F-test Фишера**, сравнив модель с фиксированными эффектами на врем. периоды и pooled-спецификацию без учета панельной структуры:

```
    pFtest(fe_time, pooled)
```

```
##
##  F test for time effects
##
## data:  lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity - mean(lndensity)) +  ...
## F = 2.996, df1 = 6, df2 = 620, p-value = 0.006814
## alternative hypothesis: significant effects
```

  Итак, мы можем увидеть маленькое значение $p-value$, позволяющее нам отвергнуть нулевую гипотезу о незначимости эффектов в пользу альтернативы на уровне значимости 0.01. При данном сравнении выбор, конечно же, делается в пользу FE-модели на врем. периоды.

- Далее переоценим выбранную модель с поправкой на гетероскедастичность. Для этого в начале проведем формальный **тест Бреуша-Пагана** на гетероскедастичность:

```
    bptest(LSDV_time)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  LSDV_time
## BP = 239.35, df = 9, p-value < 2.2e-16
```

  Логично, что мы отвергаем нулевую гипотезу о гомоскедастичности на любом уровне значимости. Гетероскедастичность есть.

  Далее **сделаем поправки по формуле HC3**: она позволяет работать с влиятельными наблюдениями.

```
    coeftest(LSDV_time, vcov = vcovHC, type = "HC3")

##
## t test of coefficients:
##
##                                 Estimate Std. Error   t value  Pr(>|t|)
## (Intercept)                   -3.5312761  0.0325690 -108.4244 < 2.2e-16 ***
## I(lnpolice - mean(lnpolice))   0.1442284  0.0773701    1.8641  0.062775 .
## I(lndensity - mean(lndensity)) 0.4864682  0.0230528   21.1023 < 2.2e-16 ***
## I(nonwhite - mean(nonwhite))   0.2196858  0.0222956    9.8533 < 2.2e-16 ***
## factor(year)82                -0.0063727  0.0474812   -0.1342  0.893277
## factor(year)83                -0.0954445  0.0528017   -1.8076  0.071153 .
## factor(year)84                -0.1544219  0.0517917   -2.9816  0.002980 **
## factor(year)85                -0.1585559  0.0528470   -3.0003  0.002806 **
## factor(year)86                -0.0999716  0.0545600   -1.8323  0.067383 .
## factor(year)87                -0.0308779  0.0476503   -0.6480  0.517218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    summary(LSDV_time)

##
## Call:
## lm(formula = lncrime ~ I(lnpolice - mean(lnpolice)) + I(lndensity -
##     mean(lndensity)) + I(nonwhite - mean(nonwhite)) + factor(year),
##     data = panel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94317 -0.18825  0.00802  0.20157  1.82448
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -3.531276   0.038396 -91.969  < 2e-16 ***
## I(lnpolice - mean(lnpolice))   0.144228   0.027752   5.197 2.75e-07 ***
## I(lndensity - mean(lndensity)) 0.486468   0.018776  25.909  < 2e-16 ***
## I(nonwhite - mean(nonwhite))   0.219686   0.015246  14.410  < 2e-16 ***
## factor(year)82                -0.006373   0.054270  -0.117  0.90656
## factor(year)83                -0.095444   0.054283  -1.758  0.07919 .
## factor(year)84                -0.154422   0.054306  -2.844  0.00461 **
## factor(year)85                -0.158556   0.054323  -2.919  0.00364 **
## factor(year)86                -0.099972   0.054298  -1.841  0.06607 .
## factor(year)87                -0.030878   0.054320  -0.568  0.56994
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.364 on 620 degrees of freedom
## Multiple R-squared:  0.6019,Adjusted R-squared:  0.5961
## F-statistic: 104.2 on 9 and 620 DF,  p-value: < 2.2e-16
```

Сравним две модели: с поправками и без. Можно заметить, что значимость оценок коэф-в при предикторах логарифм плотности населения и доли небелого населения **значимо не изменились**, а вот оценка коэф-та при предикторе логарифм числа полицейских на д.н. осталась значимой *только на уровне значимости 0.1*. Видимо, данный показатель не столь существеннен при рассмотрении его изменения в пространственной перспективе между округами. Это заставляет задуматься о корректности данной спецификации; возможно данный показатель стоило бы видоизменить.

Кроме того, *значимость оценок коэф-в при дамми-переменных на врем. периоды никак не изменилась*, что является положительным фактом, подтверждающим факт хорошей интерпретации данной модели.

Задание 7

Протестируем, *устойчива ли выбранная модель*. Переоценим ее на усеченной выборке: без единиц анализа, в которых корреляция пред- сказанного отклика и наблюдаемого мала.

```
    y_pred <- LSDV_time$fitted
    panel1 <- data.frame(panel, y_pred)

    merged <- panel1 %>% group_by(year)%>% summarize(., cor(lncrime, y_pred))%>% merge(panel1,
    head(merged)
```

```
##   year county   lncrime  lnpolice  lndensity nonwhite ypredicted    y_pred
## 1   81      1 -3.221757 -6.327340  0.8360171 3.006608  -3.340398 -3.072795
## 2   81    119 -2.505320 -6.278997  2.0559671 3.351517  -2.427975 -2.396584
## 3   81     13 -3.504916 -6.719617 -0.6980016 3.471327  -3.447006 -3.773532
## 4   81     81 -2.757219 -6.131665  1.5902070 3.273140  -2.886623 -2.619131
## 5   81    193 -3.952288 -6.984615 -0.2409203 1.780208  -3.916892 -3.960911
## 6   81     41 -3.309443 -6.360438 -0.3677247 3.752842  -3.777333 -3.499214
##   cor(lncrime, y_pred)
## 1            0.8314829
## 2            0.8314829
## 3            0.8314829
## 4            0.8314829
## 5            0.8314829
## 6            0.8314829
```

Отлично, мы рассчитали корреляцию между предсказанными и наблюдаемыми значениями для каждого года. Теперь создадим **индикатор**, показывающий, что корреляция меньше 0.3 по модулю, это будет свидетельствовать о том, что для данного года наблюдается слабая связь между предсказанными и наблюдаемыми значениями.

```
    merged$new <- ifelse(abs(merged$`cor(lncrime, y_pred)`)<0.3,1,0)
    unique(merged$year[merged$new==1])
```

```
## numeric(0)
```

```
    unique(merged$year[merged$new==0])
```

```
## [1] 81 82 83 84 85 86 87
```

Отлично, мы можем видеть, что у нас нет таких временных периодов, для которых наблюдалась бы слабая связь между предсказанными и наблюдаемыми значениями отклика. Соответственно, ***результаты никак не изменятся***: удалять попросту нечего. Это логично, учитывая, что в каждый из временных периодов попадает большое число наблюдений (по 90 — числу округов).

Но на всякий случай, проведем формальную процедуру:

```
    LSDV_time_2 <- lm(lncrime~I(lnpolice-mean(lnpolice)) +
                      I(lndensity-mean(lndensity)) +
                      I(nonwhite-mean(nonwhite)) +
                      factor(year), merged[merged$new == 0,])
    coeftest(LSDV_time_2, vcov = vcovHC, type = "HC3")

##
## t test of coefficients:
##
##                                  Estimate Std. Error   t value  Pr(>|t|)
## (Intercept)                    -3.5312761  0.0325690 -108.4244 < 2.2e-16 ***
## I(lnpolice - mean(lnpolice))    0.1442284  0.0773701    1.8641  0.062775 .
## I(lndensity - mean(lndensity))  0.4864682  0.0230528   21.1023 < 2.2e-16 ***
## I(nonwhite - mean(nonwhite))    0.2196858  0.0222956    9.8533 < 2.2e-16 ***
## factor(year)82                 -0.0063727  0.0474812   -0.1342  0.893277
## factor(year)83                 -0.0954445  0.0528017   -1.8076  0.071153 .
## factor(year)84                 -0.1544219  0.0517917   -2.9816  0.002980 **
## factor(year)85                 -0.1585559  0.0528470   -3.0003  0.002806 **
## factor(year)86                 -0.0999716  0.0545600   -1.8323  0.067383 .
## factor(year)87                 -0.0308779  0.0476503   -0.6480  0.517218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Действительно, результаты никак не изменились.

Задание 8

Задание 8.1

Покажем, какие округа получили **наибольший вес** в формировании оценки коэффициента при предикторе "логарифм числа полицейских на д.н.„, а какие — наименьший.

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

Ранее уже говорилось о том, как получить оценку коэффициента при каком-либо предикторе на основе оценок коэф-тов регрессионных моделей, оцененных на отдельных подгруппах (см. задание 3). Кратко повторим общую логику:

1. 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:

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

2. Далее мы забираем интересующие нас оценки коэффициентов при предикторе $\hat{a}_{1_{\{i\}}}$ и берем их со следующим весом:

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

, то есть в качестве веса выступает доля условной вариации у той или иной пространственной единицы.

3. В результате искомая оценка $\hat{\beta}_1$ получается следующим образом:

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

Понятно, что подгруппа не будет учитываться, если условная вариация предиктора внутри нее нулевая, т.е. все значения одинаковые. Также, если значения предиктора внутри подгруппы практически не будут меняться, оценка по этой подгруппе пойдет с минимальным весом.

Реализуем это в R:

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

```
## # A tibble: 90 x 2
##    county `var(lnpolice)`
##     <dbl>           <dbl>
##  1      1         0.00100
##  2      3         0.00638
##  3      5         0.0381
##  4      7         0.00135
##  5      9         0.00452
##  6     11         0.0724
##  7     13         0.00498
##  8     15         0.00416
##  9     17         0.00172
## 10     19         0.00380
## # ... with 80 more rows
```

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

```
## # A tibble: 90 x 2
## # Groups:   county [90]
##    county    beta
##     <dbl>   <dbl>
##  1      1 -1.02
##  2      3  0.488
##  3      5  0.509
##  4      7  2.06
##  5      9  2.51
##  6     11 -0.384
##  7     13  0.497
##  8     15  0.682
##  9     17 -1.00
## 10     19 -0.0661
## # ... with 80 more rows
```

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

```
##    county         beta var(lnpolice)
## 1       1 -1.019824344  0.0010002156
## 2       3  0.487670696  0.0063813191
## 3       5  0.508708021  0.0380569365
## 4       7  2.060532956  0.0013495698
## 5       9  2.512664501  0.0045180665
## 6      11 -0.384240907  0.0723888462
## 7      13  0.497277293  0.0049767034
## 8      15  0.682461567  0.0041566748
## 9      17 -1.003018277  0.0017226457
## 10     19 -0.066080902  0.0037999697
## 11     21  1.443183468  0.0018185369
## 12     23  0.255709566  0.0006236064
## 13     25 -0.524293919  0.0061560306
## 14     27 -2.013507500  0.0005153843
## 15     33 -0.028066474  0.0197770935
## 16     35 -0.366986421  0.0021496146
## 17     37  0.561059626  0.0046211511
## 18     39  0.920525872  0.0049294971
## 19     41  1.004300285  0.0034139455
## 20     45  0.973282519  0.0005319867
## 21     47 -0.090240122  0.0243902412
## 22     49 -0.447790612  0.0080651707
## 23     51  1.243548842  0.0112288335
## 24     53 -1.062939098  0.0085307646
## 25     55  1.922387008  0.0129893833
## 26     57  1.863113182  0.0003689250
## 27     59  1.018709767  0.0090058871
## 28     61 -0.991321469  0.0047858694
## 29     63 -0.027129062  0.0043859309
## 30     65  0.006129102  0.1244454472
## 31     67 -0.187166400  0.0044509408
## 32     69 -0.617188499  0.0013322435
## 33     71 -0.121347979  0.0019743326
## 34     77  1.473075578  0.0024060028
## 35     79  0.512852517  0.0125877781
## 36     81 -0.455630559  0.0022594891
## 37     83 -0.059246972  0.0029166191
## 38     85  0.410395015  0.0005815071
## 39     87 -0.665640183  0.0050817256
## 40     89 -0.123552014  0.0131155555
## 41     91  1.437843080  0.0028309280
## 42     93  1.434099862  0.0024176587
## 43     97  1.677206949  0.0005079605
## 44     99 -0.259078141  0.0335636290
## 45    101  1.843599643  0.0032156920
## 46    105  0.214193074  0.0012978997
```

```
## 47    107  1.163007952  0.0005663301
## 48    109  0.707906808  0.0032656440
## 49    111  0.052667368  0.0168744567
## 50    113  0.464839007  0.0042520118
## 51    115 -0.627811072  0.0729138369
## 52    117  0.517920084  0.1197225426
## 53    119  0.348909776  0.0073130676
## 54    123 -0.475676513  0.0062303183
## 55    125 -1.662213013  0.0035163243
## 56    127 -0.817760507  0.1181128747
## 57    129 -0.009309414  0.0104416839
## 58    131  0.579084057  0.0385860718
## 59    133  1.215075334  0.0162356883
## 60    135  0.076548595  0.0092759200
## 61    137  0.327888206  0.0173690304
## 62    139 -0.568096649  0.0187670567
## 63    141  0.355005515  2.9269363460
## 64    143  1.042968198  0.0081970523
## 65    145 -0.835692377  0.0010509877
## 66    147 -0.165139435  0.0160928698
## 67    149  0.235957648  0.3913301709
## 68    151  0.189926216  0.0264503373
## 69    153 -1.322305575  0.0025580909
## 70    155 -2.107765868  0.0037443046
## 71    157 -2.586563108  0.0009356849
## 72    159  0.085603832  0.0042190536
## 73    161 -0.442357078  0.0035419112
## 74    163  1.031215577  0.0006007651
## 75    165 -0.122108914  0.0034323515
## 76    167 -0.633679042  0.0016211387
## 77    169 -0.084764634  0.0218483456
## 78    171  1.072397727  0.0023518945
## 79    173  0.621922727  0.0948187191
## 80    175  0.238516400  0.2903861697
## 81    179  0.299327957  0.0005128167
## 82    181  0.895958437  0.0021007384
## 83    183 -0.838199590  0.0025257550
## 84    185  0.205473010  1.4746349719
## 85    187  0.635430730  0.0133783043
## 86    189 -0.125693011  0.0104377021
## 87    191  0.058254670  0.0010660729
## 88    193  0.296866888  0.0106012679
## 89    195 -0.831897848  0.2207182538
## 90    197 -0.931539499  0.0147332635
```

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

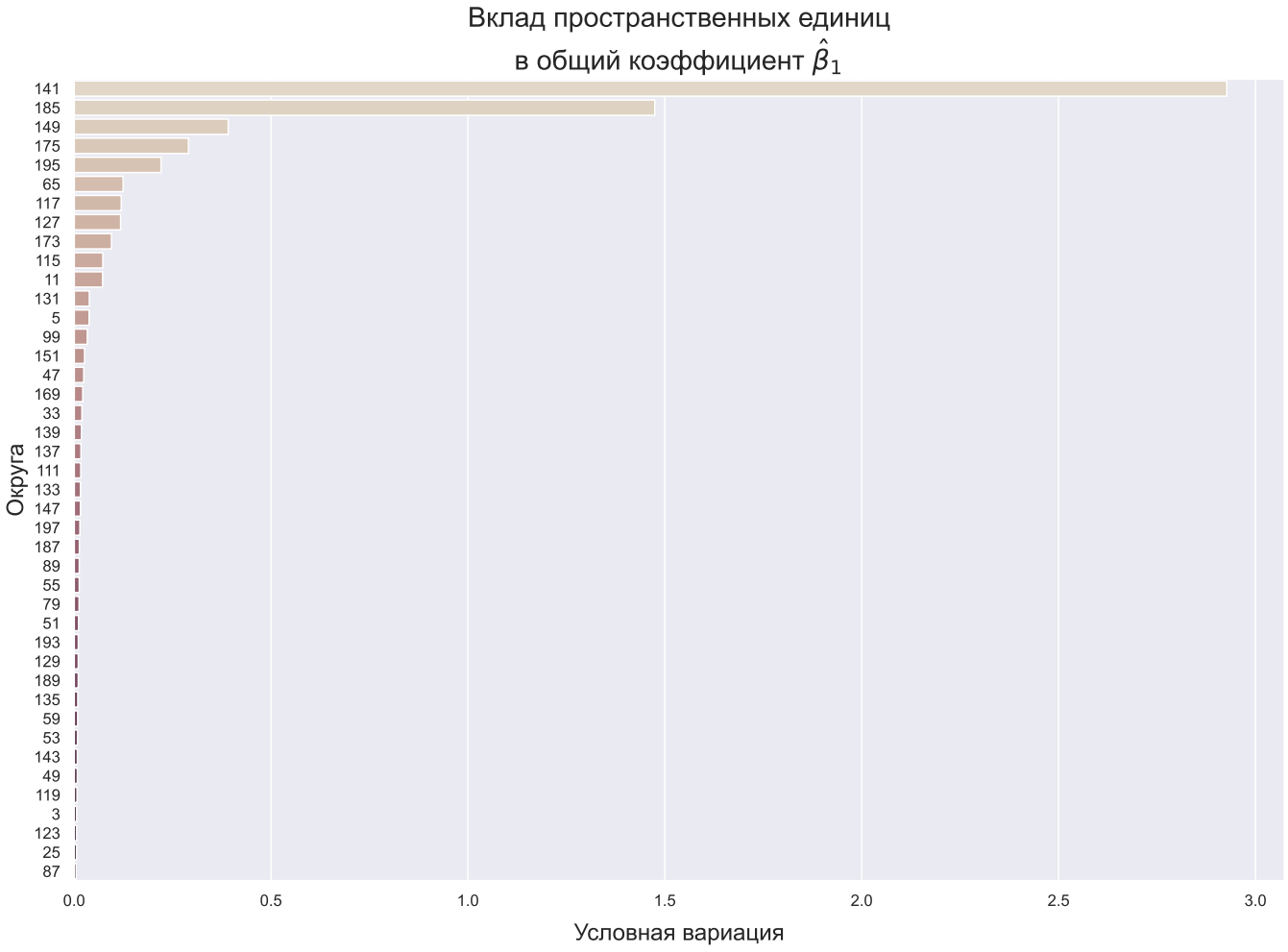
```
## [1] 0.2131733
```

```
    sum(m$`var(lnpolice)`)
```

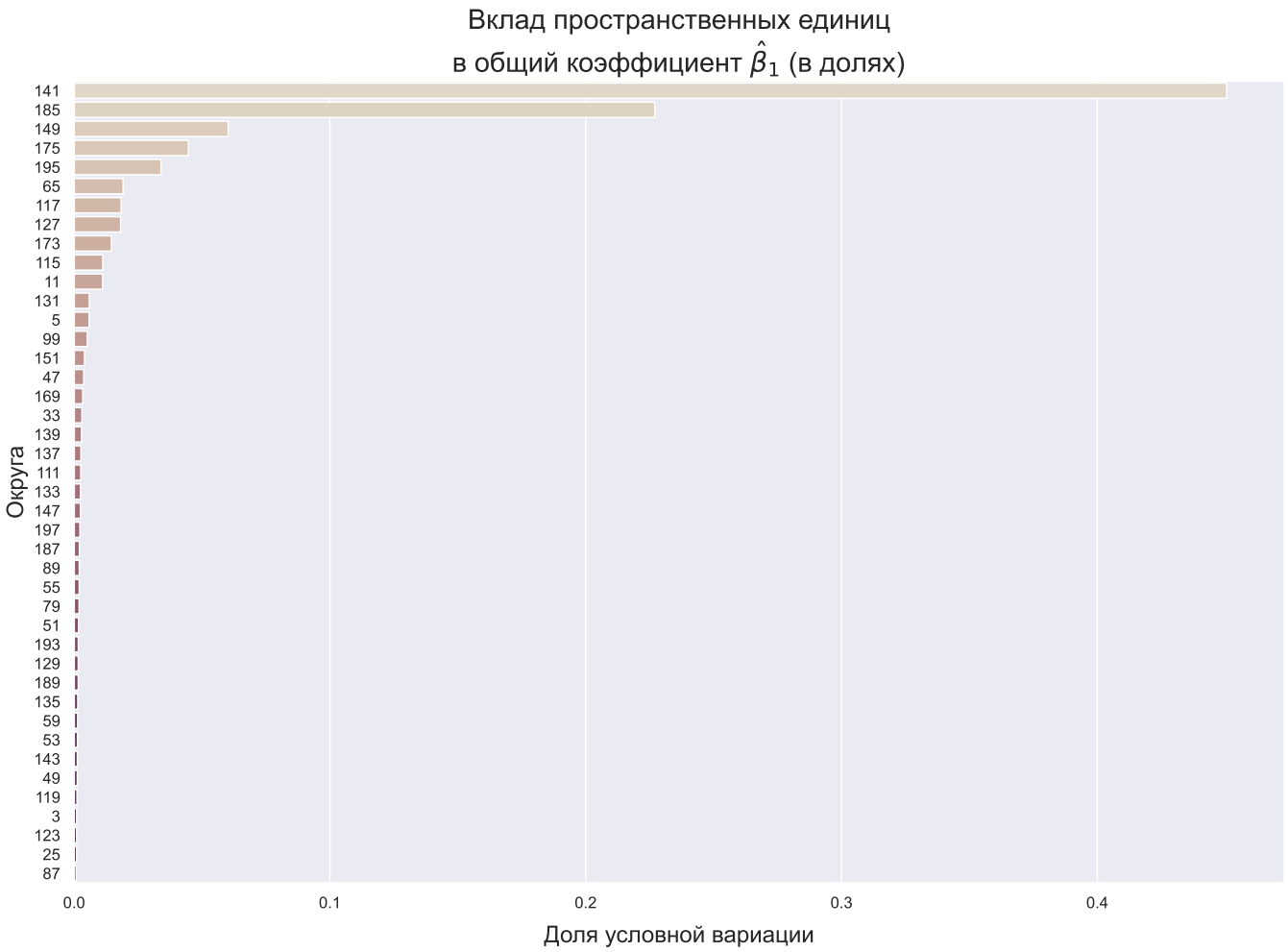
```
## [1] 6.495892
```

Теперь **проиллюстрируем** это более наглядно на графиках.

Итак, в начале *выведем условную вариацию* предиктора “логарифм числа полицейских на д.н.„, которая в дальнейшем будет выступать “весом„ для оценки коэффициента на каждый округ. Для удобства представления на графике будут отражены только те округа, условная вариация которых больше 0.005 (всего 42 из 90), т.к. значения для других неотличимы от 0 и не видны на графике:

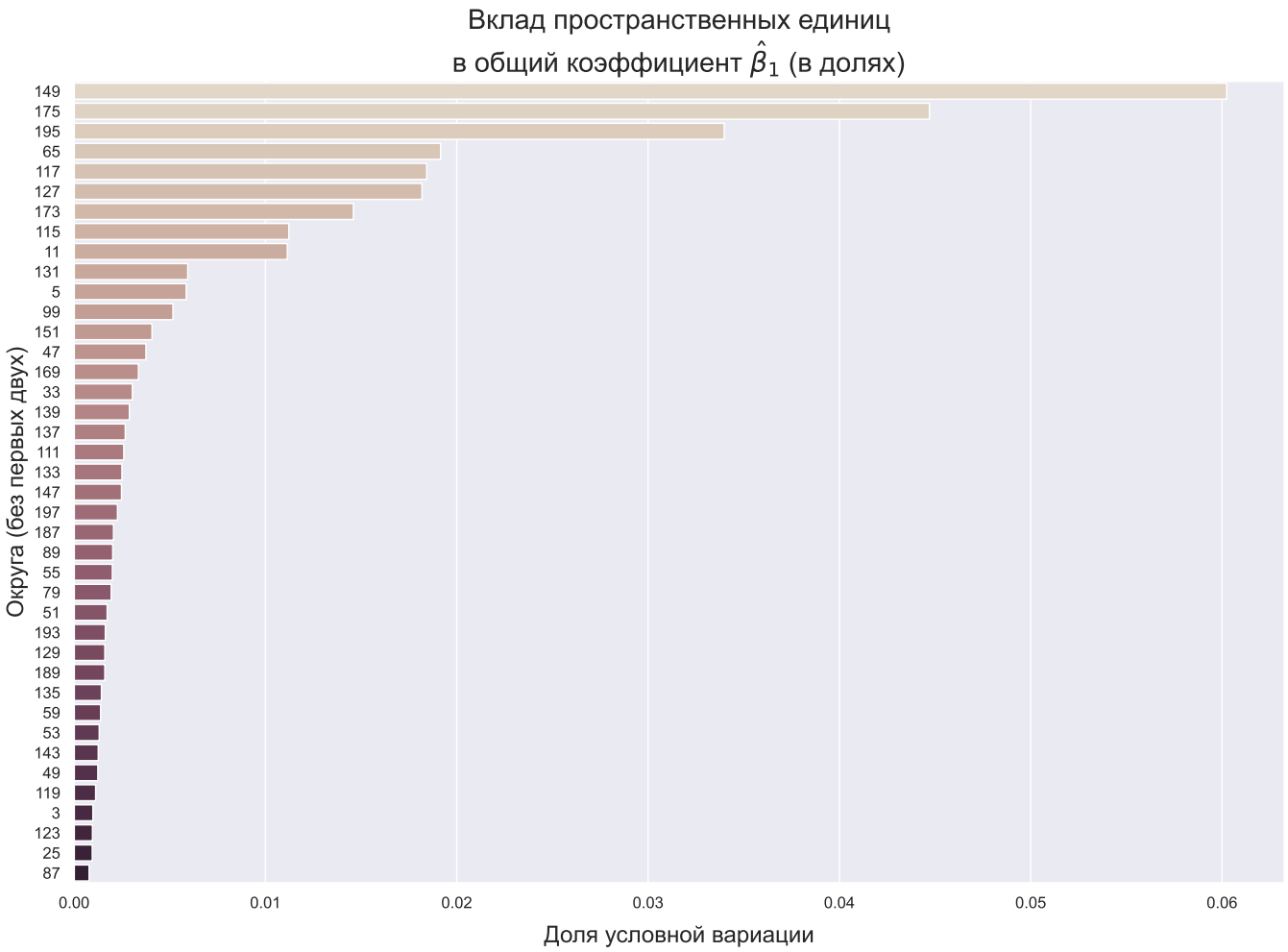


Для удобства можно также вывести те же значения, но **в терминах долей**, т.е. уже непосредственно весов в явном виде (разделив все значения на $\sum_{i=1}^{N} Var(lnpolice|county = i)$):



Видим **не самую хорошую вещь**: доля лишь одного 141 округа во всей итоговой оценке составляет больше 40%, а округа 185 — больше 20%, т.е. в сумме практически 70%! При этом, возможно, их условная дисперсия столь высока из-за **ошибок в данных или неправдоподобной информации**. Запомним это.

Учитывая тот факт, что округа 141 и 185 имеют очень большой вес относительно других округов, можно *временно исключить их из данного графика*, увеличив масштаб:



Видно, что масштаб очень маленький: начиная с 4-го по величине места вклад равен меньше 5%. Данное постепенно затухающее распределение было бы неплохим, если бы не 2 значения выше, которые вносят огромную долю.

Таким образом, мы видим, что **наибольший вклад** в оценку $\hat{\beta}_1$ с большим отрывом будут вносить такие округа, как 141 и 185.

Посмотрим на них:

```
    panel[panel$county==141, ]

## # A tibble: 7 x 7
##   county  year lncrime lnpolice lndensity nonwhite ypredicted
##    <dbl> <dbl>   <dbl>    <dbl>     <dbl>    <dbl>      <dbl>
## 1    141    81   -3.55    -6.89     -1.35     3.69      -3.26
## 2    141    82   -3.62    -6.94     -1.34     3.69      -3.27
## 3    141    83   -2.40    -4.52     -1.32     3.69      -2.76
## 4    141    84   -2.47    -3.48     -1.30     3.69      -2.53
## 5    141    85   -2.80    -3.34     -1.28     3.69      -2.51
## 6    141    86   -1.81    -3.54     -1.24     3.69      -2.55
## 7    141    87   -3.46    -6.67     -1.20     3.69      -3.22

    panel[panel$county==185, ]

## # A tibble: 7 x 7
```

```
##   county  year lncrime lnpolice lndensity nonwhite ypredicted
##    <dbl> <dbl>   <dbl>    <dbl>     <dbl>    <dbl>      <dbl>
## 1    185    81   -4.53    -4.44    -0.969     4.16      -4.47
## 2    185    82   -4.21    -4.36    -0.975     4.16      -4.45
## 3    185    83   -4.55    -4.37    -0.969     4.16      -4.45
## 4    185    84   -4.19    -4.43    -0.957     4.16      -4.47
## 5    185    85   -4.85    -4.29    -0.957     4.16      -4.44
## 6    185    86   -5.39    -7.01    -0.951     4.16      -5.02
## 7    185    87   -4.52    -6.71    -0.945     4.16      -4.95
```

Действительно, можно заметить **очень большую изменчивость** по предиктору логарифм числа полицейских на д.н., учтивая, что это степень числа e. Эти два округа вносят очень большую долю в итоговую оценку. Теперь понятно, почему за этими округами стоит наибольшая **условная вариация** данного предиктора. По сути, итоговая оценка в основном будет опираться на них (всего лишь на 2 из 90 округов).

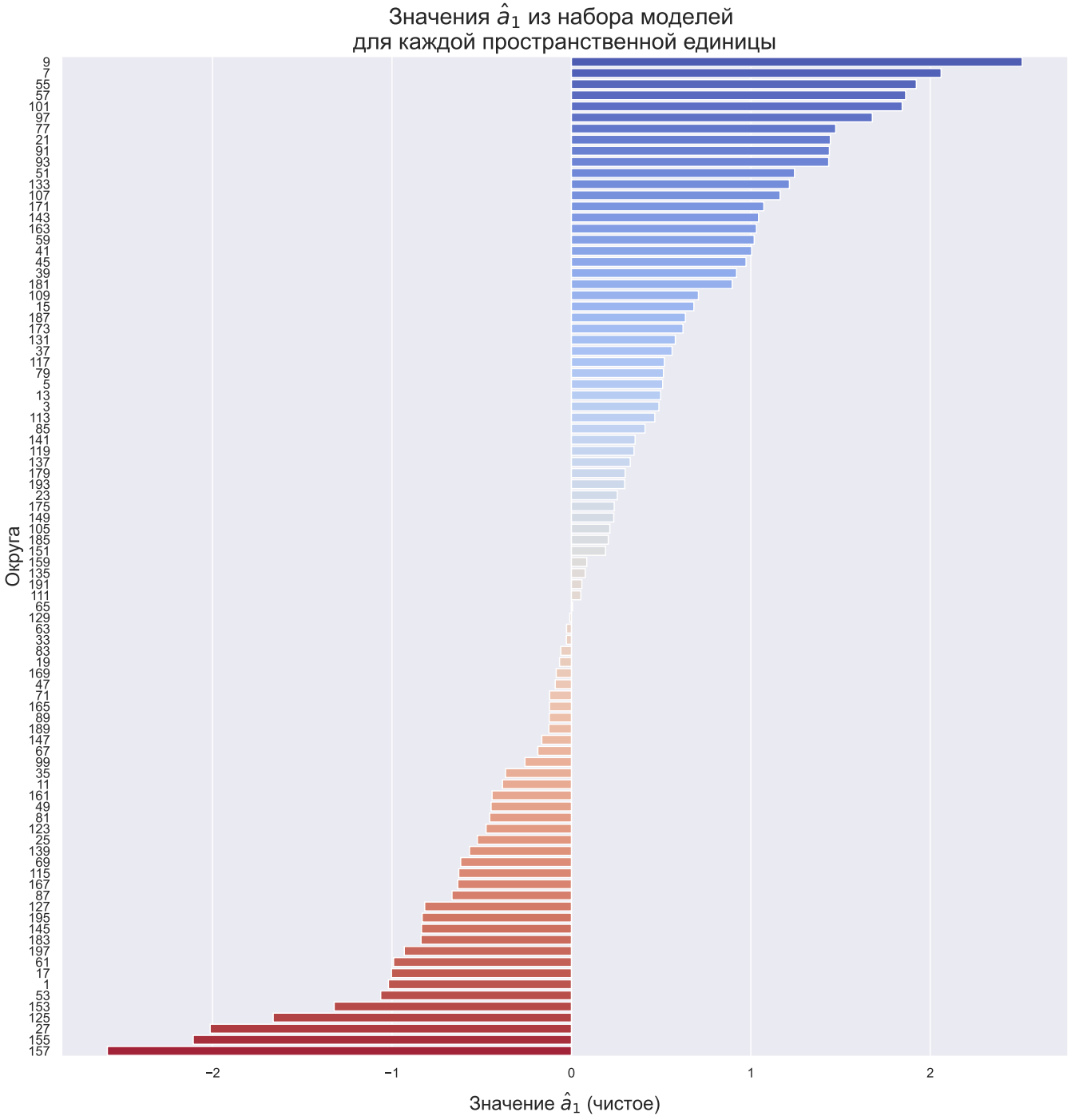
Посмотрим также на противоположный пример:

```
panel[panel$county==57, ]
```

```
## # A tibble: 7 x 7
##   county  year lncrime lnpolice lndensity nonwhite ypredicted
##    <dbl> <dbl>   <dbl>    <dbl>     <dbl>    <dbl>      <dbl>
## 1     57    81   -3.30    -6.59     0.736     2.40      -3.52
## 2     57    82   -3.47    -6.61     0.744     2.40      -3.53
## 3     57    83   -3.54    -6.64     0.749     2.40      -3.54
## 4     57    84   -3.66    -6.61     0.763     2.40      -3.53
## 5     57    85   -3.65    -6.61     0.772     2.40      -3.53
## 6     57    86   -3.58    -6.58     0.789     2.40      -3.53
## 7     57    87   -3.51    -6.60     0.812     2.40      -3.53
```

Действительно, **условная вариация практически нулевая**. Наблюдаются минимальные колебания. При этом, возможно, сам коэффициент при оценивании регрессии только на этой подвыборке будет очень большим, т.к. зависимая переменная меняется гораздо сильнее. Но в итоге **данная подгруппа** (как и многие другие) **практически не будет учитываться** при расчете оценки коэф-та модели с фикс. эффектами.

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

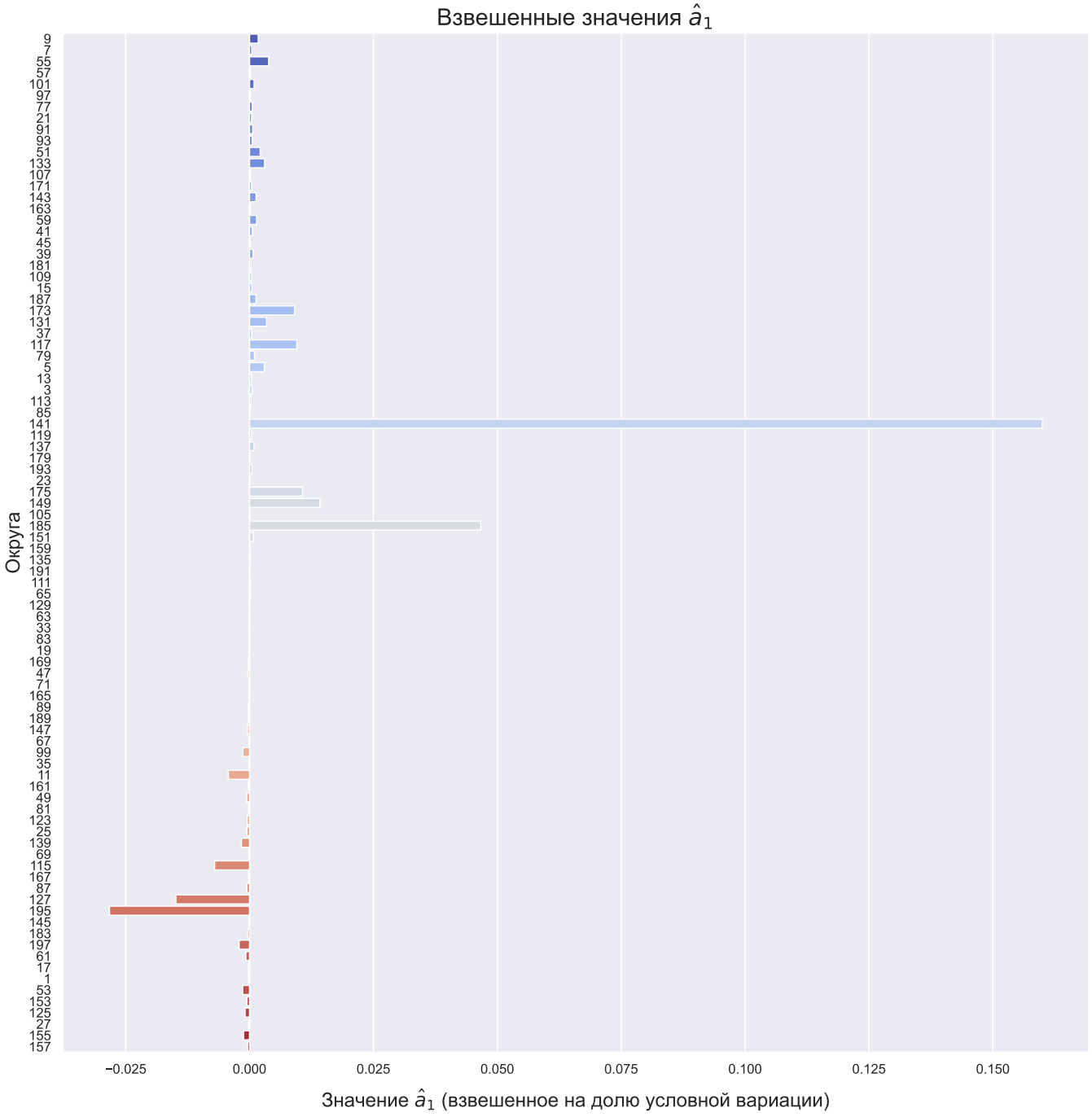


Здесь значения $\hat{a}_{1_{\{i\}}}$ отсортированы по убыванию. Можно предположить, что они соответствуют связи предиктора "логарифм числа полицейских на д.н.„ и отклика — логарифма числа преступлений на человека. Удивительно, но мы видим, что примерно в половине округов связь противоположная, нежели в другой половине. Возможно, мы столкнулись с проблемой **эндогенности**. **Ведь не до конца понятно, что первичнее — увеличение числа преступлений в округе или увеличение числа полицейских.** Однако это также может быть связано с маленькими (незначимыми) колебаниями.

По всей видимости, наибольшая связь между этими двумя показателями прослеживается в 9, 7, 55 округах, а наименьшая — в 157, 155, 27. 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 округа в нашей LSDV-модели.* Точно так же, как и эффект округов с сильной отрицательной оценкой коэффициента. 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. Ранее мы видели, каким минимальным может быть отклонение предиктора. Однако стоит помнить, что *всего два округа будут вносить с отрывом большую часть вклада в итоговую оценку.* Здесь они ближе к середине графика.

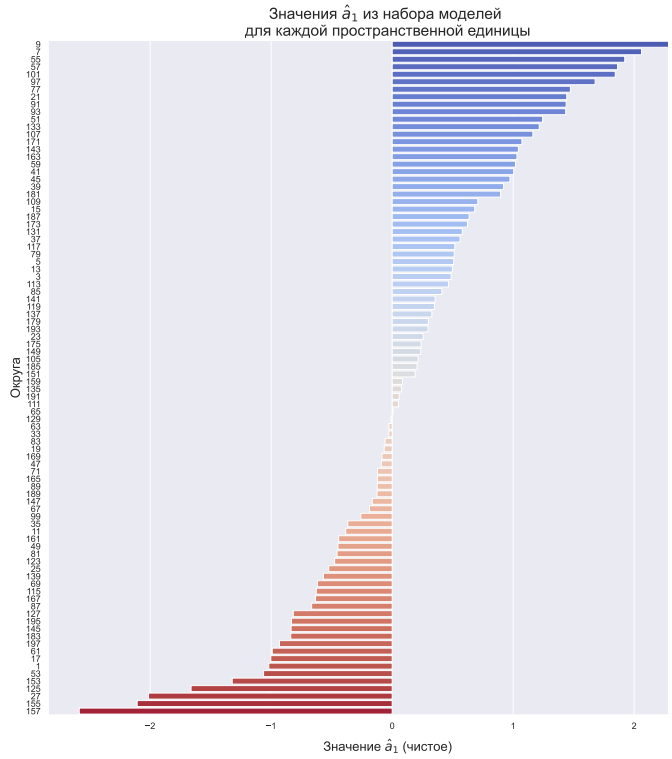
Далее выведем **взвешенные значения** $\hat{a}_{1_{\{i\}}}$, сохранив тот же порядок округов:



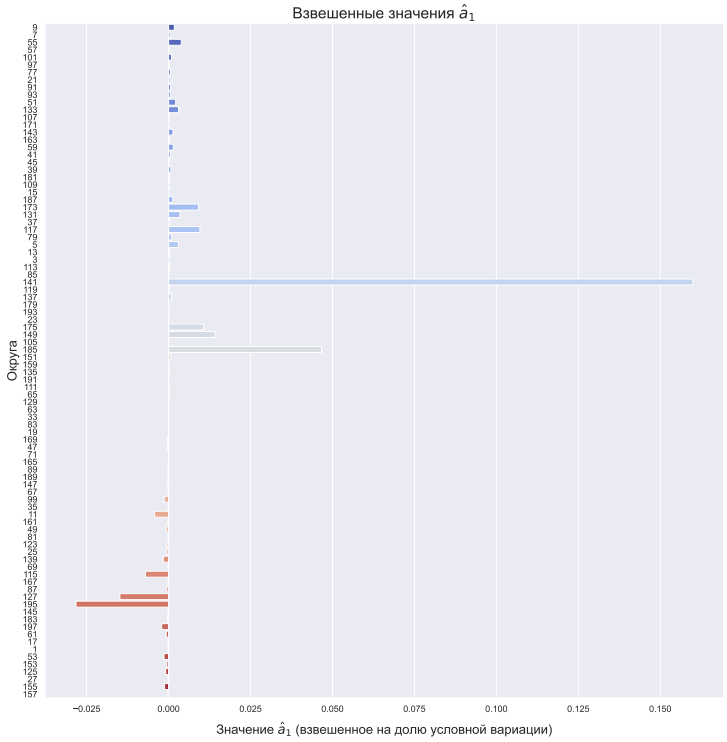
Видим, как два округа, вносящие максимальный вклад, **“выстрелили„** на этом графике.

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

Мы видим, как сильно возросли и упали итоговые значения оценок, которые учитывались моделью, в особенности смотря на бывшие максимальные и минимальные значения. Но и не нужно забывать смотреть на масштаб взвешенных оценок: он едва превосходит 0.01 для большинства округов. При этом основной вклад в итоговой оценке пришелся на два округа где-то из “середины„ графика по силе эффекта. Возможно, это делает ясным то, почему оценка при данном коэффициенте в некоторых спецификациях была незначимой.



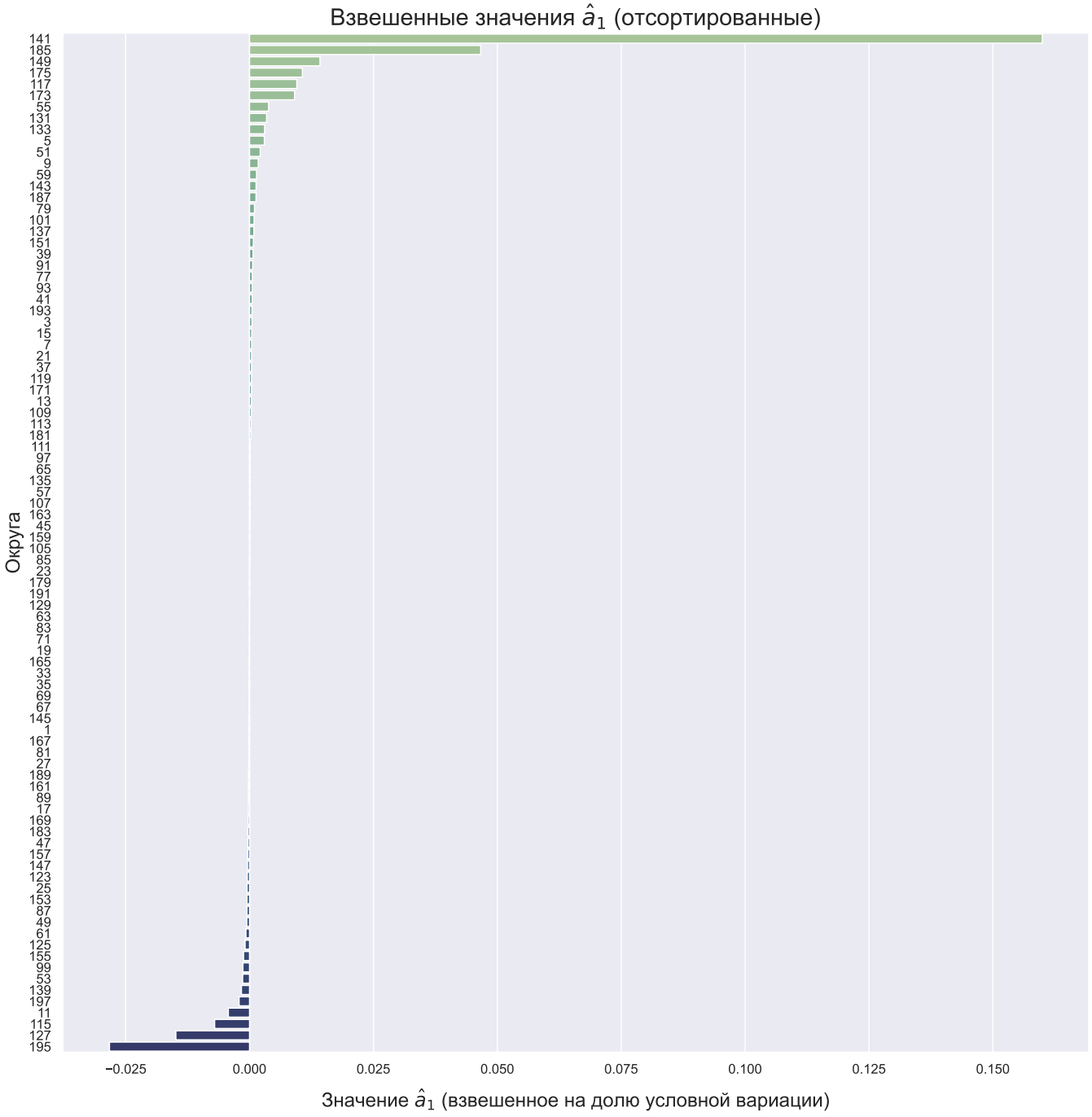
(a) Чистые значения



(b) Взвешенные значения

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

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



Видим, что по большей части вклад минимален, и это с учетом того, что мы отсекли 48 стран, в которых условная вариация практически нулевая.

Задание 8.2

Реализуем **взвешивание в случае с контрольной переменной** "натуральный логарифм плотности населения,,, т.е. уже в случае множественной регрессии.

В начале очистим $lncrime_{it}$ от эффекта контрольной переменной $lndensity_{it}$. Для этого сохраним остатки регрессии $lncrime_{it}$ на $lndensity_{it}$, а также по такому же принципу очистим предиктор $lnpolice_{it}$

от эффекта $lndensity_{it}$. Добавим остатки в исходный набор данных:

```
    reg_crime_dens = lm(lncrime~lndensity +
                        factor(county), data=panel)
    y_clear = reg_crime_dens$residuals

    reg_crime_pol = lm(lnpolice~lndensity +
                           factor(county), data=panel)
    x_clear = reg_crime_pol$residuals

    panel$y_clear = y_clear
    panel$x_clear = x_clear
```

Далее повторим процедуру взвешивания, но вместо изначальных зависимой переменной логарифм числа преступлений на человека и предиктора число полицейских на д.н. будем использовать сохраненные остатки (их **очищенный эффект**). *Именно так реализуется "взвешивание„ в случае множественной регрессии:*

```
    # считаем
    var2 <- summarize(group_by(panel, county), var(x_clear))
    var2
```

```
## # A tibble: 90 x 2
##    county `var(x_clear)`
##     <dbl>          <dbl>
##  1      1       0.000813
##  2      3       0.00502
##  3      5       0.0373
##  4      7       0.00117
##  5      9       0.00369
##  6     11       0.0675
##  7     13       0.00468
##  8     15       0.00458
##  9     17       0.00169
## 10     19       0.0149
## # ... with 80 more rows
```

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

```
## # A tibble: 90 x 2
## # Groups:   county [90]
##    county   beta
##     <dbl>  <dbl>
##  1      1 -0.842
##  2      3  1.03
##  3      5  0.513
##  4      7  1.98
##  5      9  2.52
##  6     11 -0.399
```

```
##  7     13  0.483
##  8     15  0.739
##  9     17 -1.06
## 10     19  0.256
## # ... with 80 more rows
```

```
    # запишем значения чистых оценок и условную вариацию
    m <- as.data.frame(merge(a, var2, by ="county"))

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

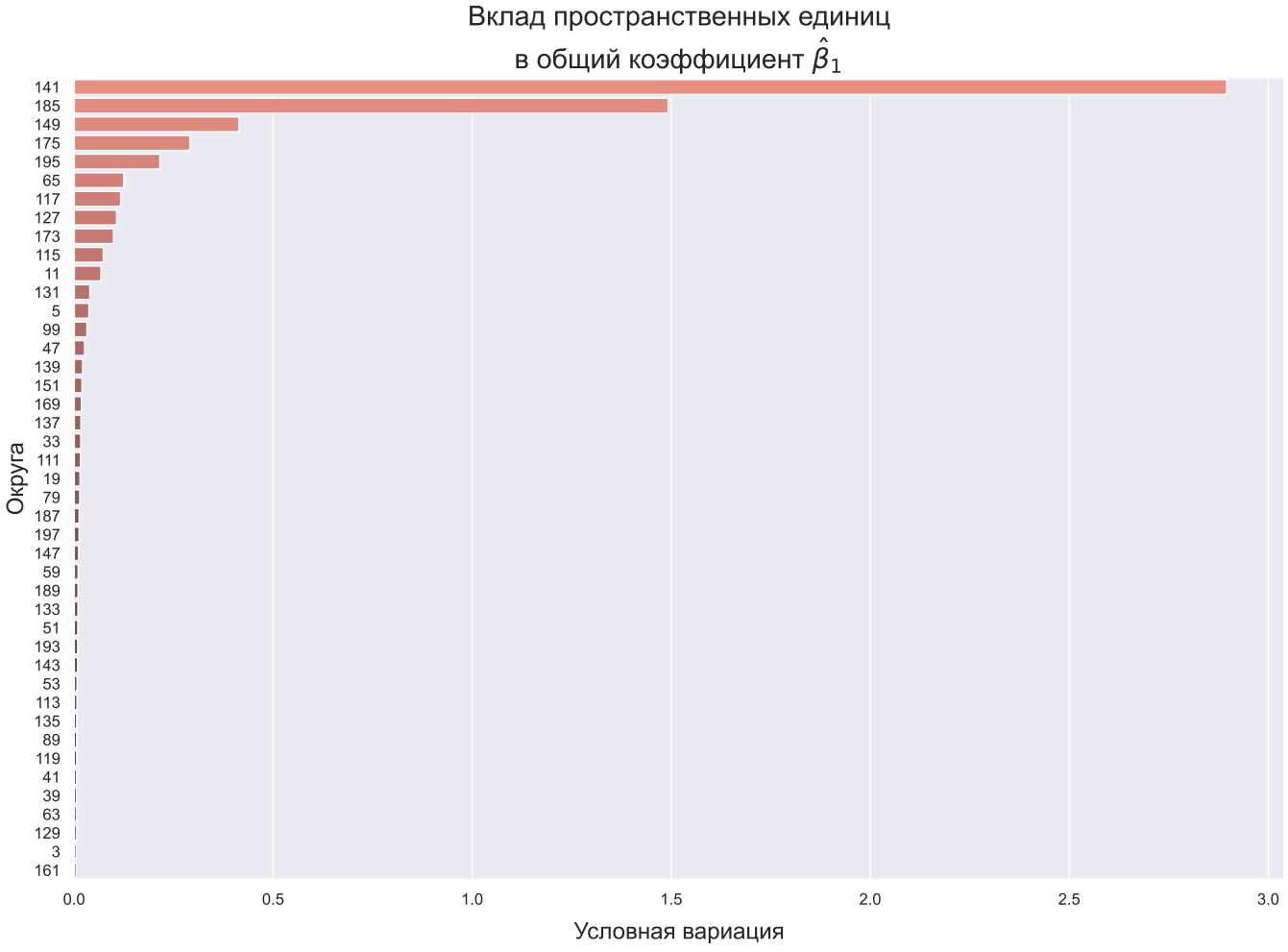
```
## [1] 0.2138657
```

```
    sum(m$`var(x_clear)`)
```

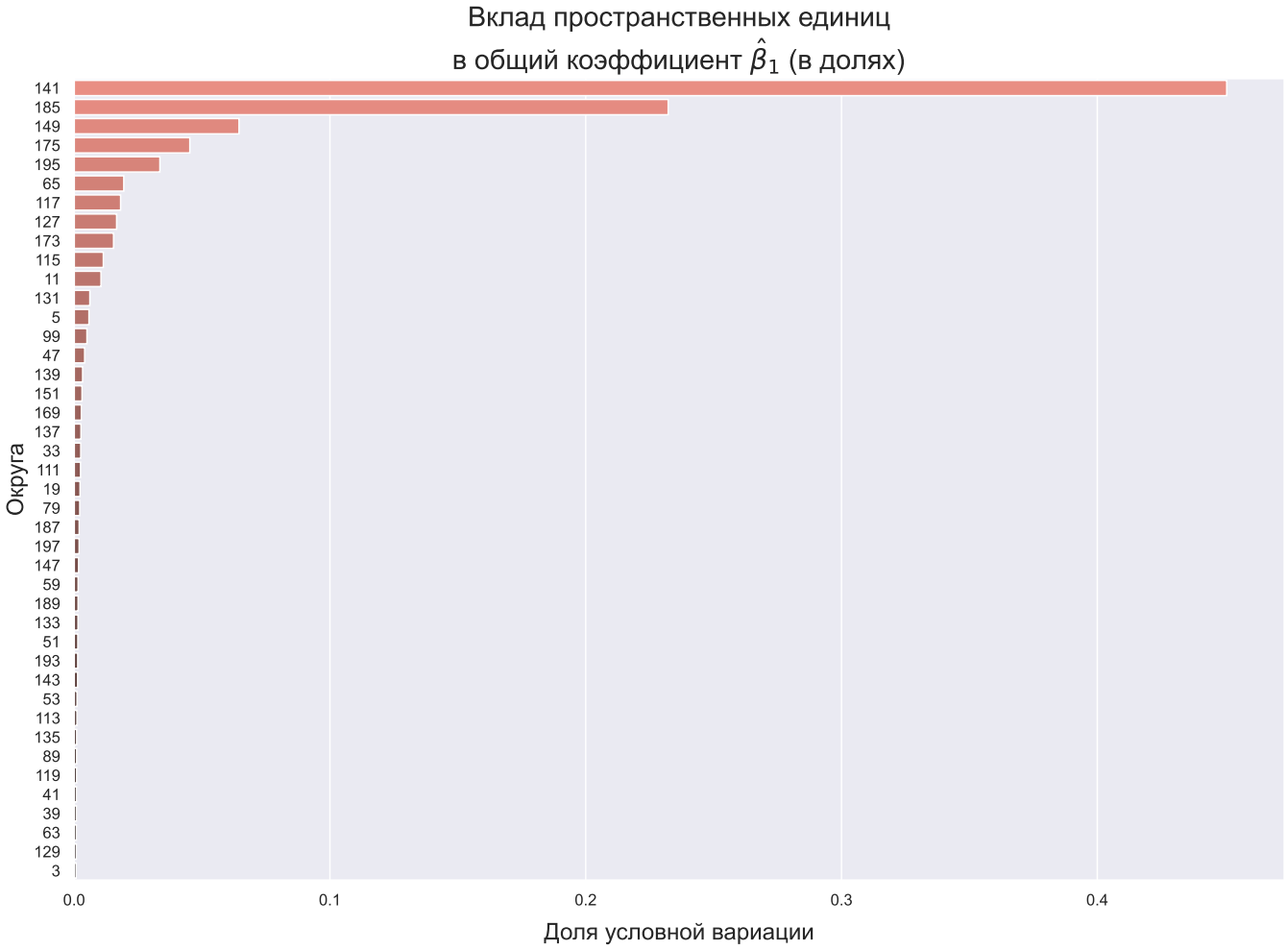
```
## [1] 6.426239
```

Теперь самое интересное: посмотрим на визуализацию.

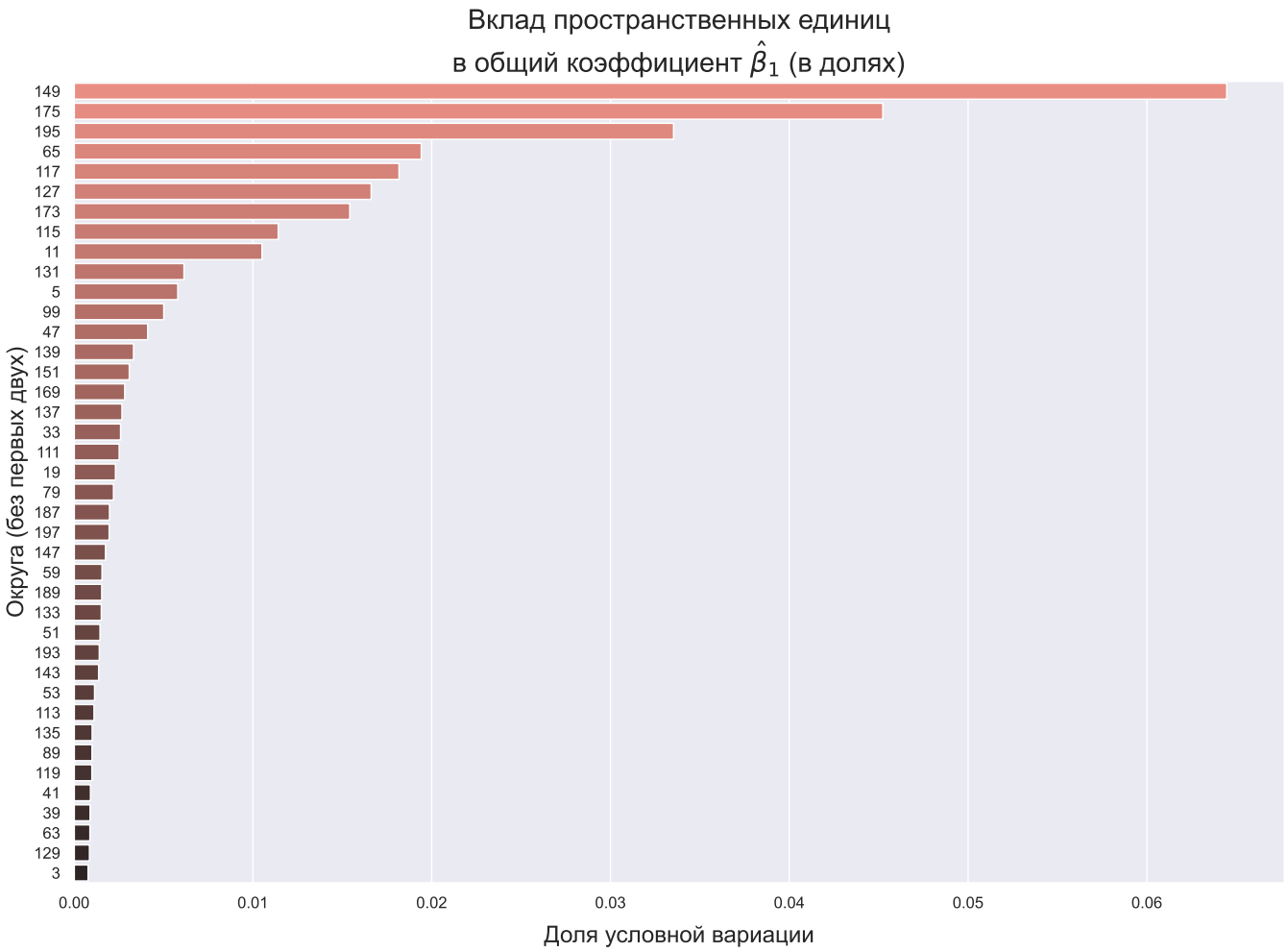
В начале, как и прежде, выведем округа, у которых условная вариация независимой переменной больше 0.005:



Теперь для удобства посмотрим на те же значения, но **в терминах долей**, т.е. уже непосредственно весов в явном виде (разделив все значения на $\sum_{i=1}^{N} Var(x\_clear|county = i)$):

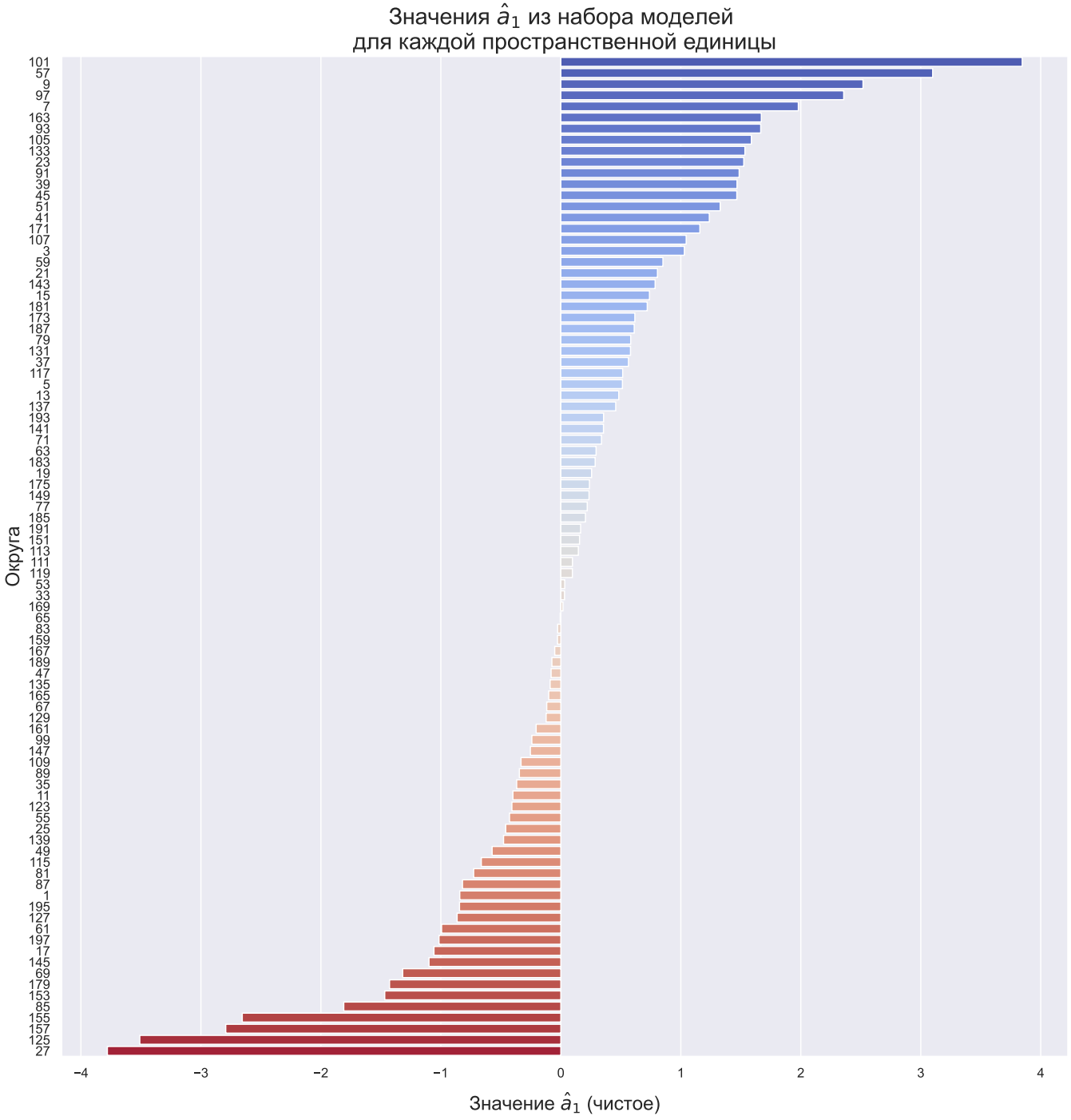


Видим мало изменений. Округа 141 и 185 все еще имеют очень большой вес относительно других округов (более 60% в сумме), можно временно исключить их из данного графика, увеличив масштаб:



Что нового можно заметить: некоторые округа "посередине„ и "в хвосте„ поменялись местами, но не сильно. **Фактически особенных изменений нет**. Особенно учитывая тот факт, что уже в середине графика вес меньше 1%.

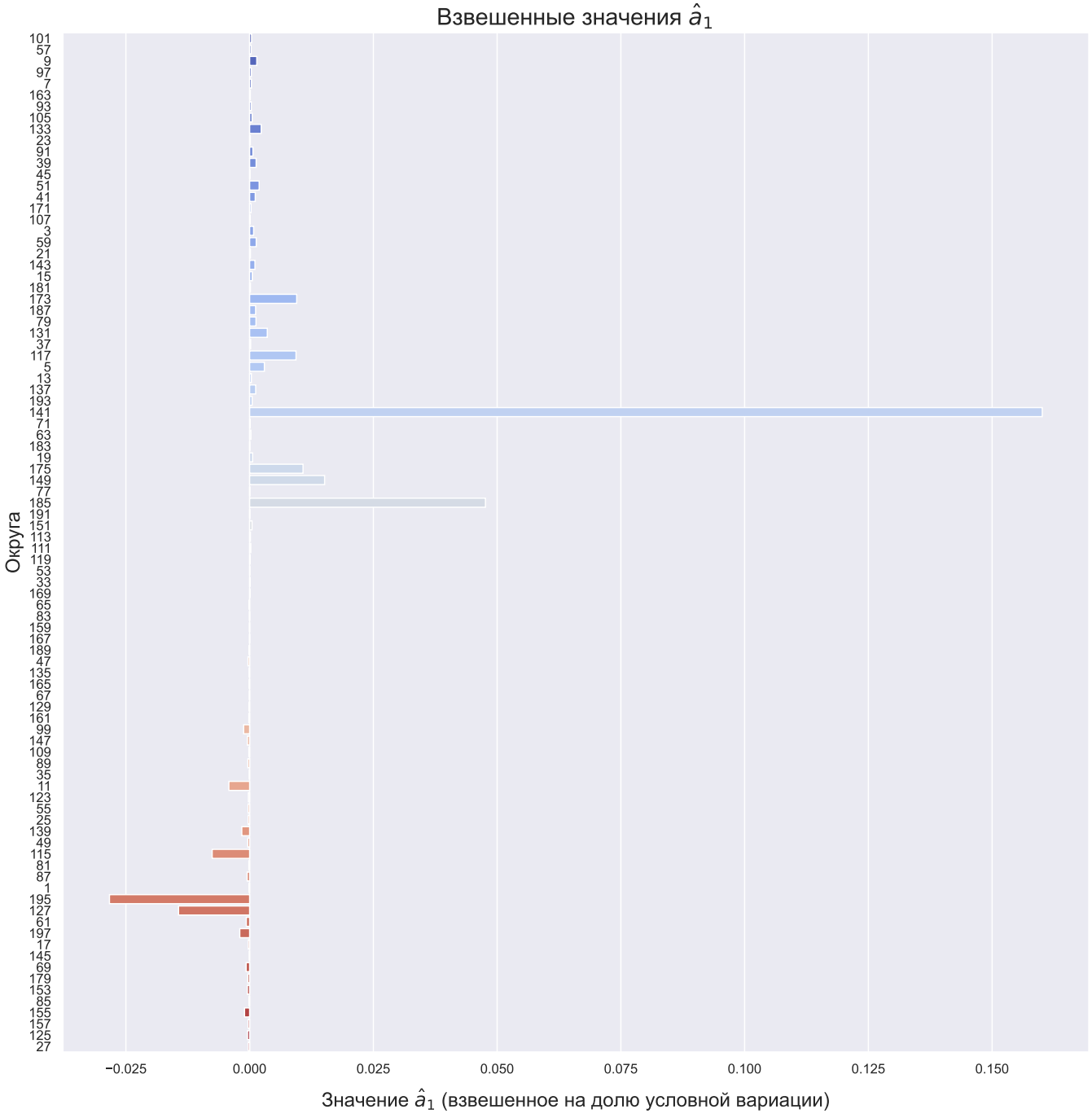
Теперь посмотрим на **"чистые„ значения** $\hat{a}_{1_{\{i\}}}$, которые получаются при оценивании отдельных моделей на *каждую пространственную единицу* i:



Здесь уже видим *более существенные изменения* после включения контрольно переменной натуральный логарифм плотности населения. Некоторые округа, как "в топе„, так и "внизу„ рейтинга (то есть с самыми большими и маленькими значениями оценок) поменялись местами на 1-3 места. Однако это вряд ли приведет к сильному изменению итоговой оценки, ведь вес оценок каждой из подгрупп практически не изменился, и первые (те же самые) два округа "весят„ более 60%, как и ранее.

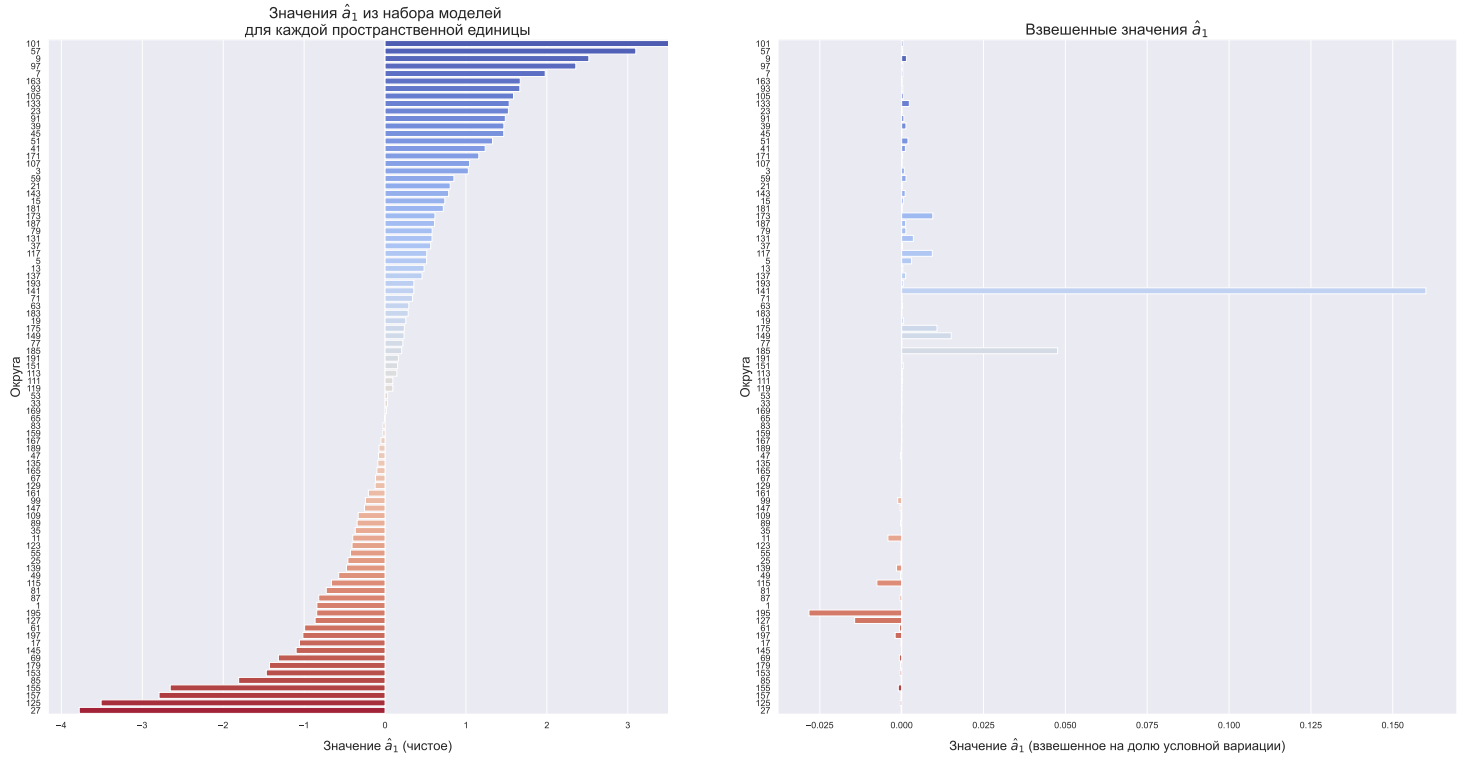
Хотя стоит отметить, что эффект внутри подгрупп несколько сместился после включения контрольной переменной.

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



Наблюдаем такую же картинку: **141 и 185 округа вносят максимальный вклад в итоговую оценку, несмотря на добавление контрольной переменной.**

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



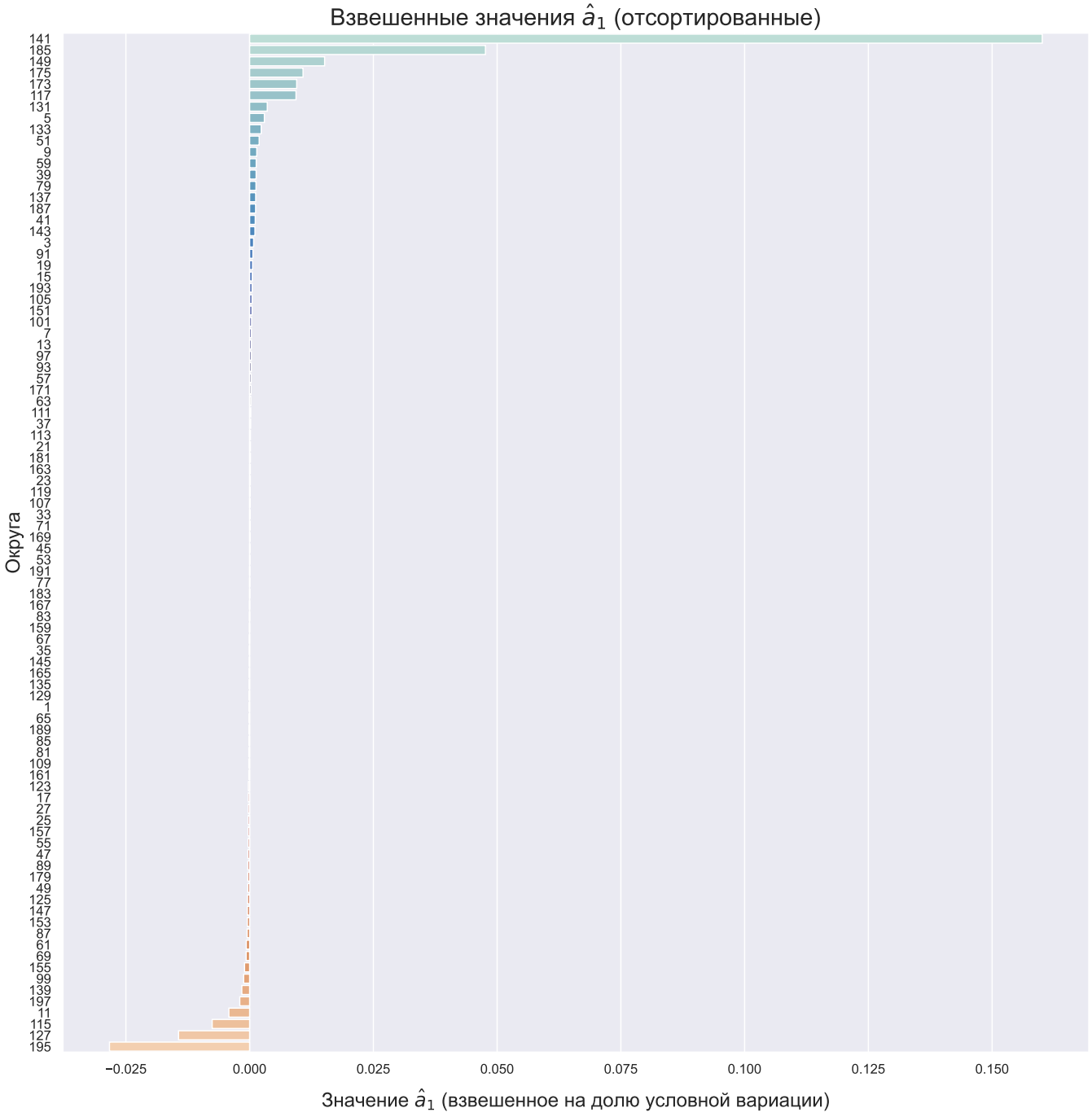
(a) Чистые значения

(b) Взвешенные значения

Рис. 2: Сравнение чистых и взвешенных оценок коэффициентов при предикторе

В целом, картина аналогичная.

И, наконец, представим итоговый **“рейтинг,,** округов по значению взвешенных оценок, т.е. вкладу в итоговую оценку $\hat{\beta}_1$:



Топ-4 по вкладу в итоговую оценку не поменялся. Округа на 5 и 6 местах поменялись друг с другом, однако это не столь значительно. Интересно, что округ 55 (“экс„ топ-7) “улетел„ далеко вниз к нулевым значениям. В целом же именно топ-4 округов по доле условной вариации составляет топ-4 по взвешенным значениям, которые идут в итоговую общую оценку (вместе их вклад составляет 79% общей вариации).

Итак, подумаем над итогом данной процедуры. В начале убедимся в том, что значения до и после включения контрольной переменной все-таки не идентичны:

```r
    # тест на совпадение
    var1 <- summarize(group_by(panel, county), var(lnpolice))
    a1 <- group_by(panel, county) %>%
      do(data.frame(beta1 = coef(lm(lncrime ~ lnpolice, data = .))[2]))
    m1 <- as.data.frame(merge(a1, var1, by ="county"))
    m1$coef1 <- m1$beta1*(m1$`var(lnpolice)`/sum(m1$`var(lnpolice)`))

    reg_crime_dens = lm(lncrime~lndensity + factor(county), data=panel)
    y_clear = reg_crime_dens$residuals
    reg_crime_pol = lm(lnpolice~lndensity + factor(county), data=panel)
    x_clear = reg_crime_pol$residuals
    panel$y_clear = y_clear
    panel$x_clear = x_clear
    var2 <- summarize(group_by(panel, county), var(x_clear))
    a2 <- group_by(panel, county) %>%
      do(data.frame(beta2 = coef(lm(y_clear ~ x_clear, data = .))[2]))
    m2 <- as.data.frame(merge(a2, var2, by ="county"))
    m2$coef2 <- m2$beta2*(m2$`var(x_clear)`/sum(m2$`var(x_clear)`))

    head(m1)     # без контролей
```

```
##   county      beta1 var(lnpolice)         coef1
## 1      1 -1.0198243   0.001000216 -0.0001570291
## 2      3  0.4876707   0.006381319  0.0004790693
## 3      5  0.5087080   0.038056936  0.0029803247
## 4      7  2.0605330   0.001349570  0.0004280910
## 5      9  2.5126645   0.004518066  0.0017476252
## 6     11 -0.3842409   0.072388846 -0.0042818991
```

```r
    sum(m1$coef1) # без контролей
```

```
## [1] 0.2131733
```

```r
    head(m2)     # с контролями
```

```
##   county      beta2 var(x_clear)         coef2
## 1      1 -0.8423332 0.0008130388 -0.0001065708
## 2      3  1.0294645 0.0050224574  0.0008045828
## 3      5  0.5133889 0.0372569420  0.0029764378
## 4      7  1.9793289 0.0011680694  0.0003597739
## 5      9  2.5184766 0.0036873500  0.0014450916
## 6     11 -0.3993365 0.0675365206 -0.0041968244
```

```r
    sum(m2$coef2) # с контролями
```

```
## [1] 0.2138657
```

Итак, мы видим, что подсчеты были не ошибочными.
Какие **выводы** можно сделать из них?

- Возможно, мы столкнулись с некоторым аналогом “выбросов„ в сфере анализа панельных данных. Из 90 территориальных единиц всего лишь 4 из них вносят более 79% вклада в общую оценку. Логично предположить, что такая экстраполяция может быть ошибочной.

- Даже если исключить только два округа, которые вносят максимальный (более 60%) вклад в итоговую оценку, относительное распределение “весов„ уже кажется адекватным и неплохим. Возможно, столь высокие и резкие изменения в независимой переменной в этих двух округах были вызваны ошибками в данных при их изначальном составлении. Стоит задуматься об осмысленности использования данного набора данных.

- В итоге мы выяснили, что модель, при формировании общей оценки коэф-та при предикторе “натуральный логарифм числа полицейских на душу населения„ в основном ориентируется на первые 10 округов (более 90% общей изменчивости), а по большей части — вообще на 2 округа. Это не совсем то, что мы бы хотели видеть от анализа панельных данных на основе 90 округов. Это может вызывать смещение и неверную экстраполяцию.

- Теперь понятно, почему в некоторых спецификациях оценка данного коэф-та становилась незначимой. Так же, как и очевидно, что все-таки действительно лучше использовать FE-модель на временные периоды, ведь **aggregation bias** в ней однозначно должен быть меньше.