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

Домашнее задание 2
Рубанов Владислав, БПТ 201

Задание 1

Загрузим данные, определим штат как факторную переменную и удалим пропущенные значения:

```
library(haven)
library(psych)
library(dplyr)
library(arm)
library(multilevel)
library(lattice)
library(ggplot2)
library(lmerTest)
library(influence.ME)
library(sjPlot)
library(glmmTMB)
library(GGally)
library(car)
library(plm)
library(plotly)

ME <- read_dta("RAPDC_lab3_2022.dta")

ME$state_id <- as.factor(ME$state)
ME <- na.omit(ME)
```

Далее посмотрим на **описательные статистики** имеющихся данных и их распределение:

```
# посмотрим на данные описательные статистики
head(ME)

## # A tibble: 6 x 8
##        house state lastname  votepct          party money acres state_id
##    <dbl+lbl> <chr> <chr>       <dbl>      <dbl+lbl> <dbl> <dbl> <fct>
## 1 0 [Senate] AK    Murkowski   0.842 1 [Republican]  9.17     0 AK
## 2 0 [Senate] AK    Stevens     0.846 1 [Republican]  0        0 AK
## 3 1 [House]  AK    Young       0.571 1 [Republican] 23.5      0 AK
## 4 0 [Senate] AL    Sessions    0.872 1 [Republican] -0.8      0 AL
## 5 0 [Senate] AL    Shelby      0.641 1 [Republican] 24.2      0 AL
## 6 1 [House]  AL    Aderholt    0.9   1 [Republican] 35        0 AL

describe(ME)

##           vars   n   mean     sd median trimmed    mad  min    max  range  skew
## house        1 527   0.83   0.38   1.00    0.91   0.00  0.0   1.00   1.00 -1.71
```

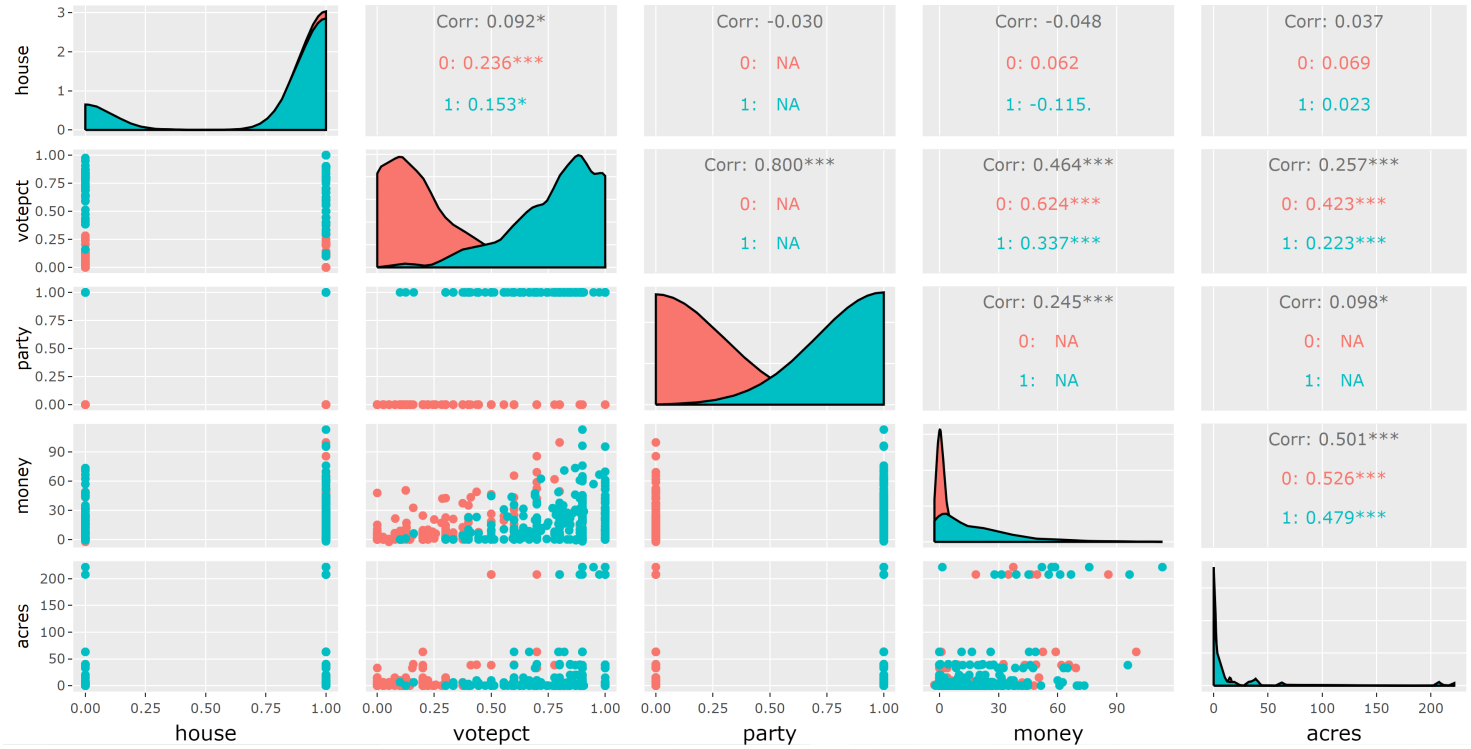
```
## state*       2 527  24.47  14.49  24.00   24.29  20.76  1.0  50.00  49.00  0.02
## lastname*    3 527 245.95 140.26 246.00  246.35 180.88  1.0 486.00 485.00 -0.01
## votepct      4 527   0.52   0.35   0.59    0.52   0.45  0.0   1.00   1.00 -0.10
## party        5 527   0.52   0.50   1.00    0.53   0.00  0.0   1.00   1.00 -0.09
## money        6 527  12.96  18.44   4.50    9.16   6.67 -2.5 113.10 115.60  1.99
## acres        7 527  14.28  41.50   0.00    4.16   0.00  0.0 221.65 221.65  4.18
## state_id*    8 527  24.47  14.49  24.00   24.29  20.76  1.0  50.00  49.00  0.02
##           kurtosis   se
## house         0.92 0.02
## state*       -1.36 0.63
## lastname*    -1.20 6.11
## votepct      -1.50 0.02
## party        -2.00 0.02
## money         4.59 0.80
## acres        17.01 1.81
## state_id*    -1.36 0.63
```

```
ggplotly(ggpairs(ME[,-c(2,3,8)],ggplot2::aes(colour=as.factor(party))))
```

```
## Error in loadNamespace(name): there is no package called 'webshot'
```

```
options(scipen = 999)
```

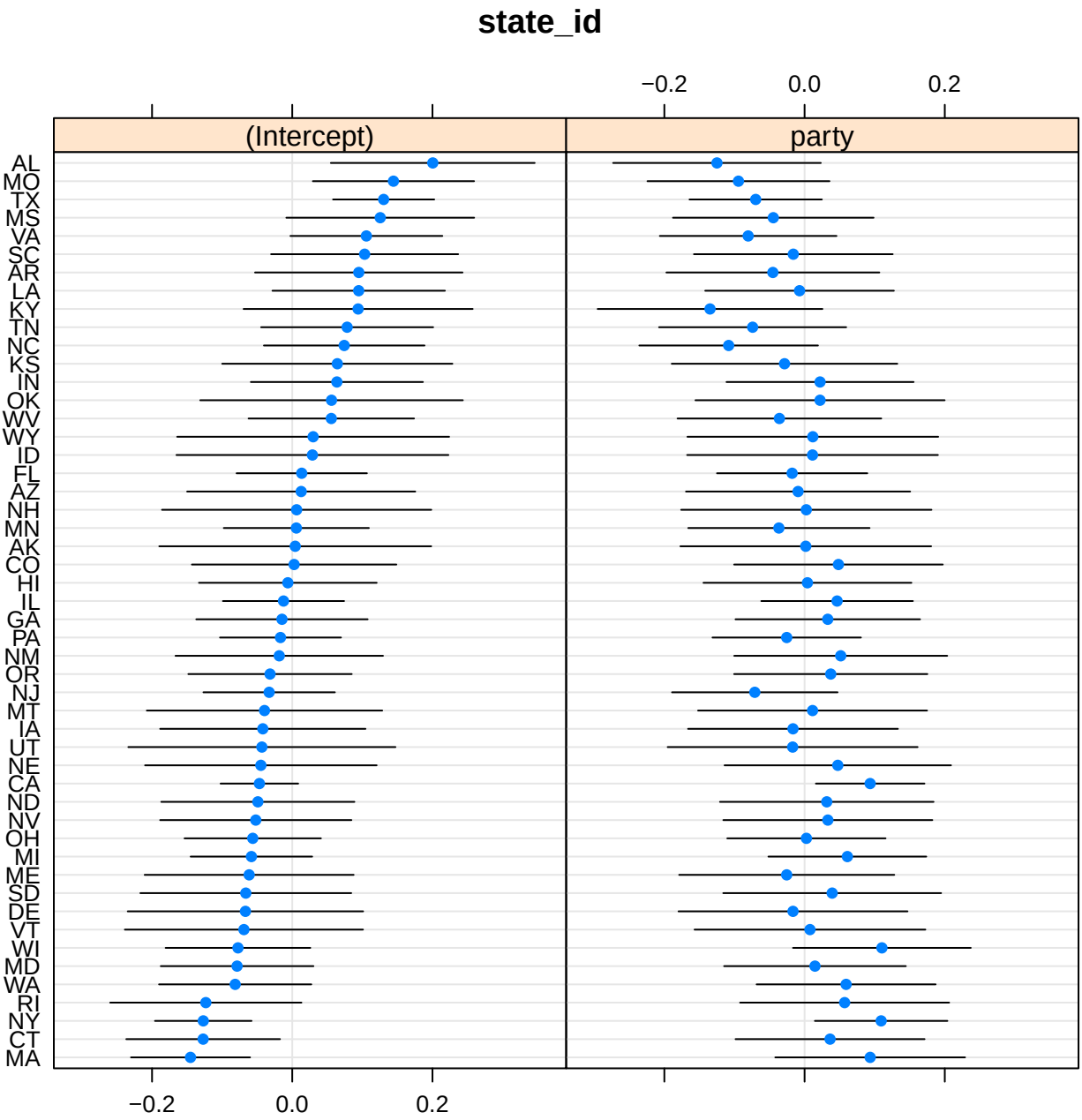
График из ggplotly не строится, поэтому я вставлю его картинкой из R:



На последнем графике можно особенно явно заметить, что **доля голосований за табачную индустрию хорошо "делится,, по политической партии**. Можно предположить, что **бОльшая часть изменчивости объясняется просто дамми-переменной на партию**. Кроме того, уже по этому графику можно заметить разницу в эффекте для зависимой переменной money — по размеру финансирования от табачных компаний — для разных партий.

Определим **модель 4.2.1**, с которой мы будем работать далее, и посмотрим на ее выдачу. Также посмотрим на выгрузку случайных эффектов в виде графика:

```
model4.2.1 <- lmer(votepct ~ money + acres + party +
    (1 + party|state_id), REML = FALSE, data = ME)
summary(model4.2.1)

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: votepct ~ money + acres + party + (1 + party | state_id)
##    Data: ME
##
##      AIC      BIC   logLik deviance df.resid
##   -312.0   -277.8    164.0   -328.0      519
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7460 -0.5764 -0.0002  0.5632  4.5800
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  state_id (Intercept) 0.010816 0.10400
##           party       0.008485 0.09211  -0.73
##  Residual             0.027284 0.16518
## Number of obs: 527, groups:  state_id, 50
##
## Fixed effects:
##                Estimate  Std. Error          df t value             Pr(>|t|)
## (Intercept)   0.2047760   0.0209332  34.7334789   9.782  0.00000000001630022
## money         0.0040650   0.0004947 510.9469193   8.217  0.00000000000000173
## acres         0.0005924   0.0003174  36.9971747   1.866                0.0699
## party         0.4882487   0.0222164  37.8990776  21.977 < 0.0000000000000002
##
## (Intercept) ***
## money       ***
## acres       .
## party       ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) money  acres
## money -0.164
## acres -0.142 -0.325
## party -0.669 -0.137 -0.001

# ranef(model4.2.1)
dotplot(ranef(model4.2.1, condVar=TRUE))

## $state_id
```

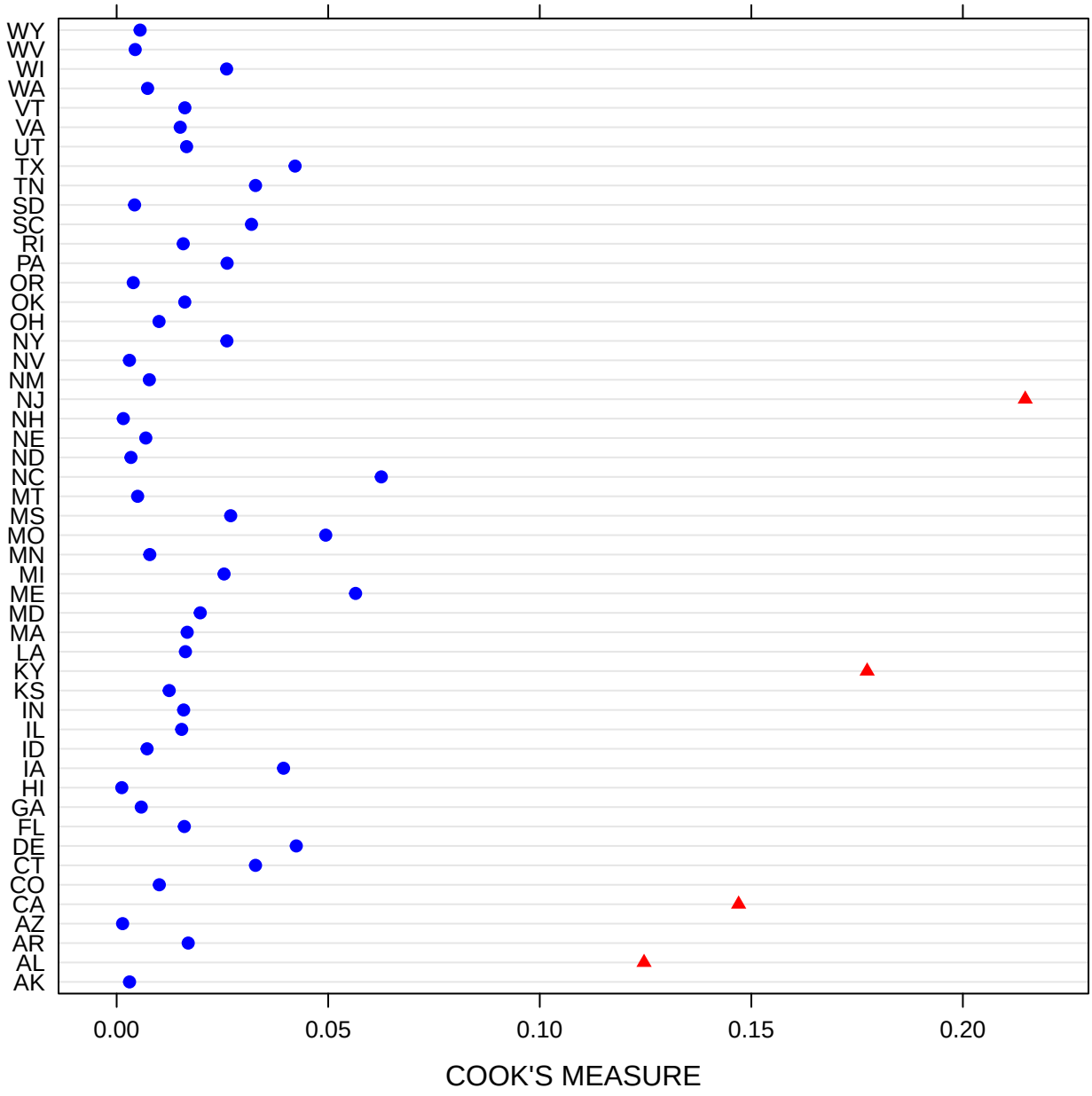


Нам необходимо поработать с влиятельными наблюдениями второго уровня в контексте ME-модели. Для этого посмотрим, какие из штатов превышают **меру Кука**:

```
# влиятельные наблюдения
state_unique <- unique(ME$state_id)
inf <- influence(model4.2.1, group = "state_id", data = ME)
cooks.distance.estex(inf, sort=TRUE)

##           [,1]
## HI 0.001217661
## AZ 0.001411733
## NH 0.001579653
## NV 0.003029974
## AK 0.003049995
## ND 0.003386937
## OR 0.003928506
```

```
## SD 0.004243326
## WV 0.004385738
## MT 0.004964767
## WY 0.005544986
## GA 0.005819324
## NE 0.006892953
## ID 0.007197633
## WA 0.007318332
## NM 0.007731376
## MN 0.007834657
## OH 0.010019280
## CO 0.010087407
## KS 0.012425073
## VA 0.015032938
## IL 0.015375128
## RI 0.015739676
## IN 0.015836567
## FL 0.015999107
## OK 0.016102770
## VT 0.016125567
## LA 0.016264229
## UT 0.016534082
## MA 0.016677455
## AR 0.016915885
## MD 0.019763735
## MI 0.025390620
## WI 0.025992456
## NY 0.026063781
## PA 0.026110747
## MS 0.026955503
## SC 0.031873206
## CT 0.032825236
## TN 0.032827785
## IA 0.039443721
## TX 0.042176282
## DE 0.042460790
## MO 0.049419741
## ME 0.056475249
## NC 0.062552613
## AL 0.124653048
## CA 0.147008538
## KY 0.177384662
## NJ 0.214744838
```

```
plot(inf, which = "cook", xlab = "COOK'S MEASURE", cutoff = 4/length(state_unique))
```




```
# AL, CA, KY, NJ
```

Итак, мы видим, что нетипичных штата *четыре*:

1. **Алабама** (AL)
2. **Калифорния** (CA)
3. **Кентукки** (KY)
4. **Нью Джерси** (NJ)

Создадим новые подвыборки поочередно без каждого из штатов, а также без всех четырех сразу:

```
# удалим влиятельные штаты
subset1 = subset(ME, ME$state_id != 'NJ')
subset2 = subset(ME, ME$state_id != 'KY')
```

```r
subset3 = subset(ME, ME$state_id != 'CA')
subset4 = subset(ME, ME$state_id != 'AL')
subset5 = subset(ME, ME$state_id != 'NJ' & ME$state_id != 'KY' & ME$state_id != 'CA' & ME$stat
```

Далее попробуем *поочередно оценить модели на полученных подвыборках* и сравнить их с моделью на полной выборке, будем двигаться от максимального отклонения (Нью Джерси) к минимальному (Алабама):

```r
model4.2.1.1 <- lmer(votepct ~ money + acres + party + (1 + party|state_id), REML = FALSE, dat
summary(model4.2.1.1)

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: votepct ~ money + acres + party + (1 + party | state_id)
##    Data: subset1
##
##      AIC      BIC   logLik deviance df.resid
##   -300.3   -266.4    158.1   -316.3      504
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6719 -0.5689 -0.0127  0.5954  4.5369
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  state_id (Intercept) 0.012252 0.11069
##           party       0.007659 0.08752  -0.86
##  Residual             0.027825 0.16681
## Number of obs: 512, groups:  state_id, 49
##
## Fixed effects:
##               Estimate  Std. Error          df t value             Pr(>|t|)
## (Intercept)   0.2084016   0.0220863  33.7609315   9.436  0.00000000005418023
## money         0.0039888   0.0004972 494.4575573   8.023  0.00000000000000754
## acres         0.0004068   0.0002957  32.8228404   1.376                0.178
## party         0.4968166   0.0220777  35.3949018  22.503 < 0.0000000000000002
##
## (Intercept) ***
## money       ***
## acres
## party       ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) money  acres
## money -0.160
## acres -0.124 -0.342
## party -0.728 -0.143  0.008
```

Можно заметить, что **оценки модели без Нью Джерси не изменились** значимым образом, но

корреляция между случайными эффектами на константу и партию **стала сильнее** на 0,13 (всего -0,86).

```
model4.2.1.2 <- lmer(votepct ~ money + acres + party + (1 + party|state_id), REML = FALSE, dat
summary(model4.2.1.2)

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: votepct ~ money + acres + party + (1 + party | state_id)
##    Data: subset2
##
##      AIC      BIC   logLik deviance df.resid
##   -316.6   -282.5    166.3   -332.6      512
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7279 -0.6015 -0.0033  0.5595  4.5533
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  state_id (Intercept) 0.009548 0.09771
##           party       0.006372 0.07983  -0.68
##  Residual             0.027044 0.16445
## Number of obs: 520, groups:  state_id, 49
##
## Fixed effects:
##                Estimate  Std. Error          df t value             Pr(>|t|)
## (Intercept)   0.1957927   0.0202234  38.0337250   9.681     0.00000000000825
## money         0.0044102   0.0005064 507.0401085   8.709 < 0.0000000000000002
## acres         0.0006395   0.0004113  31.3540126   1.555                 0.13
## party         0.4940461   0.0209990  41.9359905  23.527 < 0.0000000000000002
##
## (Intercept) ***
## money       ***
## acres
## party       ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) money  acres
## money -0.174
## acres -0.167 -0.266
## party -0.637 -0.145  0.010
```

Оценки без Кентукки не изменились значимым образом, но корреляция между случайными эффектами на константу и партию стала слабее на 0,05.

```
model4.2.1.3 <- lmer(votepct ~ money + acres + party + (1 + party|state_id), REML = FALSE, dat
summary(model4.2.1.3)

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
```

```
##   method [lmerModLmerTest]
## Formula: votepct ~ money + acres + party + (1 + party | state_id)
##    Data: subset3
##
##      AIC      BIC   logLik deviance df.resid
##   -301.2   -268.0    158.6   -317.2      464
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3148 -0.5910  0.0016  0.5656  3.8123
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  state_id (Intercept) 0.012185 0.11038
##           party       0.008896 0.09432  -0.72
##  Residual             0.025287 0.15902
## Number of obs: 472, groups:  state_id, 49
##
## Fixed effects:
##                Estimate  Std. Error          df t value             Pr(>|t|)
## (Intercept)   0.2107711   0.0218520  35.5109802   9.645    0.000000000018610
## money         0.0037409   0.0004899 450.4468054   7.636    0.000000000000135
## acres         0.0006656   0.0003287  35.9919591   2.025                0.0503
## party         0.4819649   0.0226145  35.6339719  21.312 < 0.0000000000000002
##
## (Intercept) ***
## money       ***
## acres       .
## party       ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) money  acres
## money -0.159
## acres -0.153 -0.309
## party -0.665 -0.129 -0.006
```

Оценки и корреляция без Калифорнии **не изменились значимым образом**.

```
model4.2.1.4 <- lmer(votepct ~ money + acres + party + (1 + party|state_id), REML = FALSE, dat
summary(model4.2.1.4)

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: votepct ~ money + acres + party + (1 + party | state_id)
##    Data: subset4
##
##      AIC      BIC   logLik deviance df.resid
##   -313.4   -279.4    164.7   -329.4      510
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7717 -0.5848 -0.0029  0.5527  4.5945
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  state_id (Intercept) 0.008067 0.08982
##           party       0.007895 0.08886  -0.62
##  Residual             0.027085 0.16458
## Number of obs: 518, groups:  state_id, 49
##
## Fixed effects:
##                Estimate  Std. Error          df t value             Pr(>|t|)
## (Intercept)   0.1936615   0.0193344  37.0602627  10.016  0.00000000000430880
## money         0.0040935   0.0004969 504.9971051   8.238  0.00000000000000152
## acres         0.0007429   0.0003165  38.0126729   2.347               0.0242
## party         0.4937753   0.0219781  38.0925978  22.467 < 0.0000000000000002
##
## (Intercept) ***
## money       ***
## acres       *
## party       ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) money  acres
## money -0.170
## acres -0.152 -0.333
## party -0.614 -0.142 -0.004
```

Оценки без Алабамы **не изменились значимым образом**, при этом коэффициент при предикторе “акры„ стали значимыми на уровне значимости 0.05, а корреляция между случайными эффектами на константу и партию стала слабее на 0.1 (-0.62 в целом).

```
model4.2.1.5 <- lmer(votepct ~ money + acres + party + (1 + party|state_id), REML = FALSE, dat
summary(model4.2.1.5)

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: votepct ~ money + acres + party + (1 + party | state_id)
##    Data: subset5
##
##      AIC      BIC   logLik deviance df.resid
##   -294.2   -261.4    155.1   -310.2      433
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3796 -0.5899 -0.0077  0.5981  3.8076
##
## Random effects:
```

```
##  Groups   Name        Variance Std.Dev. Corr
##  state_id (Intercept) 0.009203 0.09593
##           party       0.004646 0.06816  -0.70
##  Residual             0.025295 0.15904
## Number of obs: 441, groups:  state_id, 46
##
## Fixed effects:
##                Estimate  Std. Error          df t value             Pr(>|t|)
## (Intercept)   0.1934464   0.0205746  40.0351262   9.402   0.0000000000109835
## money         0.0040424   0.0005067 424.9530412   7.979   0.0000000000000139
## acres         0.0006845   0.0004047  26.8469550   1.691                0.102
## party         0.5016881   0.0205482  36.6895043  24.415 < 0.0000000000000002
##
## (Intercept) ***
## money       ***
## acres
## party       ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) money  acres
## money -0.172
## acres -0.175 -0.266
## party -0.630 -0.153  0.006
```

Можно утверждать, что без всех четырех нетипичных штатов **оценки модели и корреляция между случайными эффектами на константу и партию не изменились значимым образом**. **Значит, эти нетипичные наблюдения на втором уровне не вносят смещение в нашу модель.**

Задание 2

Очевидно, что в моделях со смешанными эффектами методика работы в со влиятельными наблюдениями отличается от работы с ними в других моделях. Учитывая **иерархическую структуру данных**, сейчас штаты включают в себя некоторое (разное) количество наблюдений, нужно исходить из этого. При этом наша выборка не является сбалансированной, это тоже нужно иметь в виду.

Я бы предложил разделить возможные методы работы со влиятельными наблюдениями на две части: непосредственно на работу с данными и теоретическую работу:

Работа с данными:

- сбалансировать панель (однако в Калифорнии и так было 55 конгрессменов, для этого штата такой вариант явно не подойдет, к тому же наша выборка обладает определенной временной спецификой: нельзя взять конгрессменов там, где их не было);
- увеличение выборки на втором уровне; учитывая тот факт, что число штатов фиксированное, возможно, можно было бы перейти на другой, более мелкий, второй уровень: например, на уровень районов;
- преобразовать данные так, чтобы отклонения не выбивались сильно по абсолютному значению и не способствовали смещению и переобучению:
  - центрировать и нормировать (привести к параметрам среднее $= 0$, стандартное отклонение $= 1$);

– логарифмировать (заодно попробовать приблизить к нормальному виду);

– использовать нелинейное квантильное преобразование (QuantileTransformer в Python) к стандартному нормальному распределению.

Поработать над теорией:

- улучшить спецификацию: возможно, имеют место значимые пропущенные факторы (переменные);
- попытаться объяснить на теоретическом уровне, почему, возможно, такой результат может являться приемлемым и/или ожидаемым (особенности/специфика штата): возможно, наделение некоторой единицы бОльшим “весом„ является оправданным.

Задание 3

Проверим, можно ли говорить о **нормальности распределения** ошибок первого уровня и случайных эффектов, с помощью формальных тестов.

Для этого в начале получим остатки первого уровня, а также случайные эффекты для константы и партии:

```
# сохраним предсказанные значения (fixed + random part)
pr_re <- predict(model4.2.1)
resid <- ME$votepct - pr_re

# сохраним случайные эффекты
raf <- ranef(model4.2.1)
raf_Intercept <- raf$state_id[,1]
raf_party <-raf$state_id[,2]
```

Предлагаю в начале использовать классический тест для этой задачи — **тест Шапиро-Уилка** на нормальность. Нулевая гипотеза для него означает, что данные имеют нормальное распределение. Также стоит отметить, что он чувствителен к размеру выборки: при большом n может быть сложно отвергнуть нулевую гипотезу.

```
# H0 : the sample is normally distributed
shapiro.test(resid)

##
##  Shapiro-Wilk normality test
##
## data:  resid
## W = 0.98217, p-value = 0.000004691

shapiro.test(raf_Intercept)

##
##  Shapiro-Wilk normality test
##
## data:  raf_Intercept
## W = 0.97129, p-value = 0.261

shapiro.test(raf_party)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  raf_party
## W = 0.97903, p-value = 0.512
```

Итак, можно заметить, что, согласно тесту Шапиро-Уилка, говорить о нормальности распределения можно *только для случайных эффектов* на константу и политическую партию (на большом уровне доверия). Для остатков же мы отвергаем нулевую гипотезу о нормальности распределения на любом уровне значимости.

При этом стоит отметить **разницу в числе наблюдений**, из-за которых могли быть получены такие противоречивые результаты: 527 остатков и по 50 значений для случайных эффектов (т.к. они считаются для каждого штата).

Далее посмотрим на еще один популярный критерий для определения нормальности распределения данных — **тест Колмогорова-Смирнова**. Он, наоборот, свидетельствует о ненормальности распределения данных при верной нулевой гипотезе и говорит об обратном при отвержении нулевой гипотезы в пользу альтернативы.

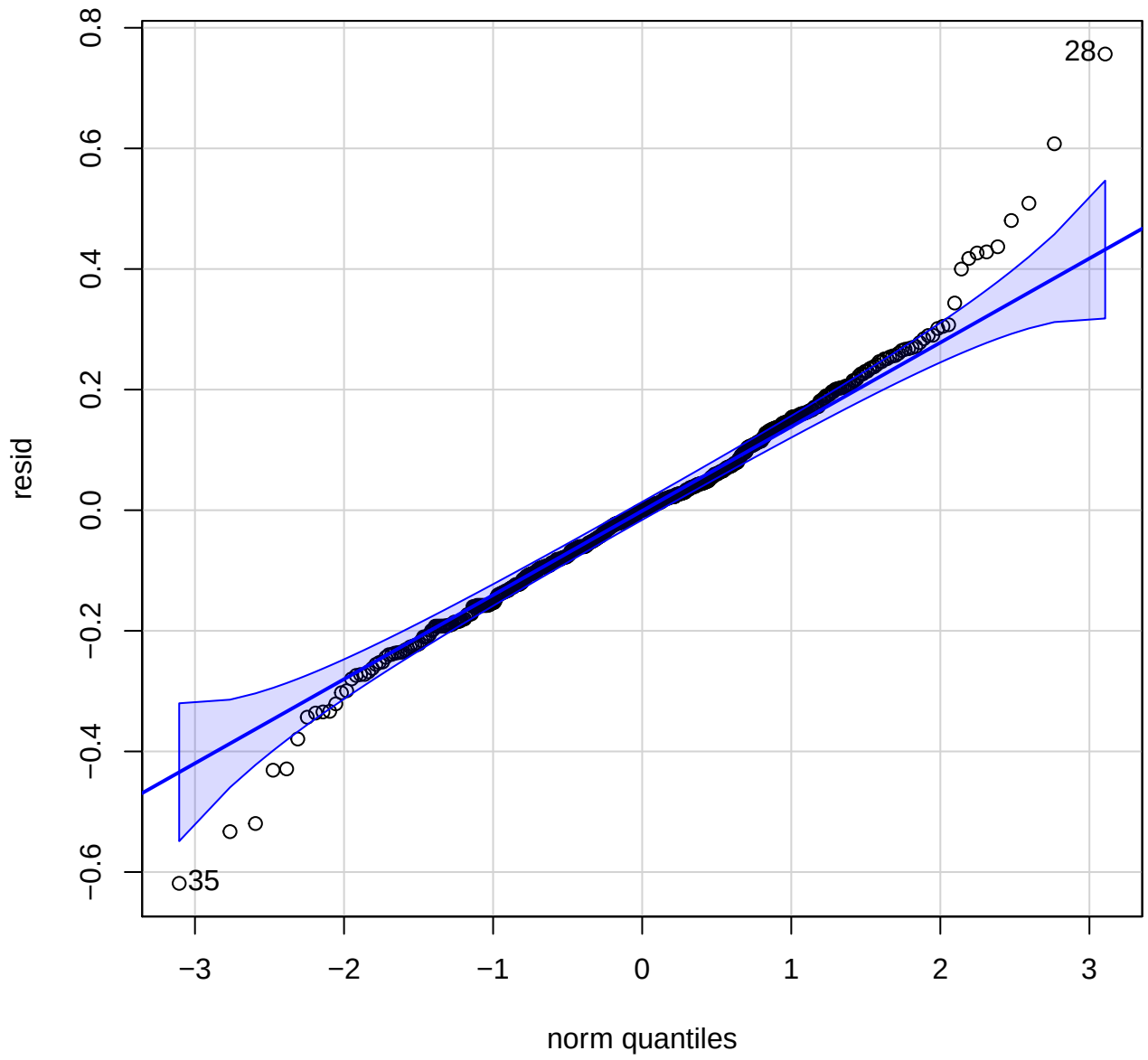
```
# H0 : the sample is not normally distributed
ks.test(unique(resid), "pnorm")

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  unique(resid)
## D = 0.36312, p-value < 0.00000000000000022
## alternative hypothesis: two-sided

length(resid) - length(unique(resid))

## [1] 43

ks.test(raf_Intercept, "pnorm")

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  raf_Intercept
## D = 0.44231, p-value = 0.000000001981
## alternative hypothesis: two-sided

ks.test(raf_party, "pnorm")

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  raf_party
## D = 0.4561, p-value = 0.0000000005008
## alternative hypothesis: two-sided
```

Стоит отметить, что данный тест требует на вход только уникальные значения, поэтому выборку для остатков пришлось немного сократить (на 43 значения остатка из 527).

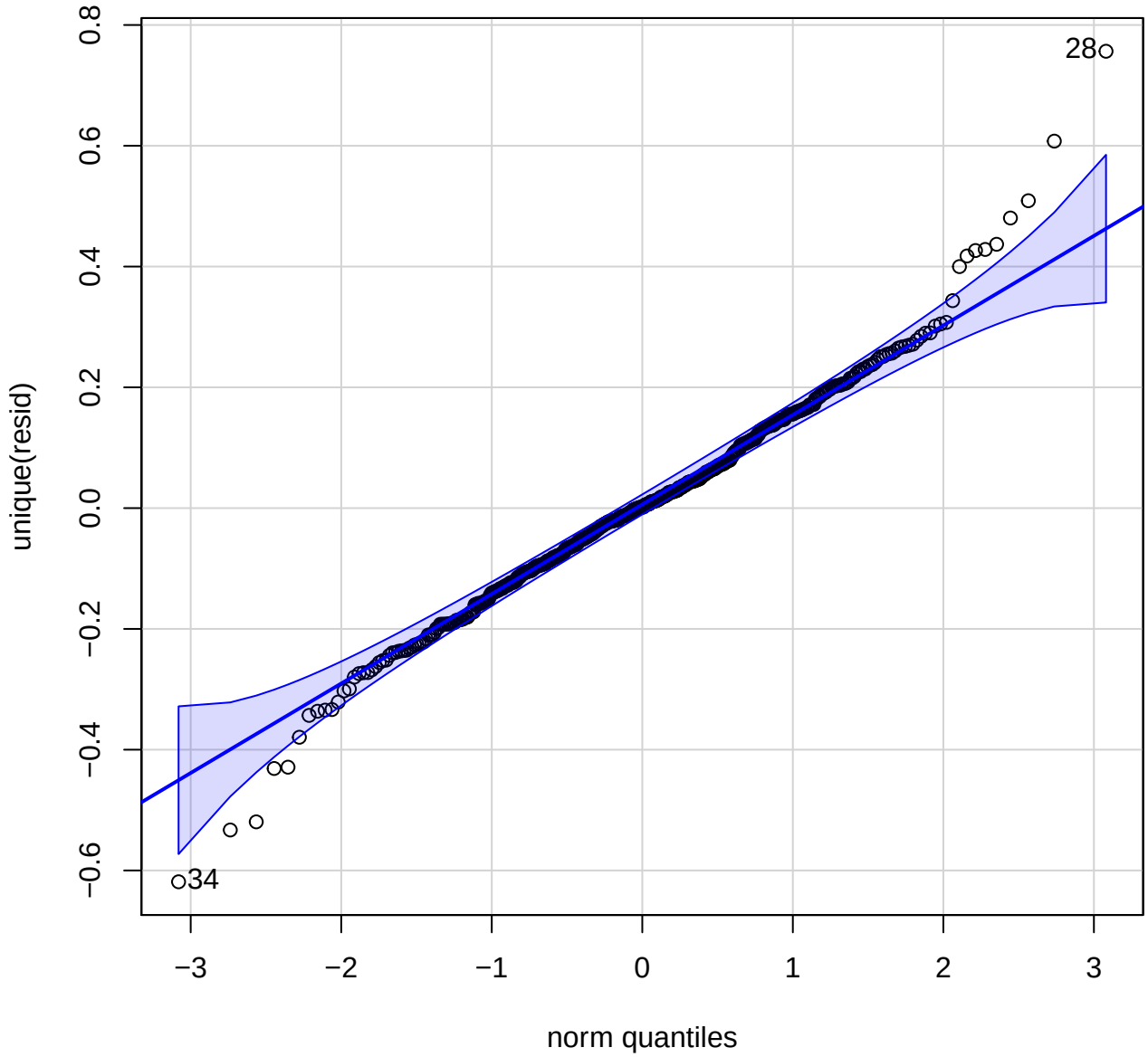
Как можно заметить, как для остатков первого уровня, так и для обоих случайных эффектов мы можем **отвергнуть нулевую гипотезу** на любом адекватном уровне значимости. Это говорит о том, что они прошли данный тест, мы можем говорить о нормальности распределения остатков первого уровня и случайных эффектов, согласно тесту Колмогорова-Смирнова.

Далее хотелось бы дополнительно привести небольшую ***визуализацию*** данных величин, чтобы иметь некоторое представление об их связи с нормальным распределением. В начале я представлю так называемый **Q-Q Plot**, а затем — **гистограммы** для них.

```
# визуализация
# qqplot
qqPlot(resid)
```

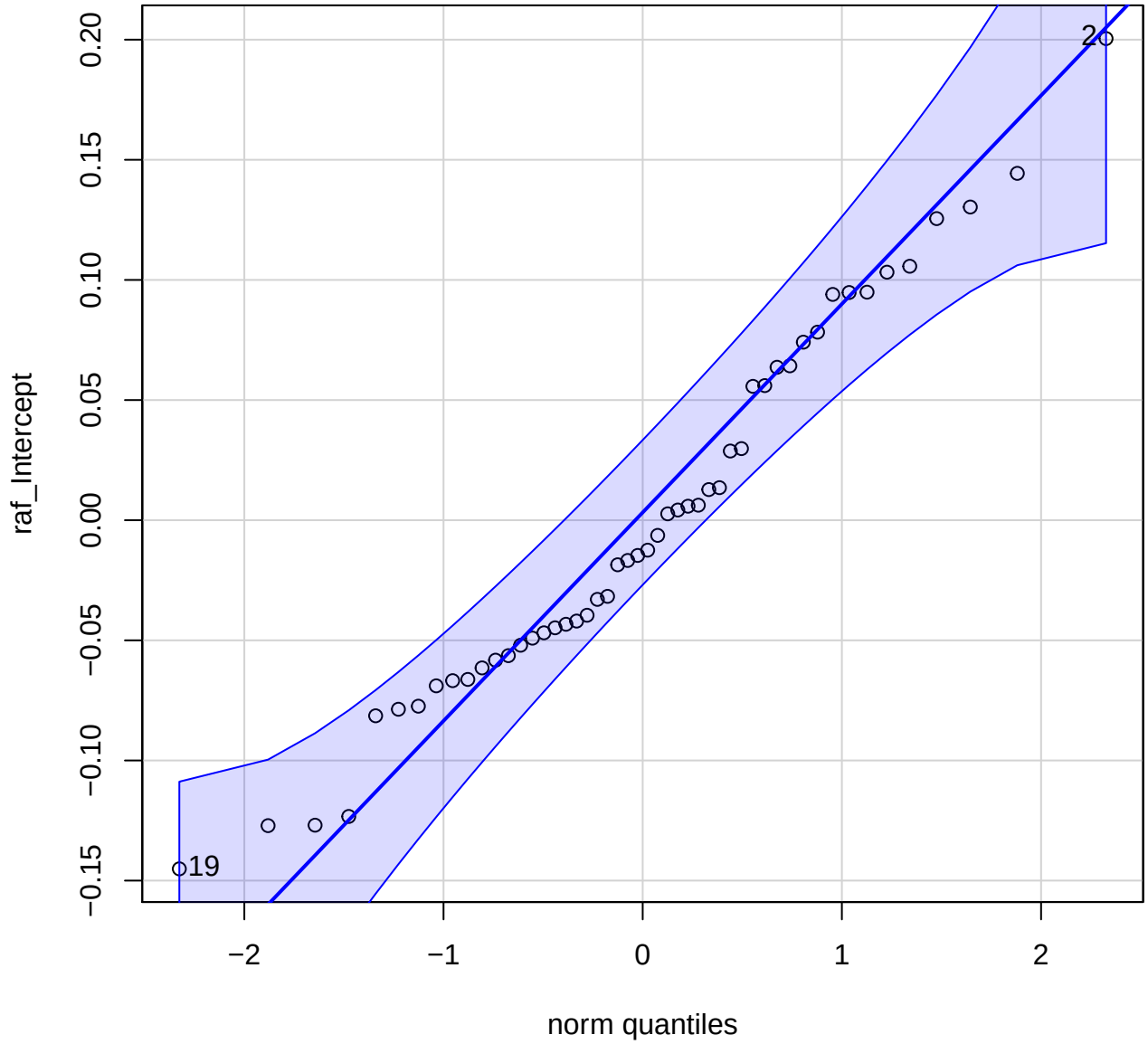



```
## [1] 28 35

qqPlot(unique(resid))
```



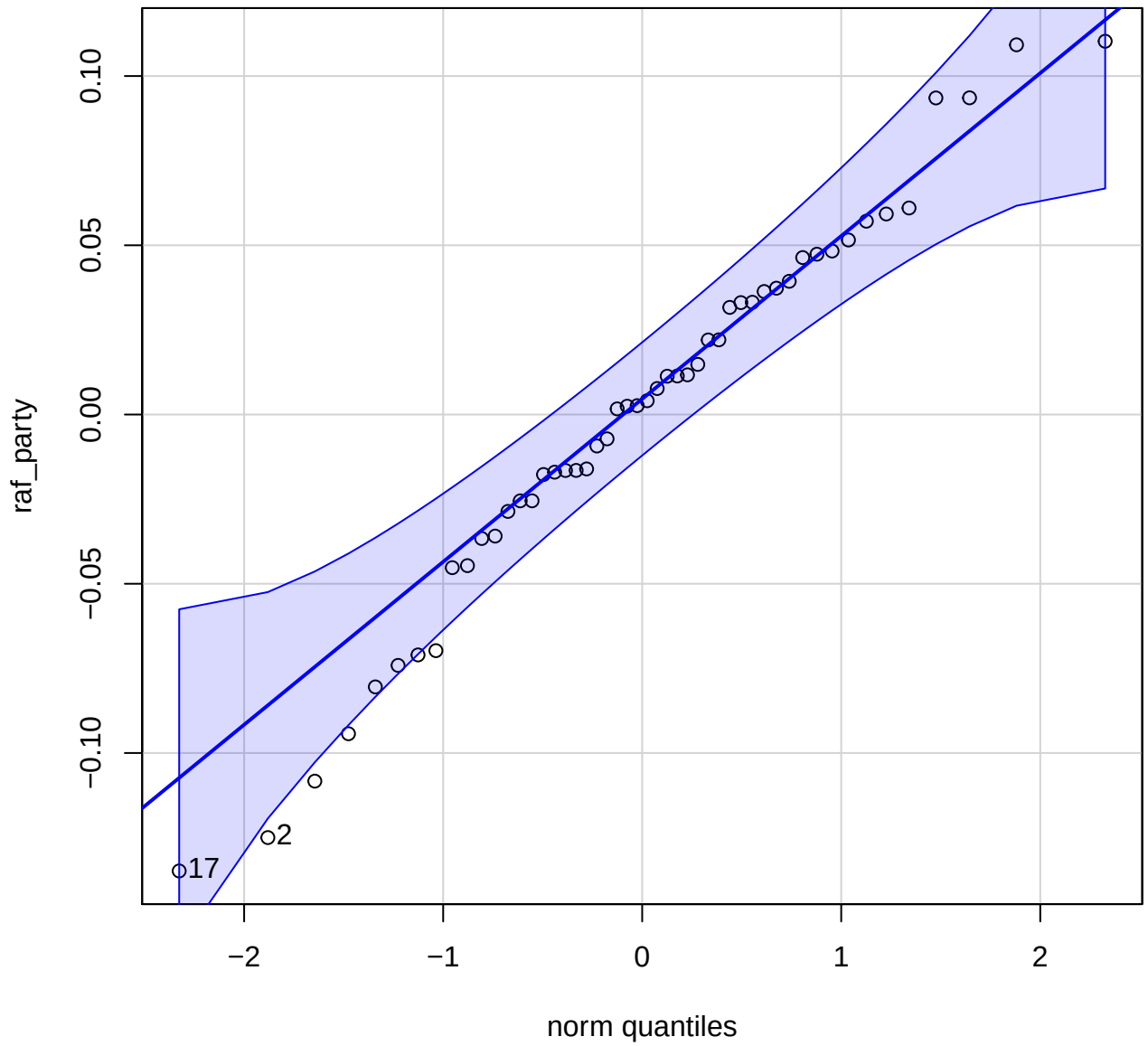

```
## [1] 28 34
```

```
qqPlot(raf_Intercept)
```




```
## [1]  2 19
```

```
qqPlot(raf_party)
```



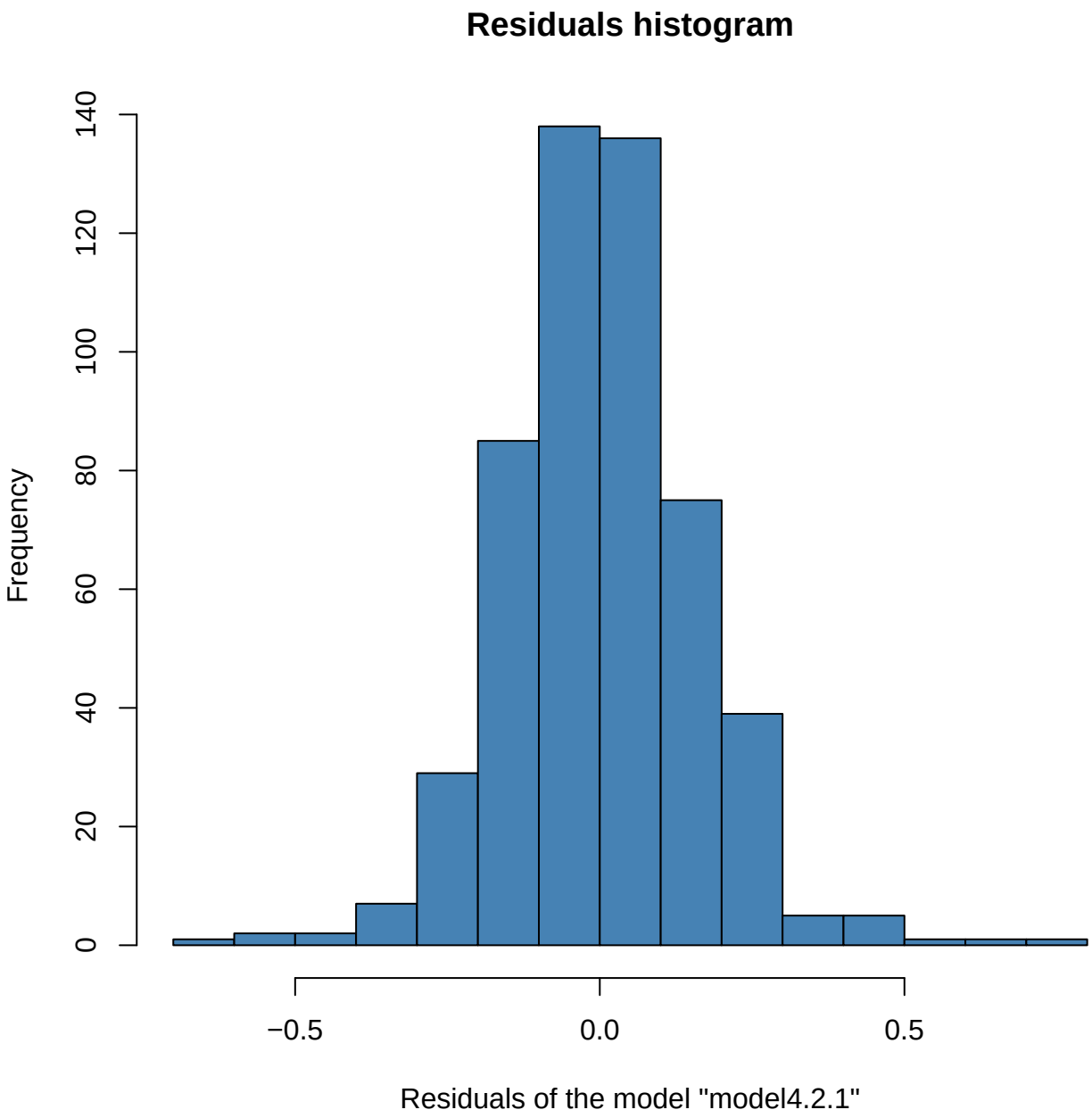

```
## [1] 17  2
```

Говоря о **Q-Q Plot**, важно, чтобы все точки (по крайней мере, близко к середине графика) лежали в пределах линии под 45 градусов или в рамках построенного доверительного интервала. Это свидетельствует о нормальности распределения данной величины.

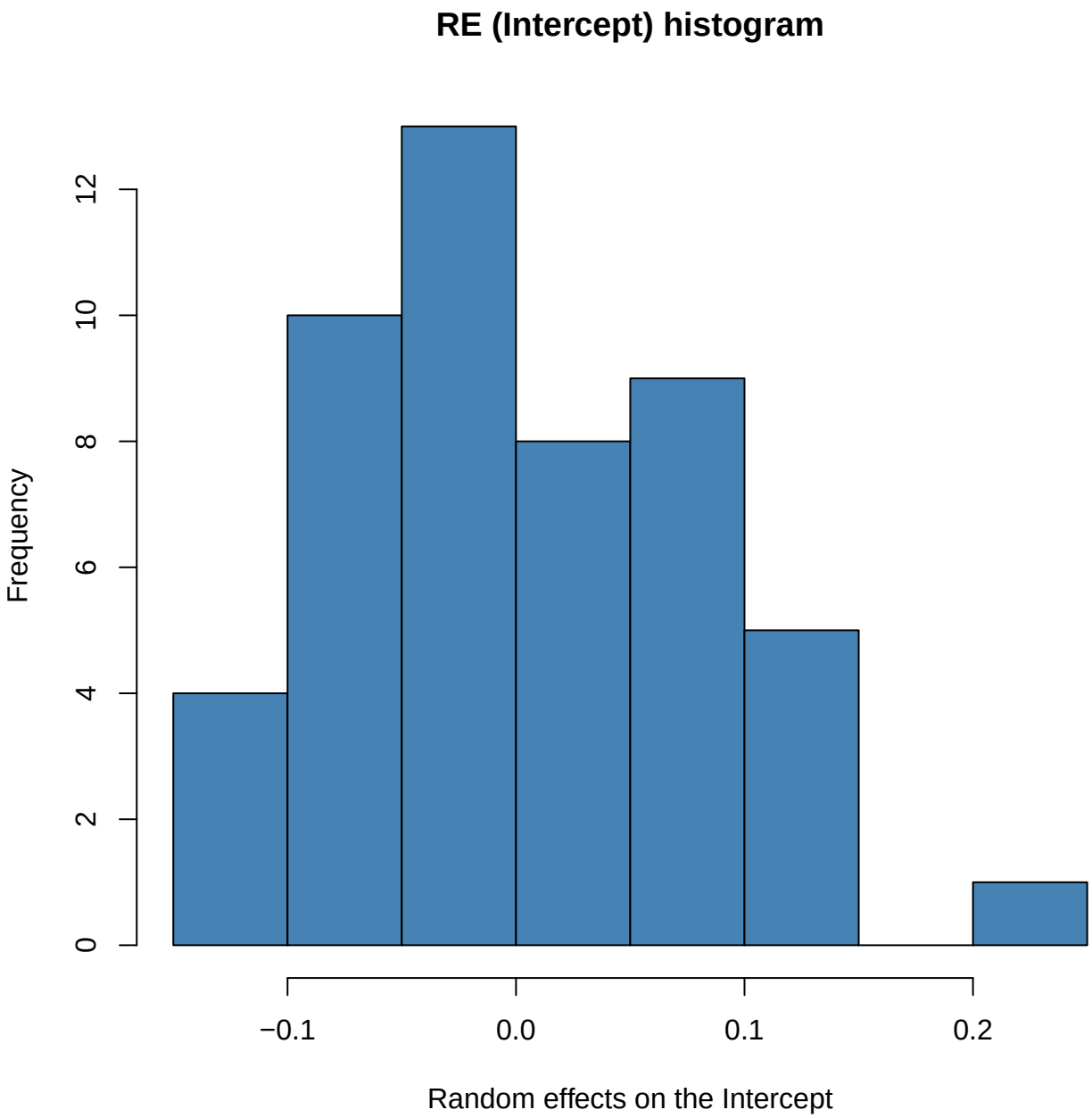
Можно сказать, что *все распределения визуально похожи на нормальные*. Также на данных графиках особенно заметна разница в числе наблюдений. Однако в некоторых случаях (на концах распределения) точки выходят за пределы ДИ, но в этом нет ничего страшного.

Далее посмотрим на гистограммы:

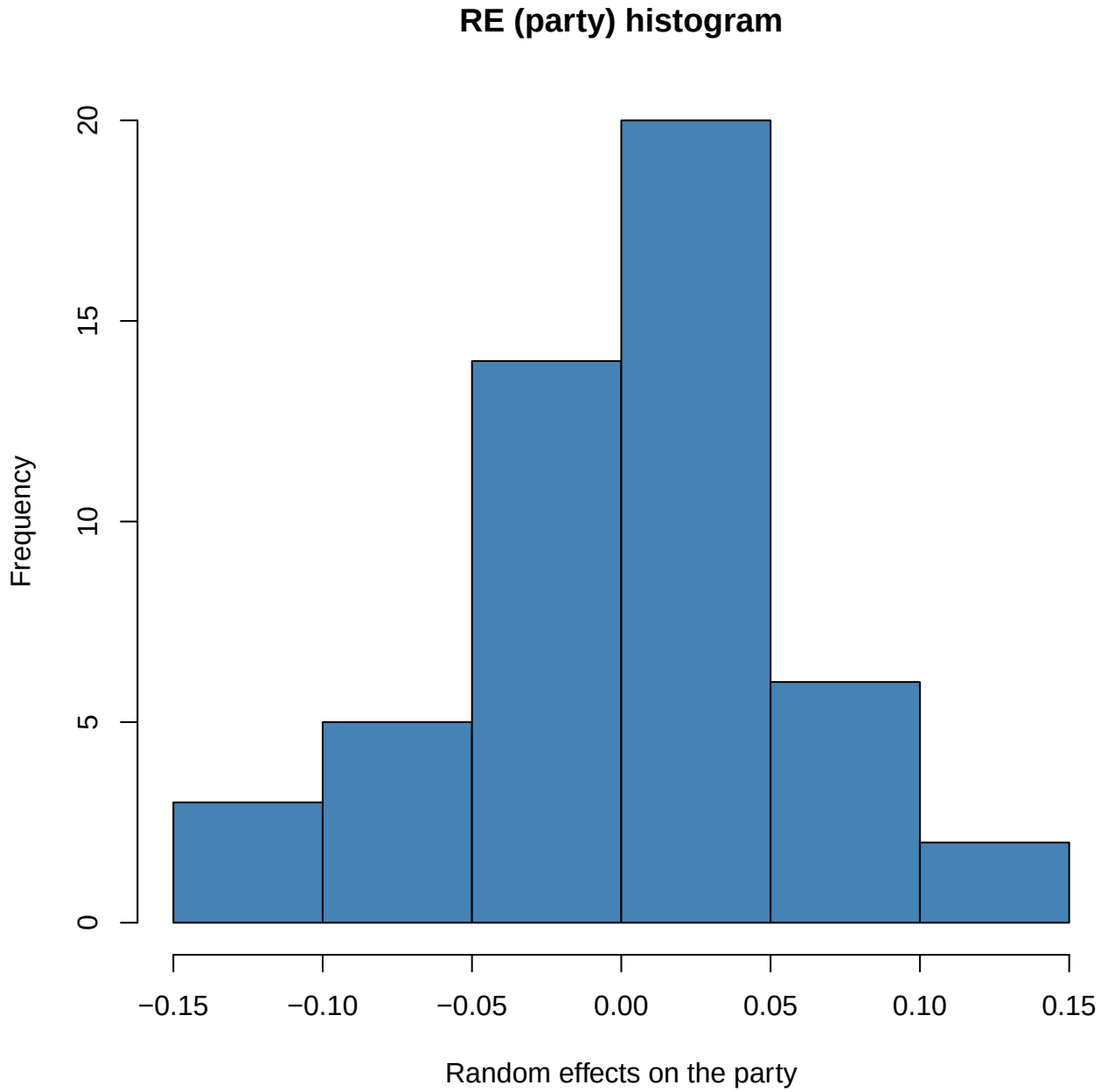
```
# гистограммы
hist(resid, col='steelblue', main='Residuals histogram', breaks=15,
     xlab='Residuals of the model "model4.2.1"')
```




```
hist(raf_Intercept, col='steelblue', main='RE (Intercept) histogram', breaks=8,
     xlab='Random effects on the Intercept')
```




```
hist(raf_party, col='steelblue', main='RE (party) histogram', breaks=8,
     xlab='Random effects on the party')
```



Для остатков первого уровня можно говорить о гистограмме, напоминающей *идеальное нормальное распределение*. Для случайных же эффектов сказывается небольшое число наблюдений — в целом они напоминают нормальное распределение, но видны некоторые отклонения (хотя случайные эффекты и прошли оба теста, в отличие от остатков первого уровня).

Говоря о ***преобразованиях***, которые можно было бы выполнить для большего приближения к нормальному распределению, я бы выделил следующие методы:

- логарифмирование переменных;
- возведение в квадратный или кубический корень[1];
- использовать нелинейное квантильное преобразование по типу QuantileTransformer в Python, о котором говорилось ранее.

[1] www.statology.org/test-for-normality-in-r/

Задание 4

В начале сделаем преобразование данных одним из базовых способов — **логарифмированием**, а именно логарифмируем следующие зависимые и независимую переменные:

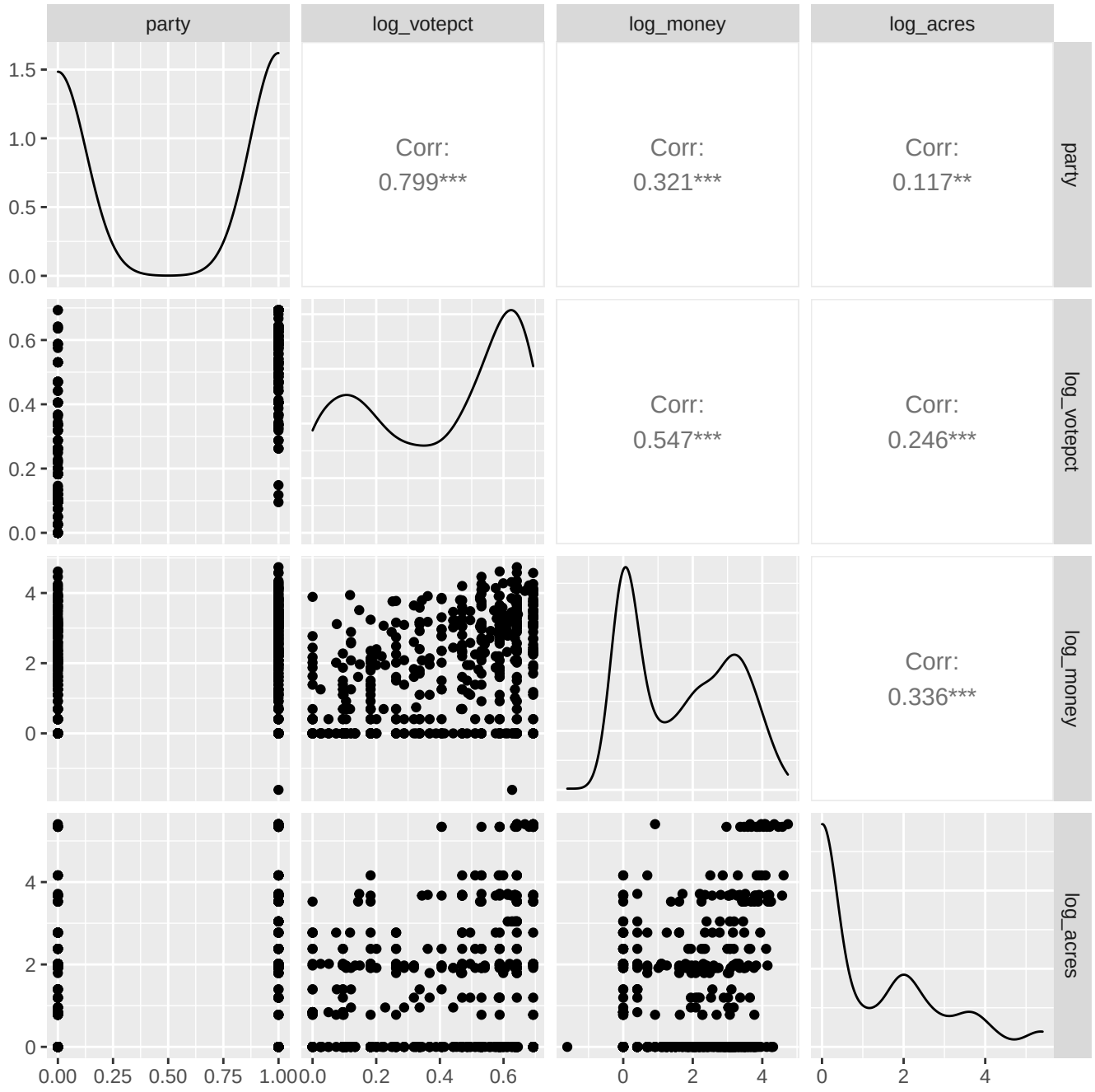
- *votepct* — доля голосов, отданных членом Конгресса в поддержку табачной индустрии (зависимая переменная);
- *money* — размер финансирования от PAC табачных корпораций (независимая переменная);
- *acres* — площадь табачных плантаций (независимая переменная).

Также, учитывая, что среди значений данных переменных присутствуют нулевые значения, я воспользуюсь функцией *log1p* в R. Кроме того, вместо двух отрицательных значении для money (связанных с некоторыми компенсациями) я возьму значение 0. Таким образом, их относительное ранжирование между друг другом сохранится.

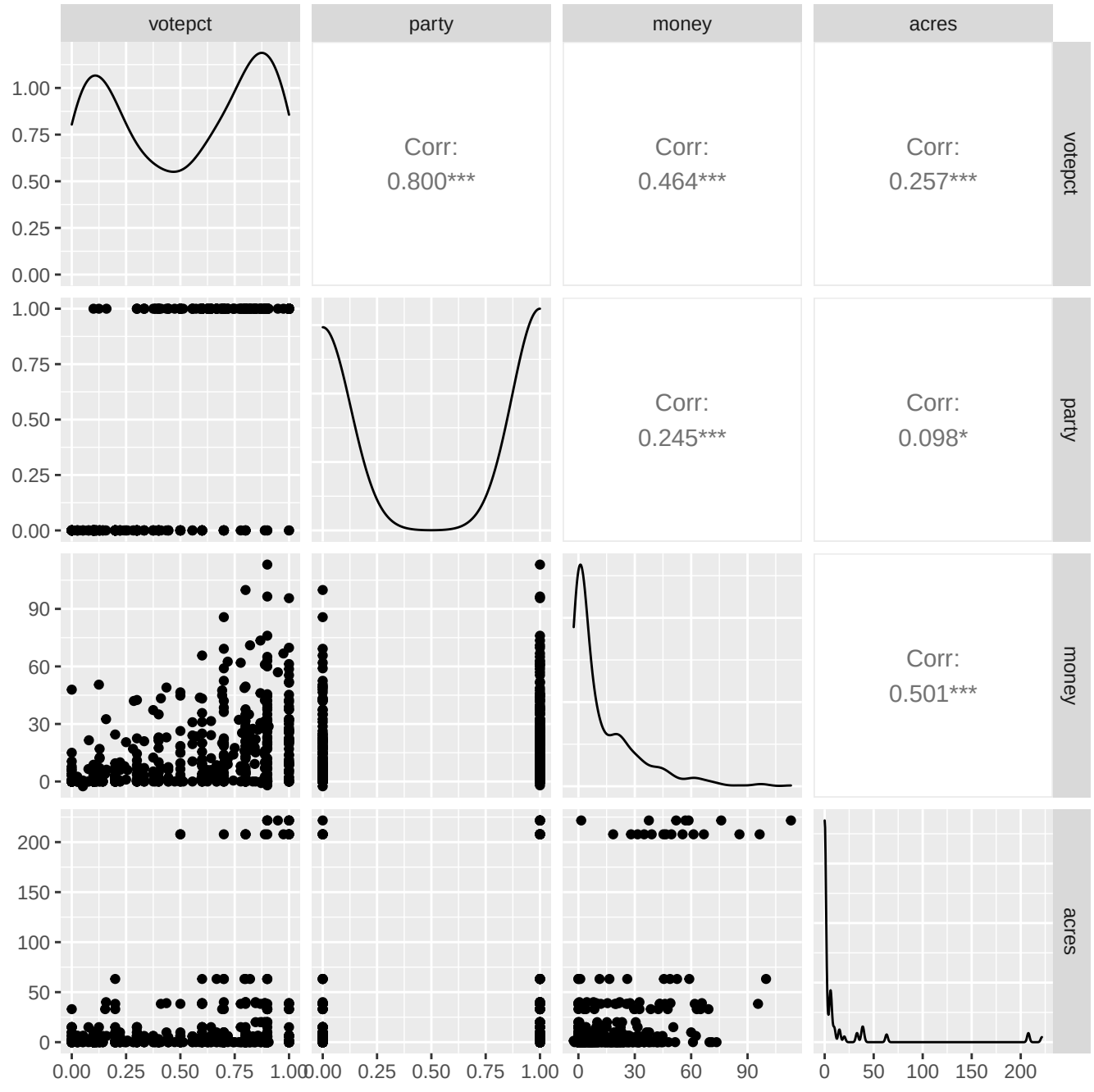
```
# преобразование
ME$log_votepct <- log1p(ME$votepct)
ME$log_money <- log1p(ME$money)
# меняем NaN на 0, т.к. они были близки к нему
ME$log_money[85] = 0
ME$log_money[208] = 0
# проверяем
ME$log_money[is.nan(ME$log_money)]

## numeric(0)

ME$log_acres <- log1p(ME$acres)
```

Далее **посмотрим на распределение** интересующих переменных до логарифмирования и после:

```
# посмотрим на распределение до логарифмирования
ggpairs(ME[,-c(1:4, 6:8)])
```




```
# посмотрим на распределение после логарифмирования
ggpairs(ME[,-c(1:3, 8:11)])
```



Видно, что *распределение не стало нормальным*. Этого сложно добиться, учитывая, что данных относительно немного, и они имеют сильно смещенное распределение — но стало лучше.

Кроме того, можно заметить интересную **особенность** — дамми-переменная party имеет **очень высокую корреляцию (0.8) с зависимой переменной votepct**.

Далее **протестируем две модели**: **базовую**, включающую только независимые переменные *money*, *party* (помним, что в качестве 0 закодирована Демократическая партия, а в качестве 1 — Республиканская) и их взаимодействие, т.е. разный эффект. Будем отталкиваться от нее.

Также оценим аналогичную **модель**, но с дополнительной зависимой переменной *acres*. Авторы предоставили достаточно теоретических обоснований для ее включения, по крайней мере, в качестве контрольной переменной.

Таким образом, моя модель будет включать **фиксированные эффекты** на уровень финансирования от табачных компаний, политическую партию (Демократическая или Республиканская) и площадь табачных полей в штате конгрессмена.

Что касается **случайных эффектов**, я считаю, что нам следует ограничиться включением случайных эффектов на константу и партию. Я не считаю, что включение случайного эффекта на уровень

финансирования имеет смысл, т.к., как было показано во время практикума на занятии, нам недостаточно информации (вариации) для идентификации данного эффекта, поскольку он слабо отличается от среднего по каждой из партий. Кроме того, включение случайного эффекта на площадь табачных полей в штате явно является бессмысленной операцией, поскольку данный показатель изменяется только по штатам: у него нет внутригрупповой изменчивости на данный момент. Также я считаю, что включение случайного эффекта на партию необходимо, поскольку это соответствует проверяемой содержательной гипотезе.

```
# тестируем модели
model5.0 <- lmer(votepct ~ money + party + party*money + (1 + party|state_id), REML = FALSE, c
summary(model5.0)

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: votepct ~ money + party + party * money + (1 + party | state_id)
##    Data: ME
##
##      AIC      BIC   logLik deviance df.resid
##   -336.5   -302.4    176.3   -352.5      519
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8890 -0.5784  0.0147  0.5492  4.2160
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  state_id (Intercept) 0.007541 0.08684
##           party       0.003853 0.06207  -0.55
##  Residual             0.026541 0.16291
## Number of obs: 527, groups:  state_id, 50
##
## Fixed effects:
##                Estimate  Std. Error          df t value             Pr(>|t|)
## (Intercept)   0.1747376   0.0192033  39.7562878   9.099      0.0000000000289
## money         0.0077697   0.0007552 282.2146651  10.288 < 0.0000000000000002
## party         0.5520859   0.0221020  55.1256564  24.979 < 0.0000000000000002
## money:party  -0.0050740   0.0009108 331.8919413  -5.571      0.0000000524171
##
## (Intercept) ***
## money       ***
## party       ***
## money:party ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) money  party
## money       -0.381
## party       -0.628  0.318
## money:party  0.300 -0.795 -0.510

model5.1 <- lmer(votepct ~ money + party + party*money + acres + (1 + party|state_id), REML =
```

```
summary(model5.1)

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: votepct ~ money + party + party * money + acres + (1 + party |
##     state_id)
##    Data: ME
##
##      AIC      BIC   logLik deviance df.resid
##   -339.3   -300.9    178.7   -357.3      518
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9260 -0.5758  0.0207  0.5401  4.2622
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  state_id (Intercept) 0.006485 0.08053
##           party       0.004185 0.06469  -0.50
##  Residual             0.026388 0.16244
## Number of obs: 527, groups:  state_id, 50
##
## Fixed effects:
##                Estimate  Std. Error          df t value             Pr(>|t|)
## (Intercept)   0.1669402   0.0187190  41.4971297   8.918      0.0000000000339
## money         0.0075145   0.0007644 344.1623042   9.831 < 0.0000000000000002
## party         0.5533958   0.0222196  54.9466972  24.906 < 0.0000000000000002
## acres         0.0006932   0.0003106  40.9922284   2.232               0.0312
## money:party  -0.0051575   0.0009084 334.8866691  -5.678      0.0000000296446
##
## (Intercept) ***
## money       ***
## party       ***
## acres       *
## money:party ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) money  party  acres
## money       -0.346
## party       -0.614  0.307
## acres       -0.153 -0.206  0.001
## money:party  0.304 -0.776 -0.506 -0.003
```

Учитывая, что модели являются **вложенными** — model5.1 может быть получена из model5.0 посредством добавления нового параметра на втором уровне, их можно **сравнить** напрямую в R:

```
# сравним
anova(model5.0, model5.1)

## Data: ME
```

```
## Models:
## model5.0: votepct ~ money + party + party * money + (1 + party | state_id)
## model5.1: votepct ~ money + party + party * money + acres + (1 + party | state_id)
##          npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## model5.0    8 -336.54 -302.40 176.27  -352.54
## model5.1    9 -339.33 -300.92 178.66  -357.33 4.7863  1    0.02869 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

"Штраф„ BIC у второй модели (model5.1) несколько меньше, также можно заметить, что значение p-value ниже конвенционального уровня 0.05, следовательно, мы **можем выбрать более полную спецификацию**: она дает реальное приращение в объясненной информации.

Далее оценим такую же спецификацию, но используя логарифмированные переменные:

```
# добавим логарифмированные переменные
model5.2 <- lmer(log_votepct ~ log_money + party + party*log_money + log_acres + (1 + party|st
summary(model5.2)

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: log_votepct ~ log_money + party + party * log_money + log_acres +
##     (1 + party | state_id)
##    Data: ME
##
##      AIC      BIC   logLik deviance df.resid
##   -783.1   -744.7    400.6   -801.1      518
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9124 -0.5614  0.0310  0.6342  4.2162
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  state_id (Intercept) 0.004504 0.06711
##           party       0.001970 0.04438  -0.88
##  Residual             0.011432 0.10692
## Number of obs: 527, groups:  state_id, 50
##
## Fixed effects:
##                   Estimate Std. Error         df t value             Pr(>|t|)
## (Intercept)       0.109962   0.015786  52.715012   6.966        0.00000000525
## log_money         0.066234   0.005716 390.855351  11.588 < 0.0000000000000002
## party             0.395332   0.017471  93.930959  22.628 < 0.0000000000000002
## log_acres         0.011311   0.005142  36.221498   2.200               0.0343
## log_money:party  -0.043560   0.007215 436.064938  -6.037        0.00000000336
##
## (Intercept)     ***
## log_money       ***
## party           ***
## log_acres       *
## log_money:party ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) lg_mny party  lg_crs
## log_money   -0.439
## party       -0.655  0.408
## log_acres   -0.326 -0.092  0.015
## lg_mny:prty  0.374 -0.759 -0.676 -0.038
```

Остановимся на этой модели. Видно, что все ее оценки коэффициентов значимы на уровне значимости 0.05 (все, кроме коэф-та при log_acres значимы на гораздо большем уровне доверия).

Какие выводы мы можем сделать из полученных результатов[2]:

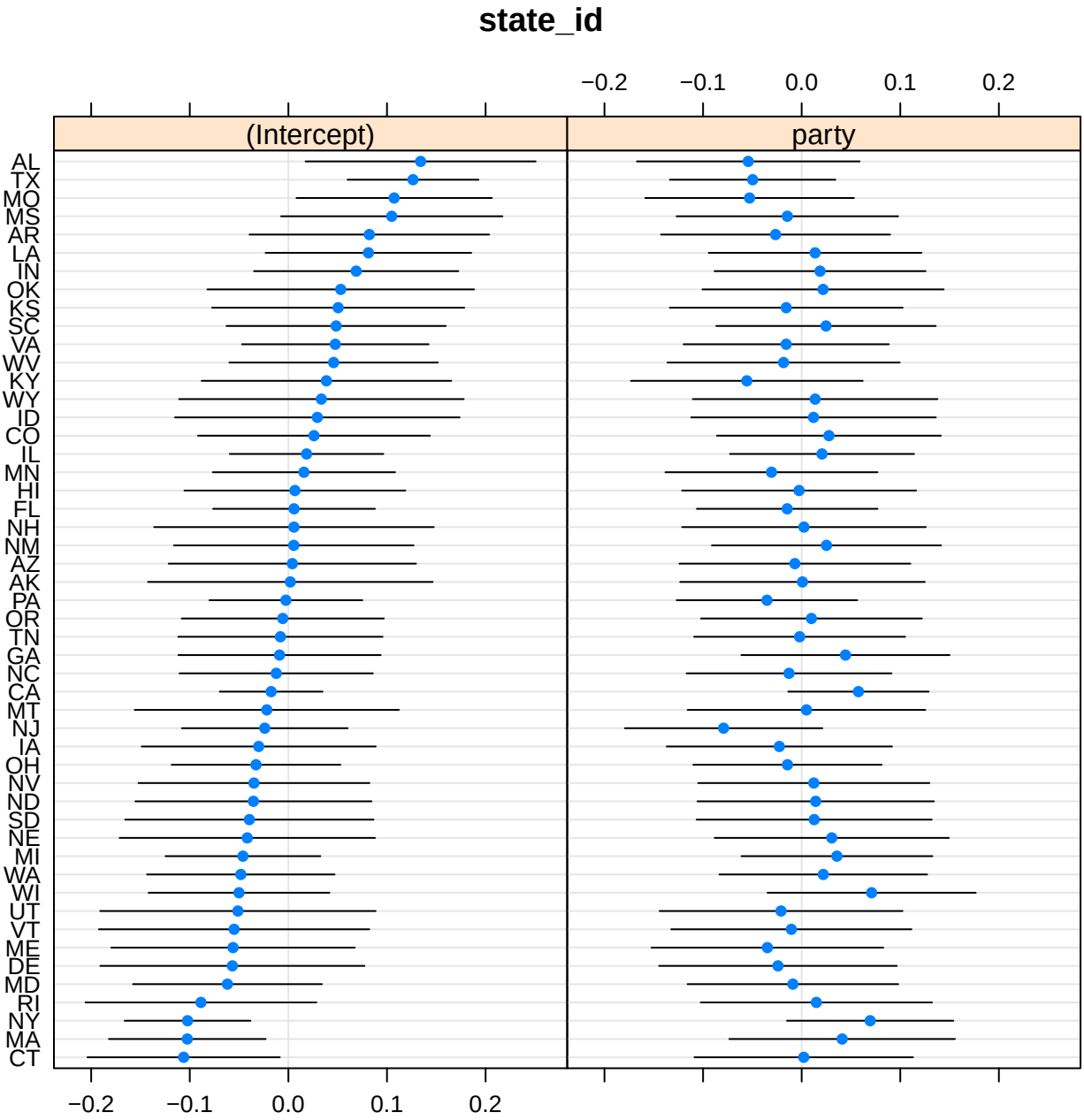
- зависимые переменные логарифм финансирования от табачных компаний, политическая партия (Республиканская) и площадь табачных полей в штате имеют; **положительную связь** с зависимой переменной — долей голосов за табачную промышленность;
- доля голосов в пользу табачной промышленности от представителей Республиканской партии в среднем выше на 40% при прочих равных;
- с увеличением размера финансирования от табачных компаний на 1%, значение зависимой переменной (доли голосов) увеличивается в среднем на 0.066%;
- с увеличением размера площади табачных полей в штате конгрессмена на 1%, значение зависимой переменной (доли голосов) увеличивается в среднем на 0.011%;
- для представителей Республиканской партии взаимосвязь между размером финансирования и долей голосов слабее, чем для представителей Демократической партии.

Таким образом, **можно предположить**, что для представителей Республиканской партии средняя доля голосов изначально (т.е. без финансирования в принципе) выше, чем для представителей Демократической партии. Но эффект финансирования от табачных компаний у последних выше, чем у первых.

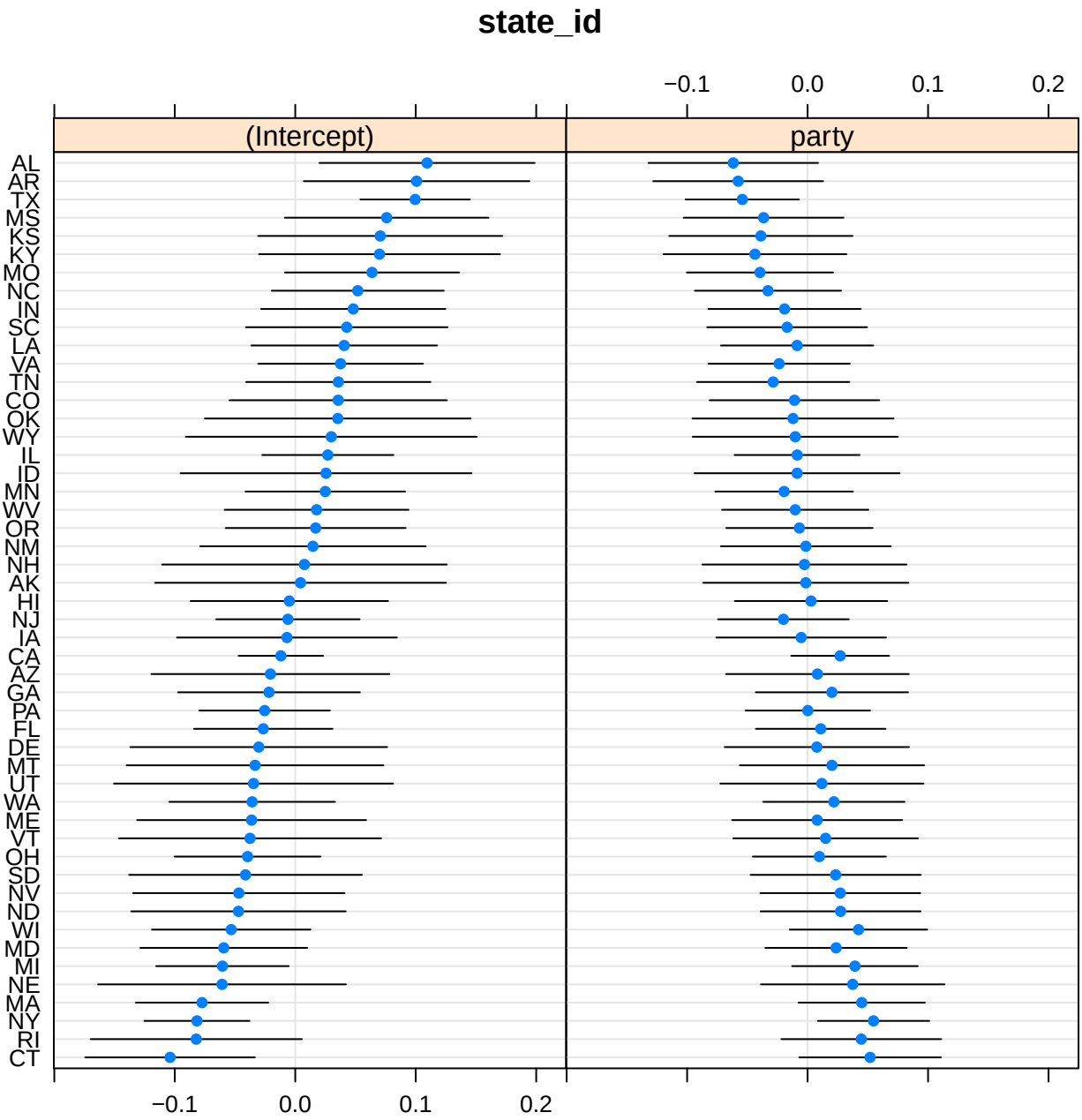
Визуализируем случайные эффекты на константу и партию для обеих моделей:

```
dotplot(ranef(model5.1, condVar=TRUE))

## $state_id
```

[2]Интерпретация на основе Gujarathi, D. M. (2004). Gujarati: Basic Econometrics. McGraw-hill, p. 175.

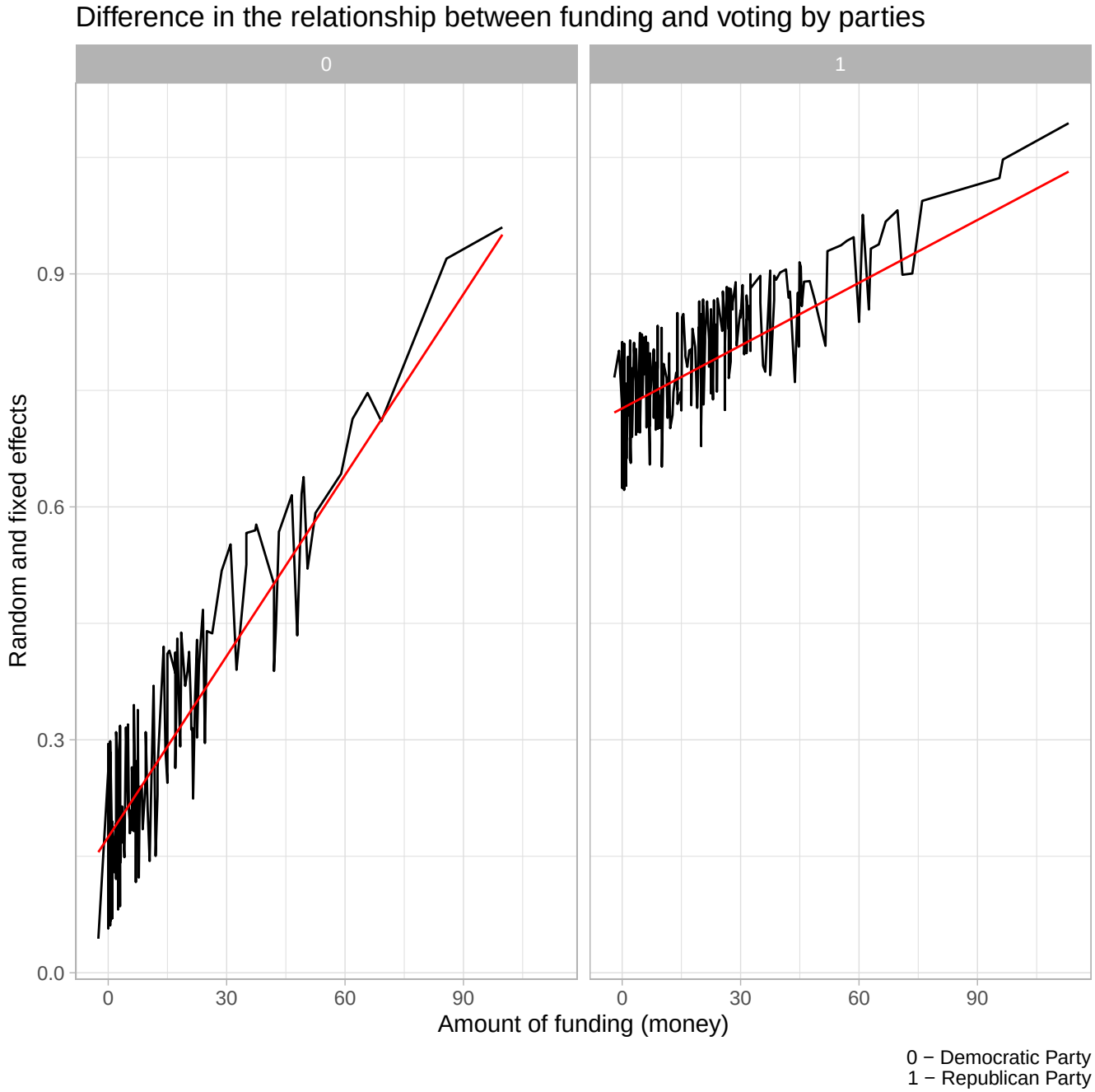



```
dotplot(ranef(model5.2, condVar=TRUE))

## $state_id
```

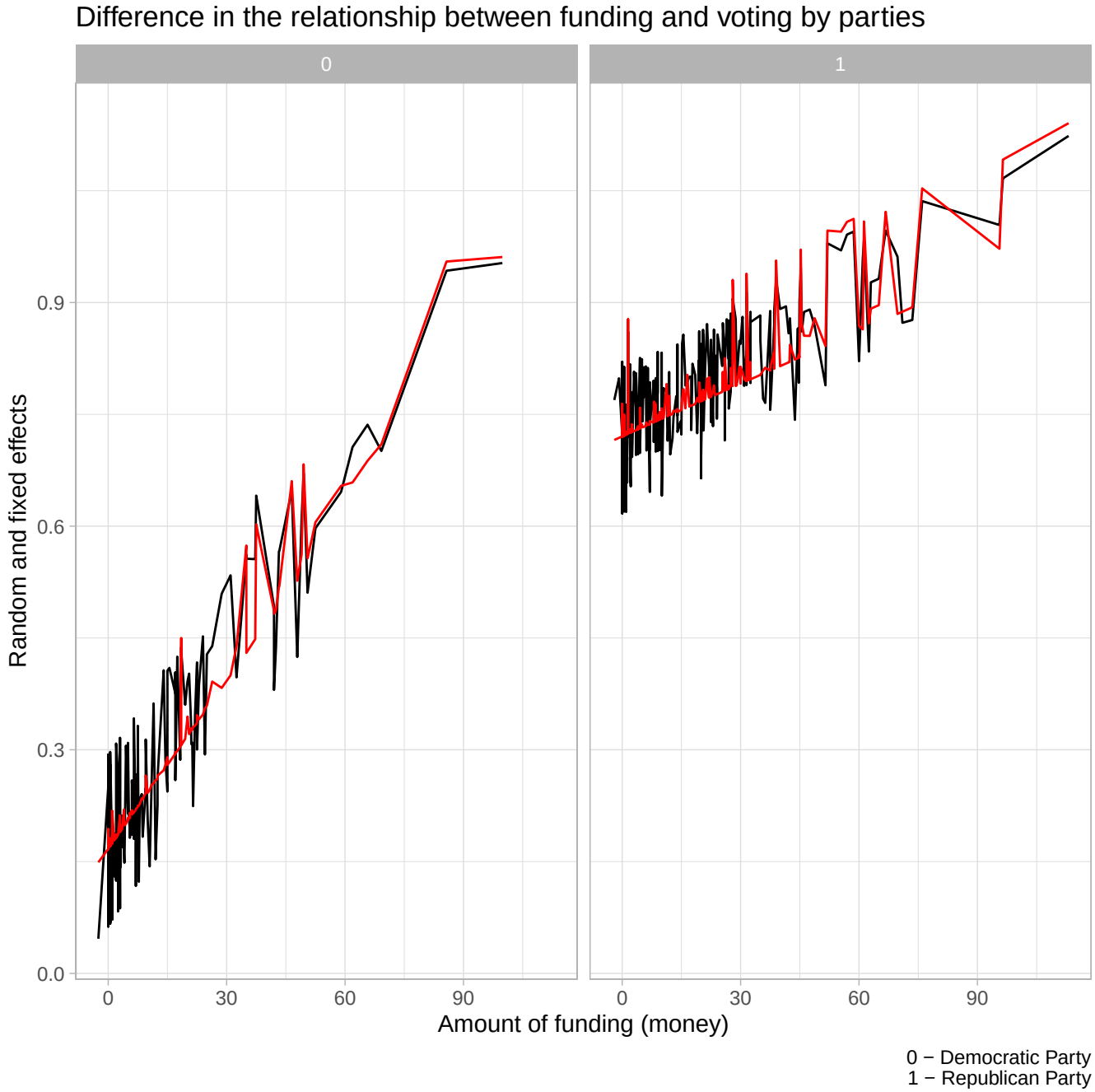


Далее я попробую **визуализировать** обнаруженную *разницу в эффекте* для двух партий. Посмотрим на разницу в предсказанных фиксированном и случайном эффектах в зависимости от партии. В начале — на линейную связь, а затем — на связь при добавлении контрольной переменной и логафримирования.

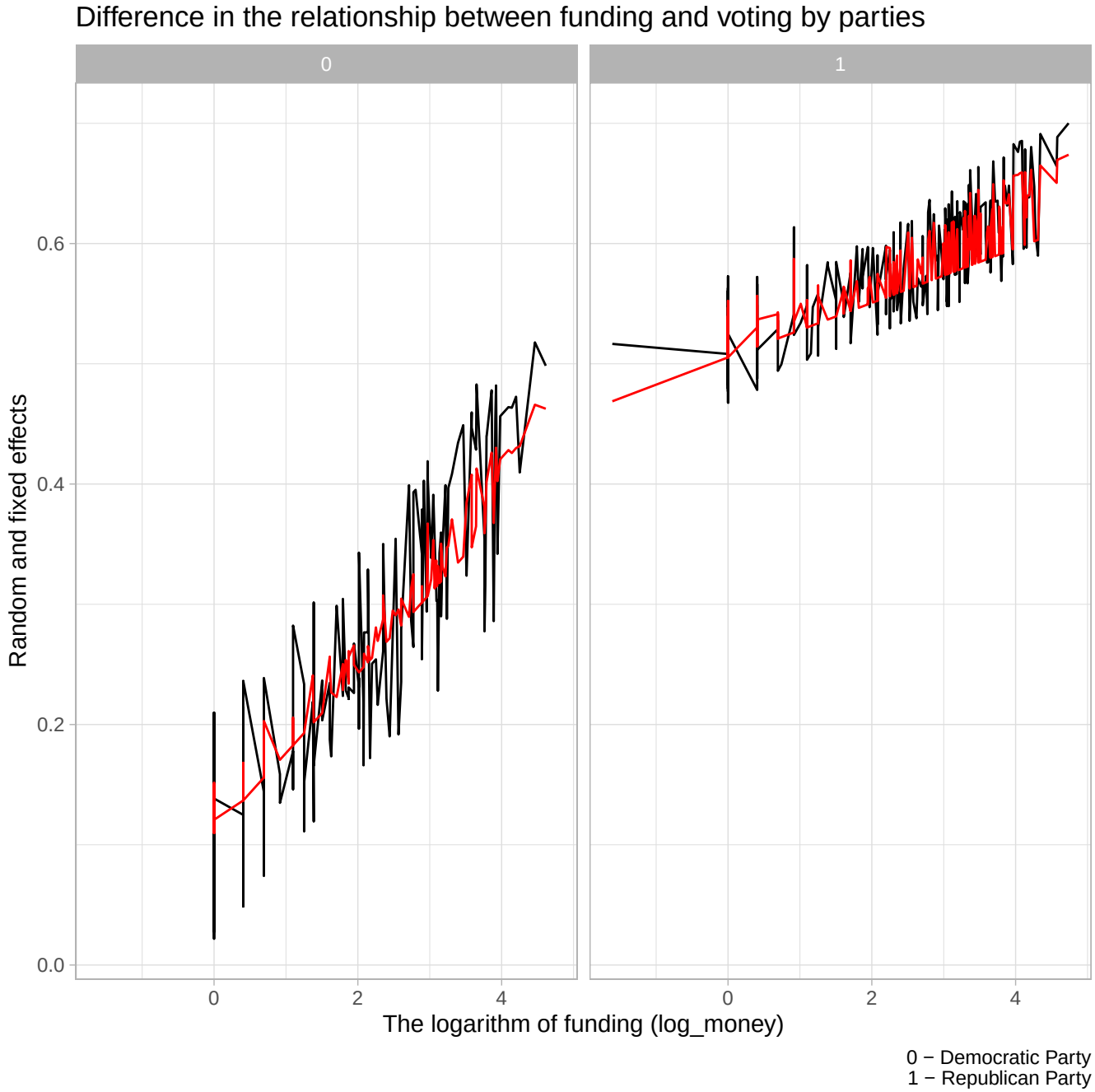
```
# визуализация различий
# простой вариант линейной связи
ME %>%
  mutate(pr_re = predict(model5.0), pr_fe = predict(model5.0, re.form = NA)) %>%
  ggplot(aes(x=money, y=pr_re, group = party)) + theme_light() + geom_line() +
  geom_line(color = "red", aes(money, pr_fe)) + facet_wrap(~party) +
  ggtitle("Difference in the relationship between funding and voting by parties") +
  xlab("Amount of funding (money)") + ylab("Random and fixed effects") +
  labs(caption = "0 - Democratic Party\n1 - Republican Party")
```




```
# добавление контрольной переменной
ME %>%
  mutate(pr_re = predict(model5.1), pr_fe = predict(model5.1, re.form = NA)) %>%
  ggplot(aes(x=money, y=pr_re, group = party)) + theme_light() + geom_line() +
  geom_line(color = "red", aes(money, pr_fe)) + facet_wrap(~party) +
  ggtitle("Difference in the relationship between funding and voting by parties") +
  xlab("Amount of funding (money)") +
  ylab("Random and fixed effects") +
  labs(caption = "0 - Democratic Party\n1 - Republican Party")
```



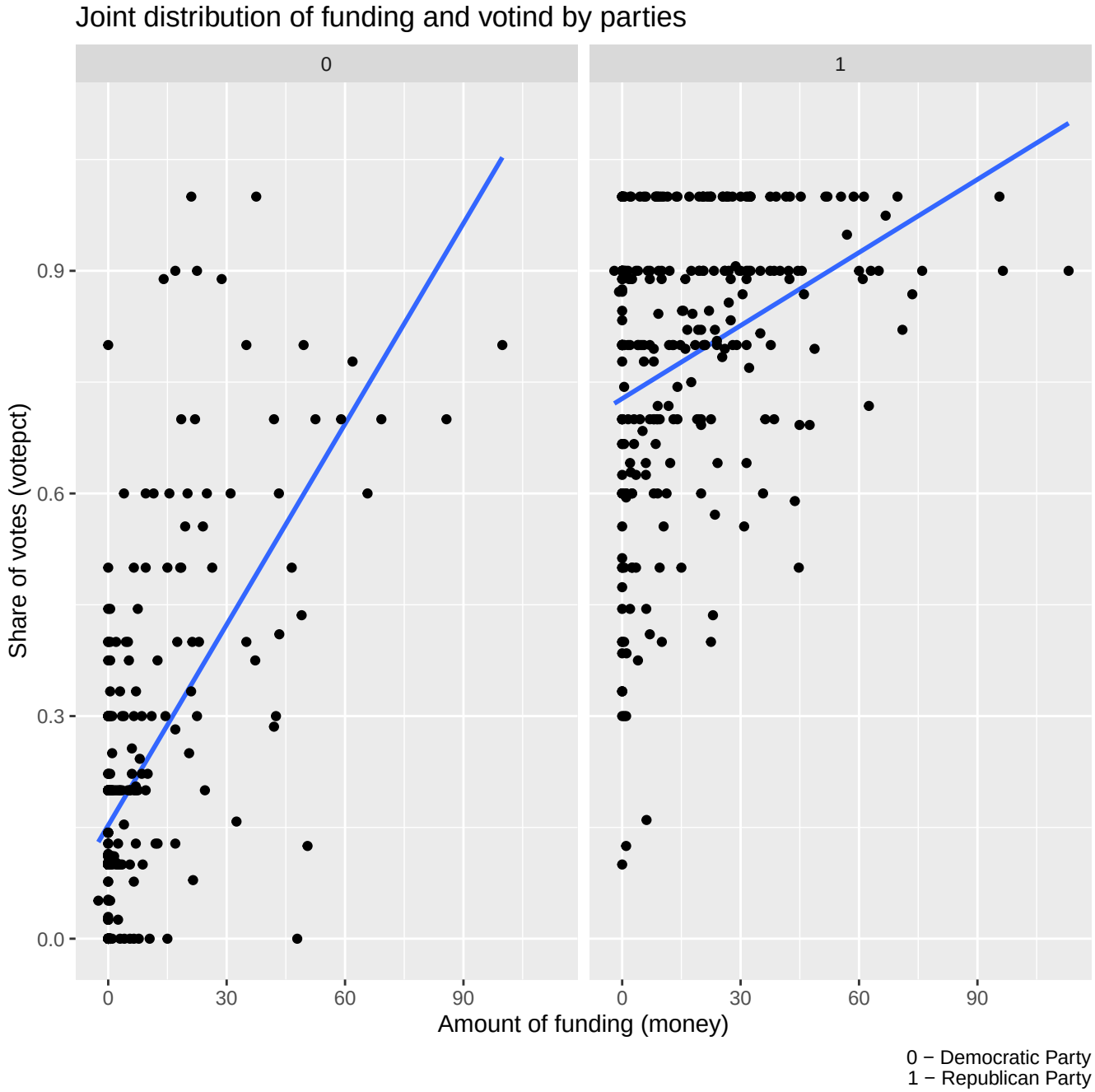

```
# логарифмирование
ME %>%
  mutate(pr_re = predict(model5.2), pr_fe = predict(model5.2, re.form = NA)) %>%
  ggplot(aes(x=log_money, y=pr_re, group = party)) + theme_light() + geom_line() +
  geom_line(color = "red", aes(log_money, pr_fe)) + facet_wrap(~party) +
  ggtitle("Difference in the relationship between funding and voting by parties") +
  xlab("The logarithm of funding (log_money)") +
  ylab("Random and fixed effects") +
  labs(caption = "0 - Democratic Party\n1 - Republican Party")
```



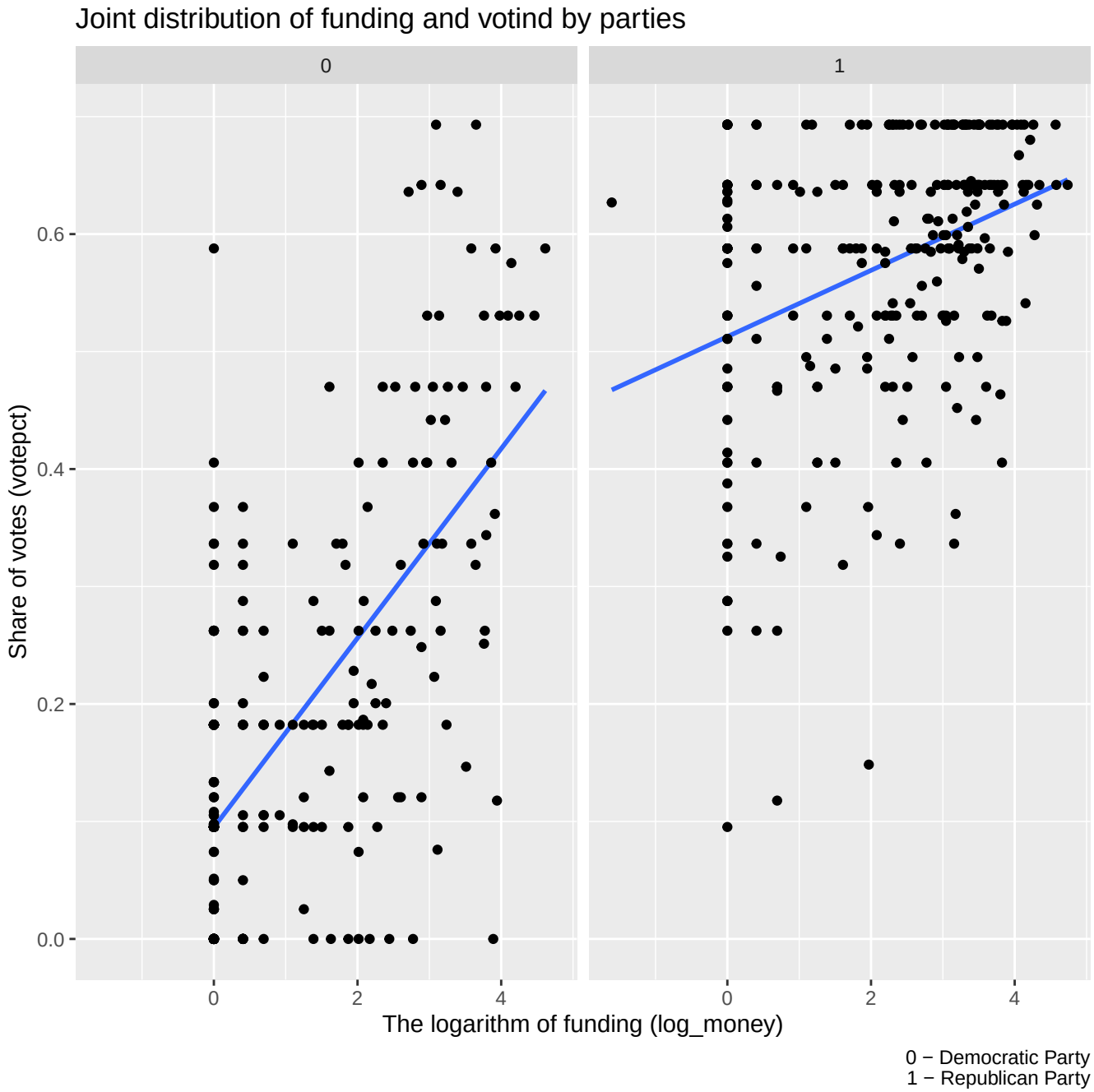
Можно заметить **явную разницу в стартовых условиях и силе эффекта**. К тому же, здесь проявляется **отрицательная корреляция** между случайными эффектами на константу и партию: чем выше стартовое значение — тем ниже эффект, как было отмечено.

Также посмотрим на совместное распределение с регрессионной линией для двух партий:

```
# простая визуализация связи
ggplot(ME, aes(x=money, y=votepct)) +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point() +
  facet_wrap("party") +
  ggtitle("Joint distribution of funding and votind by parties") +
  xlab("Amount of funding (money)") +
  ylab("Share of votes (votepct)") +
  labs(caption = "0 - Democratic Party\n1 - Republican Party")
```




```
# log
ggplot(ME, aes(x=log_money, y=log_votepct)) +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point() +
  facet_wrap("party") +
  ggtitle("Joint distribution of funding and votind by parties") +
  xlab("The logarithm of funding (log_money)") +
  ylab("Share of votes (votepct)") +
  labs(caption = "0 - Democratic Party\n1 - Republican Party")
```



Действительно можно заметить разницу во взаимосвязи — даже на совместном распределении.

Задание 5

Мы выдвинули ряд аргументов против использования модели со смешанными эффектами применительно к данному исследованию.

Говоря о возможных **иных спецификациях** для проверки выдвинутых авторами гипотез, я бы предложил **две разных альтернативных спецификации** для их проверки. Учитывая тот факт, что мы работаем со специфической выборкой, говорить об использовании модели со случайными эффектами речи не идет. Таким образом, выбор будет сделан в пользу модели с фиксированными эффектами. Далее я представлю спецификации, исходя из гипотез.

1. *Проверка гипотезы о разной взаимосвязи финансирования табачных компаний и доли голосов в поддержку табачной промышленности в зависимости от политической партии.*

   Хорошей идеей был бы предварительный исследовательский этап работы, в ходе которого можно было бы некоторым образом **кластеризовать штаты** и агрегировать их по какому-то принципу.

Однако на данный момент для этого нет возможности, поэтому все штаты будут рассматриваться по отдельности, как в оригинальном исследовании.

Во-первых, самым простым решением было бы использовать **LSDV-модель** с дамми на N-1 штатов. Хорошим решением для проверки данной гипотезы была бы **Varying-slope** модель. Можно начать с этого варианта:

```
## гипотеза о связи money и party
# LSDV
LSDV_1 <- lm(votepct~money + party + acres +
               party*money +
               state_id, data = ME)
summary(LSDV_1)


##
## Call:
## lm(formula = votepct ~ money + party + acres + party * money +
##     state_id, data = ME)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60674 -0.09372  0.00108  0.09196  0.67011
##
## Coefficients: (1 not defined because of singularities)
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)   0.1824033   0.0971914   1.877              0.06117 .
## money         0.0070820   0.0007594   9.325 < 0.0000000000000002 ***
## party         0.5489124   0.0194962  28.155 < 0.0000000000000002 ***
## acres         0.0460104   0.0762357   0.604              0.54645
## state_idAL    0.1467757   0.1104858   1.328              0.18467
## state_idAR    0.1262809   0.1210984   1.043              0.29757
## state_idAZ   -0.2368139   0.3171406  -0.747              0.45561
## state_idCA   -0.0015857   0.0986117  -0.016              0.98718
## state_idCO    0.0676816   0.1121860   0.603              0.54660
## state_idCT   -0.3269006   0.1760462  -1.857              0.06395 .
## state_idDE   -0.2096842   0.1353850  -1.549              0.12210
## state_idFL   -0.2753398   0.3726293  -0.739              0.46033
## state_idGA   -1.4663016   2.4410484  -0.601              0.54834
## state_idHI   -0.0003694   0.1275249  -0.003              0.99769
## state_idIA   -0.0876959   0.1144356  -0.766              0.44386
## state_idID    0.1021123   0.1351947   0.755              0.45045
## state_idIL    0.0259414   0.1026229   0.253              0.80054
## state_idIN   -0.1830164   0.4268560  -0.429              0.66830
## state_idKS    0.0651822   0.1170933   0.557              0.57802
## state_idKY  -10.0421987  16.8219725  -0.597              0.55081
## state_idLA    0.1323453   0.1093071   1.211              0.22659
## state_idMA   -0.2126831   0.0798203  -2.665              0.00797 **
## state_idMD   -0.4079647   0.4267086  -0.956              0.33952
## state_idME   -0.1910897   0.1267298  -1.508              0.13226
## state_idMI   -0.7294392   1.0696754  -0.682              0.49562
## state_idMN   -0.0049601   0.1096952  -0.045              0.96395
## state_idMO    0.0028730   0.1257643   0.023              0.98178
```

```
## state_idMS    0.1634621   0.1145111   1.427              0.15410
## state_idMT   -0.0546715   0.1358843  -0.402              0.68762
## state_idNC   -9.4392266  15.7657112  -0.599              0.54965
## state_idND   -0.0951441   0.1361273  -0.699              0.48494
## state_idNE   -1.8612785   2.9756046  -0.626              0.53194
## state_idNH    0.0116094   0.1264552   0.092              0.92689
## state_idNJ   -0.0945611   0.1051953  -0.899              0.36916
## state_idNM    0.0319239   0.1211004   0.264              0.79219
## state_idNV   -0.0805107   0.1270180  -0.634              0.52648
## state_idNY   -0.0946895   0.1004187  -0.943              0.34619
## state_idOH   -0.5052279   0.6748818  -0.749              0.45446
## state_idOK   -0.7932390   1.4521925  -0.546              0.58516
## state_idOR   -0.0167690   0.1149190  -0.146              0.88405
## state_idPA   -0.3170908   0.4025219  -0.788              0.43123
## state_idRI   -0.2065218   0.1270925  -1.625              0.10483
## state_idSC   -1.6635728   2.8985762  -0.574              0.56629
## state_idSD   -0.1045991   0.1356638  -0.771              0.44108
## state_idTN   -2.8763366   4.7405444  -0.607              0.54431
## state_idTX    0.1113965   0.1004197   1.109              0.26786
## state_idUT   -0.1510617   0.1209555  -1.249              0.21232
## state_idVA   -1.6793680   2.8446174  -0.590              0.55523
## state_idVT   -0.2400176   0.1513839  -1.585              0.11352
## state_idWA   -0.0668941   0.1082421  -0.618              0.53687
## state_idWI   -0.0910136   0.0781493  -1.165              0.24476
## state_idWV           NA          NA      NA                  NA
## state_idWY    0.1195997   0.1353339   0.884              0.37729
## money:party  -0.0050695   0.0008942  -5.669         0.000000025 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1655 on 474 degrees of freedom
## Multiple R-squared:  0.7964,Adjusted R-squared:  0.7741
## F-statistic: 35.66 on 52 and 474 DF,  p-value: < 0.00000000000000022
```

Выдача получилась большой, все интересующие коэффициенты при предикторах — **значимы** на любом уровне значимости. Однако важно отметить, что большинство коэф-в при дамми на штаты оказались **незначимы**. Они не интересовали нас, к тому же, из этого можно сделать некоторые содержательные выводы. Также, вероятно, значение R^2 получилось **завышенным** (почти 80%) из-за большой спецификации.

Таким образом, от LSDV-модели можно перейти к **модели с внутригрупповым преобразованием**. Учитывая тот факт, что нам больше интересны сами коэф-ты при предикторах, это избавит нас от громоздкой выдачи и скорректирует значение R^2:

```
# внутригруп. преобр.
fe_1 <- plm(votepct~money + party + acres +
              party*money, data = ME, index=c("state_id"), effect = "individual", model="
summary(fe_1)


## Oneway (individual) effect Within Model
##
```

```
## Call:
## plm(formula = votepct ~ money + party + acres + party * money,
##     data = ME, effect = "individual", model = "within", index = c("state_id"))
##
## Unbalanced Panel: n = 50, T = 2-55, N = 527
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.6067425 -0.0937168  0.0010848  0.0919587  0.6701064
##
## Coefficients:
##                Estimate  Std. Error t-value              Pr(>|t|)
## money        0.00708200  0.00075943  9.3254 < 0.00000000000000022 ***
## party        0.54891241  0.01949622 28.1548 < 0.00000000000000022 ***
## money:party -0.00506949  0.00089421 -5.6692         0.00000002496 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    44.057
## Residual Sum of Squares: 12.99
## R-Squared:      0.70516
## Adj. R-Squared: 0.67282
## F-statistic: 377.89 on 3 and 474 DF, p-value: < 0.000000000000000222
```

Такая модель выглядит гораздо лучше.

Далее можно оценить такую же спецификацию, но используя **преобразованные переменные**:

```
# log
fe_1.2 <- plm(log_votepct~log_money + party + log_acres +
              party*log_money, data = ME, index=c("state_id"), effect = "individual", mod
summary(fe_1.2)


## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_votepct ~ log_money + party + log_acres + party *
##     log_money, data = ME, effect = "individual", model = "within",
##     index = c("state_id"))
##
## Unbalanced Panel: n = 50, T = 2-55, N = 527
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.4097363 -0.0635614  0.0039008  0.0662137  0.4428054
##
## Coefficients:
##                   Estimate Std. Error t-value              Pr(>|t|)
## log_money        0.0653424  0.0056498 11.5655 < 0.00000000000000022 ***
## party            0.3972542  0.0159743 24.8684 < 0.00000000000000022 ***
```

```
## log_money:party -0.0469627  0.0071733 -6.5469        0.000000000153 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    20.686
## Residual Sum of Squares: 5.6302
## R-Squared:      0.72783
## Adj. R-Squared: 0.69797
## F-statistic: 422.515 on 3 and 474 DF, p-value: < 0.000000000000000222
```

Результаты получились схожими с ME-моделью. Знак и сила полученных оценок очень похожи: вновь логарифм финансирования и площадь табачных полей оказывают положительное воздействие на долю голосования, а эффект финансирования на долю голосов больше у представителей Демократической партии.

Однако стоит помнить о том, что такая FE-модель в большей степени подвержена смещению из-за влиятельных переменных. Обозначенные в начале штаты могут влиять на результаты гораздо больше.

2. Далее попробуем проверить *гипотезу о разной взаимосвязи площади табачных полей в штате конгрессмена и доли голосов в поддержку табачной промышленности в зависимости от политической партии.*

   Поступим похожим образом: в начале оценим простую **Varying-slope** модель, изменив переменную взаимодействия на партию и площадь:

```
## гипотеза о связи party и acres
# LSDV
LSDV_2 <- lm(votepct~money + party + acres +
               party*acres +
               state_id, data = ME)
summary(LSDV_2)


##
## Call:
## lm(formula = votepct ~ money + party + acres + party * acres +
##     state_id, data = ME)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58340 -0.10058 -0.00593  0.09056  0.72720
##
## Coefficients: (1 not defined because of singularities)
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)   0.2022452   0.0985917   2.051              0.04078 *
## money         0.0037293   0.0005045   7.392    0.000000000000658 ***
## party         0.5103772   0.0173856  29.356 < 0.0000000000000002 ***
## acres         0.0568896   0.0775216   0.734              0.46340
## state_idAL    0.1526353   0.1122416   1.360              0.17451
## state_idAR    0.1206975   0.1230050   0.981              0.32697
## state_idAZ   -0.2799203   0.3225097  -0.868              0.38586
## state_idCA   -0.0079479   0.1001615  -0.079              0.93679
```

```
## state_idCO    0.0570700   0.1139305   0.501              0.61666
## state_idCT   -0.3547036   0.1789462  -1.982              0.04804 *
## state_idDE   -0.2288722   0.1374528  -1.665              0.09655 .
## state_idFL   -0.3261153   0.3788992  -0.861              0.38984
## state_idGA   -1.8089039   2.4819068  -0.729              0.46646
## state_idHI   -0.0063813   0.1295596  -0.049              0.96074
## state_idIA   -0.1048239   0.1161771  -0.902              0.36737
## state_idID    0.0939096   0.1373234   0.684              0.49440
## state_idIL    0.0110212   0.1041756   0.106              0.91579
## state_idIN   -0.2494649   0.4340427  -0.575              0.56573
## state_idKS    0.0636195   0.1189442   0.535              0.59299
## state_idKY  -12.2016924  17.1025318  -0.713              0.47592
## state_idLA    0.1339254   0.1110527   1.206              0.22843
## state_idMA   -0.2474444   0.0808023  -3.062              0.00232 **
## state_idMD   -0.4746392   0.4338679  -1.094              0.27452
## state_idME   -0.1925728   0.1287422  -1.496              0.13537
## state_idMI   -0.8751420   1.0876179  -0.805              0.42143
## state_idMN   -0.0183337   0.1113749  -0.165              0.86932
## state_idMO   -0.0071164   0.1278404  -0.056              0.95563
## state_idMS    0.1632302   0.1163341   1.403              0.16124
## state_idMT   -0.0882113   0.1378366  -0.640              0.52250
## state_idNC  -11.4836439  16.0293778  -0.716              0.47409
## state_idND   -0.0848115   0.1383780  -0.613              0.54024
## state_idNE   -2.2237272   3.0252952  -0.735              0.46268
## state_idNH    0.0051237   0.1284487   0.040              0.96820
## state_idNJ   -0.0993044   0.1068597  -0.929              0.35321
## state_idNM    0.0285077   0.1230144   0.232              0.81684
## state_idNV   -0.0777977   0.1290691  -0.603              0.54696
## state_idNY   -0.0926662   0.1020400  -0.908              0.36427
## state_idOH   -0.6179588   0.6862298  -0.901              0.36830
## state_idOK   -0.9810865   1.4764804  -0.664              0.50671
## state_idOR   -0.0256008   0.1167170  -0.219              0.82648
## state_idPA   -0.3882681   0.4092940  -0.949              0.34329
## state_idRI   -0.2142154   0.1290931  -1.659              0.09770 .
## state_idSC   -2.0058198   2.9470167  -0.681              0.49644
## state_idSD   -0.1168135   0.1377719  -0.848              0.39693
## state_idTN   -3.4519799   4.8197642  -0.716              0.47421
## state_idTX    0.1069937   0.1020093   1.049              0.29478
## state_idUT   -0.1399792   0.1228501  -1.139              0.25510
## state_idVA   -2.0394492   2.8922377  -0.705              0.48107
## state_idVT   -0.2415361   0.1537842  -1.571              0.11694
## state_idWA   -0.0761982   0.1099269  -0.693              0.48854
## state_idWI   -0.0990530   0.0793784  -1.248              0.21270
## state_idWV           NA          NA      NA                  NA
## state_idWY    0.0999519   0.1374236   0.727              0.46738
## party:acres  -0.0016243   0.0003998  -4.063    0.000056701566776 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1682 on 474 degrees of freedom
## Multiple R-squared:  0.7899,Adjusted R-squared:  0.7669
```

```
## F-statistic: 34.28 on 52 and 474 DF,  p-value: < 0.00000000000000022
```

Ситуация аналогичная, но оценка при коэффициенте при предикторе площадь табачных полей незначима (об этом ниже). Теперь можно оценить модель с **внутригрупповым преобразованием** на изначальных данных:

```
# внутригруп. преобр.
fe_2.1 <- plm(votepct~money + party + acres +
              party*acres, data = ME, index=c("state_id"), effect = "individual", model="
summary(fe_2.1)

## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = votepct ~ money + party + acres + party * acres,
##     data = ME, effect = "individual", model = "within", index = c("state_id"))
##
## Unbalanced Panel: n = 50, T = 2-55, N = 527
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.5834038 -0.1005821 -0.0059267  0.0905571  0.7272014
##
## Coefficients:
##                Estimate  Std. Error t-value              Pr(>|t|)
## money        0.00372928  0.00050447  7.3924    0.0000000000006578 ***
## party        0.51037721  0.01738562 29.3563 < 0.00000000000000022 ***
## party:acres -0.00162432  0.00039979 -4.0630    0.0000567015667764 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    44.057
## Residual Sum of Squares: 13.404
## R-Squared:      0.69577
## Adj. R-Squared: 0.66239
## F-statistic: 361.339 on 3 and 474 DF, p-value: < 0.000000000000000222
```

И также модель с внутригрупповым преобразованием на **преобразованных данных**:

```
# log
fe_1.2 <- plm(log_votepct~log_money + party + log_acres +
                party*log_acres, data = ME, index=c("state_id"), effect = "individual", m
summary(fe_1.2)

## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_votepct ~ log_money + party + log_acres + party *
##     log_acres, data = ME, effect = "individual", model = "within",
##     index = c("state_id"))
```

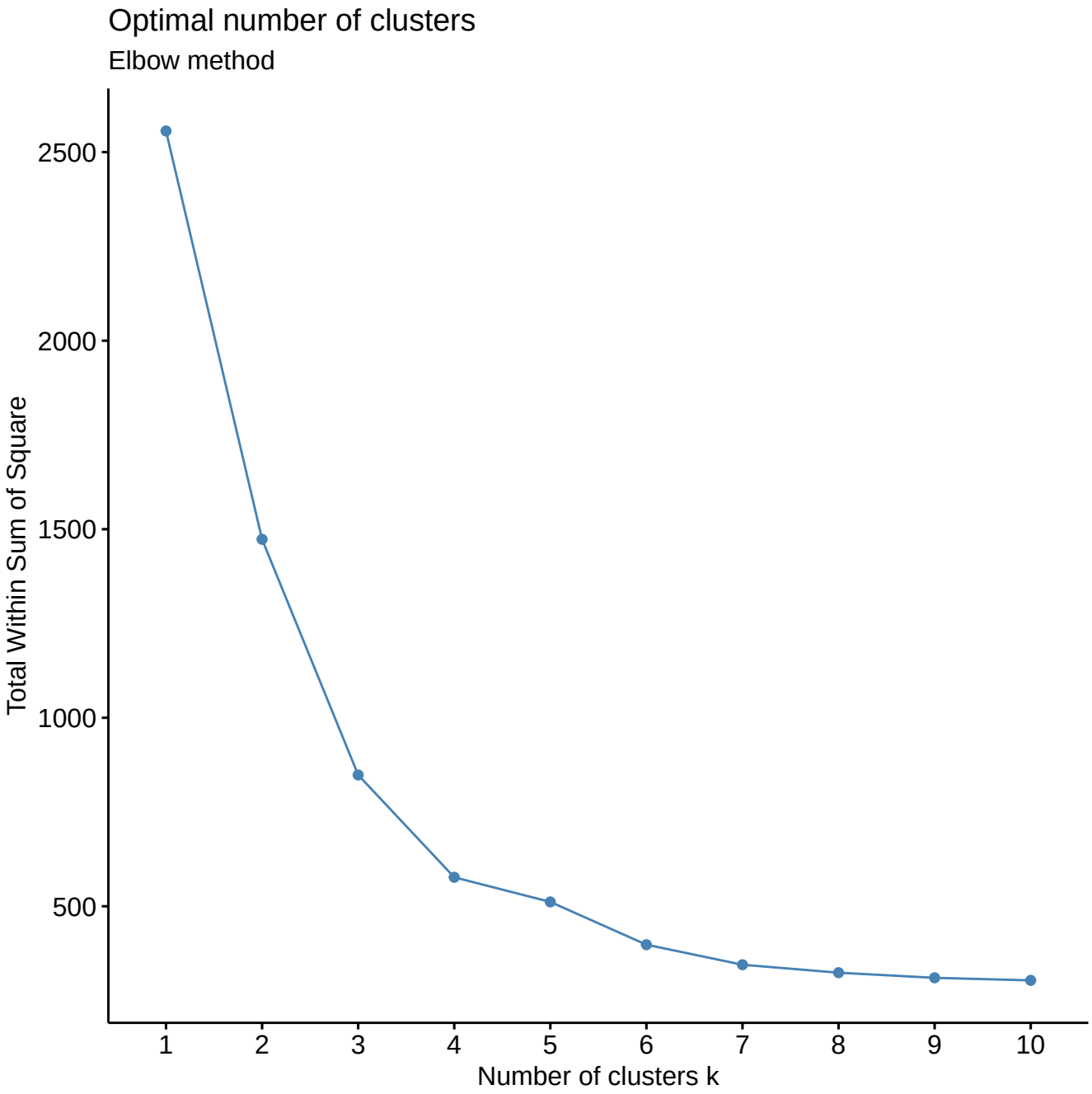
```
##
## Unbalanced Panel: n = 50, T = 2-55, N = 527
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.3919596 -0.0667420 -0.0012045  0.0685504  0.4440551
##
## Coefficients:
##                   Estimate Std. Error t-value            Pr(>|t|)
## log_money        0.0365591  0.0040381  9.0536 <0.0000000000000002 ***
## party            0.3483983  0.0141800 24.5697 <0.0000000000000002 ***
## party:log_acres -0.0228478  0.0071098 -3.2136              0.0014 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    20.686
## Residual Sum of Squares: 6.0084
## R-Squared:      0.70954
## Adj. R-Squared: 0.67768
## F-statistic: 385.974 on 3 and 474 DF, p-value: < 0.000000000000000222
```

Результат получился похожим на тот, что был ранее для партии в качестве модератора: логарифм финансирования и Республиканская партия вновь оказывают положительное воздействие на долю голосования. Несмотря на то, что, согласно выдаче, эффект площади табачных полей меньше для представителей Республиканской партии, сам коэффициент для эффекта табачных полей на долю голосов выявить не удалось идентифицировать в моделях с внутригрупповым преобразованием, поскольку данный предиктор не имеет внутригрупповой вариации. Это получилось сделать лишь технически для LSDV-модели, но, тем не менее, он сильно не значим, и это логично.

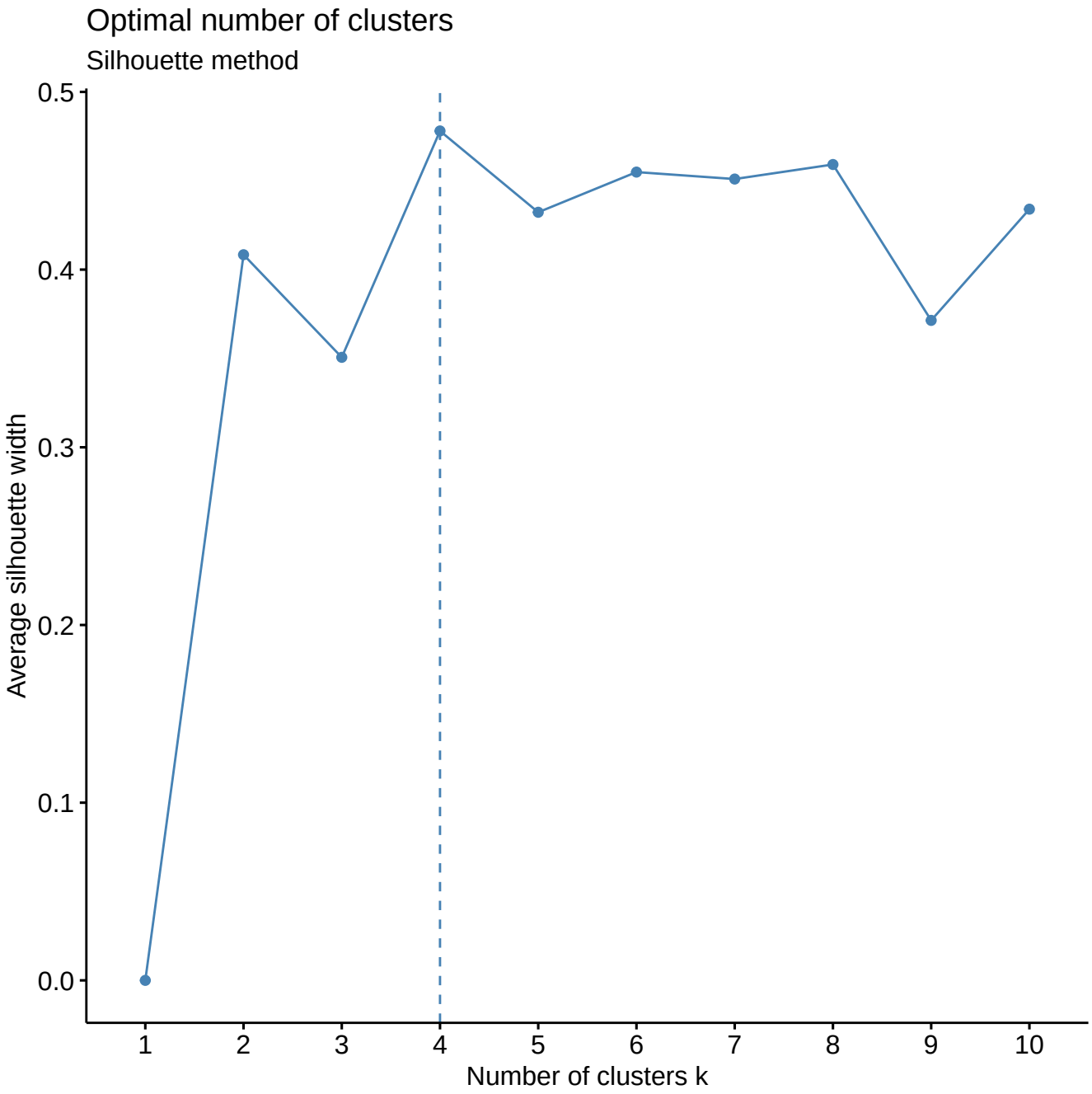
Единственной возможностью получить явный значимый эффект площади табачных полей на долю голосов в поддержку табачной промышленности является “создание„ этой вариации — т.е. группировать штаты некоторым образом, о чем говорилось ранее.

Предлагаю попробовать кластеризовать штаты (перед этим — определить оптимальное число кластеров с помощью разных алгоритмов), а затем — оценить спецификацию, используя разделение не на штаты, а на новые кластеры:

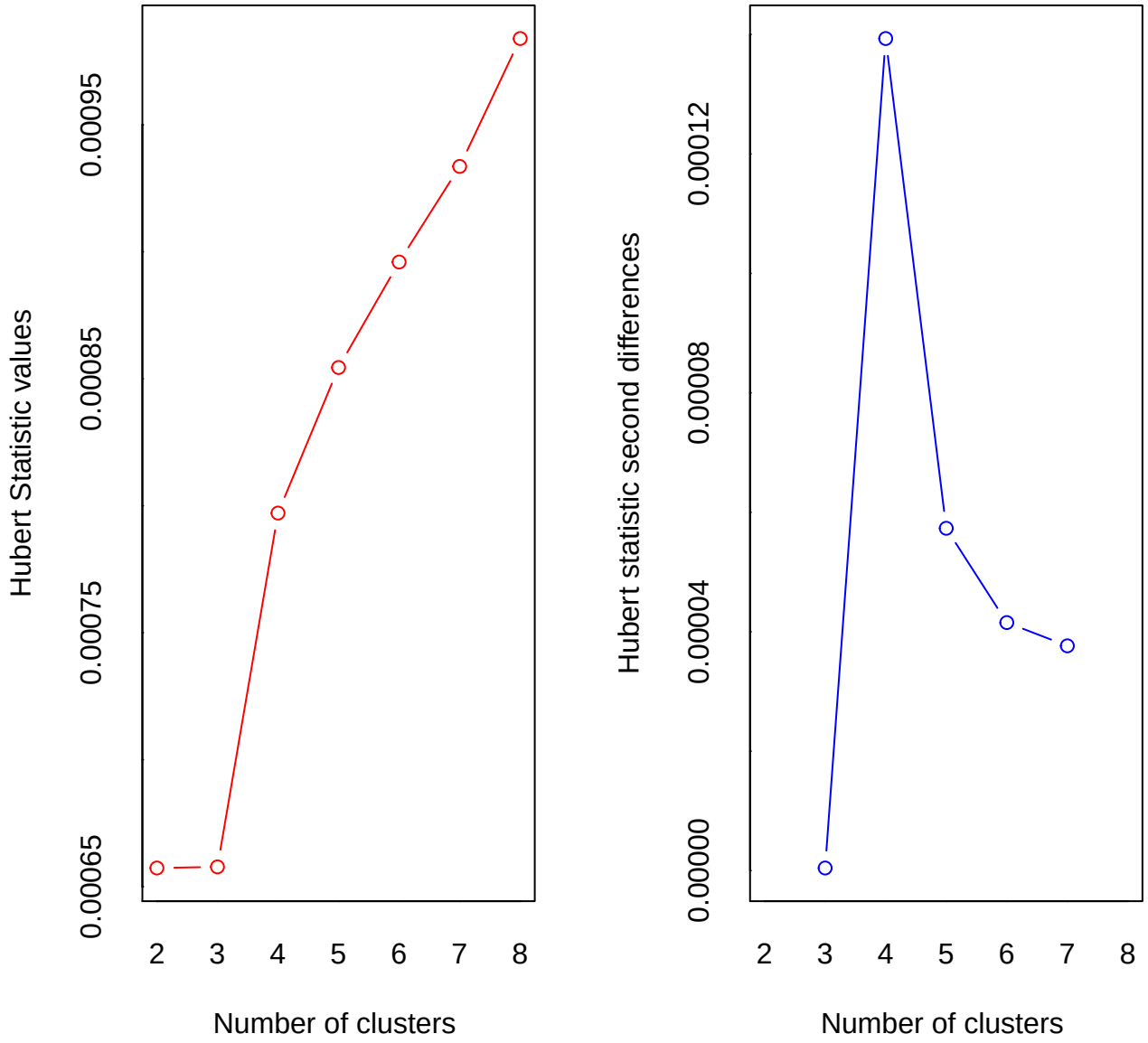
```
# cluster
to_clust <- ME %>% dplyr::select(log_votepct, log_money, party, log_acres)

library(factoextra)
fviz_nbclust(to_clust, kmeans, method = "wss") +
  labs(subtitle = "Elbow method")
```



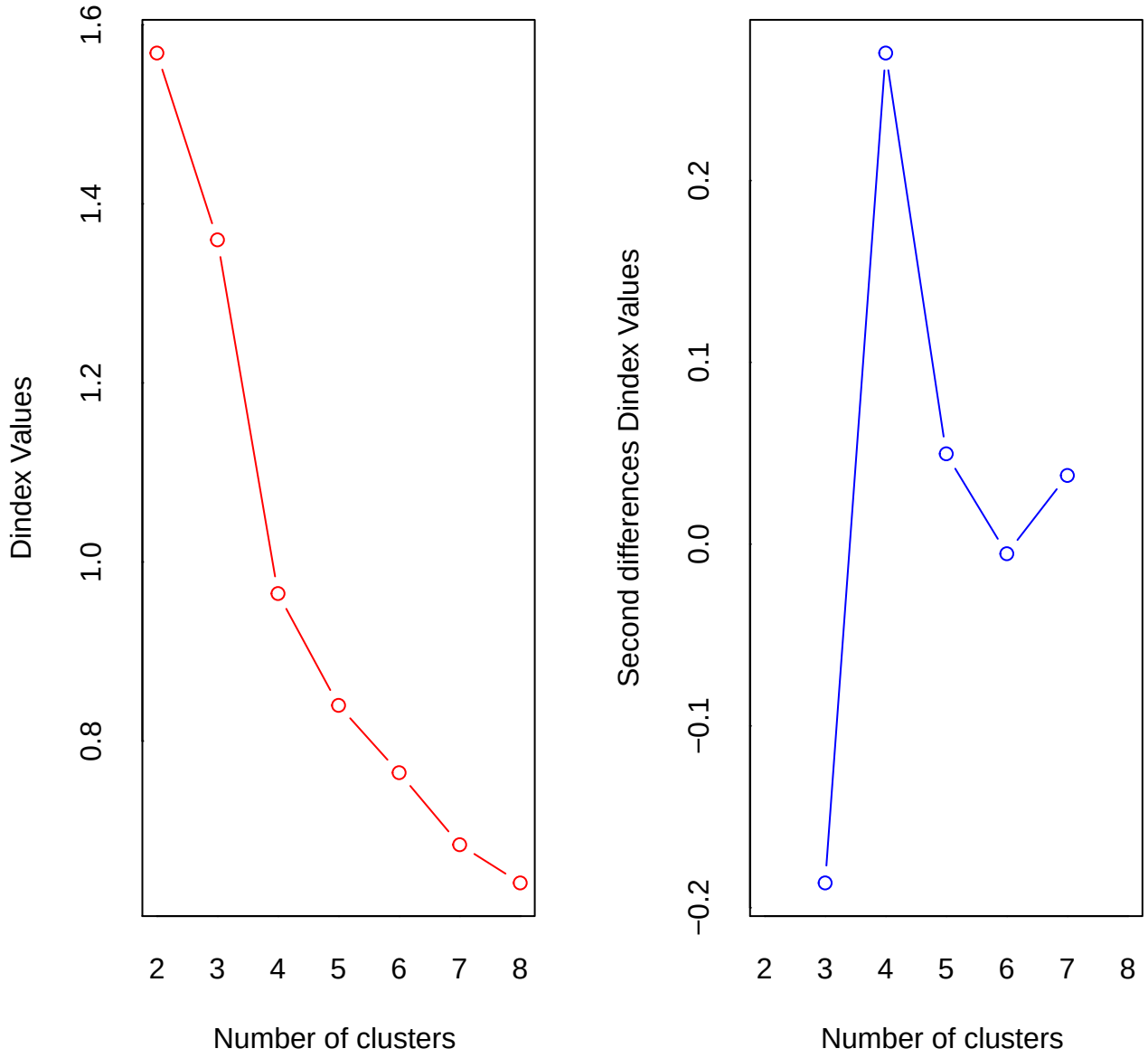

```
fviz_nbclust(to_clust, kmeans, method = "silhouette") +
  labs(subtitle = "Silhouette method")
```




```
library(NbClust)
res <- NbClust(to_clust, min.nc = 2, max.nc = 8,
               method = "kmeans")
```




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that correspon
##                 significant increase of the value of the measure i.e the significant p
##                 index second differences plot.
##
```

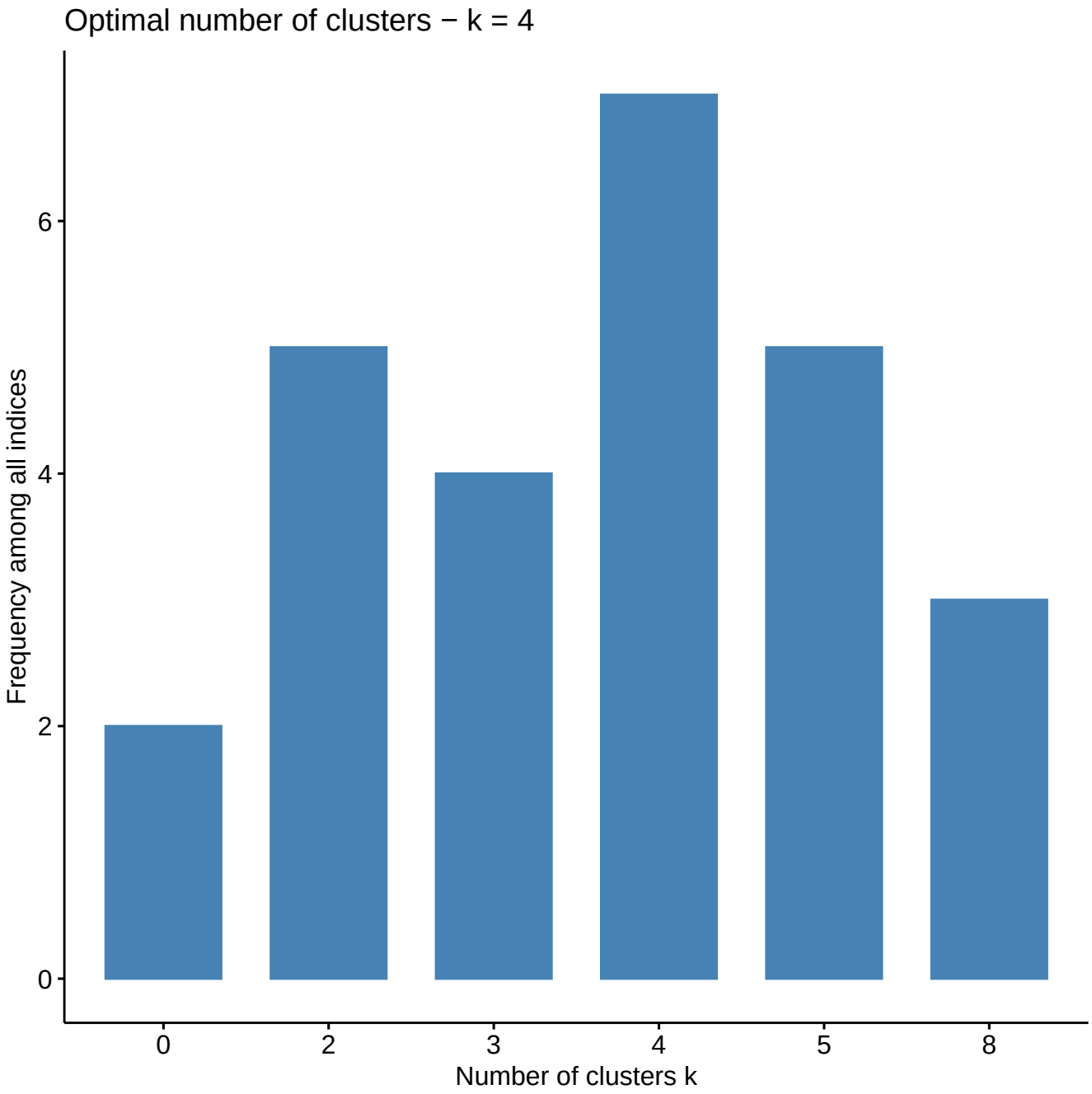



```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant pe
##                 second differences plot) that corresponds to a significant increase of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 5 proposed 2 as the best number of clusters
## * 4 proposed 3 as the best number of clusters
## * 7 proposed 4 as the best number of clusters
## * 5 proposed 5 as the best number of clusters
## * 3 proposed 8 as the best number of clusters
##
##                   ***** Conclusion *****
##
```

```
## * According to the majority rule, the best number of clusters is  4
##
##
## *******************************************************************
```

```
fviz_nbclust(res)
```

```
## Among all indices:
## ===================
## * 2 proposed  0 as the best number of clusters
## * 5 proposed  2 as the best number of clusters
## * 4 proposed  3 as the best number of clusters
## * 7 proposed  4 as the best number of clusters
## * 5 proposed  5 as the best number of clusters
## * 3 proposed  8 as the best number of clusters
##
## Conclusion
## ==========================
## * According to the majority rule, the best number of clusters is  4 .
```




```
kmeans_clust <- kmeans(to_clust, 4)
to_clust$klust <- factor(kmeans_clust$cluster)

ME$k <- factor(kmeans_clust$cluster)
```

Согласно статистическим тестам, я решил выбрать число кластеров $k = 4$. Теперь оценим новую модель:

```
fe_2.3 <- plm(log_votepct~log_money + party + log_acres +
                party*log_acres, data = ME, index=c("k"), effect = "individual", model="w
summary(fe_2.3)

## Oneway (individual) effect Within Model
##
## Call:
```

```
## plm(formula = log_votepct ~ log_money + party + log_acres + party *
##     log_acres, data = ME, effect = "individual", model = "within",
##     index = c("k"))
##
## Unbalanced Panel: n = 4, T = 76-183, N = 527
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -0.381152 -0.074599 -0.012102  0.072292  0.470324
##
## Coefficients:
##                   Estimate Std. Error t-value              Pr(>|t|)
## log_money        0.0277788  0.0086241  3.2211             0.0013573 **
## party            0.3622169  0.0143688 25.2085 < 0.00000000000000022 ***
## log_acres        0.0361483  0.0093448  3.8683             0.0001236 ***
## party:log_acres -0.0232492  0.0074258 -3.1309             0.0018412 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    21.493
## Residual Sum of Squares: 7.6854
## R-Squared:      0.64242
## Adj. R-Squared: 0.63759
## F-statistic: 233.102 on 4 and 519 DF, p-value: < 0.000000000000000222
```

Итак, теперь **удалось получить оценку для коэф-та при предикторе log_acres** — он положительный, значимый, и равен 0.036.

Гипотеза подтверждается: для Демократической партии эффект развитости табачной идустрии в родном штате меньше, чем у представителей Республиканой партии. Это хорошо сочетается с бОльшим эффектом финансирования для первых. Грубо говоря, им "все равно,,, каких представителей табачной сферы лоббировать в случае финансовой поддержки.

Ради интереса можно посмотреть, какие штаты попали в какие кластеры:

```
subset1 = subset(ME, ME$k == 1)
subset2 = subset(ME, ME$k == 2)
subset3 = subset(ME, ME$k == 3)
subset4 = subset(ME, ME$k == 4)

unique(subset1$state)

##  [1] "AK" "AL" "AR" "CA" "CO" "DE" "HI" "IA" "ID" "IL" "KS" "LA" "MA" "ME" "MN"
## [16] "MO" "MS" "ND" "NH" "NJ" "NM" "NV" "NY" "OR" "RI" "SD" "TX" "UT" "VT" "WA"
## [31] "WI" "WV" "WY"

unique(subset2$state)

##  [1] "AK" "AL" "AR" "AZ" "CA" "CO" "CT" "DE" "FL" "HI" "IA" "ID" "IL" "IN" "KS"
## [16] "LA" "MD" "ME" "MN" "MO" "MS" "MT" "ND" "NH" "NJ" "NM" "NV" "NY" "PA" "SD"
## [31] "TX" "UT" "WA" "WI" "WV" "WY"
```

```r
unique(subset3$state)
```

```
##  [1] "AZ" "CT" "FL" "GA" "IN" "MD" "MI" "NE" "OH" "OK" "PA" "SC" "TN" "VA"
```

```r
unique(subset4$state)
```

```
##  [1] "GA" "KY" "MI" "NC" "NE" "OH" "OK" "SC" "TN" "VA"
```