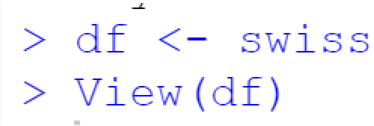
# Программирование в среде R

Шевцов Василий Викторович,

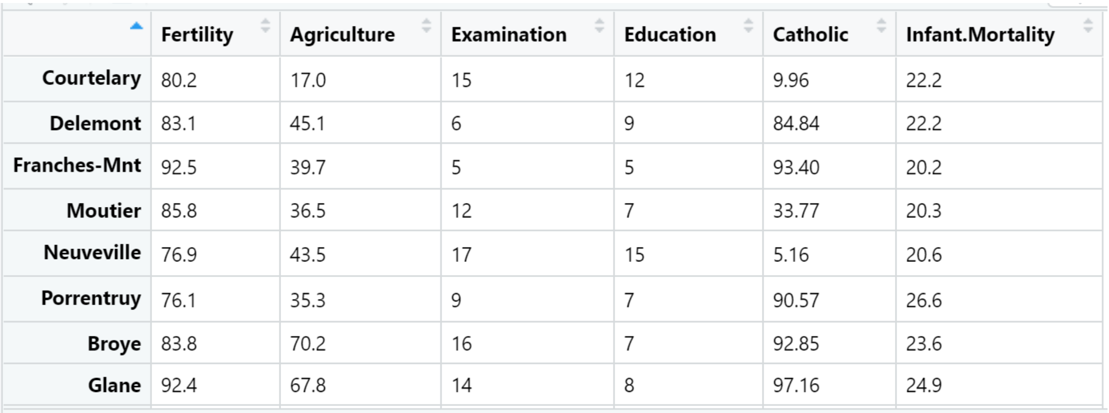
директор ДИТ РУДН, shevtsov_vv@rudn.university

# Множественная линейная регрессия.
# Отбор моделей.

# swiss



```
> df <- swiss
> View(df)
```

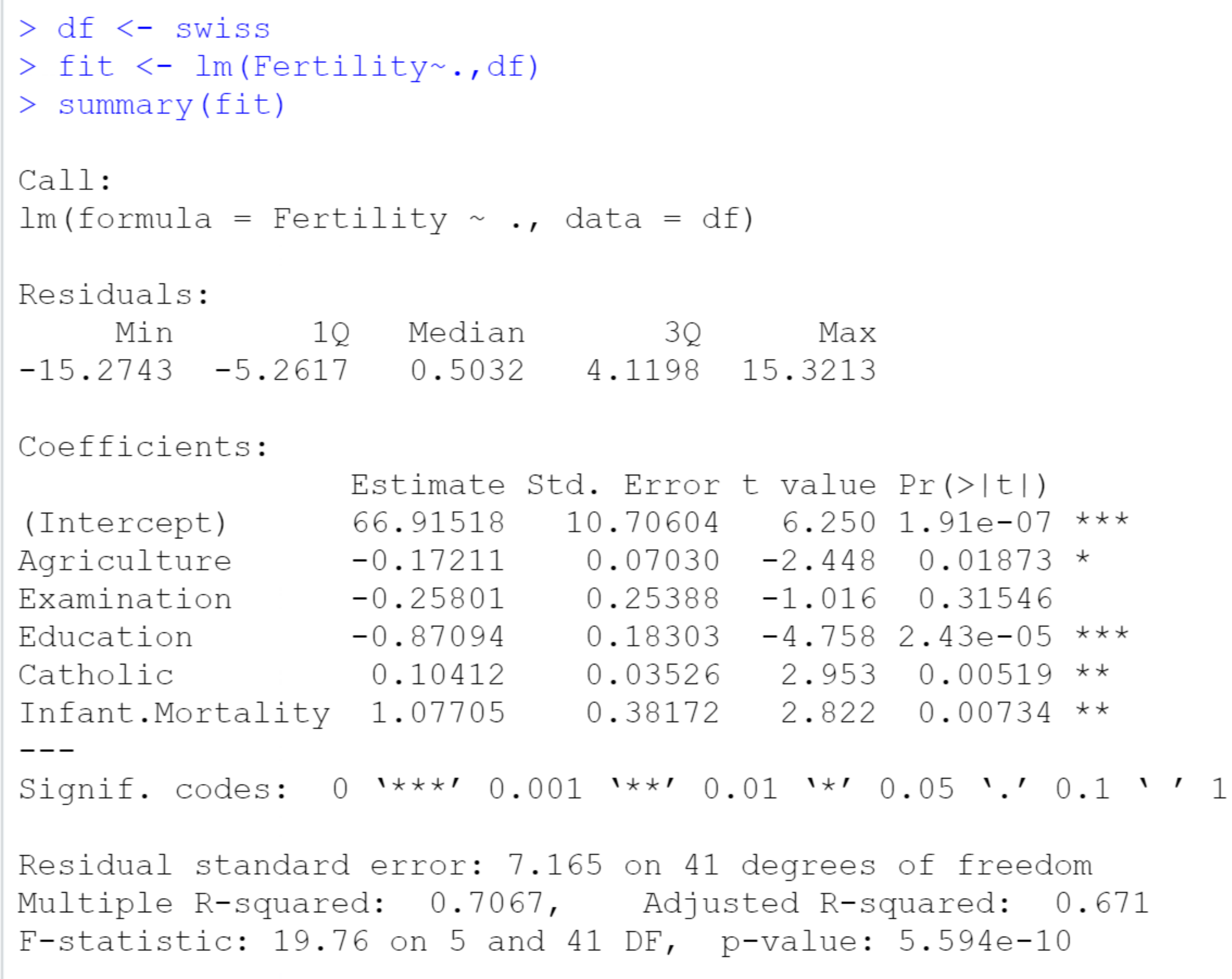


|  | Fertility | Agriculture | Examination | Education | Catholic | Infant.Mortality |
|---|---|---|---|---|---|---|
| **Courtelary** | 80.2 | 17.0 | 15 | 12 | 9.96 | 22.2 |
| **Delemont** | 83.1 | 45.1 | 6 | 9 | 84.84 | 22.2 |
| **Franches-Mnt** | 92.5 | 39.7 | 5 | 5 | 93.40 | 20.2 |
| **Moutier** | 85.8 | 36.5 | 12 | 7 | 33.77 | 20.3 |
| **Neuveville** | 76.9 | 43.5 | 17 | 15 | 5.16 | 20.6 |
| **Porrentruy** | 76.1 | 35.3 | 9 | 7 | 90.57 | 26.6 |
| **Broye** | 83.8 | 70.2 | 16 | 7 | 92.85 | 23.6 |
| **Glane** | 92.4 | 67.8 | 14 | 8 | 97.16 | 24.9 |

- Данные 1888-го года по регионам,
- Fertility — это количество детей до пяти лет, делённое на количество женщин до 50-ти лет и отмасштабированное, на 1000 домноженное;
- Agriculture — это процент мужчин, занятых в сельском хозяйстве;
- Examination — процент тех, кто получил высокий результат оценки на призывном пункте
- Catholic — процент католиков (?)
- Infant.Mortality — Смертность младенцев

# Учет всех параметров

```
fit <- lm(Fertility ~ ., df)
```

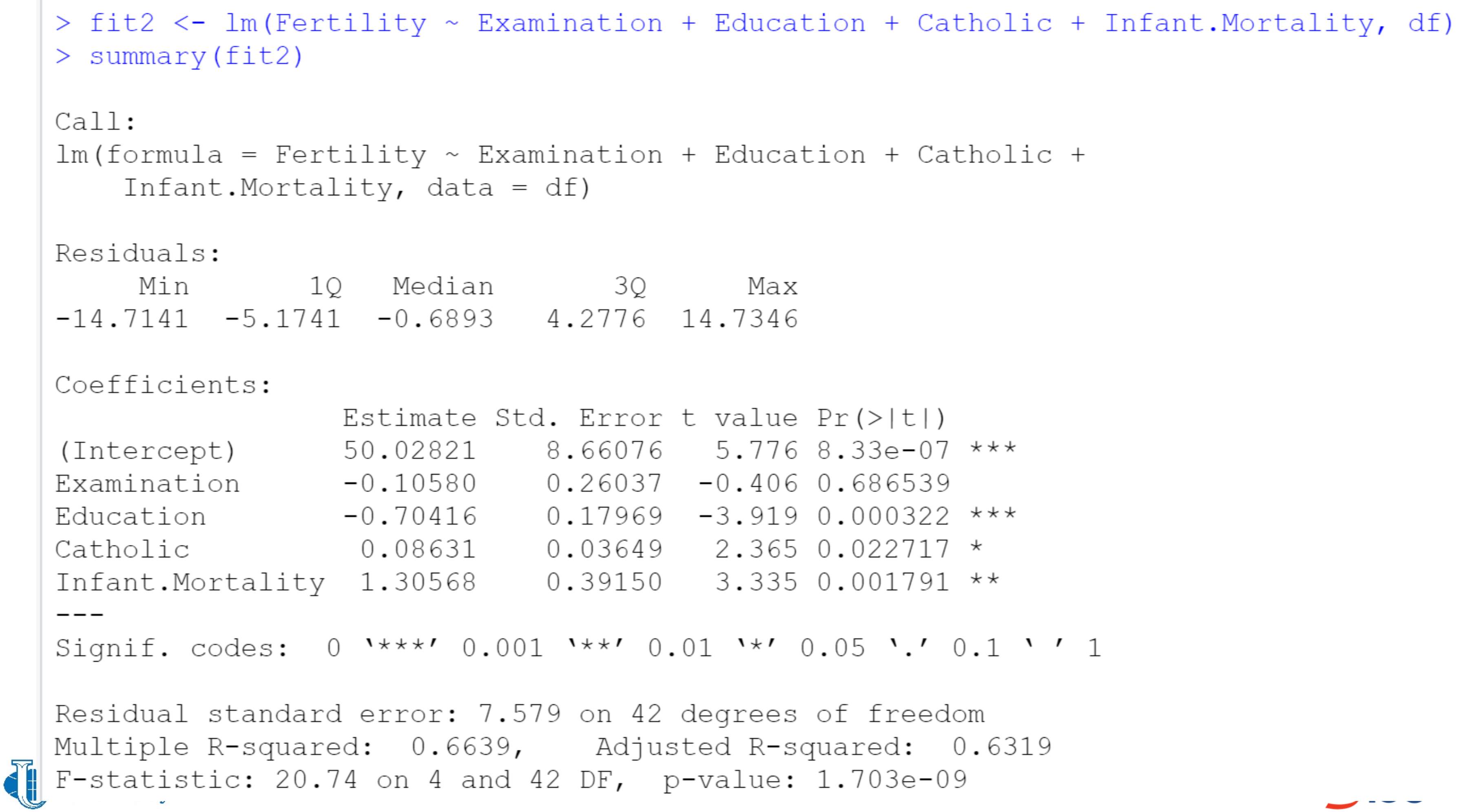


```
> df <- swiss
> fit <- lm(Fertility~.,df)
> summary(fit)

Call:
lm(formula = Fertility ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2743  -5.2617   0.5032   4.1198  15.3213 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *  
Examination      -0.25801    0.25388  -1.016  0.31546    
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 ** 
Infant.Mortality  1.07705    0.38172   2.822  0.00734 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,    Adjusted R-squared:  0.671 
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

# Исключаем параметр

fit2 <- lm(Fertility ~ Examination + Education + Catholic + Infant.Mortality, df)

исключено Agriculture



```
> fit2 <- lm(Fertility ~ Examination + Education + Catholic + Infant.Mortality, df)
> summary(fit2)

Call:
lm(formula = Fertility ~ Examination + Education + Catholic + 
    Infant.Mortality, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.7141  -5.1741  -0.6893   4.2776  14.7346 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      50.02821    8.66076   5.776 8.33e-07 ***
Examination      -0.10580    0.26037  -0.406 0.686539    
Education        -0.70416    0.17969  -3.919 0.000322 ***
Catholic          0.08631    0.03649   2.365 0.022717 *  
Infant.Mortality  1.30568    0.39150   3.335 0.001791 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.579 on 42 degrees of freedom
Multiple R-squared:  0.6639,	Adjusted R-squared:  0.6319 
F-statistic: 20.74 on 4 and 42 DF,  p-value: 1.703e-09
```

# Дисперсионный анализ

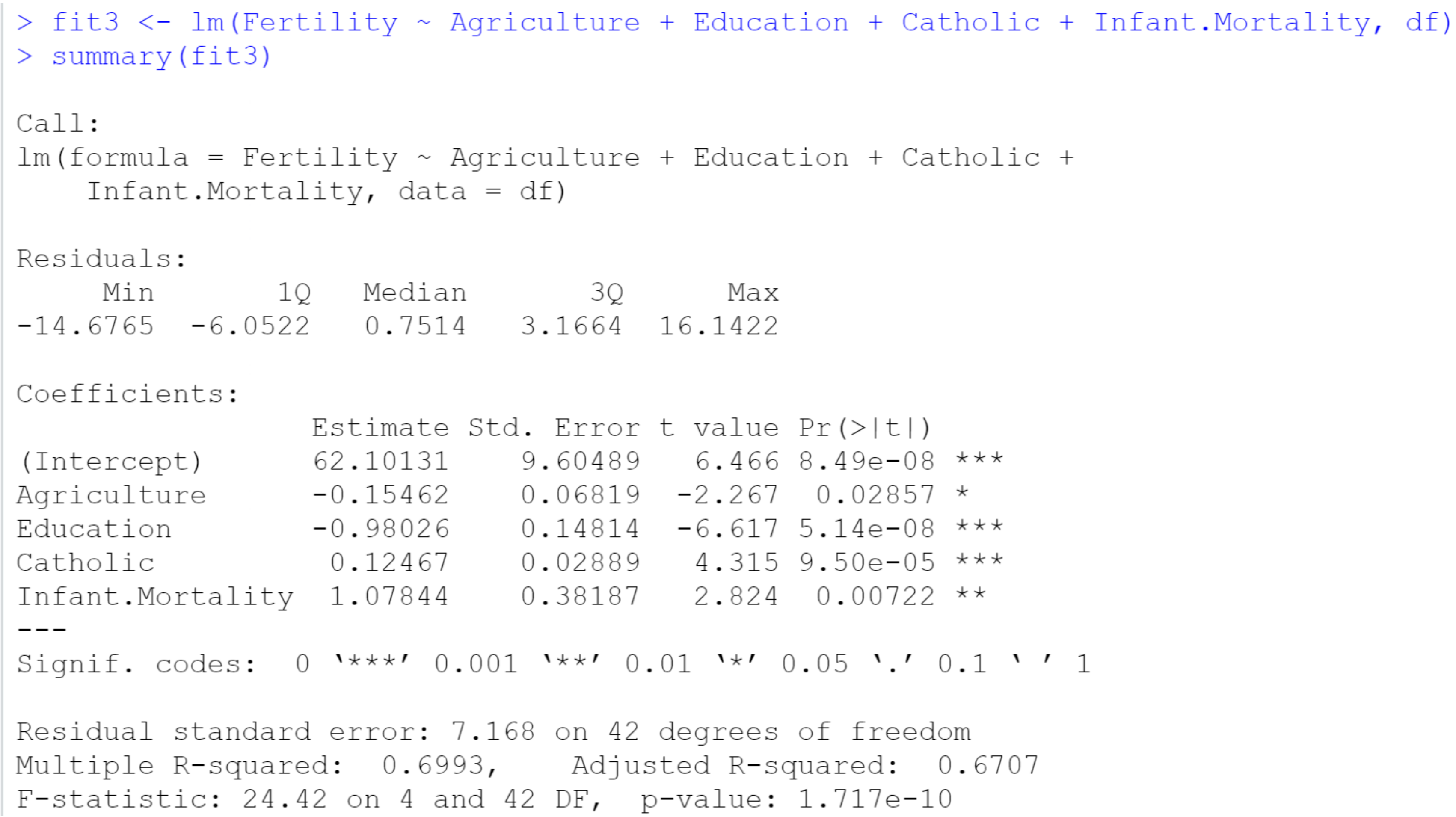
```
> anova(fit,fit2)
Analysis of Variance Table

Model 1: Fertility ~ Agriculture + Examination + Education + Catholic + 
    Infant.Mortality
Model 2: Fertility ~ Examination + Education + Catholic + Infant.Mortality
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
1     41 2105.0                              
2     42 2412.8 -1   -307.72 5.9934 0.01873 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Модели значимо различаются

# Исключаем параметр

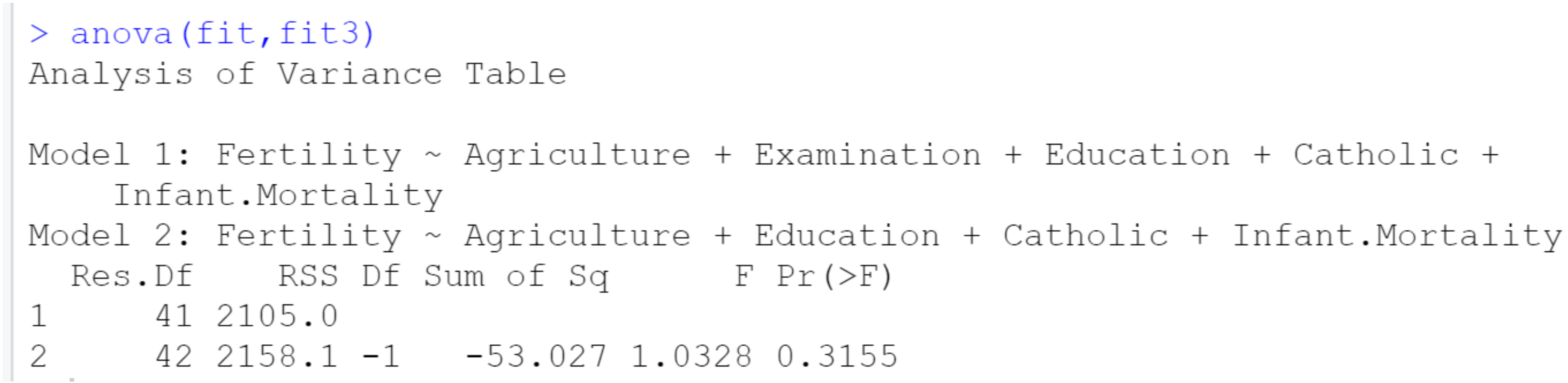
fit3 <- lm(Fertility ~ Agriculture + Education + Catholic + Infant.Mortality, df)

исключено Examination



```
> fit3 <- lm(Fertility ~ Agriculture + Education + Catholic + Infant.Mortality, df)
> summary(fit3)

Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic + 
    Infant.Mortality, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.6765  -6.0522   0.7514   3.1664  16.1422 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
Agriculture      -0.15462    0.06819  -2.267  0.02857 *  
Education        -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic          0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality  1.07844    0.38187   2.824  0.00722 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-squared:  0.6993,	Adjusted R-squared:  0.6707 
F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

# Дисперсионный анализ



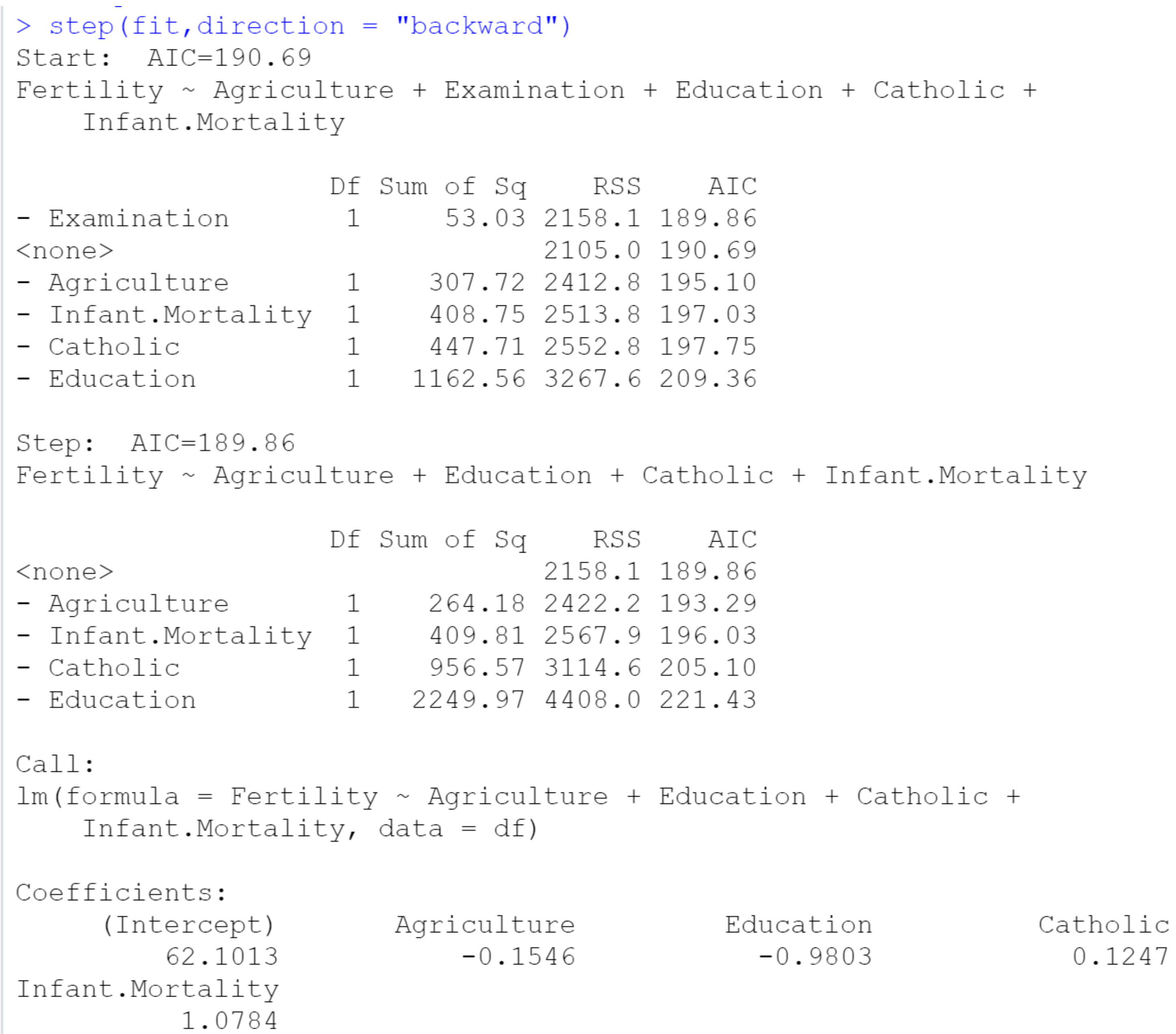
```
> anova(fit,fit3)
Analysis of Variance Table

Model 1: Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality
Model 2: Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     41 2105.0
2     42 2158.1 -1   -53.027 1.0328 0.3155
```

Модели не различаются

# Автоматический подбор предикторов – step()

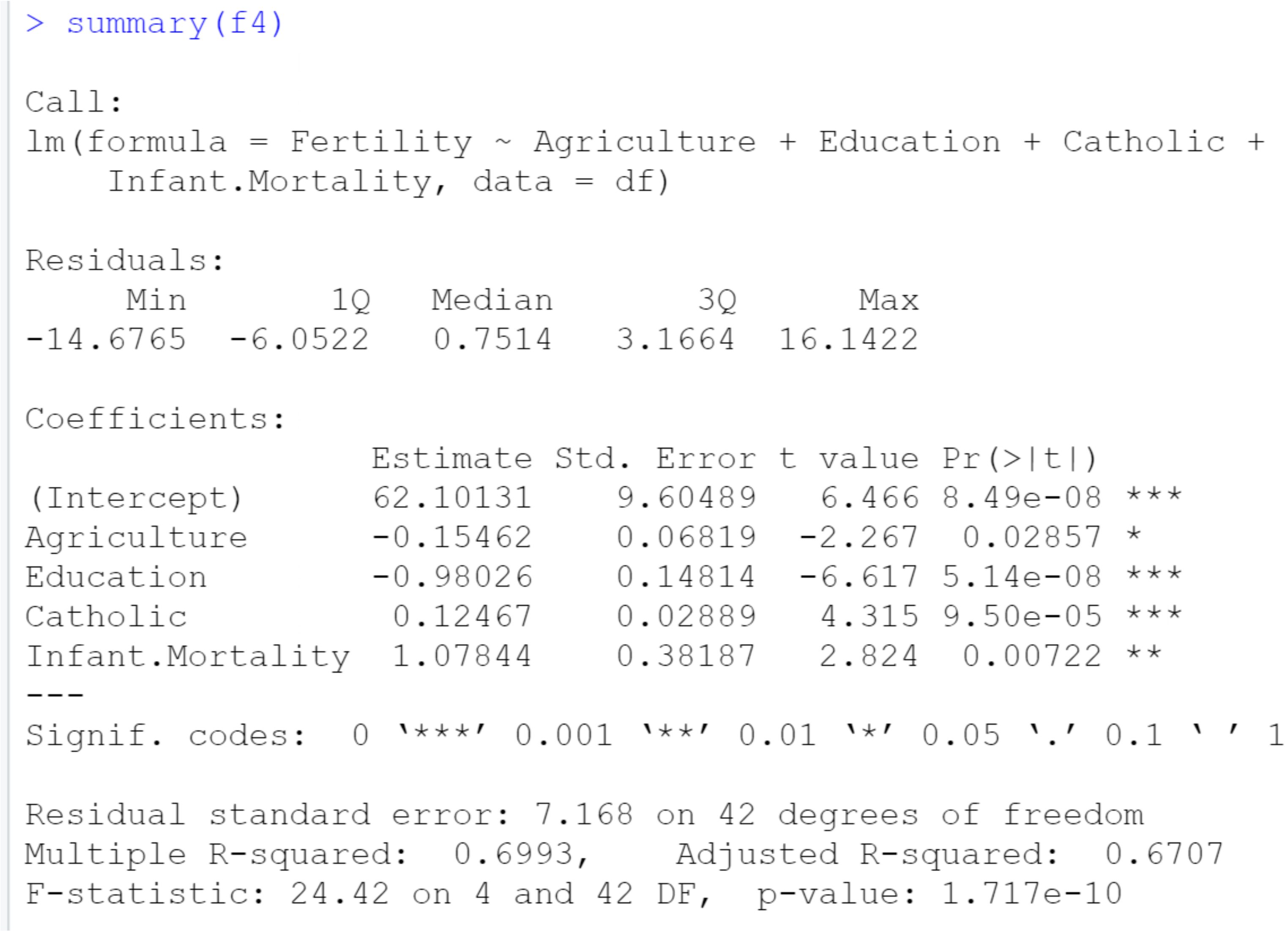
```
step(object, scope, scale = 0,
     direction = c("both", "backward", "forward"),
     trace = 1, keep = NULL, steps = 1000, k = 2, ...)
```

```
step(fit,direction = "backward")
```



```
> step(fit,direction = "backward")
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic + 
    Infant.Mortality

                   Df Sum of Sq    RSS    AIC
- Examination       1     53.03 2158.1 189.86
<none>                          2105.0 190.69
- Agriculture       1    307.72 2412.8 195.10
- Infant.Mortality  1    408.75 2513.8 197.03
- Catholic          1    447.71 2552.8 197.75
- Education         1   1162.56 3267.6 209.36

Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality

                   Df Sum of Sq    RSS    AIC
<none>                          2158.1 189.86
- Agriculture       1    264.18 2422.2 193.29
- Infant.Mortality  1    409.81 2567.9 196.03
- Catholic          1    956.57 3114.6 205.10
- Education         1   2249.97 4408.0 221.43

Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic + 
    Infant.Mortality, data = df)

Coefficients:
     (Intercept)       Agriculture         Education          Catholic  
         62.1013           -0.1546           -0.9803            0.1247  
Infant.Mortality  
          1.0784  
```

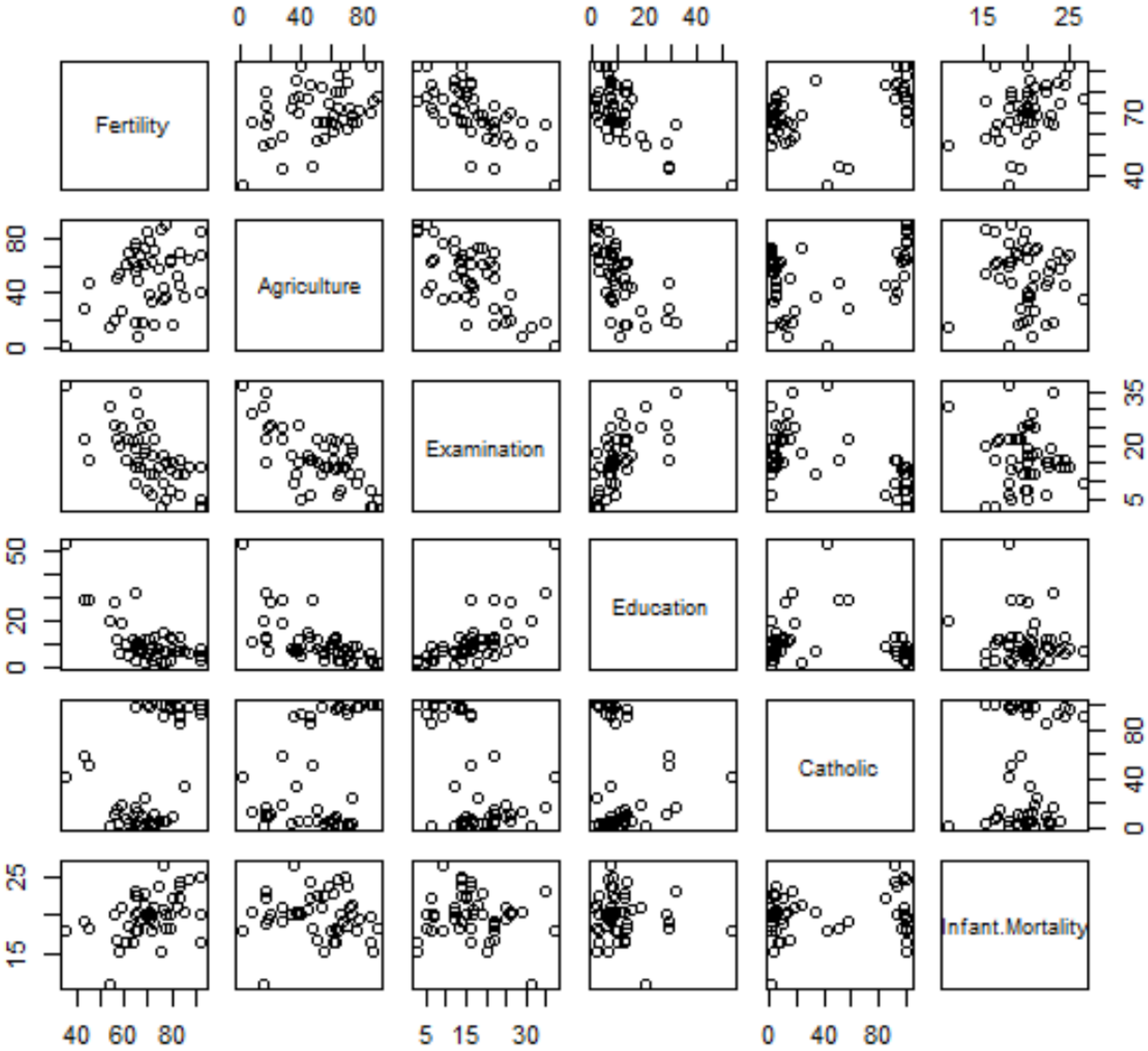
# Запись результатов в переменную



```
> summary(f4)

Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic + 
    Infant.Mortality, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.6765  -6.0522   0.7514   3.1664  16.1422 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
Agriculture      -0.15462    0.06819  -2.267  0.02857 *  
Education        -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic          0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality  1.07844    0.38187   2.824  0.00722 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-squared:  0.6993,    Adjusted R-squared:  0.6707 
F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

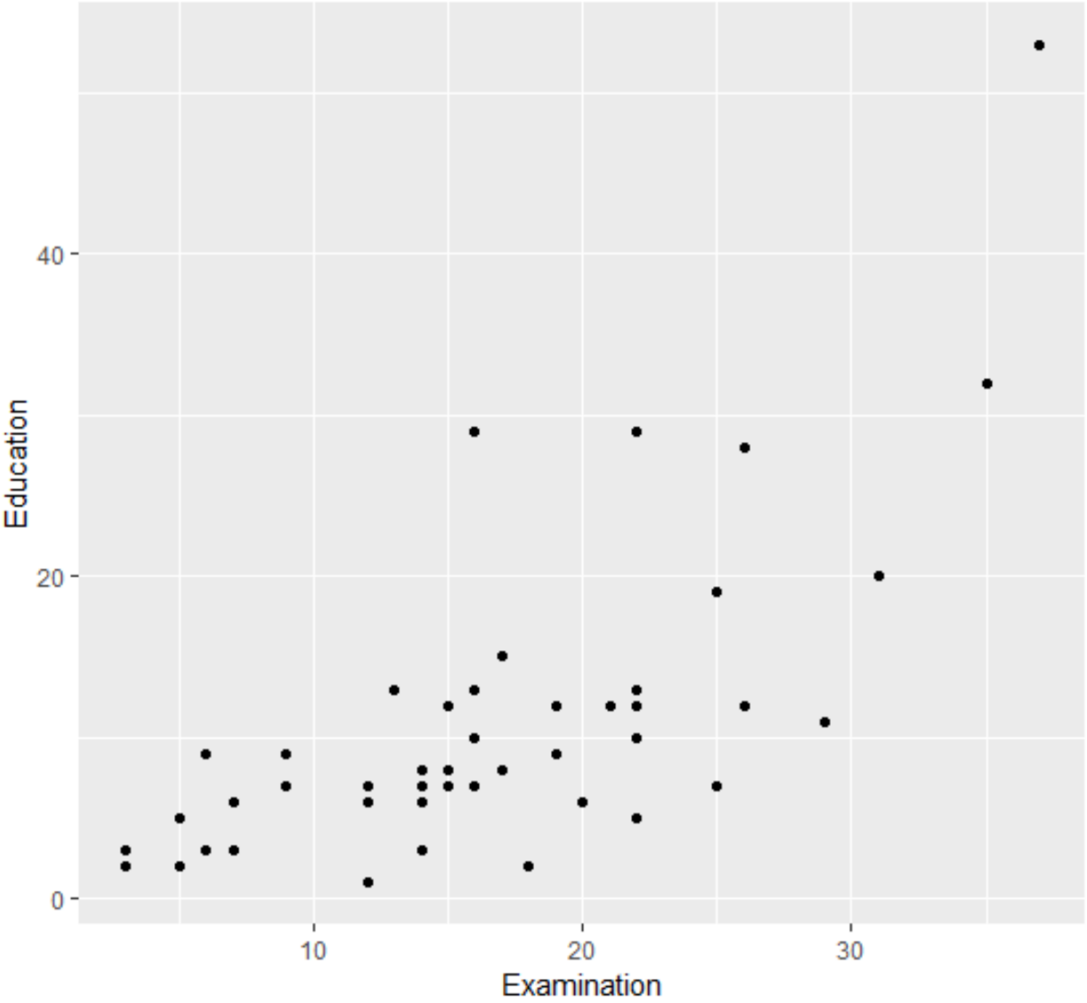
# Диагностика модели

pairs(df)


Fertility
Agriculture
Examination
Education
Catholic
Infant.Mortality
0
40
80
0
20
40
15
25
70
40
80
40
0
35
20
5
50
20
0
80
40
0
25
15
40
60
80
5
15
30
0
40
80

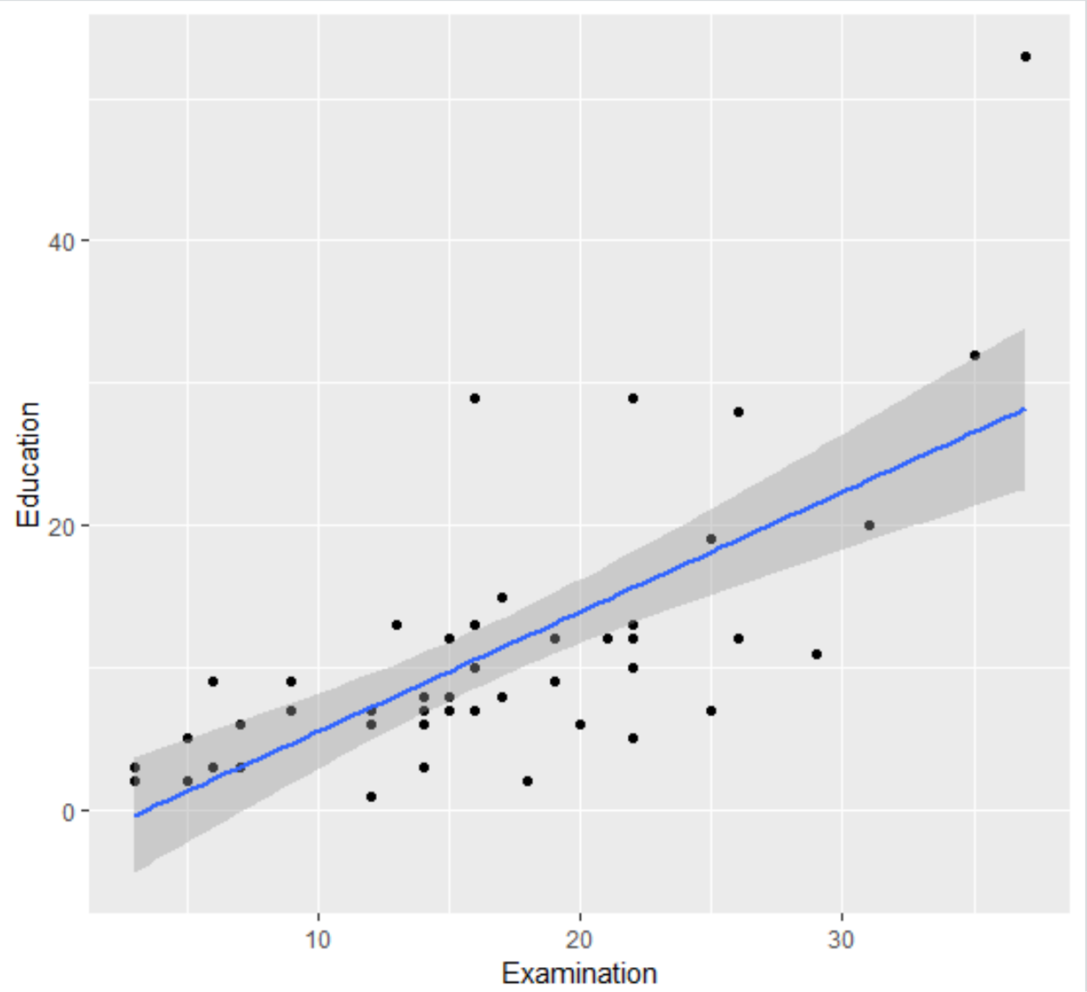
# Два параметра

```
library(ggplot2)
ggplot(df, aes(x = Examination, y = Education))+
  geom_point()
```



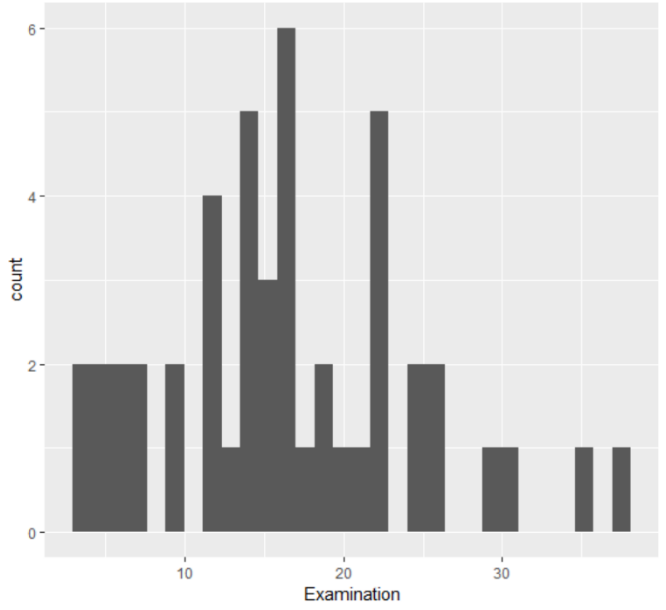
# Два параметра

```
ggplot(df, aes(x = Examination, y = Education))+
  geom_point()+
  geom_smooth(method = "lm")
```

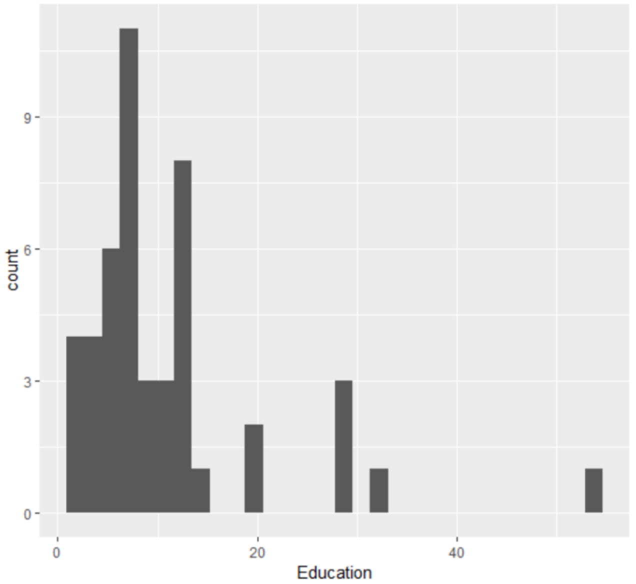


# Анализ распределения

```
ggplot(df, aes(x = Examination))+
geom_histogram()
```

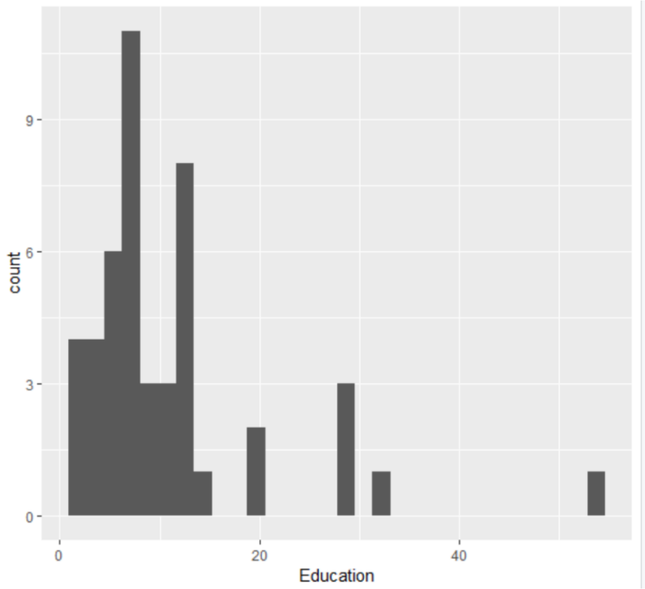


```
ggplot(df, aes(x = Education))+
geom_histogram()
```

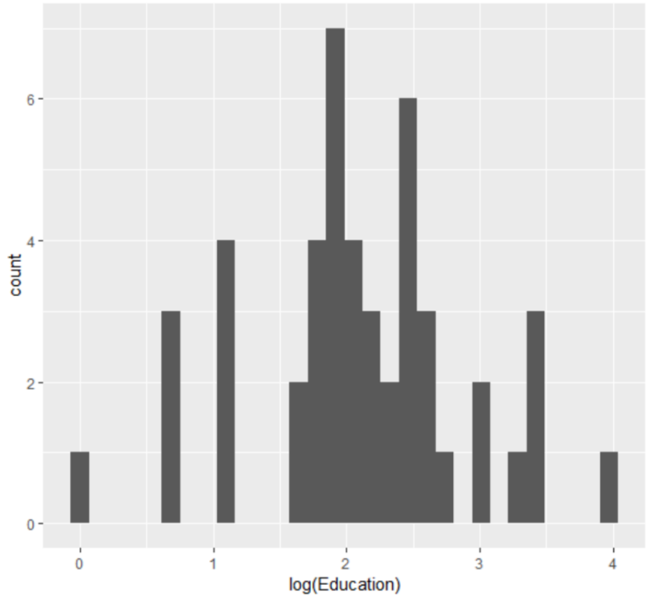


# Улучшение распределения

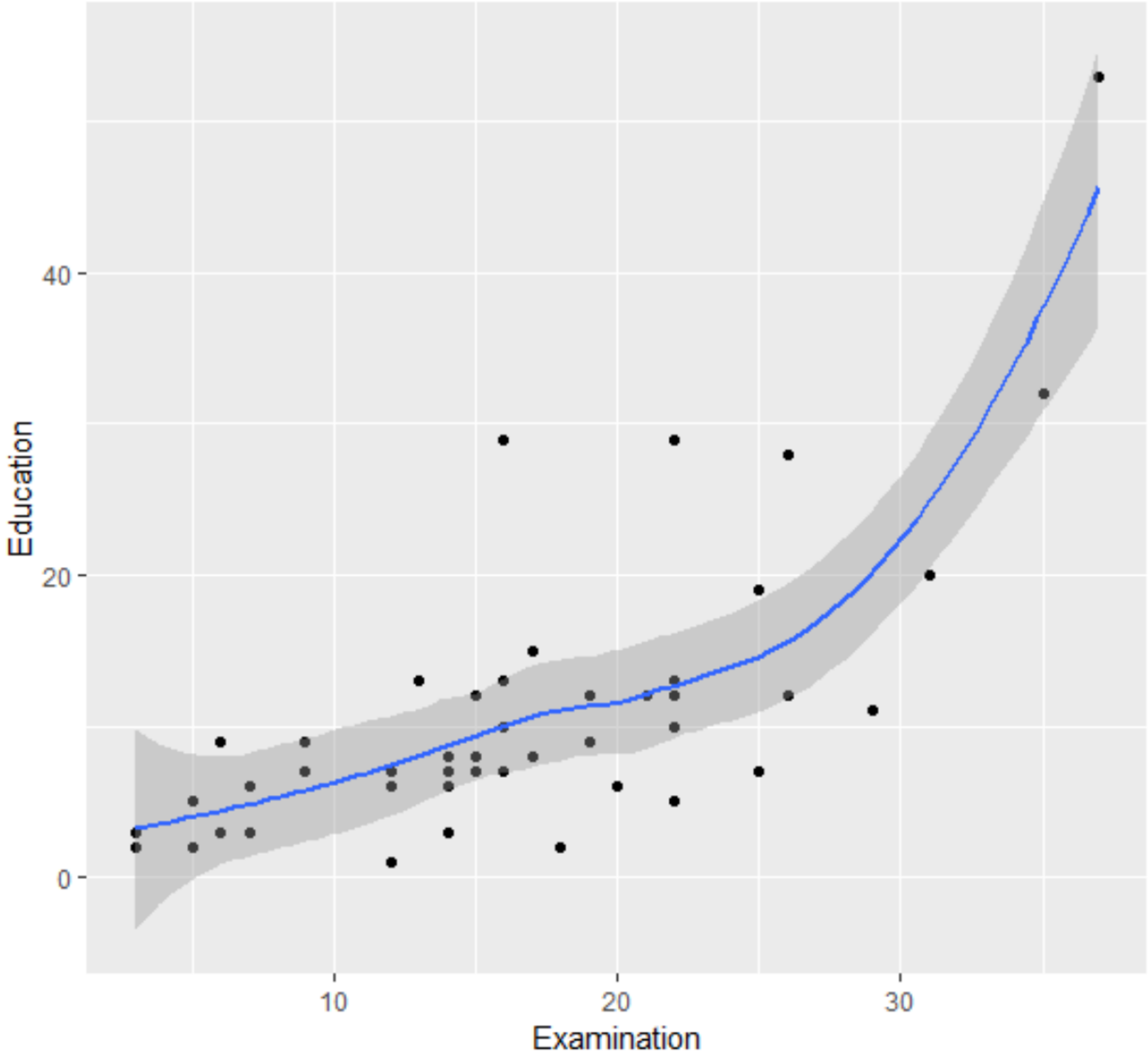
```
ggplot(df, aes(x = Education))+
geom_histogram()
```



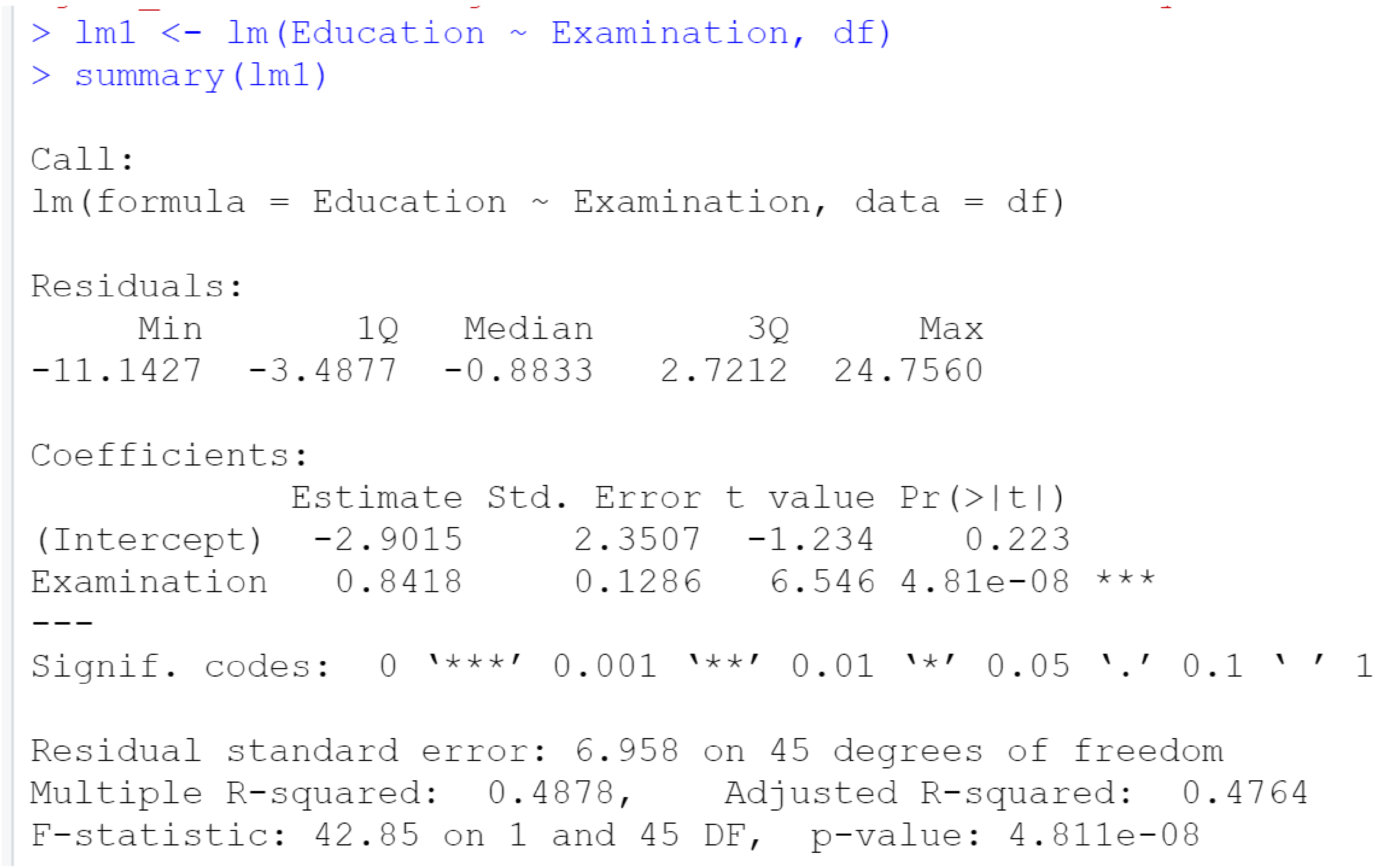
```
ggplot(df, aes(x = log(Education)))+
  geom_histogram()
```



```
ggplot(df, aes(x = Examination, y = Education))+
  geom_point()+
  geom_smooth()
```

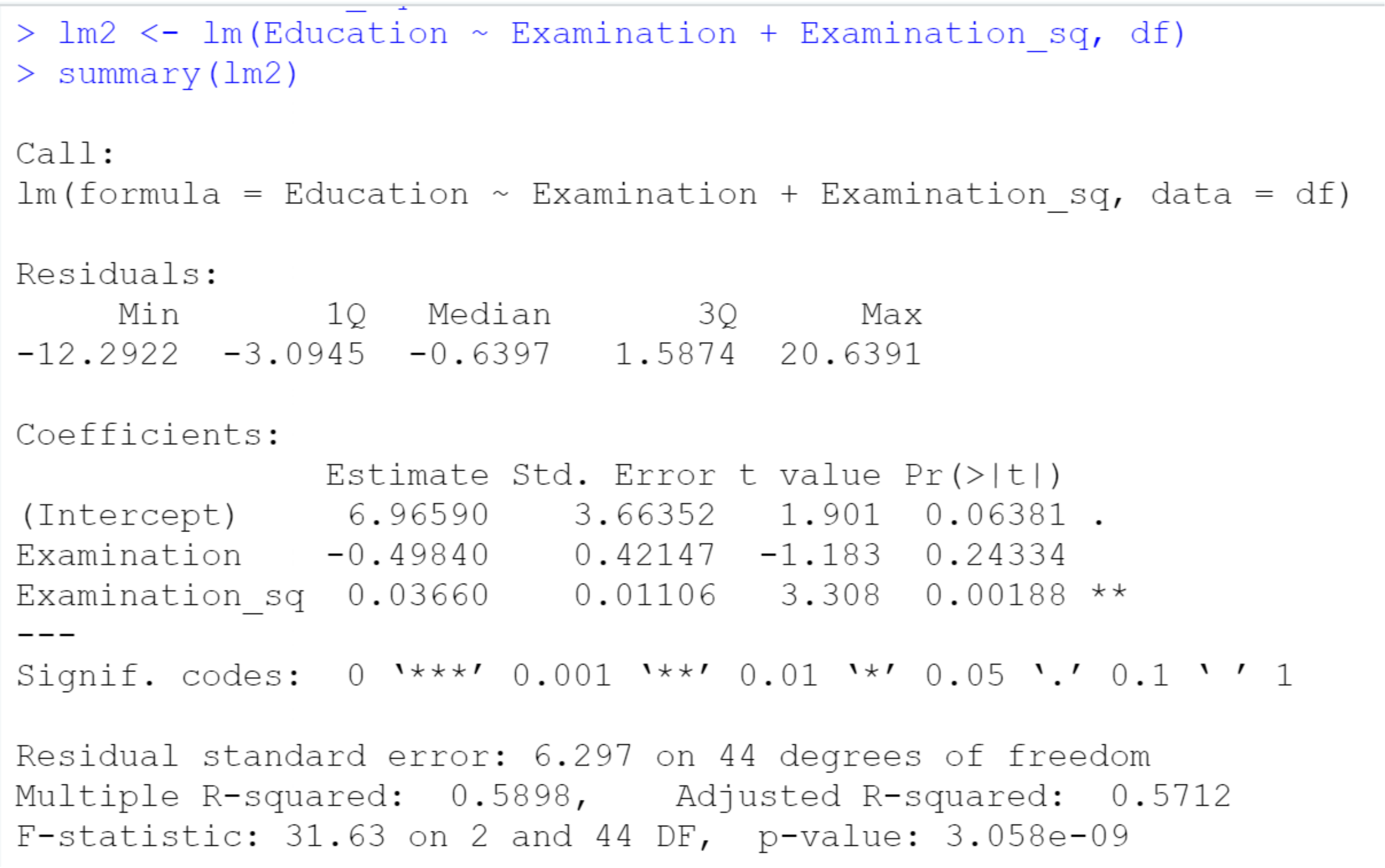


# Базовая модель



```
> lm1 <- lm(Education ~ Examination, df)
> summary(lm1)

Call:
lm(formula = Education ~ Examination, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1427  -3.4877  -0.8833   2.7212  24.7560 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.9015     2.3507  -1.234    0.223    
Examination   0.8418     0.1286   6.546 4.81e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.958 on 45 degrees of freedom
Multiple R-squared:  0.4878,    Adjusted R-squared:  0.4764 
F-statistic: 42.85 on 1 and 45 DF,  p-value: 4.811e-08
```

# Улучшенная модель



```
> lm2 <- lm(Education ~ Examination + Examination_sq, df)
> summary(lm2)

Call:
lm(formula = Education ~ Examination + Examination_sq, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.2922  -3.0945  -0.6397   1.5874  20.6391 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)   
(Intercept)     6.96590    3.66352   1.901  0.06381 . 
Examination    -0.49840    0.42147  -1.183  0.24334   
Examination_sq  0.03660    0.01106   3.308  0.00188 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.297 on 44 degrees of freedom
Multiple R-squared:  0.5898,	Adjusted R-squared:  0.5712 
F-statistic: 31.63 on 2 and 44 DF,  p-value: 3.058e-09
```

# Дисперсионный анализ

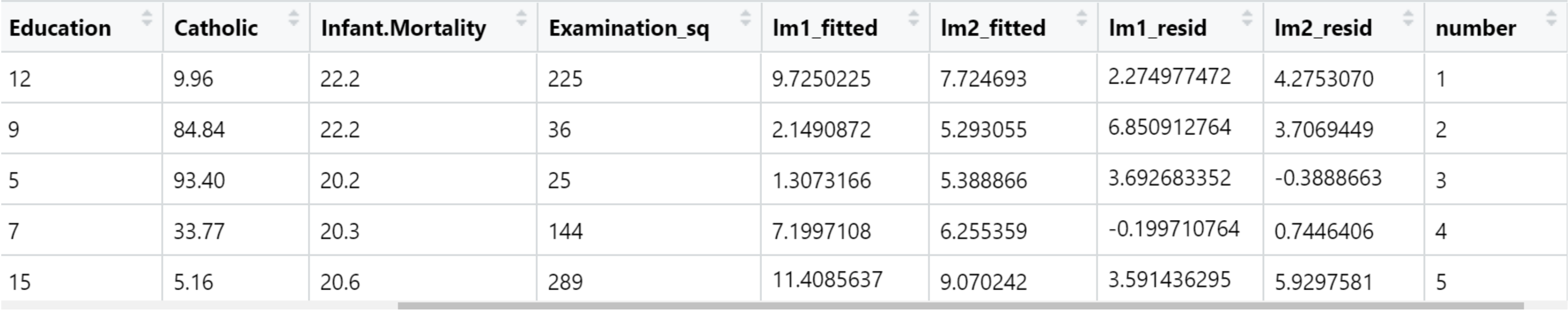
```
> anova(lm2,lm1)
Analysis of Variance Table

Model 1: Education ~ Examination + Examination_sq
Model 2: Education ~ Examination
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     44 1744.5
2     45 2178.4 -1   -433.95 10.945 0.001877 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Модели значимо различаются

# Добавление значений

Формирование набора из предсказанных значений и разницы

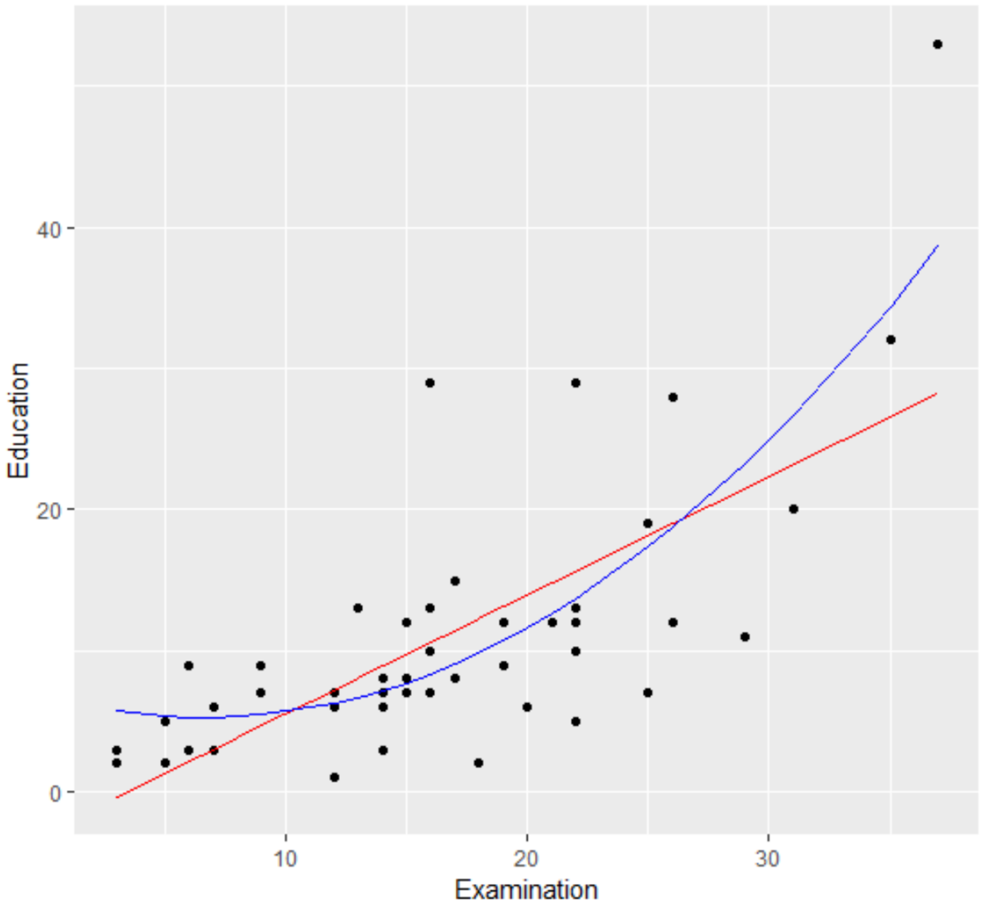
```
df$lm1_fitted <- lm1$fitted
df$lm2_fitted <- lm2$fitted
df$lm1_resid <- lm1$resid
df$lm2_resid <- lm2$resid
df$number <- 1:nrow(df)
```



| Education | Catholic | Infant.Mortality | Examination_sq | lm1_fitted | lm2_fitted | lm1_resid | lm2_resid | number |
|---|---|---|---|---|---|---|---|---|
| 12 | 9.96 | 22.2 | 225 | 9.7250225 | 7.724693 | 2.274977472 | 4.2753070 | 1 |
| 9 | 84.84 | 22.2 | 36 | 2.1490872 | 5.293055 | 6.850912764 | 3.7069449 | 2 |
| 5 | 93.40 | 20.2 | 25 | 1.3073166 | 5.388866 | 3.692683352 | -0.3888663 | 3 |
| 7 | 33.77 | 20.3 | 144 | 7.1997108 | 6.255359 | -0.199710764 | 0.7446406 | 4 |
| 15 | 5.16 | 20.6 | 289 | 11.4085637 | 9.070242 | 3.591436295 | 5.9297581 | 5 |

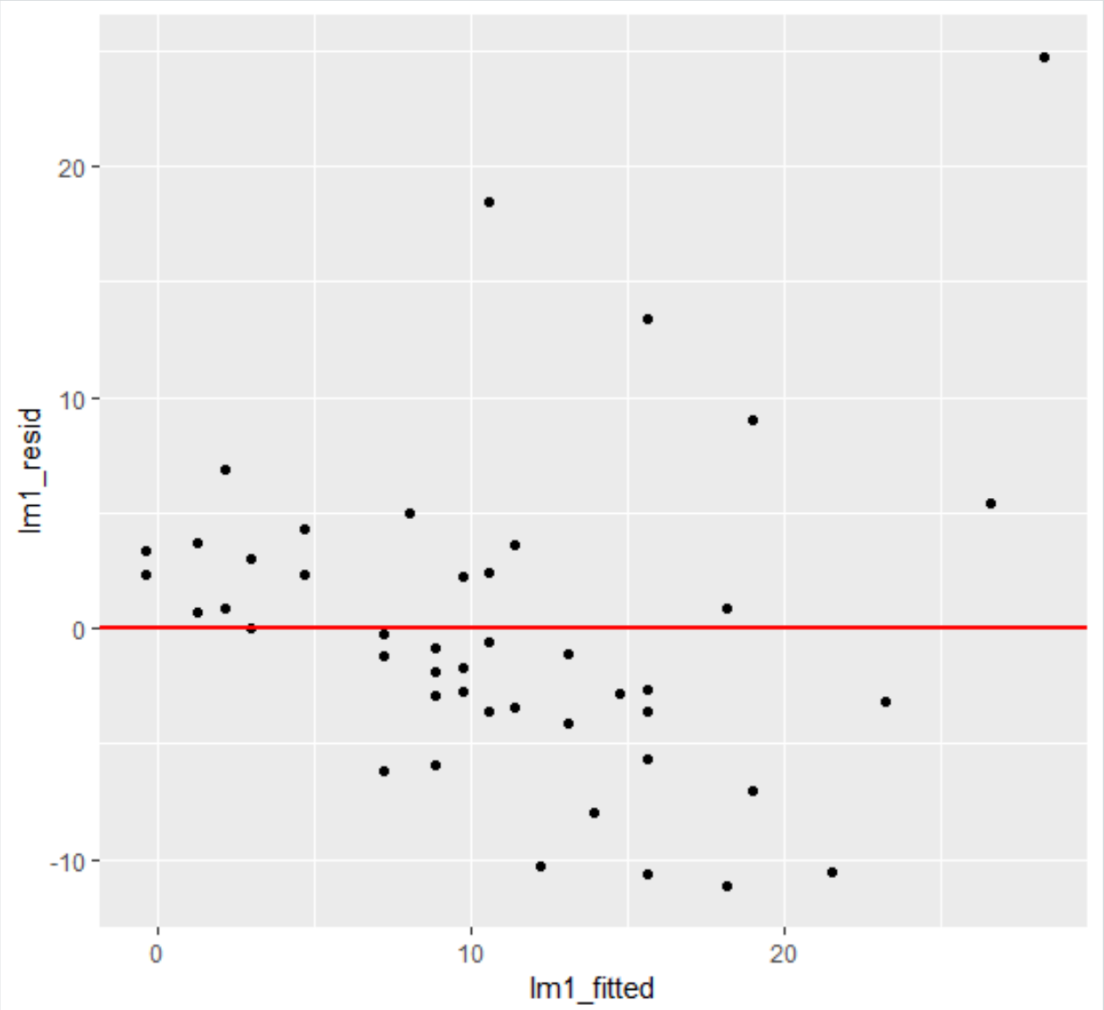
# Реальные значения + предсказанные

```
ggplot(df, aes(x = Examination, y = Education))+
  geom_point()+
  geom_line(aes(x = Examination, y = lm1_fitted), col = "red")+
  geom_line(aes(x = Examination, y = lm2_fitted), col = "blue")
```



# Остатки

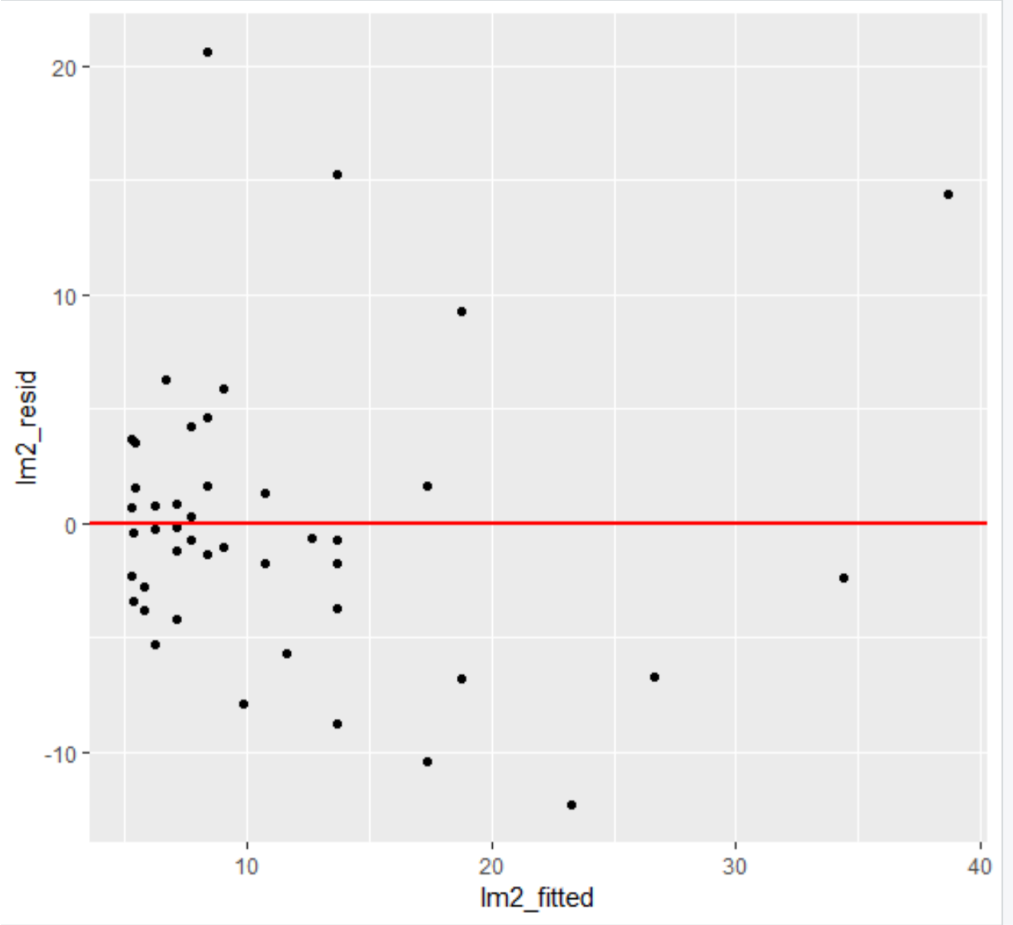
```
ggplot(df, aes(x = lm1_fitted, y = lm1_resid)) +
  geom_point() +
  geom_hline(yintercept=0,col="red", lwd=1)
```



Остатки распределяются неравномерно, есть выраженная тенденция

# Исправленные остатки

```
ggplot(df, aes(x = lm2_fitted, y = lm2_resid)) +
  geom_point() +
  geom_hline(yintercept=0,col="red", lwd=1)
```

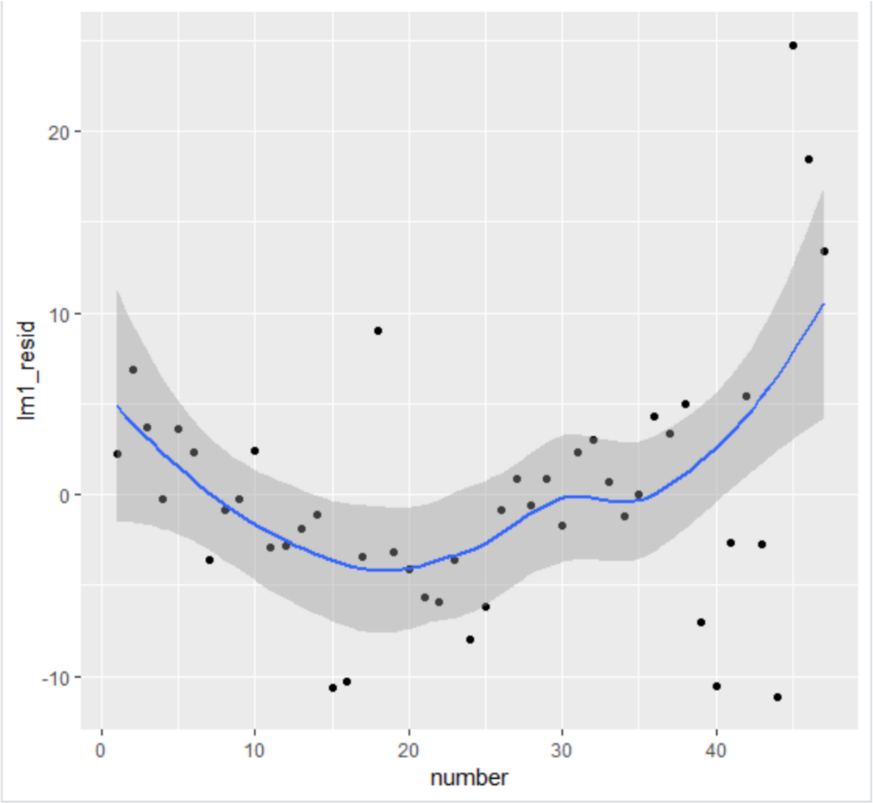


# Независимость остатков

Остатки могут быть сгруппированы из-за:

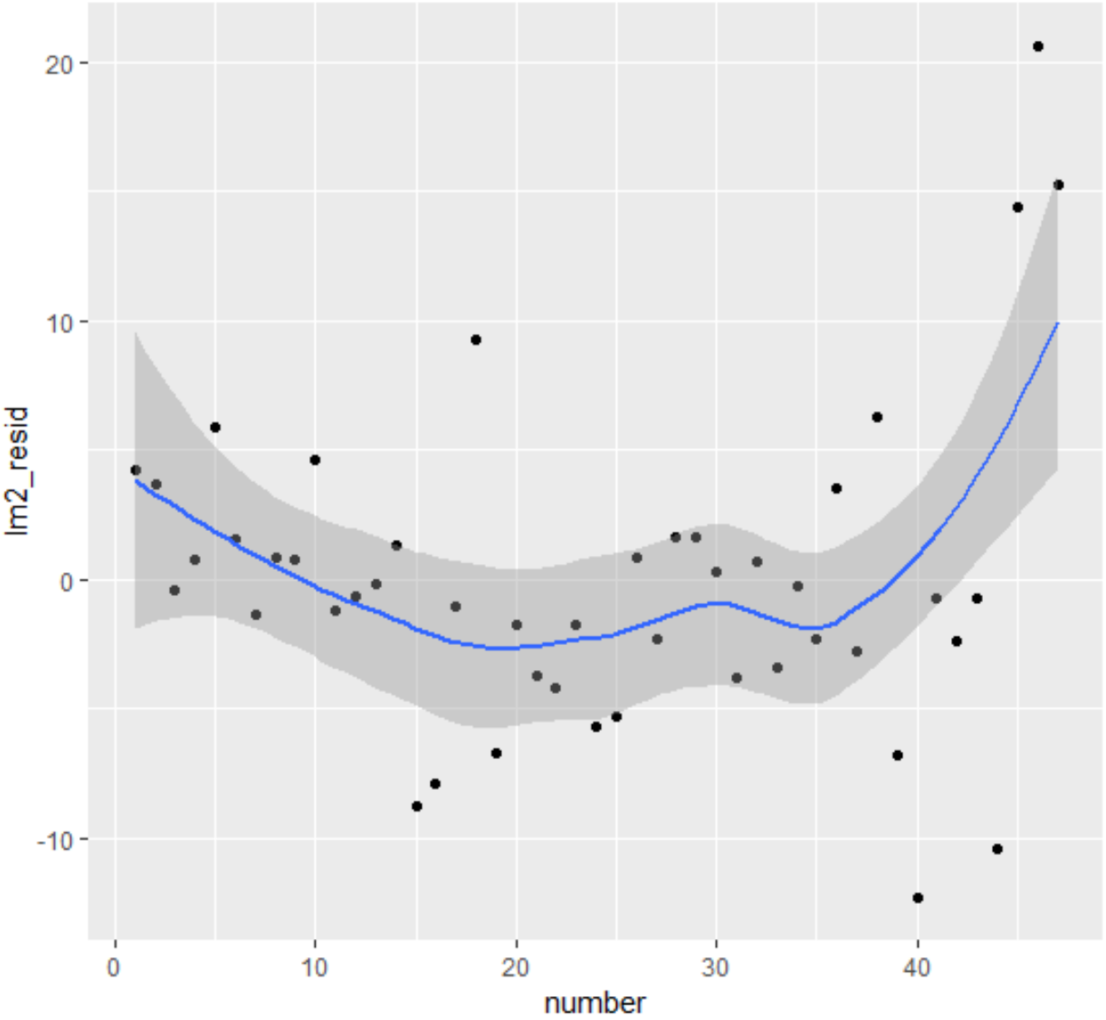
- исследования двух разных наблюдения
- не группировки данных по испытуемым

```
ggplot(df, aes(x = number, y = lm1_resid)) +
geom_point() +
geom_smooth()
```



# Независимость остатков

```
ggplot(df, aes(x = number, y = lm1_resid)) +
geom_point() +
geom_smooth()
```

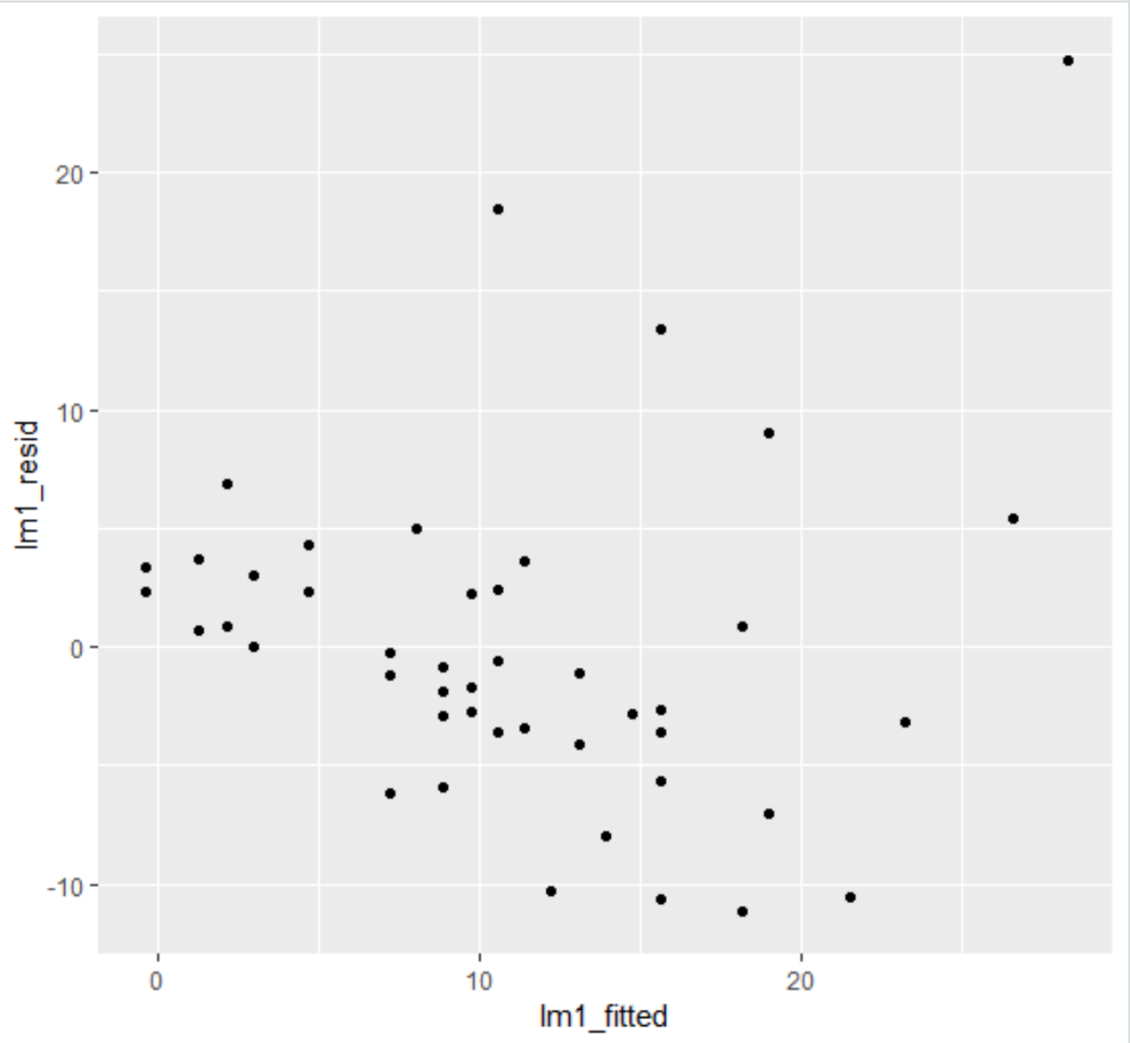


Распределение более сглаженное

# Разброс остатков

```
ggplot(df, aes(x = lm1_fitted, y = lm1_resid)) +
  geom_point()
```

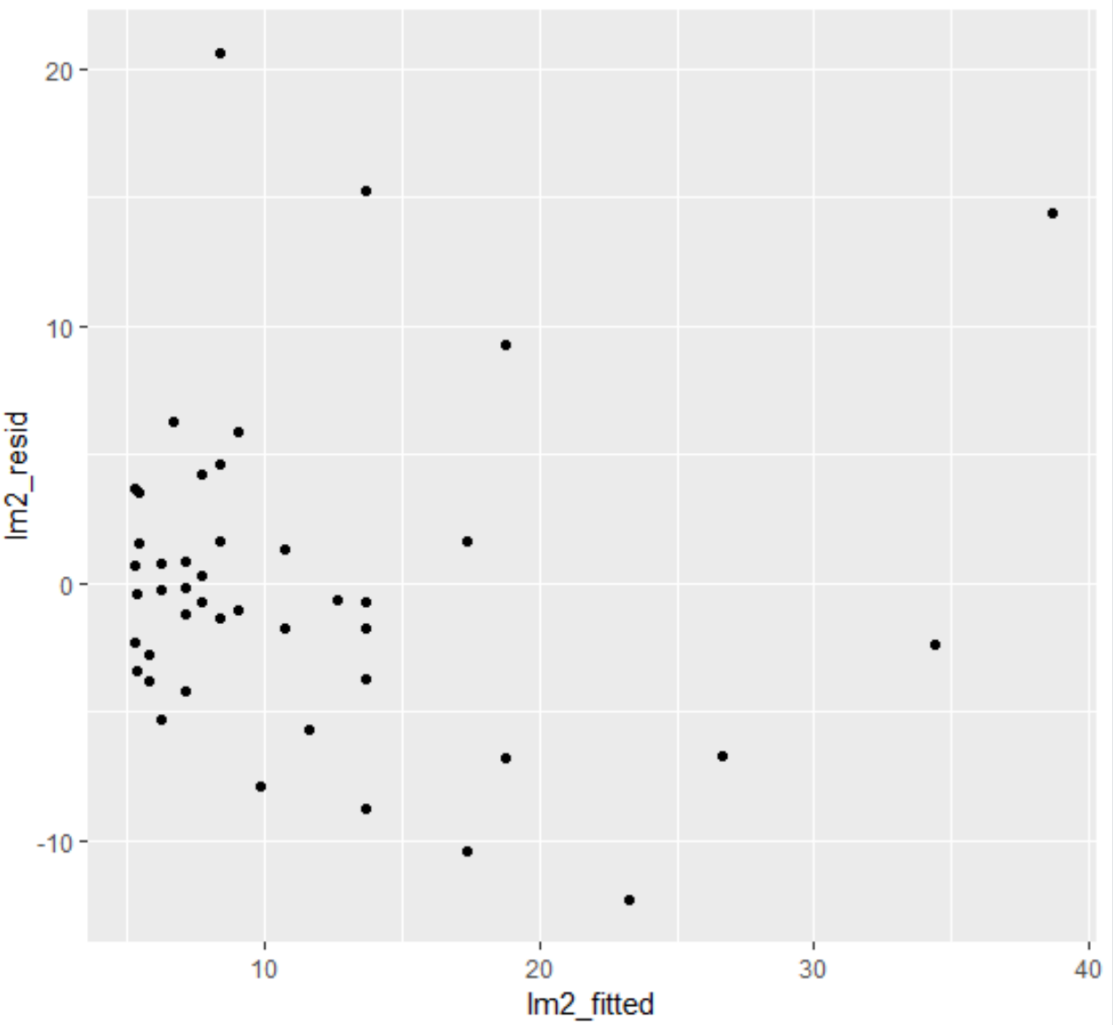
В разбросах нет равномерности



# Разброс остатков

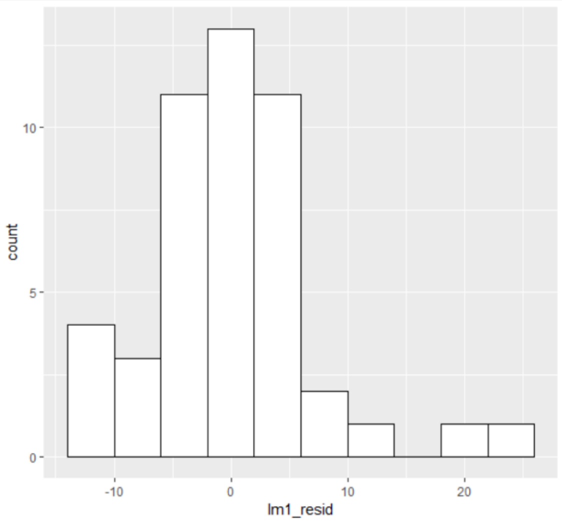
```
ggplot(df, aes(x = lm2_fitted, y = lm2_resid)) +
  geom_point()
```

Распределение более упорядоченное



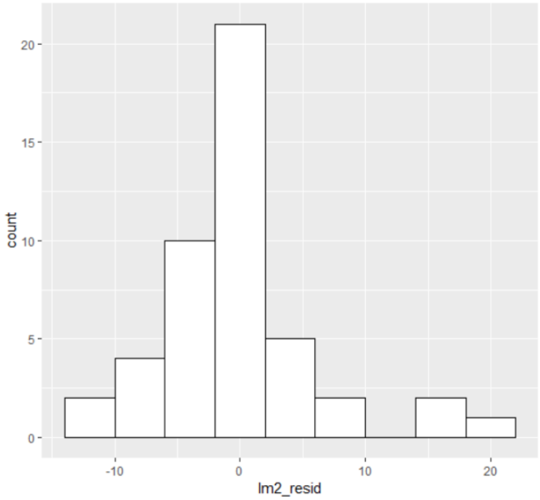
# Распределение остатков

```
ggplot(df, aes(x = lm1_resid)) +
  geom_histogram(binwidth=4,fill="white",col="black'
```



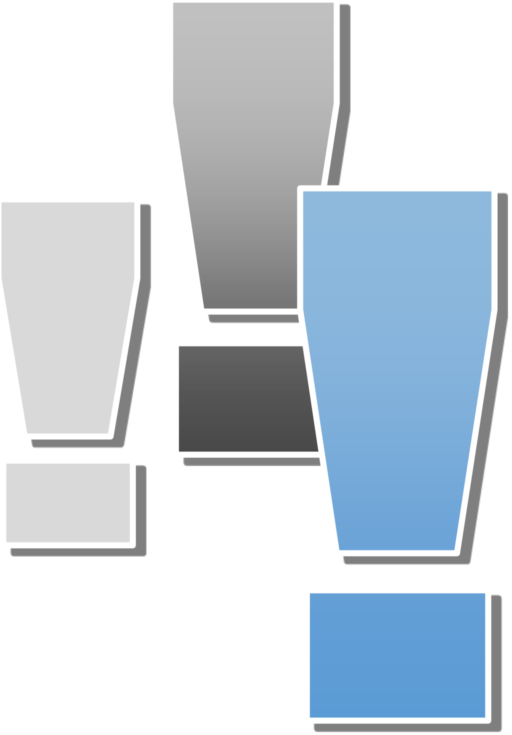
Распределение
скошено влево

```
ggplot(df, aes(x = lm2_resid)) +
  geom_histogram(binwidth=4,fill="white",col="black")
```



Распределение улучшилось, хотя не идеально

# Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university
+7(903)144-53-57