НИУ ВШЭ, ОП «Политология», 2022
Курс «Теория вероятностей и математическая статистика»

Домашнее задание 2
Владислав Рубанов БПТ201

Задание 1.

1.
```
marks <- read.table("reg_hw.txt", header = TRUE, sep = ";")
group <- rep(1:4, c(25,35,35,25))
data <- data.frame(marks, group)

data <- transform(data, group=LETTERS[group])

head(data)
```

```
##   math philosophy english hist_ec micro group
## 1    7          8       7       7     8     A
## 2    8          8       8       8     7     A
## 3    9          7       7       8     8     A
## 4    5          7       6       6     6     A
## 5   10          8       6       6     5     A
## 6    6          8       6       8     7     A
```

   Мы создали в массиве новую переменную, показывающую принадлежность студента к учебной групппе (A, B, C и D).

2. Выберем оценки по микроэкономике ("micro"в нашем датасете) и опишем их.

```
library(psych)
library(dplyr)

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(car)

## Loading required package: carData
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
## The following object is masked from 'package:psych':
##
##     logit

library(ggplot2)

##
## Attaching package: 'ggplot2'
## The following objects are masked from 'package:psych':
##
##     %+%, alpha

group_by(data, group) %>% summarise(mean = mean(micro), var = var(micro), IQR = IQR

## # A tibble: 4 x 7
##   group  mean   var   IQR median   min   max
##   <chr> <dbl> <dbl> <dbl>  <int> <int> <int>
## 1 A      6.56  1.26     1      7     5     8
## 2 B      6.74  2.02     2      7     4    10
## 3 C      6.86  1.60     2      7     5    10
## 4 D      6.44  0.84     1      7     5     8

summary(data$micro)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.000   6.000   7.000   6.675   7.250  10.000

var(data$micro)

## [1] 1.481723

IQR(data$micro)

## [1] 1.25

leveneTest(micro ~ as.factor(group), data)

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  1.1985 0.3136
##       116

summary(data)
```

```
##       math          philosophy      english        hist_ec
##  Min.   : 4.000   Min.   : 5.0   Min.   :4.00   Min.   :4.000
##  1st Qu.: 7.000   1st Qu.: 7.0   1st Qu.:6.00   1st Qu.:6.000
##  Median : 7.000   Median : 8.0   Median :7.00   Median :7.000
##  Mean   : 7.392   Mean   : 8.2   Mean   :6.95   Mean   :6.958
##  3rd Qu.: 8.000   3rd Qu.: 9.0   3rd Qu.:8.00   3rd Qu.:8.000
##  Max.   :10.000   Max.   :10.0   Max.   :9.00   Max.   :9.000
##      micro           group
##  Min.   : 4.000   Length:120
##  1st Qu.: 6.000   Class :character
##  Median : 7.000   Mode  :character
##  Mean   : 6.675
##  3rd Qu.: 7.250
##  Max.   :10.000
```

Интерпретация

Как мы можем видеть, внутригрупповые средние четырех групп находятся в интервале от 6.44 до 6.86. Групповые средние во всех группах не сильно отличаются от среднего арифметического по всем группам (6.675). Медианное значение у всех групп одинаковое: это 7. Данное значение также соответствует медиане по всему массиву.

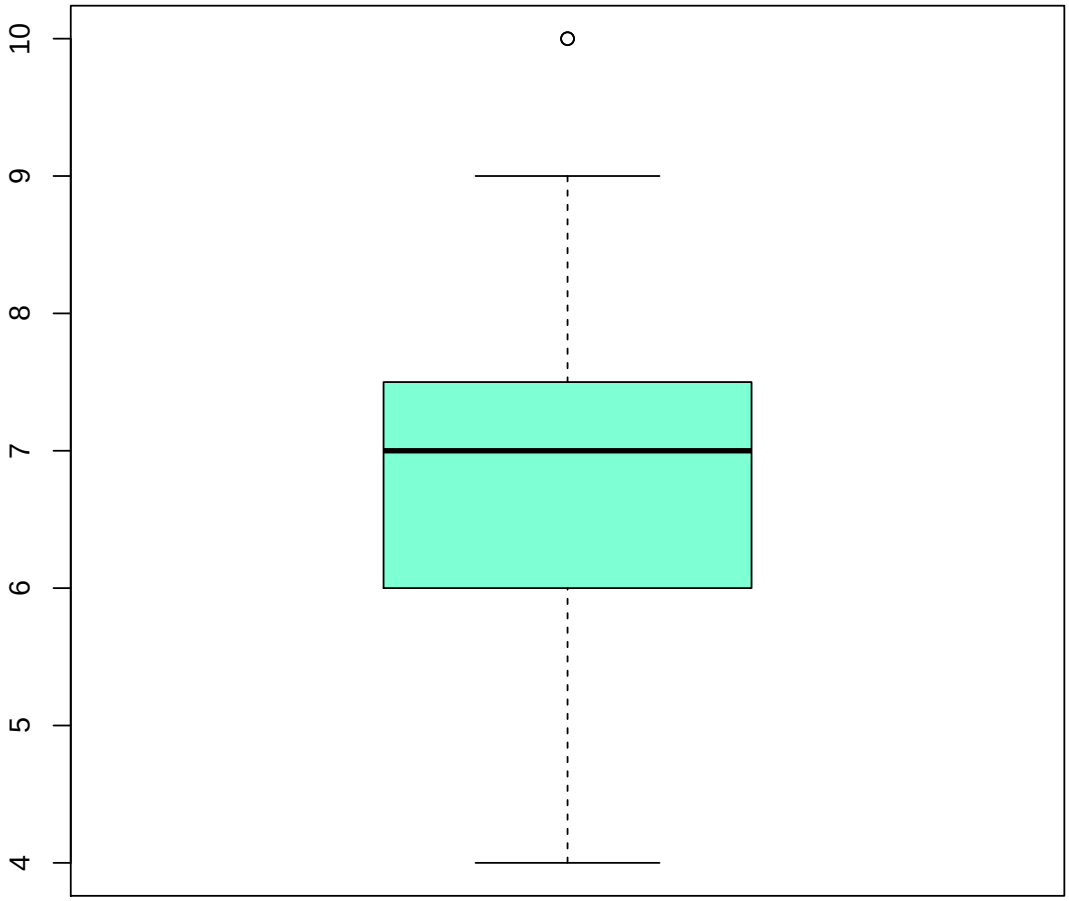
Дисперсия же различается больше. Так, она колеблется от 0.84 в группе D до 2.02 в группе B. IQR (аналог дисперсии) также показывает различие дисперсий между группами A и D с B и C в два раза (значение по всему массиву = 1.25. Значение дисперсии по всему массиву ≈ 1.48. Однако расчет теста Левена (полученное значение p-value = 0.3136) говорит о том, что мы можем говорить о равенстве дисперсий между подвыборками на статистическом уровне, что является важным в контексте ANOVA. Большое значение N позволяет нам рассчитать данный тест, основывающийся на нормальности распределения или асимптотике.

Минимальные и максимальные значения оценок по микроэкономике по группам показывают, что ни в одной из групп нет оценок ниже 4 (в трех из четырех групп минимальным баллом является 5). Также, важно отметить, что максимальное значение (10) присутствует только в двух группах (B и C), в остальных группах — это 8.

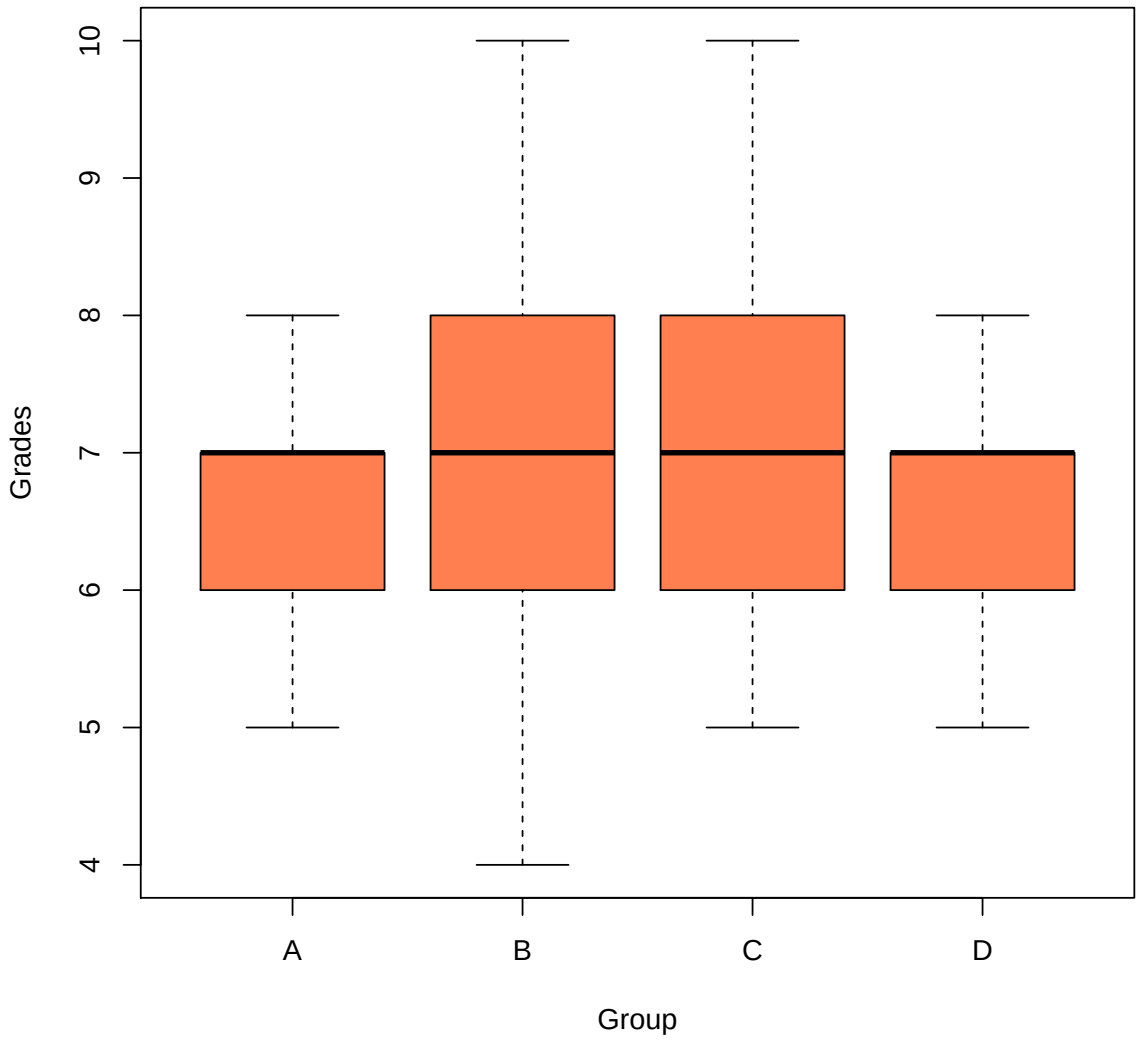
Как видно из описательных статистик по всем дисциплинам из массива, средний балл по микроэкономике — самый низкий среди всех дисциплин в приведенном датасете.

Межквартильный размах оценок по микроэкономике по всему массиву составляет 1.25 (Q3 - Q1). Таким образом, верхние и нижние границы характерных значений составляют [4.75; 8.5].

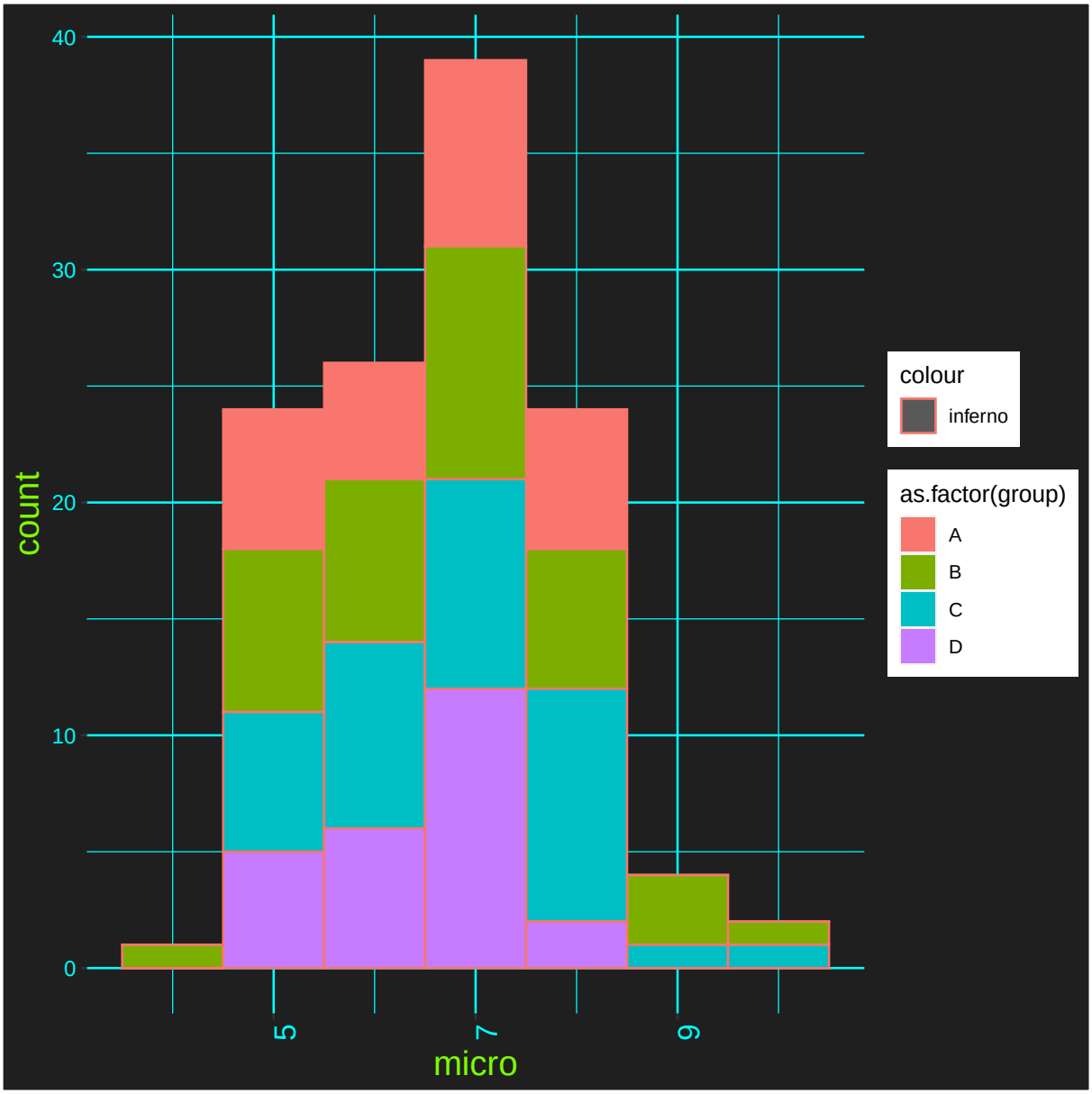
```
boxplot(data$micro, col = "aquamarine")
```




```
boxplot(data$micro ~data$group, xlab="Group", ylab="Grades", col = "coral")
```




```
data %>%
  ggplot(aes(micro, fill=as.factor(group),color="inferno"))+
  geom_histogram(binwidth = 1)+
  theme(axis.text.x = element_text(angle=90))+
  theme(panel.background = element_rect(fill="gray12",colour="gray12")) +
  theme(plot.background = element_rect(fill = "gray12"))+
  theme(panel.grid.minor = element_line(color = 'cyan1'))+
  theme(panel.grid.major = element_line(color = 'cyan1'))+
  theme(plot.title = element_text(colour = "chartreuse1",size=14))+
  theme(axis.title.x = element_text(colour = "chartreuse1",size=16, vjust=0.5),axis
  theme(axis.text.x = element_text( colour ="cyan1",size=14))+
  theme(axis.text.y = element_text( colour ="cyan1",size=10))
```



Из первого ящика с усами заметно, что, применительно ко всему массиву, оценка 10 является *нехарактерной*, т.к. верхняя граница графика лежит на значении 9. Во всем массиве находится всего две оценки 10 в группах B и C (по одной в каждой). Также видно, что на графике по всему массиву медиана (7) смещена ближе к Q3 (7.25), чем к значению Q1 (6).

Говоря о ящиках с усами по подвыборкам, можно отметить, что самое симметричное распределение оценок по микроэкономике наблюдается в группе B. Так, значения Q1 и Q3 (6 и 8 соответственно) равноотдалены от медианы (7). "Хвосты"графика также равноотдалены от Q1 и Q3 на 2 значения (4 и 10) вплоть до минимальных и максимальных значений в группе. Группа C близка к B по симметричности. Тем временем, в группах A и D медиана лежит ровно на значении Q3 (7), однако "хвосты"также одинаково отдалены от Q1 и Q3.

Как было сказано ранее, межквартильный размах по всему массиву составляет 1.25, что говорит о довольно большой скученности оценок в интервале от 6 до 7 — особенно это заметно на примере групп A и D.

Гистограмма по всему массиву с долей каждой из групп наглядно показывает пре-

валирование оценок 7 (всего 39 штук из 120) в массиве. На гистограмме особенно заметно малое количество оценок 9 и 10 (4 и 2 соответственно), а также 4 (1 раз) на курсе, особенно сравнительно с другими оценками. Также гистограмма показывает, что оценки 5, 6, 8 были выставлены практически равное число раз (24, 26 и 28 раз соответственно). Таким образом, правая часть графика относительно медианы (8-10) "падает"или "снижается"гораздо стремительнее, чем левая (4-6). Доли групп в полученных оценках визуально примерно равны для оценок 5, 6, 7. Для оценки 8 доля соответствующих оценок в группе C заметно увеличена, а D, наоборот, уменьшена относительно групп A и B. Как говорилось ранее, оценка 4 есть только в группе B, а оценки 9 и 10 — только в группах B и C. Оданко важно помнить, что в группах B и C обучается по 35 студентов, а в группах A и D — по 25 студентов.

В целом, график отдаленно напоминает нормальное распределение (особенно если бы оценок 5 было выставлено заметно меньше, чем 6).

3. Протестируем, различается ли средняя успеваемость по микроэкономике в учебных группах.

   $H_0 : a_1 = a_2 = a_3 = a_4 = a$

   $H_1 : a_j \neq a$

```
summary(aov(micro ~ as.factor(group), data))
```

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(group)    3   3.03   1.011   0.677  0.568
## Residuals         116 173.29   1.494
```

Логика, на которой базируется статистика критерия для однофакторного дисперсионного анализа, основана на идее соотношения межгрупповой дисперсии (Var_B) или информации, объясненной введенным фактором (в нашем случае — студенческой группой) к внутригрупповой дисперсии (Var_W) или необъясненной фактором информации. Эта статистика критерия имеет распределение Фишера с (k-1, N-k) степенями свободы, где k — число групп (подвыборок), а N — число наблюдений в общем массиве.

Так,

$$S = \frac{\frac{\sum n_j(\bar{x}_j - \bar{x})^2}{k-1}}{\frac{\sum\sum(x_{ij} - \bar{x}_j)^2}{N-k}} \sim F(k-1, N-k)$$

Для того, чтобы сделать вывод, необходимо либо найти квантиль, соответствующий некоторому уровню значимости и посмотреть, попадет ли значение S в зону отвержения, либо рассчитать p-value.

В нашем случае, p-value — это вероятность превысить найденное значение статистики критерия, то есть попасть правее него на графике (в силу того, что распределение Фишера не имеет отрицательных значений).

Таким образом, значение F value из выдачи R (0.677) = 1.011 / 1.494. Последние значения были получены путем нормирования сумм квадратов для каждой строки на соответствующее число степеней свободы (3.03 / 3) и (173.29 / 116).

R выдал следующее значение p-value = 0.568

Но p-value также можно было рассчитать вручную, исходя из значения статистики:

```r
pf(0.677, 3, 116, lower.tail = F)

## [1] 0.5678408
```

Как видно, результат схож.

Таким образом, значение p-value = 0.568 больше конвенционального уровня значимости = 0.05. Следовательно, у нас нет оснований отвергнуть H_0 в пользу альтернативы.

Значит, средние значения полученных оценок по микроэкономике равны по группам, а также равны среднему a по массиву. То есть средняя успеваемость по микроэкономике не отличается в учебных группах A, B, C, D. Или это означает, что наш фактор (учебные группы) плохо разделил ("объяснил") информацию — подвыборки оказались очень похожими.

```r
oneway.test(micro ~ as.factor(group), data)

##
##  One-way analysis of means (not assuming equal variances)
##
## data:  micro and as.factor(group)
## F = 0.82371, num df = 3.000, denom df = 63.048, p-value = 0.4857
```

Выше представлен еще один тест (с опущенным допущением о равенстве дисперсий), который также говорит о равенстве средних значений в группах (p-value = 0.4857). Однако ранее, благодаря тесту Левена, уже было показано, что мы можем говорить о равенстве дисперсий в группах.

Говоря об ***ограничениях*** *проведенного анализа*, на качественном уровне можно сказать о том, что, чаще всего, учебные группы довольно однородны относительно учебного курса. Из-за этого выделение учебной группы в качестве фактора создает практически однородные подвыборки, которые сложно поддаются анализу и в среднем не отличаются друг от друга. Это было видно заметно как из графиков, так и из описательных статистик. Например, медиана, среднее значение и дисперсия были равны по подвыборкам. При этом, для учебного курса сложно придумать иное разделение — только если не брать в качестве фактор бэкграунд студентов (оценки в школе, результаты ЕГЭ, уклон/профиль школы). Также, возможно, исследование было бы эффективнее, если бы сам курс был неоднородным (то есть можно было явно выделить некоторый *фактор*): например, на нем существовало расделение на разные профили подготовки или для некоторых групп требовался дополнительный отбор ("академическая группа"). В последнем случае фактор "учебная группа"можно было бы рассмотреть. Таким образом, в приведенной ситуации мы наблюдаем *искусственную кластеризацию*.

Задание 2.

Оценим в R линейную регрессию, в которой откликом является переменная «оценки по микроэкономике», предиктором — «оценки по математическому анализу» без разделения на учебные группы.

```
lab1 <- dplyr::select(data, math, micro)
describe(lab1)

##       vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
## math     1 120 7.39 1.20      7    7.43 1.48   4  10     6 -0.23    -0.01 0.11
## micro    2 120 6.67 1.22      7    6.65 1.48   4  10     6  0.19    -0.41 0.11

m1 <- lm(micro ~ math, data = lab1)
summary(m1)

##
## Call:
## lm(formula = micro ~ math, data = lab1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5813 -0.8440  0.1136  0.7662  2.7662
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.10664    0.65854   6.236 7.23e-09 ***
## math         0.34747    0.08796   3.950 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.149 on 118 degrees of freedom
## Multiple R-squared:  0.1168,Adjusted R-squared:  0.1093
## F-statistic: 15.61 on 1 and 118 DF,  p-value: 0.0001332
```

1. Итак, спецификация модели:

$$y_i = 4.10664 + 0.34747 x_i + \hat{\varepsilon}_i$$

2. Еще раз выведем оценки коэффициентов:

```
m1$coefficients

## (Intercept)        math
##   4.1066387   0.3474671
```

Интерпретация:

$\hat{\beta}_0 = 4.1066387$: среднее значение зависимой переменной (оценка по микроэкономике) $= 4.1066387$ при условии того, что все предикторы (оценка по математическому анализу) $= 0$.

$\hat{\beta}_1 = 0.3474671$: значение зависимой переменной (оценка по микроэкономике) в среднем увеличится на 0.3474671 при увеличении предиктора (оценка по математическому анализу) на единицу измерения при прочих равных условиях.

3. Значимость оценок регрессии рассчитывается исходя из t-статистики Стьюдента.

   Статистика критерия основывается на идее соотношения полученной оценки коэффициента и его стандартного отклонения: $\frac{\hat{\beta}_0}{se(\hat{\beta}_0)} \sim t(N-k)$ и $\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t(N-k)$. Из выдачи регрессии в R видно, что наблюдаемое значение t-статистики для коэф. следующее: $t_{observed}(\hat{\beta}_0) = 6.236$ и $t_{observed}(\hat{\beta}_1) = 3.950$.

   Далее, для рассчета значения p-value нам необходимо найти вероятность попадения правее, либо левее данной точки на теоретической функции t-распределения, т.к. наша альтернативная гипотеза — двусторонняя.

   Произведем расчет в R:

```
pt(6.236, df = 120 - 2, lower.tail = F)*2
```

```
## [1] 7.225868e-09
```

```
pt(3.950, df = 120 - 2, lower.tail = F)*2
```

```
## [1] 0.0001334492
```

   Как мы можем увидеть, рассчитанные вручную значения p-value совпали с соответствующей выдачей к регрессии.

   По умолчанию, проверяется гипотеза о равенстве оценки коэф-тов 0: $H_0 : \beta_0 = 0$ и $H_0 : \beta_1 = 0$ против двусторонней альтернативы: $H_1 : \beta_0 \neq 0$ $H_1 : \beta_1 \neq 0$.

   Таким образом, оба зачения p-value значительно меньше конвенционального значения 0.05. У нас есть основание отвергнуть обе нулевых гипотезы H_0 в пользу двусторонней альтернативы H_1.

   Так, оценки $\hat{\beta}_0 = 4.1066387$ и $\hat{\beta}_1 = 0.3474671$ статистически значимы.

4. Построим в R 95%-ые доверительные интервалы для коэффициентов (константы и коэффициента при предикторе).

   Их общий принцип построения ДИ для коэф-в регрессии: вычесть и прибавить к оценкам $\hat{\beta}_0$ и $\hat{\beta}_1$ соответствующий квантиль t-статистики Стьюдента, умноженный на стандартную оценку оценки.

```
m1$coefficients[1] - qt(0.975, dim(lab1)[1]-length(m1$model))*sqrt(diag(vcov(m1)))[
```

```
## (Intercept)
##    2.802542
```

```
m1$coefficients[1] + qt(0.975, dim(lab1)[1]-length(m1$model))*sqrt(diag(vcov(m1)))[
```

```
## (Intercept)
##    5.410735
```

```
m1$coefficients[2] - qt(0.975, dim(lab1)[1]-length(m1$model))*sqrt(diag(vcov(m1)))[
```

```
##      math
## 0.1732906
```

```
m1$coefficients[2] + qt(0.975, dim(lab1)[1]-length(m1$model))*sqrt(diag(vcov(m1)))[
```

```
##      math
## 0.5216437
```

```
confint(m1)
```

```
##                 2.5 %    97.5 %
## (Intercept) 2.8025423 5.4107350
## math        0.1732906 0.5216437
```

Итак, в начале мы рассчитали ДИ вручную, а затем — через автоматическую функцию.

Получили следующие значения:

95%-ный ДИ для $\hat{\beta}_0$: [2.8025423; 5.4107350]

95%-ный ДИ для $\hat{\beta}_1$: [0.1732906; 0.5216437]

Необходимо отметить, что оба доверительных интервала **не накрывают значение 0**.

Таким образом, оценки коэфициентов регрессии (константы и коэффициента при предикторе) можно считать статистически **значимыми** при уровне доверия 95%.

Интерпретация:

С 95%-ной уверенностью мы можем утверждать, что истинное значение β_0 и β_1 лежит в интервале [2.8025423; 5.4107350] и [0.1732906; 0.5216437], соответственно. Если мы будем проводить аналогичное исследование на выборках одного и того же размера много раз и независимо друг от друга, 95% доверительных интервалов будут включать истинное значение β_0 и β_1 (в предположении о том, что предельная ошибка выборки/стандартная ошибка не изменяется от выборки к выборке).

5. Из выдачи в R видно, что коэфициент детерминации (или R^2)нашей модели составил 0.1168.

   Таким образом, модель смогла объяснить 0.1168 от общей информации или $\approx 11{,}7\%$ через выделенные нами предикторы (оценка по математическому анализу).

6. ```
   anova(m1)
   ```

   ```
   ## Analysis of Variance Table
   ##
   ## Response: micro
   ##            Df  Sum Sq Mean Sq F value    Pr(>F)
   ## math        1  20.596 20.5961  15.606 0.0001332 ***
   ## Residuals 118 155.729  1.3197
   ## ---
   ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
   ```
   ```

Статистика критерия для определения значимости коэфициента детерминации базируется на идее соотношения ESS (explained sum of squares) и RSS (residual sum of squares), нормированных на соответствующее им число степеней свободы и имеет распределение Фишера с (k-1, N-k) степенями свободы: $\frac{\frac{ESS}{k-1}}{\frac{RSS}{N-k}} \sim F(k-1, N-k)$.

$H_0 : R^2 = 0$

$H_1 : R^2 > 0$

Далее для расчета p-value необходимо найти вероятность попадания правее наблюдаемого значения F-статистики Фишера (т.к. оно принимает только положительные значения, следовательно, наша альтернативная гипотеза односторонняя).

Из выдачи в R видно, что мы получили значение F-статистики = 15.606 при df = 1 (k-1) и 118 (N-k), а значение p-value = 0.0001332.

Его также можно было получить вручную:

В начале проверим наблюдаемое значение F-статистики: $\frac{\frac{20.5961}{2-1}}{\frac{155.729}{120-2}} \approx 15.606$

Далее рассчитаем p-value:

```
pf(15.606, 1, 118, lower.tail = F)

## [1] 0.0001332313
```

Как видно, значения получились идентичные.

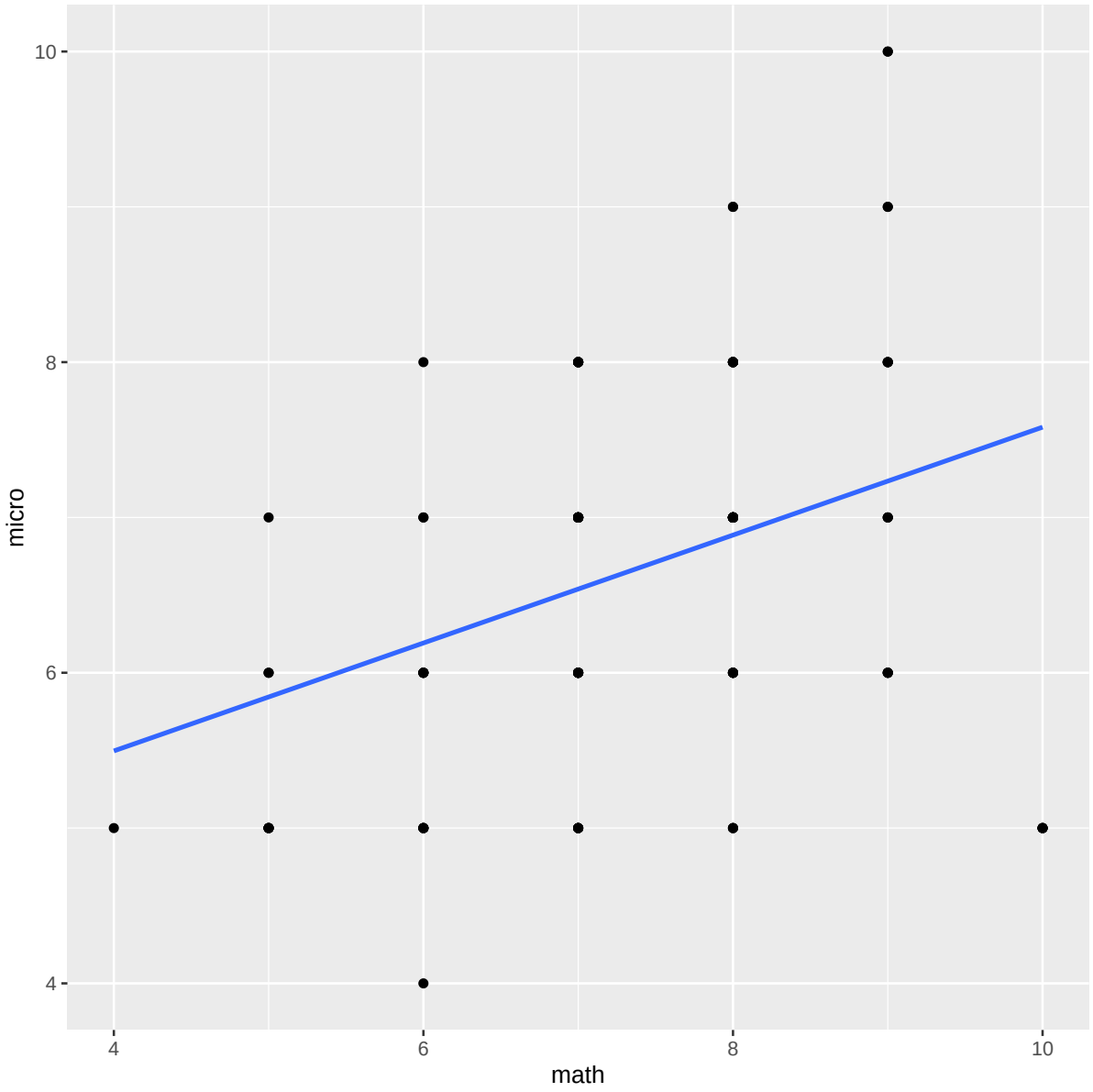
Значение p-value = 0.0001329828 меньше конвенционального уровня значимости 0.05.

Таким образом, мы имеем основания отвергнуть гипотезу H_0 в пользу альтернативы H_1. Так, коэфициент детерминации можно считать значимым.

7. Дополнительная визуализации модели для лучшего понимания.

```
ggplot(data = lab1, aes(x = math, y = micro)) +
  geom_smooth(method="lm",se=F) +
  geom_point()

## `geom_smooth()` using formula 'y ~ x'
```



Также проведем расчет корреляции:

```
cor.test(lab1$math, lab1$micro)


##
##  Pearson's product-moment correlation
##
## data:  lab1$math and lab1$micro
## t = 3.9505, df = 118, p-value = 0.0001332
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1731363 0.4909381
## sample estimates:
##       cor
## 0.3417714
```

Как можно понять из графика, относительно низкое значение корреляции (≈ 0.342), как и маленькое значение R^2, о котором говорилось ранее, связано с наличием 4

студентов, получивших 10 по мат. анализу и 5 по микроэкономике (крайняя точка справа снизу на графике).

Также примечательно, что значение корреляции оценок по мат. анализу и микроэкономике очень близко к полученному значению $\hat{\beta}_1$