Дисперсионный анализ, часть 2

Линейные модели…

Марина Варфоломеева, Вадим Хайтов

Кафедра Зоологии беспозвоночных, Биологический факультет, СПбГУ

Многофакторный дисперсионный анализ

- Модель многофакторного дисперсионного анализа
- Взаимодействие факторов
- Несбалансированные данные, типы сумм квадратов
- Многофакторный дисперсионный анализ в R
- Дисперсионный анализ в матричном виде

Вы сможете

- Проводить многофакторный дисперсионный анализ и интерпретировать его результаты с учетом взаимодействия факторов

Данные

Пример: Удобрение и беспозвоночные

Влияет ли добавление азотных и фосфорных удобрений на беспозвоночных?

Небольшие искуственные субстраты экспонировали в течение разного времени в верхней части сублиторали (Hall et al., 2000).

Зависимая переменная:

- `richness` — Число видов

Факторы:

- `time` — срок экспозиции (2, 4 и 6 месяцев)
- `treat` — удобрения (добавляли или нет)

Планировали сделать 5 повторностей для каждого сочетания факторов

Данные из Quinn, Keough, 2002

Знакомимся с данными

```
fert <- read.csv(file="data/hall.csv")
str(fert)
```

```
# 'data.frame': 29 obs. of  3 variables:
#  $ TREAT   : Factor w/ 2 levels "control","nutrient": 1 1 1 1 1 1 1 1 1 1 .
#  $ TIME    : int  2 2 2 2 2 4 4 4 4 4 ...
#  $ RICHNESS: int  5 7 5 7 5 20 18 20 18 17 ...
```

```
# Для удобства названия переменных маленькими буквами
colnames(fert) <- tolower(colnames(fert))
# Время делаем фактором
fert$time <- factor(fert$time)
levels(fert$time)
```

```
# [1] "2" "4" "6"
```

Пропущенные значения

```
sum(is.na(fert))
```

```
# [1] 0
```

```
sapply(fert, function(x)sum(is.na(x)))
```

```
#    treat     time richness
#        0        0        0
```

- Нет пропущенных значений

Объемы выборок в группах

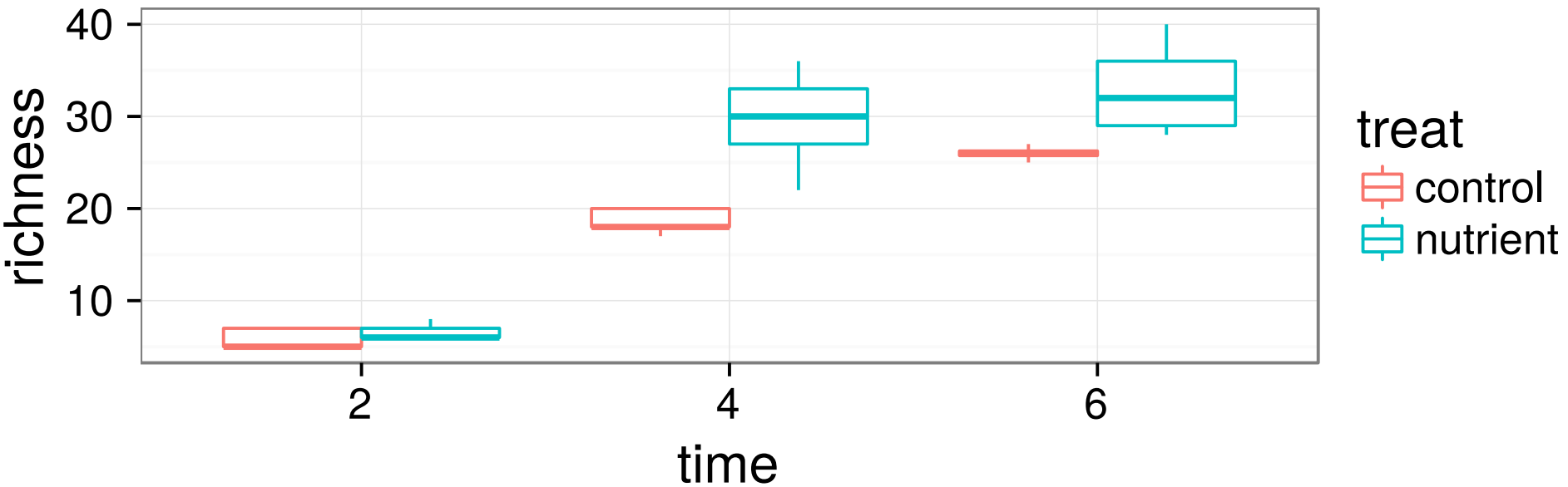
```
table(fert$time, fert$treat)
```

```
#
#     control nutrient
#   2       5        5
#   4       5        5
#   6       4        5
```

- Группы разного размера

Посмотрим на боксплот

```
library(ggplot2)
theme_set(theme_bw(base_size = 18) + theme(legend.key = element_blank()))
gg_rich <- ggplot(data = fert, aes(x = time, y = richness, colour = treat)) +
  geom_boxplot()
gg_rich
```



- Вполне возможно, здесь есть гетерогенность дисперсий.

Преобразовываем данные

На боксплоте видна гетерогенность дисперсий. И это неспроста!

Зависимая переменная `richness` – это счетная величина. Она подчиняется распределению Пуассона (и чем больше ее среднее значение, тем больше дисперсия).

Правильно было бы воспользоваться обобщенными линейными моделями с Пуассоновским распределением ошибок вместо нормального. Но пока что мы попробуем преобразовать зависимую переменную, чтобы ее распределение стало больше походить на нормальное. Это может помочь, а может и нет.

```
fert$log_rich <- log10(fert$richness + 1)
```

Модель многофакторного дисперсионного анализа

Линейные модели с разным числом дискретных предикторов

- Два фактора A и B, двухфакторное взаимодействие

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

- Три фактора A, B и C, двухфакторные взаимодействия, трехфакторное взаимодействия

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}$$

Линейные модели с разным числом дискретных предикторов

- Два фактора A и B, двухфакторное взаимодействие

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

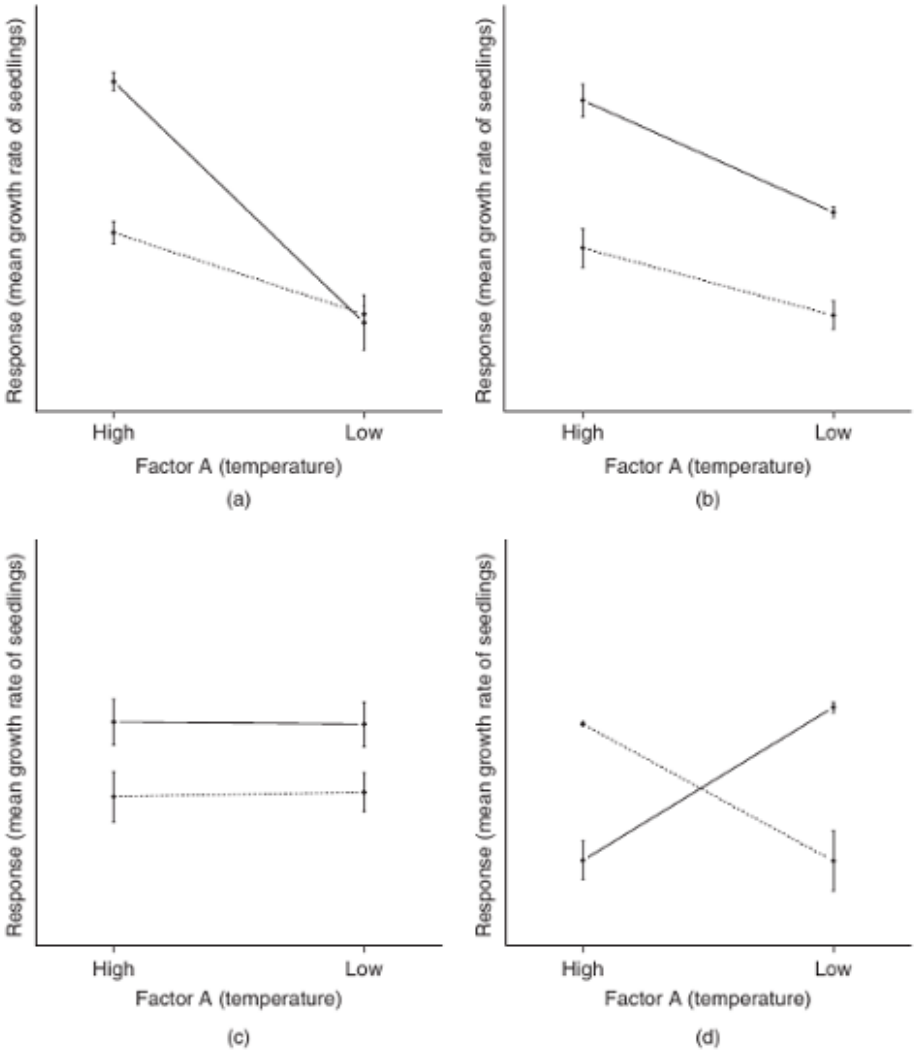
- Три фактора A, B и C, двухфакторные взаимодействия, трехфакторное взаимодействия

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}$$

α, β и γ — это группы коэффициентов, кодирующие соответствующие факторы.

Что такое взаимодействие дискретных предикторов

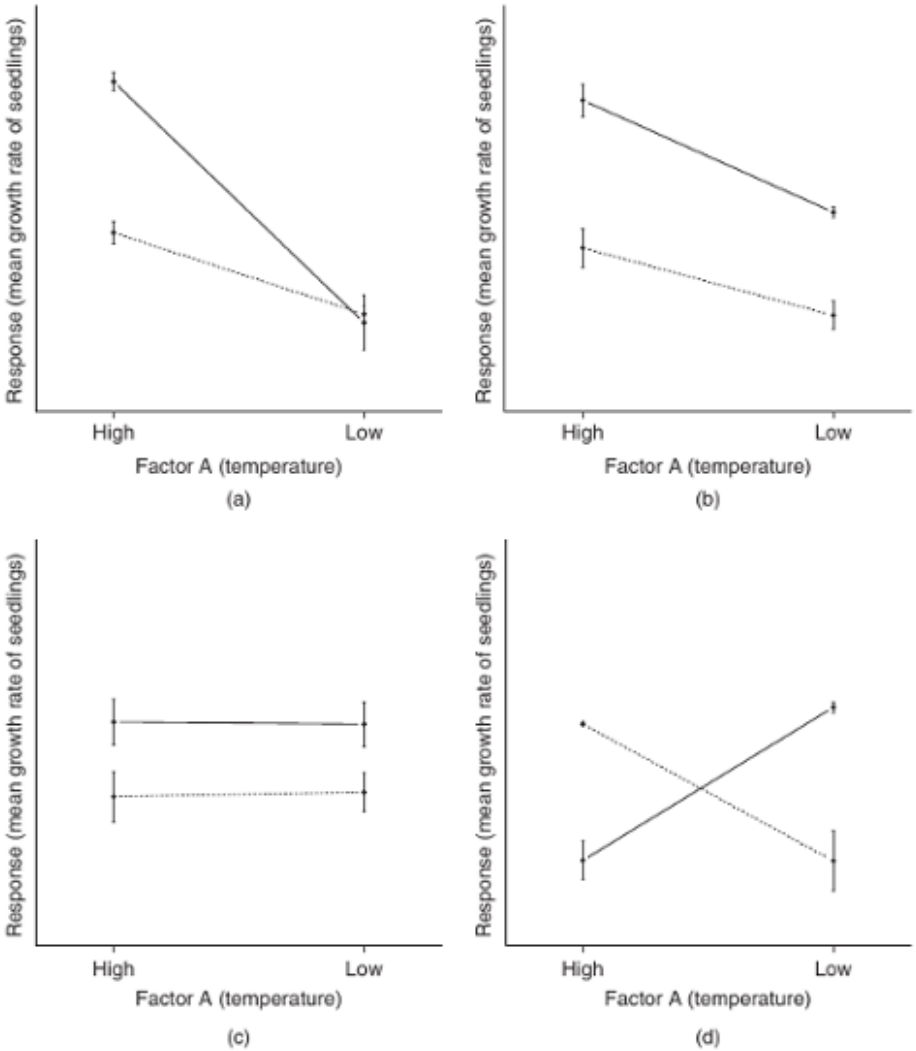
Взаимодействие факторов - когда эффект фактора B разный в зависимости от уровней фактора A и наоборот



На каких рисунках есть взаимодействие факторов?

Что такое взаимодействие дискретных предикторов

Взаимодействие факторов - когда эффект фактора B разный в зависимости от уровней фактора A и наоборот



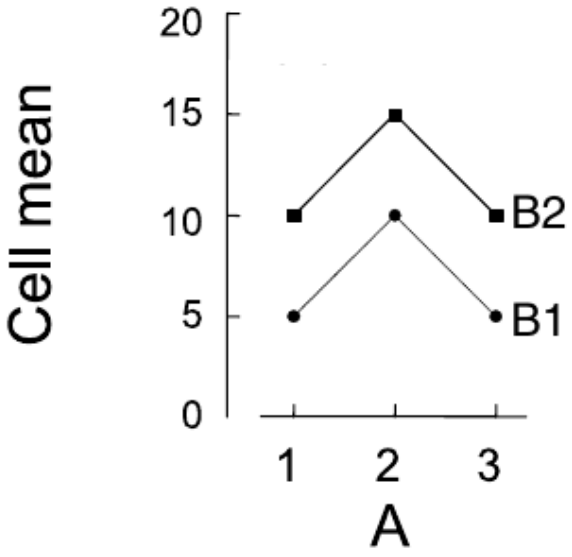
На каких рисунках есть взаимодействие факторов?

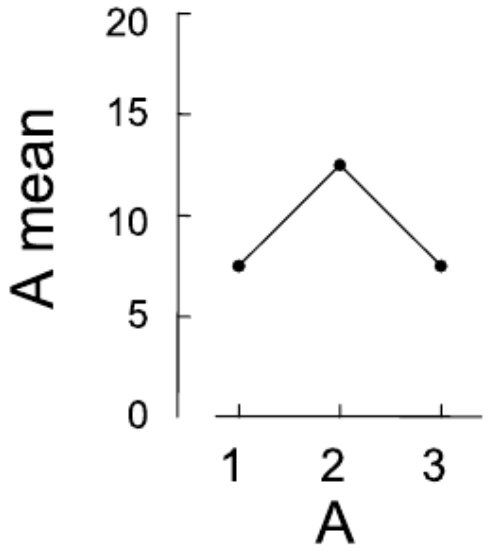
- b, c - нет взаимодействия (эффект фактора B одинаковый для групп по фактору A, линии для разных групп по фактору B на графиках расположены параллельно)
- a, d - есть взаимодействие (эффект фактора B разный для групп по фактору A, на графиках линии для разных групп по фактору B расположены под наклоном).

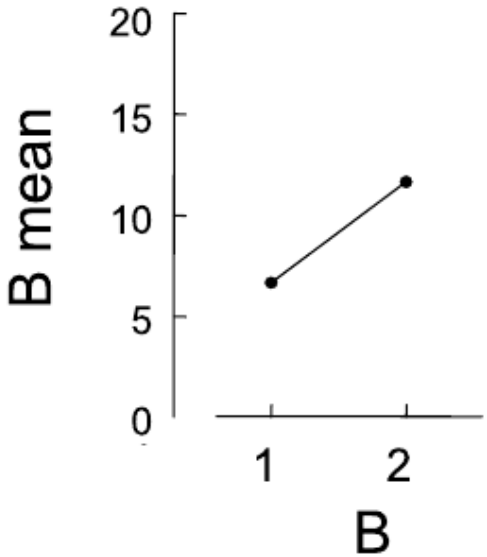
Рисунок из Logan, 2010, fig.12.2

Влияют ли главные эффекты и взаимодействие?

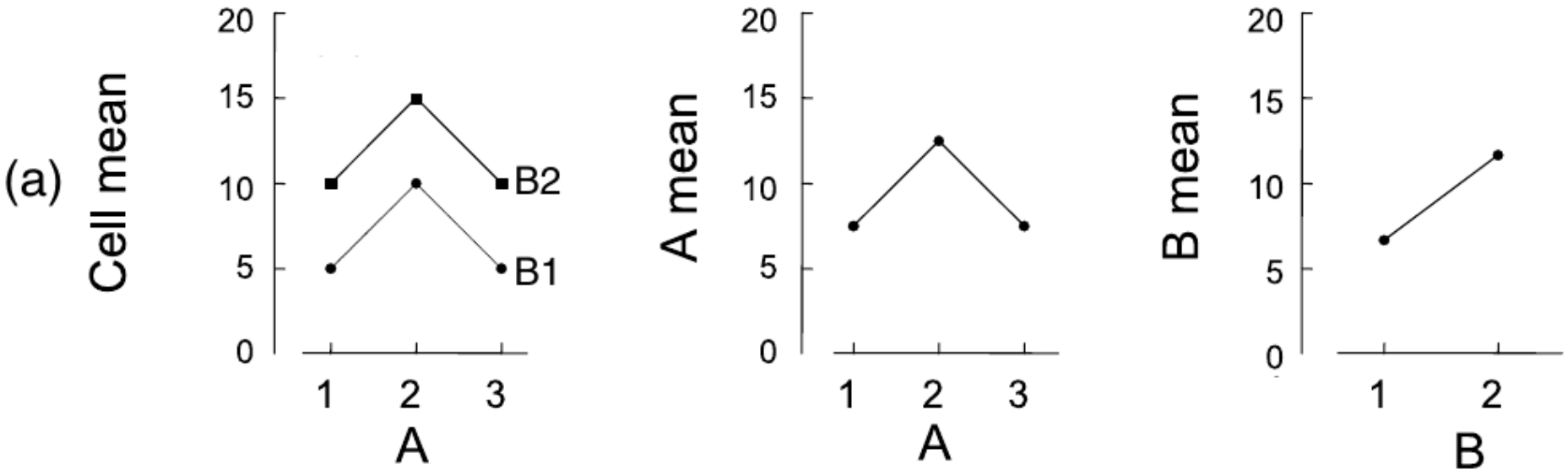
(a)







Влияют ли главные эффекты и взаимодействие?

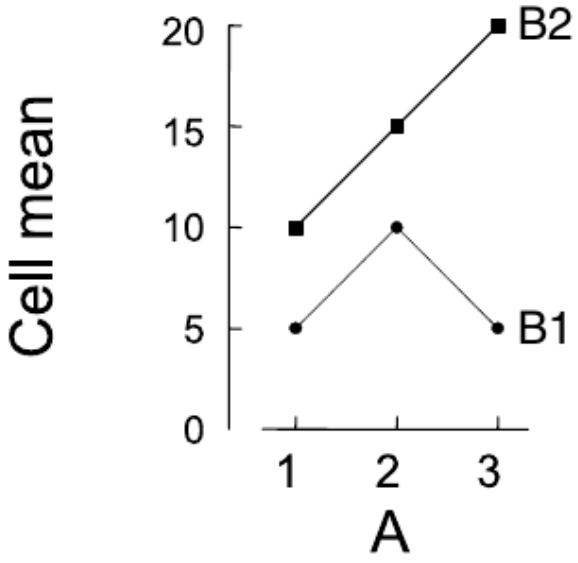


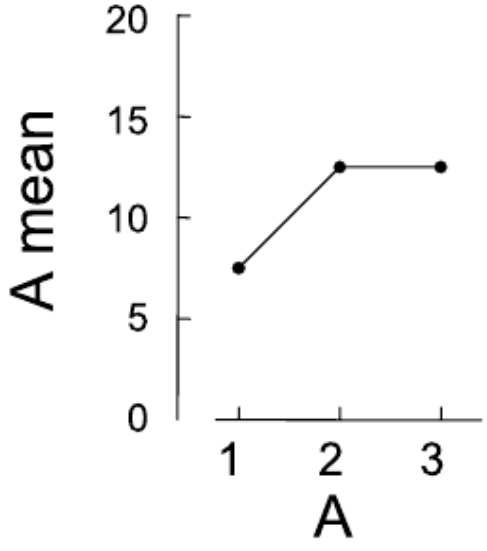
- взаимодействие недостоверно
- фактор А влияет
- фактор В влияет

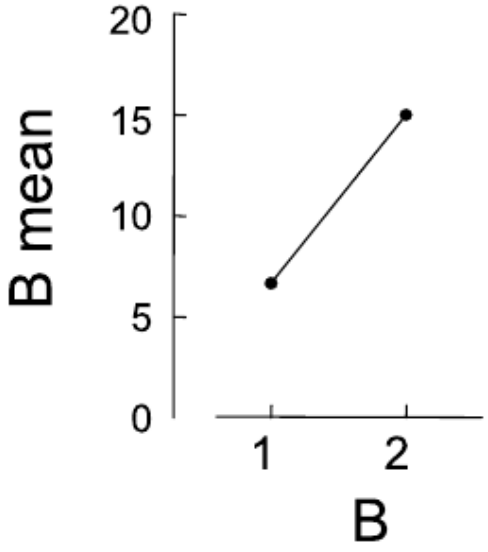
Рисунок из Quinn, Keough, 2002, fig.9.3

Влияют ли главные эффекты и взаимодействие?

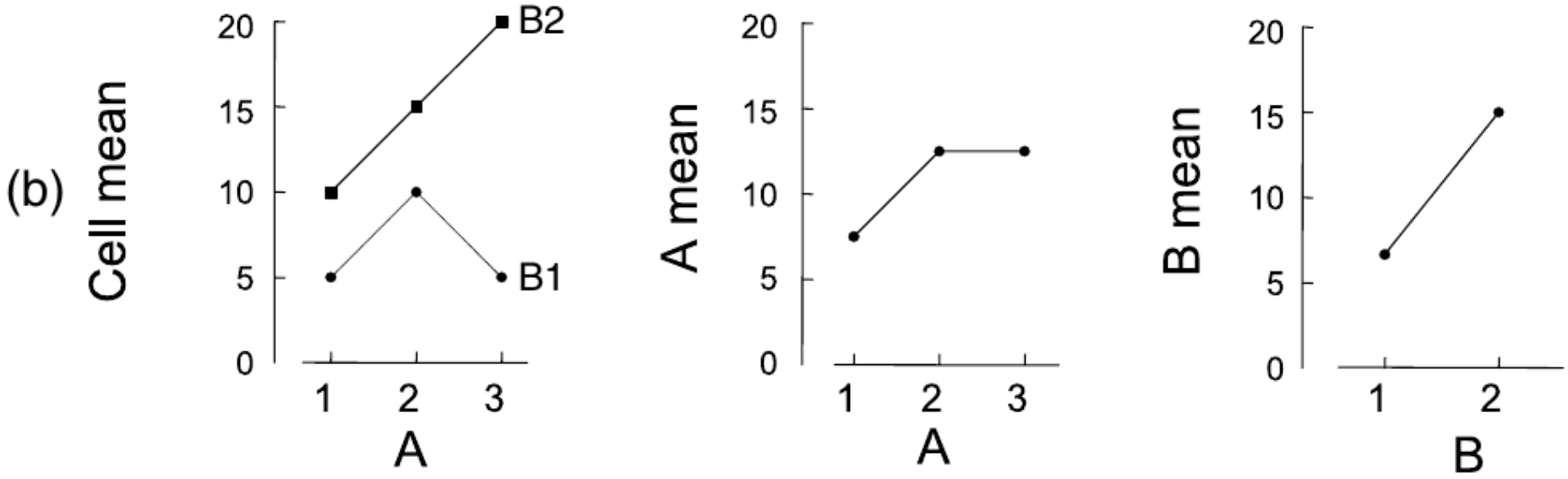
(b)







Влияют ли главные эффекты и взаимодействие?

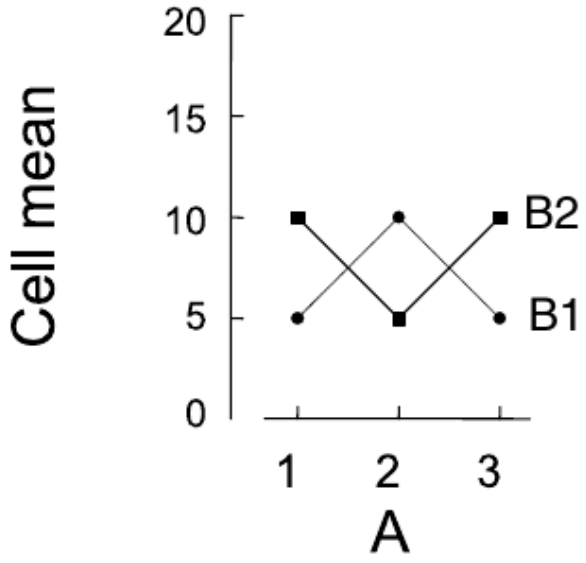


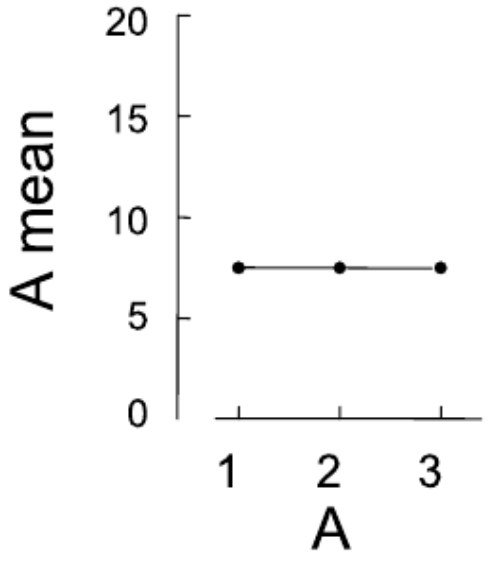
- взаимодействие достоверно и мешает интерпретировать влияние факторов отдельно:
  - для B2 зависимая переменная возрастает с изменением уровня A
  - для B1 зависимая переменная возрастает только на A2, но не различается на A1 и A3
- если смотреть на главные эффекты, можно сделать неправильные выводы:
  - фактор A влияет, группы A2 и A3 не отличаются
  - фактор B влияет, в группе B2 зависимая переменная больше, чем в B1

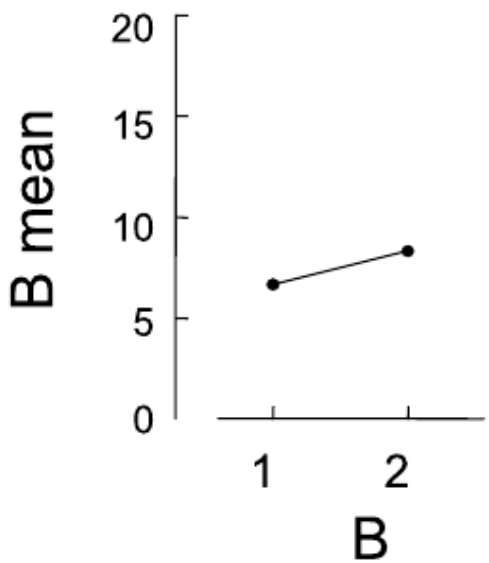
Рисунок из Quinn, Keough, 2002, fig.9.3

Влияют ли главные эффекты и взаимодействие?

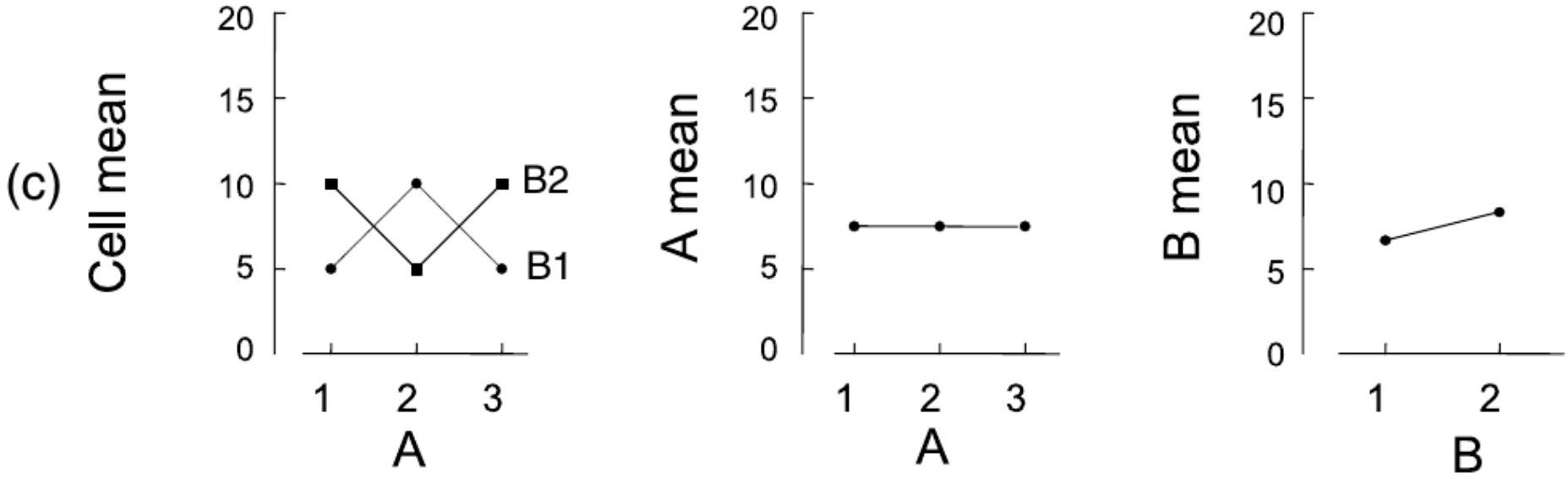
(c)







Влияют ли главные эффекты и взаимодействие?



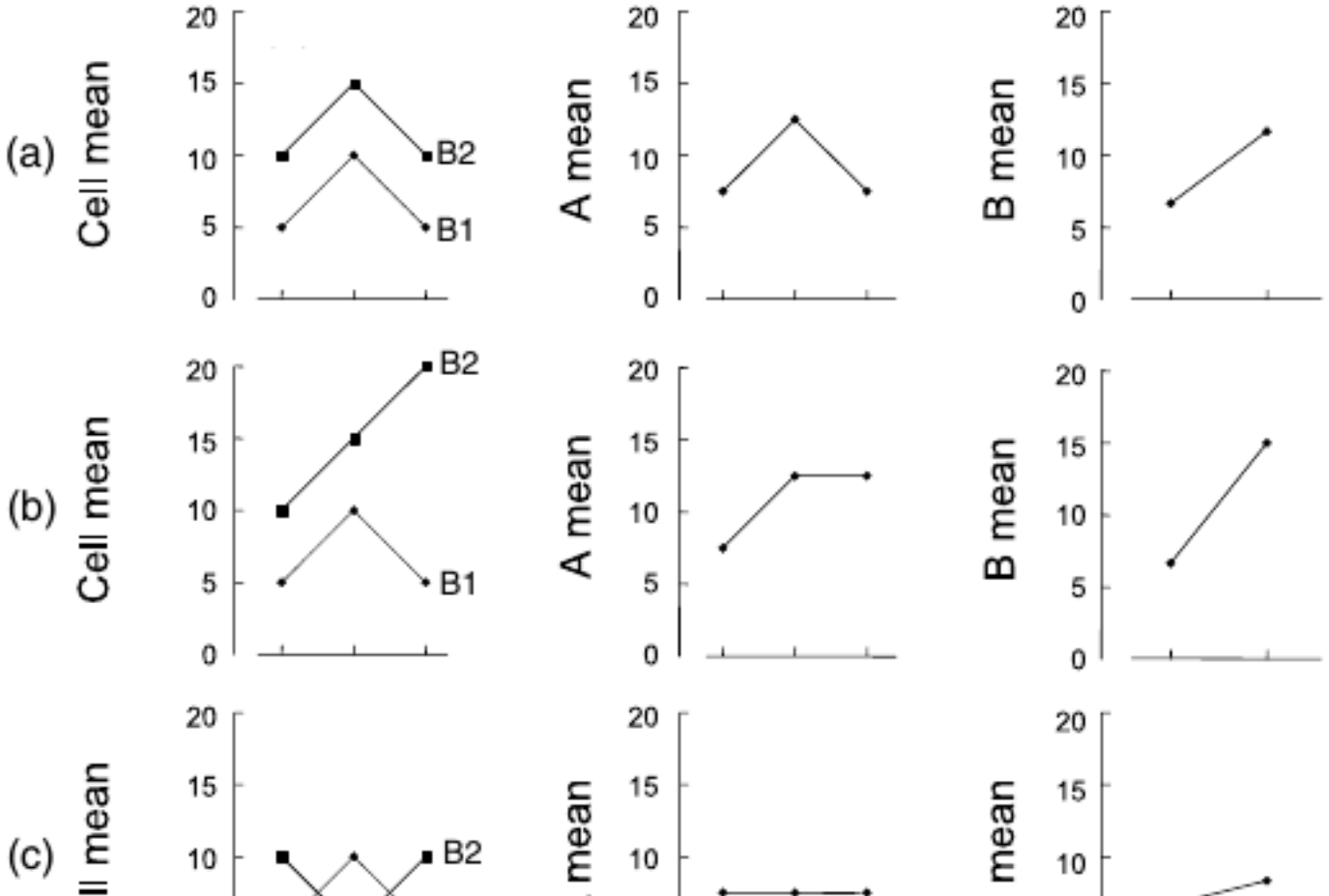
- взаимодействие достоверно и мешает интерпретировать влияние факторов отдельно:
  - A1B2, A3B2 и A2B1 не различаются, значение зависимой переменной в этих группах выше, чем в остальных
  - A1B1, A3B1 и A2B2 не различаются
- если смотреть на главные эффекты, можно сделать неправильные выводы:
  - факторы A и B не влияют

Рисунок из Quinn, Keough, 2002, fig.9.3

Взаимодействие факторов может маскировать главные эффекты

Если есть значимое взаимодействие

- главные эффекты обсуждать не имеет смысла
- пост хок тесты проводятся только для ваимодействия



Двухфакторный дисперсионный анализ в матричном виде

Двухфакторный дисперсионный анализ в матричном виде (contr.treatment)

Уравнение линейной модели для этого примера (в параметризации фиктивных переменных, contr.treatment):

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_5 x_{5i} + \varepsilon_i$$

- Здесь $i = 1, \ldots, n$, т.е. порядковый номер наблюдения,
- $x_{1i}, \ldots, x_{5i}$ — переменные-болванки
- β_0 - видовое богатство в контроле (при экспозиции 2)
- β_1 - изменение видового богатства при добавлении удобрений (при экспозиции 2)
- β_2 и β_3 - изменение видового богатства в контроле при экспозиции 4 и 6 соответственно
- β_4 и β_5 - изменение видового богатства при добавлении удобрений при экспозиции 4 и 6 соответственно

Остальное все так же как в предыдущем примере с однофакторным анализом.

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

Для параметризации эффектов (contr.sum) дискретные предикторы кодируются иначе

В модели эффектов (contr.sum) переменные-болванки будут закодированы при помощи -1, 0 и 1, так, чтобы сумма кодов для возможных состояний одной переменной была равна нулю.

Уровни фактора “удобрения” закодированы с помощью одной переменной:

```
contr.sum(levels(fert$treat))
```

```
#           [,1]
# control     1
# nutrient   -1
```

Уровни фактора “время экспозиции” — с помощью двух переменных:

```
contr.sum(levels(fert$time))
```

```
#   [,1] [,2]
# 2    1    0
# 4    0    1
# 6   -1   -1
```

Модельная матрица целиком

```
X_sum <- model.matrix(~ treat*time, fert,
  contrasts = list(treat = ”contr.sum”, time = ”contr.sum”))
```

Двухфакторный дисперсионный анализ в матричном виде (contr.sum)

Уравнение линейной модели для этого примера (в параметризации фиктивных переменных, contr.sum):

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_5 x_{5i} + \varepsilon_i$$

- Здесь $i = 1, \ldots, n$, т.е. порядковый номер наблюдения,
- $x_{1i}, \ldots, x_{5i}$ — переменные-болванки
- β_0 - средний уровень видового богатства во всех пробах
- β_1 - фактор “тритмент”, отклонение тритментов от общего среднего (в модельной матрице опыт закодирован как +1, удобрения -1, (сумма кодов 0))
- $\beta_2 + \beta_3$ - фактор “время”, отклонение видового богатства при разной экспозиции от общего среднего
- $\beta_4 + \beta_5$ - взаимодействие, отклонение видового богатства в тритментах при разных экспозициях от общего среднего

Несбалансированные данные, типы сумм квадратов

Несбалансированные данные - когда численности в группах по факторам различаются

Например так,

	A1	A2	A3
B1	5	5	5
B2	5	4	5

или так,

	A1	A2	A3
B1	3	8	4
B2	4	7	4

Проблемы несбалансированных дизайнов

- Оценки средних в разных группах с разным уровнем точности (Underwood 1997)
- ANOVA менее устойчив к отклонениям от условий применимости (особенно от гомогенности дисперсий) при разных размерах групп (Quinn Keough 2002, section 8.3)
- Проблемы с рассчетом мощности. Если $\sigma^2_\epsilon > 0$ и размеры выборок разные, то $\frac{MS_{factor}}{MS_{residuals}}$ не следует F-распределению (Searle et al. 1992).

Проблемы несбалансированных дизайнов

- Оценки средних в разных группах с разным уровнем точности (Underwood 1997)
- ANOVA менее устойчив к отклонениям от условий применимости (особенно от гомогенности дисперсий) при разных размерах групп (Quinn Keough 2002, section 8.3)
- Проблемы с рассчетом мощности. Если $\sigma^2_\epsilon > 0$ и размеры выборок разные, то $\frac{MS_{factor}}{MS_{residuals}}$ не следует F-распределению (Searle et al. 1992).

- Старайтесь *планировать* группы равной численности!
- Но если не получилось - не страшно:
  - Для фикс. эффектов неравные размеры - проблема только если значения доверительной вероятности *p* близки к выбранному критическому уровню значимости α

Если несбалансированные данные, выберите правильный тип сумм квадратов

- SSe и SSab также как в сбалансированных
- SSa, SSb - три способа расчета
- Для сбалансированных дизайнов - результаты одинаковы
- Для несбалансированных дизайнов рекомендуют **суммы квадратов III типа** если есть взаимодействие факторов (Maxwell & Delaney 1990, Milliken, Johnson 1984, Searle 1993, Yandell 1997)

Порядок тестирования гипотез в дисперсионном анализе

”Типы сумм квадратов”	I тип	II тип	III тип
Название	Последовательная	Без учета взаимодействий высоких порядков	Иерархическая
SS	SS(A) SS(B | A) SS(AB | B, A)	SS(A | B) SS(B | A) SS(AB | B, A)	SS(A | B, AB) SS(B | A, AB) SS(AB | B, A)
Величина эффекта зависит от выборки в группе	Да	Да	Нет
Результат зависит от порядка включения факторов в модель	Да	Нет	Нет
Команда R	aov()	Anova() (пакет car)	Anova() (пакет car)

Многофакторный дисперсионный анализ в R

Дисперсионный анализ со II типом сумм квадратов

Сначала тестируем взаимодействие, затем тестируем факторы в модели без взаимодействия

```
fmod2 <- lm(log_rich ~ treat * time, data = fert)
library(car)
Anova(fmod2, type = "II")
```

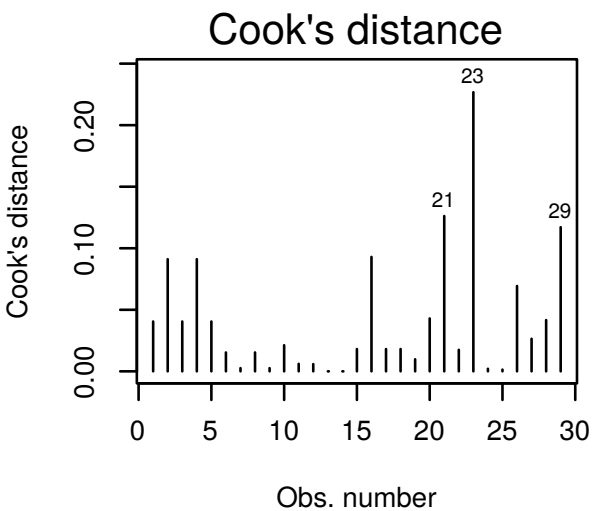
```
# Anova Table (Type II tests)
#
# Response: log_rich
#            Sum Sq Df F value   Pr(>F)
# treat       0.091  1   28.42 0.000021 ***
# time        2.216  2  344.88  < 2e-16 ***
# treat:time  0.025  2    3.84    0.036 *
# Residuals   0.074 23
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

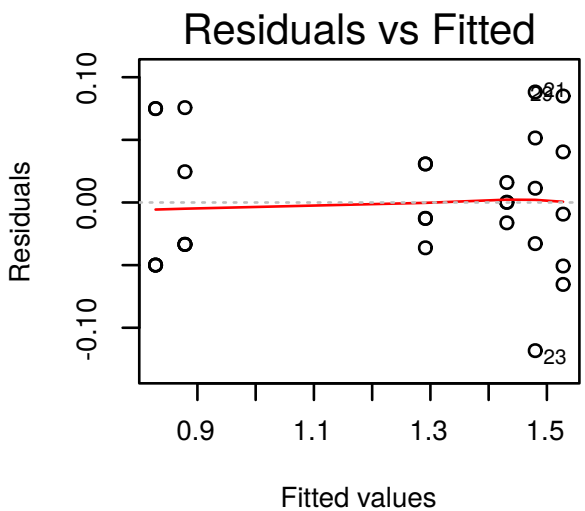
Задание

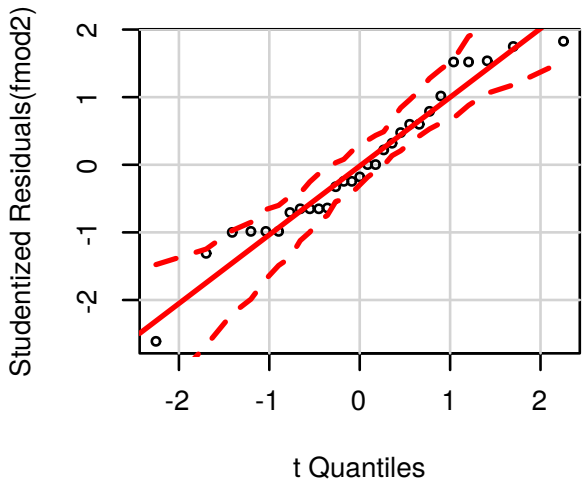
Проверьте условия применимости дисперсионного анализа

Решение

```
library(car)
op <- par(mfrow = c(1, 3))
plot(fmod2, which = 4)
plot(fmod2, which = 1)
qqPlot(fmod2)
par(op)
```



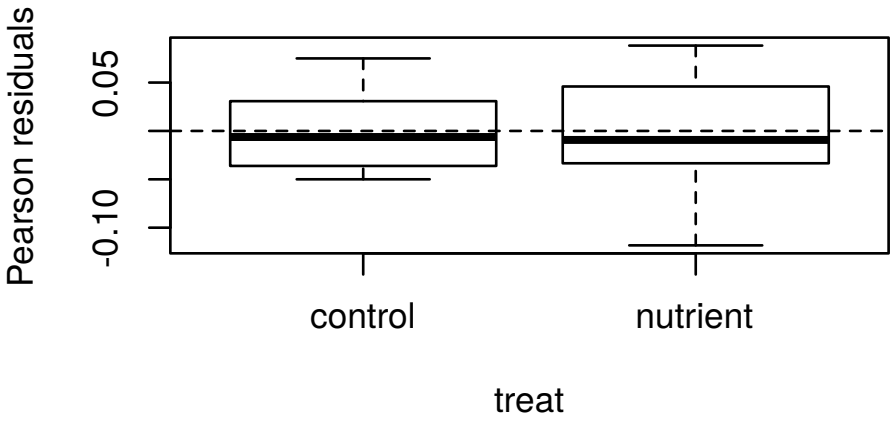


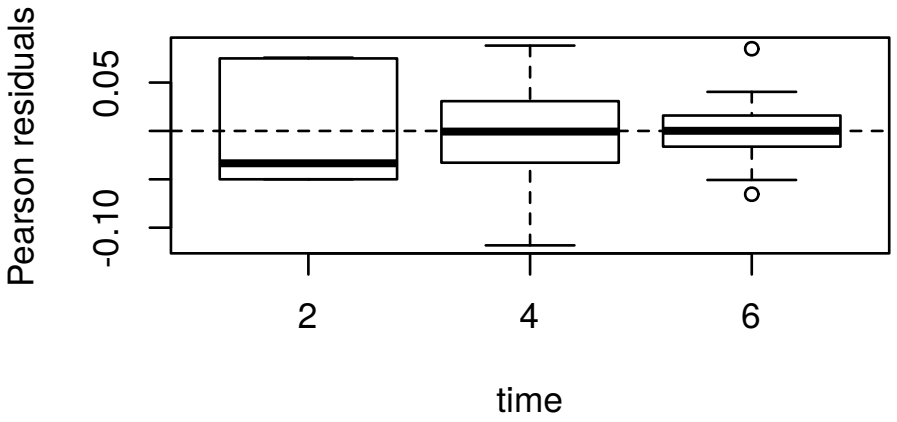


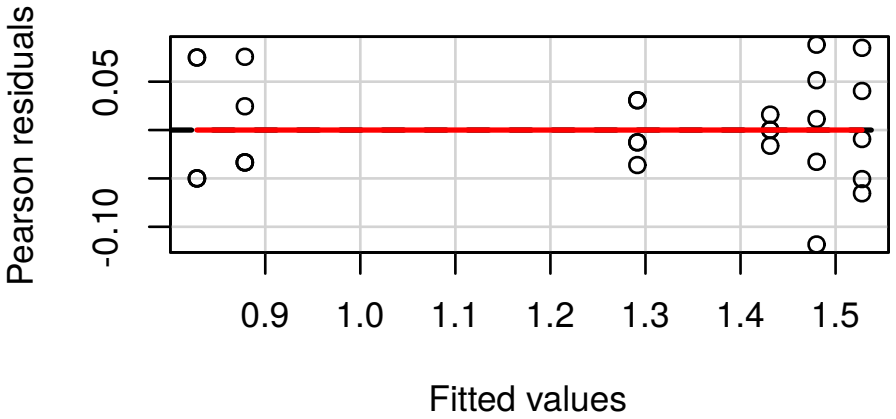
- Выбросов нет
- Дисперсии почти одинаковые. Может быть, в одной из групп больше
- Остатки нормально распределены

Графики остатков от переменных в модели

```
residualPlots(fmod2)
```







- По-видимому, с увеличением продолжительности экспозиции дисперсия остатков уменьшается.

Результаты дисперсионного анализа

```
Anova(fmod2, type = 2)
```

```
# Anova Table (Type II tests)
#
# Response: log_rich
#            Sum Sq Df F value   Pr(>F)
# treat       0.091  1   28.42 0.000021 ***
# time        2.216  2  344.88  < 2e-16 ***
# treat:time  0.025  2    3.84    0.036 *
# Residuals   0.074 23
# ---
# Signif. codes:
#   0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1  ’ ’ 1
```

Дисперсионный анализ с III типом сумм квадратов

На случай, если вдруг вам понадобится воспроизвести в R дисперсионный анализ с III типом сумм квадратов.

Тестируем взаимодействие, когда все другие факторы есть в модели. Затем тестируем факторы, когда все другие факторы и взаимодействие есть в модели.

Внимание: при использовании III типа сумм квадратов, нужно **обязательно указывать тип контрастов для факторов** (`contrasts=list(фактор_1 = contr.sum, фактор_2=contr.sum)`).

```
fmod3 <- lm(log_rich ~ treat * time, data = fert, contrasts = list(treat = co
Anova(fmod3, type = 3)
```

```
# Anova Table (Type III tests)
#
# Response: log_rich
#             Sum Sq Df  F value   Pr(>F)
# (Intercept)   44.2  1 13776.11  < 2e-16 ***
# treat          0.1  1    28.04 0.000022 ***
# time           2.2  2   344.75  < 2e-16 ***
# treat:time     0.0  2     3.84    0.036 *
# Residuals      0.1 23
# ---
# Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Пост хок тест для взаимодействия факторов

Пост хок тесты в многофакторном дисперсионном анализе

- Поскольку взаимодействие достоверно, факторы отдельно можно не тестировать. Проведем пост хок тест по взаимодействию, чтобы выяснить, какие именно группы различаются
- Если бы взаимодействие было недостоверно, мы бы провели пост хок тест по тем факторам, влияние которых было бы достоверно. Как? См. предыдущую презентацию.

Пост хок тест для взаимодействия факторов

Пост хок тест для взаимодействия факторов делается легче всего “обходным путем”

1. Создаем переменную-взаимодействие
2. Подбираем модель без свободного члена
3. Делаем пост хок тест для этой модели

```
fert$treat_time <- interaction(fert$treat, fert$time)
fit_inter <- lm(log_rich ~ treat_time - 1, data = fert)
library(multcomp)
dat_tukey <- glht(fit_inter, linfct = mcp(treat_time = ”Tukey”))
summary(dat_tukey)
```

Результаты пост хок теста

```
#
#    Simultaneous Tests for General Linear Hypotheses
#
# Multiple Comparisons of Means: Tukey Contrasts
#
#
# Fit: lm(formula = log_rich ~ treat_time - 1, data = fert)
#
# Linear Hypotheses:
#                              Estimate Std. Error t value Pr(>|t|)
# nutrient.2 - control.2 == 0    0.0504     0.0358    1.41    0.723
# control.4 - control.2 == 0     0.4633     0.0358   12.93   <0.001 ***
# nutrient.4 - control.2 == 0    0.6519     0.0358   18.19   <0.001 ***
# control.6 - control.2 == 0     0.6031     0.0380   15.86   <0.001 ***
# nutrient.6 - control.2 == 0    0.6997     0.0358   19.52   <0.001 ***
# control.4 - nutrient.2 == 0    0.4129     0.0358   11.52   <0.001 ***
# nutrient.4 - nutrient.2 == 0   0.6015     0.0358   16.78   <0.001 ***
# control.6 - nutrient.2 == 0    0.5527     0.0380   14.54   <0.001 ***
# nutrient.6 - nutrient.2 == 0   0.6493     0.0358   18.11   <0.001 ***
# nutrient.4 - control.4 == 0    0.1885     0.0358    5.26   <0.001 ***
# control.6 - control.4 == 0     0.1398     0.0380    3.68    0.014 *
# nutrient.6 - control.4 == 0    0.2364     0.0358    6.59   <0.001 ***
# control.6 - nutrient.4 == 0   -0.0488     0.0380   -1.28    0.791
# nutrient.6 - nutrient.4 == 0   0.0478     0.0358    1.33    0.763
# nutrient.6 - control.6 == 0    0.0966     0.0380    2.54    0.153
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Данные для графика при помощи predict()

```
MyData <- expand.grid(treat = levels(fert$treat),
                      time = levels(fert$time))
MyData <- data.frame(MyData,
  predict(fmod2, newdata = MyData,
          interval = "confidence"))
# Обратная трансформация
MyData$richness <- 10^MyData$fit
MyData$LWR <- 10^MyData$lwr
MyData$UPR <- 10^MyData$upr
MyData
```

```
#       treat time   fit   lwr   upr richness   LWR   UPR
# 1   control    2 0.828 0.776 0.881     6.73  5.97  7.60
# 2 nutrient    2 0.879 0.826 0.931     7.56  6.70  8.53
# 3   control    4 1.291 1.239 1.344    19.56 17.34 22.07
# 4 nutrient    4 1.480 1.428 1.532    30.20 26.76 34.07
# 5   control    6 1.431 1.373 1.490    26.99 23.58 30.89
# 6 nutrient    6 1.528 1.475 1.580    33.71 29.88 38.04
```

Задание:

Создайте MyData вручную:

- предсказанные значения
- стандартные ошибки
- верхнюю и нижнюю границы доверительных интервалов

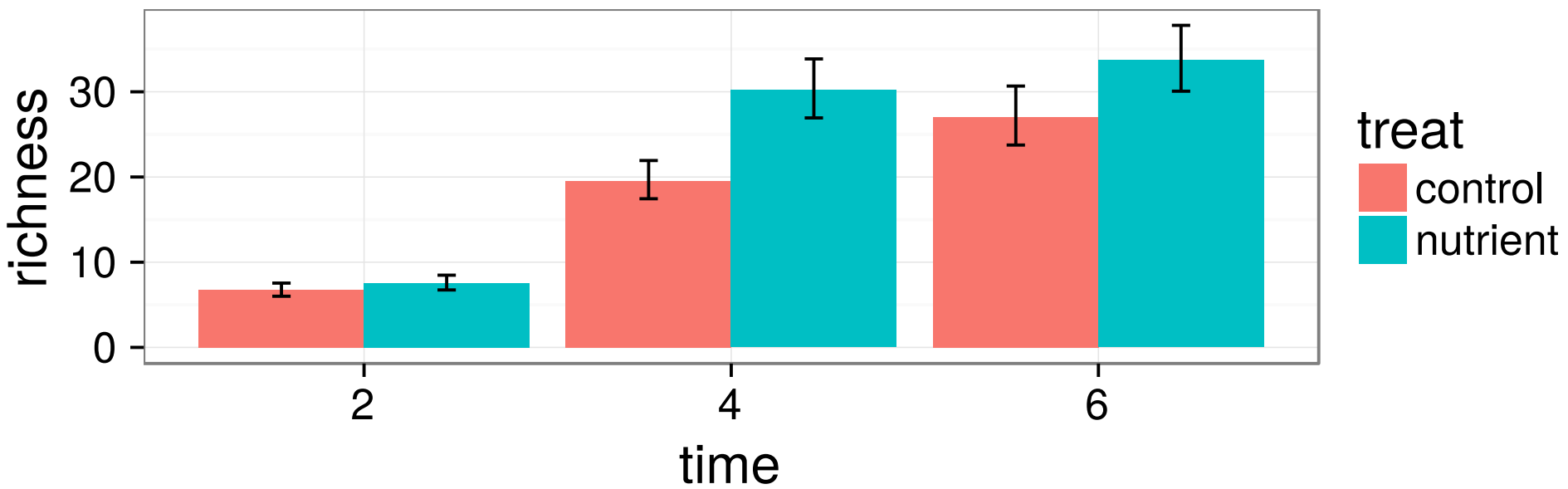
Решение:

```
MyData <- expand.grid(treat = levels(fert$treat),
                      time = levels(fert$time))
X <- model.matrix(~ treat * time, data = MyData)
betas <- coef(fmod2)
MyData$fit <- X %*% betas
MyData$se <- sqrt(diag(X %*% vcov(fmod2) %*% t(X)))
MyData$lwr <- MyData$fit - qnorm(0.975) * MyData$se
MyData$upr <- MyData$fit + qnorm(0.975) * MyData$se
# Обратная трансформация
MyData$richness <- 10^MyData$fit
MyData$LWR <- 10^MyData$lwr
MyData$UPR <- 10^MyData$upr
MyData
```

```
#       treat time   fit     se   lwr   upr richness   LWR   UPR
# 1   control    2 0.828 0.0253 0.778 0.878     6.73  6.00  7.55
# 2 nutrient    2 0.879 0.0253 0.829 0.928     7.56  6.74  8.48
# 3   control    4 1.291 0.0253 1.242 1.341    19.56 17.45 21.93
# 4 nutrient    4 1.480 0.0253 1.430 1.530    30.20 26.93 33.86
# 5   control    6 1.431 0.0283 1.376 1.487    26.99 23.75 30.67
# 6 nutrient    6 1.528 0.0253 1.478 1.577    33.71 30.07 37.80
```

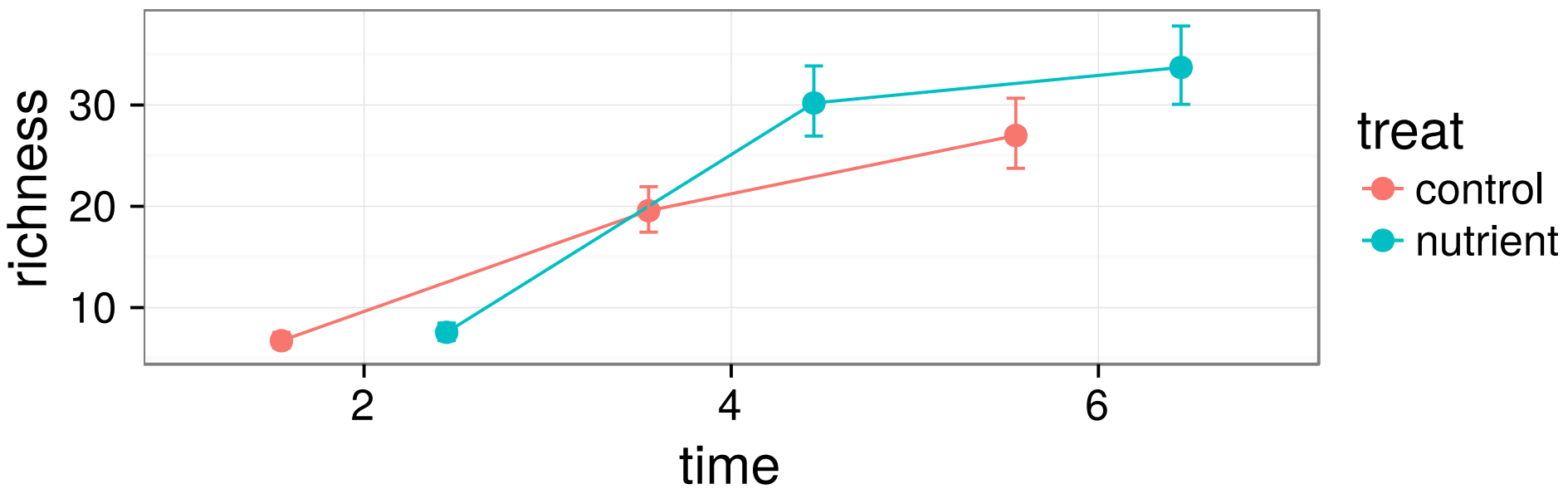
Графики для результатов: Столбчатый график

```
pos <- position_dodge(width = 0.9)
gg_barp <- ggplot(data = MyData, aes(x = time, y = richness,
             ymin = LWR,  ymax = UPR, fill = treat)) +
  geom_bar(stat = "identity", position = pos) +
  geom_errorbar(width = 0.1, position = pos)
gg_barp
```



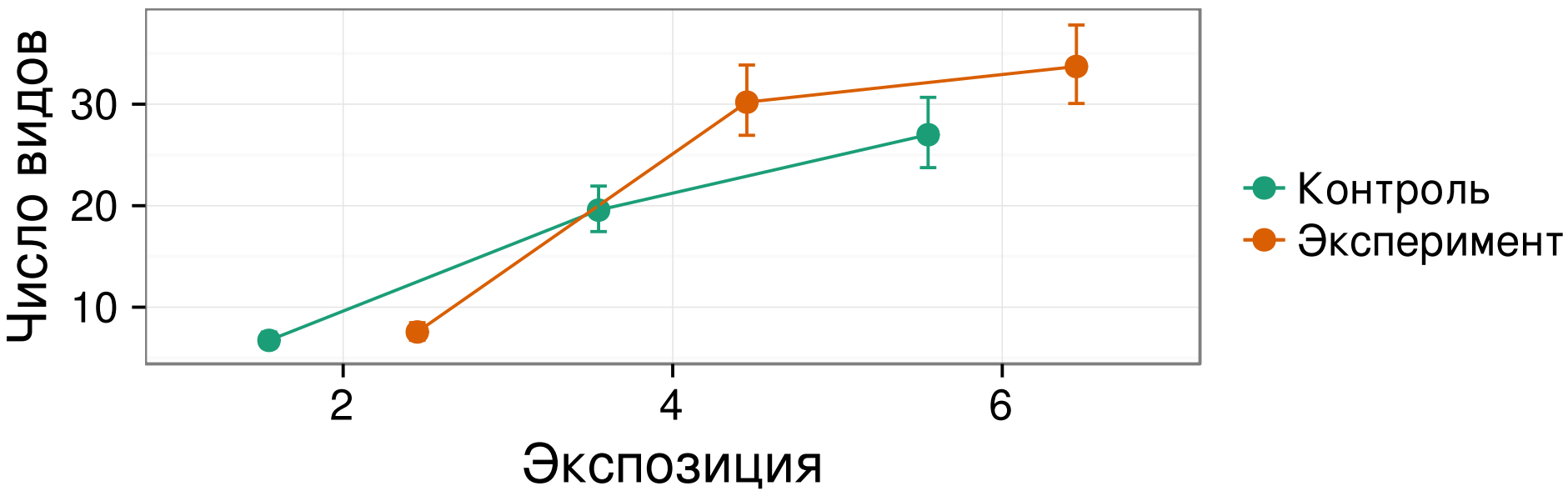
Графики для результатов: Линии с точками

```
gg_linep <- ggplot(data = MyData, aes(x = time, y = richness,
                ymin = LWR,  ymax = UPR, colour = treat)) +
  geom_point(size = 3, position = pos) +
  geom_line(aes(group = treat), position = pos) +
  geom_errorbar(width = 0.1, position = pos)
gg_linep
```



Приводим понравившийся график в приличный вид

```
gg_final <- gg_linep + labs(x = "Экспозиция",  y = "Число видов") +
  scale_colour_brewer(name = "", palette = "Dark2",
    labels = c("Контроль", "Эксперимент"))
gg_final
```



Take home messages

- Многофакторный дисперсионный анализ позволяет оценить взаимодействие факторов. Если оно значимо, то лучше воздержаться от интерпретации их индивидуальных эффектов
- Если численности групп равны - получаются одинаковые результаты с использованием I, II, III типы сумм квадратов
- В случае, если численности групп неравны (несбалансированные данные) по разному тестируется значимость факторов (I, II, III типы сумм квадратов)

Дополнительные ресурсы

- Quinn, Keough, 2002, pp. 221-250
- Logan, 2010, pp. 313-359
- Sokal, Rohlf, 1995, pp. 321-362
- Zar, 2010, pp. 246-266