

# Линейные модели для счетных данных

Линейные модели...

Вадим Хайтов, Марина Варфоломеева



## Мы рассмотрим

- ▶ Различные варианты анализа, применяющегося для тех случаев, когда зависимая переменная - счетная величина (целые неотрицательные числа)

## Вы сможете

- ▶ Объяснить особенности разных типов распределений, принадлежащих экспоненциальному семейству.
- ▶ Построить пуасоновскую и квази-пуассоновскую линейную модель
- ▶ Объяснить проблемы, связанные с избыточностью дисперсии в модели
- ▶ Построить модель, основанную на отрицательном биномиальном распределении



## Различные типы распределений



## Распределение

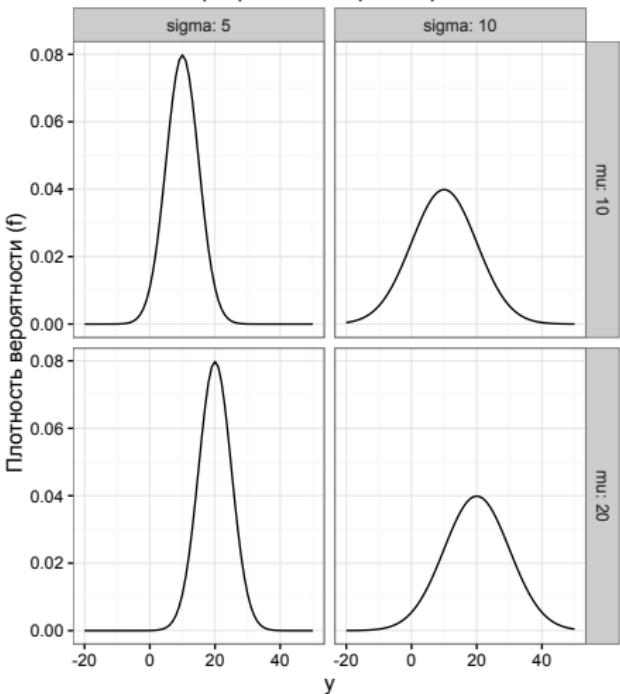
То, что мы в быту привыкли называть **распределением** - это функция плотности вероятности.

**Плотность вероятности** - это функция, описывающая вероятность получения разных значений случайной величины

# Нормальное распределение

$$f(y; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Нормальное распределение  
при разных параметрах



**Два параметра** ( $\mu, \sigma$ )

Среднее:  $E(Y) = \mu$

Дисперсия:  $\text{var}(Y) = \sigma^2$

**Пределы варьирования**

$-\infty \leq Y \leq +\infty$



# Распределение Пуассона

$$f(y; \mu) = \frac{\mu^y \times e^{-\mu}}{y!}$$

**Один параметр ( $\mu$ )**

Среднее:  $E(Y) = \mu$

Дисперсия:  $\text{var}(Y) = \mu$

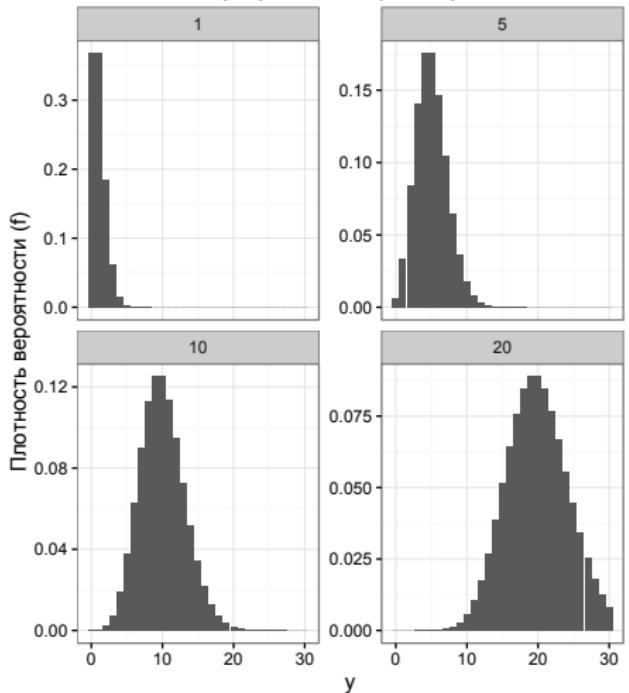
**Важное свойство:** При увеличении значения  $\mu$  увеличивается размах варьирования

**Пределы варьирования**

$0 \leq Y \leq +\infty$ ,

$Y$  целочисленные!

Распределение Пуассона  
при разных параметрах



# Гамма-распределение

$$f(y; \mu, \nu) = \frac{1}{\Gamma(\nu)} \times \left(\frac{\nu}{\mu}\right)^\nu \times y^{\nu-1} \times e^{\frac{-y \times \nu}{\mu}}$$

**Два параметра ( $\mu, \nu$ )**

Среднее:  $E(Y) = \mu$

Дисперсия:  $var(Y) = \frac{\mu^2}{\nu}$

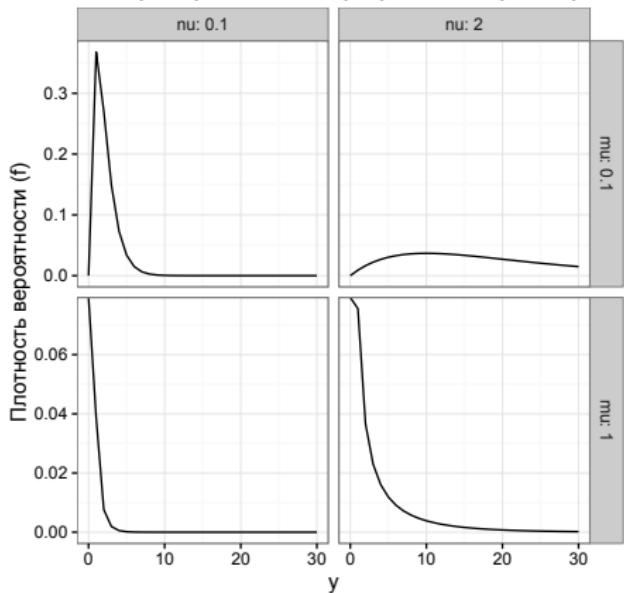
Параметр  $\nu$  определяет степень избыточности дисперсии

**Пределы варьирования**

$0 < Y \leq +\infty$

Внимание!  $Y$  строго больше 0

Гамма распределение при разных параметрах



# Отрицательное биномиальное распределение

$$f(y; k, \mu) = \frac{\Gamma(y + k)}{\Gamma(k) \times \Gamma(y + 1)} \times \left(\frac{k}{\mu + k}\right)^k \times \left(1 - \frac{k}{\mu + k}\right)^y$$

Это смесь Пуассоновского и Гамма распределений:  $Y$  демонстрируют распределение Пуассона с  $\mu$ , подчиняющимися Гамма-распределению.

**Два параметра** ( $\mu, k$ )

Среднее:  $E(Y) = \mu$  Дисперсия:

$$\text{var}(Y) = \mu + \frac{\mu^2}{k}$$

Параметр  $k$  определяет степень избыточности дисперсии.

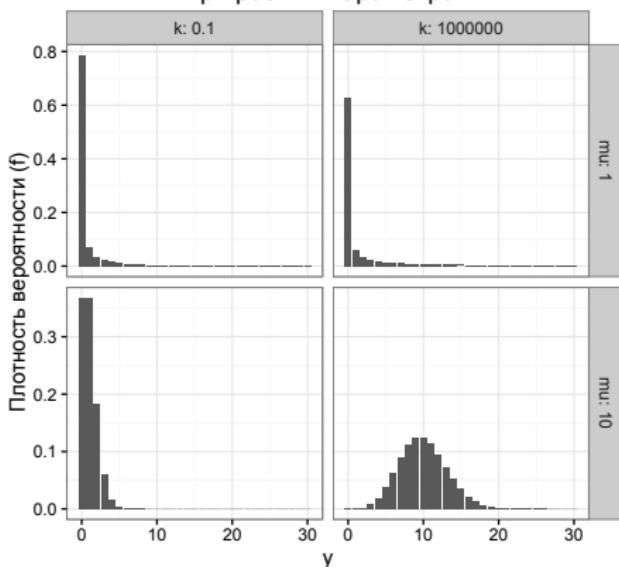
**Важное свойство:** Приближается к распр. Пуассона при очень больших  $k$ .

**Пределы варьирования**

$$0 \leq Y \leq +\infty,$$

$Y$  целочисленные

Отрицательное биномиальное распределение при разных параметрах



# Биномиальное распределение

$$f(y; N, \pi) = \frac{N!}{y! \times (N - y)!} \times \pi^y \times (1 - \pi)^{N-y}$$

**Два параметра ( $N, \pi$ )**

Среднее:  $E(Y) = N \times \pi$  Дисперсия:

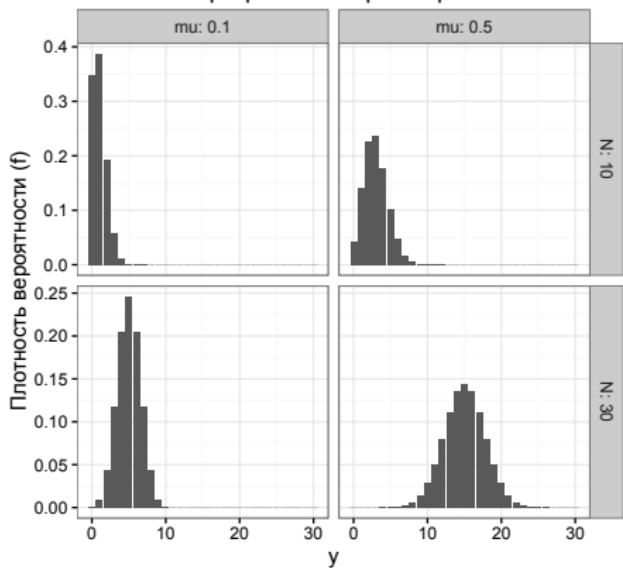
$\text{var}(Y) = N \times \pi \times (1 - \pi)$

Параметр  $N$  определяет количество объектов в испытании Параметр  $\pi$  - вероятность события ( $y = 1$ )

**Пределы варьирования**

$0 \leq Y \leq +\infty$   $Y$  целочисленные

Биномиальное распределение  
при разных параметрах



Распределение зависимой переменной и линия  
регрессии



## Связующая функция (Link function)

Предсказанные значения, т.е.  $E(Y_i) = \mu_i$ , лежат на линии регрессии.

Линейные модели основаны на линейной связи между зависимой переменной и предиктором. Но для некоторых типов распределений это невозможно по природе величин (например, они лежат в интервале  $[0,1]$ ).

Поэтому для каждого типа распределения наиболее “естественным” будет свой особый характер связи.

Функция, описывающая связь между  $\mu_i$  и значениями предикторов, называется *связующей функцией*.

Эта функция преобразует нелинейную зависимость в линейную.



# Наиболее распространенные (канонические) связующие функции

Характер величин	Распределение	Связующая функция (link function)
Непрерывные величины, потенциально варьирующие в пределах $-\infty, +\infty$	Гауссовское (Нормальное)	$\text{identity } X\beta = \mu$
Бинарные величины (1; 0), или количество (доля) объектов, относящихся к одному из двух классов	Биномиальное распределение	$\text{logit } X\beta = \ln(\frac{\mu}{1-\mu})$
Счетные величины (0, 1, 2, 3...)	Распределение Пуассона или Отрицательное биномиальное распределение	$\log X\beta = \ln(\mu)$

**NB!** Есть и другие связующие функции

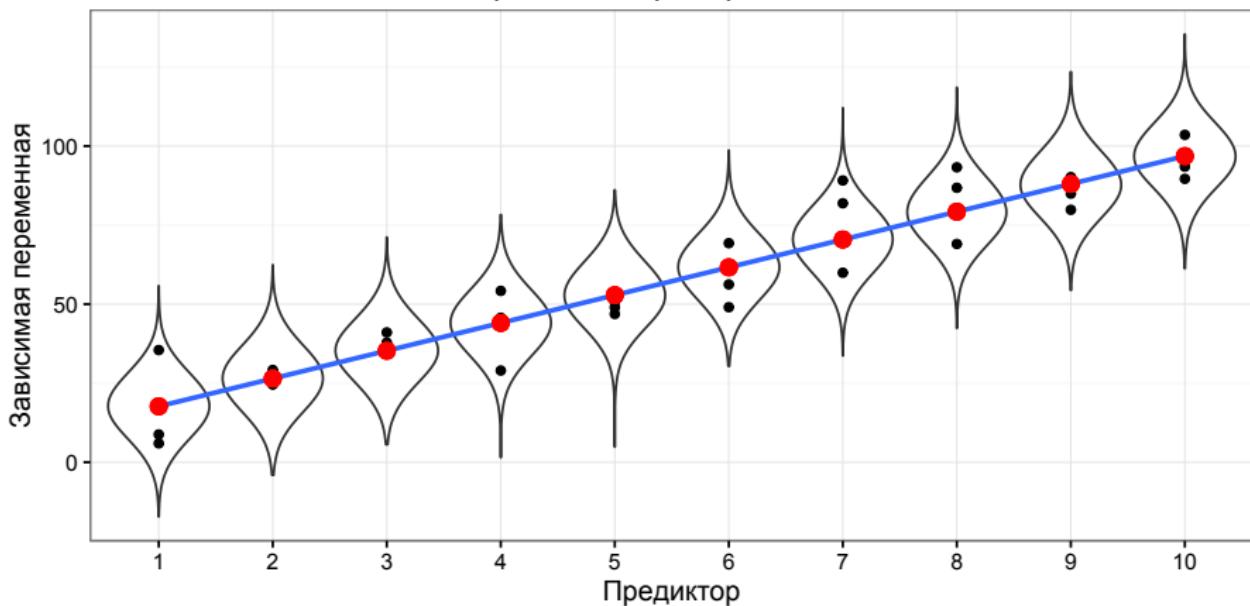


## Связующая функция (Link function)

В случае с нормальным распределением, предсказанные значения лежат на прямой линии, связывающей значения предиктора  $x_i$  и предсказанное значение  $\mu_i$ .

link = "identity"

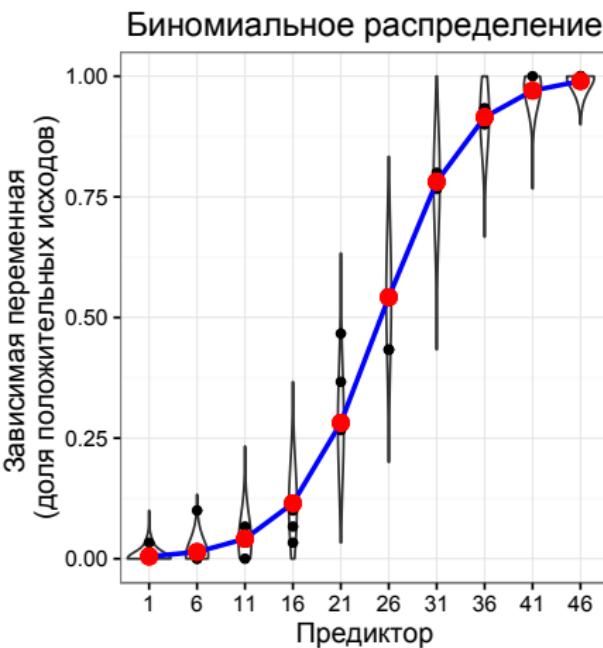
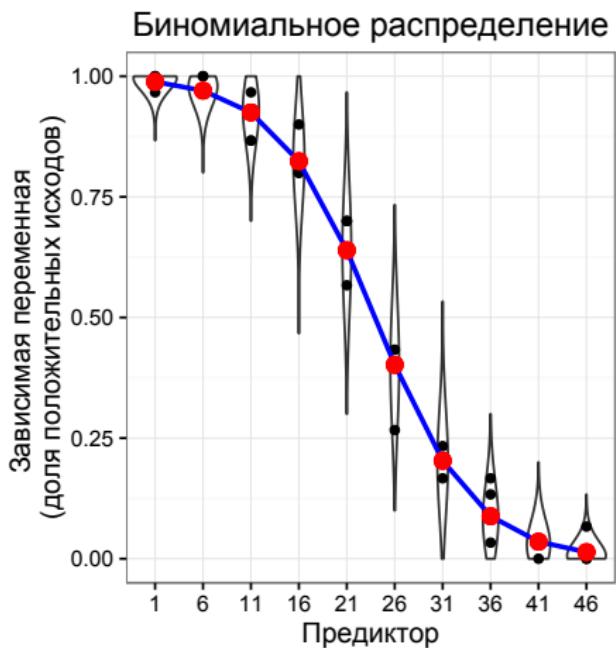
Нормальное распределение



## Связующая функция (Link function)

В случае с биномиальным распределением, предсказанные значения лежат на логистической кривой, связывающей значения предиктора  $x_i$  и предсказанное значение  $\mu_i$ .

link = "logit"

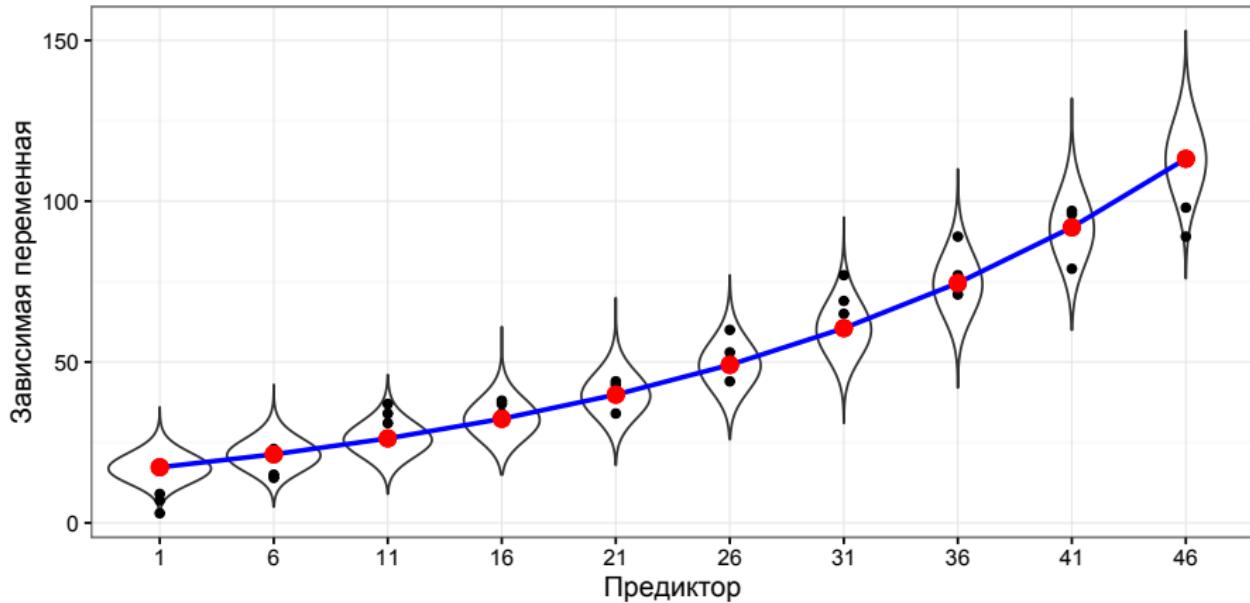


## Связующая функция (Link function)

В случае с пуассоновским распределением, предсказанные значения лежат на экспоненциальной кривой, связывающей значения предиктора  $x_i$  и предсказанное значение  $\mu_i$ .

link = "log"

Пуассоновское распределение



Модели, основанные на распределении Пуассона и  
отрицательном биномиальном распределении



# Способствуют ли взрослые мидии притоку молоди?



```
juv_ad <- read.table("data/mussel_juv_ad.csv")
head(juv_ad, 12)
```

#	Year	Bank	Sample	Juv	Adult
# 1	1997	vor4		1 90	42
# 2	1997	vor4		2 134	25
# 3	1997	vor4		3 166	27
# 4	1997	vor4		4 168	48
# 5	1997	vor4		5 102	25
# 6	1997	vor4		6 69	57
# 7	1998	vor4		1 410	21
# 8	1998	vor4		2 73	76
# 9	1998	vor4		3 347	0
# 10	1998	vor4		5 402	5
# 11	1998	vor4		6 158	28
# 12	1999	vor4		1 282	0

Данные взяты из работы Khaitov, 2013



В этих данных ожидается проблема автокорреляции остатков

Вопрос:

Как можно учесть в модели многолетний характер сбора материала?



## Построим простую линейную модель

```
M1 <- glm(Juv ~ Adult * factor(Year), data = juv_ad)
drop1(M1, test = "F")
```

```
# Single term deletions
#
# Model:
# Juv ~ Adult * factor(Year)
#           Df Deviance AIC F value Pr(>F)
# <none>            256799 1024
# Adult:factor(Year) 14    361458 1026     1.72  0.076 .
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



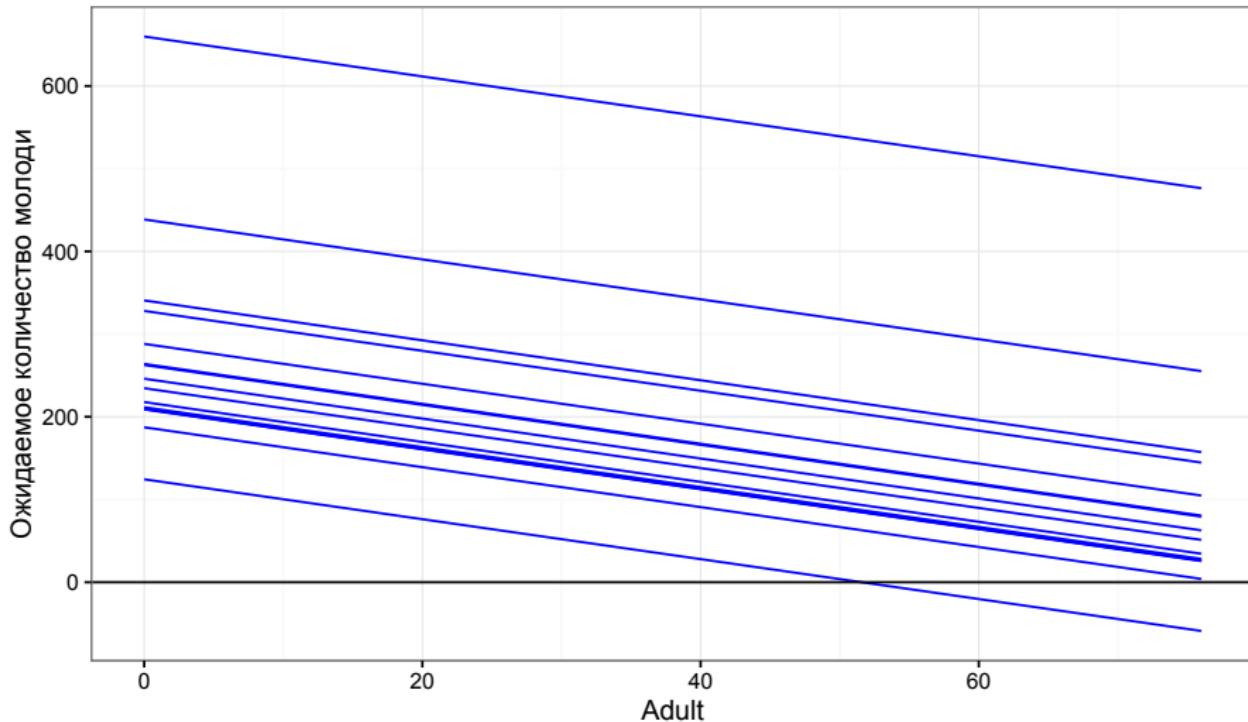
## Можно убрать взаимодействие

```
M2 <- lm(Juv ~ Adult + factor(Year), data = juv_ad)
library(car)
Anova(M2)
```

```
# Anova Table (Type II tests)
#
# Response: Juv
#           Sum Sq Df F value    Pr(>F)
# Adult      63886  1   12.9 0.00059 ***
# factor(Year) 1357772 14   19.6 < 2e-16 ***
# Residuals   361458 73
# ---
# Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



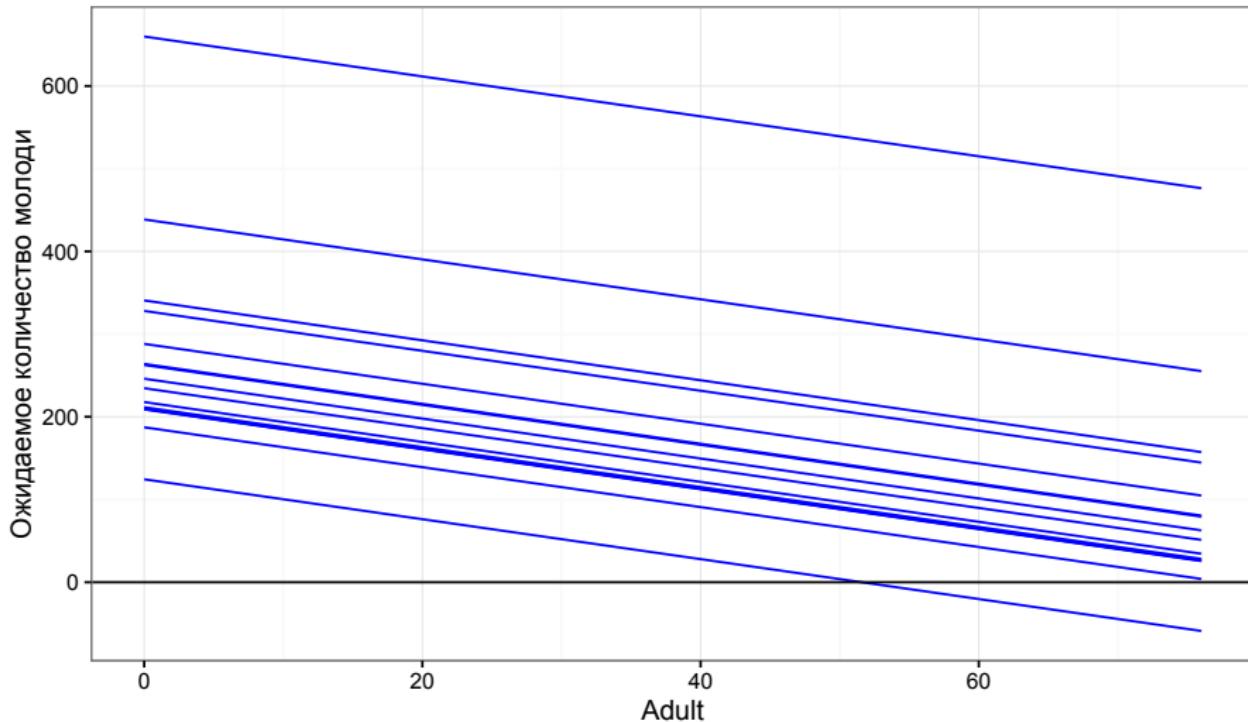
## Посмотрим на предсказания этой модели



- ▶ Модель предсказывает, что взрослые негативно влияют на обилие молоди.



## Посмотрим на предсказания этой модели



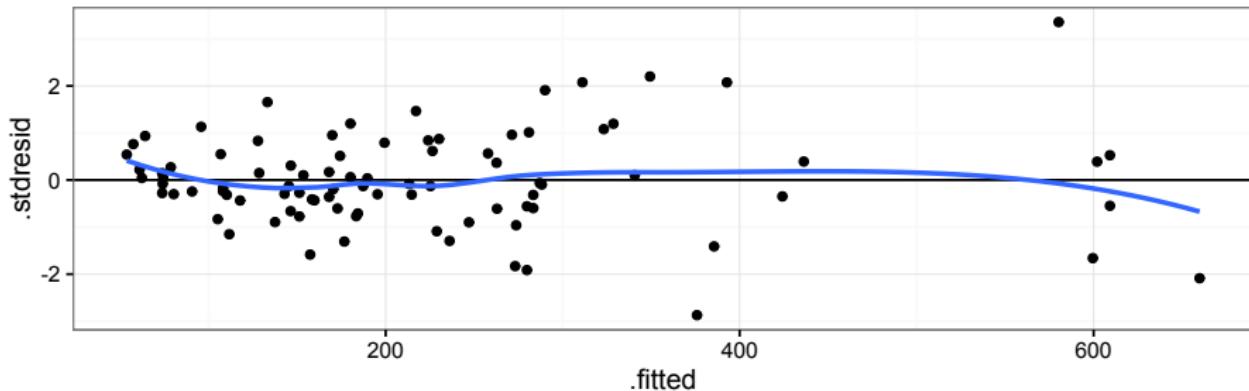
- ▶ Модель предсказывает, что взрослые негативно влияют на обилие молоди.
- ▶ Модель предсказывает отрицательные значения!



## Диагностика модели

```
M2_diag <- fortify(M2)
```

```
ggplot(M2_diag, aes(x = .fitted, y = .stdresid)) + geom_point() +  
  geom_hline(yintercept = 0) + geom_smooth(se = FALSE)
```



- ▶ Явные признаки гетероскедастичности!



## Два способа решения проблем с моделью

1. Провести логарифмирование зависимой переменной и построить модель для логарифмированных величин.
2. Построить модель, основанную на распределении Пуассона.



## Модель, основанная на распределении Пуассона

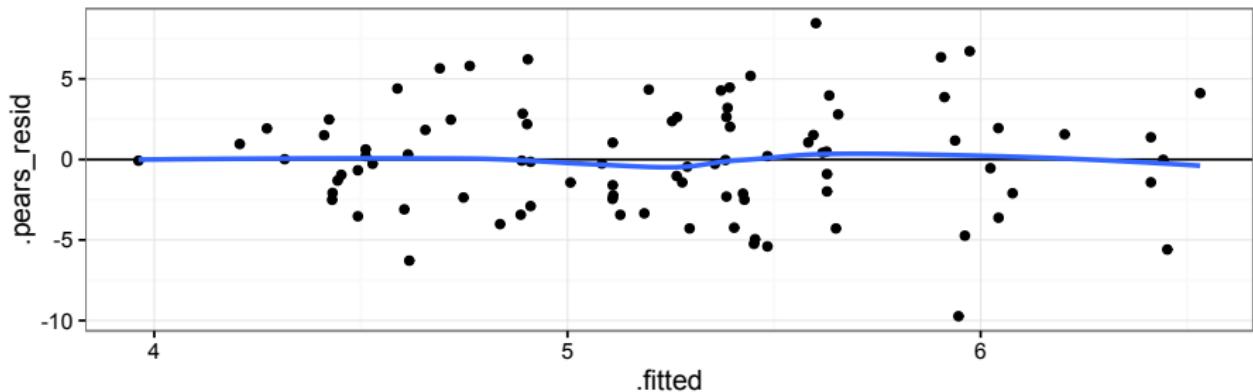
```
M3 <- glm(Juv ~ Adult * factor(Year),  
           data = juv_ad, family = "poisson")  
library(car)  
Anova(M3)
```

```
# Analysis of Deviance Table (Type II tests)  
#  
# Response: Juv  
#  
#          LR Chisq Df Pr(>Chisq)  
# Adult      272    1    <2e-16 ***  
# factor(Year) 5031   14    <2e-16 ***  
# Adult:factor(Year) 439   14    <2e-16 ***  
# ---  
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



## Диагностика модели

```
M3_diag <- data.frame(.fitted = predict(M3),  
                      .pears_resid = residuals(M3, type = "pearson"))
```



Рассеяние остатков выглядит лучше!



## Избыточность дисперсии (Overdispersion)

В Пуассоновской регрессии мы моделируем изменение распределения Пуассона в зависимости от каких-то предикторов.

В распределении Пуассона  $E(Y) = \mu$  и  $\text{var}(Y) = \mu$

Если в имеющихся данных  $\text{var}(Y) > \mu$ , то нарушается условие применимости пуассоновской регрессии.



# Избыточность дисперсии (Overdispersion)

Первый способ оценки избыточности дисперсии

```
Resid_M3 <- resid(M3, type = "pearson") # Пирсоновские остатки  
  
N <- nrow(juv_ad) # Объем выборки  
  
p <- length(coef(M3)) # Число параметров в модели  
  
df <- (N - p) # число степеней свободы  
  
fi <- sum(Resid_M3^2) / df # Величина fi показывает во сколько раз в среднем S  
  
fi  
  
# [1] 16.8
```

Дисперсия в 16.756 раза больше среднего!



# Избыточность дисперсии (Overdispersion)

Второй способ оценки избыточности дисперсии

```
library(qcc)
qcc.overdispersion.test(juv_ad$Juv, type = "poisson")
```

```
#
# Overdispersion test  Obs.Var/Theor.Var Statistic p-value
#      poisson data          103      9079      0
```



# Избыточность дисперсии (Overdispersion)

Очень маленькие стандартные ошибки (и все очень достоверно) - это явный признак избыточности дисперсии

summary(M3)

```
#  
# Call:  
# glm(formula = Juv ~ Adult * factor(Year), family = "poisson",  
#       data = juv_ad)  
#  
# Deviance Residuals:  
#      Min        1Q    Median        3Q       Max  
# -10.779   -2.370   -0.074    2.127    7.858  
#  
# Coefficients:  
#                               Estimate Std. Error z value Pr(>|z|)  
# (Intercept)                 5.14952   0.11551  44.58 < 2e-16 ***  
# Adult                     -0.00955   0.00305  -3.13  0.00173 **  
# factor(Year)1998              0.89375   0.12042   7.42  1.2e-13 ***  
# factor(Year)1999              0.46774   0.12279   3.81  0.00014 ***  
# factor(Year)2000              -0.67585   0.20462  -3.30  0.00096 ***  
# factor(Year)2001               0.09944   0.23147   0.43  0.66748  
# factor(Year)2002               0.16164   0.25066   0.64  0.52526
```



# Источники избыточности дисперсии

## 1. Мнимая избыточность дисперсии

- ▶ Наличие отскакивающих значений
- ▶ Как следствие пропущенных ковариат или взаимодействий предикторов
- ▶ Наличие внутригрупповых корреляций (нарушение независимости выборок)
- ▶ Нелинейный характер взаимосвязи между ковариатами и зависимой переменной
- ▶ Неверно подобранная связывающая функция
- ▶ Количество нулей больше, чем предсказывает распределение Пуассона (Zero inflation)

## 2. Истинная избыточность дисперсии, как следствие природы данных.



## Как бороться с избыточностью дисперсии

Если избыточность дисперсии *минимая*, то ее надо устраниить, введя в модель соответствующие поправки.

Если избыточность дисперсии *истинная*, то необходима более серьезная коррекция модели.



## Два пути решения проблемы при истинной избыточности дисперсии

1. Построить квази-пуассоновскую модель
2. Построить модель, основанную на отрицательном биномиальном распределении



## Квази-пуассоновские модели

Отличие от пуассоновской модели заключается лишь в том, что в квази-пуассоновских моделях вводится поправка для связи дисперсии и матожидания.

В этой модели матожидание  $E(Y) = \mu$  и дисперсия  $var(Y) = \phi \times \mu$

Величина  $\phi$  показывает во сколько раз дисперсия превышает матожидание.

$$\phi = \frac{var(Y)}{\mu} = \frac{\sum_{N-p} (\epsilon_i)^2}{\mu} = \frac{\sum (\epsilon_{pearson})^2}{N - p}$$



## Квази-пуассоновские модели

Отличие от пуассоновской модели заключается лишь в том, что в квази-пуассоновских моделях вводится поправка для связи дисперсии и матожидания.

В этой модели матожидание  $E(Y) = \mu$  и дисперсия  $\text{var}(Y) = \phi \times \mu$

Величина  $\phi$  показывает во сколько раз дисперсия превышает матожидание.

$$\phi = \frac{\text{var}(Y)}{\mu} = \frac{\frac{\sum (\epsilon_i)^2}{N-p}}{\mu} = \frac{\sum (\epsilon_{\text{pearson}})^2}{N-p}$$

Модель, по сути, остается той же, что и пуассоновская, но изменяются стандартные ошибки оценок параметров, они домножаются на  $\sqrt{\phi}$



## Квази-пуассоновские модели

Отличие от пуассоновской модели заключается лишь в том, что в квази-пуассоновских моделях вводится поправка для связи дисперсии и матожидания.

В этой модели матожидание  $E(Y) = \mu$  и дисперсия  $\text{var}(Y) = \phi \times \mu$

Величина  $\phi$  показывает во сколько раз дисперсия превышает матожидание.

$$\phi = \frac{\text{var}(Y)}{\mu} = \frac{\sum_{N-p} (\epsilon_i)^2}{\mu} = \frac{\sum (\epsilon_{\text{pearson}})^2}{N - p}$$

Модель, по сути, остается той же, что и пуассоновская, но изменяются стандартные ошибки оценок параметров, они домножаются на  $\sqrt{\phi}$

Для квази-пуассоновских моделей не определена функция максимального правдоподобия и, следовательно, нельзя вычислить AIC



## Квази-пуассоновская модель

```
M4 <- glm(Juv ~ Adult * factor(Year), data = juv_ad, family = "quasipoisson")
Anova(M4)
```

```
# Analysis of Deviance Table (Type II tests)
#
# Response: Juv
#           LR Chisq Df Pr(>Chisq)
# Adult      16.2   1  0.000056 ***
# factor(Year) 300.3  14  < 2e-16 ***
# Adult:factor(Year) 26.2  14     0.024 *
# ---
# Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Важно:** Распределение разности девианс лишь приблизительно описывается распределением  $\chi^2$ . Поэтому уровни значимости близкие к 0.05 нельзя рассматривать как надежные.

Уровень значимости для взаимодействия в данной модели близок к 0.05. Можно подумать об упрощении модели!



## Квази-пуассоновская модель

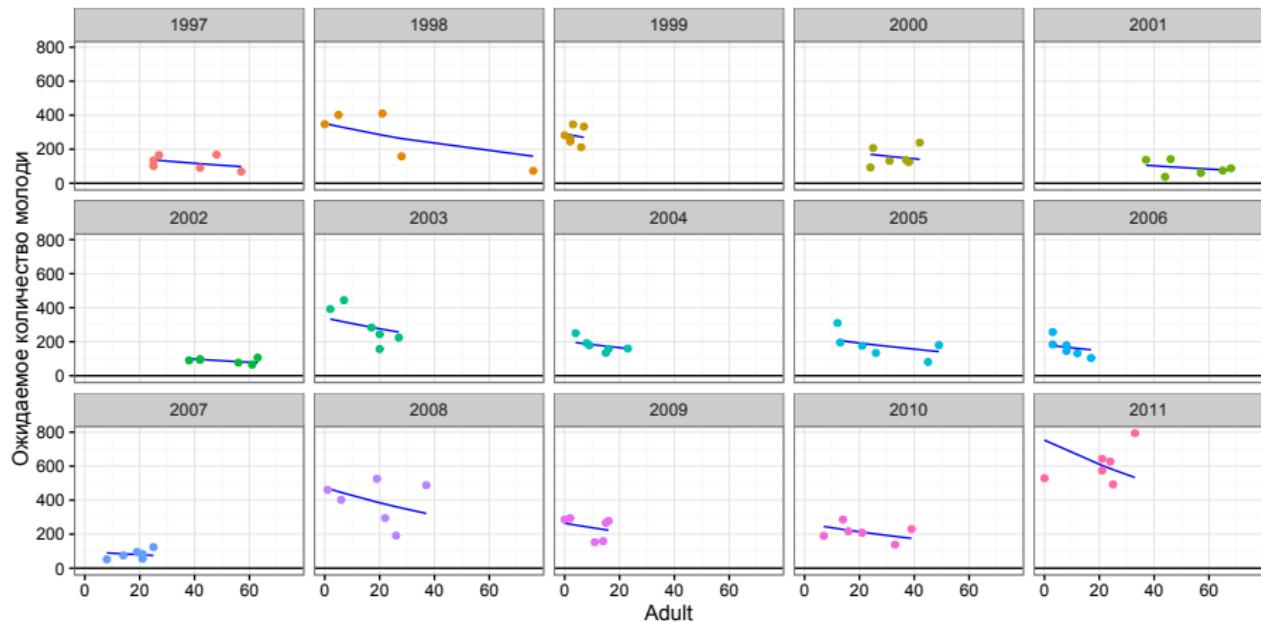
### Упрощенная модель

```
M4a <- glm(Juv ~ Adult + factor(Year), data = juv_ad, family = "quasipoisson")
Anova(M4a)
```

```
# Analysis of Deviance Table (Type II tests)
#
# Response: Juv
#           LR Chisq Df Pr(>Chisq)
# Adult      13.8   1    0.0002 ***
# factor(Year) 254.9 14    <2e-16 ***
# ---
# Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



# Предсказания упрощенной модели



- **Биологический вывод: взрослые мидии препятствуют пополнению молодью**



## Модель, основанная на отрицательном биномиальном распределении

```
library(MASS)
M5 <- glm.nb(Juv ~ Adult*factor(Year) , data = juv_ad, link = "log")
Anova(M5)
```

```
# Analysis of Deviance Table (Type II tests)
#
# Response: Juv
#           LR Chisq Df Pr(>Chisq)
# Adult          27   1 0.00000016 ***
# factor(Year)    370  14   < 2e-16 ***
# Adult:factor(Year) 34  14     0.0023 **
# ---
# Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Уровень значимости для взаимодействия заметно ниже 0.05. Взаимодействие факторов отбросить нельзя!



## Задание

Проверьте на избыточность дисперсии модель, основанную на отрицательном биномиальном распределении



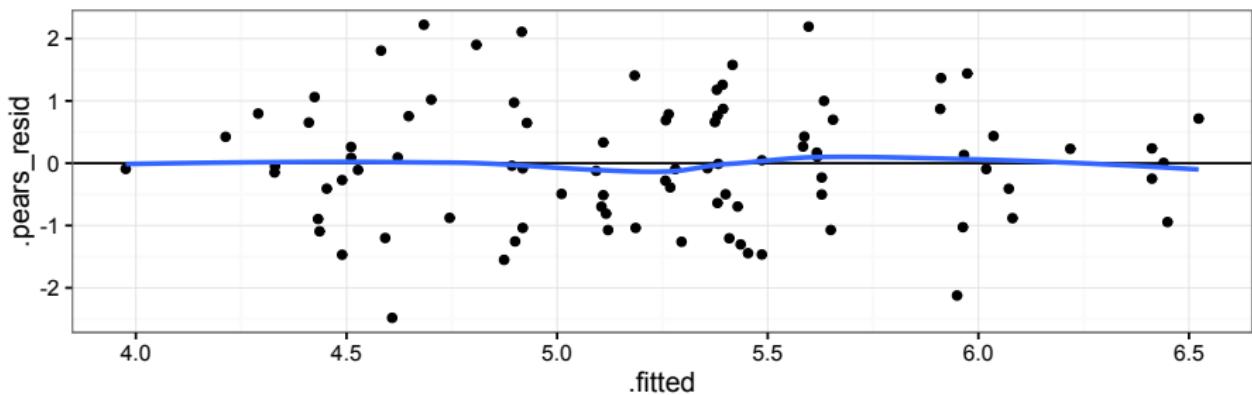
## Решение

```
Resid_M5 <- resid(M5, type = "pearson") # Пирсоновские остатки
N <- nrow(juv_ad) # Объем выборки
p <- length(coef(M5)) +1 # Число параметров в модели
df <- (N - p) # число степеней свободы
fi <- sum(Resid_M5^2) /df # Величина fi показывает
# во сколько раз в среднем  $\sigma$  >  $\tau_i$  для данной модели
fi
# [1] 1.47
```



## Диагностика модели

```
M5_diag <- data.frame(.fitted = predict(M5),  
                      .pears_resid = residuals(M5, type = "pearson"))  
  
ggplot(M5_diag, aes(x=.fitted, y = .pears_resid)) + geom_point() +  
  geom_hline(yintercept = 0) + geom_smooth(se = F)
```

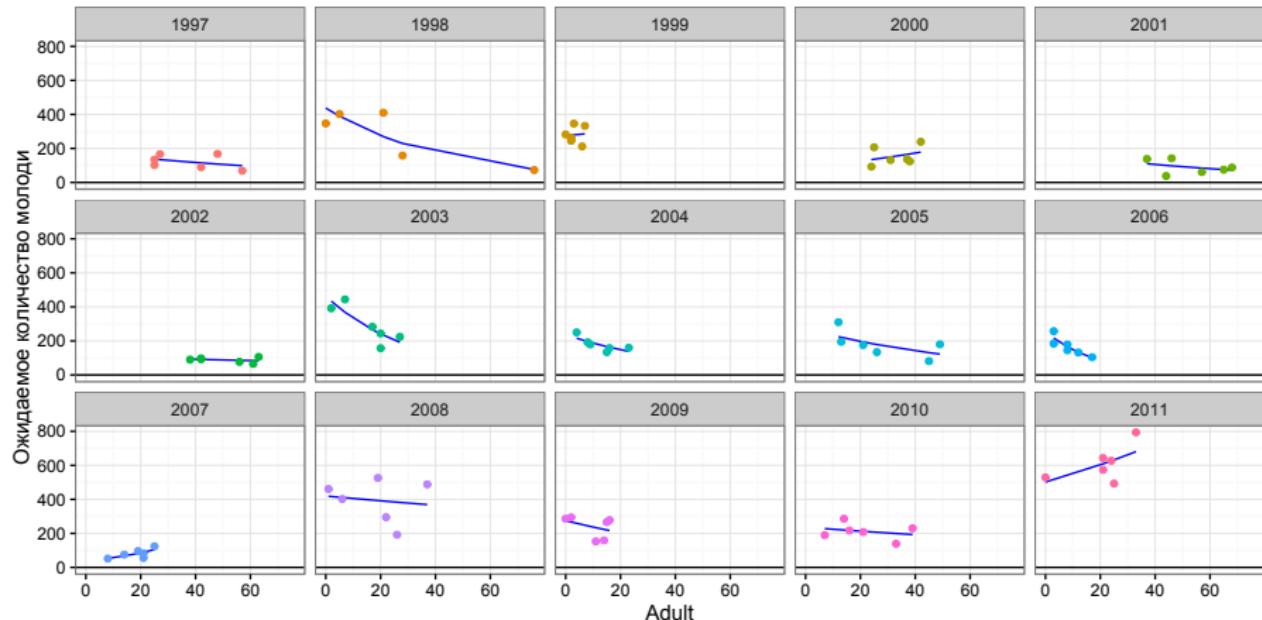


## Задание

Визуализируйте предсказания модели, основанной на отрицательном биномиальном распределении



# Визуализируем предсказание модели



- **Биологический вывод: В разные годы характер взаимосвязи молоди и взрослых мидий может быть разным**



## Код для графика

```
MyData <- unique(juv_ad)

MyData$Predicted <- predict(M5, newdata = MyData, type = "response")
ggplot(MyData, aes(x = Adult, y = Predicted, group = Year)) +
  geom_line(color = "blue") + geom_hline(yintercept = 0) +
  ylab("Ожидаемое количество молоди") + geom_point(data = juv_ad,
  aes(x = Adult, y = Juv, color = factor(Year))) +
  facet_wrap(~Year, ncol = 5) + guides(color = FALSE)
```



## Таким образом

1. Модели, основанные на неподходящем типе распределения, могут давать нереалистичные предсказанные значения.
2. В зависимости от того, как сконструирована модель, можно получить результаты, которые позволят сформулировать принципиально разные биологические выводы.

## Выбор оптимальной модели



# Какие факторы определяют супружескую неверность?

```
data(Affairs, package = "AER")
af <- Affairs
```

affairs - Количество внебрачных связей за последний год

gender - пол

age - возраст

yearsmarried - сколько лет в браке

children - наличие детей

religiousness - уровень религиозности

education - уровень образования

rating - самооценка ощущений от брака



## Задание:

1. Постройте оптимальную модель, описывающую зависимость количества внебрачных связей в зависимости от пола, времени, проведенного в браке, наличия детей, уровня религиозности и уровня образованности.
2. Проверьте валидность данной модели



## Решение

```
Mod <- glm(affairs ~ gender * yearsmarried * children * religiousness +
    education, data = af, family = "poisson")
```

```
summary(Mod)
```

```
#  
# Call:  
# glm(formula = affairs ~ gender * yearsmarried * children * religiousness +  
#       education, family = "poisson", data = af)  
#  
# Deviance Residuals:  
#      Min        1Q    Median        3Q       Max  
# -3.27    -1.78    -1.26    -0.60     6.74  
#  
# Coefficients:  
#  
# (Intercept)          0.6798   0.5380  
# gendermale         -1.2262   0.7071  
# yearsmarried        0.2747   0.0666  
# childrenyes        0.7479   0.6325  
# religiousness      -0.5268   0.1915  
# education          -0.0274   0.0149  
# gendermale:yearsmarried  0.0325   0.0796  
# genderfemale:childrenyes  1.7251   0.8711
```

## Проверка на избыточность дисперсии

```
# Проверка на Overdispersion
Resid_Mod <- resid(Mod, type = "pearson")
N <- nrow(af)
p <- length(coef(Mod))
df <- (N - p)
fi <- sum(Resid_Mod^2) / df
fi
```

```
# [1] 6.44
```



# Строим квази-пуассоновскую модель

```
Mod1 <- glm(affairs ~ gender * yearsmarried * children * religiousness +
             education, data = af, family = "quasipoisson")
```

```
summary(Mod1)
```

```
#  
# Call:  
# glm(formula = affairs ~ gender * yearsmarried * children * religiousness +  
#       education, family = "quasipoisson", data = af)  
#  
# Deviance Residuals:  
#      Min        1Q    Median        3Q       Max  
# -3.27   -1.78   -1.26   -0.60    6.74  
#  
# Coefficients:  
#  
# (Intercept)          0.6798   1.3649  
# gendermale         -1.2262   1.7941  
# yearsmarried        0.2747   0.1689  
# childrenyes         0.7479   1.6048  
# religiousness       -0.5268   0.4858  
# education           -0.0274   0.0378  
# gendermale:yearsmarried 0.0325   0.2021  
# gendermale:childrenyes 1.7851   2.2109  
# yearsmarried:childrenyes -0.2238   0.1877  
# gendermale:religiousness 0.5167   0.6608  
# yearsmarried:religiousness -0.0183   0.0612  
# childrenyes:religiousness 0.1421   0.6101  
# gendermale:yearsmarried:childrenyes -0.0833   0.2287  
# gendermale:yearsmarried:religiousness -0.0282   0.0731  
# gendermale:childrenyes:religiousness -0.7454   0.8159  
# yearsmarried:childrenyes:religiousness 0.0216   0.0676  
# gendermale:yearsmarried:childrenyes:religiousness 0.0499   0.0823  
# t value Pr(>|t|)  
# (Intercept) 0.50   0.62  
# gendermale   -0.68   0.49  
# yearsmarried 1.63   0.10  
# childrenyes  0.47   0.64  
# religiousness -1.08   0.28  
# education    -0.72   0.47  
# gendermale:yearsmarried 0.16   0.87  
# gendermale:childrenyes 0.81   0.42
```



## Подбираем оптимальную модель

```
step(Mod1) #Для квази-пуассоновских моделей эта функция работать  
# не будет, так как не определен AIC
```

## Строим модель, основанную на отрицательном биномиальном распределении

```
Mod_nb <- glm.nb(affairs ~ gender * yearsmarried * children *  
    religiousness + education, data = af)  
Anova(Mod_nb)
```

```
# Analysis of Deviance Table (Type II tests)
```

```
#
```

```
# Response: affairs
```

```
#
```

```
# gender
```

	LR	Chisq	Df	Pr(>Chisq)	
# gender	0.12	1		0.7282	

```
# yearsmarried
```

# yearsmarried	17.87	1		0.000024	***
----------------	-------	---	--	----------	-----

```
# children
```

# children	1.24	1		0.2647	
------------	------	---	--	--------	--

```
# religiousness
```

# religiousness	15.06	1		0.0001	***
-----------------	-------	---	--	--------	-----

```
# education
```

# education	1.93	1		0.1645	
-------------	------	---	--	--------	--

```
# gender:yearsmarried
```

# gender:yearsmarried	0.22	1		0.6366	
-----------------------	------	---	--	--------	--

```
# gender:children
```

# gender:children	0.02	1		0.8975	
-------------------	------	---	--	--------	--

```
# yearsmarried:children
```

# yearsmarried:children	7.98	1		0.0047	**
-------------------------	------	---	--	--------	----

```
# gender:religiousness
```

# gender:religiousness	0.01	1		0.9179	
------------------------	------	---	--	--------	--

```
# yearsmarried:religiousness
```

# yearsmarried:religiousness	0.28	1		0.5965	
------------------------------	------	---	--	--------	--

```
# children:religiousness
```

# children:religiousness	0.02	1		0.8758	
--------------------------	------	---	--	--------	--

```
# gender:yearsmarried:children
```

# gender:yearsmarried:children	0.82	1		0.3662	
--------------------------------	------	---	--	--------	--

```
# gender:yearsmarried:religiousness
```

# gender:yearsmarried:religiousness	0.80	1		0.3715	
-------------------------------------	------	---	--	--------	--

```
# gender:children:religiousness
```

# gender:children:religiousness	2.25	1		0.1337	
---------------------------------	------	---	--	--------	--



## Подбираем оптимальную модель

```
step(Mod_nb)
```



## Смотрим на результаты оптимальной модели

```
Mod_nb_final <- glm.nb(formula = affairs ~ yearsmarried +  
  children + religiousness + yearsmarried:children,  
  data = af, init.theta = 0.1346363532, link = log)
```

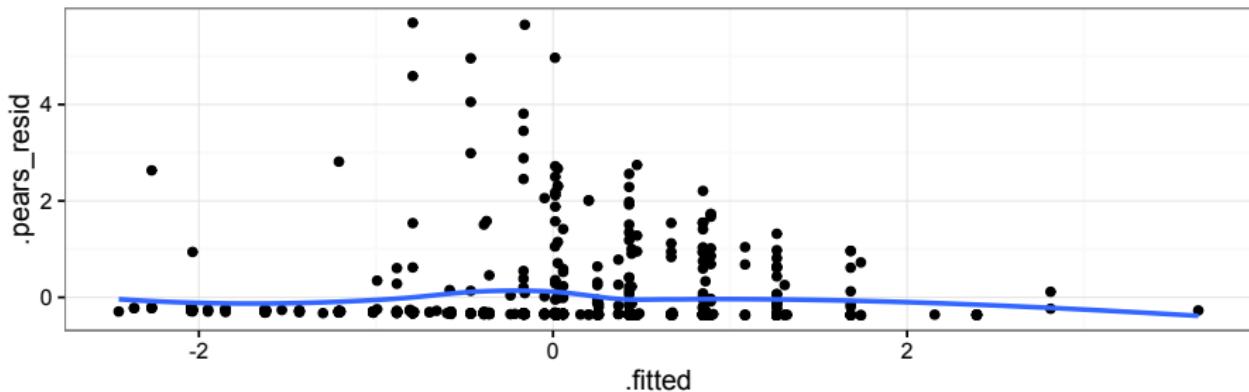
```
summary(Mod_nb_final)
```

```
#  
# Call:  
# glm.nb(formula = affairs ~ yearsmarried + children + religiousness +  
#   yearsmarried:children, data = af, init.theta = 0.1346359871,  
#   link = log)  
#  
# Deviance Residuals:  
#   Min     1Q Median     3Q    Max  
# -1.090  -0.824  -0.700  -0.372   1.910  
#  
# Coefficients:  
#                               Estimate Std. Error z value Pr(>|z|)  
# (Intercept)                 -0.4047    0.4309  -0.94    0.3476  
# yearsmarried                  0.2978    0.0631   4.72 0.0000024 ***  
# childrenyes                  1.3847    0.4621   3.00    0.0027 **  
# religiousness                 -0.4171    0.1064  -3.92 0.0000878 ***  
# yearsmarried:childrenyes   -0.2232    0.0690  -3.24    0.0012 **
```



## Проводим диагностику оптимальной модели

```
Mod_nb_test <- data.frame(.fitted = predict(Mod_nb_final),  
                           .pears_resid = residuals(Mod_nb, type = "pearson"))  
  
ggplot(Mod_nb_test, aes(x=.fitted, y = .pears_resid)) + geom_point() +  
  geom_smooth(se = FALSE)
```



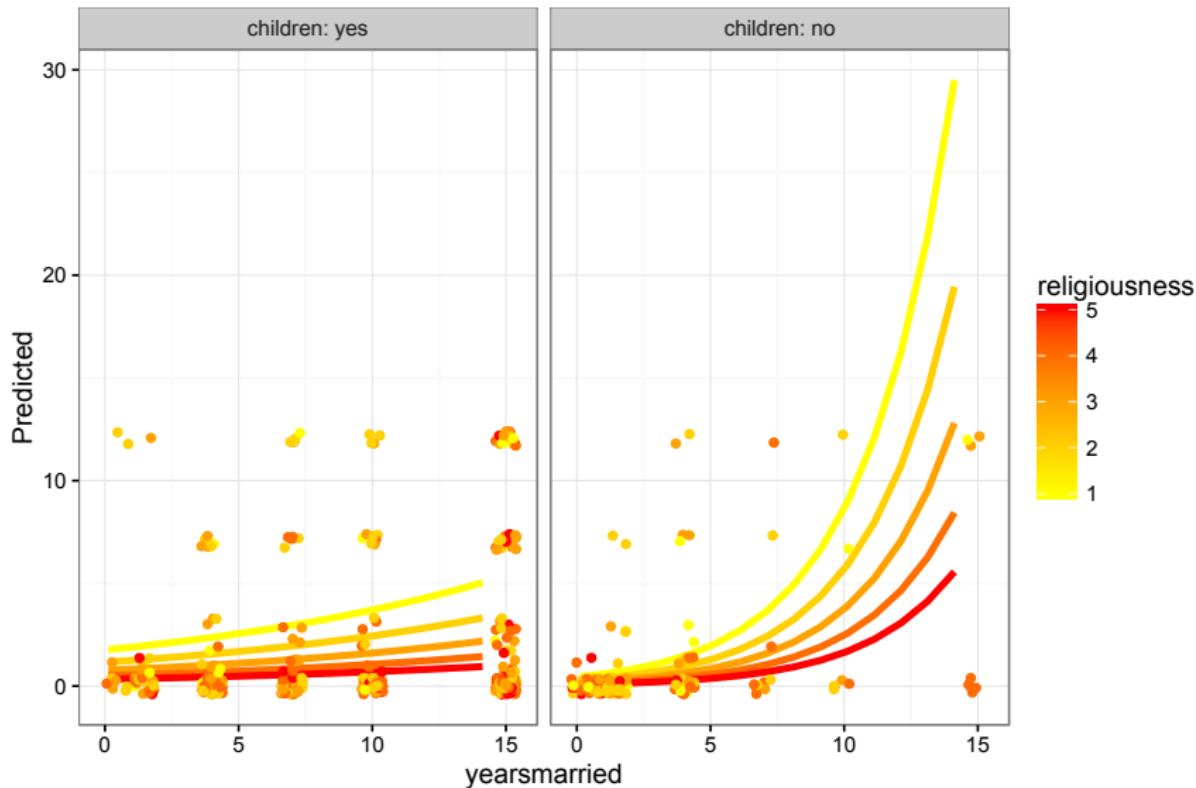
## Проверим на избыточность дисперсии

```
Resid_Mod <- resid(Mod_nb_final, type = "pearson")
N <- nrow(af)
p <- length(coef(Mod_nb_final))
df <- (N - p)
fi <- sum(Resid_Mod^2) / df
fi

# [1] 0.723
```



# Визуализируем предсказание модели



## Код для графика

```
MyData <- expand.grid(yearsmarried = seq(min(af$yearsmarried),  
    max(af$yearsmarried)), children = c("yes", "no"),  
    religiousness = seq(min(af$religiousness), max(af$religiousness)))  
MyData$Predicted <- predict(Mod_nb_final, newdata = MyData,  
    type = "response")  
  
ggplot(MyData, aes(x = yearsmarried, y = Predicted,  
    color = religiousness)) + geom_line(aes(group = religiousness),  
    size = 1.5) + facet_grid(~children, labeller = label_both) +  
    scale_color_gradient(low = "yellow", high = "red") +  
    geom_point(data = af, aes(x = yearsmarried, y = affairs),  
    position = position_jitter(width = 1, height = 1))
```



## Summary

- ▶ В случае счетных зависимых переменных (неотрицательных целочисленных величин) применяются модели, основанные на распределении Пуассона или отрицательном биномиальном распределении.

## Summary

- ▶ В случае счетных зависимых переменных (неотрицательных целочисленных величин) применяются модели, основанные на распределении Пуассона или отрицательном биномиальном распределении.
- ▶ Важным ограничивающим условием применения этих моделей является отсутствие избыточности дисперсии.



## Summary

- ▶ В случае счетных зависимых переменных (неотрицательных целочисленных величин) применяются модели, основанные на распределении Пуассона или отрицательном биномиальном распределении.
- ▶ Важным ограничивающим условием применения этих моделей является отсутствие избыточности дисперсии.
- ▶ Избыточность дисперсии может быть истинной и мнимой.



## Summary

- ▶ В случае счетных зависимых переменных (неотрицательных целочисленных величин) применяются модели, основанные на распределении Пуассона или отрицательном биномиальном распределении.
- ▶ Важным ограничивающим условием применения этих моделей является отсутствие избыточности дисперсии.
- ▶ Избыточность дисперсии может быть истинной и мнимой.
- ▶ При истинной избыточности дисперсии модель можно скорректировать, построив квази-пуассоновскую модель (вводятся поправки для ошибок оценок коэффициентов модели).



## Summary

- ▶ В случае счетных зависимых переменных (неотрицательных целочисленных величин) применяются модели, основанные на распределении Пуассона или отрицательном биномиальном распределении.
- ▶ Важным ограничивающим условием применения этих моделей является отсутствие избыточности дисперсии.
- ▶ Избыточность дисперсии может быть истинной и мнимой.
- ▶ При истинной избыточности дисперсии модель можно скорректировать, построив квази-пуассоновскую модель (вводятся поправки для ошибок оценок коэффициентов модели).
- ▶ Другой подход - построение моделей, основанных на отрицательном биномиальном распределении.



## Что почитать

- ▶ Кабаков Р.И. R в действии. Анализ и визуализация данных на языке R. М.: ДМК Пресс, 2014.
- ▶ Zuur, A.F. et al. 2009. Mixed effects models and extensions in ecology with R. - Statistics for biology and health. Springer, New York, NY.

