

# Обобщенные линейные модели с нормальным распределением остатков

Линейные модели...

Марина Варфоломеева, Вадим Хайтов

Кафедра Зоологии беспозвоночных, Биологический факультет, СПбГУ



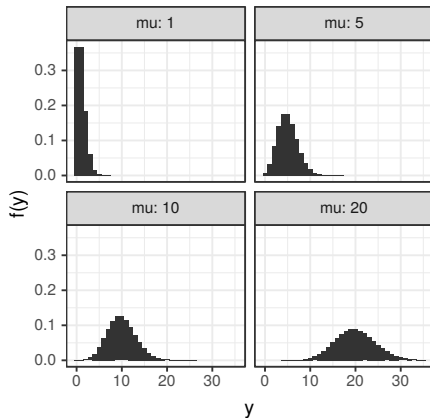
# Мы рассмотрим

- ▶ Варианты анализа для случаев, когда зависимая переменная — счетная величина (целые неотрицательные числа)

## Вы сможете

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

# Распределение Пуассона



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

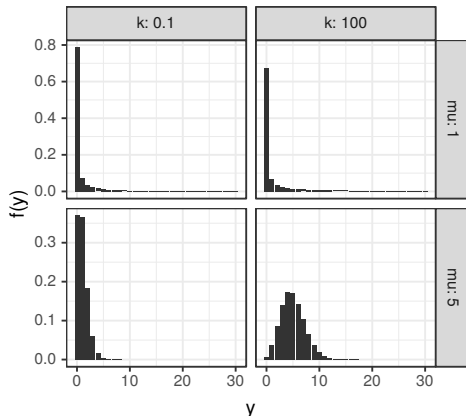
Параметр:

- ▶  $\mu$  – задает среднее и дисперсию

Свойства:

- ▶  $E(y) = \mu$  — мат.ожидание
- ▶  $var(y) = \mu$  — дисперсия
- ▶  $0 \leq y \leq +\infty, y \in \mathbb{N}$  — значения

# Отрицательное биномиальное распределение



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

Параметры:

- ▶  $\mu$  – среднее
- ▶  $k$  – определяет степень избыточности дисперсии

Свойства:

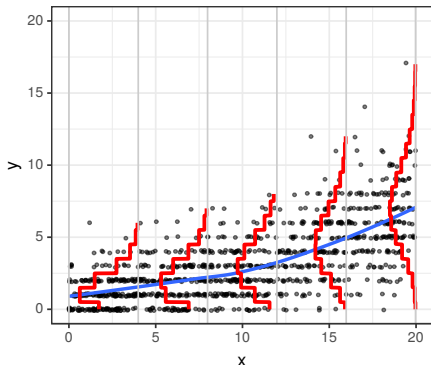
- ▶  $E(y) = \mu$  – мат. ожидание
- ▶  $var(y) = \mu + \frac{\mu^2}{k}$  – дисперсия
- ▶  $0 \leq y \leq +\infty, y \in \mathbb{N}$  – значения

# GLM в "туннеле" из распределений

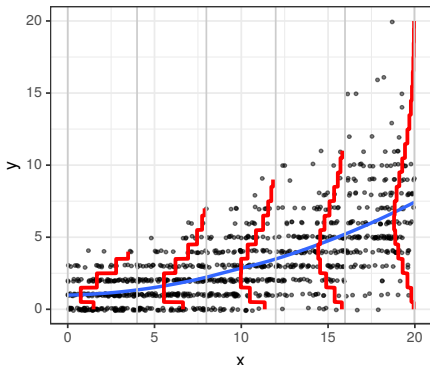
В каждом конкретном случае при анализе данных нам предстоит выяснить, какое из распределений больше подходит для моделирования отклика.

- ▶ Достаточно ли возможностей распределения Пуассона, чтобы описать связь среднего и дисперсии?
- ▶ Если нет, то справится ли отрицательное биномиальное распределение, у которого есть параметр, для описания этой связи?

GLM с Пуассоновским распределением отклика



GLM с отрицательным биномиальным распределением отклика



## Гадючий лук, копеечник и визиты опылителей

Гадючий лук (мускари, *Leopoldia comosa*) — представитель родной флоры острова Менорка. В 18-19вв на остров завезли копеечник венечный (*Hedysarum coronarium*), который быстро натурализовался. Оба вида цветут одновременно и нуждаются в опылении насекомыми.

Как зависит число визитов опылителей на цветки мускари от присутствия вселенца и разнообразия флоры в ближайшей окрестности? (Данные Montero-Castaño, Vilà, 2015)



Muscari à toupet (Muscari comosum), Dordogne, France — Père Igor



French-honeysuckle. Close to Santadi Basso, Sardinia, Italy — Hans Hillewaert

# Дизайн исследования

Подсчитывали число визитов опылителей на выбранное растение гадючьего лука (в пунктирной рамке) на трех типах участков.

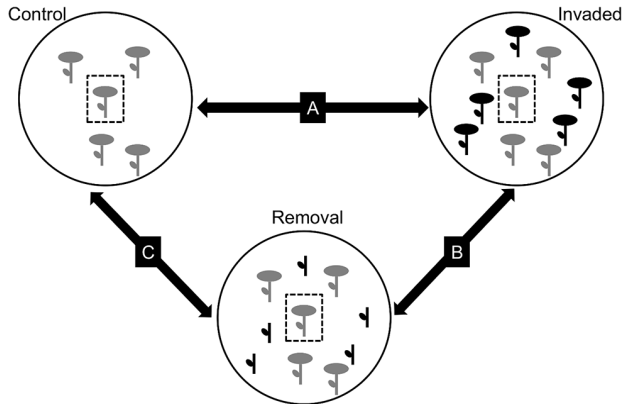


Fig.2 из Montero-Castaño, Vilà, 2015

<https://doi.org/10.1371/journal.pone.0128595>

# Переменные

- ▶ Visits — число визитов всех опылителей на цветок *Leopoldia*
- ▶ Treatment — тип площадки, тритмент (фактор с 3 уровнями):
  - ▶ Invaded — *Leopoldia* в смеси с видом-вселенцем;
  - ▶ Removal — *Leopoldia* в смеси с видом-вселенцем с удаленными цветками;
  - ▶ Control — *Leopoldia* без вида вселенца.
- ▶ DiversityD\_1 — Разнообразие флоры на площадке ( $\exp(H')$ , где  $H'$  — индекс Шеннона-Уивера)  
(на луг с более разнообразной растительностью прилетит больше опылителей).
- ▶ Flowers — число цветков *Leopoldia* на площадке (чем больше, тем больше опылителей).
- ▶ Hours — продолжительность наблюдений (чем дольше, тем больше насчитали).

Другие переменные:

- ▶ Total\_1 — общая плотность цветков
- ▶ Visits\_NO\_Apis — посещения опылителей без учета пчел
- ▶ Fruit — число цветов с плодами через месяц
- ▶ No\_Fruit — число цветов без плодов через месяц



# Открываем из знакомимся с данными

```
library(readxl)
pol <- read_excel("data/Pollinators_Montero-Castano, Vila, 2015.xlsx", sheet = 1)
head(pol)
```

```
# # A tibble: 6 x 10
#   Individual Treatment DiversityD_1 Visits Visits_NO_Apis Total_1
#   <dbl> <chr> <dbl> <dbl> <dbl> <dbl>
# 1 1 Removal 2.66 4 4 53.9
# 2 2 Removal 1 6 6 2.45
# 3 3 Removal 1.44 5 5 41.6
# 4 4 Removal 2.21 1 1 58.8
# 5 5 Removal 2.83 6 6 19.6
# 6 6 Removal 2.40 2 2 78.4
# # ... with 4 more variables: Fruit <dbl>, No_Fruit <dbl>,
# # Flowers <dbl>, Hours <dbl>
```

Сколько пропущенных значений?

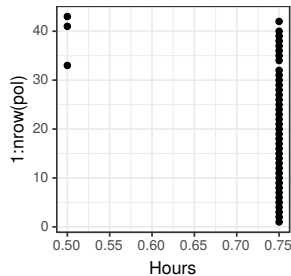
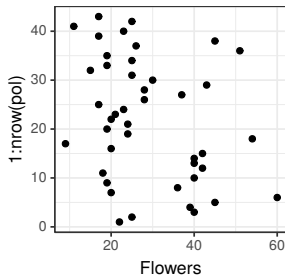
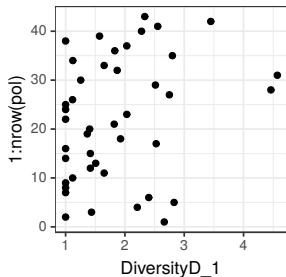
```
colSums(is.na(pol))
```

```
#   Individual      Treatment DiversityD_1      Visits
#           0           0           0           0
# Visits_NO_Apis      Total_1      Fruit      No_Fruit
#           0           0           0           0
#   Flowers      Hours
#           0           0
```

# Есть ли выбросы?

```
library(cowplot)
library(ggplot2)
theme_set(theme_bw())
```

```
dot_plot <- ggplot(pol, aes(y = 1:nrow(pol))) + geom_point()
plot_grid(dot_plot + aes(x = DiversityD_1), dot_plot + aes(x = Flowers),
          dot_plot + aes(x = Hours), nrow = 1)
```



Выбросов нет.

Периоды наблюдений имеют разную продолжительность. Нужно это учесть в модели.

# Каков объем выборки?

```
table(pol$Treatment)
```

```
#  
# Control Invaded Removal  
#      14      11      18
```

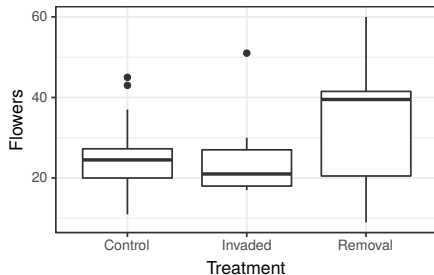
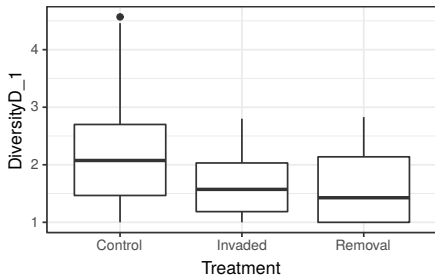
Как распределены короткие периоды наблюдений по тритментам?

```
table(pol$Hours, pol$Treatment)
```

```
#  
#      Control Invaded Removal  
#    0.5        2        1        0  
#    0.75       12       10       18
```

# Коллинеарны ли непрерывные и дискретные предикторы?

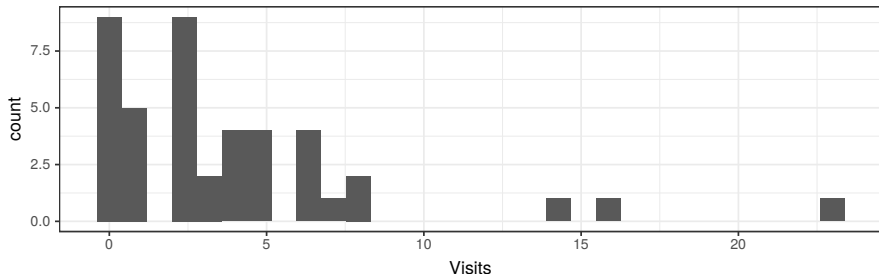
```
box_plot <- ggplot(pol, aes(x = Treatment)) + geom_boxplot()  
  
plot_grid(box_plot + aes(y = DiversityD_1),  
          box_plot + aes(y = Flowers), nrow = 1)
```



Возможно, есть коллинеарность.

# Как распределена переменная-отклик?

```
ggplot(pol, aes(x = Visits)) + geom_histogram()
```



```
mean(pol$Visits == 0) # Какова пропорция нулей?
```

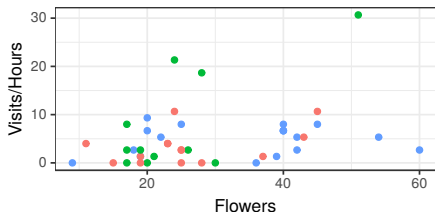
```
# [1] 0.2093023
```

Число визитов насекомых – счетная переменная. Для ее моделирования нужно использовать подходящее распределение.

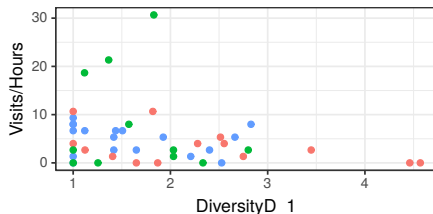
Примерно 21 % наблюдений – нули. Иногда из-за избытка нулей (Zero inflation) в модели может появиться избыточность дисперсии. Будем иметь это в виду.

# Линейна ли связь между предикторами и откликом?

```
gg_shape <- ggplot(pol, aes(y = Visits/Hours, colour = Treatment)) +  
  theme(legend.position = 'bottom')  
plot_grid(  
  gg_shape + geom_point(aes(x = Flowers)),  
  gg_shape + geom_point(aes(x = DiversityD_1)),  
  nrow = 1)
```



Treatment    Control    Invaded    Removal



Treatment    Control    Invaded    Removal

Связь практически линейна.

## Если мы (ошибочно) подберем GLM с нормальным распределением отклика?

$$Visits_i \sim N(\mu_i, \sigma)$$

$$E(Visits_i) = \mu_i, \text{ var}(Visits_i) = \sigma^2$$

$\mu_i = \eta_i$  – функция связи “идентичность”

$$\eta_i = b_0 + b_1 Treatment_{Invaded\ i} + b_2 Treatment_{Removal\ i} + b_3 DiversityD1_i + b_4 Flowers_i + b_5 Hours_i$$

```
M_norm <- glm(Visits ~ Treatment + DiversityD_1 + Flowers + Hours, data = pol)
coef(M_norm)
```

#	(Intercept)	TreatmentInvaded	TreatmentRemoval	DiversityD_1
#	-3.254511	2.890799	-1.061720	-1.204077
#	Flowers	Hours		
#	0.143856	6.654556		

```
sigma(M_norm)
```

```
# [1] 4.173941
```

# Данные для графика предсказаний простой линейной модели

```
library(dplyr)
NewData <- pol %>% group_by(Treatment)%>%
  do(data.frame(Flowers = seq(min(. $Flowers), max(. $Flowers), length.out=50))) %>%
  mutate(DiversityD_1 = mean(pol$DiversityD_1),
         Hours = mean(pol$Hours))

# Модельная матрица и коэффициенты
X <- model.matrix(~ Treatment + DiversityD_1 + Flowers + Hours, data = NewData)
b <- coef(M_norm)
# Предсказания в масштабе функции связи (eta) совпадают с масштабом отклика (mu)
NewData$mu <- X %*% b
NewData$SE_mu <- sqrt(diag(X %*% vcov(M_norm) %*% t(X))) # SE

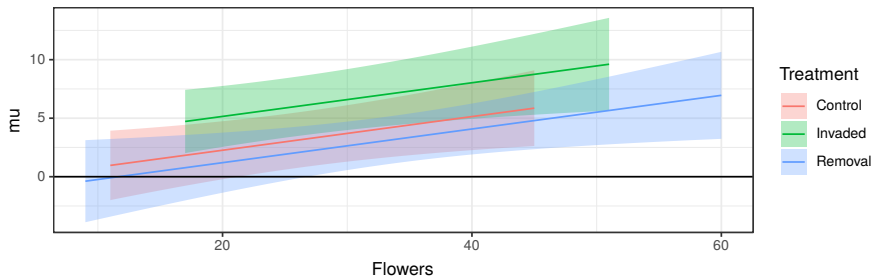
head(NewData, 3)

## # A tibble: 3 x 6
## # Groups:   Treatment [1]
##   Treatment Flowers DiversityD_1 Hours      mu SE_mu
##   <chr>      <dbl>      <dbl> <dbl> <dbl> <dbl>
## 1 Control      11          1.86 0.733 0.966  1.48
## 2 Control     11.7          1.86 0.733 1.07   1.46
## 3 Control     12.4          1.86 0.733 1.17   1.44
```



# График предсказаний

```
ggplot(NewData, aes(x = Flowers, y = mu, fill = Treatment)) +  
  geom_ribbon(aes(ymin = mu - 2 * SE_mu, ymax = mu + 2 * SE_mu), alpha=0.3) +  
  geom_line(aes(colour = Treatment)) +  
  geom_hline(yintercept = 0)
```



# Смотрим на результаты подбора модели

```
summary(M_norm)
```

```
#
# Call:
# glm(formula = Visits ~ Treatment + DiversityD_1 + Flowers + Hours,
#      data = pol)
#
# Deviance Residuals:
#      Min       1Q   Median       3Q      Max
# -7.4320  -2.3611  -0.3929   1.0335  13.2385
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  -3.25451    7.73105  -0.421   0.6762
# TreatmentInvaded  2.89080    1.76262   1.640   0.1095
# TreatmentRemoval -1.06172    1.67758  -0.633   0.5307
# DiversityD_1    -1.20408    0.78510  -1.534   0.1336
# Flowers         0.14386    0.05851   2.458   0.0188 *
# Hours          6.65456   10.64239   0.625   0.5356
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for gaussian family taken to be 17.42178)
#
#      Null deviance: 891.86  on 42  degrees of freedom
# Residual deviance: 644.61  on 37  degrees of freedom
# AIC: 252.45
#
# Number of Fisher Scoring iterations: 2
```

## Анализ девиансы для модели с нормальным распределением отклика

```
drop1(M_norm, test = 'Chi')
```

```
# Single term deletions
#
# Model:
# Visits ~ Treatment + DiversityD_1 + Flowers + Hours
#
#           Df Deviance      AIC scaled dev. Pr(>Chi)
# <none>          644.61 252.45
# Treatment      2   744.02 254.62      6.1672  0.04579 *
# DiversityD_1   1   685.58 253.10      2.6502  0.10354
# Flowers        1   749.91 256.95      6.5062  0.01075 *
# Hours          1   651.42 250.90      0.4520  0.50138
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Число визитов опылителей на цветки гадючьего лука:

- ▶ НЕ зависит от присутствия вселенца и его цветов,
- ▶ НЕ зависит от разнообразия флоры на участке,
- ▶ зависит от числа цветов самого гадючьего лука.

Можем ли мы доверять этим результатам? Пока не известно.

# Нет ли коллинеарности предикторов

```
library(car)  
vif(M_norm)
```

#		GVIF	Df	$GVIF^{(1/(2*Df))}$
# Treatment		1.356502	2	1.079208
# DiversityD_1		1.162061	1	1.077989
# Flowers		1.226279	1	1.107375
# Hours		1.133910	1	1.064852

Коллинеарности нет.

# Задание 1

Постройте график пирсоновских остатков от предсказанных значений для модели `M_norm`.

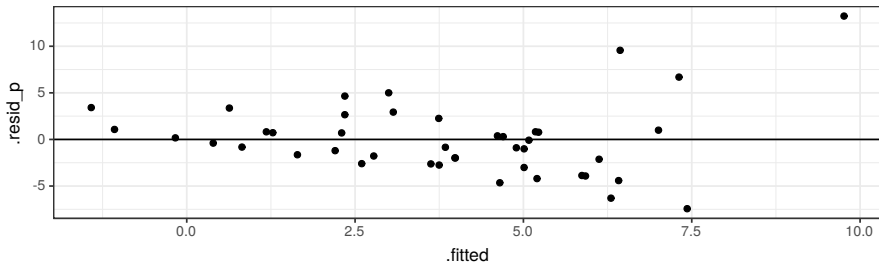
Какие нарушения условий применимости вы на нем видите?

Дополните код:

```
M_norm_diag <- data.frame(.fitted = fitted(),  
                           .resid_p = residuals())  
  
ggplot(data = , aes()) + geom_hline( = 0) +  
  geom_point()
```

## График остатков от предсказанных значений

```
M_norm_diag <- data.frame(.fitted = predict(M_norm, type = "response"),  
                          .resid_p = residuals(M_norm, type = "pearson"))  
  
ggplot(M_norm_diag, aes(y = .resid_p)) + geom_hline(yintercept = 0) +  
  geom_point(aes(x = .fitted))
```



Гетерогенность дисперсий остатков.

Отрицательные предсказания!

# Модель с нормальным распределением отклика не подходит

Два способа решения проблем с моделью:

1. Грубый способ: логарифмировать зависимую переменную и построить модель для нее.
2. Лучше построить модель, основанную на распределении, подходящем для счетных данных:
  - ▶ распределение Пуассона,
  - ▶ отрицательное биномиальное распределение.

# GLM с Пуассоновским распределением отклика

$$Visits_i \sim \text{Poisson}(\mu_i)$$

$$E(Visits_i) = \mu_i, \text{var}(Visits_i) = \mu_i$$

$\ln(\mu_i) = \eta_i$  — функция связи логарифм

$$\eta_i = b_0 + b_1 \text{Treatment}_{Invaded\ i} + b_2 \text{Treatment}_{Removal\ i} + \\ + b_3 \text{DiversityD1}_i + b_4 \text{Flowers}_i + b_5 \text{Hours}_i$$

```
M_pois <- glm(Visits ~ Treatment + DiversityD_1 + Flowers + Hours, data = pol,  
              family = "poisson")
```



# Уравнение модели с Пуассоновским распределением отклика

$$Visits_i \sim \text{Poisson}(\mu_i)$$

$$E(Visits_i) = \mu_i, \text{var}(Visits_i) = \mu_i$$

$$\ln(\mu_i) = \eta_i$$

$$\eta_i = -2.66 + 0.71Treatment_{Invaded\ i} - 0.22Treatment_{Removal\ i} - \\ - 0.46DiversityD1_i + 0.04Flowers_i + 4.69Hours_i$$

`coef(M_pois)`

#	(Intercept)	TreatmentInvaded	TreatmentRemoval
#	-2.66090631	0.71341797	-0.21537935
#	DiversityD_1	Flowers	Hours
#	-0.45740225	0.03731004	4.68668983

# Смотрим на результаты подбора модели

```
summary(M_pois)
```

```
#
# Call:
# glm(formula = Visits ~ Treatment + DiversityD_1 + Flowers + Hours,
#      family = "poisson", data = pol)
#
# Deviance Residuals:
#      Min       1Q   Median       3Q      Max
# -4.0675  -1.3189  -0.3523   0.7068   3.2346
#
# Coefficients:
#              Estimate Std. Error z value Pr(>|z|)
# (Intercept)  -2.660906   2.174439  -1.224   0.221058
# TreatmentInvaded  0.713418   0.214155   3.331   0.000864 ***
# TreatmentRemoval -0.215379   0.222648  -0.967   0.333368
# DiversityD_1    -0.457402   0.128586  -3.557   0.000375 ***
# Flowers         0.037310   0.006835   5.459 0.0000000479 ***
# Hours           4.686690   2.903254   1.614   0.106465
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for poisson family taken to be 1)
#
#      Null deviance: 189.75  on 42  degrees of freedom
# Residual deviance: 119.88  on 37  degrees of freedom
# AIC: 238.78
#
# Number of Fisher Scoring iterations: 6
```

Угловые коэффициенты показывают, на сколько единиц меняется значение логарифма отклика, если соответствующий предиктор изменяется на единицу.

Это значит, что сам отклик изменяется в  $e^{b_k}$  раз при изменении предиктора на единицу.



## Анализ девиансы для модели с Пуассоновским распределением отклика

```
drop1(M_pois, test = 'Chi')
```

```
# Single term deletions
#
# Model:
# Visits ~ Treatment + DiversityD_1 + Flowers + Hours
#
```

	Df	Deviance	AIC	LRT	Pr(>Chi)	
# <none>		119.88	238.78			
# Treatment	2	145.29	260.19	25.4091	0.00000303722	***
# DiversityD_1	1	134.31	251.21	14.4290	0.0001455	***
# Flowers	1	148.67	265.56	28.7871	0.00000008079	***
# Hours	1	123.64	240.53	3.7561	0.0526148	.
# ---						
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Число визитов опылителей на цветки гадючьего лука:

- ▶ зависит от присутствия вида вселенца и его цветов,
- ▶ зависит от разнообразия флоры на данном участке,
- ▶ зависит от числа цветов самого гадючьего лука.

Можем ли мы доверять этим результатам? Пока не известно.

# Данные для предсказаний

```
NewData <- pol %>% group_by(Treatment)%>%  
  do(data.frame(Flowers = seq(min(.$Flowers), max(.$Flowers), length.out=50))) %>%  
  mutate(DiversityD_1 = mean(pol$DiversityD_1),  
         Hours = mean(pol$Hours))
```

Давайте получим предсказания двумя способами:

- ▶ при помощи операций с матрицами,  
чтобы своими глазами увидеть работу функции связи,
- ▶ при помощи функции predict().

```
?predict.glm
```

# Предсказания модели при помощи операций с матрицами

```
# Модельная матрица и коэффициенты
X <- model.matrix(~ Treatment + DiversityD_1 + Flowers + Hours, data = NewData)
b <- coef(M_pois)

# Предсказанные значения и стандартные ошибки...
# ...в масштабе функции связи (логарифм)
NewData$fit_eta <- X %*% b
NewData$SE_eta <- sqrt(diag(X %*% vcov(M_pois) %*% t(X)))

# ...в масштабе отклика (применяем функцию, обратную функции связи)
NewData$fit_mu <- exp(NewData$fit_eta)
NewData$lwr <- exp(NewData$fit_eta - 2 * NewData$SE_eta)
NewData$upr <- exp(NewData$fit_eta + 2 * NewData$SE_eta)

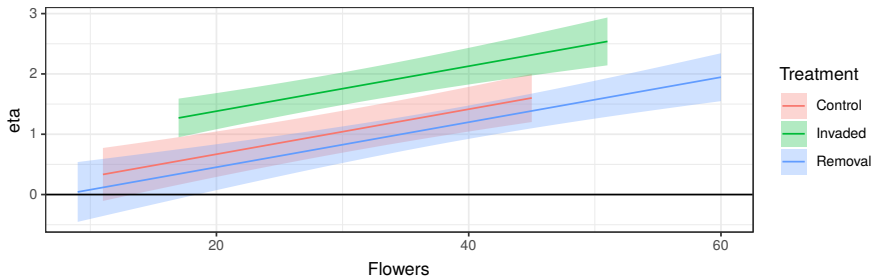
head(NewData, 2)
```

```
# # A tibble: 2 x 9
# # Groups:   Treatment [1]
#   Treatment Flowers DiversityD_1 Hours fit_eta SE_eta fit_mu lwr
#   <chr>      <dbl>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
# 1 Control      11          1.86 0.733  0.333  0.220  1.40 0.899
# 2 Control     11.7          1.86 0.733  0.359  0.217  1.43 0.928
# # ... with 1 more variable: upr <dbl>
```

## График предсказаний в масштабе функции связи

```
predict_eta <- predict(M_pois, newdata = NewData, se.fit = TRUE)
NewData$eta <- predict_eta$fit
NewData$SE_eta <- predict_eta$se.fit
```

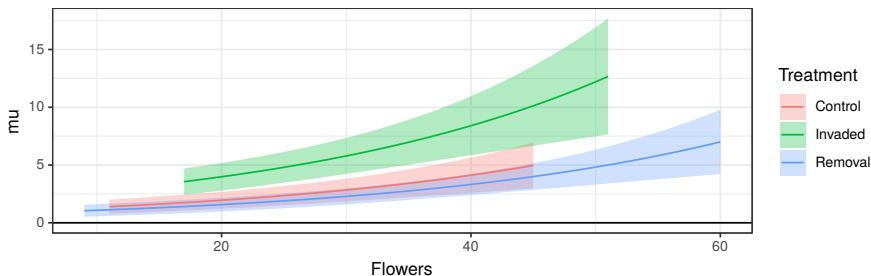
```
ggplot(NewData, aes(x = Flowers, y = eta, fill = Treatment)) +
  geom_ribbon(aes(ymin = eta - 2 * SE_eta, ymax = eta + 2 * SE_eta), alpha=0.3) +
  geom_line(aes(colour = Treatment)) + geom_hline(yintercept = 0)
```



В масштабе функции связи мы моделируем линейную зависимость логарифмов мат. ожидания отклика от предикторов.

# График предсказаний в масштабе переменной-отклика

```
predict_mu <- predict(M_pois, newdata = NewData,  
                      se.fit = TRUE, type = 'response')  
NewData$mu <- predict_mu$fit  
NewData$SE_mu <- predict_mu$se.fit  
  
ggplot(NewData, aes(x = Flowers, y = mu, fill = Treatment)) +  
  geom_ribbon(aes(ymin = mu - 2 * SE_mu, ymax = mu + 2 * SE_mu), alpha=0.3) +  
  geom_line(aes(colour = Treatment)) + geom_hline(yintercept = 0)
```

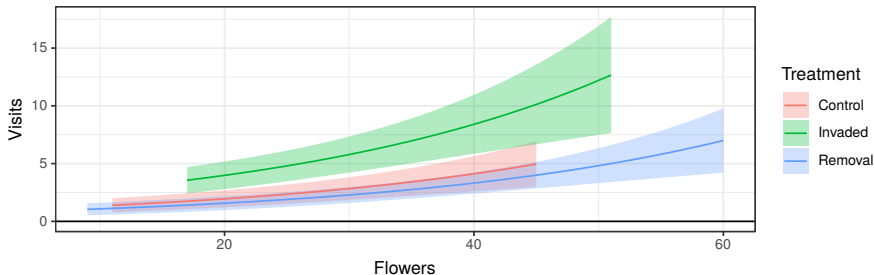


GLM с Пуассоновским распределением отклика моделирует его нелинейную связь предикторами за счет функции связи  $\log()$ .

# Возможные проблемы GLM с Пуассоновским распределением отклика

GLM с Пуассоновским распределением отклика учитывает гетерогенность дисперсии ( $var(y_i) = \mu_i = E(y_i)$ ). Стандартные ошибки возрастают с увеличением предсказанного значения.

Но достаточно ли этого для моделирования данных? Нет ли здесь сверхдисперсии?



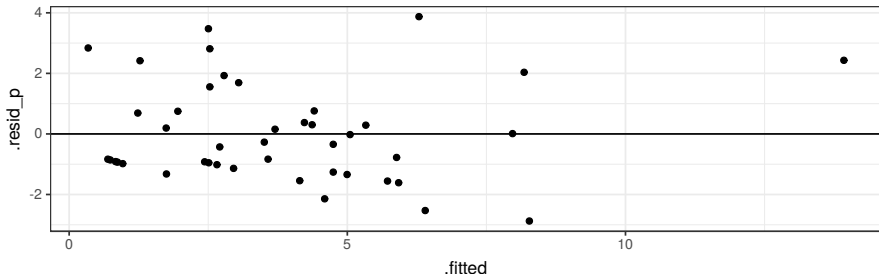


## Условия применимости GLM с Пуассоновским распределением отклика

- ▶ Случайность и независимость наблюдений внутри групп.
- ▶ Отсутствие сверхдисперсии. (Дисперсия остатков равна мат.ожиданию при каждом уровне значений предикторов).
- ▶ Отсутствие коллинеарности предикторов.

## График остатков

```
M_pois_diag <- data.frame(.fitted = predict(M_pois, type = "response"),  
                           .resid_p = residuals(M_pois, type = "pearson"))  
ggplot(M_pois_diag, aes(x = .fitted, y = .resid_p)) +  
  geom_point() +  
  geom_hline(yintercept = 0)
```



## Избыточность дисперсии (overdispersion)

Если данные подчиняются распределению Пуассона, то дисперсия должна быть равна среднему значению.

$$\begin{aligned} \blacktriangleright E(y_i) &= \mu_i \\ \blacktriangleright \text{var}(y_i) &= \mu_i \end{aligned}$$

Если это не так, то мы не сможем доверять результатам. Это будет значить, что мы применяем модель, основанную на Пуассоновском распределении, к данным, которые не подчиняются этому распределению.

# Проверка на сверхдисперсию

Используем предложенную Беном Болкером функцию проверки на сверхдисперсию

# Функция для проверки наличия сверхдисперсии в модели (автор Ben Bolker)

# <http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html>

```
overdisp_fun <- function(model) {  
  rdf <- df.residual(model) # Число степеней свободы N - p  
  rp <- residuals(model,type="pearson") # Пирсоновские остатки  
  Pearson.chisq <- sum(rp^2) # Сумма квадратов остатков  
  prat <- Pearson.chisq/rdf # Степень избыточности дисперсии  
  pval <- pchisq(Pearson.chisq, df=rdf, lower.tail=FALSE) # Уровень значимости  
  c(chisq=Pearson.chisq, ratio=prat, rdf=rdf, p=pval) # Вывод результатов  
}
```

Ben Bolker's glmmFAQ

<http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html>

```
overdisp_fun(M_pois)
```

```
#          chisq          ratio          rdf          p  
# 1.115831e+02 3.015760e+00 3.700000e+01 2.079463e-09
```

Избыточность дисперсии есть! Дисперсия в 3 раза выше среднего.

## Если есть избыточность дисперсии...

Пуассоновские модели недооценивают (приуменьшают) “раздувшиеся” стандартные ошибки.

Если данные подчиняются распределению Пуассона, то:

$$\text{var}(y_i) = \mu_i$$

$$\text{var}(E(y_i)) = \mu_i/n$$

$$SE_{E(y_i)} = \sqrt{\text{var}(E(y_i))}$$

Если данные не подчиняются распределению Пуассона, и дисперсия в  $\varphi$  раз больше среднего ( $\varphi > 1$ ), то:

$$\text{var}^*(y_i) = \varphi \mu_i$$

Тогда дисперсии и стандартные ошибки “раздуты”:

$$\text{var}(E(y_i)) = \varphi \mu_i/n$$

$$SE_{E(y_i)} = \sqrt{\varphi \text{var}(E(y_i))}$$

# Проблемы из-за неучтенной избыточности дисперсии

Когда есть избыточность дисперсии, использование распределения Пуассона приведет к проблемам:

- ▶ Доверительная зона предсказаний модели будет заужена из-за того, что оценки стандартных ошибок занижены.
- ▶ Тесты Вальда для коэффициентов модели дадут неправильные результаты из-за того, что оценки стандартных ошибок занижены. Уровень значимости будет занижен.
- ▶ Тесты, основанные на сравнении правдоподобий дадут смещённые результаты, т.к. соотношение девианс уже не будет подчиняться  $\chi^2$ -распределению.

## Причины избыточности дисперсии

- ▶ Наличие выбросов.
- ▶ В модель не включен важный предиктор или взаимодействие предикторов.
- ▶ Нарушена независимость выборок (есть внутригрупповые корреляции).
- ▶ Нелинейная связь между ковариатами и зависимой переменной.
- ▶ Выбрана неподходящая связывающая функция.
- ▶ Количество нулей больше, чем предсказывает выбранное распределение отклика (Zero inflation) .
- ▶ Выбрана неподходящая функция распределения для отклика.

# Причины избыточности дисперсии

- ▶ Наличие выбросов.
- ▶ В модель не включен важный предиктор или взаимодействие предикторов.
- ▶ Нарушена независимость выборок (есть внутригрупповые корреляции).
- ▶ Нелинейная связь между ковариатами и зависимой переменной.
- ▶ Выбрана неподходящая связывающая функция.
- ▶ Количество нулей больше, чем предсказывает выбранное распределение отклика (Zero inflation) .
- ▶ Выбрана неподходящая функция распределения для отклика.

## Как бороться с избыточностью дисперсии

Взвесив все, что известно о данных, можно решить, как именно усовершенствовать модель.

Для модели числа визитов опылителей мы попробуем два варианта действий:

- ▶ Можно построить квази-пуассоновскую модель.
- ▶ Можно построить модель, основанную на отрицательном биномиальном распределении.



# Квази-пуассоновские модели

$$Visits_i \sim Quasipoisson(\mu_i)$$

$$E(Visits_i) = \mu_i, \text{ var}(y_i) = \varphi \mu_i$$

$\ln(\mu_i) = \eta_i$  — функция связи логарифм

$$\eta_i = b_0 + b_1 Treatment_{Invaded\ i} + b_2 Treatment_{Removal\ i} + \\ + b_3 DiversityD1_i + b_4 Flowers_i + b_5 Hours_i$$

В этих моделях используется распределение Пуассона, но вводится поправка на степень избыточности дисперсии  $\varphi$ .

Величина  $\varphi$  показывает, во сколько раз дисперсия превышает среднее.

$\varphi$  оценивается по данным.

Помните, не бывает  
“квази-пуассоновского  
распределения”!

## Особенности квази-пуассоновской GLM

- ▶ Оценки параметров  $\beta$  такие же как в Пуассоновской GLM.
- ▶ Стандартные ошибки оценок коэффициентов домножены на  $\sqrt{\varphi}$ .
- ▶ Доверительные интервалы к оценкам коэффициентов домножены на  $\sqrt{\varphi}$ .
- ▶ Логарифмы правдоподобий уменьшаются в  $\varphi$  раз.

# Особенности квази-пуассоновской GLM

- ▶ Оценки параметров  $\beta$  такие же как в Пуассоновской GLM.
- ▶ Стандартные ошибки оценок коэффициентов домножены на  $\sqrt{\varphi}$ .
- ▶ Доверительные интервалы к оценкам коэффициентов домножены на  $\sqrt{\varphi}$ .
- ▶ Логарифмы правдоподобий уменьшаются в  $\varphi$  раз.

## Особенности работы с квази-моделями

1. В тестах параметров используются  $t$ -тесты (и  $t$ -распределение) вместо  $z$ -тестов Вальда (и стандартного нормального распределения).
2. Для анализа девиансы используются  $F$ -тесты.
3. Для квази-пуассоновских моделей не определена функция максимального правдоподобия, поэтому нельзя вычислить AIC (но иногда считают квази-AIC = QAIC).

# Подбираем квази-пуассоновскую модель

$$Visits_i \sim Quasipoisson(\mu_i)$$

$$E(Visits_i) = \mu_i, \text{ var}(Visits_i) = \varphi \mu_i$$

$\ln(\mu_i) = \eta_i$  — функция связи логарифм

$$\eta_i = b_0 + b_1 Treatment_{Invaded\ i} + b_2 Treatment_{Removal\ i} + \\ + b_3 DiversityD1_i + b_4 Flowers_i + b_5 Hours_i$$

```
M_quasi <- glm(Visits ~ Treatment + DiversityD_1 + Flowers + Hours, data = pol,  
              family = "quasipoisson")
```

# Уравнение квази-пуассоновской модели

$$Visits_i \sim Quasipoisson(\mu_i)$$

$$E(Visits_i) = \mu_i, \text{ var}(Visits_i) = 3.016 \mu_i$$

$$\ln(\mu_i) = \eta_i$$

$$\eta_i = -2.66 + 0.71Treatment_{Invaded\ i} - 0.22Treatment_{Removal\ i} - \\ - 0.46DiversityD1_i + 0.04Flowers_i + 4.69Hours_i$$

```
coef(M_quasi)
```

#	(Intercept)	TreatmentInvaded	TreatmentRemoval	DiversityD_1
#	-2.66090631	0.71341797	-0.21537935	-0.45740225
#	Flowers	Hours		
#	0.03731004	4.68668983		

```
summary(M_quasi)$dispersion
```

```
# [1] 3.01578
```

# Смотрим на результаты подбора модели

```
summary(M_quasi)
```

```
#
# Call:
# glm(formula = Visits ~ Treatment + DiversityD_1 + Flowers + Hours,
#      family = "quasipoisson", data = pol)
#
# Deviance Residuals:
#      Min       1Q   Median       3Q      Max
# -4.0675  -1.3189  -0.3523   0.7068   3.2346
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  -2.66091    3.77613  -0.705  0.48543
# TreatmentInvaded  0.71342    0.37190   1.918  0.06282 .
# TreatmentRemoval -0.21538    0.38665  -0.557  0.58086
# DiversityD_1    -0.45740    0.22330  -2.048  0.04767 *
# Flowers         0.03731    0.01187   3.143  0.00328 **
# Hours          4.68669    5.04179   0.930  0.35862
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for quasipoisson family taken to be 3.01578)
#
#      Null deviance: 189.75  on 42  degrees of freedom
# Residual deviance: 119.88  on 37  degrees of freedom
# AIC: NA
#
# Number of Fisher Scoring iterations: 6
```

# Анализ девиансы для квази-пуассоновской модели

```
drop1(M_quasi, test = "F")
```

```
# Single term deletions
#
# Model:
# Visits ~ Treatment + DiversityD_1 + Flowers + Hours
#
#           Df Deviance F value    Pr(>F)
# <none>                119.88
# Treatment           2    145.29   3.9211 0.02854 *
# DiversityD_1         1    134.31   4.4533 0.04166 *
# Flowers              1    148.67   8.8848 0.00506 **
# Hours                1    123.64   1.1593 0.28859
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Число визитов опылителей на цветки гадючьего лука:

- ▶ зависит от присутствия вида вселенца и его цветов,
- ▶ зависит от разнообразия флоры на данном участке,
- ▶ зависит от числа цветов самого гадючьего лука.

Можем ли мы доверять этим результатам? Это приблизительные результаты.  
Не стоит доверять  $p$  близким к  $\alpha = 0.05$ .

# GLM с отрицательным биномиальным распределением отклика

$$Visits_i \sim NB(\mu_i, k)$$

$$E(Visits_i) = \mu_i, \text{ var}(Visits_i) = \mu_i + \frac{\mu_i^2}{k}$$

$\ln(\mu_i) = \eta_i$  – функция связи логарифм

$$\eta = b_0 + b_1 Treatment_{Invaded\ i} + b_2 Treatment_{Removal\ i} + b_3 DiversityD1_i + b_4 Flowers_i + b_5 Hours_i$$

```
library(MASS)
```

```
M_nb <- glm.nb(Visits ~ Treatment + DiversityD_1 + Flowers + Hours, data = pol,  
               link = "log")
```



# Уравнение модели с отрицательным биномиальным распределением отклика

$$Visits_i \sim NB(\mu_i, 1.936)$$

$$E(Visits_i) = \mu_i, \text{ var}(Visits_i) = \mu_i + \frac{\mu_i^2}{1.936}$$

$$\ln(\mu_i) = \eta_i$$

$$\eta_i = -1.97 + 0.57Treatment_{Invaded\ i} - 0.11Treatment_{Removal\ i} - 0.49DiversityD1_i + 0.03Flowers_i + 4.10Hours_i$$

```
coef(M_nb)
```

#	(Intercept)	TreatmentInvaded	TreatmentRemoval	DiversityD_1
#	-1.97122318	0.56873105	-0.10895602	-0.48867762
#	Flowers	Hours		
#	0.03092964	4.10245668		

```
summary(M_nb)$theta
```

```
# [1] 1.93593
```

# Смотрим на результаты подбора модели

```
summary(M_nb)
```

```
#
# Call:
# glm.nb(formula = Visits ~ Treatment + DiversityD_1 + Flowers +
#       Hours, data = pol, link = "log", init.theta = 1.935929584)
#
# Deviance Residuals:
#      Min       1Q   Median       3Q      Max
# -2.4604  -0.9716  -0.2443   0.4706   1.5442
#
# Coefficients:
#              Estimate Std. Error z value Pr(>|z|)
# (Intercept)  -1.97122    2.45300  -0.804   0.4216
# TreatmentInvaded  0.56873    0.38823   1.465   0.1429
# TreatmentRemoval -0.10896    0.37690  -0.289   0.7725
# DiversityD_1    -0.48868    0.19901  -2.456   0.0141 *
# Flowers         0.03093    0.01279   2.419   0.0156 *
# Hours          4.10246    3.29490   1.245   0.2131
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for Negative Binomial(1.9359) family taken to be 1)
#
#      Null deviance: 70.826  on 42  degrees of freedom
# Residual deviance: 48.891  on 37  degrees of freedom
# AIC: 208.85
#
# Number of Fisher Scoring iterations: 1
#
#              Theta:  1.936
#             Std. Err.:  0.723
#
# 2 x log-likelihood:  -194.849
```

Угловые коэффициенты показывают, на сколько единиц меняется значение логарифма отклика, если соответствующий предиктор изменяется на единицу.

Это значит, что сам отклик изменяется в  $e^{b_k}$  раз при изменении предиктора на единицу.

## Анализ девиансы модели с отрицательным биномиальным распределением отклика

```
drop1(M_nb, test = 'Chi')
```

```
# Single term deletions
#
# Model:
# Visits ~ Treatment + DiversityD_1 + Flowers + Hours
#           Df Deviance    AIC    LRT Pr(>Chi)
# <none>             48.891 206.85
# Treatment         2   53.389 207.35 4.4981  0.10550
# DiversityD_1       1   54.732 210.69 5.8414  0.01565 *
# Flowers            1   55.402 211.36 6.5110  0.01072 *
# Hours              1   50.384 206.34 1.4927  0.22180
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Число визитов опылителей на цветки гадючьего лука:

- ▶ не зависит от присутствия вида вселенца и его цветов,
- ▶ зависит от разнообразия флоры на данном участке,
- ▶ зависит от числа цветов самого гадючьего лука.

Можем ли мы доверять этим результатам? Это нужно еще проверить.

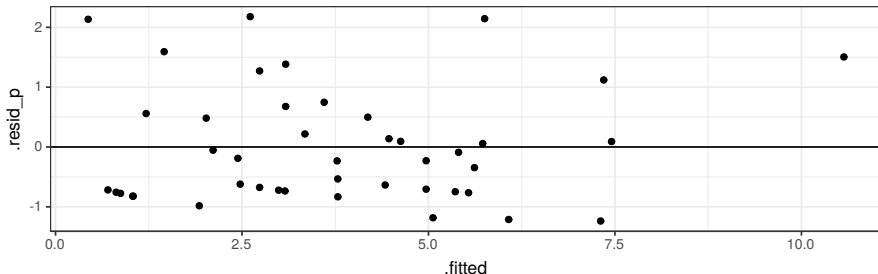
## Задание 2

Проведите диагностику модели  $M_{nb}$ .

Видите ли вы какие-нибудь нарушения условий применимости?

## График остатков

```
M_nb_diag <- data.frame(.fitted = predict(M_nb, type = "response"),  
                        .resid_p = residuals(M_nb, type = "pearson"),  
                        pol)  
gg_resid <- ggplot(M_nb_diag, aes(y = .resid_p)) + geom_hline(yintercept = 0)  
gg_resid + geom_point(aes(x = .fitted))
```



# Проверка на сверхдисперсию

Обратите внимание, у моделей с отрицательным биномиальным распределением добавляется еще один параметр

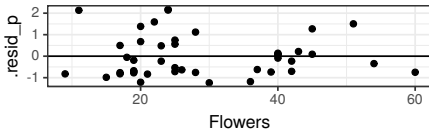
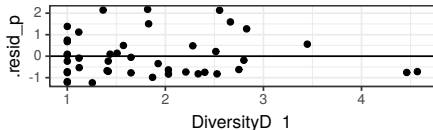
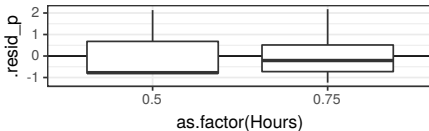
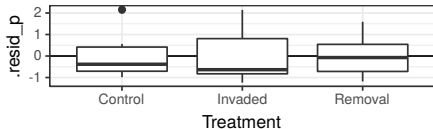
```
overdisp_fun(M_nb)
```

```
#      chisq      ratio      rdf      p
# 38.981747  1.053561 37.000000 0.380700
```

Избыточности дисперсии нет

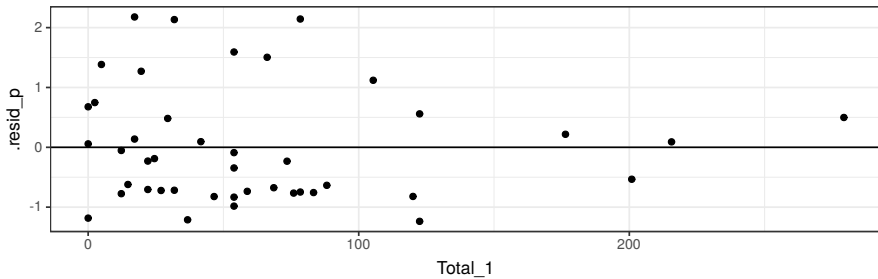
# Графики остатков от переменных, которые есть в модели

```
plot_grid(gg_resid + geom_boxplot(aes(x = Treatment)),  
          gg_resid + geom_boxplot(aes(x = as.factor(Hours))),  
          gg_resid + geom_point(aes(x = DiversityD_1)),  
          gg_resid + geom_point(aes(x = Flowers)),  
          nrow = 2)
```



## Графики остатков от переменных, которых нет в модели

```
gg_resid + geom_point(aes(x = Total_1))
```





# Данные для предсказаний

```
NewData <- pol %>% group_by(Treatment)%>%  
  do(data.frame(Flowers = seq(min(.$Flowers), max(.$Flowers), length.out=50))) %>%  
  mutate(DiversityD_1 = mean(pol$DiversityD_1),  
         Hours = mean(pol$Hours))
```

Как и в прошлый раз, мы получим предсказания двумя способами:

- ▶ при помощи операций с матрицами, чтобы своими глазами увидеть работу функции связи,
- ▶ при помощи функции `predict.glm()`.

# Предсказания модели при помощи операций с матрицами

```
# Модельная матрица и коэффициенты
X <- model.matrix(~ Treatment + DiversityD_1 + Flowers + Hours, data = NewData)
b <- coef(M_nb)

# Предсказанные значения и стандартные ошибки...
# ...в масштабе функции связи (логарифм)
NewData$fit_eta <- X %*% b
NewData$SE_eta <- sqrt(diag(X %*% vcov(M_nb) %*% t(X)))

# ...в масштабе отклика (применяем функцию, обратную функции связи)
NewData$fit_mu <- exp(NewData$fit_eta)
NewData$lwr <- exp(NewData$fit_eta - 2 * NewData$SE_eta)
NewData$upr <- exp(NewData$fit_eta + 2 * NewData$SE_eta)

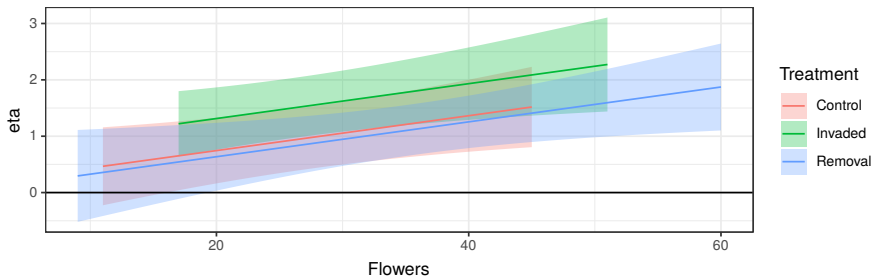
head(NewData, 2)
```

```
# # A tibble: 2 x 9
# # Groups:   Treatment [1]
#   Treatment Flowers DiversityD_1 Hours fit_eta SE_eta fit_mu lwr
#   <chr>      <dbl>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
# 1 Control      11          1.86 0.733  0.467  0.346  1.59 0.798
# 2 Control     11.7          1.86 0.733  0.488  0.341  1.63 0.824
# # ... with 1 more variable: upr <dbl>
```

## График предсказаний в масштабе функции связи

```
predict_eta <- predict(M_nb, newdata = NewData, se.fit = TRUE)
NewData$eta <- predict_eta$fit
NewData$SE_eta <- predict_eta$se.fit
```

```
ggplot(NewData, aes(x = Flowers, y = eta, fill = Treatment)) +  
  geom_ribbon(aes(ymin = eta - 2 * SE_eta, ymax = eta + 2 * SE_eta), alpha = 0.3) +  
  geom_line(aes(colour = Treatment)) + geom_hline(yintercept = 0)
```

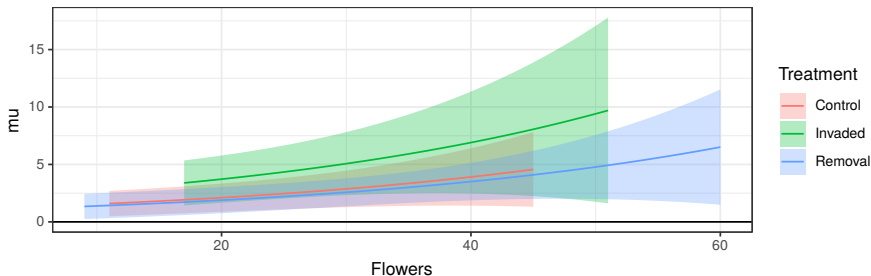


В масштабе функции связи мы моделируем линейную зависимость логарифмов мат. ожидания отклика от предикторов.

## График предсказаний в масштабе переменной-отклика

```
predict_mu <- predict(M_nb, newdata = NewData, se.fit = TRUE, type = 'response')
NewData$mu <- predict_mu$fit
NewData$SE_mu <- predict_mu$se.fit
```

```
ggplot(NewData, aes(x = Flowers, y = mu, fill = Treatment)) +  
  geom_ribbon(aes(ymin = mu - 2 * SE_mu, ymax = mu + 2 * SE_mu), alpha = 0.3) +  
  geom_line(aes(colour = Treatment)) + geom_hline(yintercept = 0)
```

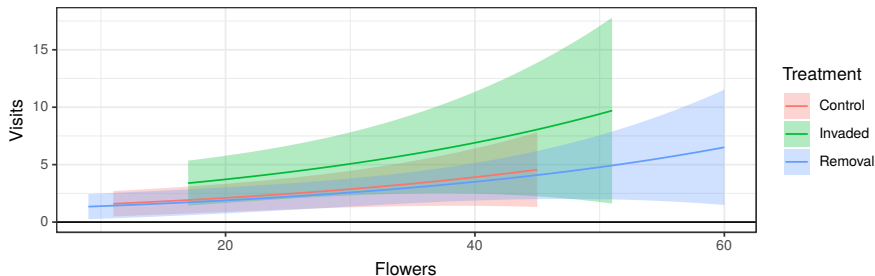


GLM с отрицательным биномиальным распределением отклика моделирует его нелинейную связь предикторами за счет функции связи  $\log()$ .

# GLM с отрицательным биномиальным распределением отклика

GLM с отрицательным биномиальным распределением отклика учитывает гетерогенность дисперсии ( $E(y_i) = \mu_i$ ,  $var(y_i) = \mu_i + \frac{\mu_i^2}{k}$ ). Стандартные ошибки возрастают с увеличением предсказанного значения даже сильнее, чем это было у Пуассоновской модели.

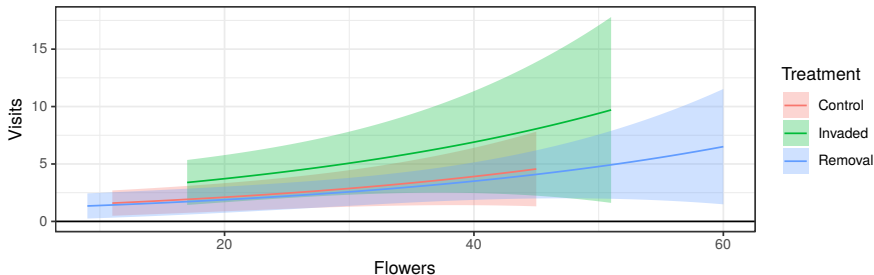
Этого оказалось вполне достаточно для моделирования данных (сверхдисперсии здесь нет).



## Выводы.

Число визитов опылителей на цветки гадючьего лука зависит не от присутствия вида вселенца или его цветов, а от разнообразия флоры на данном участке (тест отношения правдоподобий,  $p = 0.02$ ).

При этом, чем больше цветов самого гадючьего лука, тем больше прилетает опылителей (тест отношения правдоподобий,  $p = 0.01$ ).



## Take-home messages

Очень важно правильно формулировать модель для данных.

Для моделирования счетных зависимых переменных применяются модели, основанные на распределении Пуассона или отрицательном биномиальном распределении.

Одно из условий применимости этих моделей — отсутствие избыточности дисперсии.

Избыточность дисперсий может возникать в силу разных причин, поэтому единого рецепта борьбы с ней нет.

Квази-пуассоновские модели решают проблему сверхдисперсии в Пуассоновской GLM внося поправки для стандартных ошибок оценок коэффициентов модели.

Модели, основанные на отрицательном биномиальном распределении, учитывают избыточность дисперсии при помощи отдельного параметра.

## Что почитать

- ▶ Zuur, A.F. and Ieno, E.N., 2016. A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution*, 7(6), pp.636-645.
- ▶ Кабаков Р.И. R в действии. Анализ и визуализация данных на языке R. М.: ДМК Пресс, 2014
- ▶ Zuur, A., Ieno, E.N. and Smith, G.M., 2007. *Analyzing ecological data*. Springer Science & Business Media.