Обобщенные линейные модели с нормальным распределением остатков

Линейные модели...

Марина Варфоломеева, Вадим Хайтов

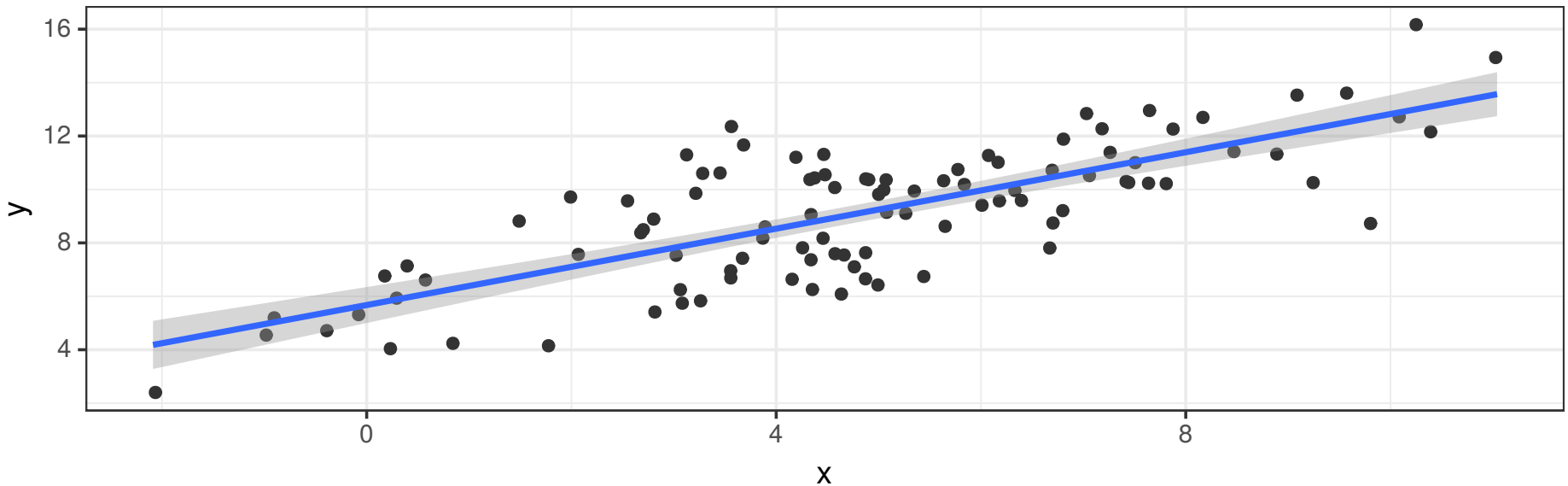
Кафедра Зоологии беспозвоночных, Биологический факультет, СПбГУ

Общая линейная модель — удобный инструмент описания связей

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i} + \varepsilon_i,$$

где $\varepsilon_i \sim N(0, \sigma)$.

Предикторы в такой модели могут быть дискретными и непрерывными.



Применимость общих линейных моделей ограничена

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i} + \varepsilon_i,$$

где $\varepsilon_i \sim N(0, \sigma)$.

Т.е. на самом деле мы имеем ввиду, что переменная-отклик подчиняется нормальному распределению:

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

$$E(y_i) = \mu_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$$

Если отклик подчиняется другому распределению, такие модели не годятся.

Последний из двух вариантов записи модели (распределение отклика и линейный предиктор) мы будем использовать дальше.

Обобщенные линейные модели

Обобщенные линейные модели (Generalized Linear Models, GLM, GLZ, GLIM) позволяют моделировать зависимости не только для нормально-распределенных величин, но и для других распределений.

Не путайте обобщенные линейные модели с общими (General Linear Models, тоже сокращаются как GLM).

Важнейшие распределения из семейства экспоненциальных

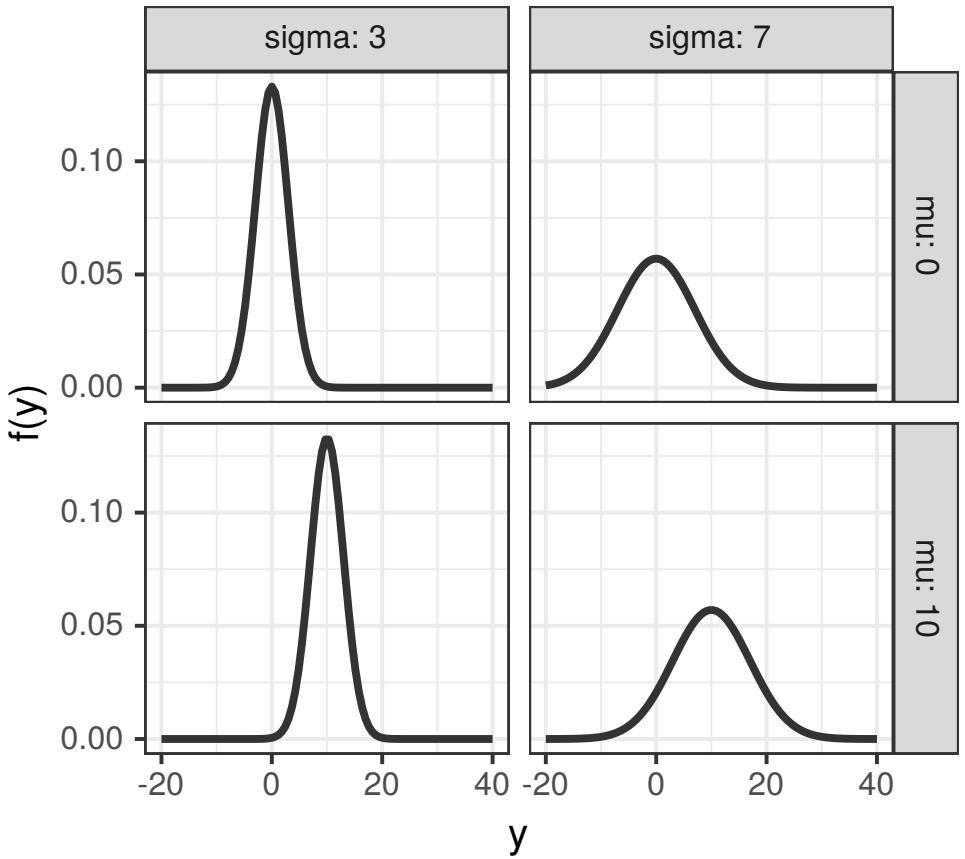
Для непрерывных величин

- Нормальное распределение
- Гамма распределение

Для дискретных величин

- Биномиальное распределение
- Распределение Пуассона
- Отрицательное биномиальное распределение

Нормальное распределение



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

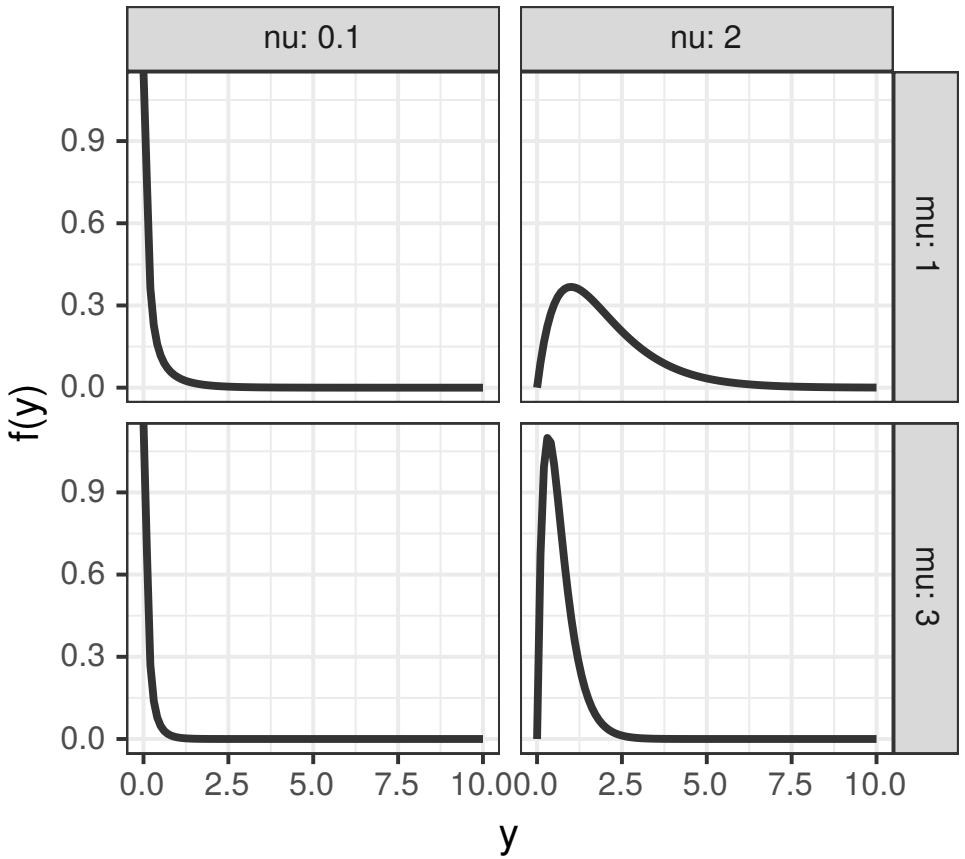
Параметры:

- μ – среднее
- σ – стандартное отклонение

Свойства:

- $E(y) = \mu$ – мат.ожидание
- $var(y) = \sigma^2$ – дисперсия
- $-\infty \leq y \leq +\infty$, $y \in \mathbb{R}$ – значения

Гамма-распределение



$$f(y) = \frac{1}{\Gamma(\nu)} \cdot (\frac{\nu}{\mu})^{\nu} \cdot y^{\nu-1} \cdot e^{\frac{y \cdot \nu}{\mu}}$$

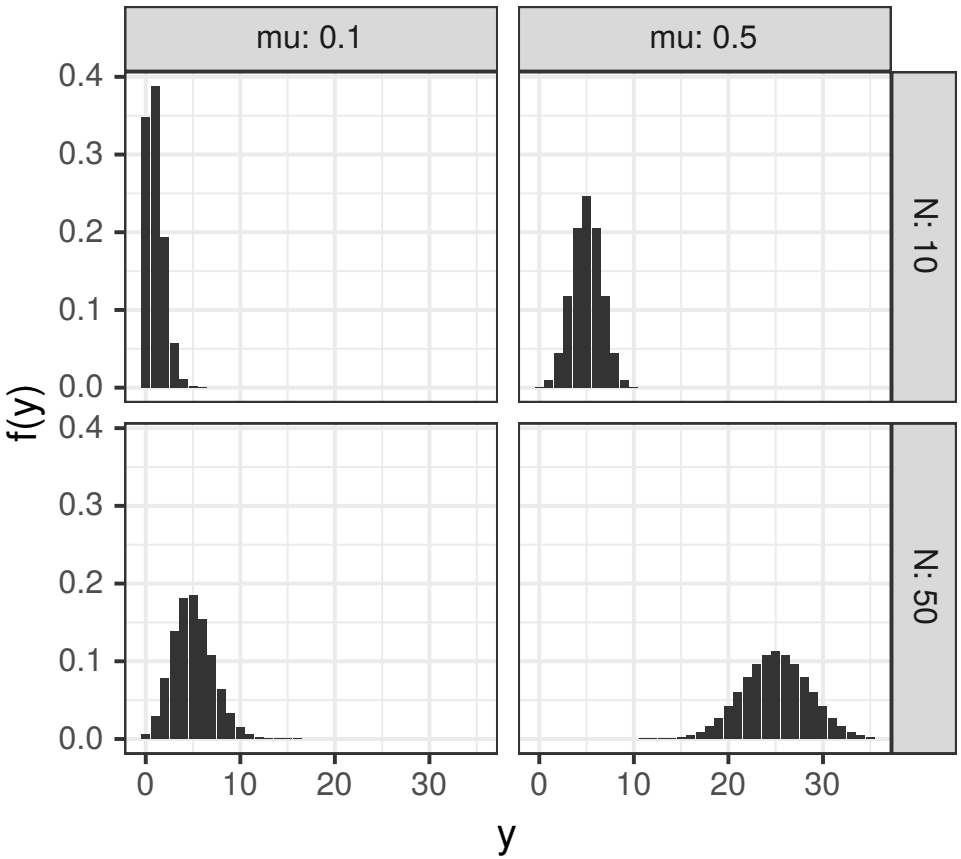
Параметры:

- μ – среднее
- ν – определяет степень избыточности дисперсии

Свойства:

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

Биномиальное распределение



$$f(y) = \frac{n!}{y! \cdot (n-y)!} \cdot \pi^y \cdot (1-\pi)^{n-y}$$

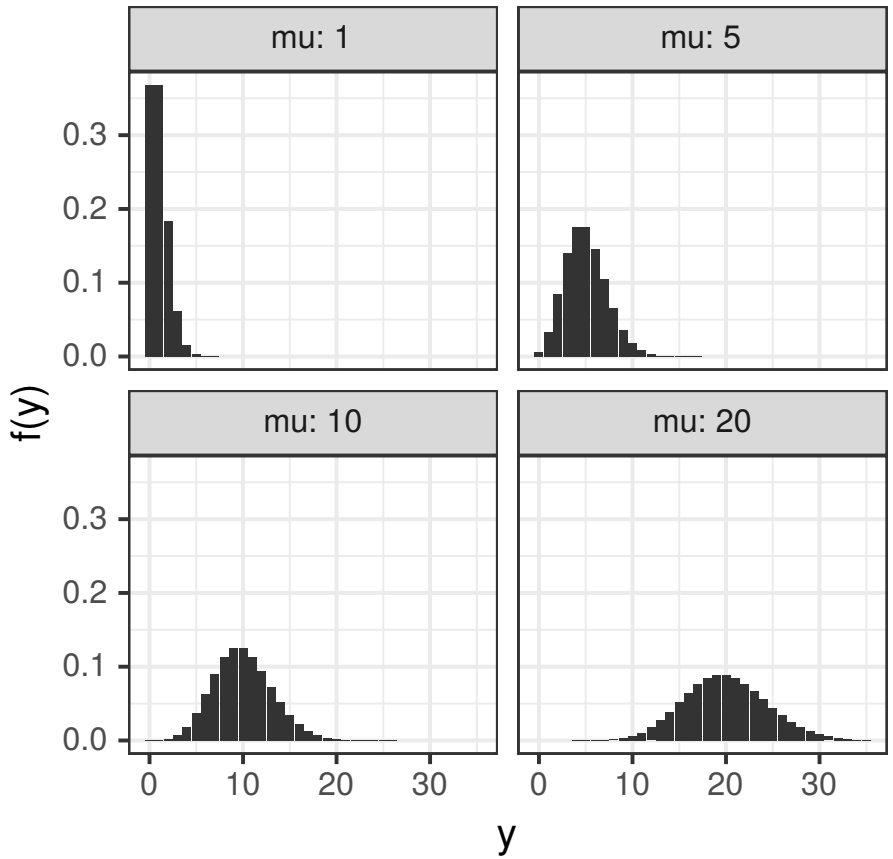
Параметры:

- n – число объектов в испытании
- π – вероятность того, что $y = 1$

Свойства:

- $E(y) = n \cdot \pi$ – мат.ожидание
- $var(y) = n \cdot \pi \cdot (1-\pi)$ – дисперсия
- $0 \leq y \leq +\infty$, $y \in \mathbb{N}$ – значения

Распределение Пуассона



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

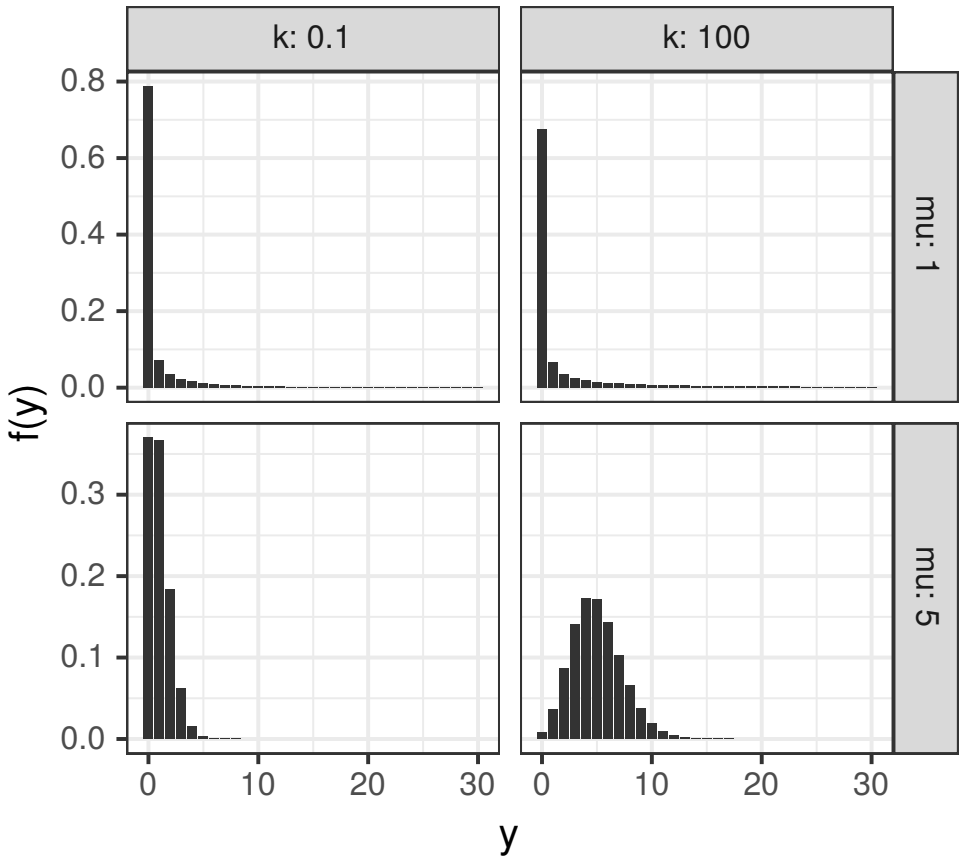
Параметр:

- μ – 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



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

Параметры:

- μ – 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

Обобщенные линейные модели (Generalized Linear Models)

Обобщенные линейные модели (Nelder, Weddlerburn, 1972)

- Позволяют моделировать не только величины, подчиняющиеся нормальному распределению, но и другим распределениям из семейства экспоненциальных.
- Подбор коэффициентов делается методом максимального правдоподобия.

Чем отличается обобщенная линейная модель от общей?

На примере нормального распределения $f(y_i|\theta) = N(\mu_i, \sigma)$

Общая линейная модель состоит из двух компонентов

$y_i \sim f(y_i|\theta)$
$E(y_i) = \mu_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Чем отличается обобщенная линейная модель от общей?

На примере нормального распределения $f(y_i|\theta) = N(\mu_i, \sigma)$

Общая линейная модель состоит из двух компонентов

$y_i \sim f(y_i|\theta)$
$E(y_i) = \mu_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Обобщенная линейная модель состоит из трех компонентов

$y_i \sim f(y_i|\theta)$
$E(y_i) = \mu_i = g^{-1}(\beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i})$
$g(\mu_i) = \eta_i$
$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

- Случайная часть — $y_i \sim f(y_i|\theta)$ — распределение из семейства экспоненциальных с параметрами θ.
- Фиксированная часть.

Компонент, который появляется в GLM:

- $g()$ — функция связи, которая трансформирует мат. ожидание отклика в линейный предиктор (обратная функция обозначается $g^{-1}()$).

Функция связи

$y_i \sim f(y_i|\theta)$

$E(y_i) = \mu_i = g^{-1}(\beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i})$

$g(\mu_i) = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Функция связи $g(\mu)$ используется разная в зависимости от распределения $f(y_i|\theta)$.

Функция связи

$y_i \sim f(y_i|\theta)$

$E(y_i) = \mu_i = g^{-1}(\beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i})$

$g(\mu_i) = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Функция связи $g(\mu)$ используется разная в зависимости от распределения $f(y_i|\theta)$.

Например:

Идентичность:

$E(y_i) = \mu_i$

$\mu_i = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Функция связи

$y_i \sim f(y_i|\theta)$

$E(y_i) = \mu_i = g^{-1}(\beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i})$

$g(\mu_i) = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Функция связи $g(\mu)$ используется разная в зависимости от распределения $f(y_i|\theta)$.

Например:

Идентичность:

$E(y_i) = \mu_i$

$\mu_i = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Логарифм:

$E(y_i) = e^{\beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}}$

$ln(\mu_i) = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Функция связи

$y_i \sim f(y_i|\theta)$

$E(y_i) = \mu_i = g^{-1}(\beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i})$

$g(\mu_i) = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Функция связи $g(\mu)$ используется разная в зависимости от распределения $f(y_i|\theta)$.

Например:

Идентичность:

$E(y_i) = \mu_i$

$\mu_i = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Логарифм:

$E(y_i) = e^{\beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}}$

$ln(\mu_i) = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Логит:

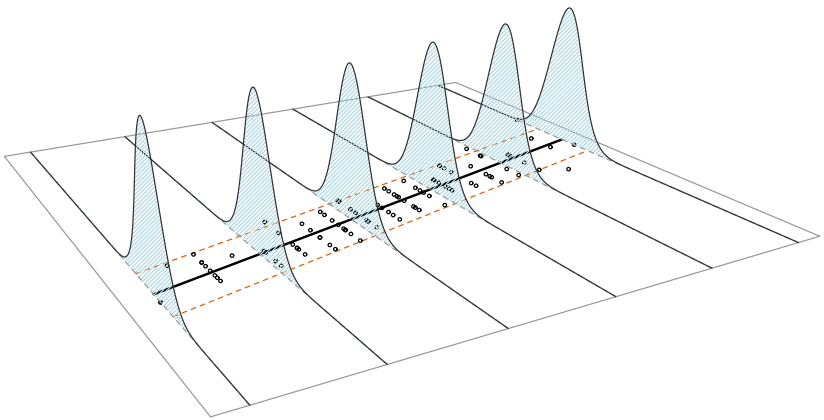
$E(y_i) = \frac{e^{\beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}}}$

$ln(\frac{\mu_i}{1-\mu_i}) = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Параметры обобщенных линейных моделей подбирают методом максимального правдоподобия

Правдоподобие (likelihood) — измеряет соответствие реально наблюдаемых данных тем, что можно получить из модели при определенных значениях параметров.



Это произведение вероятностей получения каждой из точек данных:

$$L(\theta|\text{data}) = \Pi_{i=1}^{n} f(\text{data}|\theta)$$

- $f(\text{data}|\theta)$ — функция распределения с параметрами θ

Нужно найти параметры θ, которые максимизирут правдоподобие:

$$L(\theta|\text{data}) \longrightarrow \max$$

Вычислительно проще работать с логарифмами правдоподобий (loglikelihood):

$$lnL(\theta|\text{data}) \longrightarrow \max$$

Редко можно решить аналитически, обычно используются численные решения.

Пример – энергетическая ценность икры

Один из показателей, связанных с жизнеспособностью икры – доля сухого вещества. Она пропорциональна количеству запасенной энергии.

Отличается ли энергетическая ценность икры большой озерной форели в сентябре и ноябре?

Данные: Lantry et al., 2008
Источник: пакет Stat2Data

Открываем данные

```
library(Stat2Data)
data(FishEggs)
```

Все ли правильно открылось?

```
str(FishEggs)
```

```
# 'data.frame': 35 obs. of  4 variables:
#  $ Age  : int  7 8 8 9 9 9 9 10 10 11 ...
#  $ PctDM: num  37.4 38 37.5 39 37.9 ...
#  $ Month: Factor w/ 2 levels "Nov","Sep": 1 1 1 1 1 1 1 1 1 1 ...
#  $ Sept : int  0 0 0 0 0 0 0 0 0 0 ...
```

Нет ли пропущенных значений?

```
colSums(is.na(FishEggs))
```

```
#   Age PctDM Month  Sept
#     0     0     0     0
```

Меняем порядок уровней факторов

Уровни факторов в исходных данных:

```
levels(FishEggs$Month)
```

```
# [1] ”Nov” ”Sep”
```

Делаем базовым уровнем сентябрь.

```
FishEggs$Month <- relevel(FishEggs$Month, ref = 'Sep')
```

Теперь уровни в хронологическом порядке:

```
levels(FishEggs$Month)
```

```
# [1] ”Sep” ”Nov”
```

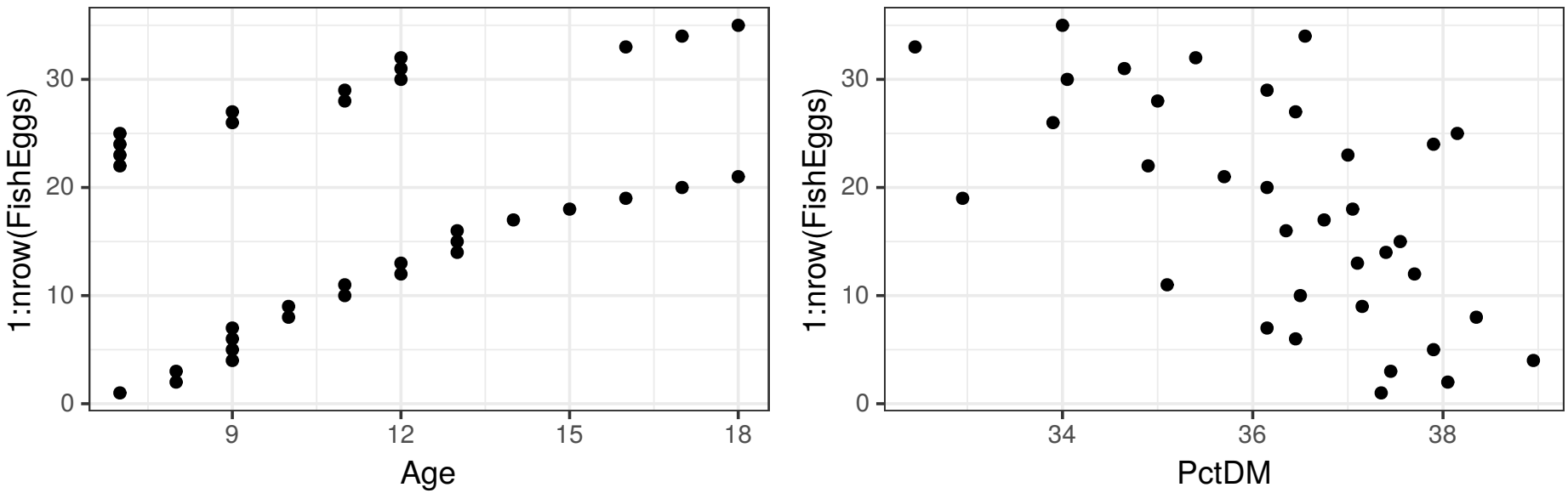
Объемы выборок

```
table(FishEggs$Month)
```

```
#
# Sep Nov
#  14  21
```

Ищем выбросы

```
library(cowplot)
library(ggplot2)
theme_set(theme_bw())
gg_dot <- ggplot(FishEggs, aes(y = 1:nrow(FishEggs))) + geom_point()
plot_grid(gg_dot + aes(x = Age),
          gg_dot + aes(x = PctDM), nrow = 1)
```



Модель для описания питательной ценности икры

GLM с нормальным распределением отклика

$y_i \sim N(\mu_i, \sigma)$
$E(y_i) = \mu_i$
$\mu_i = \eta_i$ — функция связи “идентичность” (identity)
$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$

Модель зависимости в примере

Как зависит питательная ценность икры от месяца вылова рыбы (сентябрь или ноябрь) с учетом ковариаты (возраста рыбы) и их взаимодействия.

$PctDM_i \sim N(\mu_i, \sigma)$
$E(PctDM_i) = \mu_i$
$\mu_i = \eta_i$
$\eta_i = \beta_0 + \beta_1 Age_i + \beta_2 Month_{Nov\,i} + \beta_3 Age_i Month_{Nov\,i}$

- $PctDM_i$ – содержание сухого вещества в икре
- Age_i – возраст рыбы
- $Month_{Nov\,i}$ – переменная-индикатор для уровня $Month_{Nov\,i} = 1$

Подбираем модель

```
mod <- glm(PctDM ~ Age * Month, data = FishEggs)
mod
```

```
#
# Call:  glm(formula = PctDM ~ Age * Month, data = FishEggs)
#
# Coefficients:
#  (Intercept)           Age      MonthNov  Age:MonthNov
#     38.12111      -0.23965       1.27623       0.02144
#
# Degrees of Freedom: 34 Total (i.e. Null);  31 Residual
# Null Deviance:        83.96
# Residual Deviance: 47.83  AIC: 120.3
```

```
# Чтобы записать модель, нужна еще сигма.
sigma(mod)
```

```
# [1] 1.242124
```

$$PctDM_i \sim N(\mu_i, 1.24)$$

$$E(PctDM_i) = \mu_i$$

$$\mu_i = \eta_i$$

$$\eta_i = 38.12 - 0.24 Age_i + 1.28 Month_{Nov\,i} + 0.02 Age_i Month_{Nov\,i}$$

Диагностика модели

Разновидности остатков в GLM

Остатки в масштабе отклика (response residuals)

$$e_i = y_i - E(y_i)$$

- Это разница между наблюдаемым и предсказанным значениями.
  - Аналог “сырых” остатков в простой линейной регрессии.

```
resid(mod, type = 'response')[1:5]
```

```
#           1           2           3           4           5
# -0.5198421  0.3983712 -0.2016288  1.5165844  0.4665844
```

Разновидности остатков в GLM

Остатки в масштабе отклика (response residuals)

$$e_i = y_i - E(y_i)$$

- Это разница между наблюдаемым и предсказанным значениями.
  - Аналог “сырых” остатков в простой линейной регрессии.

```
resid(mod, type = 'response')[1:5]
```

```
#           1           2           3           4           5
# -0.5198421  0.3983712 -0.2016288  1.5165844  0.4665844
```

Пирсоновские остатки (Pearson’s residuals)

$$r_{p\,i} = \frac{y_i - E(y_i)}{\sqrt{var(y_i)}}$$

- Это обычные остатки, деленные на стандартную ошибку предсказанного значения.
- Аналог стандартизованных остатков в простой линейной регрессии.

```
resid(mod, type = 'pearson')[1:5]
```

```
#           1           2           3           4           5
# -0.5198421  0.3983712 -0.2016288  1.5165844  0.4665844
```

Для GLM с нормальным распределением отклика оба типа остатков одинаковы.

Условия применимости GLM с нормальным распределением отклика

- Случайность и независимость наблюдений внутри групп.
- Нормальное распределение остатков.
- Гомогенность дисперсий остатков.
- Отсутствие коллинеарности предикторов.

Проверка на коллинеарность

Мы и так знаем, что в параметризации индикаторных переменных взаимодействие будет коллинеарно со своими составляющими (и это нормально).

Важно проверить, будут ли коллинеарны другие предикторы.

```
library(car)
vif(update(mod, . ~ . - Age:Month))
```

```
#      Age    Month
# 1.006666 1.006666
```

Коллинеарности нет.

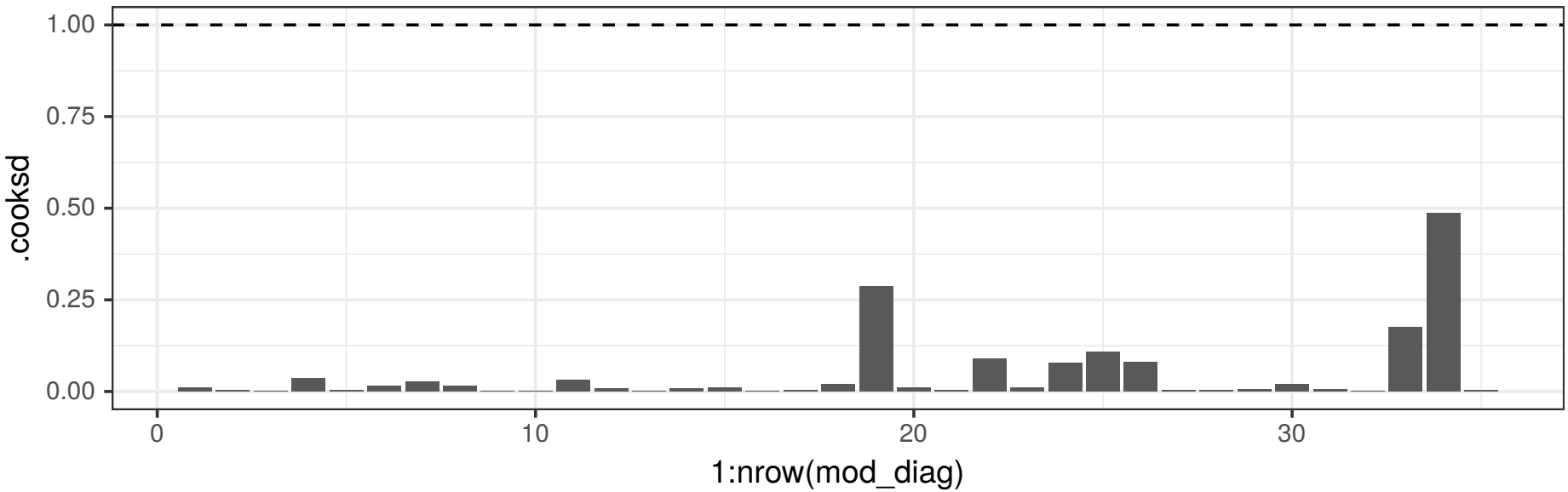
Данные для анализа остатков

```
mod_diag <- fortify(mod) # функция из пакета ggplot2
head(mod_diag)
```

```
#   PctDM Age Month       .hat   .sigma      .cooksd  .fitted     .resid
# 1 37.35   7   Nov 0.15819348 1.258412 0.0097750037 37.86984 -0.5198421
# 2 38.05   8   Nov 0.11549852 1.260286 0.0037963407 37.65163  0.3983712
# 3 37.45   8   Nov 0.11549852 1.262050 0.0009725112 37.65163 -0.2016288
# 4 38.95   9   Nov 0.08316881 1.229097 0.0368743995 37.43342  1.5165844
# 5 37.90   9   Nov 0.08316881 1.259518 0.0034902153 37.43342  0.4665844
# 6 36.45   9   Nov 0.08316881 1.248655 0.0155047929 37.43342 -0.9834156
#    .stdresid
# 1 -0.4561423
# 2  0.3410151
# 3 -0.1725990
# 4  1.2751372
# 5  0.3923020
# 6 -0.8268513
```

График расстояния Кука

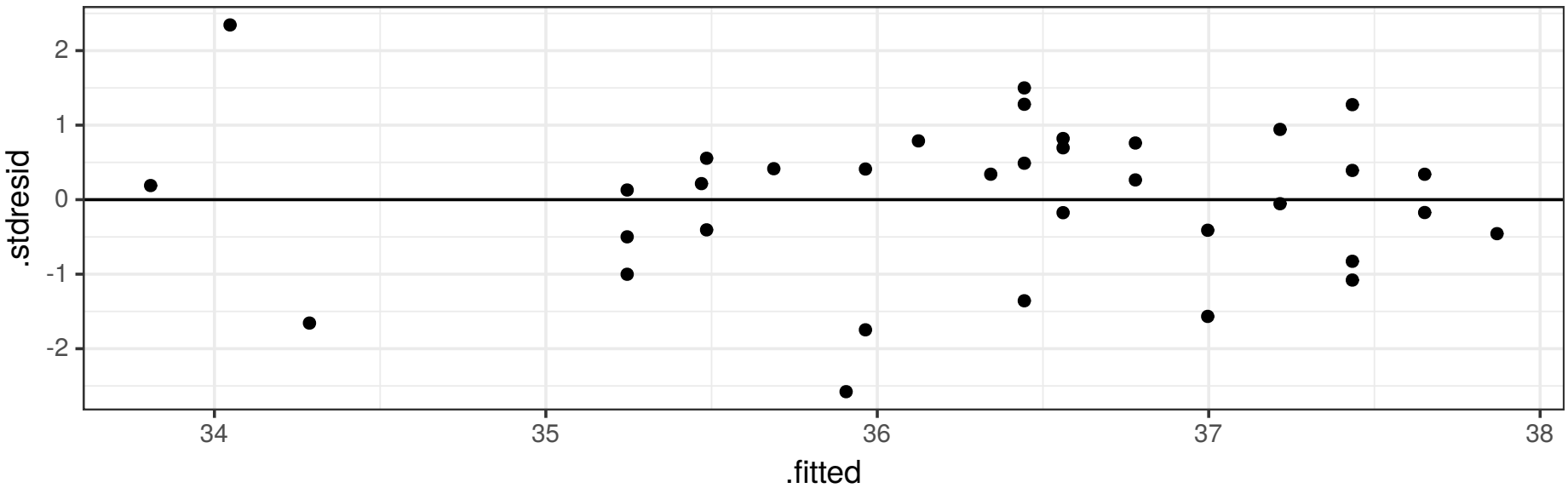
```
ggplot(mod_diag, aes(x = 1:nrow(mod_diag), y = .cooksd)) +
  geom_bar(stat = 'identity') +
  geom_hline(yintercept = 1, linetype = 2)
```



Влиятельных наблюдений нет.

График остатков от предсказанных значений

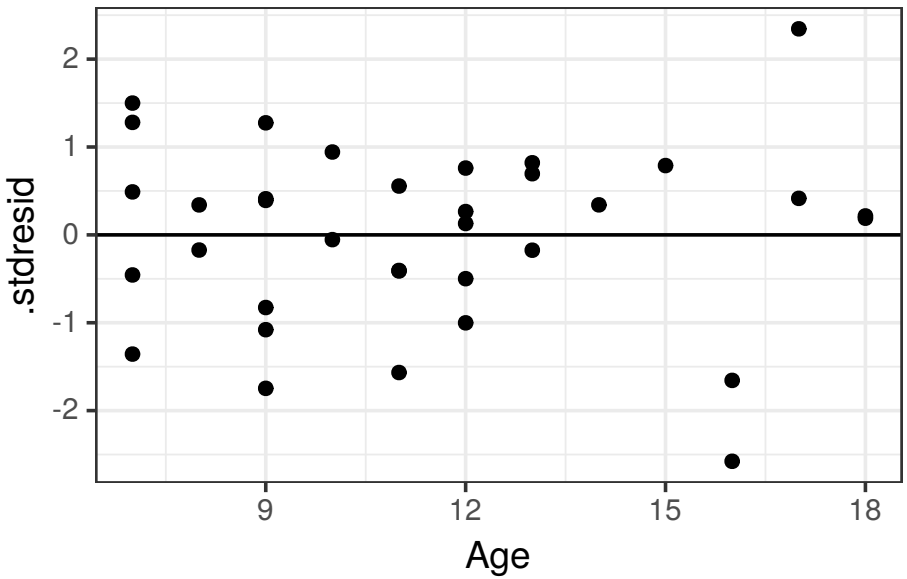
```
gg_resid <- ggplot(data = mod_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point() + geom_hline(yintercept = 0)
gg_resid
```

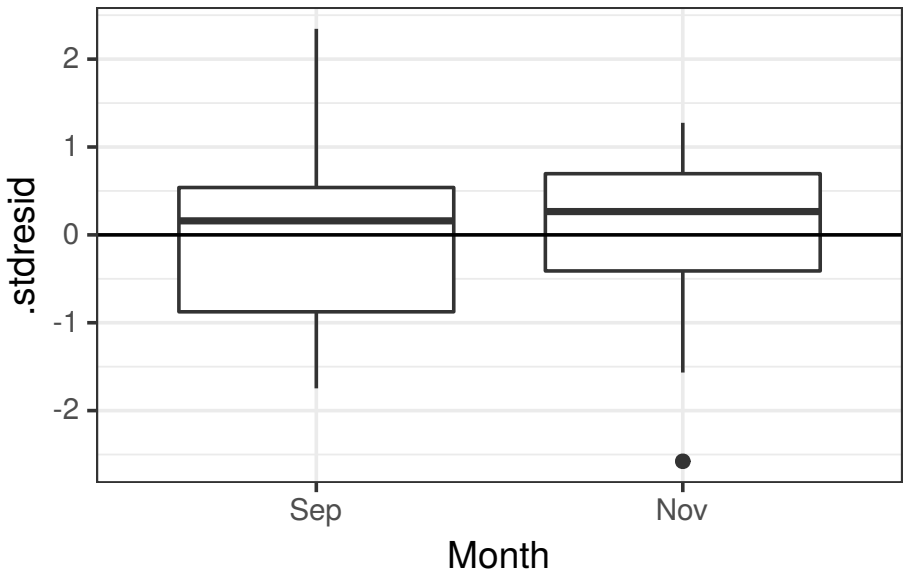


Влиятельных наблюдений нет

График зависимости остатков от предикторов в модели

```
ggplot(data = mod_diag, aes(x = Age, y = .stdresid)) +
  geom_point() + geom_hline(yintercept = 0)
ggplot(data = mod_diag, aes(x = Month, y = .stdresid)) +
  geom_boxplot() + geom_hline(yintercept = 0)
```





Гетерогенности дисперсий нет.

Тестирование значимости коэффициентов

Тест Вальда для коэффициентов GLM

$H_0 : \beta_k = 0$

$H_A : \beta_k \neq 0$

$$\frac{b_k - \beta_k}{SE_{b_k}} = \frac{b_k}{SE_{b_k}} \sim N(0,1)$$

- b_k — оценка коэффициента GLM.

Хорошо работает только на больших выборках.

Если приходится оценивать σ, то $\frac{b_k}{SE_{b_k}} \sim t_{(df=n-p)}$

- n — объем выборки.
- p — число параметров модели.

В summary() записаны результаты теста Вальда

```
summary(mod)
```

```
#
# Call:
# glm(formula = PctDM ~ Age * Month, data = FishEggs)
#
# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -2.9559  -0.5576   0.2305   0.7522   2.5029
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  38.12111    1.06436  35.816   <2e-16 ***
# Age          -0.23965    0.09134  -2.624   0.0134 *
# MonthNov      1.27623    1.51190   0.844   0.4051
# Age:MonthNov  0.02144    0.12782   0.168   0.8679
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for gaussian family taken to be 1.542872)
#
#     Null deviance: 83.962  on 34  degrees of freedom
# Residual deviance: 47.829  on 31  degrees of freedom
# AIC: 120.26
#
# Number of Fisher Scoring iterations: 2
```

В summary() записаны результаты теста Вальда

```
summary(mod)
```

```
#
# Call:
# glm(formula = PctDM ~ Age * Month, data = FishEggs)
#
# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -2.9559  -0.5576   0.2305   0.7522   2.5029
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  38.12111    1.06436  35.816   <2e-16 ***
# Age          -0.23965    0.09134  -2.624   0.0134 *
# MonthNov      1.27623    1.51190   0.844   0.4051
# Age:MonthNov  0.02144    0.12782   0.168   0.8679
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for gaussian family taken to be 1.542872)
#
#     Null deviance: 83.962  on 34  degrees of freedom
# Residual deviance: 47.829  on 31  degrees of freedom
# AIC: 120.26
#
# Number of Fisher Scoring iterations: 2
```

С увеличением возраста рыбы на год процент сухого вещества в икре снижается на 0.24 % (тест Вальда, $t_{df=31} = -2.62$, p = 0.013). Это происходит одинаково в сентябре и ноябре. Энергетическая ценность икры в сентябре и ноябре не различается.

Анализ девианcы

Насыщенная и нулевая модели задают шкалу для сравнений с предложенной

Насыщенная модель – каждое уникальное наблюдение (сочетание значений предикторов) описывается одним из n параметров.

$lnL_{saturated} = 0$
$df_{saturated} = n - p_{saturated} = n - n = 0$

Предложенная модель – модель, подобранная в данном анализе
$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{p-1} x_{p-1\,i}$.

$lnL_{model} \neq 0$
$df_{model} = n - p_{model}$

Нулевая модель – все наблюдения описываются одним параметром (средним)
$\hat{y}_i = \beta_0$.

$lnL_{null} \neq 0$
$df_{null} = n - p_{null} = n - 1$

Девианса

Это мера различия правдоподобий двух моделей (оценка разницы логарифмов правдоподобий).

Остаточная девианса

$$d_{residual} = 2(lnL_{saturated} - lnL_{model}) = -2lnL_{model}$$

Нулевая девианса

$$d_{null} = 2(lnL_{saturated} - lnL_{null}) = -2lnL_{null}$$

Сравнение нулевой и остаточной девианс позволяет судить о статистической значимости модели в целом (при помощи теста отношения правдоподобий).

$$d_{null} - d_{residual} = -2(lnL_{null} - lnL_{model}) = 2(lnL_{model} - lnL_{null})$$

Тест отношения правдоподобий (Likelihood Ratio Test)

Используется для сравнения правдоподобий

$$LRT = 2ln\left(\frac{L_{M_1}}{L_{M_2}}\right) = 2(lnL_{M_1} - lnL_{M_2})$$

- M_1 и M_2 — вложенные модели (M_1 — более полная, M_2 — уменьшенная),
- L_{M_1}, L_{M_2} - правдоподобия,
- lnL_{M_1}, lnL_{M_2} - логарифмы правдоподобий.

Сравниваются **вложенные модели, подобранные методом максимального проавдоподобия**.

Разница логарифмов правдоподобий имеет распределение, которое можно апроксимировать распределением χ^2 с числом степеней свободы $df = df_{M_2} - df_{M_1}$.

LRT используется для сравнения моделей

Для тестирования значимости модели целиком:

$$LRT = 2ln\left(\frac{L_{model}}{L_{null}}\right) = 2(lnL_{model} - lnL_{null}) = d_{null} - d_{residual}$$

$$df = df_{null} - df_{model} = (n-1) - (n - p_{model}) = p_{model} - 1$$

Для тестирования значимости предикторов:

$$LRT = 2ln\left(\frac{L_{model}}{L_{reduced}}\right) = 2(lnL_{model} - lnL_{reduced})$$

$$df = df_{reduced} - df_{model} = (n - p_{reduced}) - (n - p_{model}) = p_{model} - p_{reduced}$$

Тестируем значимость модели целиком при помощи LRT

```
null_model <- glm(PctDM ~ 1, data = FishEggs)
anova(null_model, mod, test = 'Chi')
```

```
# Analysis of Deviance Table
#
# Model 1: PctDM ~ 1
# Model 2: PctDM ~ Age * Month
#   Resid. Df Resid. Dev Df Deviance   Pr(>Chi)
# 1        34     83.962
# 2        31     47.829  3   36.133 0.00003302 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Тестируем значимость предикторов при помощи LRT

Используем II тип тестов (“II тип сумм квадратов”):

1. Тестируем значимость взаимодействия

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

```
# Single term deletions
#
# Model:
# PctDM ~ Age * Month
#           Df Deviance    AIC scaled dev. Pr(>Chi)
# <none>         47.829 120.26
# Age:Month  1   47.872 118.29    0.031735   0.8586
```

2. Тестируем значимость предикторов, когда взаимодействие удалено

```
mod_no_int <- update(mod, . ~ . -Age:Month)
drop1(mod_no_int, test = 'Chi')
```

```
# Single term deletions
#
# Model:
# PctDM ~ Age + Month
#        Df Deviance    AIC scaled dev.  Pr(>Chi)
# <none>      47.872 118.29
# Age     1   67.638 128.38      12.097 0.0005051 ***
# Month   1   67.133 128.12      11.835 0.0005813 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Запись результатов LRT-тестов значимости предикторов

Содержание сухого вещества в икре зависит от возраста рыбы и месяца ($p < 0.01$, тест отношения правдоподобий, табл. 1). Характер зависимости энергетической ценности икры от возраста одинаков в сентябре и ноябре.

Анализ девианс для модели зависимости энергетической ценности икры от возраста рыбы, месяца и их взаимодействия. Тесты II типа. df — число степеней свободы, D — девианса, p — уровень значимости.

Предиктор	df	D	p
-		47.829	
Возраст:Месяц	1	47.872	0.86
Возраст	1	67.638	<0.01
Месяц	1	67.133	<0.01

Описание качества подгонки GLM

Доля объясненной девиансы

Аналог R^2, одна из характеристик качества подгонки модели.

Остаточная девианса

$$d_{residual} = 2(lnL_{saturated} - lnL_{model}) = -2lnL_{model}$$

Нулевая девианса

$$d_{null} = 2(lnL_{saturated} - lnL_{null}) = -2lnL_{null}$$

Доля объясненной девиансы

$$\frac{d_{null} - d_{residual}}{d_{null}}$$

Долю объясненной девиансы легко вычислить

```
(mod$null.deviance - mod$deviance) / mod$null.deviance
```

```
# [1] 0.4303481
```

Модель объясняет 43% девиансы.

Задание 1

Постройте график предсказаний модели.

Данные для графика предсказаний

```
library(dplyr)
New_Data <- FishEggs %>% group_by(Month) %>%
  do(data.frame(Age = seq(min(.$Age), max(.$Age), length.out = 100)))
head(New_Data)
```

```
# # A tibble: 6 x 2
# # Groups:   Month [1]
#   Month   Age
#   <fct> <dbl>
# 1 Sep    7
# 2 Sep    7.11
# 3 Sep    7.22
# 4 Sep    7.33
# 5 Sep    7.44
# 6 Sep    7.56
```

Предсказания при помощи predict()

```
Predictions <- predict(mod, newdata = New_Data, se.fit = TRUE)
New_Data$fit <- Predictions$fit  # Предсказанные значения
New_Data$se <- Predictions$se.fit # Стандартные ошибки
t_crit <- qt(0.975, df = nrow(FishEggs) - length(coef(mod))) # t для дов. инт.
New_Data$lwr <- New_Data$fit - t_crit * New_Data$se
New_Data$upr <- New_Data$fit + t_crit * New_Data$se

head(New_Data)
```

```
# # A tibble: 6 x 6
# # Groups:   Month [1]
#   Month   Age   fit    se   lwr   upr
#   <fct> <dbl> <dbl> <dbl> <dbl> <dbl>
# 1 Sep    7     36.4 0.499  35.4  37.5
# 2 Sep    7.11  36.4 0.491  35.4  37.4
# 3 Sep    7.22  36.4 0.484  35.4  37.4
# 4 Sep    7.33  36.4 0.476  35.4  37.3
# 5 Sep    7.44  36.3 0.469  35.4  37.3
# 6 Sep    7.56  36.3 0.462  35.4  37.3
```

Данные для графика вручную

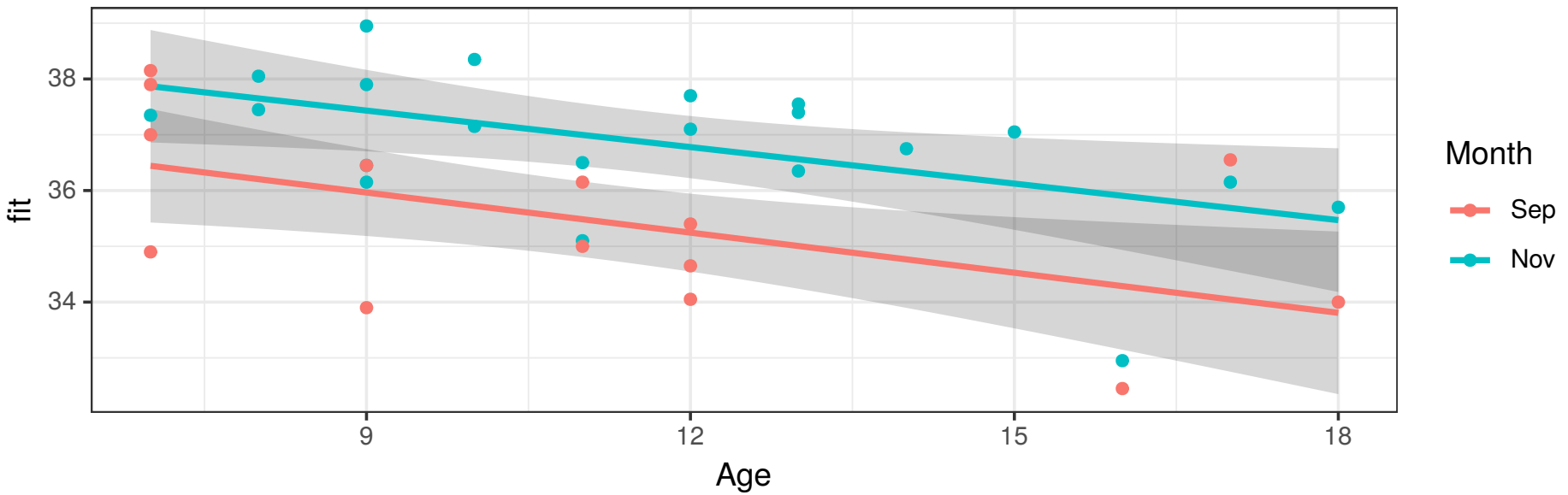
```
X <- model.matrix(~ Age * Month, data = New_Data) # Модельная матрица
betas <- coef(mod) # Коэффициенты
New_Data$fit <- X %*% betas  # Предсказанные значения
New_Data$se <- sqrt(diag(X %*% vcov(mod) %*% t(X))) # Стандартные ошибки
t_crit <- qt(0.975, df = nrow(FishEggs) - length(coef(mod))) # t для дов. инт.
New_Data$lwr <- New_Data$fit - t_crit * New_Data$se
New_Data$upr <- New_Data$fit + t_crit * New_Data$se

head(New_Data)
```

```
# # A tibble: 6 x 6
# # Groups:   Month [1]
#   Month   Age fit[,1]    se lwr[,1] upr[,1]
#   <fct> <dbl>   <dbl> <dbl>   <dbl>   <dbl>
# 1 Sep    7       36.4 0.499    35.4    37.5
# 2 Sep    7.11    36.4 0.491    35.4    37.4
# 3 Sep    7.22    36.4 0.484    35.4    37.4
# 4 Sep    7.33    36.4 0.476    35.4    37.3
# 5 Sep    7.44    36.3 0.469    35.4    37.3
# 6 Sep    7.56    36.3 0.462    35.4    37.3
```

График предсказаний модели

```
Plot_egg <- ggplot(New_Data, aes(x = Age, y = fit)) +
  geom_ribbon(alpha = 0.2, aes(ymin = lwr, ymax = upr, group = Month)) +
  geom_line(aes(colour = Month), size = 1) +
  geom_point(data = FishEggs, aes(x = Age, y = PctDM, colour = Month))
Plot_egg
```



Отбор GLM

Подбор “оптимальной” модели

В подобранной можно протестировать значимость влияния предикторов и на этом остановиться.

Или можно упростить модель, аналогично тому, как мы это делали со множественной линейной регрессией, но использовав LRT вместо F-теста.

Пытаемся сократить модель

```
mod_no_int <- update(mod, . ~ . -Age:Month)
drop1(mod_no_int, test = 'Chi')
```

```
# Single term deletions
#
# Model:
# PctDM ~ Age + Month
#        Df Deviance    AIC scaled dev.  Pr(>Chi)
# <none>      47.872 118.29
# Age     1   67.638 128.38      12.097 0.0005051 ***
# Month   1   67.133 128.12      11.835 0.0005813 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Если мы решили “сократить” модель, то теперь придется описать модель `mod_no_int`

Уравнение сокращенной модели

Как зависит питательная ценность икры от месяца вылова рыбы (сентябрь или ноябрь) с учетом ковариаты (возраста рыбы) и их взаимодействия.

```
mod_no_int
```

```
#
# Call:  glm(formula = PctDM ~ Age + Month, data = FishEggs)
#
# Coefficients:
# (Intercept)          Age     MonthNov
#     37.9999      -0.2287        1.5193
#
# Degrees of Freedom: 34 Total (i.e. Null);  32 Residual
# Null Deviance:        83.96
# Residual Deviance: 47.87  AIC: 118.3
```

```
# Чтобы записать модель, нужна сигма.
sigma(mod_no_int)
```

```
# [1] 1.223116
```

$$PctDM_i \sim N(\mu_i, 1.22)$$

$$E(PctDM_i) = \mu_i$$

$$\mu_i = \eta_i$$

$$\eta_i = 38.0 - 0.23Age_i + 1.52Month_{Nov\,i}$$

Информационные критерии — еще один способ сравнения или упрощения моделей

Информационный критерий Акаике (Akaike Information Criterion, AIC)

$$AIC = -2logLik + 2p$$

- $logLik$ - логарифм правдоподобия для модели
- $2p$ - штраф за введение в модель p параметров, т.е. за "сложность" модели

AIC — это мера потери информации, которая происходит, если реальность описывать этой моделью (Akaike 1974)

AIC — относительная мера качества модели. Т.е. не бывает какого-то "хорошего" AIC. Значения AIC можно интерпретировать только в сравнении с AIC для других моделей: чем меньше AIC — тем лучше модель.

Важно! Информационные критерии можно использовать для сравнения **даже для невложенных моделей**. Но модели должны быть **подобраны с помощью ML** и **на одинаковых данных**!

Некоторые другие информационные критерии

Критерий	Название	Формула
AIC	Информационный критерий Акаике	$AIC = -2logLik + 2p$
BIC	Баесовский информационный критерий	$BIC = -2logLik + p \cdot ln(n)$
AICc	Информационный критерий Акаике с коррекцией для малых выборок (малых относительно числа параметров: $n/p < 40$, Burnham, Anderson, 2004)	$AIC_c = -2logLik + 2p + \frac{2p(p+1)}{n-p-1}$

- $logLik$ - логарифм правдоподобия для модели
- p - число параметров
- n - число наблюдений

Как рассчитать AIC в GLM?

$$AIC = -2logLik + 2p$$

В GLM с нормальным распределением отклика число параметров — это число коэффициентов + 1, т.к. появился дополнительный параметр σ

$PctDM_i \sim N(\mu_i, 1.22)$

$E(PctDM_i) = \mu_i$

$\mu_i = \eta_i$

$\eta_i = 38.0 - 0.23Age_i + 1.52Month_{Nov\,i}$

```
(p <- length(coef(mod_no_int)) + 1) # число параметров
```

```
# [1] 4
```

```
logLik(mod_no_int)          # Логарифм правдоподобия
```

```
# 'log Lik.' -55.1437 (df=4)
```

```
as.numeric(-2 * logLik(mod_no_int) + 2 * p)
```

```
# [1] 118.2874
```

```
# Есть готовая функция
AIC(mod_no_int)
```

```
# [1] 118.2874
```

AIC удобно использовать для сравнения моделей, даже невложенных

Пусть у нас есть несколько моделей:

```
mod <-         glm(formula = PctDM ~ Age * Month, data = FishEggs)
mod_no_int <- glm(formula = PctDM ~ Age + Month, data = FishEggs)
mod_age <-    glm(formula = PctDM ~ Age,         data = FishEggs)
mod_month <-  glm(formula = PctDM ~ Month,       data = FishEggs)
```

```
AIC(mod, mod_no_int, mod_age, mod_month)
```

```
#             df      AIC
# mod          5 120.2557
# mod_no_int   4 118.2874
# mod_age      3 128.1223
# mod_month    3 128.3842
```

Судя по AIC, лучшая модель `mod_no_int`. Если значения AIC различаются всего на 1-2 единицу — такими различиями можно пренебречь и выбрать более простую модель (`mod_no_int`).

Take-home messages

- Обобщенные линейные модели (Generalized Linear Models, GLM) позволяют моделировать зависимости для откликов с распределением из семейства экспоненциальных
- Для тестирования предикторов в GLM используются:
  - Тесты Вальда для коэффициентов (плохо на малых выборках).
  - Тесты отношения правдоподобий вложенных моделей (более точно).
- Доля объясненной девиансы оценивает качество подгонки GLM
- Сравнивая модели можно отбраковать переменные, включение которых в модель не улучшает ее
- **Метод сравнения моделей нужно выбрать заранее, еще до анализа**

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.
- Quinn G.P., Keough M.J. 2002. Experimental design and data analysis for biologists
- Logan M. 2010. Biostatistical Design and Analysis Using R. A Practical Guide