Линейные модели

- Описывают линейную связь независимых переменных со средней величиной зависящей переменной
- Определяют если включение дополнительных независимых переменных позволяет учитывать большую часть вариации чем простая средняя
- Позволяют сделать выводы о популяциях (различаются ли средние зависимости между переменными) на основании выборки
- Используются для описания, предсказания или контроля

Однофакторный ANOVA (1 фактор - качественный)

$$Y_{ij} = \beta_0 + \beta_i + \varepsilon_{ij}$$

Двухфакторный ANOVA (2 factors)

$$Y_{ij} = \beta_0 + \beta_{ti} + \beta_{di} + \varepsilon_{ij}$$

Простая регрессионная модель (одна численная переменная)

$$Y_{ij} = \beta_0 + \beta_1 X + \varepsilon_{ij}$$

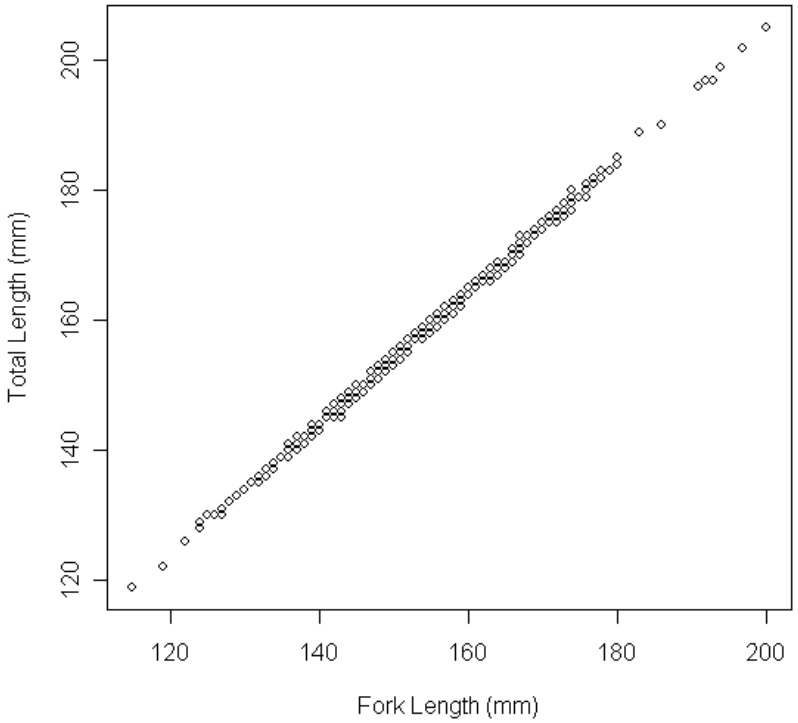
Модель множественной регрессии (2 или больше численной переменной)

$$Y_{ij} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_{ij}$$

Анализ ковариации (численные переменные и факторы)

$$Y_{ij} = \beta_0 + \beta_1 t + \beta_2 X + \varepsilon_{ij}$$

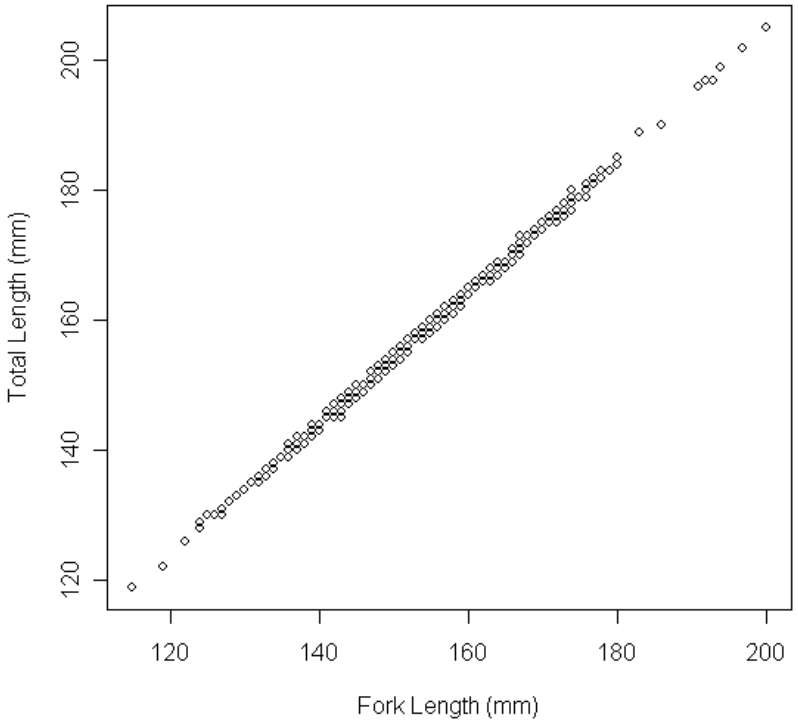
Простая линейная регрессия



Create a Fork Length to Total Length Conversion Equation

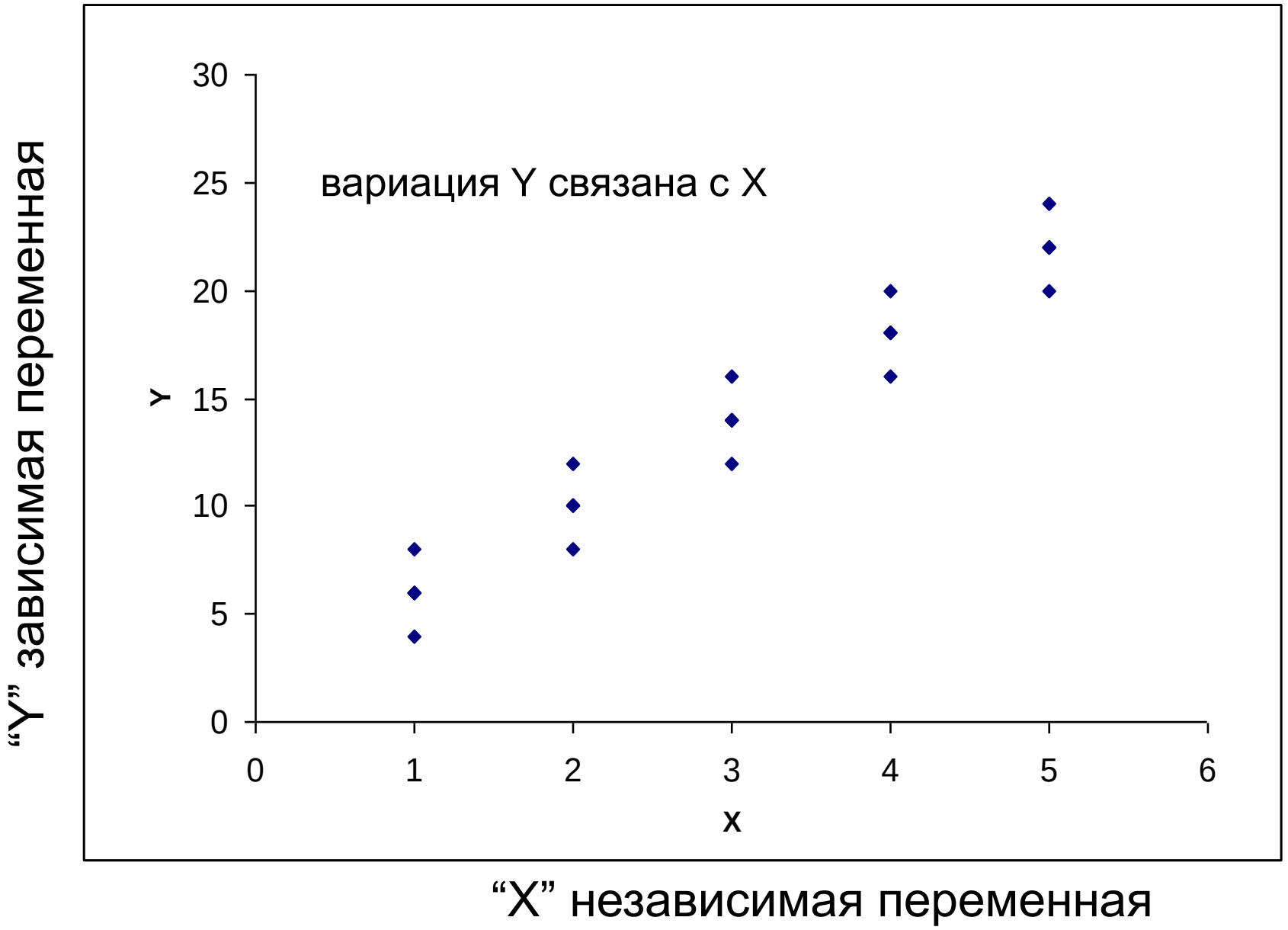
$TL=\beta_0+ \beta_1*FL$

Простая линейная регрессия

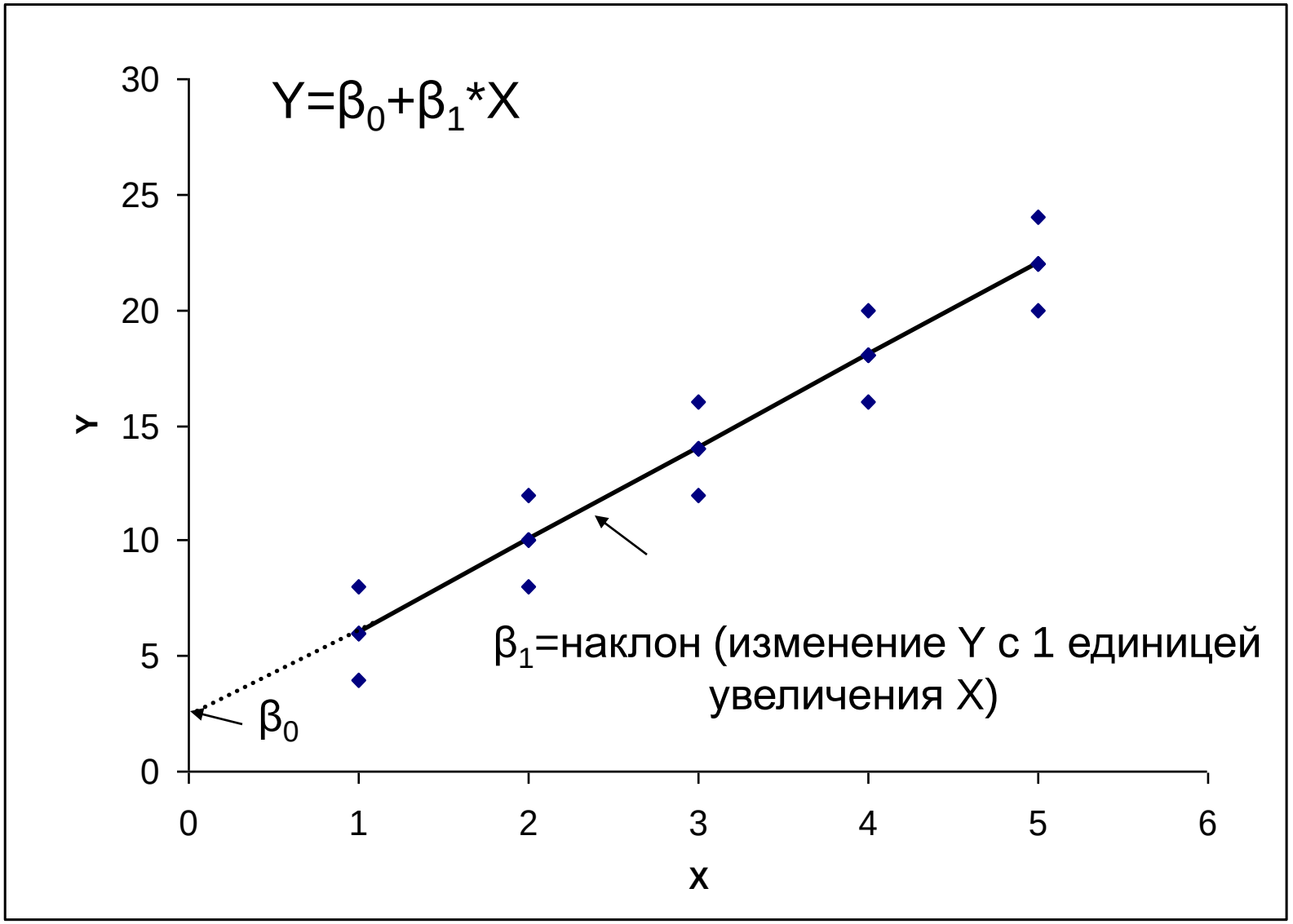


Create a Fork Length to Total Length Conversion Equation

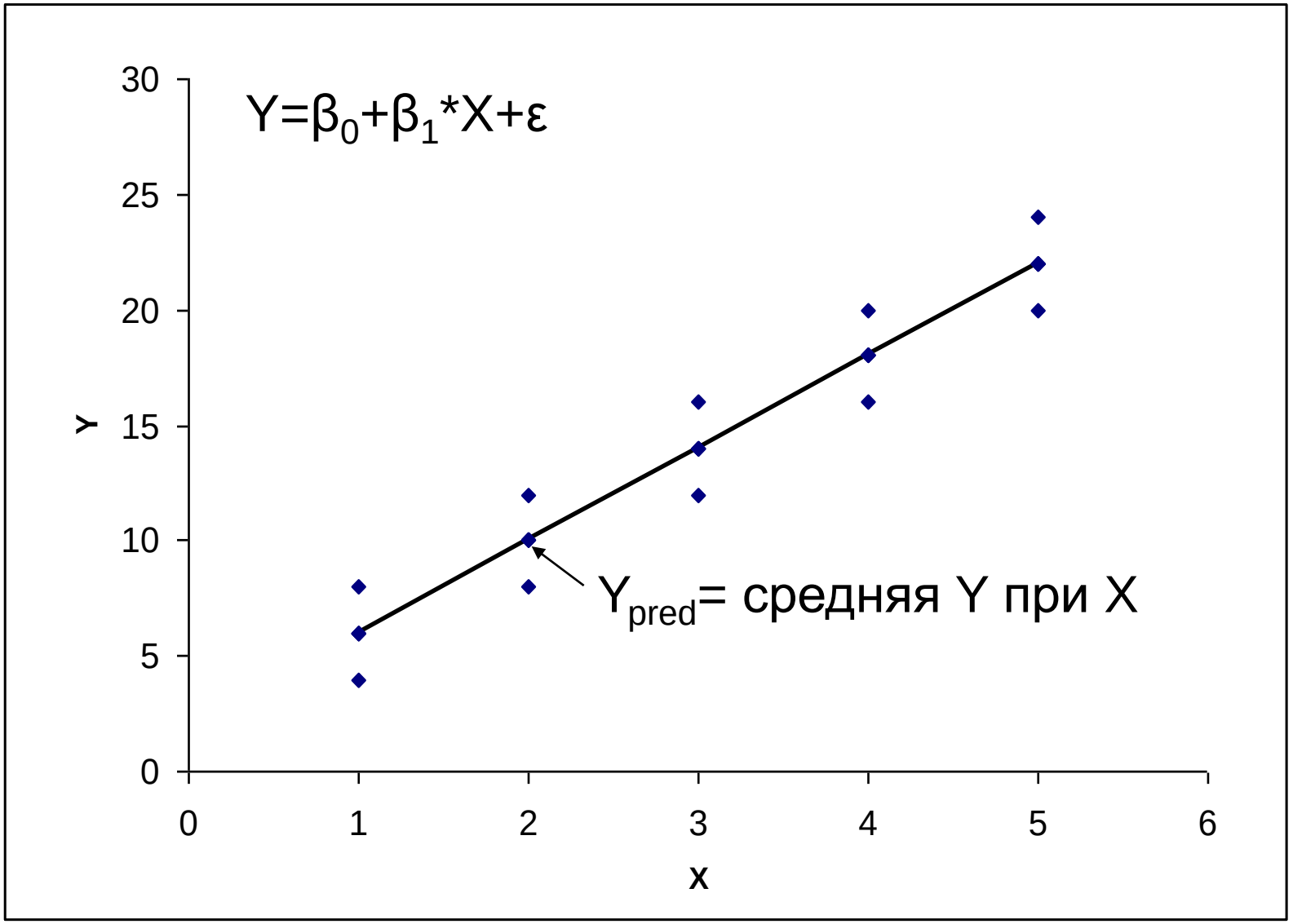
$TL=\beta_0+ \beta_1*FL$


вариация Y связана с X
30
25
20
15
10
5
0
Y
0
1
2
3
4
5
6
x
“Y” зависимая переменная
“X” независимая переменная

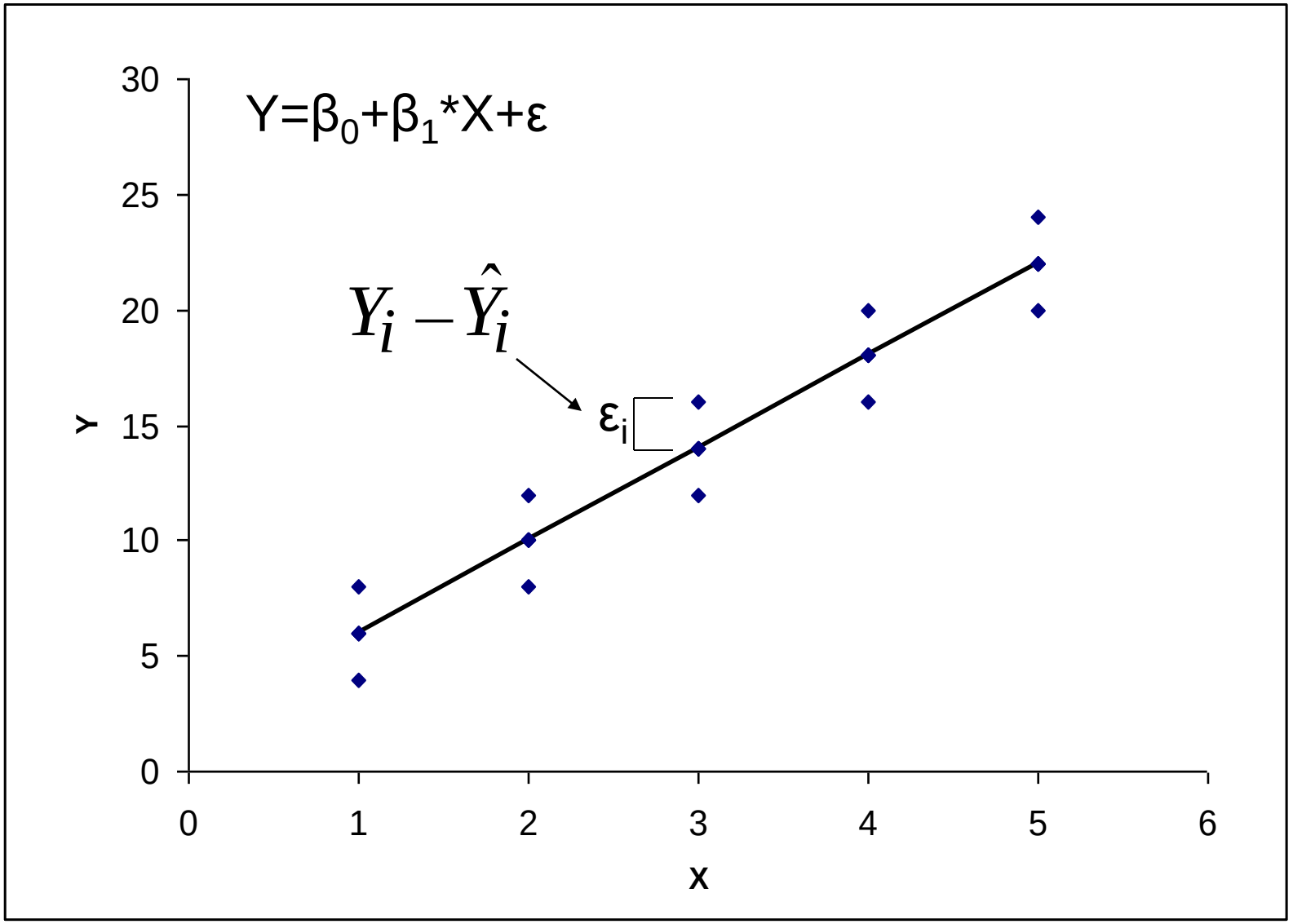
Наклон и пересечение



Предсказанная величина регрессии



отклонения

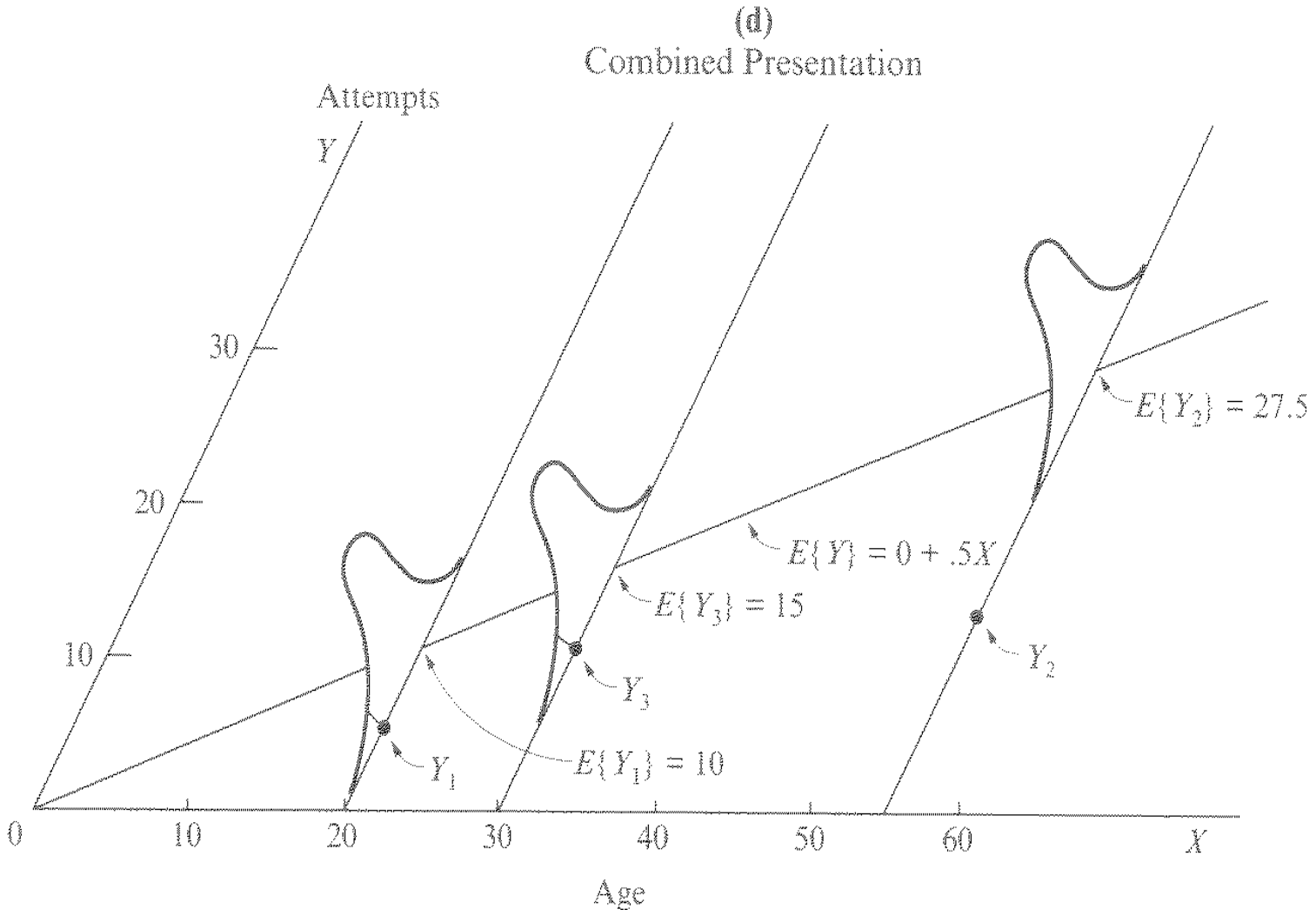


Предположения относительно ошибок

- Каждая ошибка принадлежит нормальгому распределению
- Постоянная дисперсия для всех уровней X
- Ошибки распределены независимо

Эти предпосылки важны для заключений о значительности модели и параметров!

Ошибки имеют нормальное распределение вокруг средней величины



Условия линейных моделей

- Y независимы (выбор одной не влияет на выбор другой величины, обеспечивается случайной выборкой)

  (f correlations, generalized estimating equation)

- X зафиксированы и измерены без ошибки

  (random effects models; measurement error models)

- Мат ожидание (средняя) Y для любого данного значения X описывается линейной функцией

  (if nonlinear, generalized additive models)

- Для каждого X_i, Ys независимы и нормально распределены

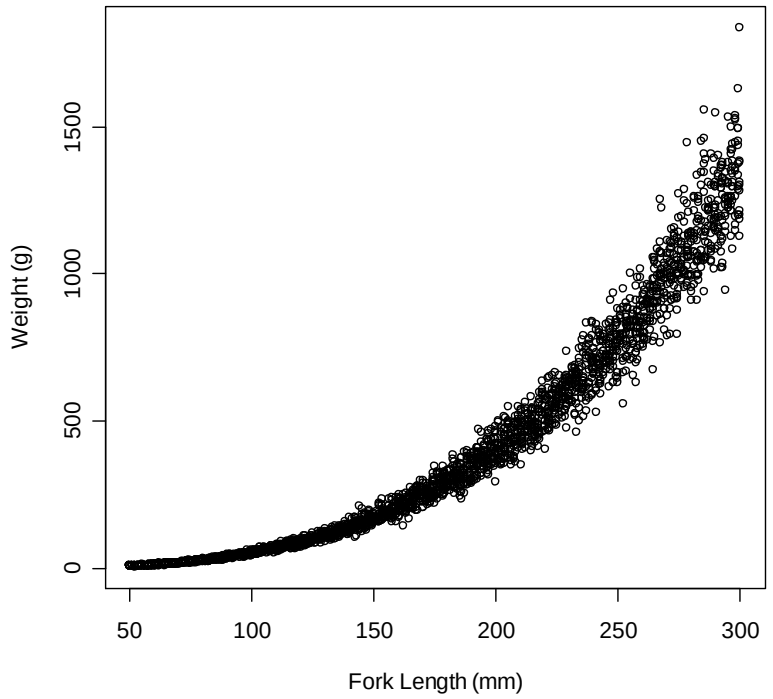
  (if not normal, generalized linear models)

- Y при данных X имеют постоянную дисперсию (гомосцедастичны)

  (if heteroscedastic, generalized linear models)

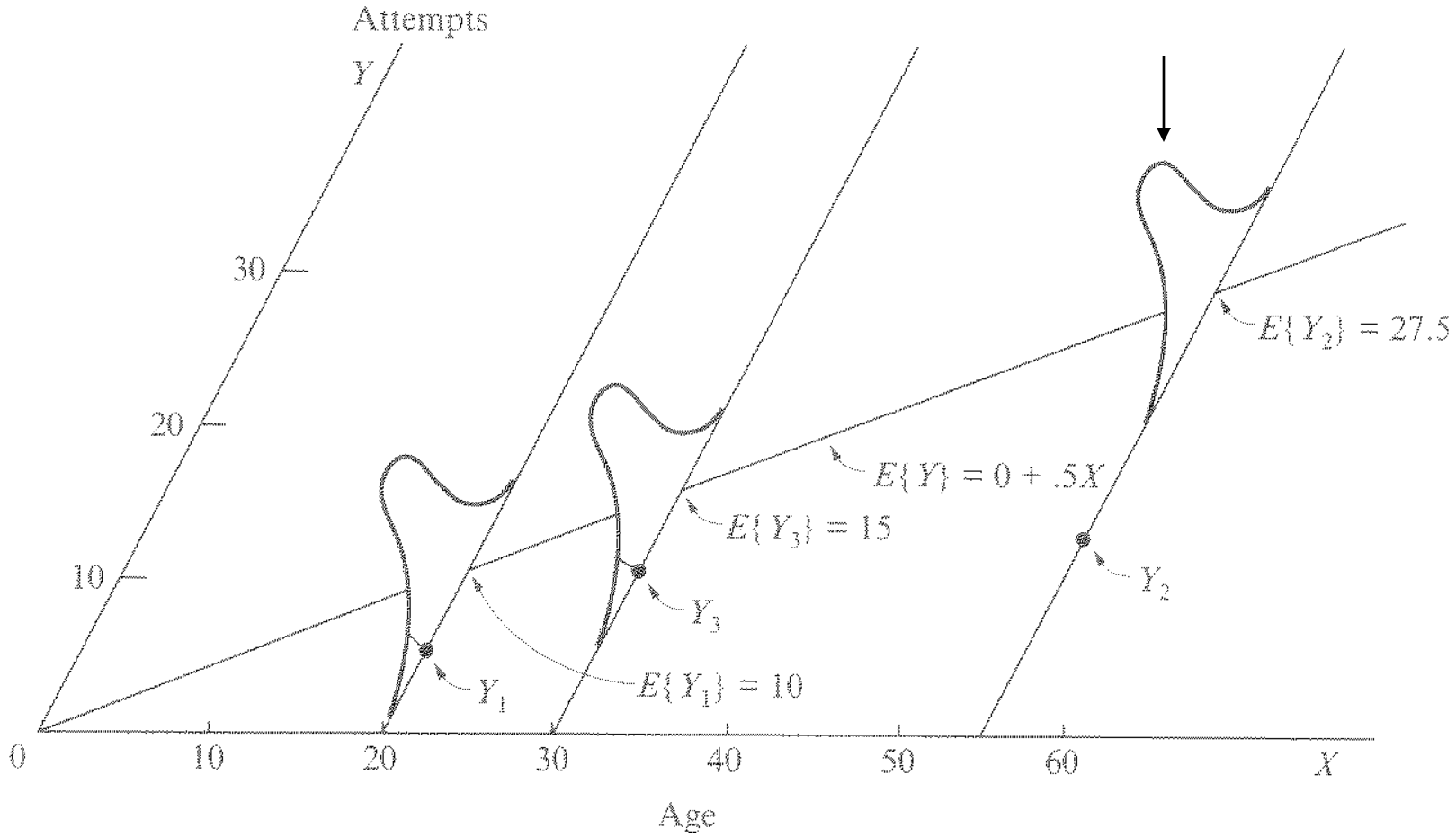
Обобщенные линейные модели

- Расширение линейных моделей которые позволяют работать с распределениями отличными от нормального и учитывать нелинейность в структуре модели



- Для оценки параметров использует метод максимального правдоподобия

Распределение ошибок может иметь различные формы

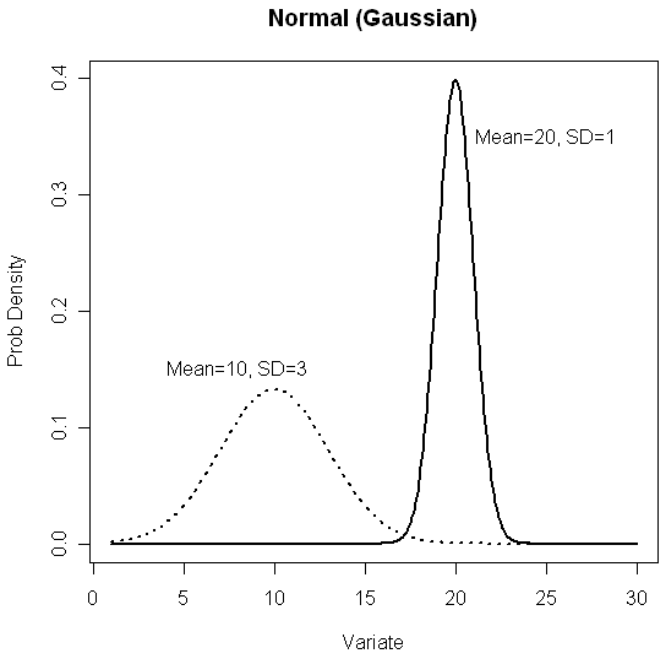


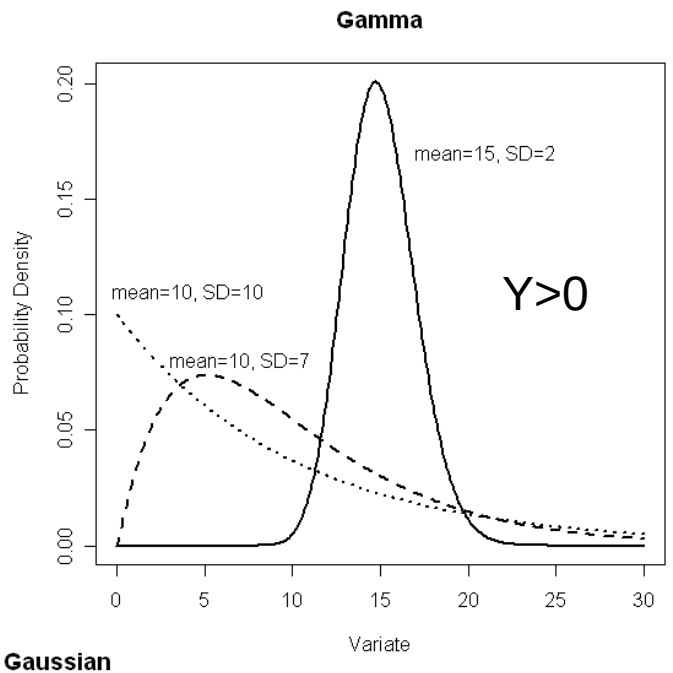
Случайная компонента (распределение ошибки)

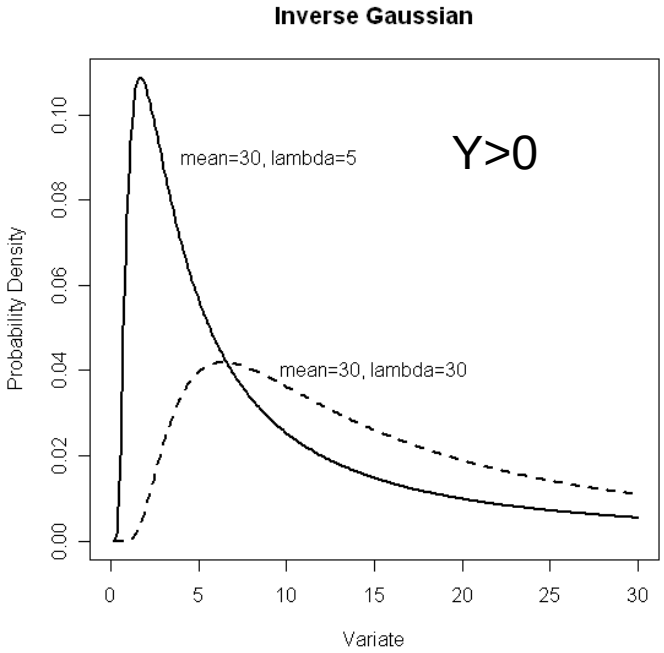
- Непрерывные и дискретные распределения экспоненциального семейства

Распределение	тип данных
Normal	непрерывное
Gamma	непрерывное
Inverse Gaussian	непрерывное
Binomial	дискретное
Poisson	дискретное

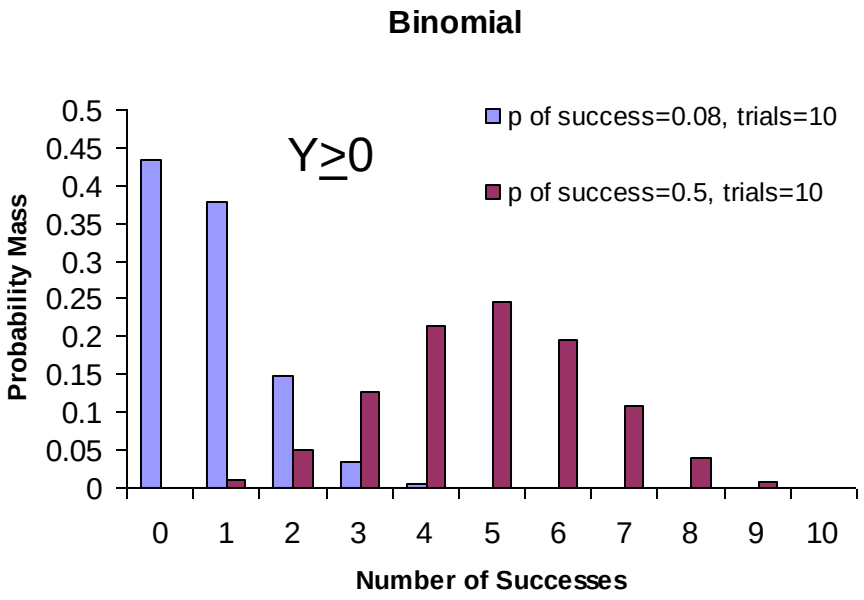
Continuous Distributions

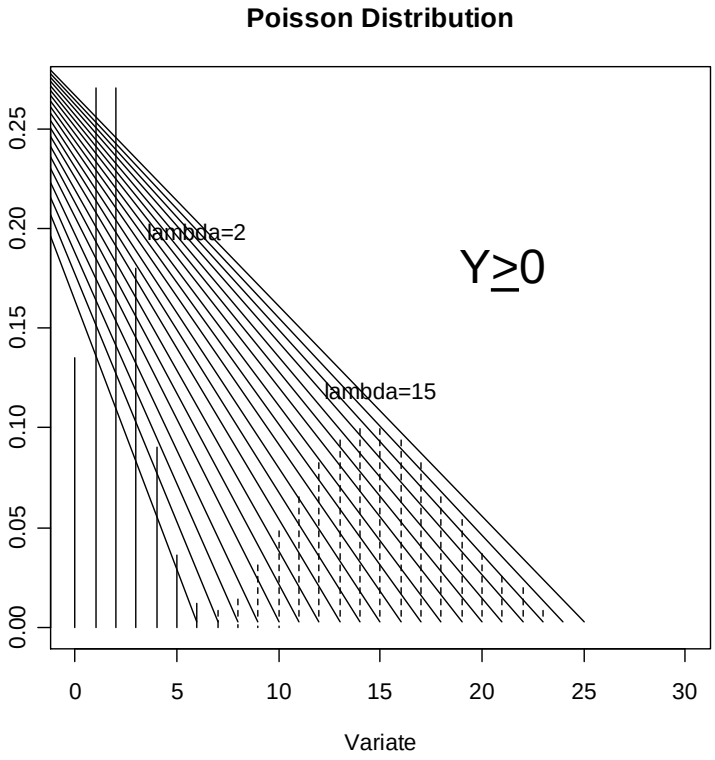




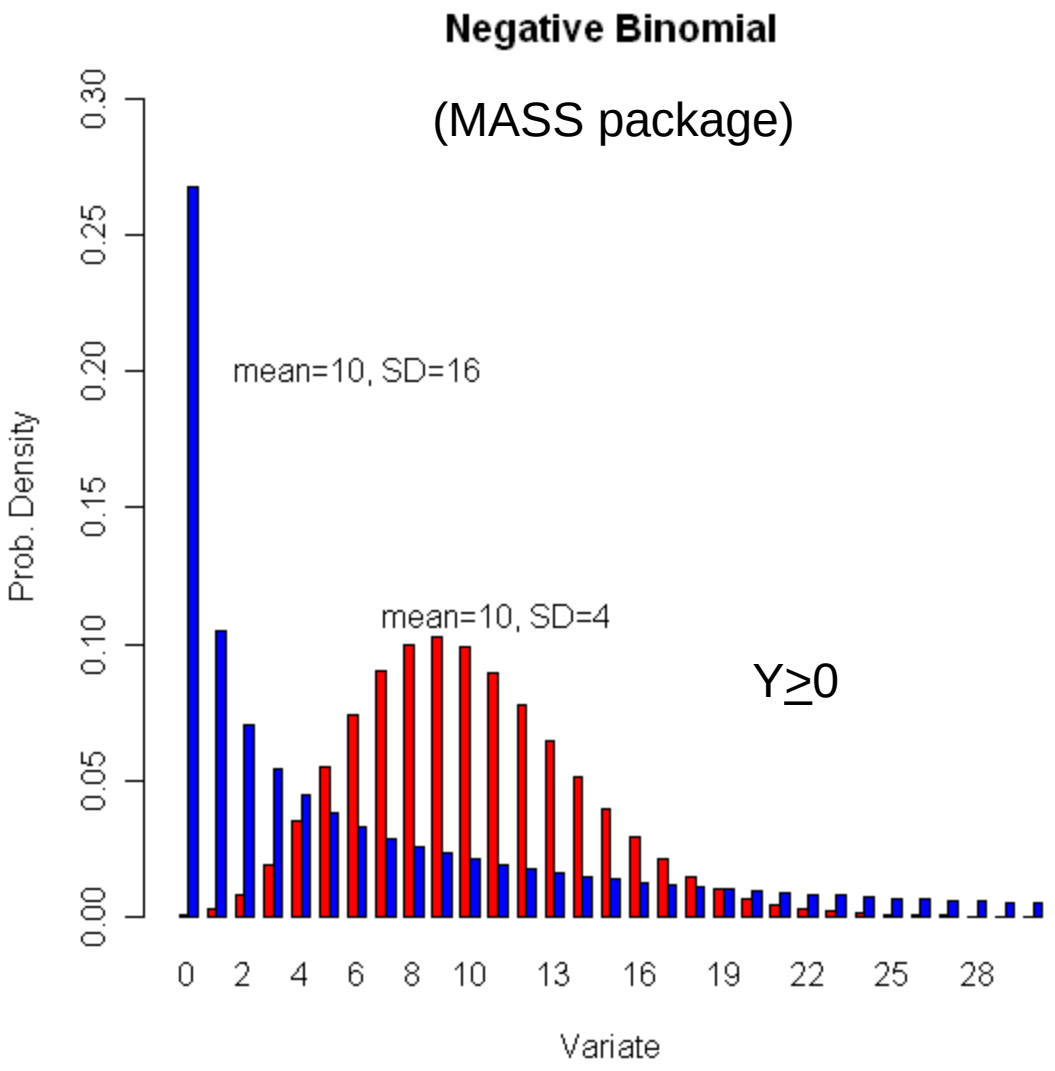


Discrete Distributions





Available Discrete Distribution



See Distributions.xls

Обобщенные линейные модели

- Обобщенная линейная модель состоит из трех компонентов :
  - Случайная компонента (ошибки) определяет условное распределение зависимой еременной для данных независимых переменных (предполагает что ошибки суммируются)
  - систематическая, линейная компонета независимых переменных

$$g(\mu) = \eta = \beta_0 + t + \beta_1 X_1 + \beta_2 t * X_1$$

  - Связная функция (g) которая линеаризует связь между средней величинойзависомй переменной и независимых переменных

Функции связи

пример: логистическая регрессия – p зависит от независимыз переменых (ковариванты)

$$p_i = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$$

- Можно оценить с помощью нелинейных методов

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta X_i$$ **Logit link function**

Функции связи

Family		“Natural” and Alternate Links
Normal		**identity,** log, 1/μ
Gamma	common →	log**, 1/μ**, identity
Inverse Gaussian		**1/μ²**, inverse, identity, log
Binomial		**logit,** probit, cloglog**,** cauchit, log
Poisson		**log**, identify, sqrt

Natural links are the technical link associated with the error distribution. Slight information is lost if other links are used.

Link functions add another layer of diagnostics!

Дисперсия в GLMs

- Кадое распределение имеет определенную связь между средней и дисперсией средней при разных X

Распределение	дисперсия
Normal	константа
Gamma	μ^2/v $(v=1/CV^2)$
Inverse Gaussian	μ^3/λ
Binomial	$\mu(1-\mu)$
Poisson	μ
Neg. Binomial	$\mu+\mu^2/k$

Функция дисперсии добавляет новый уровень диагностики!

- Designate Model and Error Relationship

$$\mu = \alpha X^{\beta} + \varepsilon \qquad \text{- additive error}$$

$$\mu = \alpha X^{\beta} \varepsilon \qquad \text{- multiplicative error}$$

Если ошибки перемножаются, трансформировать все уравнение чтобы линеаризовать модель а ошибки складывающимися

$$\log_{10} W = a + \beta * \log_{10}(L) + \varepsilon$$

- выбрать функцию связи

если ошибки складываются, выбрать функцию связи чтобы сделать линейной X (исследовать литературу, использовать натуральную связь)

- выбрать распределение ошибки

Select Models and Relationships

Family	Model	Link	Variance
Gaussian	$\mu = \beta_0 + \beta_1 X_1 + \varepsilon$	identity	constant
Gamma	$\mu = e^{\beta_0 + \beta_1 X_1} + \varepsilon$	log	μ^2
	$\mu = \frac{\alpha_0 X_1}{h + X_1} + \varepsilon$	inverse	μ^2
Poisson	$\mu = e^{\beta_0 + \beta_1 X_1} + \varepsilon$	log	μ
	$\mu = te^{\beta_0 + \beta_1 X_1} + \varepsilon$	log	μ
Binomial	$\mu = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}} + \varepsilon$	logit	$\mu(1-\mu)$