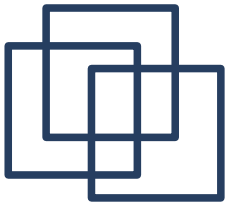


Линейная регрессия

Регрессионный анализ. Задачи

Регрессионный анализ – статистический мотод, который изучает влияние одной или нескольких независимых переменных на зависимую

- Выявление зависимости
- Предсказание значений зависимой переменной по значениям независимых
- Вклад отдельных независимых переменных в изменение зависимой

Регрессия

Основная идея: наблюдаемые значения зависимой переменной – измерения, которые содержат шум

$$y = f(x, b) + e$$

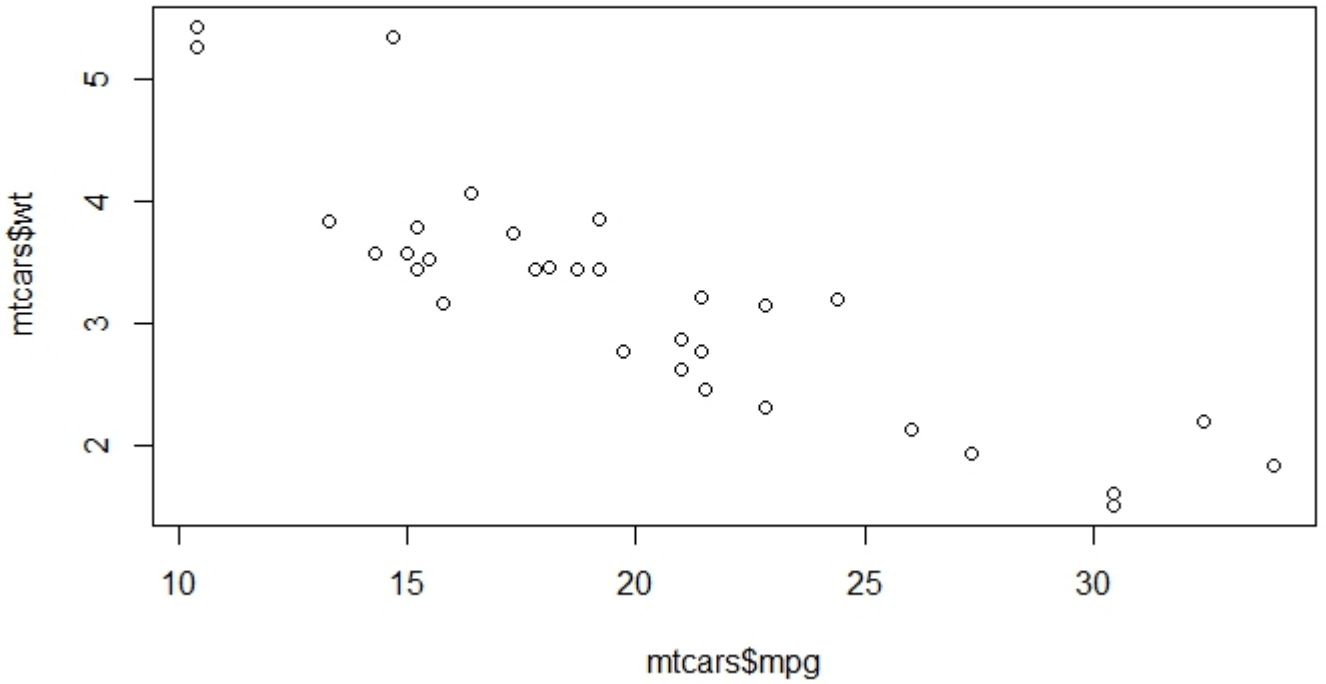
b_i – параметры модели

x_i – предикаторы (независимые переменные)

e – ошибка (все, что мы не можем измерить и учесть в модели)

Данные

```
> plot(mtcars$mpg, mtcars$wt)
```



Линейная регрессия

$$y = f(x, b) + e$$

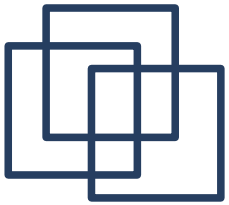
$$f(x, b) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

b_i – параметры модели

x_i – предикторы (независимые переменные)

e – ошибка (все, что мы не можем измерить и учесть в модели)

e распределено нормально!



Линейная регрессия

В случае одной независимой переменной

a – константа

b – коэффициент

$$f(x, b) = a + b\,x$$

Как это выглядит

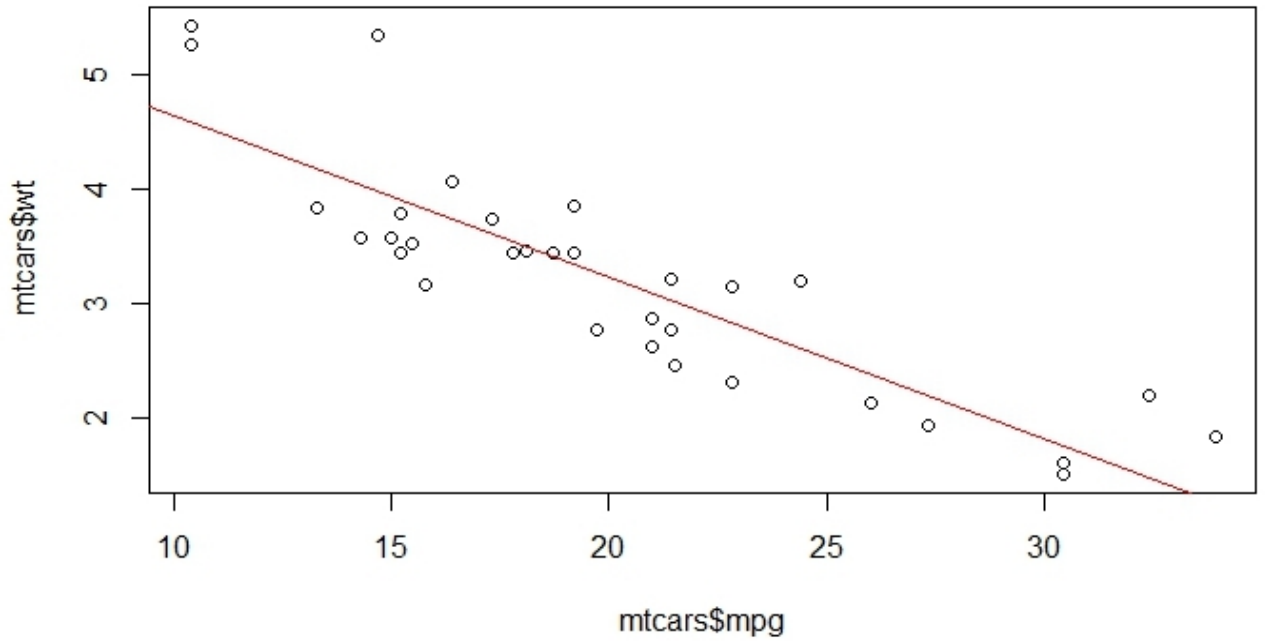
```
> lm1<-lm(mtcars$wt~mtcars$mpg)
> lm1

Call:
lm(formula = mtcars$wt ~ mtcars$mpg)

Coefficients:
(Intercept)   mtcars$mpg
     6.0473      -0.1409
```

Как это выглядит

```
> plot(mtcars$mpg, mtcars$wt)
> abline(lm1, col='red')
```



Оценка качества линейной регрессионной модели

> summary(lm1)

```
Call:
lm(formula = mtcars$wt ~ mtcars$mpg)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6516 -0.3490 -0.1381  0.3190  1.3684
```

Квантили для остатков (остаток=отклонение наблюдаемого значения от модели)
В идеале должны быть симметричны относительно 0

Оценка качества линейной регрессионной модели

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.04726    0.30869  19.590  < 2e-16 ***
mtcars$mpg  -0.14086    0.01474  -9.559 1.29e-10 ***
---
```

t-статистика (оценка/стандартная ошибка)

p-value

```
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
```

Оценка качества линейной регрессионной модели

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)}$$

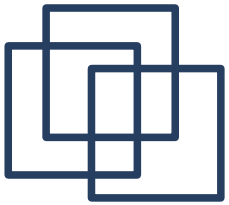
Доля объясненной дисперсии (чем ближе к 1, тем лучше)

```
Multiple R-squared: 0.7528,Adjusted R-squared: 0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

F-статистика (отношение объясненной дисперсии к ошибочной)

$$F = \frac{Var(\hat{Y})}{Var(error)}$$

P-value для всей модели



Доверительные интервалы

Даем параметру не точечную (одно значение), а интервальную оценку с заданным уровнем надежности. **Такая оценка предпочтительна при небольшом обьеме выборки!**

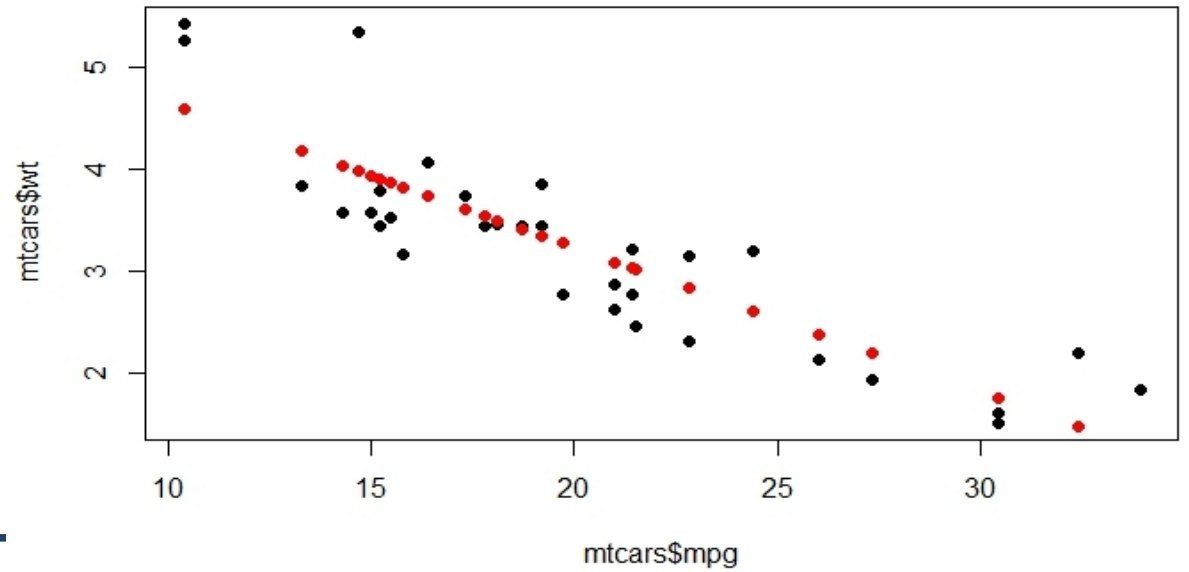
```
> confint(lm1)# confint(lm1, level=0.95)
                2.5 %     97.5 %
(Intercept)  5.4168245  6.6776856
mtcars$mpg  -0.1709569 -0.1107671
```

Уровень надежности (95%) означает вероятность того, что значение параметра попадет в этот интервал.

Полезные значения

Значения по модели:

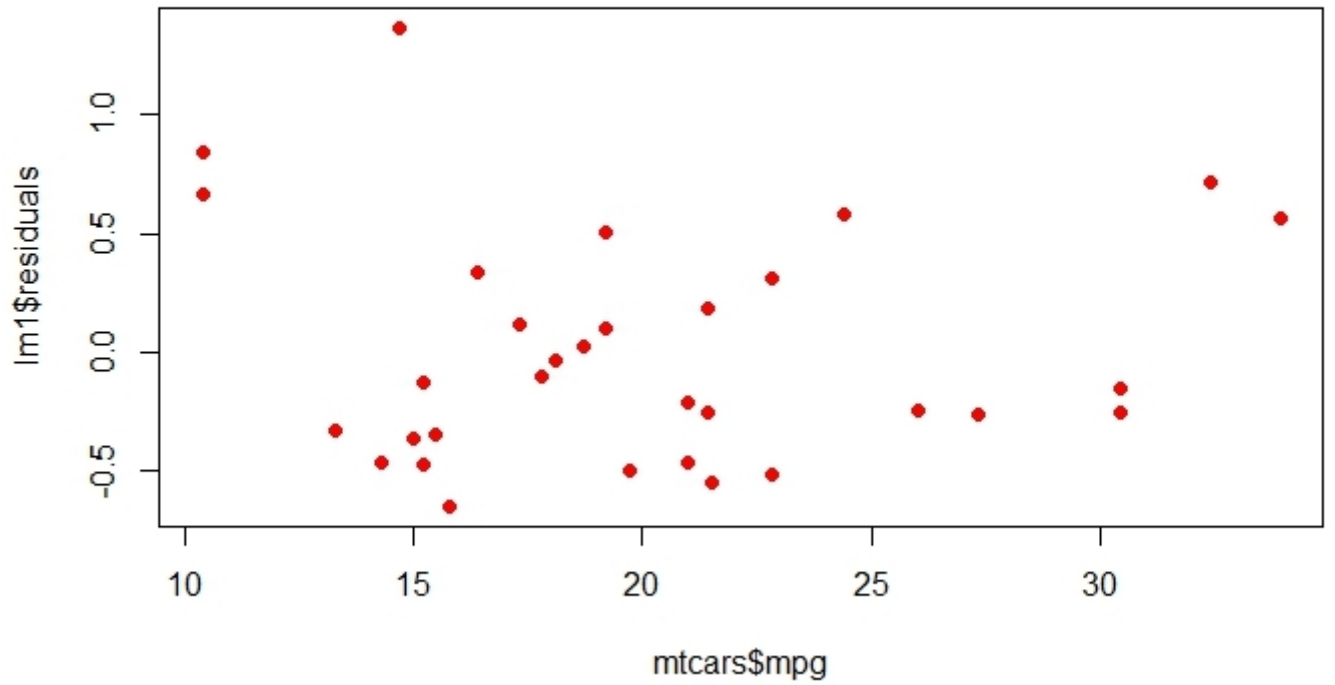
```
> plot(mtcars$mpg, mtcars$wt, pch=19)
> lm1<-lm(mtcars$wt~mtcars$mpg)
> points(mtcars$mpg, lm1$fitted, pch=19, col='red')
```

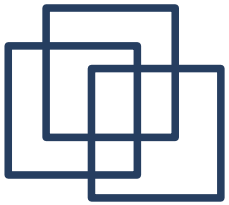


Полезные значения

Остатки:

> plot(mtcars$mpg, **lm1$residuals**, pch=19, col='red')





Полезные значения

Коэффициенты:

```
> lm1$coefficients
(Intercept)  mtcars$mpg
   6.047255   -0.140862
> lm1$coefficients[1]
(Intercept)
   6.047255
```

Обобщенные линейные модели

- Пуассонова (log-linear model)
- Биномиальная (логистическая)
- Гамма
- Гауссова
- Обратная Гауссова
- Квази
- Квазибиномиальная
- Квазипуассонова

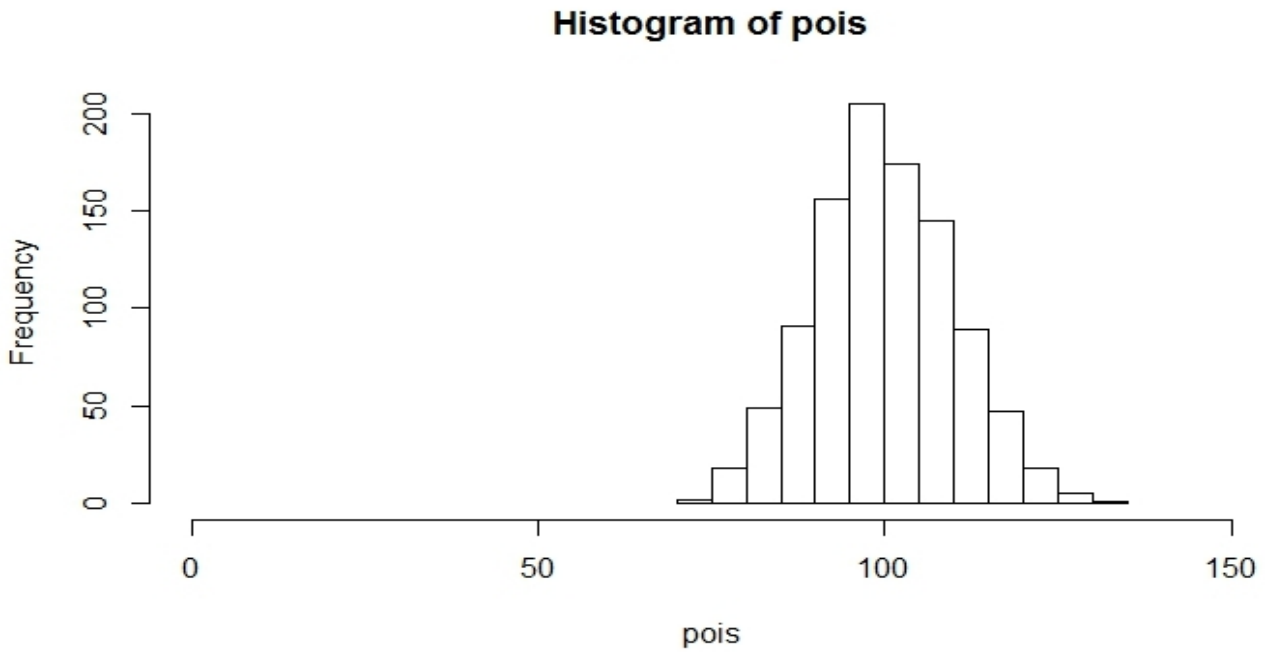
Пуассонова регрессия

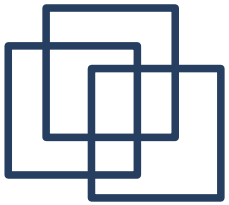
- Используется для работы с **количественными данными**
- Предполагается, что зависимая переменная имеет распределение Пуассона (редкие события, например, появление автобусов на остановке за определенный промежуток времени, количество звонков на коммутатор за день и т.п.). События независимы, но происходят с некоторой фиксированной средней интенсивностью
- тогда логарифм ожидаемого значения зависимой переменной (например, количество автобусов) является линейной комбинацией независимых переменных (например, времени)

$$\log(E(y)|x, b_{0,} b_1) = b_0 + b_1 x$$

Распределение Пуассона

```
> pois<-rpois(1000, lambda=100)
> hist(pois, xlim=c(0, 150))
```





Задача: определить зависимость

Данные: количество посещений сайта (для удобства превратим дату в единое число)

```
> load('gaData.rda')

> gaData$julian<-julian(gaData$date)

> head(gaData)
```

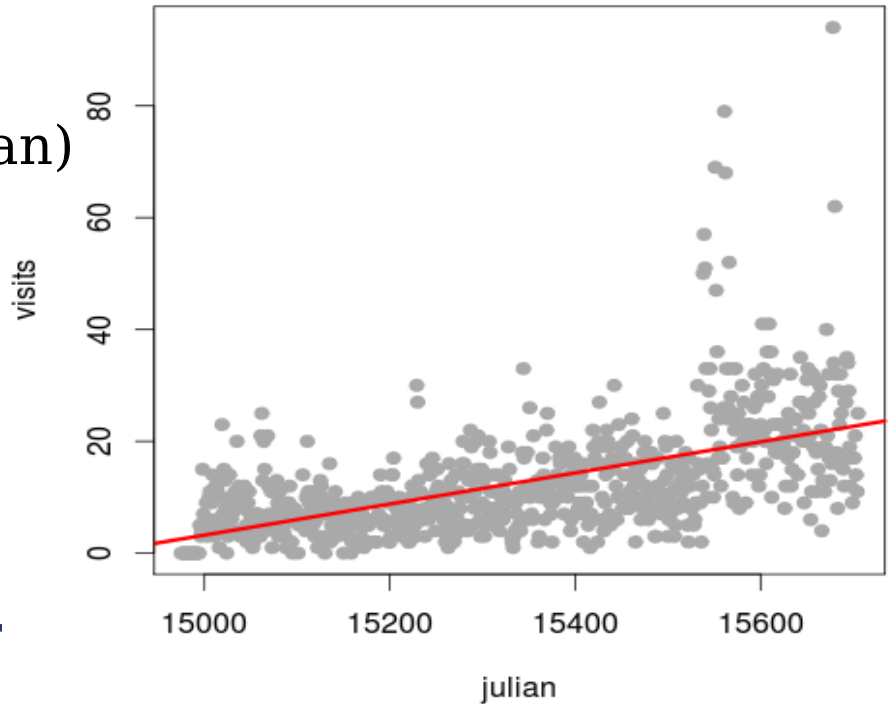
	date	visits	simplystats	julian
1	2011-01-01	0	0	14975
2	2011-01-02	0	0	14976
3	2011-01-03	0	0	14977
4	2011-01-04	0	0	14978
5	2011-01-05	0	0	14979
6	2011-01-06	0	0	14980

Задача: определить зависимость

```
> plot(gaData$julian, gaData$visits, xlab="julian", ylab="visits",
pch=19, col="darkgrey")
```

Сначала построим линейную модель (**В этом случае мы считаем, что число посещений всегда растет с постоянной скоростью!**)

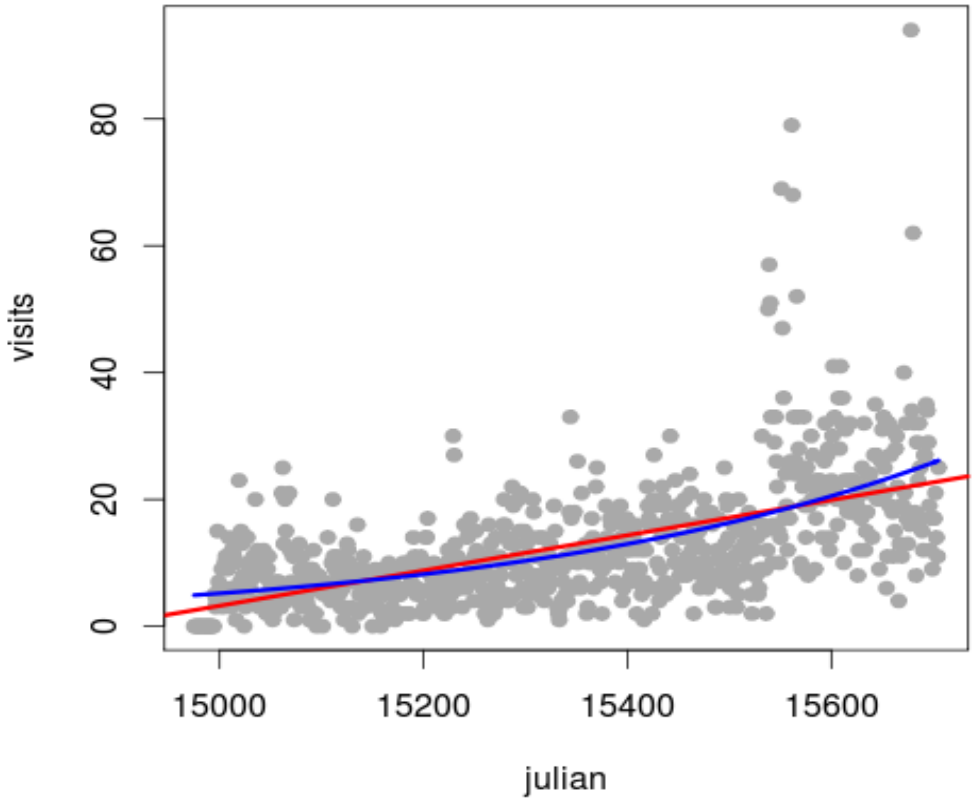
```
> lm1<-lm(gaData$visits~gaData$julian)
> abline(lm1, col='red', lwd=3)
```

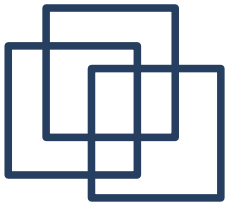


Пуассонова регрессия

Другая постановка задачи: рассматриваем посещения сайта как случайный процесс с некоторой средней интенсивностью!

```
> glm1<-glm(gaData$visits~gaData$julian, family='poisson')
> lines(gaData$julian, glm1$fitted, col='blue', lwd=3)
```





Логистическая регрессия

Зависимая переменная принимает два значения:

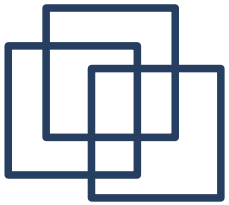
болен/здоров

жив/мертв

успех/неудача

И т.п.

Требуется вычислить вероятности для значений зависимой переменной



Логистическая регрессия

Примечание: вероятность всегда от 0 до 1, поэтому линейная регрессия не годится :-)

Пример: таблица выигрышей команды. Хотим вычислять вероятность выигрыша в зависимости от количества очков

```
> load("ravensData.rda")

> head(ravensData)
  ravenWinNum ravenWin ravenScore opponentScore
1           1        W         24             9
2           1        W         38            35
3           1        W         28            13
4           1        W         34            31
5           1        W         44            13
6           0        L         23            24
```

Пробуем линейную модель:

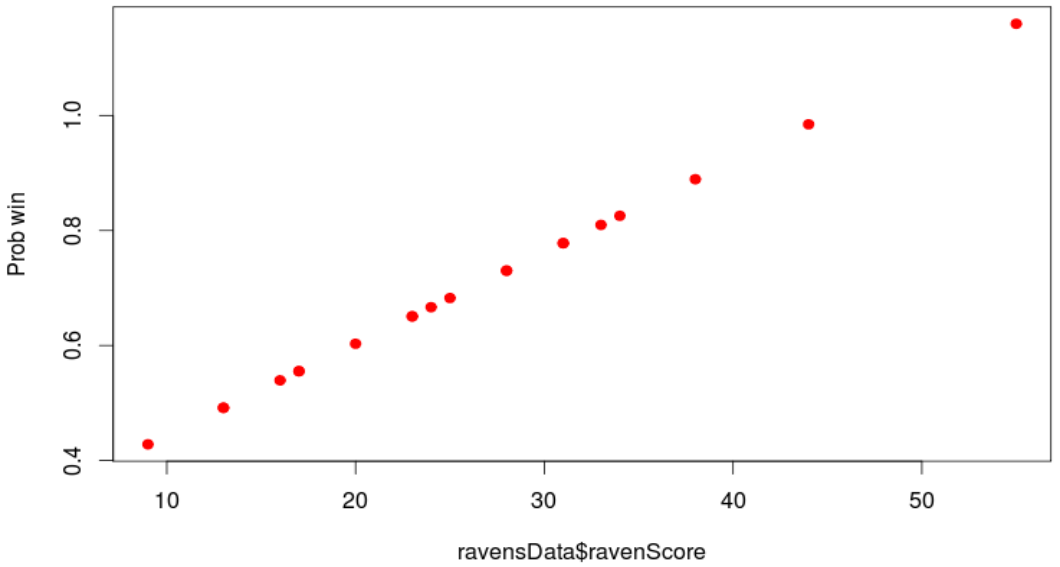
$$RW_i = b_0 + b_1 RS_i$$

где RW_i – выигрыш (0 или 1)

RS_i – количесво очков

```
> lmRav<-
lm(ravensData$ravenWinNum~ravensData$ravenScore)

> plot(ravensData$ravenScore, lmRav$fitted, pch=19,
col='red', ylab="Prob win")
```



Для некоторых значений получили вероятность >1!!!

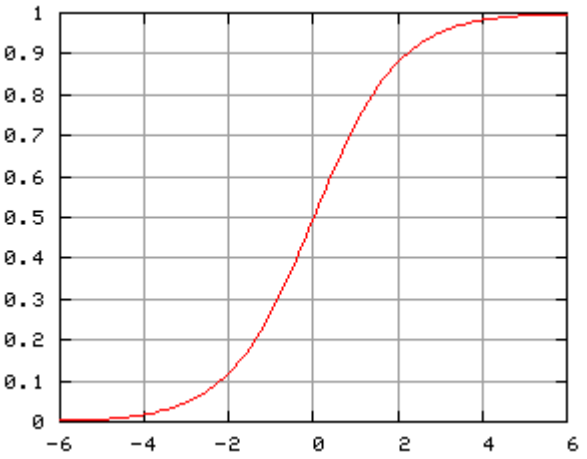
Логистическая регрессия

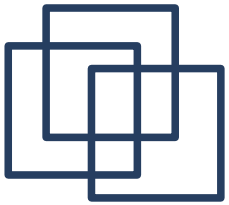
логарифм отношения правдоподобия линейно зависит от независимой переменной

$$\log\left(\frac{Pr(RW_i|RS_i, b_0, b_1)}{1-Pr(RW_i|RS_i, b_{0,} b_1)}\right)=b_0+b_1 RS_i$$

где RW_i – выигрыш (0 или 1)

RS_i – количесво очков





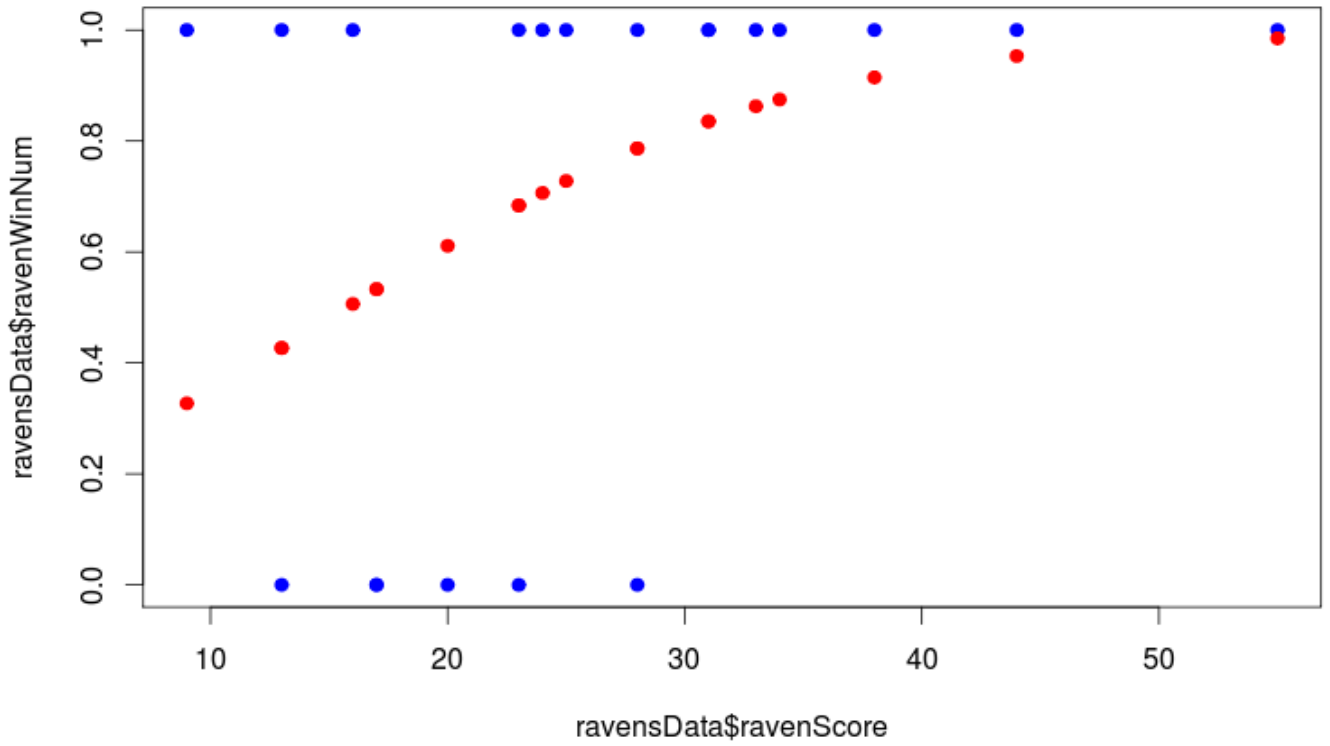
Логистическая регрессия

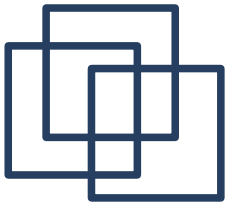
```
> logReg<-
glm(ravensData$ravenWinNum~ravensData$ravenScore,
family="binomial")
> logReg

Call:  glm(formula = ravensData$ravenWinNum ~
ravensData$ravenScore,
    family = "binomial")
Coefficients:
          (Intercept)  ravensData$ravenScore
              -1.6800                 0.1066
Degrees of Freedom: 19 Total (i.e. Null);  18
Residual
Null Deviance:      24.43
Residual Deviance: 20.89        AIC: 24.89
```

Логистическая регрессия

```
> plot(ravensData$ravenScore, ravensData$ravenWinNum,
pch=19, col='blue')
> points(ravensData$ravenScore, logReg$fitted, pch=19,
col='red')
```





Конец