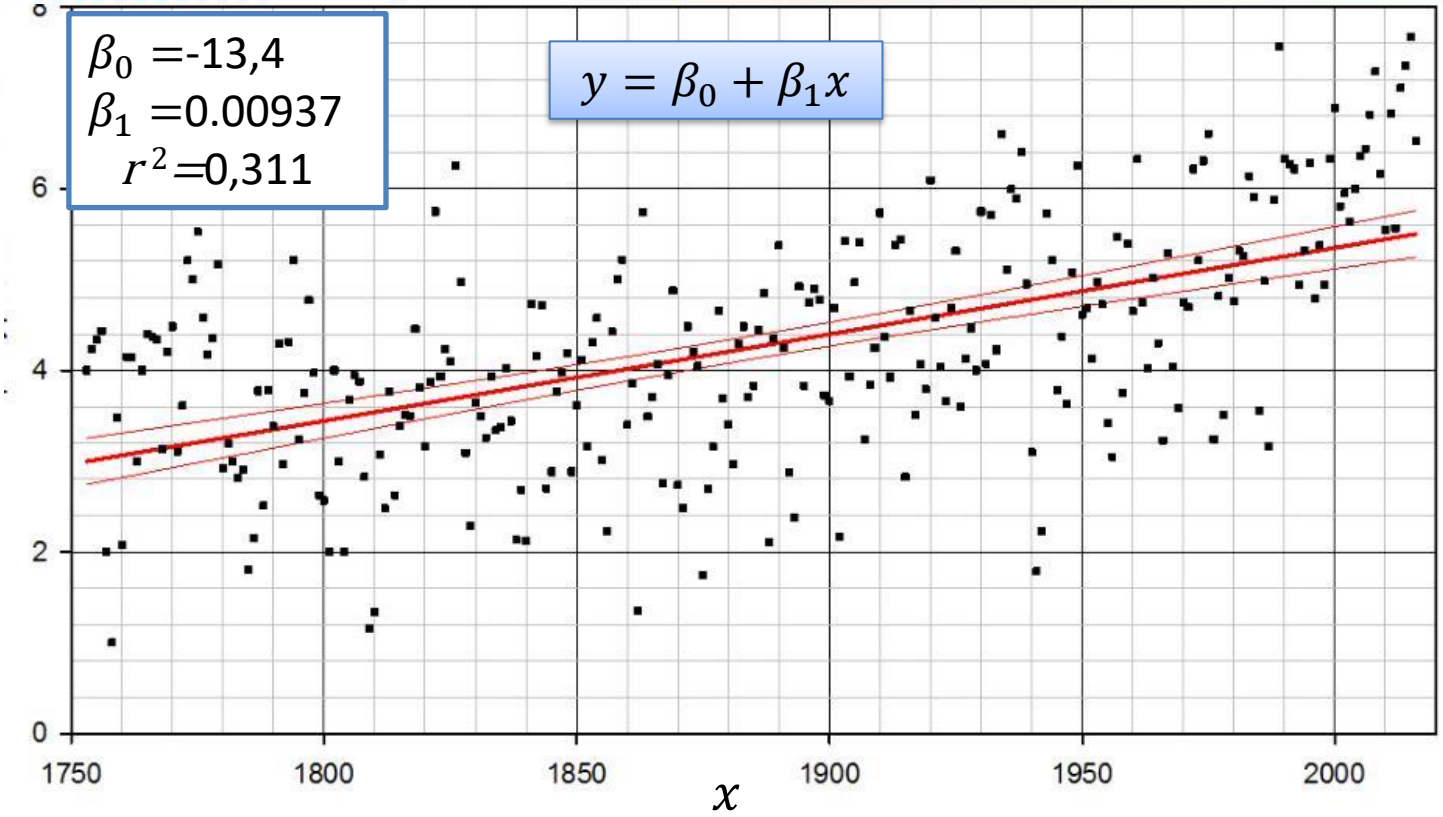
СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 18. Регрессионный анализ. Часть 2

Зависимость среднегодовых температур СПб от времени

```
> f<-scan("SPB.txt",what="numeric")
Read 264 items
> y<-as.numeric(f)
> n<-length(y)
> x<-c(1753:(1752+n))
> data1<-data.frame(year=x,temperature=y)
> M<-lm(y~x,data=data1)
> summary(M)
Call:
lm(formula = y ~ x, data = data1)
Residuals:
    Min      1Q  Median      3Q     Max
-3.00281 -0.63548  0.05369  0.86254  2.54936
Coefficients:
             Estimate          Std. Error    t value    Pr(>|t|)
(Intercept) -1.339e+01  1.623e+00   -8.252     7.65e-15 ***
x            9.367e-03         8.606e-04   10.884     < 2e-16 ***
Residual standard error: 1.066 on 262 degrees of freedom
Multiple R-squared:  0.3114,    Adjusted R-squared:  0.3087
```

Зависимость среднегодовых температур СПб от времени



Нелинейные модели

В общем случае рассматривается регрессионная модель вида:

$$y_i = f(x_i, \beta) + e_i, \qquad i = 1, \dots, n; \qquad \beta = (\beta_1, \dots, \beta_l).$$

Задача нахождения значений векторного параметра β сводится к нелинейной задаче МНК в виде задачи нелинейного программирования:

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - f(x_i, \beta)\right)^2 \to \min_{\beta}.$$

Решение таких задач рассматривается в курсе по методам оптимизации.

Нелинейные модели

Нелинейные регрессионные модели, сводящиеся к простой линейной регрессии при помощи замены переменных.

1. Модели, в которых преобразуются только факторы (замена переменной фактора).

Пусть коэффициенты β входят линейно, а фактор X– нелинейно:

$$Y = \beta_0 + \beta_1 F(X) + e,$$

Положив $Z = F(X)$, мы приходим к уравнению линейной регрессии

$$Y = \beta_0 + \beta_1 Z + e.$$

Примеры:

$$1)\ Y = \beta_0 + \beta_1 (1/X) + e;\ \ Z = 1/X.$$

$$2)\ Y = \beta_0 + \beta_1 \ln X + e;\ \ Z = \ln X.$$

$$3)\ Y = \beta_0 + \beta_1 X^{1/2} + e;\ \ Z = X^{1/2}.$$

Нелинейные модели

2. Нелинейные модели, в которых наряду с факторами преобразуется и отклик

Мультипликативная модель: $Y = \alpha X^{\gamma} \varepsilon,$

α, γ – неизвестные параметры, ε – мультипликативная случайная ошибка, распределенная на положительной полуоси с математическим ожиданием, равным единице и конечной дисперсией.

Логарифмирование исходного уравнения :

$\ln Y = \ln\alpha + \gamma \ln X + \ln\varepsilon$ или $V = \beta_0 + \beta_1 Z + e,$

где $V = \ln Y,\ Z = \ln X,\ \beta_0 = \ln\alpha,\ \beta_1 = \gamma,\ e = \ln\varepsilon.$

Нелинейные модели

Экспоненциальная модель:

$$Y = \alpha \exp(\gamma X)\varepsilon,\ \ \ln Y = \ln\alpha + \gamma X + \ln\varepsilon,\ V = \beta_0 + \beta_1 Z + e,$$

Здесь $V = \ln Y,\ Z = X,\ \beta_0 = \ln\alpha,\ \beta_1 = \gamma,\ e = \ln\varepsilon$

Обратная модель:

$$Y = \frac{1}{\beta_0 + \beta_1 F(X) + e},\ \frac{1}{Y} = \beta_0 + \beta_1 F(X) + e,\ V = 1/Y$$

Обратная экспоненциальная модель:

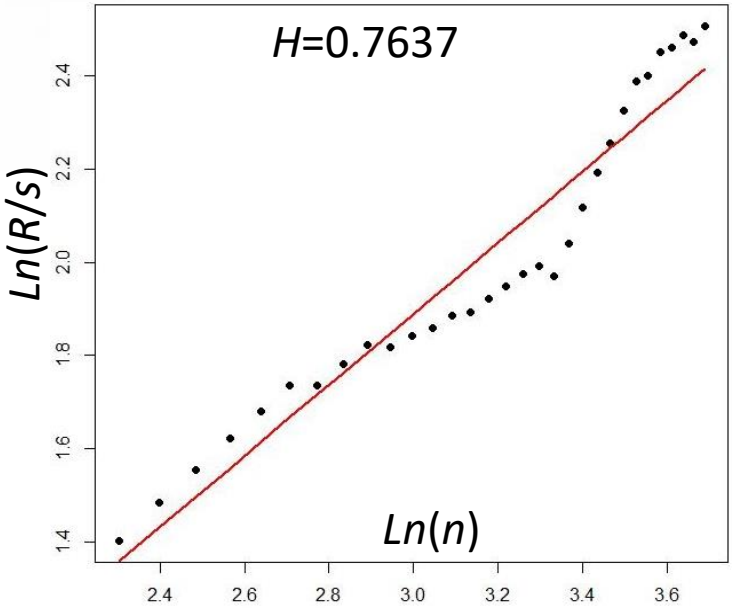
$$Y = \frac{1}{1 + \alpha \exp(\beta_1 X + e)}, \qquad \ln(1/Y - 1) = \ln\alpha + \beta_1 X + e.$$

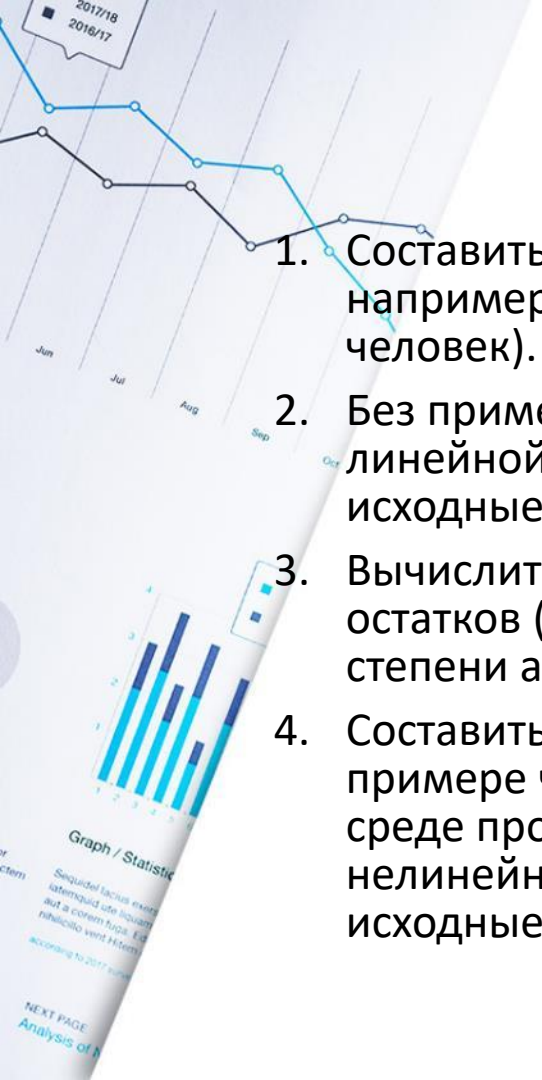
$$V = \ln(1/Y - 1)$$

Пример.Расчет показателя Херста

```
1.  f<-scan("SPB.txt",what="numeric")
2.  y<-as.numeric(f)
3.  RS<-vector("numeric",length=31)
4.  LN<-vector("numeric",length=31)
5.  for (i in 10:40)
6.  {
7.  y1<-vector("numeric",length=i)
8.  x<-vector("numeric",length=i)
9.  z<-vector("numeric",length=i)
10. y1<-y[1:i]
11. m<-mean(y1)
12. s<-sd(y1)
13. x<-y1-m
14. for (k in 1:i) { Z[k]<-sum(X[1:k]) }
15. R<-max(Z)-min(Z)
16. RS[i-9]<-log(R/s)
17. LN[i-9]<-log(i)
18. }
19. data1<-data.frame(LN,RS)
20. M<-lm(RS~LN,data=data1)
21. H<-M$coefficient[2]
```

Первые 40 точек ряда среднегодовых температур СПб





Контрольные вопросы и задания

1. Составить таблицу сопоставляющую рост и вес взрослых людей например, членов семьи, друзей, членов учебной группы и т.п. (5-10 человек).
2. Без применения программных средств построить модель простой линейной регрессии зависимости веса от роста. Нанести на график исходные точки и линию регрессии.
3. Вычислить суммы квадратов остатков (RSS), стандартной ошибки остатков (RSE) и корреляционное отношение η^2 (R^2). Сделать выводы о степени адекватности линейной модели.
4. Составить таблицу сопоставляющую возраст и вес различных людей на примере членов семьи, родственников и знакомых (10-20 значений). В среде программирования R построить мультипликативную модель нелинейной регрессии зависимости веса от возраста. Нанести на график исходные точки и линию регрессии. Оценить адекватность модели.