Регресионни модели

Лекция 10

Simple Linear Regression

Стохастични модели

Обща форма

Y = Детерминистичен компонент+грешка

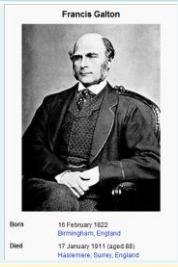
Регресия: Y= a + bX +u

X,Y са непрекъснати променливи

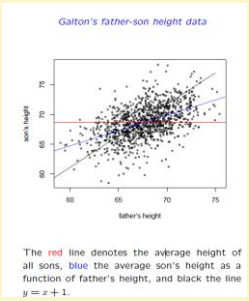
Y – зависима променлива (dependent)

X – независима променлива (independent)

Франсис Галтон и Карл Пирсън

Сър Франсис Галтон	Карл Пирсън
	

“Регресия”

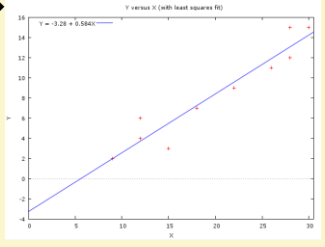


The red line denotes the average height of all sons, blue the average son's height as a function of father's height, and black the line $y = x + 1$.

Пример 1.

- Y – Зависима променлива (лв)
- X – Независима променлива (кг)
- Корелограма→

	Y	X
1	2	9
2	4	12
3	3	15
4	7	18
5	6	12
6	9	22
7	11	26
8	15	28
9	12	28
10	15	30



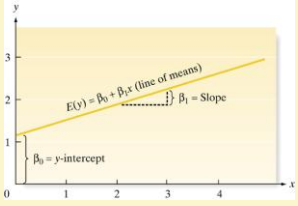
Регресионен модел

```
Model 1: OLS, using observations 1-10
Dependent variable: Y

              coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       -3.28498      1.36964     -2.398     0.0433   **
  X            0.584249     0.0642378    9.095     1.72e-05 ***

Mean dependent var   8.400000   S.D. dependent var   4.765618
Sum squared resid    18.02454   S.E. of regression   1.501022
R-squared            0.911817   Adjusted R-squared   0.900794
F(1, 8)              82.72075   P-value(F)           0.000017
Log-likelihood      -17.13513   Akaike criterion     38.27026
Schwarz criterion    38.87543   Hannan-Quinn         37.60639
```

- Статистическа значимост на модела (ДА) p=0.000017
- (Но: Моделът не е стат. значим На: Моделът е стат. значим)
- Коефициент на детерминация – R-squared=0.911817 или 91%
- Регресионни коефициенти
  Свободен член (constant) β_0 = -3.28 Регресионен коефициент β_1 =0.58
- Интерпретация: β_0 : Когато X=0 кг. то Y= -3.28 лв (намаление от 3 лв и 28 ст.)
- Интерпретация: β_1 : Ако X се увеличи с 1 кг. то Y се увеличава с 0.58 лв (58 ст.)

Проста линейна регресия

$$y = \beta_0 + \beta_1 x + \varepsilon$$



β_0 = свободен член (constant, intercept)

β_1 = регресионен коефициент (slope)

ε = грешка

$\varepsilon = y - \hat{y}$;

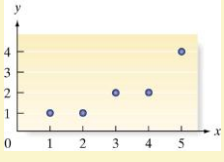
y – реална стойност

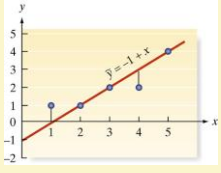
$\hat{y}$ – изгладена стойност

Оценка на модела. Метод на най-малките квадрати (МНМК)

(Ordinary Least Squares, OLS)

Време за реакция при лекарство		
□□□□□□□	□□□□□□□□□□□ □□ □□□□□□□□□ x (%)	□□□□□ □□ □□□□□□□ y (□□□□□□□)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4





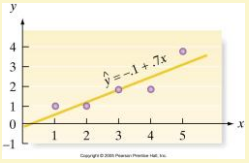
Пример 2.

Preliminary Computations

	x_i	y_i	x_i^2	$x_i y_i$
	1	1	1	1
	2	1	4	2
	3	2	9	6
	4	2	16	8
	5	4	25	20
Totals	$\sum x_i = 15$	$\sum y_i = 10$	$\sum x_i^2 = 55$	$\sum x_i y_i = 37$

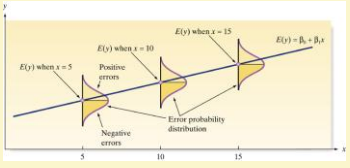
Comparing Observed and Predicted Values for the Least Squares Prediction Equation

x	y	$\hat{y} = -.1 + .7x$	$(y - \hat{y})$	$(y - \hat{y})^2$
1	1	.6	.4	.16
2	1	1.3	-.3	.09
3	2	2.0	0.0	.00
4	2	2.7	-.7	.49
5	4	3.4	.6	.36
			Sum of Errors = 0	SSE = 1.10



Изисквания

1. Средната грешка ε = 0
2. Дисперсията на ε е константа за всички стойности на *x*
3. Разпределението на ε е нормално
4. Стойностите на ε са независими една от друга



МНМК

Регресионна линия: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

•Сумата от грешките (отклоненията) (SE) = 0 $\varepsilon = y - \hat{y}$

•Сумата от грешките на квадрат (SSE) =min

Формули:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

Проверка на хипотези за β_1

Проста линейна регресия

Едностранен тест	Двустранен тест
H_0: β_1=0	H_0: β_1=0
H_a: β_1<0 (or H_a: β_1>0)	H_a: $\beta_1 \neq 0$

Критерий: *t*

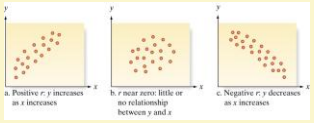
Използува се равнището p за проверка на хипотезата за значимост на регресионните коефициенти.

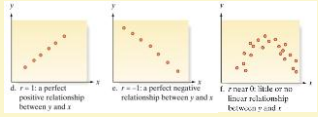
Корелационен коефициент

Измерва силата на зависимостта между x и y

$-1 \le r \le 1$

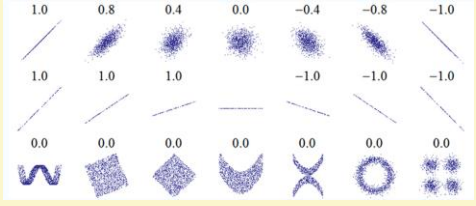
1. По абсолютна стойност колкото е по-близо до 1 толкова връзката е по-силна
2. Положителна и отрицателна корелация
   Ако r<0 връзката е обратна: Когато X нараства Y намалява
   Ако r>0 връзката е права: Когато X нараства Y нараства
3. Корелационният коефициент на Пирсън (Pearson) измерва само линейна корелация





13

Корелационен коефициент



14

Коефициент на детерминация

- Равен на квадрата на корелационния коефициент (няма знак)
  $0 \le r^2 \le 1$
- Интерпретация – измерва вариацията дължаща се или обяснена от модела
- Пример $r^2 = 0.80$
  Моделът обяснява 80% от цялата вариация на Y

15

Пример 1 продължение

```
Model 1: OLS, using observations 1-10
Dependent variable: Y

              coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       -3.28498      1.36964     -2.398    0.0433   **
  X            0.584249     0.0642378    9.095    1.72e-05 ***

Mean dependent var   8.400000   S.D. dependent var   4.765618
Sum squared resid    18.02454   S.E. of regression   1.501022
R-squared            0.911817   Adjusted R-squared   0.900794
F(1, 8)              82.72075   P-value(F)           0.000017
Log-likelihood      -17.13513   Akaike criterion     38.27026
Schwarz criterion    38.87543   Hannan-Quinn         37.60639
```

- Моделът е статистически значим p=0.000017 (обикновено α=0.05)
- Коефициент на детерминация R-squared=0.91
  91% от цялата вариация на Y е обяснена от модела (Браво на нас!)
- Регресионни коефициенти
- Интерпретация: β_0 : Когато X=0 кг. то Y= -3.28 лв (намаление от 3 лв и 28 ст.)
- Интерпретация: β_1 : Ако X се увеличи с 1 кг. то Y се увеличава с 0.58 лв (58 ст.)
- Корелационният коефициент на Пирсън $r = \sqrt{0.91} = +0.95$
  Знакът на корел. коефициент се взема от знака на регрес. коеф. +0.58

16

Пример 1 край

```
Model 1: OLS, using observations 1-10
Dependent variable: Y

              coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       -3.28498      1.36964     -2.398    0.0433   **
  X            0.584249     0.0642378    9.095    1.72e-05 ***

Mean dependent var   8.400000   S.D. dependent var   4.765618
Sum squared resid    18.02454   S.E. of regression   1.501022
R-squared            0.911817   Adjusted R-squared   0.900794
F(1, 8)              82.72075   P-value(F)           0.000017
Log-likelihood      -17.13513   Akaike criterion     38.27026
Schwarz criterion    38.87543   Hannan-Quinn         37.60639
```

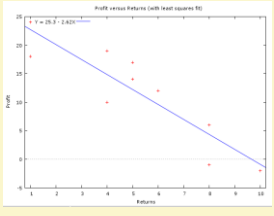
- Проверка на значимостта на регресионните коефициенти
  β_0 : p=0.0433 Означава статистически значим при α=0.05
  β_1 : p=0.0000172 Означава статистически значим при α=0.05

17

Пример 3.

- Y – Печалба (х.лв.)
- X – Рекламации, върнати продукти (%)
- Корелограма→

	Profit	Returns
1	24	1
2	18	1
3	17	5
4	19	4
5	14	5
6	12	6
7	10	4
8	6	8
9	−1	8
10	−2	10



18

Пример 3 продължение

Model 1: OLS, using observations 1-10
Dependent variable: Profit

	coefficient	std. error	t-ratio	p-value	
const	25.3299	2.68871	9.421	1.32e-05	***
Returns	-2.62113	0.455779	-5.751	0.0004	***

Mean dependent var	11.70000	S.D. dependent var	8.577101
Sum squared resid	128.9613	S.E. of regression	4.014993
R-squared	0.805224	Adjusted R-squared	0.780877
F(1, 8)	33.07277	P-value(F)	0.000429
Log-likelihood	-26.97402	Akaike criterion	57.94805
Schwarz criterion	58.55322	Hannan-Quinn	57.28418

- Моделът е статистически значим p=0.000429 (обикновено α=0.05)
- Коефициент на детерминация R-squared=0.81
  81% от цялата вариация на печалбата е обяснена от модела
- Регресионни коефициенти
- β_0 : Когато няма рекламации (=0%) то печалбата е 25.3 х.лв
- β_1 : Ако рекламациите се увеличат с 1% то печалбата намалява с 2.6 х.лв.
- Корелационният коефициент на Пирсън $r = \sqrt{0.81} = -0.9$
  Знакът на корел. коефициент се взема от знака на регрес. коеф. -2.62

19

Пример 3 край

Model 1: OLS, using observations 1-10
Dependent variable: Profit

	coefficient	std. error	t-ratio	p-value	
const	25.3299	2.68871	9.421	1.32e-05	***
Returns	-2.62113	0.455779	-5.751	0.0004	***

Mean dependent var	11.70000	S.D. dependent var	8.577101
Sum squared resid	128.9613	S.E. of regression	4.014993
R-squared	0.805224	Adjusted R-squared	0.780877
F(1, 8)	33.07277	P-value(F)	0.000429
Log-likelihood	-26.97402	Akaike criterion	57.94805
Schwarz criterion	58.55322	Hannan-Quinn	57.28418

- Проверка на значимостта на регресионните коефициенти
  β_0 : p=0.0000132 Означава статистически значим при α=0.05
  β_1 : p=0.0004 Означава статистически значим при α=0.05

20