Regression - Non parametric testing

May 22, 2019

Regression

Introduction

- Suppose there is an input x that takes values $\{x_1, \ldots, x_n\}$ (non random)
- and a random variable Y with samples $\{y_1, \ldots, y_n\}$
- Want to know if there is any relation between x and Y
- Especially if Y is a function of X.

Linear regression model

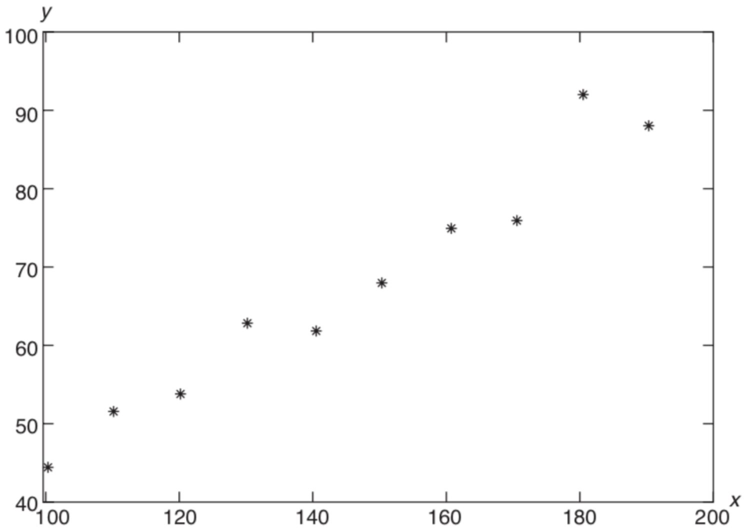
- Simplest case: Y is a linear function of x
- x is independent variable (non random)
- $Y = \alpha + \beta x + e$
- e is random error with mean 0

Example

Y: percent yield of a lab experiment

x: temperature of experiment

i	x_i	y_i	i	x_i	y_i
1	100	45	6	150	68
2	110	52	7	160	75
3	120	54	8	170	76
4	130	63	9	180	92
5	140	62	10	190	88


y
100
90
80
70
60
50
40
100
120
140
160
180
200
x

Least square estimators

- Want to estimate α and β from the data by A and B.
- The line $y = A + Bx$ should be "close" to the data
- For each x_i, the error is $Y_i - A - Bx_i$
- Minimize the sum of square errors

$$SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

Least square estimators

Solve the optimization problem

$$\frac{\partial SS}{\partial A} = 0; \quad \frac{\partial SS}{\partial B} = 0$$

Solution

- $B = \dfrac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$
- $A = \bar{Y} - B\bar{x}$

  where $\bar{Y} = \sum_{i=1}^{n} Y_i/n, \; \bar{x} = \sum_{i=1}^{n} x_i/n$

Better formula

$$S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

then

$$B = \frac{S_{xY}}{S_{xx}}$$

$$A = \bar{Y} - B\bar{x}$$

$$SS = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

Regression line

- The line $y = A + Bx$ is called the regression line
- Can use the line to predict the expected value of Y for some missing x_0

$$E(Y_0) \approx A + Bx_0$$

Example

Measurements of the relative humidity in the storage location and the moisture content of a sample of the raw material were taken over 15 days

Relative humidity	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture content	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

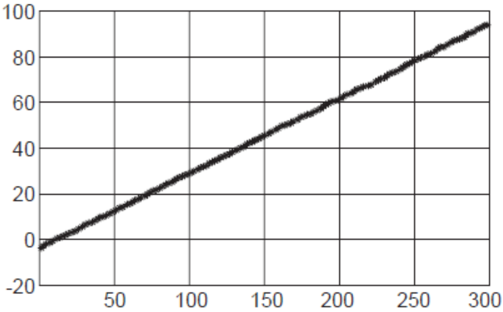
Result

The least squares estimators are as follows:

$a = -2.51$ Average x value = 46.13

$b = 0.32$ Sum of squares of the x values = 33212.0

The estimated regression line is $Y = -2.51 + 0.32x$



$S(x, Y) = 416.2$

$S(x, x) = 1287.73$

$S(Y, Y) = 147.6$

$SS_R = 13.08$

Example

EXAMPLE 9.3a The following data relate x, the moisture of a wet mix of a certain product, to Y, the density of the finished product.

x_i	y_i
5	7.4
6	9.3
7	10.6
10	15.4
x_i	y_i
12	18.1
15	22.2
18	24.1
20	24.8

Fit a linear curve to these data. Also determine SS_R.

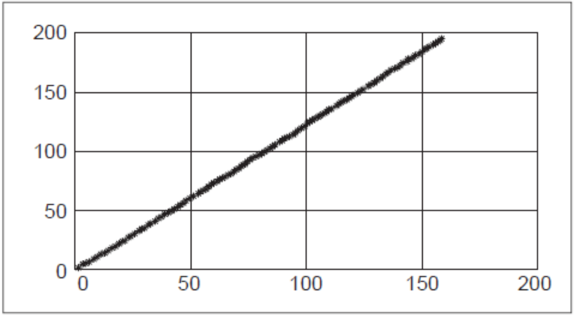
Result

The least squares estimators are as follows:

$a = 2.46$ Average x value = 11.63

$b = 1.21$ Sum of squares of the x values = 1303.0

The estimated regression line is $Y = 2.46 + 1.21x$



$S(x, Y) = 267.66$
$S(x, x) = 221.88$
$S(Y, Y) = 332.37$
$SS_R = 9.47$

Distribution of estimators

- Suppose $e \sim \mathcal{N}(0, \sigma^2)$, σ unknown
- then $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$
- $B = \dfrac{\sum_{i=1}^{n}(x_i - n\bar{x})Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$
  then B is also normal distributed
- $A = \bar{Y} - B\bar{x}$ is also normal

- $E(B) = \beta, \ \mathrm{Var}(B) = \dfrac{\sigma^2}{S_{xx}}$
- $B \sim \mathcal{N}(\beta, \dfrac{\sigma^2}{S_{xx}})$
- $E(A) = \alpha, \ \mathrm{Var}(A) = \dfrac{\sigma^2 \sum_i x_i^2}{n S_{xx}^2}$
- $A \sim \mathcal{N}(\alpha, \dfrac{\sigma^2 \sum_i x_i^2}{n S_{xx}^2})$

We can also prove that

$$\frac{SS}{\sigma^2} \sim \chi^2_{n-2}$$

and SS is independent of B and A,
so we can use SS to replace σ^2 in hypothesis test

Test about β

- Suppose want to show that Y has no relation with x
- $H_0 : \beta = 0$
- $H_1 : \beta \neq 0$
- Use $\dfrac{B - \beta}{\sqrt{\sigma^2 / S_{xx}}} \sim \mathcal{N}(0, 1)$
- but σ is unknown

Because $\dfrac{\mathcal{N}}{\sqrt{\frac{1}{n}\chi_n^2}} \sim t_n$

then

$$\frac{\sqrt{S_{xx}}(B-\beta)/\sigma}{\sqrt{\frac{SS}{\sigma^2(n-2)}}} = \sqrt{\frac{(n-2)S_{xx}}{SS}}(B-\beta) \sim t_{n-2}$$

If H_0 is true then $\beta = 0$ and

$$\sqrt{\frac{(n-2)S_{xx}}{SS}}B \sim t_{n-2}$$

The test

- Significance level γ
- Test statistic $T = \sqrt{\frac{(n-2)S_{xx}}{SS}}|B|$
- t is observed value of T

- p-value
  $= P(|T_{n-2}| > t) = 2P(T_{n-2} > t)$
- Reject H_0 if $\gamma \geq$ p-value
- or if $t > t_{\gamma/2,n-2}$

Example

A man claims that the fuel consumption of his car does not depend on how fast the car is driven. To test this hypothesis, the car was driven at various speeds between 45 and 70 miles per hour and the miles per gallon attained at each of these speeds was recorded. Is the claim correct?

Speed	Miles per Gallon
45	24.2
50	25.0
55	23.3
60	22.0
65	21.5
70	20.6
75	19.8

Solution

- Y: fuel consumption, x: speed
- $Y = \alpha + \beta x + e$
- $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

- $S_{xx} = 700, \ S_{YY} = 21.757, \ S_{xY} = -119$
- $B = -.17, SS = 1.527$
- $t = 8.139 > t_{0.005,5} = 4.032$
- Reject H_0 at level 1%
- Claim is not correct

Non parametric testing

Non parametric

- Want to test some hypothesis about 1 set of data
- Don't know about the distribution of data
- Make no assumption about the distribution
- No parameter !!!

The sign test

- Sample $X_, \ldots, X_n$ from a cdf F
- Want to test about the median m
- $H_0 : m = m_0$ vs $H_1 : m \neq m_0$

Convert to Binomial

- If H_0 is true then $F(m_0) = 1/2$
- Define

$$I_i = \begin{cases} 1 & \text{if } X_i < m_0 \\ 0 & \text{if } X_i \geq m_0 \end{cases}$$

- $T = \sum_{i=1}^{n} I_i \sim \text{ Bino}(n, 1/2)$

Test

- t is observed value of T
- Reject if $P(\text{ Bino}(n, 1/2) \leq t) < \alpha/2$ or $P(\text{ Bino}(n, 1/2) \geq t) < \alpha/2$
- p-value $= 2\min\{P(\text{ Bino}(n, 1/2) \leq t),$ $P(\text{ Bino}(n, 1/2) \geq t)\}$
- Reject H_0 if $\gamma \geq$ p-value

Example

If a sample of size 200 contains 120 values that are less than m_0 and 80 values that are greater, what is the p-value of the test of the hypothesis that the median is equal to m_0?

Solution

- $t = 80$
- p-value:

$$2P(\text{ Bino}(200, .5) \leq 80) = .00528$$

- Reject H_0 even at level 1%

2 sample sets

- Can use sign test to check if 2 sets of data come from the same distribution
- If so then the median of the difference is 0
- Do sign test on set $\{z_i = x_i - y_i\}$
- $H_0 : m = 0$ vs $H_1 : m \neq 0$.

Homework

Chapter 9: 5, 6, 11, 12

Chapter 11: 1, 2, 3