Serik Sagitov, Chalmers Tekniska Högskola, August 16, 2004

6. Simple linear regression

Relation between two continuous variables

X = explanatory variable, Y = dependent variable

data: n paired observations (x_i, y_i)

Ex 1: heights of fathers and sons

http://www.scc.ms.unimelb.edu.au/discday/dyk/faso.html

X = father's height, Y = son's height

6.1 Least square method

Random response to a known independent variable value

$Y = \beta_0 + \beta_1 x + \epsilon$

random noise $\epsilon \sim \mathrm{N}(0, \sigma^2)$ independent of x

model parameters: β_0, β_1, σ^2

Regression lines

unknown true line $y = \beta_0 + \beta_1 x$

fitted line $y = b_0 + b_1 x$ found from the data (x_i, y_i)

Responses

observed y_i and predicted $\hat{y}_i = b_0 + b_1 x_i$

Least square method leading to MLEs

find b_0 and b_1 by minimizing $\mathrm{SSE} = \Sigma(y_i - \hat{y}_i)^2$

Least square regression line $y = \bar{y} + r \cdot \frac{s_y}{s_x}(x - \bar{x})$

sample correlation coefficient $r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$

$s_x^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$, $s_y^2 = \frac{1}{n-1}\Sigma(y_i - \bar{y})^2$

Least square estimates

slope $b_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} = r \cdot \frac{s_y}{s_x}$

intercept $b_0 = \bar{y} - b_1\bar{x}$

In contrast to correlation coefficient r, regression coefficient b_1 is neither symmetric nor scale free

6.2 Variance estimation $\boxed{\text{SST} = \text{SSR} + \text{SSE}}$

Total sum of squares

$$\text{SST} = \Sigma(y_i - \bar{y})^2 = (n-1)s_y^2$$

Regression sum of squares

$$\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2 = (n-1)b_1^2 s_x^2$$

Error sum of squares

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2$$

$$\boxed{\text{Corrected MLE of } \sigma^2\text{: sample variance } s^2 = \frac{\text{SSE}}{n-2}}$$

Coefficient of determination $r^2 = \frac{\text{SSR}}{\text{SST}}$

proportion of variation in y_i explained by x_i variation

Ex 1: heights of fathers and sons

Point estimates in inches (1 inch = 2.54 cm)

$\bar{x} = 68$, $s_x = 2.7$, $\bar{y} = 69$, $s_y = 2.7$

Fitted regresion line $y = 35 + 0.5 \cdot x$

$r = b_1 \cdot \frac{s_x}{s_y} = 0.5$

coefficient of determination is 25%

6.3 CI and hypothesis testing

Estimates of β_0 and β_1 are unbiased and consistent

$b_1 \sim \mathrm{N}(\beta_1, \frac{\sigma_1^2}{n-1})$, $\sigma_1^2 = \sigma^2/s_x^2$

$b_0 \sim \mathrm{N}(\beta_0, \frac{\sigma_0^2}{n-1})$, $\sigma_0^2 = \sigma_1^2 \cdot \frac{1}{n} \Sigma\, x_i^2$

negative covariance $\mathrm{Cov}(b_0, b_1) = -\, \frac{\sigma^2 \cdot \bar{x}}{(n-1) \cdot s_x^2}$

Estimated standard errors

$s_{b_1} = \frac{s}{s_x \sqrt{n-1}}$, $s_{b_0} = s_{b_1} \sqrt{\frac{1}{n} \Sigma\, x_i^2}$

Exact 100(1–α)% CI for $\beta_i = b_i \pm t_{\alpha/2, n-2} \times s_{b_i}$

two t-distributions $\frac{b_0 - \beta_0}{s_{b_0}} \sim t_{n-2}$, $\frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}$

Hypothesis testing

test H_0: $\beta_1 = \beta_{10}$, using test statistic $T = \frac{b_1 - \beta_{10}}{s_{b_1}}$

null distribution $T \sim t_{n-2}$

Model utility test H_0: $\beta_1 = 0$ (no relationship) test statistic $T = b_1/s_{b_1}$, null distribution: $T \sim t_{n-2}$

Ex 1: heights of fathers and sons

$\mathrm{SST} = (n-1)s_y^2 = 7851$

$\mathrm{SSE} = \mathrm{SST}(1 - r^2) = 5888.5$

$s^2 = \frac{\mathrm{SSE}}{n-2} = 5.47$, $s = 2.34$

$s_{b_1} = \frac{s}{s_x \sqrt{n-1}} = 0.026$

99% CI for β_1 is

$0.5 \pm 2.58 \cdot 0.026 = 0.5 \pm 0.07$

model utility test: $T = \frac{b_1}{s_{b_1}} = 18.9$, reject H_0

6.4 Prediction interval

New observation of independent variable for a given x_{n+1}

$Y_{n+1} = \beta_0 + \beta_1 \cdot x_{n+1} + \epsilon_{n+1}$

Expected value of the new observation

true mean $\mu_{n+1} = \beta_0 + \beta_1 \cdot x_{n+1}$

estimated mean $\hat{\mu}_{n+1} = b_0 + b_1 \cdot x_{n+1}$

$\text{Var}(\hat{\mu}_{n+1}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n-1} \cdot \frac{(x_{n+1}-\bar{x})^2}{s_x^2}$

Estimated s.e. of $\hat{\mu}_{n+1}$: $s_{n+1} = s\sqrt{\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{(n-1)s_x^2}}$

Exact 100(1–α)% CI for the mean μ_{n+1}

$b_0 + b_1 \cdot x_{n+1} \pm t_{\alpha/2,n-2} \cdot s_{n+1}$

Exact 100(1–α)% prediction interval for Y_{n+1}

$$b_0 + b_1 \cdot x_{n+1} \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s_{n+1}^2}$$

Two sources of prediction uncertainty

$\text{Var}(Y_{n+1} - \hat{\mu}_{n+1}) = \text{Var}(\hat{\mu}_{n+1}) + \sigma^2$

Ex 2: my son's height

Estimated mean height of my son $\hat{\mu}_{n+1} = 35{+}0.5{\cdot}72 = 71$

estimated s.e. of $\hat{\mu}_{n+1}$: $s_{n+1} = 0.11$

95% CI for the mean height of my son $= 71 \pm 0.22$

95% PI for the height of my son is

71 ± 4.6 or between 169 cm and 192 cm

actual heights 68.9 (175 cm) and 71.6 (182 cm)

7. Chi-square tests

approximate tests for discrete and categorical data

7.1 Pearson's chi-square test: simple H_0

One sample from population distribution assigning probabilities $(p_1, \ldots, p_J)$ to j distinct values (cells)

Test a simple H_0 against complimentary H_1

H_0: $(p_1, \ldots, p_J) = (p_1^0, \ldots, p_J^0)$

H_1: $(p_1, \ldots, p_J) \neq (p_1^0, \ldots, p_J^0)$

Observed counts $(O_1, \ldots, O_J) \sim \mathrm{Mn}(n; p_1, \ldots, p_J)$

Chi-square test statistic: $X^2 = \Sigma_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j}$

expected counts $E_j = \mathrm{E}(O_j | H_0) = np_j^0$

Approximate null distribution of X^2 is χ^2_{J-1}

GLRT: reject H_0 for large values of $2\Delta \approx X^2$

Critical values for χ^2-distribution with df $= m$, $\alpha = 5\%$

m	2	3	4	5	10	20	30	60
$\chi^2_m(0.05)$	5.99	7.81	9.49	11.07	18.3	31.4	43.8	79.1

Ex 1: gender ratio

Saxony 1889: $n = 6115$ families with 12 children

data: $Y_1, \ldots, Y_n$ numbers of boys in each family

$J = 13$ cells, observed cell counts $O_1, \ldots, O_{13}$

Model M_1: number of boys in a family $Y \sim \mathrm{Bin}(12, 0.5)$

simple H_0: $p_j = \binom{12}{j-1} \cdot 2^{-12}$, $j = 1, \ldots, 13$

$X^2 = 249.2$, df $= 12$, $\chi^2_{12}(0.005) = 28.3$, reject H_0

y	cell j	O_j	E_j for M_1	$\frac{(O_j-E_j)^2}{E_j}$	E_j for M_2	$\frac{(O_j-E_j)^2}{E_j}$
0	1	7	1.5	20.2	2.3	9.6
1	2	45	17.9	41.0	26.1	13.7
2	3	181	98.5	69.1	132.8	17.5
3	4	478	328.4	68.1	410.0	11.3
4	5	829	739.0	11.0	854.2	0.7
5	6	1112	1182.4	4.2	1265.6	18.6
6	7	1343	1379.5	1.0	1367.3	0.4
7	8	1033	1182.4	18.9	1085.2	2.5
8	9	670	739.0	6.4	628.1	2.8
9	10	286	328.4	5.5	258.5	2.9
10	11	104	98.5	0.3	71.8	14.4
11	12	24	17.9	2.1	12.1	11.7
12	13	3	1.5	1.5	0.9	4.9

7.2 Pearson's chi-square test: composite H_0

Composite H_0: $(p_1, \ldots, p_J) = (p_1(\lambda), \ldots, p_J(\lambda))$
unknown parameter $\lambda = (\lambda_1, \ldots, \lambda_r)$, $\dim(\Omega_0) = r$

Expected cell counts
$E_j = n \cdot p_j(\hat{\lambda})$ with $\hat{\lambda}$ = MLE of λ under H_0

Approximate null distribution of X^2 is χ^2_{J-1-r}

$\mathrm{df}(X^2) = \#\{\text{cells}\} - \#\{\text{samples}\}$
$- \#\{\text{independent parameters estimated from the data}\}$

Ex 1: gender ratio

Test a more flexible model M_2: $Y \sim \mathrm{Bin}(12, p)$
composite H_0: $p_j = \binom{12}{j-1} \cdot p^{j-1} \cdot q^{13-j}$, $j = 1, \ldots, 13$

Expected cell counts for model M_2

$E_j = 6115 \cdot \binom{12}{j-1} \cdot \hat{p}^{j-1} \cdot \hat{q}^{13-j}$ based on MLE

$\hat{p} = \frac{\text{number of boys}}{\text{number of children}} = \frac{1\cdot 45 + 2\cdot 181 + \ldots + 12\cdot 3}{6115\cdot 12} = 0.4808$

Observed test statistic

$X^2 = 110.5$, df $= 11$, $\chi^2_{11}(0.005) = 26.76$

Reject H_0 at 0.5% level

observed variation is larger than expected

possible explanation : p differs from family to family

7.3 Chi-square test of independence

One sample cross-classified for two factors

observed counts $\|n_{jk}\| \sim \mathrm{Mn}(n_{..}; \|p_{jk}\|)$ matrix $J \times K$

marginal distributions $(p_{1\cdot}, \ldots, p_{J\cdot})$ and $(p_{\cdot 1}, \ldots, p_{\cdot K})$

Test of independence

H_0: $\|p_{jk}\| = \|p_{j\cdot} \times p_{\cdot k}\|$ (independence)

H_1: $\|p_{jk}\| \neq \|p_{j\cdot} \times p_{\cdot k}\|$ (dependence)

$$X^2 = \Sigma_{j=1}^{J} \Sigma_{k=1}^{K} \frac{(n_{jk} - E_{jk})^2}{E_{jk}}, \; E_{jk} = \frac{n_{j\cdot} \times n_{\cdot k}}{n_{..}}$$

$E_{jk} = n_{..}\hat{p}_{j\cdot}\hat{p}_{\cdot k}$ based on MLEs $\hat{p}_{j\cdot} = \frac{n_{j\cdot}}{n_{..}}$, $\hat{p}_{\cdot k} = \frac{n_{\cdot k}}{n_{..}}$

df $= JK - 1 - [(J-1) + (K-1)] = (J-1)(K-1)$

Chi-square test with df $= 1$

the approximate null distribution of $\sqrt{X^2}$ is N(0,1)

Ex 2: marital status and education

H_0: no relationship between
educational level and marital status of women

Contingency table of cross-classification:

Education	Married Once	Married ≥ 2	Total
College	550 (523.8)	61(87.2)	$n_{1\cdot}$=611
No College	681(707.2)	144(117.8)	$n_{2\cdot}$=825
Total	$n_{\cdot 1}$=1231	$n_{\cdot 2}$=205	$n_{\cdot\cdot}$=1436

$X^2 = 16.01$, df = 1, $\sqrt{16.01} = 4.001$, P < 0.1%
dependence: educated women marry smarter

7.4 Chi-square test of homogeneity

Data: K independent samples of sizes $n_{\cdot k}$, $k = 1, \ldots, K$
from K population distributions $(p_{1k}, \ldots, p_{Jk})$

Observed counts
$(n_{1k}, \ldots, n_{Jk}) \sim \text{Mn}(n_{\cdot k}; p_{1k}, \ldots, p_{Jk})$

Homogeneity means all K distributions are equal
H_0: $(p_{1k}, \ldots, p_{Jk}) = (p_{1l}, \ldots, p_{Jl})$ for all (k, l)
H_1: $p_{jk} \neq p_{jl}$ for some (j, k, l)

Single MLE for K parameters $p_{j1}, \ldots, p_{jK}$ under H_0
pooled sample proportion $\hat{p}_{jk} = n_{j\cdot}/n_{\cdot\cdot}$

The same X^2 and df as with independence test
expected cell counts $n_{\cdot k} \times \hat{p}_{jk} = n_{j\cdot} \times n_{\cdot k}/n_{\cdot\cdot}$
df $= JK - K - (J-1) = (J-1)(K-1)$

Ex 3: attitude toward small cars

Personality type:	Cautious	Midroad	Explorer	Total
Favorable	79(61.6)	58(62.2)	49(62.2)	186
Neutral	10(8.9)	8(9.0)	9(9.0)	27
Unfavorable	10(28.5)	34(28.8)	42(28.8)	86
Total	99	100	100	299

df $= (3-1)(3-1) = 4$, $\chi^2_{4,0.005} = 14.86$
$X^2 = 27.24$, reject H_0 at 0.5% level

Homogeneity =
equality of conditional distributions = independence

7.5 Grouping together small cells

Chi-square test is an approximate test
use (rather conservative) rule of thumb:
all expected counts E_j should not be less than 5
Combine small cells and
reduce the number of cells when calculating df

Ex 5: numerical example

Original data:	33(34.2)	27(25.8)
	17(16.0)	11(12.0)
	8(8.0)	6(6.0)
	3(2.9)	2(2.1)

Grouped data:	33(34.2)	27(25.8)
	17(16.0)	11(12.0)
	11(10.8)	8(8.2)

Grouped data calculation:
df $= (3-1)(2-1) = 2$, $\chi^2_{2,0.10} = 4.61$, $X^2 = 0.25$