Serik Sagitov, Chalmers Tekniska Högskola, September 12, 2005

Chapter 14. Linear least squares

X = independent variable assumed to be fixed

Y = dependent variable

1. Simple linear regression model

Random response

$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$

to a fixed value x_i of independent variable

Model assumption:

noise random variables ϵ_i are independent N(0,σ^2)

Unknown model parameters: β_0, β_1, σ^2

$L(\beta_0, \beta_1, \sigma^2) = \Pi_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}\}$

Least squares estimates

Regression lines:

true $y = \beta_0 + \beta_1 x$ and fitted $y = b_0 + b_1 x$

Responses:

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

Least squares method: minimize $S(b_0, b_1) = \Sigma(y_i - \hat{y}_i)^2$

solve $\partial S/\partial b_0 = 0$ and $\partial S/\partial b_1 = 0$

Normal equations:

$nb_0 + (\Sigma\, x_i)b_1 = \Sigma\, y_i$ and $(\Sigma\, x_i)b_0 + (\Sigma\, x_i^2)b_1 = \Sigma\, x_i y_i$

$$\boxed{\text{Slope } b_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}, \text{ intercept } b_0 = \bar{y} - b_1\bar{x}}$$

$\ln L(\beta_0, \beta_1, \sigma^2) = n\ln(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2} S(\beta_0, \beta_1)$, LSE = MLE

Least square regression line $y = \bar{y} + b_1(x - \bar{x})$

regression coefficient $b_1 = r \cdot \frac{s_y}{s_x}$ scale dependent

Sample correlation coefficient $r = \frac{s_{xy}}{s_x s_y}$

sample covariance $s_{xy} = \frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})$

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

Least square estimates are not robust against outliers

page 522: ouliers exert leverage on the fitted line

Sums of squares

SST = SSE + SSR

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

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

$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 = (n-1)s_y^2(1 - r^2)$ $\quad \text{df} = n - 2$

Corrected MLE of σ^2: $\quad s^2 = \frac{\text{SSE}}{n-2} = \frac{n-1}{n-2}\, s_y^2(1 - r^2)$

Coefficient of determination $r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$

proportion of variation in Y explained by factor X

2. CI and hypothesis testing

Unbiased and consistent estimates: $b_i \sim \text{N}(\beta_i, \sigma_i^2)$

$\sigma_0^2 = \frac{\sigma^2 \cdot \sum x_i^2}{n(n-1)s_x^2}$, $\sigma_1^2 = \frac{\sigma^2}{(n-1)s_x^2}$

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

Exact sampling distributions

$\frac{b_i - \beta_i}{s_{b_i}} \sim t_{n-2}$, $s_{b_0} = \frac{s}{s_x} \cdot \sqrt{\frac{\sum x_i^2}{n(n-1)}}$, $s_{b_1} = \frac{s}{s_x} \cdot \sqrt{\frac{1}{n-1}}$

Exact $100(1-\alpha)\%$ CI for β_i: $\quad b_i \pm t_{\alpha/2,n-2} \cdot s_{b_i}$

Hypothesis testing H_0: $\beta_1 = \beta_{10}$

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 between X and Y)

test statistic $T = b_1/s_{b_1}$, null distribution $T \sim t_{n-2}$

Zero intercept hypothesis

H_0: $\beta_0 = 0$

test statistic $T = b_0/s_{b_0}$, null distribution $T \sim t_{n-2}$

Intervals for individual observations

Given x predict $Y = \beta_0 + \beta_1 \cdot x + \epsilon$

expected value $\mu = \beta_0 + \beta_1 \cdot x$

least square estimate $\hat{\mu} = b_0 + b_1 \cdot x$

Standard error of $\hat{\mu}$

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

Exact $100(1-\alpha)\%$ CI for the mean μ

$b_0 + b_1 x \pm t_{\alpha/2,n-2} \cdot s\sqrt{\frac{1}{n} + \frac{1}{n-1}(\frac{x-\bar{x}}{s_x})^2}$

Exact $100(1-\alpha)\%$ prediction interval

$b_0 + b_1 x \pm t_{\alpha/2,n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{1}{n-1}(\frac{x-\bar{x}}{s_x})^2}$

The latter are wider limits since

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

Draw confidence bands around the regression line

both for the individual observation Y and its mean μ

3. Assessing the fit

Properties of the least square residuals $e_i = y_i - \hat{y}_i$

$e_1 + \ldots + e_n = 0$, $e_1^2 + \ldots + e_n^2$ at minimum

$x_1 e_1 + \ldots + x_n e_n = 0$, e_i are uncorrelated with x_i

$\hat{y}_1 e_1 + \ldots + \hat{y}_n e_n = 0$, e_i are uncorrelated with $\hat{y}_i$

Residual e_i has normal distribution with zero mean

$\mathrm{Var}(e_i) = \sigma^2(1 - \frac{\sum_k (x_k - x_i)^2}{n \sum (x_k - \bar{x})^2})$

$\mathrm{Cov}(e_i, e_j) = -\frac{\sum_k (x_k - x_i)(x_k - x_j)}{n \sum (x_k - \bar{x})^2}$

Standardized residuals $= e_i / s_{e_i}$, $s_{e_i} = s\sqrt{1 - \frac{\sum_k (x_k - x_i)^2}{n \sum (x_k - \bar{x})^2}}$

normal distribution plot to test normality assumption

Expected plot of the standardized residuals versus x_i:

horizontal blur (linearity)

variance does not depend on x (homoscedasticity)

Ex 1: flow rate vs stream depth

Page 517-518: scatter plot is slightly non-linear

residual plot has the U-shape

Page 518-519: scatter log-log plot is closer to linear

residual plot is horizontal

Ex 2: breast cancer

Page 520-521: absolute mortality y vs population size x

heteroscedastic residual plot

page 523: normal probability plot

Transformed variables: $\sqrt{y}$ vs $\sqrt{x}$

page 521: homoscedastic residual plot

page 524: normal probability plot is closer to linear

4. Multiple regression

Linear regression model

$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1} + \epsilon$, $\epsilon \sim \mathrm{N}(0,\sigma^2)$

n independent observations

$Y_1 = \beta_0 + \beta_1 x_{1,1} + \ldots + \beta_{p-1} x_{1,p-1} + \epsilon_1$

$\ldots$

$Y_n = \beta_0 + \beta_1 x_{n,1} + \ldots + \beta_{p-1} x_{n,p-1} + \epsilon_n$

Matrix notation $\mathbf{Y} = \mathbf{X}\beta$

$\mathbf{Y} = (y_1, \ldots, y_n)^T$

$\beta = (\beta_0, \ldots, \beta_{p-1})^T$, $\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p-1} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n,1} & \ldots & x_{n,p-1} \end{pmatrix}$

Least square estimates

$\mathbf{b} = (b_0, \ldots, b_{p-1})^T$

minimize $S(\mathbf{b}) = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$, where $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$

Normal equations $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y}$

if $\mathrm{rank}(\mathbf{X}) = p$, then $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

$$\boxed{\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}, \text{ where } \mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}$$

covariance matrix $\Sigma_{bb} = \|\mathrm{Cov}(b_i, b_j)\| = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

$s^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2/(n-p)$ unbiased estimate of σ^2

Standard errors

$s_{b_i} = s\sqrt{s_{ii}}$, where $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$

exact distributions $\frac{b_i - \beta_i}{s_{b_i}} \sim t_{n-p}$

Residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$

covariance matrix $\Sigma_{ee} = \|\mathrm{Cov}(e_i, e_j)\| = \sigma^2(\mathbf{I} - \mathbf{P})$

Standardized residuals $\frac{y_i - \hat{y}_i}{s\sqrt{1-p_{ii}}}$

Coefficient of multiple determination

$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$

$\mathrm{SSE} = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$, $\mathrm{SST} = (n-1)s_y^2$

Adjusted coefficient of multiple determination

$R_a^2 = 1 - \frac{n-1}{n-p} \cdot \frac{\mathrm{SSE}}{\mathrm{SST}}$

Ex 1: flow rate vs stream depth

Quadratic model $y = \beta_0 + \beta_1 x + \beta_2 x^2$

page 543: residuals shows no signs of systematic misfit

Linear and quadratic terms are stat. significant ($n = 10$)

Coefficient	Estimate	Standard Error	t Value
β_0	1.68	1.06	1.52
β_1	-10.86	4.52	-2.40
β_2	23.54	4.27	5.51

Emperical relationship developed in a region might break down if extrapolated to a wider region in which no data been observed

Ex 3: heart catheter

Catheter length depending on child's height and weight

page 546: pairwise scatterplots, $n = 12$

Two simple linear regressions

Estimate	Height	t Value	Weight	t Value
$b_0(s_{b_0})$	12.1(4.3)	2.8	25.6(2.0)	13.3
$b_1(s_{b_1})$	0.60(0.10)	6.0	0.28(0.04)	8.0
s	4.0		3.8	
$r^2(R_a^2)$	0.78 (0.76)		0.80 (0.78)	

page 547: plots of standardized residuals

Multiple regression model $L = \beta_0 + \beta_1 H + \beta_2 W$

$b_0 = 21$, $s_{b_0} = 8.8$, $b_0/s_{b_0} = 2.39$

$b_1 = 0.20$, $s_{b_1} = 0.36$,$b_1/s_{b_1} = 0.56$

$b_2 = 0.19$, $s_{b_2} = 0.17$, $b_2/s_{b_2} = 1.12$

$s = 3.9$, $R^2 = 0.81$, $R_a^2 = 0.77$

Can not reject neither $H_1 : \beta_1 = 0$ nor $H_2 : \beta_2 = 0$

β_1 = expected change in L

when H increased by one unit and W held constant

Height and weight are highly collinear

strong linear relationship

Fitted plane has a well resolved slope

along the line about which the (H, W) points fall

and poorly resolved slopes along the H and W axes

Page 549: stand. residuals from the multiple regression

little gain from adding W to the model with H