Serik Sagitov, Chalmers Tekniska Högskola, March 7, 2013

Chapter 14. Linear least squares

1 Simple linear regression model

A linear model for the random response $Y = Y(x)$ on an independent variable $X = x$. For a given set of values $(x_1, \dots, x_n)$ of the independent variable put

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

assuming that the noise $(\epsilon_1, \dots, \epsilon_n)$ has independent N(0,σ^2) random components. Given the data $(y_1, \dots, y_n)$, the model is characterized by the likelihood function

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

of three unknown model parameters β_0, β_1, σ^2. Summary statistics:
sample covariance $s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$,
sample variances $s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$, $s_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$,
sample correlation coefficient $r = \frac{s_{xy}}{s_x s_y}$.

Least squares estimates
Regression lines: true $y = \beta_0 + \beta_1 x$ and fitted $y = b_0 + b_1 x$. We want to find (b_0, b_1) such that the observed responses y_i are approximated by the predicted responses $\hat{y}_i = b_0 + b_1 x_i$ in an optimal way. Least squares method: find (b_0, b_1) minimizing the sum of squares $S(b_0, b_1) = \sum (y_i - \hat{y}_i)^2$.

From $\partial S/\partial b_0 = 0$ and $\partial S/\partial b_1 = 0$ we get the so-called Normal Equations:

$$\begin{cases} n b_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases} \quad \Rightarrow \quad \begin{cases} b_1 = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} = r \cdot \frac{s_y}{s_x} \\ b_0 = \bar{y} - b_1 \bar{x} \end{cases}$$

Observe that the least square estimates (b_0, b_1) are the maximum likelihood estimates of (β_0, β_1).

Least square regression line: for a given value x the predicted response is $\hat{y} = \bar{y} + r \frac{s_y}{s_x}(x - \bar{x})$.

Least square estimates are not robust against outliers: outliers exert leverage on the fitted line, p. 522.

Sums of squares SST = SSE + SSR

$\text{SST} = \sum (y_i - \bar{y})^2 = (n-1) s_y^2$	df $= n - 1$
$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2 = (n-1) b_1^2 s_x^2$	df $= 1$
$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = (n-1) s_y^2 (1 - r^2)$	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 $r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$ is the proportion of variation in Y explained by main factor X. The coefficient of determination r^2 has a more transparent meaning than correlation r.

2 Confidence intervals and hypothesis testing

Unbiased and consistent estimates: $b_0 \sim \mathrm{N}(\beta_0, \sigma_0^2)$, $\sigma_0^2 = \frac{\sigma^2 \cdot \sum x_i^2}{n(n-1)s_x^2}$; $b_1 \sim \mathrm{N}(\beta_1, \sigma_1^2)$, $\sigma_1^2 = \frac{\sigma^2}{(n-1)s_x^2}$.
Weak dependence between the two estimates $\mathrm{Cov}(b_0, b_1) = -\frac{\sigma^2 \cdot \bar{x}}{(n-1)s_x^2}$: negative, if $\bar{x} > 0$, and positive, if $\bar{x} < 0$. Exact sampling distributions

$$\frac{b_0 - \beta_0}{s_{b_0}} \sim t_{n-2}, \quad s_{b_0} = \frac{s\sqrt{\sum x_i^2}}{s_x\sqrt{n(n-1)}}, \qquad \frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}, \quad s_{b_1} = \frac{s}{s_x\sqrt{n-1}}$$

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

Hypothesis testing H_0: $\beta_1 = \beta_{10}$: test statistic $T = \frac{b_1 - \beta_{10}}{s_{b_1}}$, exact 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 the value y for the random variable $Y = \beta_0 + \beta_1 \cdot x + \epsilon$. Its expected value $\mu = \beta_0 + \beta_1 \cdot x$ has the least square estimate $\hat{\mu} = b_0 + b_1 \cdot x$. The standard error of $\hat{\mu}$ is computed as the square root of $\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n-1} \cdot (\frac{x-\bar{x}}{s_x})^2$.

Exact $100(1-\alpha)\%$ confidence interval 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 for y: $b_0 + b_1 x \pm t_{\alpha/2,n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{1}{n-1}(\frac{x-\bar{x}}{s_x})^2}$

Prediction interval has wider limits since $\mathrm{Var}(Y - \hat{\mu}) = \mathrm{Var}(\hat{\mu}) + \sigma^2 = \sigma^2(1 + \frac{1}{n} + \frac{1}{n-1} \cdot (\frac{x-\bar{x}}{s_x})^2)$.
To illustrate draw confidence bands around the regression line both for the individual observation y and the mean μ.

Assessing the fit
Properties of the least square residuals $e_i = y_i - \hat{y}_i$: $e_1^2 + \ldots + e_n^2$ is at minimum,
$e_1 + \ldots + e_n = 0$, $x_1 e_1 + \ldots + x_n e_n = 0$, $\hat{y}_1 e_1 + \ldots + \hat{y}_n e_n = 0$,
meaning that e_i are uncorrelated with x_i and e_i are uncorrelated with $\hat{y}_i$.
Residual e_i has normal distribution with zero mean and

$$\mathrm{Var}(e_i) = \sigma^2\Big(1 - \frac{\sum_k (x_k - x_i)^2}{n\sum(x_k - \bar{x})^2}\Big), \qquad \mathrm{Cov}(e_i, e_j) = -\frac{\sum_k (x_k - x_i)(x_k - x_j)}{n\sum(x_k - \bar{x})^2}$$

Standardized residuals $:= e_i/s_{e_i}$, where $s_{e_i} = s\sqrt{1 - \frac{\sum_k (x_k - x_i)^2}{n\sum(x_k - \bar{x})^2}}$. Use the normal distribution plot for standardized residuals to test normality assumption. Expected plot of the standardized residuals versus x_i: horizontal blur (linearity), variance does not depend on x (homoscedasticity)

Example: flow rate vs stream depth.
Page 517-518: the scatter plot is slightly non-linear. The residual plot has the U-shape. Page 518-519:

the scatter log-log plot is closer to linear and the residual plot is horizontal.

Example: breast cancer
Page 520-521: absolute mortality y vs population size x produces a heteroscedastic residual plot. Page 523: normal probability plot is not linear.
Transformed variables $\sqrt{y}$ vs $\sqrt{x}$: homoscedastic residual plot on page 521. Page 524: normal probability plot is closer to linear.

3 Multiple regression

Linear regression model $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1} + \epsilon$ with a homoscedastic noise $\epsilon \sim \mathrm{N}(0,\sigma^2)$. Data: observations $(y_1, \ldots, y_n)$ are realizations of n independent random variables

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

In the matrix notation the vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ is a realization of $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\mathbf{Y} = (Y_1, \ldots, Y_n)^T, \quad \boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^T, \quad \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T,$$

and $\mathbf{X}$ is the so called design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p-1} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n,1} & \ldots & x_{n,p-1} \end{pmatrix}.$$

Least square estimates $\mathbf{b} = (b_0, \ldots, b_{p-1})^T$ minimize $S(\mathbf{b}) = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$.

Normal equations $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$: if $\mathrm{rank}(\mathbf{X}) = p$, then $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Least squares multiple regression: predicted responses $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{P}\mathbf{y}$, where $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Covariance matrix for the least square estimates $\Sigma_{bb} = \Big(\mathrm{Cov}(b_i, b_j)\Big)_{i,j=0}^{p-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

An unbiased estimate of σ^2 is given by $s^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-p)$.

Standard errors $s_{b_i} = s\sqrt{s_{ii}}$, where s_{ii} are the diagonal elements of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$.

Exact sampling distributions $\frac{b_i - \beta_i}{s_{b_i}} \sim t_{n-p}$, $i = 1, \ldots, p-1$.

Residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{P})\mathbf{y}$ have a 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 $R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$, where $\mathrm{SSE} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$, $\mathrm{SST} = (n-1)s_y^2$. The problem with R^2 is that it increases even if irrelevant variables are added to the model.

Adjusted coefficient of multiple determination $R_a^2 = 1 - \frac{n-1}{n-p} \cdot \frac{\mathrm{SSE}}{\mathrm{SST}}$

is more appropriate as it punishes for irrelevant variables.

Example: 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 statistically 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

Example: 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)	12.8
$b_1(s_{b_1})$	0.60(0.10)	6.0	0.28(0.04)	7.0
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 $L = \beta_0 + \beta_1 H + \beta_2 W$ brings

$$\begin{array}{lll} b_0 = 21, & s_{b_0} = 8.8, & b_0/s_{b_0} = 2.39, \\ b_1 = 0.20, & s_{b_1} = 0.36, & b_1/s_{b_1} = 0.56, \\ b_2 = 0.19, & s_{b_2} = 0.17, & b_2/s_{b_2} = 1.12, \\ s = 3.9, & R^2 = 0.81, & R_a^2 = 0.77. \end{array}$$

Can not reject neither $H_1 : \beta_1 = 0$ nor $H_2 : \beta_2 = 0$. Different meaning of the slope parameters in the simple and multiple regression models. Here β_1 is the expected change in L when H increased by one unit and W held constant.

Collinearity problem: height and weight have a 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: standard residuals from the multiple regression. Conclusion: little or no gain from adding W to the simple regression model model with an independent variable H.