Serik Sagitov, Chalmers and GU, March 9, 2016

Chapter 14. Linear least squares

1 Simple linear regression model

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

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

assuming that the noise vector $(\epsilon_1, \ldots, \epsilon_n)$ has independent N(0,σ^2) random components. Given the data $(y_1, \ldots, y_n)$, the model is characterised by the likelihood function of three parameters β_0, β_1, σ^2

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

where $S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$. Observe that

$$n^{-1}S(\beta_0, \beta_1) = n^{-1}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \beta_0^2 + 2\beta_0\beta_1\bar{x} - 2\beta_0\bar{y} - 2\beta_1\overline{xy} + \beta_1^2\overline{x^2} + \overline{y^2}.$$

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) minimising 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:

$$\begin{cases} b_0 + b_1\bar{x} = \bar{y} \\ b_0\bar{x} + b_1\overline{x^2} = \overline{xy} \end{cases} \qquad \text{implying} \qquad \begin{cases} b_1 = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \dfrac{rs_y}{s_x} \\ b_0 = \bar{y} - b_1\bar{x} \end{cases}$$

The least square regression line $y = b_0 + b_1 x$ takes the form $y = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$.
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 covariance $s_{xy} = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$,
sample correlation coefficient $r = \frac{s_{xy}}{s_x s_y}$.
The least square estimates (b_0, b_1) are the maximum likelihood estimates of (β_0, β_1).
The least square estimates (b_0, b_1) are not robust: outliers exert leverage on the fitted line.

2 Residuals

The estimated regression line predicts the responses to the values of the explanatory variable by $\hat{y}_i = \bar{y} + r\frac{s_y}{s_x}(x_i - \bar{x})$. The noise in the observed responses y_i is represented by the residuals
$e_i = y_i - \hat{y}_i = y_i - \bar{y} - r\frac{s_y}{s_x}(x_i - \bar{x})$,
$e_1 + \ldots + e_n = 0, \qquad x_1e_1 + \ldots + x_ne_n = 0, \qquad \hat{y}_1e_1 + \ldots + \hat{y}_ne_n = 0.$

Residuals e_i have normal distributions with zero mean and

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

Error sum of squares

$$\mathrm{SSE} = \textstyle\sum_i e_i^2 = \sum_i (y_i - \bar{y})^2 - 2r\frac{s_y}{s_x} n(\overline{xy} - \bar{y}\bar{x}) + r^2\frac{s_y^2}{s_x^2}\sum_i (x_i - \bar{x})^2 = (n-1)s_y^2(1 - r^2).$$

$$\boxed{\text{Corrected maximum likelihood estimate of } \sigma^2: \quad s^2 = \tfrac{\mathrm{SSE}}{n-2} = \tfrac{n-1}{n-2}\, s_y^2(1 - r^2)}$$

Using $y_i - \bar{y} = \hat{y}_i - \bar{y} + e_i$ we obtain SST = SSR + SSE,

$\mathrm{SST} = \sum_i (y_i - \bar{y})^2 = (n-1)s_y^2$ is the total sum of squares,

$\mathrm{SSR} = \sum_i (\hat{y}_i - \bar{y})^2 = (n-1)b_1^2 s_x^2$ is the regression sum of squares.

$$\boxed{\text{Coefficient of determination } r^2 = \tfrac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \tfrac{\mathrm{SSE}}{\mathrm{SST}}.}$$

Coefficient of determination is the proportion of variation in Y explained by main factor X. Thus r^2 has a more transparent meaning than the correlation coefficient r.

To test the normality assumption use the normal distribution plot for the standardized residuals $\frac{e_i}{s_i}$, where $s_i = s\sqrt{1 - \frac{\sum_k (x_k - x_i)^2}{n(n-1)s_x^2}}$ are the estimated standard deviations of e_i.
The expected plot of the standardised residuals versus x_i is a horizontal blur (linearity), variance does not depend on x (homoscedasticity).

Example (flow rate vs stream depth)
For this example with $n = 10$, the scatter plot looks slightly non-linear. The residual plot gives a clearer picture having the U-shape. After the log-log transformation, the scatter plot is closer to linear and the residual plot has a horizontal profile.

3 Confidence intervals and hypothesis testing

The list square estimators (b_0, b_1) are unbiased and consistent. Due to the normality assumption we have the following exact distributions

$$b_0 \sim \mathrm{N}(\beta_0, \sigma_0^2), \qquad \sigma_0^2 = \tfrac{\sigma^2 \cdot \sum x_i^2}{n(n-1)s_x^2}, \qquad \tfrac{b_0 - \beta_0}{s_{b_0}} \sim t_{n-2}, \qquad s_{b_0} = \tfrac{s\sqrt{\sum x_i^2}}{s_x\sqrt{n(n-1)}},$$

$$b_1 \sim \mathrm{N}(\beta_1, \sigma_1^2), \qquad \sigma_1^2 = \tfrac{\sigma^2}{(n-1)s_x^2}, \qquad \tfrac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}, \qquad s_{b_1} = \tfrac{s}{s_x\sqrt{n-1}}.$$

Weak dependence between the two estimators: $\mathrm{Cov}(b_0, b_1) = -\frac{\sigma^2 \cdot \bar{x}}{(n-1)s_x^2}$.

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

Hypothesis testing H_0: $\beta_i = \beta_{i0}$: test statistic $T = \frac{b_i - \beta_{i0}}{s_{b_i}}$, exact null distribution $T \sim t_{n-2}$.
Model utility test and zero-intercept test

H_0: $\beta_1 = 0$ (no relationship between X and Y), test statistic $T = b_1/s_{b_1}$, null distribution $T \sim t_{n-2}$.

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

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

Prediction interval has wider limits $\mathrm{Var}(Y - \hat{\mu}) = \sigma^2 + \mathrm{Var}(\hat{\mu}) = \sigma^2(1 + \frac{1}{n} + \frac{1}{n-1} \cdot (\frac{x-\bar{x}}{s_x})^2)$, since it contains the uncertainty due the noise factors.
Compare these two formulas by drawing the confidence bands around the regression line both for the individual observation y and the mean μ.

4 Linear regression and ANOVA

Recall the two independent samples case from Chapter 11:
first sample $\mu_1 + \epsilon_1, \ldots, \mu_1 + \epsilon_n$,
second sample $\mu_2 + \epsilon_{n+1}, \ldots, \mu_2 + \epsilon_{n+m}$,
where the noise variables are independent and identically distributed $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. This setting is equivalent to the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad x_1 = \ldots = x_n = 0, \quad x_{n+1} = \ldots = x_{n+m} = 1,$$

with

$$\mu_1 = \beta_0, \quad \mu_2 = \beta_0 + \beta_1.$$

The model utility test $H_0 : \beta_1 = 0$ is equivalent to the equality test $H_0 : \mu_1 = \mu_2$.

More generally, for the one-way ANOVA setting with $I = p$ levels for the main factor and $n = pJ$ observations

$$\begin{aligned} \beta_0 + \epsilon_i, &\quad i = 1, \ldots, J, \\ \beta_0 + \beta_1 + \epsilon_i, &\quad i = J+1, \ldots, 2J, \\ \ldots & \\ \beta_0 + \beta_{p-1} + \epsilon_i, &\quad i = (p-1)J + 1, \ldots, n, \end{aligned}$$

we need a multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i, \quad i = 1, \ldots, n$$

with dummy variables $x_{i,j}$ taking values 0 and 1 so that

$$\begin{aligned} x_{i,1} &= 1 \text{ only for } i = J+1, \ldots 2J, \\ x_{i,2} &= 1 \text{ only for } i = 2J+1, \ldots 3J, \\ &\ldots \\ x_{i,p-1} &= 1 \text{ only for } i = (p-1)J + 1, \ldots, n. \end{aligned}$$

5 Multiple linear regression

Consider a linear regression model

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

with $p-1$ explanatory variables and a homoscedastic noise. This is an extension of the simple linear regression model with $p = 2$.

The corresponding data set consists of observations $(y_1, \ldots, y_n)$ with $n > p$, which are realisations of n independent random variables

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

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

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

are column vectors, and $\mathbf{X}$ is the so called design matrix

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

assumed to have rank p. Least square estimates $\mathbf{b} = (b_0, \ldots, b_{p-1})^T$ minimise $S(\mathbf{b}) = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$, where $\|\mathbf{a}\|$ is the length of a vector $\mathbf{a}$. Solving the normal equations $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$ we find the least squares estimates being

$$\mathbf{b} = \mathbf{M}\mathbf{X}^T\mathbf{y}, \quad \mathbf{M} = (\mathbf{X}^T\mathbf{X})^{-1}.$$

$$\boxed{\text{Least squares multiple regression: predicted responses } \hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{P}\mathbf{y}, \text{ where } \mathbf{P} = \mathbf{X}\mathbf{M}\mathbf{X}^T.}$$

Covariance matrix for the least square estimates $\Sigma_{bb} = \sigma^2\mathbf{M}$ is a $p \times p$ matrix with elements $\mathrm{Cov}(b_i, b_j)$. The vector of residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{P})\mathbf{y}$ have a covariance matrix $\Sigma_{ee} = \sigma^2(\mathbf{I} - \mathbf{P})$.

$$\boxed{\text{An unbiased estimate of } \sigma^2 \text{ is given by } s^2 = \tfrac{\mathrm{SSE}}{n-p}, \quad \text{where SSE} = \|\mathbf{e}\|^2.}$$

The standard error of b_i is computed as $s_{b_j} = s\sqrt{m_{jj}}$, where m_{jj} is a diagonal element of $\mathbf{M}$.

$$\boxed{\text{Exact sampling distributions } \tfrac{b_j - \beta_j}{s_{b_j}} \sim t_{n-p}, \quad j = 1, \ldots, p-1.}$$

Inspect the normal probability plot for the standardised residuals $\frac{y_i - \hat{y}_i}{s\sqrt{1-p_{ii}}}$, where p_{ii} are the diagonal elements of $\mathbf{P}$.

Coefficient of multiple determination can be computed similarly to the simple linear regression model as $R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$, where $\mathrm{SST} = (n-1)s_y^2$. The problem with R^2 is that it increases even if irrelevant variables are added to the model. To punish for irrelevant variables it is better to use the adjusted coefficient of multiple determination

$$\boxed{R_a^2 = 1 - \tfrac{n-1}{n-p} \cdot \tfrac{\text{SSE}}{\text{SST}} = 1 - \tfrac{s^2}{s_y^2}.}$$

The adjustment factor $\frac{n-1}{n-p}$ gets larger for the larger values of p.

Example (flow rate vs stream depth)
The multiple linear regression framework works for the quadratic model $y = \beta_0 + \beta_1 x + \beta_2 x^2$. The residuals show no sign of systematic misfit. Linear and quadratic terms are statistically significant

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 (catheter length)
Doctors want predictions on heart catheter length depending on child's height and weight. The pairwise scatterplots for the data of size $n = 12$ suggests 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)	

The plots of standardised residuals do not contradict the normality assumptions.

The simple regression models should be compared to the multiple regression model $L = \beta_0 + \beta_1 H + \beta_2 W$, which gives

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

In contrast to the simple models, we can not reject neither $H_1 : \beta_1 = 0$ nor $H_2 : \beta_2 = 0$. This paradox is explained by different meaning of the slope parameters in the simple and multiple regression models. In the multiple model 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. The 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.

Conclusion: since the simple model $L = \beta_0 + \beta_1 W$ gives the highest adjusted coefficient of determination, there is little or no gain from adding H to the regression model model with a single explanatory variable W.