MSc 1 DS&AI

Statistical inference part 2

Practice session 2: Simple and multiple linear regression

Exercises on simple linear regression

6.2 **(a)** Show that $E(\hat{\beta}_1) = \beta_1$ as in (6.7).

(b) Show that $E(\hat{\beta}_0) = \beta_0$ as in (6.8).

with $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

and $E[y_i] = \beta_0 + \beta_1 x_i$

6.9 **(a)** Obtain a test for $H_0 : \beta_0 = a$ versus $H_1 : \beta_0 \neq a$.

(b) Obtain a confidence interval for β_0.

Use $\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \overbrace{\sigma^2\left[1/n + \bar{x}^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2\right]}^{c}\right)$, $(n-2)\dfrac{s^2}{\sigma^2} \sim \chi^2_{(n-2)}$

and $\hat{\beta}_0$ and s^2 independent. Thus deduce the distribution of $\dfrac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{c}}$ ...

6.14 Table 6.1 (Weisberg 1985, p. 231) gives the data on daytime eruptions of Old Faithful Geyser in Yellowstone National Park during August 1–4, 1978. The variables are $x =$ duration of an eruption and $y =$ interval to the next eruption. Can x be used to successfully predict y using a simple linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$? see file geyser.csv

(a) Find $\hat{\beta}_0$ and $\hat{\beta}_1$.

(b) Test $H_0 : \beta_1 = 0$ using (6.14).

(c) Find a confidence interval for β_1.

(d) Find r^2 using (6.16).

6.2

(a) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\,,$$ we have $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\left(\beta_1(x_i-\bar{x}) + \epsilon_i - \bar{\epsilon}\right)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\epsilon_i-\bar{\epsilon})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$E[\hat{\beta}_1] = \beta_1 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})\left(\underbrace{E(\epsilon_i)}_{0} - \underbrace{E(\bar{\epsilon})}_{0}\right)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$ by linéarity of the expectation

(b) $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

$$E[\hat{\beta}_0] = E[\bar{y}] - E[\hat{\beta}_1]\cdot\bar{x} = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

Thus $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased

6.9

(a) $\dfrac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{c}} \sim \mathcal{N}(0,1)$ and $(m-2)\dfrac{s^2}{\sigma^2} \sim \chi^2_{(m-2)}$

and $\hat{\beta}_0 \perp s^2$ thus $\dfrac{\hat{\beta}_0 - \beta_0}{s\sqrt{c}} \sim t_{(m-2)}$ with $c = \dfrac{1}{m} + \dfrac{\bar{x}^2}{\sum_{i=1}^{m}(x_i - \bar{x})^2}$

Under $H_0 : \beta_0 = a$, $t := \dfrac{\hat{\beta}_0 - a}{s\sqrt{c}} \sim t_{(m-2)}$

Thus we reject H_0 at risk α if $|t| > t_{\alpha/2, m-2}$

(b) $\dfrac{\hat{\beta}_0 - \beta_0}{s\sqrt{c}} \sim t_{(m-2)}$ thus

$$P\left(-t_{\alpha/2, m-2} \leq \frac{\hat{\beta}_0 - \beta_0}{s\sqrt{c}} \leq t_{\alpha/2, m-2}\right) = 1-\alpha$$

$$\iff P\left(\hat{\beta}_0 - t_{\alpha/2, m-2}\, s\sqrt{c} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2, m-2}\, s\sqrt{c}\right) = 1-\alpha$$

Thus an $1-\alpha$ confidence interval for β_0 is

$$\boxed{\hat{\beta}_0 \pm t_{\alpha/2, m-2}\, s\sqrt{\frac{1}{m} + \frac{\bar{x}^2}{\sum_{i=1}^{m}(x_i - \bar{x})^2}}}$$

6.14 : See Rmd file.

Exercises on multiple linear regression

Proove (i) of theorem 7.6b

Theorem 7.6b. Suppose that $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is $n \times (k+1)$ of rank $k + 1 < n$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$. Then the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ given in Theorem 7.6a have the following distributional properties:

(i) $\hat{\boldsymbol{\beta}}$ is $N_{k+1}[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$.

Thus give $\mathrm{cov}(\hat{\beta})$

use that $\hat{\beta} = (X'X)^{-1} X y$

7.2 Show that (7.10) follows from (7.9). Why is $\mathbf{X}'\mathbf{X}$ positive definite, as noted below (7.10)?

7.29 (**a**) Show that R^2 in (7.55) can be written in the form $R^2 = 1 - \mathrm{SSE}/\sum_i (y_i - \bar{y})^2$.

(**b**) Replace SSE and $\sum_i (y_i - \bar{y})^2$ in part (a) by variance estimators $\mathrm{SSE}/(n-k-1)$ and $\sum_i (y_i - \bar{y})^2/(n-1)$ and show that the result is the same as R_a^2 in (7.56).

7.53 When gasoline is pumped into the tank of a car, vapors are vented into the atmosphere. An experiment was conducted to determine whether y, the amount of vapor, can be predicted using the following four variables based on initial conditions of the tank and the dispensed gasoline:

data gas.csv

$$x_1 = \text{tank temperature } (^\circ\text{F})$$
$$x_2 = \text{gasoline temperature } (^\circ\text{F})$$
$$x_3 = \text{vapor pressure in tank (psi)}$$
$$x_4 = \text{vapor pressure of gasoline (psi)}$$

The data are given in Table 7.3 (Weisberg 1985, p. 138).

(**a**) Find $\hat{\boldsymbol{\beta}}$ and s^2.

(**b**) Find an estimate of $\mathrm{cov}(\hat{\boldsymbol{\beta}})$.

~~(**c**) Find $\hat{\boldsymbol{\beta}}_1$ and $\hat{\beta}_0$ using $\mathbf{S}_{xx}$ and $\mathbf{s}_{yx}$ as in (7.46) and (7.47).~~

(**d**) Find R^2 and R_a^2.

Proof of thm 7.6

$$y \sim \mathcal{N}_n(X\beta, \sigma^2 I)$$

$$\hat{\beta} = (X'X)^{-1}X'y = Ay \quad \text{with } A = (X'X)^{-1}X'$$

Thus $\hat{\beta} \sim \mathcal{N}_{k+1}\Big(\underbrace{AX\beta}_{\underbrace{(X'X)^{-1}X'X}_{I}},\ \underbrace{A\sigma^2 I A'}_{\underbrace{\sigma^2 (X'X)^{-1}X'X(X'X)^{-1}}_{\sigma^2(X'X)^{-1}}}\Big)$ due to the properties of linear combinations of Gaussian vectors

$$\Rightarrow \quad \boxed{\hat{\beta} \sim \mathcal{N}_{k+1}\left(\beta, \sigma^2 (X'X)^{-1}\right)}$$

$\hat{\beta}$ follows a multivariate normal distribution centered in β and with variance-covariance matrix $\sigma^2 (X'X)^{-1}$

unfortunatly σ^2 is unknown ..., but it can be estimated by $s^2 = \frac{1}{n-k-1}(y - X\hat{\beta})'(y - X\hat{\beta})$

And $\boxed{(n-k-1)\frac{s^2}{\sigma^2} \sim \chi^2_{(n-k-1)}}$ and $\hat{\beta}$ and s^2 independant. This will be usefull to build tests and confidence intervals on β.

(7.2)

$$(y - Xb)'(y - Xb) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\mathbf{b}) \tag{7.9}$$

$$= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})$$

$$+ 2(\hat{\boldsymbol{\beta}} - \mathbf{b})'(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}). \tag{7.10}$$

From (7.9) we get

$$(y - X\hat{\beta})'(y - X\hat{\beta}) + (y - X\hat{\beta})'(\underbrace{X\hat{\beta} - Xb}_{X(\hat{\beta} - b)}) + (X\hat{\beta} - Xb)'(y - X\hat{\beta}) + (X\hat{\beta} - Xb)'(X\hat{\beta} - Xb)$$

$$= (y - X\hat{\beta})'(y - X\hat{\beta}) + 2(\hat{\beta} - b)'(X'y - X'X\hat{\beta}) + (\hat{\beta} - b)'X'X(\hat{\beta} - b)$$

Let $u \in \mathbb{R}^{k+1}$ $\quad u'X'Xu = v'v$ with $v = Xu$

$$= \|v\|^2$$

if $u \neq 0$ then $v \neq 0$ since X is of rank $k+1$

(if there exist $u \neq 0$ such $Xu = 0$ this would mean that a column of X would be a combination of the other which would contradict the assumption that X is of rank $k+1$)

Thus if $u \neq 0$ $\quad u'X'Xu = \|v\|^2 > 0$ with $v = Xu$

thus $X'X$ is a positive definite matrix.

(7.29)

$$R^2 = \frac{SSR}{SST}$$

with $SSR = \sum_{i=1}^{m} (\hat{y}_i - \bar{y})^2$ and $SST = \sum_{i=1}^{m} (y_i - \bar{y})^2$

$$SSE = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 .$$

Thus we need to show that $SST = SSR + SSE$ to deduce that

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

see solution p. 181

7.7 R^2 IN FIXED-x REGRESSION

In (7.39), we have $\text{SSE} = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\boldsymbol{\beta}}_1' \mathbf{X}_c' \mathbf{y}$. Thus the corrected total sum of squares $\text{SST} = \sum_i (y_i - \bar{y})^2$ can be partitioned as

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \hat{\boldsymbol{\beta}}_1' \mathbf{X}_c' \mathbf{y} + \text{SSE}, \tag{7.53}$$

$$\text{SST} = \text{SSR} + \text{SSE},$$

where $\text{SSR} = \hat{\boldsymbol{\beta}}_1' \mathbf{X}_c' \mathbf{y}$ is the *regression sum of squares*. From (7.37), we obtain $\mathbf{X}_c' \mathbf{y} = \mathbf{X}_c' \mathbf{X}_c \hat{\boldsymbol{\beta}}_1$, and multiplying this by $\hat{\boldsymbol{\beta}}_1'$ gives $\hat{\boldsymbol{\beta}}_1' \mathbf{X}_c' \mathbf{y} = \hat{\boldsymbol{\beta}}_1' \mathbf{X}_c' \mathbf{X}_c \hat{\boldsymbol{\beta}}_1$. Then $\text{SSR} = \hat{\boldsymbol{\beta}}_1' \mathbf{X}_c' \mathbf{y}$ can be written as

$$\text{SSR} = \hat{\boldsymbol{\beta}}_1' \mathbf{X}_c' \mathbf{X}_c \hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_c \hat{\boldsymbol{\beta}}_1)' (\mathbf{X}_c \hat{\boldsymbol{\beta}}_1). \tag{7.54}$$

In this form, it is clear that SSR is due to $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_k)'$.

The proportion of the total sum of squares due to regression is

$$\boxed{R^2 = \frac{\hat{\boldsymbol{\beta}}_1' \mathbf{X}_c' \mathbf{X}_c \hat{\boldsymbol{\beta}}_1}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\text{SSR}}{\text{SST}},} \tag{7.55}$$

which is known as the *coefficient of determination* or the *squared multiple correlation*. The ratio in (7.55) is a measure of model fit and provides an indication of how well the x's predict y.

The partitioning in (7.53) can be rewritten as the identity

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2 = (\hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{y} - n\bar{y}^2) + (\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{y})$$

$$= \text{SSR} + \text{SSE},$$

which leads to an alternative expression for R^2:

$$R^2 = \frac{\hat{\boldsymbol{\beta}}' \mathbf{X}' \mathbf{y} - n\bar{y}^2}{\mathbf{y}'\mathbf{y} - n\bar{y}^2}. \tag{7.56}$$

7.53 see the Rmd file.