Multiple regression

- Multiple regression. Suppose that $\{Y_i, X_{i,1}, \ldots, X_{i,k}\}_{i=1}^n$ are n observations for $(Y, X_1, \ldots, X_k)$ and

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} + \varepsilon_i \tag{1}$$

for $i = 1, \ldots, n$.

- Matrix expression of (1). let Y be the n vector whose i-th element is Y_i. Let X be the $n \times k$ matrix whose (i, j)-th element is $X_{i,j}$ and let ε be the n vector whose i-th element is ε_i. Then (1) can be expressed as

$$\mathrm{Y} = (\mathbf{1} \;\; \mathrm{X})\boldsymbol{\beta} + \varepsilon,$$

where $\mathbf{1}$ is a $n \times 1$ vector of 1's, and $\boldsymbol{\beta}$ is the $(k+1) \times 1$ vector $(\beta_0, \beta_1, \cdots, \beta_k)^T$.

- Assumptions.

  A1. ε_i's are IID $N(0, \sigma_\varepsilon^2)$.

  A2. ε and X are independent.

  A3. $(\mathbf{1} \;\; \mathrm{X})$ is of rank $k + 1$.

  Note: Assumption A3 ensures that we can estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$.

- Least square estimation. Let X^* be the $n \times (k+1)$ matrix $(\mathbf{1} \;\; \mathrm{X})$ and for $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_k)^T$, let

$$S(\boldsymbol{\gamma}) = (\mathrm{Y} - \mathrm{X}^*\boldsymbol{\gamma})^T(\mathrm{Y} - \mathrm{X}^*\boldsymbol{\gamma}) = \sum_{i=1}^n (Y_i - \gamma_0 - \gamma_1 X_{1,i} - \cdots - \gamma_k X_{k,i})^2.$$

Then $S(\boldsymbol{\gamma})$ is minimized when $\boldsymbol{\gamma} = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ can be found using the equation

$$(\mathrm{X}^*)^T(\mathrm{Y} - \mathrm{X}^*\hat{\boldsymbol{\beta}}) = 0,$$

or

$$\hat{\boldsymbol{\beta}} = ((\mathrm{X}^*)^T\mathrm{X}^*)^{-1}(\mathrm{X}^*)^T\mathrm{Y}. \tag{2}$$

$\hat{\boldsymbol{\beta}}$ is called the least square estimator of β. Let

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^T = \hat{\boldsymbol{\beta}},$$

Then $\hat{\beta}_j$ is the least square estimator of β_j for $j = 0, \ldots, k$, and

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$$

is called the estimated regression equation.

- The distribution of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_k)^T$ is the multivariate normal distribution

$$N\left(\boldsymbol{\beta}, \sigma_\varepsilon^2((\mathrm{X}^*)^T\mathrm{X}^*)^{-1}\right),$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^T = (E(\hat{\beta}_0), \ldots, E(\hat{\beta}_k))^T$ and $\sigma_\varepsilon^2((\mathrm{X}^*)^T\mathrm{X}^*)^{-1}$ is the covariance matrix of $\hat{\boldsymbol{\beta}}$.

- The distribution for $\mathrm{a}^T\hat{\beta} = a_0\hat{\beta}_0 + \ldots + a_k\hat{\beta}_k$. Suppose that $a_0, \ldots, a_k$ are known constants. Let a be the $(k+1) \times 1$ vector $(a_0, \ldots, a_k)^T$ and

$$V = \mathrm{a}^T((\mathrm{X}^*)^T\mathrm{X}^*)^{-1}\mathrm{a}, \tag{3}$$

then $Var(\mathrm{a}^T\hat{\beta}) = \sigma_\varepsilon^2 V$ and

$$\frac{a_0\hat{\beta}_0 + \ldots + a_k\hat{\beta}_k - (a_0\beta_0 + \ldots + a_k\beta_k)}{\sqrt{\sigma_\varepsilon^2 V}} \sim N(0,1).$$

- The residual sum of squares and the estimation of σ_ε^2. The residual sum of squares, denoted by RSS (or SSE), is defined as

$$RSS = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \cdots - \hat{\beta}_k X_{k,i})^2.$$

It can be shown that

$$\frac{RSS}{\sigma_\varepsilon^2} \sim \chi^2(n-k-1),$$

so we use

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{RSS}{n-k-1}}$$

to estimate σ_ε.

- Constructing confidence interval for $a_0\beta_0 + \ldots + a_k\beta_k$. Let V be the variable in (3) and let $\hat{\sigma}_\varepsilon = \sqrt{RSS/(n-k-1)}$. Then $\hat{\sigma}_\varepsilon$ and $\hat{\boldsymbol{\beta}}$ are independent and we have

$$\frac{a_0\hat{\beta}_0 + \ldots + a_k\hat{\beta}_k - (a_0\beta_0 + \ldots + a_k\beta_k)}{\sqrt{\hat{\sigma}_\varepsilon^2 V}} \sim t(n-k-1).$$

Thus a $(1-\alpha)$ C.I. for $a_0\beta_0 + \ldots + a_k\beta_k$ is given by

$$a_0\hat{\beta}_0 + \ldots + a_k\hat{\beta}_k \pm t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}_\varepsilon^2 V}.$$

- Hypothesis testing. Suppose that a_0, ..., a_k are constants. Consider the testing problem

$$H_0 : a_0\beta_0 + \ldots + a_k\beta_k = 0 \text{ versus } H_1 : a_0\beta_0 + \ldots + a_k\beta_k \neq 0.$$

Let

$$T = \frac{a_0\hat{\beta}_0 + \ldots + a_k\hat{\beta}_k}{\sqrt{\hat{\sigma}_\varepsilon^2 V}},$$

where V is given in (3). Then the test that rejects H_0 when $|T| > t_{\alpha/2,n-k-1}$ is of level α. In particular, we can test $H_0 : \beta_j = 0$ by taking $a_i = 0$ for $i \neq j$ and $a_j = 1$.

- The multiple R^2 (coefficient of multiple determination). The multiple R^2 is defined as

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{RSS}{SS_{\text{total}}} = \frac{SS_{\text{reg}}}{SS_{\text{total}}},$$

where

  - $SS_{\text{total}} = \sum_{i=1}^n (Y_i - \bar{Y})^2$ is called the total sum of squares, and
  - $SS_{\text{reg}} = SS_{\text{total}} - RSS$ is called the regression sum of squares.

It can be shown that $RSS \leq SS_{\text{total}}$, so $0 \leq R^2 \leq 1$.

- Interpretation of R^2. R^2 is interpreted as the proportion of the variation of Y that is explained by X_1, ..., X_k. To see this, suppose that in (1), $(X_{1,i}, \ldots, X_{k,i}, Y_i)$: $i = 1$, ..., n are IID observations from the distribution of $(X_1, \ldots, X_k, Y)$. By Assumptions A1-A2, we have

$$Var(Y) = Var(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) + \sigma_\varepsilon^2.$$

The proportion of variation in Y explained by X_1, ..., X_k is

$$\frac{Var(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{Var(Y)} = 1 - \frac{\sigma_\varepsilon^2}{Var(Y)} \approx 1 - \frac{RSS}{SS_{\text{total}}} = R^2$$

when n is large.

Therefore, we use R^2 to represent the proportion of variation in Y that can be explained by X_1, ..., X_k.

- Adjusted R^2. The adjusted R^2, denoted by R^2_{adj}, is the quantity

$$1 - \frac{RSS/(n-k-1)}{SS_{\text{total}}/(n-1)}.$$

R^2_{adj} is used because it better approximates the quantity $1 - \sigma_\varepsilon^2/Var(Y)$ than R^2 when $n - k - 1$ is large but k is not small comparing to n.

- R^2_{adj} 不一定 ≥ 0, 所以我們通常用R^2代表Y的變動中可由X_1, ..., X_k 解釋的比例. 然而當$n-k-1$ 大而且k/n沒有很小時, 我們會看 R^2_{adj} 來了解模型的解釋度.

- The global test. The global test is for testing

  $$H_0 : \beta_1 = \cdots = \beta_k = 0 \text{ versus } H_1 : \ \beta_1, \ldots, \beta_k \text{ are not all zero}$$

  based on the test statistic

  $$F = \frac{SS_{\text{reg}}/k}{RSS/(n-k-1)} = \frac{(n-k-1)R^2}{k(1-R^2)}.$$

  $F \sim F(k, n-k-1)$ under H_0. The global test rejects H_0 at level a if $F > f_{a,k,n-k-1}$.

- Example 1. Consider the data set

  `https://stat.walkup.tw/teaching/statistics/data/cheese.txt`

  that is originally in "Introduction to the Practice of Statistics" by Moore and McCabe (1989). The data set contains observations for each of the four variables: taste, Acetic, H2S and Lactic. The observations for Acetic, H2S and Lactic are concentration levels of acetic acid (醋酸), hydrogen sulfide (硫化氫) and lactic aid (乳酸) in the samples of cheddar cheese, and the observations for the variable taste are taste scores for the samples. Suppose that the data file is downloaded to a file named `cheese.txt` in the directory `D:\temp`, and we run

  ```
  data1 <-  read.table("D:\\temp\\cheese.txt", header = TRUE, row.names=1)
  taste <- data1[,1]
  Acetic <- data1[,2]
  H2S <- dat1a[,3]
  Lactic <- data1[,4]
  ```

  so that the observations for taste, Acetic, H2S and Lactic are read into R vectors `taste`, `Acetic`, `H2S` and `Lactic` respectively. After running the command

  ```
  summary(lm(taste~H2S+Lactic))
  ```

  we have the following R output:

```
Call:
lm(formula = taste ~ H2S + Lactic)

Residuals:
    Min      1Q  Median      3Q     Max
-17.343  -6.530  -1.164   4.844  25.618

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.592      8.982  -3.072  0.00481 **
H2S            3.946      1.136   3.475  0.00174 **
Lactic        19.887      7.959   2.499  0.01885 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.942 on 27 degrees of freedom
Multiple R-squared: 0.6517,     Adjusted R-squared: 0.6259
F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07
```

(a) Write down the estimated regression equation based on the above regression analysis.

(b) What is the proportion of the variation in taste that can be explained by H2S and Lactic?

(c) Can we conclude that the coefficients for H2S and Lactic are not all zero at level 0.05?

(d) Can we conclude that the coefficient for Lactic is not zero at level 0.01?

(e) Give an observed 95% confidence interval for the coefficient of H2S.

(f) Let n be the number of samples of cheddar cheese. What is n?

(g) Find the residual sum of squares.

(h) Find the regression sum of squares.

(i) Find the total sum of squares.

Solutions.

(a) The estimated regression equation is

$$\widehat{\text{taste}} = -27.592 + 3.946 \times \text{H2S} + 19.887 \times \text{Lactic}.$$

(b) 65.17%.

(c) Yes; the p-value is $6.551 \times 10^{-7} < 0.05$.

(d) No; the p-value is $0.01885 > 0.01$.

(e) An observed 95% confidence interval for the coefficient of H2S is

$$[3.946 - t_{0.025,27} \times 1.136, 3.946 + t_{0.025,27} \times 1.136].$$

From the table "Quantiles for t distributions", $t_{0.025,27} = 2.052$, so the observed 95% confidence interval for the coefficient of H2S is

$$[3.946 - 2.052 \times 1.136, 3.946 + 2.052 \times 1.136] = [1.614928, 6.277072].$$

(f) $n - 1 = 2 + 27$, so $n = 30$.

(g) $\sqrt{RSS/(n-k-1)} = \sqrt{RSS/27} = 9.942$, so the residual sum of squares is $RSS = 27 \times (9.942)^2 = 2668.771$.

(h) $(SS_{\text{reg}}/k)/(RSS/(n-k-1)) = (SS_{\text{reg}}/2)/(RSS/27) = 25.26$ and $RSS/27 = (9.942)^2$, so regression sum of squares is $SS_{\text{reg}} = 25.26 \times (9.942^2) \times 2 = 4993.567$.

(i) The total sum of squares is $SS_{\text{total}} = SS_{\text{reg}} + RSS = 2668.771 + 4993.567 = 7662.338$.

- In Example 1, the estimated regression coefficients can be computed using the formula in (2):

```
length(H2S)
Xstar <- cbind(rep(1, 30), H2S, Lactic)
#Xstar has three columns: a column of ones, H2S, Lactic

tXstar <- t(Xstar)
#tXstar is the transpose matrix of Xstar

tXX.inverse <- solve(tXstar%*%Xstar)
#tXX.inverse is the inverse matrix of the product of tXstar and Xstar

beta.hat <- tXX.inverse %*%  tXstar %*% taste
#beta.hat is the vector of estimated regression coefficients

beta.hat
```

In addition, the estimated standard deviations of the intercept, the coefficient of H2S, and the coefficient of Lactic can be computed using (3) and $\hat{\sigma}_\varepsilon$:

```
residual <- taste - (beta.hat[1]+beta.hat[2]*H2S+beta.hat[3]*Lactic)
```

```
RSS <- sum(residual^2)
sigma.hat <- sqrt(RSS/27)   #n=30, k=2, n-k-1=27
sigma.hat #9.942

a <- c(1,0,0)
V <- (a %*% tXX.inverse %*% a)[1,1]
sqrt(V)*sigma.hat #8.982, estimated standard deviation of the intercept

a <- c(0,1,0)
V <- (a %*% tXX.inverse %*% a)[1,1]
sqrt(V)*sigma.hat #1.136, estimated standard deviation of the coefficient of H2S

a <- c(0,0,1)
V <- (a %*% tXX.inverse %*% a)[1,1]
sqrt(V)*sigma.hat #7.959, estimated standard deviation of the coefficient of Lactic
```

- Adding a redundant explanatory variable may increase the standard errors of regression coefficient estimates.

  Example 2. For the data in Example 1, create a redundant explanatory variable `x.extra` by running the commands:

  ```
  n <- length(taste)
  set.seed(1)
  x.extra <- H2S+Lactic+rnorm(n,sd=0.01)
  ```

  Then, obtain the result of fitting a multiple regression model by running the command

  ```
  summary(lm(taste~H2S+Lactic+x.extra))
  ```

  The output is

  ```
  Call:
  lm(formula = taste ~ H2S + Lactic + x.extra)

  Residuals:
       Min       1Q   Median       3Q      Max
  -17.2283  -6.5576  -0.9878   4.9823  25.8567
  ```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.769      9.287  -2.990  0.00603 **
H2S           27.219    208.090   0.131  0.89694
Lactic        43.365    210.073   0.206  0.83807
x.extra      -23.286    208.206  -0.112  0.91181
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6519,Adjusted R-squared:  0.6117
F-statistic: 16.23 on 3 and 26 DF,  p-value: 3.797e-06
```

Comparing with the result in Example 1, the standard errors for the coefficient estimators are very large here.

- The problem of multicollinearity (共線性問題). In a multiple regression model, if an explanatory variable can be approximated well using linear combination of other explanatory variable(s), then the explanatory variable is nearly redundant and should be removed from the model. Otherwise, the regression coefficients cannot be well-estimated.

- In Example 2, the variable `x.extra` can be explained well by `H2S` and `Lactic`. The output for running

  ```
  summary(lm(x.extra~H2S+Lactic))
  ```

  shows that the multiple R^2 is almost 1, which indicates a strong collinearity. The standard errors for the estimated regression coefficients in

  ```
  summary(lm(taste~H2S+Lactic+x.extra))
  ```

  are large. However, when `x.extra` is removed, the standard errors for the estimated regression coefficients in

  ```
  summary(lm(taste~H2S+Lactic))
  ```

  are smaller.

- A simulated example of multicollinearity. 這個例子的 R output 就不貼了, 想看結果可以自己去執行相關指令.

  Example 3. Generate data:

```
 set.seed(1)
x1 <- rnorm(1000)
x2 <- rnorm(1000)
x3 <- x1 + x2 + rnorm(1000)/50
y <- 1 +  x1 + 0.5*x2+ rnorm(1000)
```

  - Run `summary(lm(y~x1+x2+x3))`. None of the 3 regression coefficients are significant, but the global test is significant.
  - Run

    ```
    summary(lm(x3~x1+x2));summary(lm(x2~x1+x3));summary(lm(x1~x2+x3))
    ```

    In each of the 3 models, the multiple R^2 is greater than 90%, which implies strong collinearity.
  - Run

    ```
    summary(lm(y~x1+x2));summary(lm(y~x1+x3));summary(lm(y~x2+x3))
    ```

    For each of the above three models, the coefficients of the explanatory variables are all significant (p-value < 0.05), so the problem of collinearity can be solved by removing an explanatory variable from `lm(y~x1+x2+x3)`. To choose one from the above models for the prediction of `y`, we can choose the model `lm(y~x1+x3)`, which gives the largest multiple R^2 and does not have a collinearity problem.

- Stepwise regression.
  - Forward selection: add the most useful explanatory variable one at a time; start with 0 variable.
  - Backward elimination: remove the least significant explanatory variable one at a time; start with all variables.
- In the collinearity example, forward selection starts with

  ```
  summary(lm(y~x1));summary(lm(y~x2));summary(lm(y~x3))
  ```

  and backward elimination starts with

  ```
  summary(lm(y~x1+x2+x3))
  ```