

Stat 324: Lecture 23

Multiple Regression

Moo K. Chung
mchung@stat.wisc.edu

April 28, 2005

1. An experiment was carried out to assess the impact of x_1 amount of extractable iron, x_2 amount of extractable aluminum on y phosphate adsorption index.

```
> xmp13.15
      index iron aluminum
1         4   61        13
2        18  175        21
...
12        62  333        88
13        40  199        54

> pairs(xmp13.15)
```

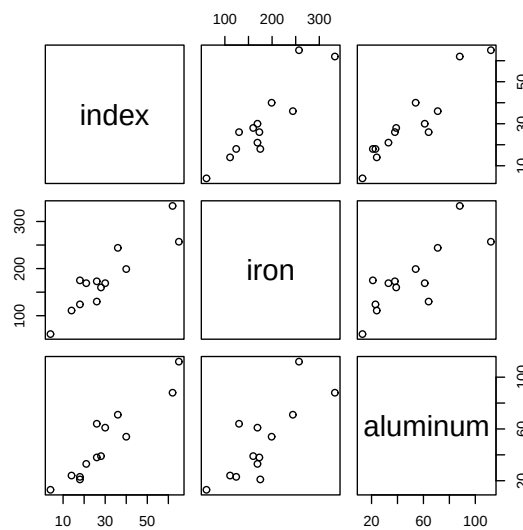


Figure 1: Pairwise scatter plots for Example 13.15.

2. The aim of *multiple regression* is to build a model that relates a dependent variable y to more than one independent variables $x_1, x_2, \dots, x_k$. We consider the following regression model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon.$$

Given multivariate data $(y_j, x_{1j}, x_{2j}, \dots, x_{kj})$, we fit the regression model and estimate $\beta_0, \dots, \beta_k$ via the least squares method. It is done by setting up a normal equation $y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_k x_{kj}$

for all $j = 1, \dots, n$ and making a matrix equation of the form $y = X\beta$. Then the least squares estimation of β is given by $\beta = (X'X)^{-1}X'y$. Inference on single coefficient β_j is based on the statistic $T = \frac{\hat{\beta}_j - \beta}{S_{\hat{\beta}_j}} \sin t_{n-k+1}$.

```
> summary(lm(index~iron+aluminum))
Call: lm(formula = index ~ iron + aluminum)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.35066      3.48467  -2.109 0.061101 .
iron          0.11273      0.02969   3.797 0.003504 **
aluminum      0.34900      0.07131   4.894 0.000628 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 4.379 on 10 degrees of freedom Multiple
R-Squared: 0.9485, Adjusted R-squared: 0.9382 F-statistic: 92.03
on 2 and 10 DF, p-value: 3.634e-07

3. *F-test on a group of predictors.* For given model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_l x_l + \beta_{l+1} x_{l+1} + \dots + \beta_k x_k + \epsilon,$$

we might be interested in testing if $H_0 : x_{l+1} = \dots = \beta_k = 0$ vs. $H_1 : \text{at least one among } \beta_{l+1}, \dots, \beta_k \text{ is not } 0$. This is equivalent to testing if null model (or reduced model):

$$H_0 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_l x_l + \epsilon$$

is statistically significant against the alternate model (or full model):

$$H_1 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_l x_l + \beta_{l+1} x_{l+1} + \dots + \beta_k x_k + \epsilon.$$

We can view this as a model or variable selection problem. Let SSE_k be the sum of squared residual for the full model and SSE_l be the corresponding SSE for the reduced model. Then $SSE_k \leq SSE_l$. To see this, note that

$$SSE_k = \min_{\beta_0, \dots, \beta_k} \sum_{j=1}^n (y_j - \mathbf{x}_j \mathbf{b}) \leq \min_{\beta_0, \dots, \beta_l, \beta_{l+1}=0, \dots, \beta_k=0} \sum_{j=1}^n (y_j - \mathbf{x}_j \mathbf{b}) = SSE_l,$$

where $\mathbf{x}_j = (1, x_{j1}, \dots, x_{jk})$ and $\mathbf{b} = (\beta_0, \beta_1, \dots, \beta_k)'$. The ratio $0 \leq \frac{SSE_l - SSE_k}{SSE_k} \leq 1$ measures the percentage increase of the goodness-of-fit from the reduced model to the full model. Hence this can be used as a test statistic. Note that

$$R_k^2 = 1 - \frac{SSE_k}{SST_k}, R_l^2 = 1 - \frac{SSE_l}{SST_l}$$

while $SST_k = SST_l = \sum_{j=1}^n (y_j - \bar{y})^2$. Hence,

$$\frac{SSE_l - SSE_k}{SSE_k} = \frac{R_k^2 - R_l^2}{1 - R_k^2}$$

The test statistic is

$$F = \frac{(SSE_l - SSE_k)/(k - l)}{SSE_k/(n - (k + 1))} = \frac{(R_k^2 - R_l^2)/(k - l)}{(1 - R_k^2)/(n - (k + 1))} \sim F_{k-l, n-(k+1)},$$

the F distribution with $k - l$ and $n - (k + 1)$ degrees of freedom. The R package has commands `qf` for computing the quantile threshold.

Example. In the previous example with $l = 0$, we are testing

$$H_0 : Y = \beta_0 + \epsilon \text{ vs. } H_1 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

The null model implies that none of the variables x_1 and x_2 has any effect on variable y . Note that $R_0^2 = 1 - \frac{SST}{SST} = 0$. $k = 2$. So the test statistic value is

$$f = \frac{R_2^2/2}{(1 - R_2^2)/(13 - 3)} = \frac{0.9485/2}{(1 - 0.9485)/10} = 92.08738.$$

We reject H_0 if $f > \text{qf}(0.95, 2, 10) = 4.10$ at level $\alpha = 0.05$. Hence we conclude that the full model is more appropriate.

4. It is possible to construct more complex models with interaction and high order terms. The model with first order predictors and interaction:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

```
> summary(lm(index~iron+aluminum+iron*aluminum))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.3678404	7.1789075	-0.330	0.749
iron	0.0827883	0.0481758	1.718	0.120
aluminum	0.2459707	0.1480639	1.661	0.131
iron:aluminum	0.0005278	0.0006610	0.799	0.445

Residual standard error: 4.461 on 9 degrees of freedom Multiple
R-Squared: 0.9519, Adjusted R-squared: 0.9358 F-statistic: 59.34
on 3 and 9 DF, p-value: 2.983e-06

Quadratic model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \epsilon.$$

```
> summary(lm(index~iron+aluminum+iron*aluminum+I(iron^2)))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-5.3376291	7.7805901	-0.686	0.512
iron	0.1785085	0.1077084	1.657	0.136
aluminum	0.0166120	0.2742567	0.061	0.953
I(iron^2)	-0.0004053	0.0004078	-0.994	0.349
iron:aluminum	0.0016883	0.0013420	1.258	0.244

Residual standard error: 4.464 on 8 degrees of freedom Multiple
R-Squared: 0.9572, Adjusted R-squared: 0.9357 F-statistic: 44.69
on 4 and 8 DF, p-value: 1.626e-05

Assigned problems. Exercise 13.44. (produce R computer output similar to SPSS output), Exercise 13.50. HW 7 due last day of class. Solve the assigned problems in Lecture 23-25 and Exercise 13.25 from Lecture 22. (generate a similar computer output using R).