Biostatistics 576B
Multiple Linear Regression - SAS

1. Introduction

"Big picture view" – Multiple linear regression is the statistical approach used when we want to determine whether there is a relationship between two variables after adjusting for other variables. In Epidemiology, it is used most often to assess whether a relationship is present after adjusting for potential confounders and effect modifiers.

Goal is to determine whether or not there is a linear relationship between a dependent variable, y, and multiple independent variable, x_1, $x_2, \ldots, x_k$

Example 1: FEV (forced expiratory volume) is an index of pulmonary function that measures the volume of air expelled after one second of constant effort. As part of a longitudinal study assessing changes in pulmonary function over time in children, FEV was determined on 654 children ages 3-19. In addition to FEV measurements, data were collected on age, height, gender and smoking status. Determine the predictors of FEV in this sample of children, i.e., is FEV related to age, height, gender or smoking status?

Statistical model for multiple linear regression is:

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k + e$$

where y = dependent variable (FEV in example)
x_1 = first independent variable (age in example)
x_2 = second independent variable (height in example)
x_k = kth independent variable
α = intercept
β_1= slope corresponding to the first independent variable (age in example)
β_2= slope corresponding to the second independent variable (height in example)
β_k= slope corresponding to the kth independent variable
e = error

Assume $e \sim N(0,\sigma^2)$

2. Estimating partial-regression coefficients, $\alpha, \beta_1, \beta_2, \ldots, \beta_k$

Some books use a and b_1, b_2, …, b_k to denote the estimates of $\alpha, \beta_1, \beta_2, \ldots, \beta_k$; other books use $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$, ..., $\hat{\beta}_k$

Estimate coefficients by minimizing sum of squared deviations from the regression surface (least squares)

Coefficients are called partial-regression coefficients because they represent the expected increase in y for a unit increase in a given x, assuming all other x's are held constant.

Example 2: What are the estimated regression coefficients for the relationship between FEV adjusted for age and height?

SAS PROC and Output:

```
* CPH 576B SAS Code for Lecture LI_1;

* FEV dataset stored in library CPH_576b;

data fev;
 set Cph_576b.Fev;
run;

proc reg data=fev corr;
 model fev = age hgt /clb pcorr2 scorr2;
run;
```

The SAS System

The REG Procedure

Number of Observations Read	654
Number of Observations Used	654

Correlation

Variable	AGE	HGT	FEV
AGE	1.0000	0.7919	0.7565
HGT	0.7919	1.0000	0.8681
FEV	0.7565	0.8681	1.0000

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: FEV

Number of Observations Read	654
Number of Observations Used	654

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	376.24494	188.12247	1067.96	<.0001
Error	651	114.67489	0.17615		
Corrected Total	653	490.91984			

Root MSE	0.41970	**R-Square**	0.7664
Dependent Mean	2.63678	**Adj R-Sq**	0.7657
Coeff Var	15.91732		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Squared Semi-partial Corr Type II**	**Squared Partial Corr Type II**
Intercept	**1**	-4.61047	0.22427	-20.56	<.0001	.	.
AGE	**1**	0.05428	0.00911	5.96	<.0001	0.01275	0.05176
HGT	**1**	0.10971	0.00472	23.26	<.0001	0.19418	0.45393

Example 3: What is the expected FEV for a 10 year old child who is 54 inches tall?

E(FEV) = -4.610 + 0.054*10 + 0.110*54 = 1.857

3. Hypothesis testing of parameters from the multiple regression line

a. Simultaneous test for whether the partial-regression coefficients are all simultaneously equal to 0

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ H_1: At least one of the $\beta_j \neq 0$

Standard analysis of variance (ANOVA) table

Test is based on the following relationship:

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Total SS = Regression SS + Residual SS

If regression plane fits the data well, expect large regression SS and a small residual SS

Then the test statistic is

$$F = \frac{\text{Regression MS}}{\text{Residual MS}} = \left(\frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 / k}{\left[\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 - \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2\right] / (n - k - 1)} \right)$$

The p-value is p = Probability($F_{k,n\text{-}k\text{-}1} > F$)

Example 4: Test whether or not there is a significant linear relationship between FEV versus age and height for these children.

b. Test for whether a particular partial-regression coefficient equals 0

H_0: $\beta_j = 0$, all other $\beta_j \neq 0$ H_1: $\beta_j \neq 0$, all other $\beta_j \neq 0$

Uses the estimated value of the partial-slope coefficient, b_j and its estimated standard error

Then the test statistic is

$$t = b_j / se(b_j)$$

The p-value is p = 2 · (area to the left of t under a t_{n-k-1} distribution) if $t < 0$
p = 2 · (area to the right of t under a t_{n-k-1} distribution) if $t \geq 0$

Example 5: Test whether or not there is a significant linear relationship between FEV and age, after adjusting for height.

4. Measures of correlation in multiple regression analysis

a. Multiple correlation coefficient (R)

R measures the degree of dependence between the dependent variable and a set of independent variables

R = Correlation between y_i and $\hat{y}_i = \sqrt{\dfrac{\text{Regression SS}}{\text{Total SS}}}$

Note that $0 \le R \le 1$

R^2 measures the proportion of variance in the dependent variable that is explained by the set of independent variables (also called the coefficient of determination)

Note that R^2 never decreases as additional independent variables are added to the model

Example 6: What proportion of the variation in FEV is explained by age and height?

The adjusted R^2 is adjusted for the number of predictors. When a predictor is added, adjusted R^2 increases only if the increment in R^2 is larger than the increment in the penalty.

$$R^2_{adj} = 1 - \frac{\left(1 - R^2\right)\cdot\left(n - 1\right)}{n - k - 1}$$

b. Partial correlation coefficient ($r_{ij\cdot k}$)

$r_{ij\cdot k}$ measures the degree of dependence between the dependent variable and a particular independent variable, after adjusting for the linear effect of the other independent variables in the model; specifically, it is the partial correlation between variables i and j after adjusting for variable k

$$r_{ij\cdot k} = \frac{r_{ij} - r_{ik}\cdot r_{jk}}{\sqrt{\left(1 - r_{ik}^2\right)\cdot\left(1 - r_{jk}^2\right)}}$$

Note that $-1 \le r_{ij\cdot k} \le 1$

Example 7: What is the partial correlation coefficient between FEV and age, after adjusting for height?

The partial correlation is the correlation between y and x_1 if the other independent variables did not vary (are held constant). The squared partial correlation represents the proportion of variance in y not explained by the other independent variables that is explained by x_1.

The semipartial correlation is the correlation that would be observed between y and x_1 after the effects of all other independent variables are removed from x_1 but not from y. The squared semipartial correlation represents the proportion of variance in y that is explained by x_1 only. This can be interpreted as the decrease in the model's R^2 value that results from removing x_1 from the full model.