Multiple linear regression

Biometry 755

Spring 2009

The multiple linear regression model

Multiple linear regression is a statistical method that allows us to find the best fitting linear relationship (response surface) between a single dependent variable, Y, and a collection of independent variables $X_1, X_2, \ldots, X_k$. We assume that the following model expresses the true relationship between Y and the set of independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

where ε is a random error term that accounts for the random deviations of data points from the response surface.

Multiple linear regression assumptions

Linearity The mean value of Y is a linear function of $X_1, X_2, \ldots, X_k$. That is to say, the true statistical model is

$$E[Y|X_1, X_2, \ldots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k.$$

NOTE: The linearity assumption does not preclude the presence of higher order terms in the model. For example, both $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$ and $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$ satisfy the assumption of linearity, even though each contains second order terms (X_1^2 and $X_1 X_2$, respectively).

Linearity means linear in the regression coefficients. Here is an example of *non*-linear model.

$$Y = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}.$$

Multiple linear regression assumptions (cont.)

Independence The Y values must be independent, i.e. form a random sample.

Homoscedasticity The variance of Y is the same for any combination of values of $X_1, X_2, \ldots, X_k$. In symbols, we write

$$Var(Y|X_1, X_2, \ldots, X_k) = \sigma^2.$$

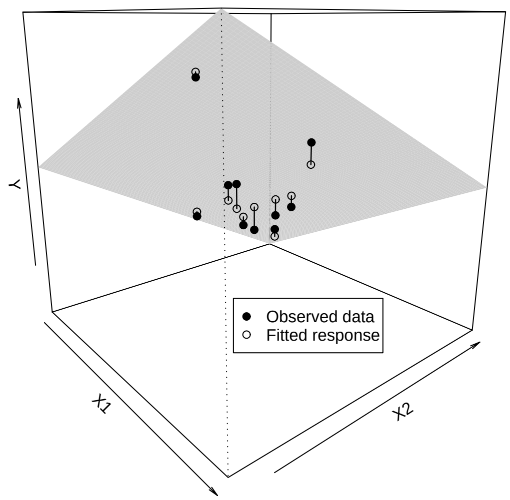
Normality Given any fixed combination of $X_1, X_2, \ldots, X_k$,

$$Y \sim \text{Normal}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k, \sigma^2),$$

or equivalently $\varepsilon \sim \text{Normal}(0, \sigma^2)$.

Determining the optimal surface

The “best” surface is that which minimizes the sum of the squared residuals. It can be shown that the $\hat{\beta}$s that result from this method have minimal variance and are unbiased.



Summarizing multiple regression results

We represent the fitted model as

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

The formulas for the fitted regression coefficients are matrix equations and require knowledge of matrix algebra (not a prerequisite for this course). Instead, we’ll rely on the computer to provide fitted values.

The ANOVA table for multiple linear regression

Source	df	SS	MS	F
Model	k	SSR = SSY - SSE	MSR $= \frac{\text{SSR}}{k}$	$\frac{\text{MSR}}{\text{MSE}}$
Error	$n-k-1$	SSE	MSE $= \frac{\text{SSE}}{n-k-1}$	
Total	$n-1$	SSY		

where k is the number of independent variables in the model. Note that this general ANOVA table is consistent with the ANOVA table presented for SLR.

The ANOVA table for MLR (cont.)

The interpretation of the components in the ANOVA table are the same as for SLR.

- SSY is the total variability in Y
- SSR is the variation in Y attributable to its linear association with $X_1, \ldots, X_k$
- SSE is the amount of variation in Y left unexplained by the model

R^2 = SSR/SSY, but unlike in SLR, R^2 does *not* equal the square of the sample correlation coefficient. However, it does measure the proportion of total variation explained by the model and varies between 0 and 1.

SENIC example: MLR analyses

R^2 values for simple and multiple linear regressions of risk of nosocomial infection on selected variables.

Model	R^2
LOS	
CULT	
BEDS	
LOS, CULT	
LOS, BEDS	
CULT, BEDS	
LOS, CULT, BEDS	

R^2 and adjusted R^2

For nested models, R^2 can never decrease. This is because SSE monotonically decreases and SSY is fixed, so the quantity

$$R^2 = \frac{\text{SSR}}{\text{SSY}} = \frac{\text{SSY} - \text{SSE}}{\text{SSY}} = 1 - \frac{\text{SSE}}{\text{SSY}}$$

can only increase. It is therefore possible to artificially inflate the value of R^2 simply by including additional variables in the multiple regression.

An alternative measure of fit is the *adjusted* R^2.

R^2 and adjusted R^2 (cont.)

Adjusted R^2 is defined as

$$R_a^2 = 1 - \frac{\left(\frac{\text{SSE}}{n-p}\right)}{\left(\frac{\text{SSY}}{n-1}\right)} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSE}}{\text{SSY}}.$$

This index divides each of the sums of squares by its associated degrees of freedom. In so doing, R_a^2 can actually decrease when a covariate is added to the model, because any decrease in SSE may be more than offset by the loss of a degree of freedom in the denominator $n - p$.

(Note: p is the number of parameters in the MLR and is always equal to $k + 1$. Therefore, $n - p = n - k - 1$, which is what is reported as the df associated with SSE in Slide 7.)

SENIC example: MLR analyses (cont.)

R^2 and adjusted R^2 values for simple and multiple linear regressions of risk of nosocomial infection on selected variables.

Model	R^2	Adjusted R^2
LOS		
CULT		
BEDS		
LOS, CULT		
LOS, BEDS		
CULT, BEDS		
LOS, CULT, BEDS		

Inference in multiple linear regression

1. Overall test of significance of the regression.
2. Test of significance for addition of a single variable.
3. Test of significance for addition of a group of variables.

Overall test

Given the linear model with k independent variables

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon,$$

does the regression of Y on $X_1, \ldots, X_k$ explain a significant proportion of the variability in Y? Formally, we state

- H_0: The regression on $X_1, \ldots, X_k$ does not explain a significant proportion of the variability in Y.
  H_A: The regression on $X_1, \ldots, X_k$ does explain a significant proportion of the variability in Y.

OR

- $H_0 : \beta_1 = \ldots = \beta_k = 0$
  H_A: At least one of $\beta_1, \ldots, \beta_k$ is different from zero.

Overall test (cont.)

Test statistic $F = \frac{\text{MSR}}{\text{MSE}} \sim F_{k,n-k-1}$ under H_0.

p-value p-value = Prob $\left(F > \frac{\text{MSR}}{\text{MSE}}\right)$ where $F \sim F_{k,n-k-1}$.

Conclusion If we reject H_0, we conclude that at least one of the independent variables significantly explains the variation in Y. If we fail to reject H_0, we conclude that there is insufficient evidence to conclude that any of the independent variables significantly explains the variation in Y.

Overall test in SAS

Consider a multiple linear regression of risk of nosocomial infection on length of stay, routine culturing ratio, and number of beds.

```
proc reg data = one;
    model infrisk = los cult beds;
run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	95.36610	31.78870	32.68	<.0001
Error	109	106.01372	0.97260		
Corrected Total	112	201.37982			

Overall test in SAS (cont.)

The value of the test statistic for the overall F test is $31.78870/0.97260 = 32.68$ which has an F distribution with 3 numerator degrees of freedom and 109 denominator degrees of freedom under H_0. The p-value is less than 0.0001. We conclude at $\alpha = 0.05$ that at least one of LOS, CULT and BEDS significantly explains the variation in INFRISK.

The significance of a single covariate

When the overall F test is rejected in multiple linear regression, additional tests called *partial F tests* are performed to investigate the importance of each of the independent variables *while controlling or adjusting for the effects of the other independent variables*. If there are k independent variables, then there are k partial F tests.

Partial sum of squares

A first approach at assessing the significance of each independent variable is to consider the *partial sum of squares for each variable*. Recall that the total variability in Y (SSY) is partitioned into two mutually exclusive components:

1. Variability explained by the linear regression model of Y on $X_1, \ldots, X_k$ (SSR)
2. Unexplained variability (SSE).

For a model containing k covariates (independent variables), the partial sum of squares for a specific variable measures the increase in the regression sum of squares by adding that variable to a model already containing the other $k-1$ covariates.

Partial sum of squares example

For example, suppose we fit a model with three independent variables, X_1, X_2, X_3. Then

- SSR$(X_1|X_2, X_3)$ measures the increase in SSR by adding X_1 to a model already containing X_2 and X_3.
- SSR$(X_2|X_1, X_3)$ measures the increase in SSR by adding X_2 to a model already containing X_1 and X_3.
- SSR$(X_3|X_1, X_2)$ measures the increase in SSR by adding X_3 to a model already containing X_1 and X_2.

Partial sum of squares in SAS

Consider the regression of INFRISK on LOS and CULT, and the regression of INFRISK on LOS, CULT and BEDS.

```
TWO VARIABLE MODEL
                                   Sum of
Source                   DF       Squares
                                              proc reg data = one;
Model                     2      90.70199        model infrisk = los cult;
Error                   110     110.67784     run;
Corrected Total         112     201.37982
/********************************************************************/
THREE VARIABLE MODEL
                                   Sum of
Source                   DF       Squares
                                              proc reg data = one;
Model                     3      95.36610        model infrisk =
Error                   109     106.01372                  los cult beds;
Corrected Total         112     201.37982     run;
```


Partial sum of squares in SAS (cont.)

- SSR(LOS,CULT) = 90.70199
- SSR(LOS,CULT,BEDS) = 95.36610
- Therefore, SSR(BEDS|LOS,CULT) = SSR(LOS,CULT,BEDS) - SSR(LOS,CULT) = 4.66411

Interpretation: The variable BEDS adds an additional 4.66411 to the sum square regression obtained from a model already containing LOS and CULT.

Question: Is that a meaningful (*significant*) addition?

Formalizing the partial F test

Although the partial sum of squares helps us quantify the effect of an individual variable on explaining the total variability in the response, we still need a formal hypothesis test to assess the significance of a variable's impact. To achieve this, we use a partial F test.
There are several ways to express the null and alternative hypotheses for partial F tests. Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon.$$

Suppose we want to test the hypothesis that a particular independent variable, X_i, explains a significant amount of the variability in Y, *given that all the other variables are in the model*. Then each of the following sets of null and alternative hypotheses are equivalent.

H_0 and H_A for the partial F test

1. $H_0 : \beta_i = 0$ (all other $\beta_j s \neq 0$)
   $H_A : \beta_i \neq 0$ (all other $\beta_j s \neq 0$)
   where β_i is the true slope associated with X_i.

   or equivalently

2. $H_0 : Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{i-1} X_{i-1} + \beta_{i+1} X_{i+1} + \ldots + \beta_k X_k + \varepsilon$
   is the better model.
   $H_A : Y =$
   $\beta_0 + \beta_1 X_1 + \ldots + \beta_{i-1} X_{i-1} + \beta_i X_i + \beta_{i+1} X_{i+1} + \ldots + \beta_k X_k + \varepsilon$
   is the better model.

H_0 and H_A for the partial F test (cont.)

The model specified in the null hypothesis is called the *reduced model*. The model specified in the alternative hypothesis is called the *full model*.

The partial F test on X_i can be thought of as a test comparing two models: the full model (which includes X_i and all other independent variables), and the reduced model (which includes all independent variables *except* X_i).

Constructing the partial F test

The comparison of a full and reduced model forms the basis for the construction of the partial F test. The test statistic and its null distribution are

$$\begin{aligned} F &= \frac{\text{SSR}(X_i|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_k)/1}{\text{MSE(full)}} \\ &= \frac{(\text{SSR(full) - SSR(reduced)})/1}{\text{MSE(full)}} \sim F_{1,n-k-1} \end{aligned}$$

where MSE(full) = SSE(full)$/(n-k-1)$.

Constructing the partial F test (cont.)

The numerator of the test statistic has 1 degree of freedom since the full and reduced models differ by a single variable. The denominator degrees of freedom is n minus the total number of parameters (intercept and slope parameters, i.e. β_0 and all the βs) estimated in the full model. The ratio provides a measure of whether the additional sum of squares explained by adding X_i are important or large in comparison to the unexplained variation, and is therefore a measure of the additional usefulness of the full model over the reduced model.

Conducting partial F tests in SAS

Fortunately, you can conduct partial F tests directly in SAS. The code below shows the statements in PROC REG needed to conduct a partial F test on the variable BEDS.

```
proc reg data = one;
    model infrisk = los cult beds;
    F_Beds: test beds = 0;
run;
```

```
   Test F_Beds Results for Dependent Variable INFRISK
                                 Mean
Source              DF         Square     F Value     Pr > F
Numerator            1        4.66412        4.80     0.0307
Denominator        109        0.97260
```

Compare with the partial sum of squares for BEDS shown on Slide 22.

Coincidence?

Partial F test for BEDS

Test F_Beds Results for Dependent Variable INFRISK

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	4.66412	4.80	0.0307
Denominator	109	0.97260		

t test for BEDS

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
BEDS	1	0.00116	0.00052963	2.19	0.0307

The t test alternative to the partial F test

Previously, we stated that

$$F_{1,\nu,1-\alpha} = t^2_{\nu,1-\alpha/2} = t^2_{\nu,\alpha/2}.$$

Since the numerator degrees of freedom for the partial F test is 1, then this principle holds, and there is a t test equivalent to the partial F test. More specifically, the one-sided partial F test is equivalent to a two-sided t test.

While the partial F test reflects the spirit of the full/reduced model null and alternative hypotheses (Slide 24, H_0 and H_A (2)), the t test reflects the spirit of assessing the significance of the appropriate β coefficient (Slide 24, H_0 and H_A (1)).

The t test

Null and alternative

$H_0 : \beta_i = 0$ (all other βs $\neq 0$)
$H_A : \beta_i \neq 0$ (all other βs $\neq 0$)

Test statistic $t = \frac{\hat{\beta}_i}{\text{SE}(\hat{\beta}_i)} \sim t_{n-k-1}$ under H_0.

p-value p-value $=$ Prob $\left(|t| > \frac{\hat{\beta}_i}{\text{SE}(\hat{\beta}_i)}\right)$, where $t \sim t_{n-k-1}$.

The t tests in SAS

The t-tests for the βs are standard output for any multiple linear regression in SAS. No special options need to be specified. Since the t-test is equivalent to the partial F-test, this is the preferred way to conduct any partial F-test

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.97491	0.48575	2.01	0.0472
LOS	1	0.22784	0.05598	4.07	<.0001
CULT	1	0.05630	0.00963	5.84	<.0001
BEDS	1	0.00116	0.00052963	2.19	0.0307

Interpreting the t tests

If all tests are conducted at the 0.05 level of significance, then

- LOS contributes significantly to a model already containing CULT and BEDS.
- CULT contributes significantly to a model already containing LOS and BEDS.
- BEDS contributes significantly to a model already containing LOS and CULT.

Multiple partial F tests

It is sometimes of interest to test for the importance of groups of independent variables. In such situations, a *multiple partial F test* is performed. For example, in a multiple linear regression containing X_1, X_2, X_3, X_4, we might want to test whether the pair of independent variables $\{X_2, X_3\}$ contributes significantly to a model already containing X_1 and X_4.

H_0 and H_A for a multiple partial F test

To test for the significance of the collection of g independent variables,

1. $H_0 : \beta_1^* = \ldots = \beta_g^* = 0$ (all other βs $\neq 0$)
   H_A : At least one of $\beta_1^*, \ldots, \beta_g^* \neq 0$ (all other βs $\neq 0$)

   or equivalently

2. $H_0 : Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$ is the better model.
   $H_A : Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \beta_1^* X_1^* + \ldots + \beta_g^* X_g^* + \varepsilon$ is the better model.
   The model specified in the null hypothesis is called the *reduced model*. The model specified in the alternative hypothesis is called the *full model*.

Formalizing the multiple partial F test

The form of the multiple partial F test is simply a generalization of the partial F test presented on Slide 26. The test statistic and its distribution under H_0 are

Test statistic

$$
\begin{aligned}
F &= \frac{\text{SSR}(X_1^*, \ldots, X_g^* | X_1, \ldots, X_k)/g}{\text{MSE(full)}} \\
&= \frac{(\text{SSR(full)} - \text{SSR(reduced)})/g}{\text{MSE(full)}} \\
&\sim F_{g, n-(\text{\# parameters in full model})}.
\end{aligned}
$$

p-value p-value $= \text{Prob}\left(F > \frac{(\text{SSR(full)} - \text{SSR(reduced)}/g)}{\text{MSE(full)}} \right)$

where $F \sim F_{g, n-(\text{\# parameters in full model})}$.

Formalizing the multiple partial F test (cont.)

Conclusion If we fail to reject H_0, then there is insufficient evidence that the collection of variables being tested contributes significantly to the model already containing the other variables. If we do reject H_0, then at least one of the independent variables in the collection being tested contributes significantly to a model already containing the other variables.

SENIC example: Multiple partial F test

Suppose we want to test the significance of the contribution of BEDS and NURSE to a model already containing LOS and CULT.

```
proc reg data = one;
    model infrisk = los cult beds nurse;
    F_beds_nurse: test beds, nurse = 0;
run;
```

```
Test F_beds_nurse Results for Dependent Variable INFRISK
                                   Mean
Source              DF           Square    F Value    Pr > F

Numerator            2          3.68062       3.85    0.0243
Denominator        108          0.95664
```

SENIC example: Multiple partial F test (cont.)

Where did the numerator and denominator for the test statistics come from?

```
TWO VARIABLE MODEL
                                  Sum of
Source                   DF      Squares
                                             proc reg data = one;
Model                     2     90.70199        model infrisk = los cult;
Error                   110    110.67784     run;
Corrected Total         112    201.37982
/*********************************************************************/
FOUR VARIABLE MODEL
                                  Sum of
Source                   DF      Squares
                                             proc reg data = one;
Model                     4     98.06324        model infrisk = los cult
Error                   108    103.31659                       beds nurse;
Corrected Total         112    201.37982        run;
```


Multiple partial F test conclusion

Since the p-value for the multiple partial F test is significant, we conclude that at least one of BEDS and NURSE contributes significantly to a model already containing LOS and CULT.