Multiple Linear Regression

1. The Multiple Regression Model in General

1. Multiple Regression Model with p - 1 Independent Variables

The multiple linear regression model with p - 1 independent variables can be written

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad i = 1,\ldots,n$$

where

- Y_i is the response for the ith case
- $X_{i,1}, X_{i,2}, \ldots, X_{i,p-1}$ are the values of p - 1 independent variables for the ith case, assumed to be known constants
- $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters
- ε_i are independent $\sim N(0, \sigma^2)$

(The independent variables are indexed 1 to p - 1 so that the total number of independent variables, including the implicit column of 1 associated with the intercept β_0, is equal to p.)

The interpretation of the parameters:

1. β_0 indicates the mean of the distribution of Y when $X_1 = X_2 = \ldots = X_{p-1} = 0$
2. β_k (k = 1, 2, ..., p - 1) indicates the change in the mean response E{Y} (measured in Y units) when X_k increases by one unit while all the other independent variables remain constant
3. σ^2 is the common variance of the distribution of Y

The β_k are sometimes called *partial regression coefficients,* but more often just *regression coefficients*, or *unstandardized regression coefficients* (to distinguish them from *standardized coefficients* discussed below.)

Recall, the regression model for the entire data set can be written

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

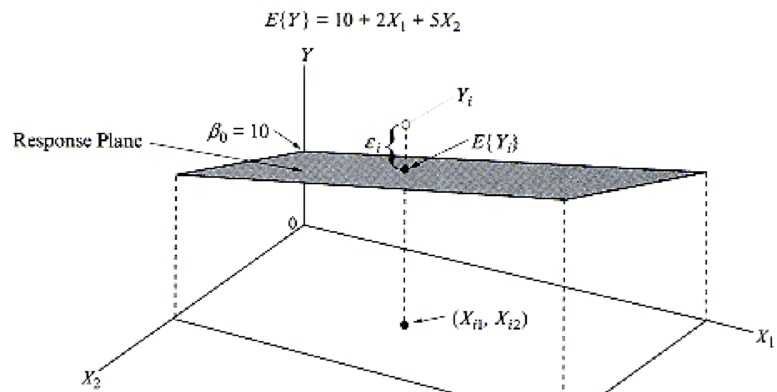
In the model

- $\mathbf{y}$ is a n x 1 vector of responses
- $\boldsymbol{\beta}$ is a p x 1 vector of parameters
- $\mathbf{X}$ is a n x p matrix of constants
- $\boldsymbol{\varepsilon}$ is a vector of independent normal random variables such that $E\{\boldsymbol{\varepsilon}\} = \mathbf{0}$ and the variance-covariance matrix $\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\} = E\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\} = \sigma^2\mathbf{I}$

2. Geometry of the First Order Multiple Regression Model

The response function (also called regression function or response surface) defines a hyperplane in p-dimensional space. When there are only 2 predictor variables (besides the constant) the response surface is a plane.

FIGURE 6.1 **Response Function is a Plane—Sales Promotion Example.**



When there are more than 2 independent variables (in addition to the constant) the regression function is a hyperplane and can no longer be visualized in 3-dimensional space.

2. Elements of the Regression Model

1. Ph.D. Example

To illustrate a typical multiple regression analysis we use the Ph.D. example with more predictors

$$PUBS = \beta_0 + \beta_1 TIME + \beta_2 CITS + \beta_3 SALARY + \beta_4 AGE + \varepsilon_i$$

The variables are defined as
(y) PUBS, Number of publications
(x_1) TIME, Years since Ph.D.
(x_2) CITS, Number of citations
(x_3) SALARY, Salary in dollars
(x_4) AGE, Age of the professor

2. Correlation Matrix

The simple correlation coefficients among variables in the multiple regression model are often presented in the form of a matrix.

3. Estimated Regression Function $\hat{y}$

The estimated regression function for the multiple regression model with p - 1 variables is

$$\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \dots + b_{p-1} x_{i,p-1}$$

where $b_0, b_1, \dots, b_{p-1}$ are estimated as the solution of the ordinary least squares normal equations

$$\mathbf{X'Xb} = \mathbf{X'Y} \quad \text{or} \quad \mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

as derived in the last set of notes.

The variance-covariance matrix of **b** is estimated as

$$\mathbf{s^2}\{\mathbf{b}\} = \text{MSE}(\mathbf{X'X})^{-1}$$

The standard errors of each estimated coefficient b_k is the square root of the corresponding diagonal element of $\mathbf{s^2}\{\mathbf{b}\}$, so that $s\{b_0\}$ is in position (1,1), $s(b_1\}$ in position (2,2), ..., and $s\{b_{p-1}\}$ in position (p,p).

On the standard multiple regression printout the estimated coefficients b_k are presented, together with the estimated standard errors $s\{b_k\}$ and the t-ratio $t^* = b_k/s\{b_k\}$ (see SAS output).

4. Analysis of Variance (ANOVA)

1. Fitted Values $\hat{y}_i$

The fitted values $\hat{y}_i$ are defined in a way analogous to simple regression as

$$\hat{y}_i = b_0 - b_1 x_{i,1} - b_2 x_{i,2} - \dots - b_{p-1} x_{i,p-1}$$

or

$$\hat{\mathbf{Y}} = \mathbf{Xb}$$

where $\hat{\mathbf{y}}$ is a n x 1 vector of fitted values. Note that $\hat{y}_i$ is a single number associated with each case, regardless of the number p - 1 of independent variables in the model.

2. Sums of Squares

The sums of squares are defined identically in simple and multiple regression, as

$\text{SSTO} = \Sigma(Y_i - \bar{Y})^2$
$\text{SSE} = \Sigma(Y_i - \hat{Y}_i)^2$
$\text{SSR} = \Sigma(\hat{Y}_i - \bar{Y})^2$
with the relation SSTO = SSR + SSE

3. Degrees of Freedom

The degrees of freedom (df) associated with various sums of squares are

- SSTO has n - 1 df
- SSE has n - p df
- SSR has p - 1 df

4. Mean Squares

Mean squares are sums of squares divided by their respective degrees of freedom (df).
In particular, MSE = SSE/(n - p) is again the estimate of σ^2, the common variance of ε and of Y.

5. ANOVA Table

Analysis of variance results are summarized in an ANOVA table analogous to the one for simple regression. Table 1 shows the general format of the ANOVA table.

Table 1. General Format of ANOVA Table for Multiple Regression

Source of variation	SS	df	MS	F Ratio
Regression	$SSR = (\hat{Y}_i - \bar{Y})^2$	p - 1	MSR = SSR/(p - 1)	F* = MSR/MSE
Error	$SSE = \Sigma(Y_i - \hat{Y}_i)^2$	n - p	MSE = SSE/(n - p)	
Total	$SSTO = \Sigma(Y_i - \bar{Y})^2$	n - 1	$s_Y^2 = SSTO/(n - 1)$	

Table 1 also shows the calculation of the *f-ratio* or *f-statistic* F*= MSR/MSE. The interpretation of F* is discussed below.

5. Coefficient of Determination R^2

1. Coefficient of Determination R^2 (SLR)

The following formulas are equivalent:
R^2 = (SSTO - SSE)/SSTO = SSR/SSTO = 1 - SSE/SSTO
where $0 \leq R^2 \leq 1$

Example: In the regression of publications since Ph.D., the R^2 can be calculated equivalently as

R^2 = 1 - (1521.515/2674.933) = .4312

Limiting cases:

- if all observations on regression line, then SSE=0 and $r^2 = 1$
- if slope $b_1 = 0$ then SSR = 0 and $r^2 = 0$

The R^2 is interpreted as the proportion of the variation in y "explained" by the regression model. That is, 43.12% of the variation in publications is explained by time since Ph.D.

2. Coefficient of Multiple Determination R^2 (MLR)

The coefficient of multiple determination R^2 is defined analogously to the simple regression R^2 as $R^2 = SSR/SSTO = 1 - (SSE/SSTO)$

where

$0 \leq R^2 \leq 1$

3. Adjusted R-Square R_a^2

The adjusted coefficient of multiple determination R_a^2 adjusts for the number of independent variables in the model (to correct the tendency of R^2 to always increase when independent variables are added to the model). It is calculated as

$R^2_a = 1 - ((n-1)/(n-p))(SSE/SSTO) = 1 - MSE/(SSTO/(n - 1))$

R^2_a can be interpreted as 1 minus the ratio of the variance of the errors (MSE) to the variance of y, SSTO/(n-1).

Example: In the Ph.D. example the adjusted r-square R^2_a is .4544 as contrasted with the ordinary (unadjusted) $R^2 = .4902$. 45.44% of the variation in publications since Ph.D. is explained by the set of independent variables included in the multiple regression.

5. Inference for Entire Model - F Test for Regression Relation

The F test for regression relation (*aka* screening test) tests the existence of a relation between the dependent variable and the *entire set* of independent variables. The test involves the hypothesis setup

H_0: $\beta_1 = \beta_2 = ... = \beta_{p-1} = 0$
H_1: Not all $\beta_k = 0$ $k = 1, 2, ..., p - 1$

The test statistic is (same as for simple linear regression)
$F^* = MSR/MSE$

which is distributed as $F(p - 1; n - p)$, the same df as the numerator and denominator, respectively, in the ratio MSR/MSE.

Using the <u>p-value method</u>, calculate the p-value $P\{F(p - 1; n - p) > F^*\}$.
For a significance level α, the decision rule is

- if p-value $< \alpha$
  - reject H_0 and conclude H_1 (not all coefficients = 0 so there is a significant statistical relation)

- if p-value $\geq \alpha$
  - fail to reject H_0 and conclude H_0 (there is no significant statistical relation)

Using the <u>critical value method</u>, calculate the critical value $F(1 - \alpha; p - 1, n - p)$. For a significance level α, the decision rule is

- if $F^* \leq F(1 - \alpha; p - 1, n - p)$,
  - fail to reject H_0 and conclude H_0
- if $F^* > F(1 - \alpha; p - 1, n - p)$
  - reject H_0 and conclude H_1

Example: Carry out the F test for the regression for the Ph.D. example.

Step 1: Set up null and alternative hypothesis

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1: Not all $\beta_k = 0$ $k = 1, 2, 3, 4$

Step 2: Choose a significance level

$\alpha = .05$

Step 3: Calculate test statistic

$F^* = 13.70$

Step 4: Determine the p-value or the critical value

P-value approach: Find the ~~2-tailed~~ p-value

```
data pvalue;
Fobs = 13.7;
ndf = 4;
ddf = 57;
prob = 1-probf(fobs,ndf,ddf);
run;
```

p-value < .0001

Critical value approach: Determine the critical value

With $\alpha = .05$, $F(1 - \alpha; p - 1, n - p) = F(0.95;4,57) = 2.54$

Step 5: Make a decision

Since $F^* = 13.7 > 2.54$ or $p < .05$, reject H_0 and conclude H_1 (not all coefficients = 0 so there is a significant statistical relation) at the .05 level.

6. Inference for Individual Regression Coefficients

Statistical inference on individual regression β_k is carried out in the same way as for simple regression, except that the t tests are now based on the Student t distribution with n - p degrees of freedom (corresponding to the n - p df associated with MSE), instead of the n - 2 df of the simple regression model. Remember p is the number of parameters in the model, NOT the number of independent variables (that is p - 1).

1. Hypothesis Tests for β_k

1. Two-Sided Tests

The most common tests concerning β_k involve the null hypothesis that $\beta_k = 0$.

H_0: $\beta_k = 0$
H_1: $\beta_k \neq 0$

The test statistic is

$$t^* = b_k/s\{b_k\}$$

where $s\{b_k\}$ is the estimated standard deviation of b_k. When $\beta_k = 0$, $t^* \sim t(n - p)$.

Example: Test the hypothesis that the coefficient of age $\beta_4 = 0$. The setup is

Step 1: Set up null and alternative hypothesis
H_0: $\beta_4 = 0$ ("null hypothesis")
H_1: $\beta_4 \neq 0$ ("alternative hypothesis")

Step 2: Choose a significance level
$\alpha = .05$

Step 3: Calculate test statistic
$t^* = b_4/s\{b_4\} = .39215/.18587 = 2.11$

Step 4: Determine the p-value or the critical value

P-value approach: Find the 2-tailed p-value

```
data pvalue;
tobs = 2.11;
df = 57;
prob = 2*(1-probt(tobs, df));
run;
```

p-value = .0393

Critical value approach: Determine the critical value

With $\alpha = .05$, $t(1 - \alpha/2; n - p)$ is $t(0.975; 57) = 2$

Step 5: Make a decision

Since $|t^*| = 2.11 > 2$ or $p < .05$, reject H_0 and conclude H_1 ($\beta_4 \neq 0$) at the .05 level.

2. One-Sided Tests

One-sided tests for a coefficient β_k are carried out by dividing the 2-sided p-value by 2, as before.

Example: Test that the coefficient of age is positive. The hypotheses are

H_0: $\beta_4 \leq 0$
H_1: $\beta_4 > 0$
Using the p-value method, find the 1-tailed p-value $P\{t(57) > 2.11\} = 0.0393/2 = 0.0197$.

Thus conclude H_1, that $\beta_4 > 0$.

Thus a 1-sided test is "easier" (more likely to yield a significant result) than a 2-sided test, as before.

2. Confidence Interval for β_k

1. Construction of CI for β_k

The $1 - \alpha$ confidence limits for a coefficient β_k of a multiple regression model are given by

$$b_k \pm t(1 - \propto; n - p)s\{b_k\}$$

Ph.D. Example: find the 95% CI for β_4, the coefficient of age.

$b_4 = .39215$
$s\{b_1\} = .18587$

Choose $\alpha = .05$; then $t(1 - \alpha/2; n - p) = t(0.975; 57) = 2$ (from statistical program or table)
Therefore the 95% CI for β_1 is

Lower bound of CI = .39215 - (2)(0.18587) = 0.020
Upper bound of CI = .39215 + (2)(0.18587) = 0.764

The 95% CI is [0.020, 0.764]. Over repeated sampling, 95 out of 100 confidence intervals will contain β_4. We are 95% confident that this interval contains β_4.

2. Equivalence of CI and 2-sided Test

The $(1 - \alpha)$ CI for β_k and 2-sided hypothesis test on β_k are equivalent in the sense that if the $(1 - \alpha)$ CI for β_k does not include 0, β_k is significant at the α-level in a 2-sided test.

7. CI for E{Y_h}

It is often important to estimate the mean response $E\{Y_h\}$ for given values of the independent variables. The values of the independent variables for which $E\{Y_h\}$ is to be estimated are denoted $X_{h,1}, X_{h,2}, ..., X_{h,\, p-1}$. (This set of values of the X variables may or may not correspond to one of the cases in the data set.)

The estimator of $E\{Y_h\}$ is

$$\hat{Y}_h = b_0 + b_1X_{h,1} + b_2X_{h,2} + \ldots + b_{p-1}X_{h,p-1}$$

The $1 - \alpha$ confidence limits for the mean response $E\{Y_h\}$ are then given by

$$\hat{Y}_h \pm t(1 - \alpha/2;\, n - p)s\{\hat{Y}_h\}$$

where $s\{\hat{Y}_h\}$ is the estimated standard deviation of $\hat{Y}_h$.

The standard error $s\{\hat{Y}_h\}$ of $\hat{Y}_h$ is estimated as

$$s\{\hat{Y}_h\} = (MSE(\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h))^{1/2}$$

$s\{\hat{Y}_h\}$ can be obtained from SAS.

Example: Given the regression for the number of publications since Ph.D., calculate a confidence interval estimate for $E\{Y_h\}$ when time = 9, cits = 41, salary = 52,926 and age = 39 (this is professor #5).

$\hat{Y}_h = -23.203 + 1.619(9) - 0.089(41) + 0.0003(52926) + .392(39) = 18.885$

$s\{\hat{Y}_h\} = 2.112$

Choose $\alpha = .05$; then $t(1 - \alpha/2;n - p) = t(0.975;\, 57) = 2$ (from statistical program or table)

Therefore the 95% CI for $E\{Y_h\}$ is

Lower bound of CI = 18.885 - (2)(2.112) = 14.661
Upper bound of CI = 18.885 + (2)(2.122) = 23.109

The 95% CI is [14.661, 23.109]. Over repeated sampling, 95 out of 100 confidence intervals will contain $E\{Y_h\}$. We are 95% confident that this interval contains $E\{Y_h\}$.

8. Prediction Interval for $Y_{h(new)}$

Given a new observation with values $\mathbf{X}_h$ of the independent variables, the predicted value $Y_{h(new)}$ is estimated as $\hat{Y}_h$, the same as for the mean response. But the variance $s^2\{pred\}$ of $Y_{h(new)}$ is different. The expression for $s^2\{pred\}$ combines the sampling variance of the mean response, estimated as $s^2\{\hat{Y}_h\}$, and the variance of individual observations around the mean response, estimated as MSE, so that

$$s^2\{pred\} = MSE + s^2\{\hat{Y}_h\} = MSE + MSE\ \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h$$

Thus the standard error $s\{pred\}$ is obtained as

$$s\{pred\} = (MSE + s^2\{\hat{Y}_h\})^{1/2} = (MSE + MSE\ \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)^{1/2}$$

The $1 - \alpha$ prediction interval for $Y_{h(new)}$ corresponding to $\mathbf{X}_h$ is

$$\hat{Y}_h \pm t(1 - \alpha/2;\ n - p)\ s\{pred\}$$

$s\{\hat{Y}_h\}$ can be obtained from SAS.

Example: Given the regression for the number of publications since Ph.D., calculate a prediction interval estimate for $E\{Y_h\}$ when time = 9, cits = 41, salary = 52,926 and age = 39 (this is professor #5).

$\hat{Y}_h = -23.203 + 1.619(9) - 0.089(41) + 0.0003(52926) + .392(39) = 18.885$

$s\{\hat{Y}_h\} = 10.557$

Choose $\alpha = .05$; then $t(1 - \alpha/2; n - p) = t(0.975;\ 57) = 2$ (from statistical program or table)

Therefore the 95% CI for $E\{Y_h\}$ is

Lower bound of CI = 18.885 - (2)(10.557) = -2.229
Upper bound of CI = 18.885 + (2)(10.557) = 39.999

The 95% CI is [-2.229, 39.999]. Over repeated sampling, 95 out of 100 confidence intervals will contain $Y_{h(new)}$. We are 95% confident that this interval contains $Y_{h(new)}$.

9. Other Elements of the Multiple Regression

1. Standardized Regression Coefficients

The standardized regression coefficient b_k^* is calculated as:

$b_k^* = b_k(s(X_k)/s(Y))$

where $s(X_k)$ and $s(Y)$ denote the sample standard deviations of X_k and Y, respectively.

The standardized coefficient b_k^* measures the change in standard deviations of Y associated with an increase of one standard deviation of X.

Standardized coefficients permit comparisons of the relative strengths of the effects of different independent variables, measured in different ***metrics*** (= units).

2. Tolerance or Variance Inflation Factor

SAS provides a diagnostic measure of the collinearity (linear association) of a predictor with the other predictors in the model, either the *tolerance* (TOL) or the *variance inflation factor* (VIF).

1. Tolerance (TOL)

$$TOL = 1 - R_k^2$$

where R_k^2 is the R-square of the regression of X_k on the other p - 2 predictors in the regression and a constant. TOL can vary between 0 and 1;

- TOL close to 1 means that R_k^2 is close to 0, indicating that X_k is not highly correlated with the other predictors in the model
- TOL close to 0 means that X_k is highly correlated with the other predictors; one then says that X_k is *collinear* with the other predictors

A common rule of thumb is that TOL < .1. This is an indication that collinearity may unduly influence the results.

2. Variance Inflation Factor

$$VIF = (TOL)^{-1} = (1 - R_k^2)^{-1}$$

The variance inflation factor is the inverse of the tolerance. Large values of VIF therefore indicate a high level of collinearity.

The corresponding rule of thumb is that VIF > 10. This is an indication that collinearity may unduly influence the results.

Collinearity is discussed further later.

10. The General Linear Model

The term *general linear model* is used for multiple regression models that include variables other than first powers of different predictors. The X variables can also represent

- different powers of a single variable (polynomial regression)
- interaction terms represented as the product of two or more variables
- qualitative (categorical) variables represented by one or more indicators (variables with values 1 or 0, *aka* "dummy variables")
- mathematical transformations of variables