Multiple Regression

Multiple Regression analysis, which is an extension of simple linear regression, allow researchers to improve their predictive power by using two or more indep. variables.

The basic equation is:

$$\hat{y} = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

The Standardized Multiple Regression Equation

In multiple regression, the independent variables are typically in different units of measure. The regression coefficients, then, necessarily incorporate differen in the units of measures, and so the b weights in the basic regression equation cannot directly be compared.

To address this issue, the regression equation is sometimes presented in the following standardized form:

$$z_{\hat{y}} = \beta_1 z_{x_1} + \beta_2 z_{x_2} + \ldots + \beta_k z_{x_k}$$

where $z_{\hat{y}}$ = predicted value of standard score for Y

β_1 to β_k: standardized reg. weights for k indep. vari

z_{x_1} to z_{x_k}: standard scores for k independent variable

Discussion of p260 example.

Adjusted R^2

$$\text{Adjust } \tilde{R}^2 = 1 - (1-R^2)\left[\frac{N-1}{N-k-1}\right]$$

where N: # of cases
k: # of predictors.

Tests of significance for Multiple Regression

Thus far we have considered multiple regression in a purely descriptive sense; the regression equation and R are specific to the sample being used. However, researchers are always interested in generalizing their results to the population, and therefore tests of significance are needed to facilitate the required inferences.

Tests of the Overall Equation & R

The most basic statistical test in multiple regress is a test of the null hypothesis that the population value of R is zero.

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

$$F = \frac{\left(SSreg / d.f.reg\right)}{SSresidual / d.f.residual}$$

Tests of the Reg. coeff.

$$t_k = \frac{\hat{\beta}_k}{s.e(\hat{\beta}_k)}$$

rule of thumb:
$|t_k| > 2$ significant predictor

Tests for Added Predictors

$$F = \frac{\left[\left(R^2_{yk_1} - R^2_{yk_2}\right) / (k_1 - k_2)\right]}{\left[\left(1 - R^2_{yk_1}\right) / (N - k_1 - 1)\right]}$$ ~ F distributed with $(k_1 - k_2, N - k_1 - 1)$ d.f.

where $R^2_{yk_1}$: R^2 with K_1 predictors

k_1 : larger of the 2 sets of predictors

$R^2_{yk_2}$: R^2 with k_2 predictors

k_2 : Smaller of the 2 sets of predictors.

See computer printout for discussion page 8.23

Three methods For entering predictors in Multiple Regression

Method 1 : Simultaneous Multiple Regression Model

This strategy is most appropriate when there is no theoretical basis for considering that any particular variable is causally prior to another, and when

8.14a

How can we tell if adding (or removing) a certain set of X variables causes a *significant* increase (or decrease) in R^2?

The Partial F Test

Consider the situation in which the personnel director is trying to determine whether to retain three variables (X_6, X_7, X_8) as predictors of a person's performance on a CPA exam. We know one thing—R^2 *will* be higher with these three variables included in the model. If we do not observe a *significant* increase, however, our advice would be to *remove* these variables from the analysis. To determine the extent of this increase, we use another F test.

We define two models—one contains X_6, X_7, X_8, and one does not.

Complete model: uses all predictor variables, including X_6, X_7, and X_8

Reduced model: uses the same predictor variables as the complete model except X_6, X_7, and X_8

Also, let

R_c^2 = the value of R^2 for the complete model

R_r^2 = the value of R^2 for the reduced model

Do X_6, X_7, and X_8 contribute to the prediction of Y? We will test

H_0: $\beta_6 = \beta_7 = \beta_8 = 0$ (they do not)

H_a: at least one of the β's $\neq 0$ (at least one of them does)

The test statistic here is

$$F = \frac{(R_c^2 - R_r^2)/v_1}{(1 - R_c^2)/v_2} \qquad (15\text{-}15)$$

where v_1 = number of β's in H_0, and $v_2 = n - 1 -$ (number of X's in the complete model).

For this illustration, $v_1 = 3$ because there are three β's contained in H_0. Assuming that there are eight variables in the complete model, then $v_2 = n - 1 - 8 = n - 9$. Here, n is the total number of observations (rows) in your data. This F statistic measures the *partial* effect of these three variables; it is a **partial F statistic.**

Equation 15-15 resembles the F statistic given in equation 15-14, which we used to test H_0: all β's = 0. If all the β's are zero, then the reduced model consists of only a constant term and the resulting R^2 will be zero; that is, $R_r^2 = 0$. Setting $R_r^2 = 0$ in equation 15-15 produces equation 15-14, where $v_1 = k$ and $v_2 = n - k - 1$.

These variables (as a group) contribute significantly if the computed partial F value in equation 15-15 exceeds F_{α, v_1, v_2} from Table A-7.

Example 15.6

The personnel director gathered data from 30 individuals using all eight of the independent variables. These data were entered into a computer, and a multiple linear regression analysis was performed. The resulting R^2 was .857.

Next, variables X_6, X_7, and X_8 were omitted, and a second regression analysis was performed. The resulting R^2 was .824. Do the variables X_6, X_7, and X_8 (height, weight, and age) appear to have any predictive ability? Use $\alpha = .10$.

Solution Here, $n = 30$ and

$R_c^2 = .857$ (complete model)

$R_r^2 = .824$ (reduced model)

Based on the previous discussion, the value of the partial F statistic is

$$F^* = \frac{(.857 - .824)/3}{(1 - .857)/(30 - 1 - 8)}$$
$$= \frac{.033/3}{.143/21}$$
$$= 1.61$$

The procedure is to reject H_0: $\beta_6 = \beta_7 = \beta_8 = 0$ if $F^* > F_{.10,3,21} = 2.36$. The computed F value does not exceed the table value, so we fail to reject H_0. We conclude that these variables should be removed from the analysis because including them in the model fails to produce a significantly larger R^2.

The partial F test also can be used to determine the effect of adding a *single* variable to the model.

Example 15.7 Using the real-estate data analyzed in example 15.2, determine whether X_2 = family size contributes to the prediction of home size, given that X_1 = income and X_3 = years of education are included in the model. Use a significance level of $\alpha = .10$.

Solution We will test the hypotheses

H_0: $\beta_2 = 0$ (if X_1 and X_3 are included)

H_a: $\beta_2 \neq 0$.

The complete model uses X_1, X_2, and X_3. Using Figure 15.4,

$R_c^2 = .905$

The reduced model uses X_1 and X_3 only. Figure 15.7 shows the MINITAB output for this, and

$R_r^2 = .801$

Figure 15.7

MINITAB output using X_1 = income and X_3 = years of education as predictors.

```
MTB > REGRESS Y IN C1 USING 2 PREDICTORS IN C2, C4 <- X3
                                                  |_ X1

The regression equation is
C1 = 10.5 + 0.373 C2 - 0.494 C4

Predictor        Coef        Stdev     t-ratio
Constant       10.465        3.148        3.32
C2            0.37315      0.07168        5.21
C4            -0.4938       0.2787       -1.77

s = 2.735          [R-sq = 80.1%] <- R^2     R-sq(adj) = 74.4%

Analysis of Variance

SOURCE        DF          SS          MS
Regression     2      210.06      105.03
Error          7       52.34        7.48
Total          9      262.40
```

research problem

Method II : Hierarchical Multiple regression , independant variables are entered into the model in a series of steps, and the order of entry is controlled by the researcher. The order of entry of predictors should be based on logical or theoretical considerations.

A common reason for using hierarchical regression is to examine the effect of certain independant variables after the effect of other variables have been controlled.

Method III : Stepwise Multiple Regression

The basic stepwise model involves successive steps in which predictors are entered, one at a time, in the order in which the increment to R is greatest. The computer, rather than the researcher, determines the order of entry of predictor variables.

Nature of the independent variables

Multiple regression is used to predict a dependant variable that is measured on an interval or ratio-level scale .

Interval & Ratio-level Indep. variables

The use of variables measured on interval and ratio scales is straightforward. The raw data values are used directly in the analysis.

Nominal-level Independent Variables

Nominal-level variables must be coded in a manner that allows for appropriate interpretation of the regression coefficients.
For a k-categories variable, one has to create k-1 dummy variables.

e.g.

Race	Original code	White X_1	African American X_2	Hispanic X_3
White	1	1	0	0
African American	2	0	1	0
Hispanic	3	0	0	1
other	4	0	0	0

Since there need to be k-1 new variables, there is always a category that is omitted & served as a reference group. In this example 'other' is the reference group.

ANOVA VIA Regression

See computer printout for discussion. (p260)

Following is an example from Polit's book Page 260.

```
-> get file='d:\stat601.14\p260.sav'.

-> list variables=all.
```

SUBJECT	GPA_U	GRE_V	GRE_Q	MOTIV	GPA_G	GROUP	FINISH
1.00	3.40	600.00	540.00	75.00	3.60	1.00	1.00
2.00	3.10	510.00	480.00	70.00	3.00	1.00	1.00
3.00	3.70	650.00	710.00	85.00	3.90	1.00	1.00
4.00	3.20	530.00	450.00	60.00	2.80	1.00	.00
5.00	3.50	610.00	500.00	90.00	3.70	1.00	1.00
6.00	2.90	540.00	620.00	60.00	2.60	2.00	.00
7.00	3.30	530.00	510.00	75.00	3.40	2.00	1.00
8.00	2.90	540.00	600.00	55.00	2.70	2.00	.00
9.00	3.40	550.00	580.00	75.00	3.30	2.00	.00
10.00	3.20	700.00	630.00	65.00	3.50	2.00	1.00
11.00	3.70	630.00	700.00	80.00	3.60	3.00	1.00
12.00	3.00	480.00	490.00	75.00	2.80	3.00	1.00
13.00	3.10	530.00	520.00	60.00	3.00	3.00	.00
14.00	3.70	580.00	610.00	65.00	3.50	3.00	1.00
15.00	3.90	710.00	660.00	80.00	3.80	3.00	1.00
16.00	3.50	500.00	480.00	75.00	3.20	4.00	1.00
17.00	3.10	490.00	510.00	60.00	2.40	4.00	.00
18.00	2.90	560.00	540.00	55.00	2.70	4.00	.00
19.00	3.20	550.00	590.00	65.00	3.10	4.00	.00
20.00	3.40	600.00	550.00	70.00	3.60	4.00	1.00

Number of cases read: 20 Number of cases listed: 20

```
-> compute white=0.

-> compute africa=0.

-> compute hispanic=0.

-> if (group=1)white=1.

-> if (group=2)africa=1.

-> if (group=3)hispanic=1.

-> freq var=group,white to africa,finish.
```

8.18

GROUP race

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
white	1.00	5	25.0	25.0	25.0
african-american	2.00	5	25.0	25.0	50.0
hispanic	3.00	5	25.0	25.0	75.0
other	4.00	5	25.0	25.0	100.0
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

WHITE

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	.00	15	75.0	75.0	75.0
	1.00	5	25.0	25.0	100.0
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

AFRICA

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	.00	15	75.0	75.0	75.0
	1.00	5	25.0	25.0	100.0
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

FINISH Finish Grad School

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
no	.00	8	40.0	40.0	40.0
yes	1.00	12	60.0	60.0	100.0
	Total	20	100.0	100.0	

Valid cases 20 Missing cases 0

```
-> oneway gpa_g by group(1,4)/statistics=all.
```

- - - - - O N E W A Y - - - - -

Variable GPA_G Graduate GPA
By Variable GROUP race

Analysis of Variance

Source	D.F.	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	3	.5460	.1820	.9180	.4544
Within Groups	16	3.1720	.1983		
Total	19	3.7180			

Group	Count	Mean	Standard Deviation	Standard Error	95 Pct Conf Int for Mean		
white	5	3.4000	.4743	.2121	2.8110	TO	3.9890
african-	5	3.1000	.4183	.1871	2.5806	TO	3.6194
hispanic	5	3.3400	.4219	.1887	2.8161	TO	3.8639
other	5	3.0000	.4637	.2074	2.4243	TO	3.5757
Total	20	3.2100	.4424	.0989	3.0030	TO	3.4170
Fixed Effects Model			.4453	.0996	2.9989	to	3.4211
Random Effects Model				.0996	2.8932	to	3.5268

Warning - between component variance is negative
It was replaced by 0.0 in computing above Random Effects Measures

Random Effects Model - estimate of between component variance -3.25E-03

8.20

GROUP	MINIMUM	MAXIMUM
white	2.8000	3.9000
african-	2.6000	3.5000
hispanic	2.8000	3.8000
other	2.4000	3.6000
TOTAL	2.4000	3.9000

Levene Test for Homogeneity of Variances

Statistic	df1	df2	2-tail Sig.
.0823	3	16	.969

ANOVA VIA Regression

```
-> regression variables=gpa_g white africa hispanic/
->        dependent=gpa_g/enter white africa hispanic.
```

* * * * M U L T I P L E   R E G R E S S I O N * * * *

Listwise Deletion of Missing Data

Equation Number 1   Dependent Variable.. GPA_G   Graduate GPA

Block Number 1. Method: Enter   WHITE   AFRICA   HISPANIC

Variable(s) Entered on Step Number
1.. HISPANIC
2.. AFRICA
3.. WHITE

Multiple R	.38321
R Square	.14685
Adjusted R Square	-.01311
Standard Error	.44525

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	3	.54600	.18200
Residual	16	3.17200	.19825

F = .91803   Signif F = .4544 ← Compare with output on Page 3 from oneway ANOVA.

------------------ Variables in the Equation ------------------

Variable	B	SE B	Beta	T	Sig T
WHITE	.400000	.281603	.401718	1.420	.1747
AFRICA	.100000	.281603	.100429	.355	.7271
HISPANIC	.340000	.281603	.341460	1.207	.2448
(Constant)	3.000000	.199123		15.066	.0000

End Block Number 1   All requested variables entered.

```
-> regression variables=gpa_u gre_v gre_q motiv gpa_g/
->    statistics=default,cha/
->    dependent=gpa_g/enter gpa_u gre_v gre_q motiv.
```

← Simultaneous Multiple Regression model

*** * * *   M U L T I P L E   R E G R E S S I O N   * * * ***

Listwise Deletion of Missing Data

Equation Number 1   Dependent Variable..   GPA_G   Graduate GPA

Block Number 1. Method: Enter   GPA_U   GRE_V   GRE_Q   MOTIV

Variable(s) Entered on Step Number

1..	MOTIV	Movitation
2..	GRE_Q	GRE-Quant
3..	GRE_V	GRE-Verbal
4..	GPA_U	Undergrad GPA

Multiple R	.94062		
R Square	.88476	R Square Change	.88476
Adjusted R Square	.85403	F Change	28.79035
Standard Error	.16901	Signif F Change	.0000

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	4	3.28953	.82238
Residual	15	.42847	.02856

F =   28.79035   Signif F = .0000

------------------ Variables in the Equation ------------------

Variable	B	SE B	Beta	T	Sig T
GPA_U	.540282	.224071	.362042	2.411	.0292
GRE_V	.003191	9.5846E-04	.469534	3.329	.0046
GRE_Q	-5.81444E-04	7.4879E-04	-.098479	-.777	.4495
MOTIV	.015213	.005803	.341521	2.622	.0192
(Constant)	-1.126417	.450169		-2.502	.0244

End Block Number   1   All requested variables entered.

8.22


```
-> regression variables=gpa_u gre_v gre_q motiv gpa_g/
->    statistics=default,cha/
->    dependent=gpa_g/enter gpa_u motiv/enter gre_v gre_q.
```

Hierarchical multiple regression.

Examine gre-v, gre-q effect on gpa-g controlling for motiv, gpa-u

```
          * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Listwise Deletion of Missing Data

Equation Number 1    Dependent Variable..   GPA_G   Graduate GPA

Block Number  1.  Method:  Enter      GPA_U     MOTIV


Variable(s) Entered on Step Number
   1..     MOTIV      Movitation
   2..     GPA_U      Undergrad GPA
```

R^2_{Yk2} →

```
Multiple R              .88334
R Square                .78029       R Square Change      .78029
Adjusted R Square       .75444       F Change           30.18656
Standard Error          .21921       Signif F Change      .0000

Analysis of Variance
                        DF      Sum of Squares      Mean Square
Regression               2 → k2        2.90110          1.45055
Residual                17              .81690           .04805

F =        30.18656       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable                 B          SE B        Beta          T  Sig T

GPA_U              .924207       .246158     .619309      3.755  .0016
MOTIV              .014461       .007348     .324641      1.968  .0656
(Constant)        -.853166       .568086                 -1.502  .1515


------------- Variables not in the Equation -------------

Variable       Beta In   Partial   Min Toler          T  Sig T

GRE_V          .402688   .674100     .349125      3.650  .0022
GRE_Q          .159262   .296466     .376611      1.242  .2322


End Block Number   1   All requested variables entered.


          * * * * * * * * * * * * * * * * * * * * * * * * * * * *
```