Correlation and simple regression

In this section we discuss the population and sample correlation coefficients. These quantities are measures of linear association between two variables. Pearson's r is appropriate when both the independent and dependent variables are measured on approximately an interval level or on a ratio level. The correlation coefficient (sometimes called the "simple correlation coeff." or the "Pearson's r" or the "product-moment correlation coefficient") is calculated as

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

e.g. Can a person's weight as a young adult be predicted from his/her birth wt.?

subject	Birth wt (oz), X	Adult (lb) Y
1	112	120
2	118	135
3	118	125
4	118	138
5	101	115
6	93	118
7	128	145
8	118	120
9	140	144
10	108	130
11	118	120
12	118	135
13	123	120
14	112	132
15	127	125
16	112	135
	134	

p8.1a

Use Minitab to perform data on Page 8.1

```
MTB > read 'p81dat' c1 c2

MTB > prin c1-c2
```

ROW	C1	C2
1	112	120
2	118	135
3	118	125
4	118	138
5	101	115
6	93	118
7	128	145
8	118	120
9	140	144
10	108	130
11	118	120
12	123	135
13	112	120
14	127	132
15	112	125
16	134	135

```
MTB > desc c1 c2
```

	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
C1	16	117.50	118.00	117.64	11.73	2.93
C2	16	128.56	127.50	128.36	9.49	2.37

	MIN	MAX	Q1	Q3
C1	93.00	140.00	112.00	126.00
C2	115.00	145.00	120.00	135.00

```
MTB > let c3=c1-117.5
MTB > let c4=c2-128.56
MTB > let c5=c3*c4
MTB > let c6=c3*c3
MTB > let c7=c4*c4
MTB > prin c1-c7
```

ROW	C1	C2	C3	C4	C5	C6	C7
1	112	120	-5.5	-8.5600	47.080	30.25	73.274
2	118	135	0.5	6.4400	3.220	0.25	41.474
3	118	125	0.5	-3.5600	-1.780	0.25	12.674
4	118	138	0.5	9.4400	4.720	0.25	89.114
5	101	115	-16.5	-13.5600	223.740	272.25	183.874
6	93	118	-24.5	-10.5600	258.720	600.25	111.514
7	128	145	10.5	16.4400	172.620	110.25	270.274
8	118	120	0.5	-8.5600	-4.280	0.25	73.274
9	140	144	22.5	15.4400	347.400	506.25	238.394
10	108	130	-9.5	1.4400	-13.680	90.25	2.074
11	118	120	0.5	-8.5600	-4.280	0.25	73.274
12	123	135	5.5	6.4400	35.420	30.25	41.474
13	112	120	-5.5	-8.5600	47.080	30.25	73.274
14	127	132	9.5	3.4400	32.680	90.25	11.834
15	112	125	-5.5	-3.5600	19.580	30.25	12.674
16	134	135	16.5	6.4400	106.260	272.25	41.474

```
MTB > let k1=sum(c5)
MTB > let k2=sum(c6)
MTB > let k3=sum(c7)
MTB > prin k1,k2,k3
K1        1274.50
K2        2064.00
K3        1349.94
MTB > let k4=k1/sqrt(k2*k3)
MTB > prin k4
K4        0.763534
MTB > corr c1 c2

Correlation of C1 and C2 = 0.764

MTB > regr c2 1 c1

The regression equation is
C2 = 56.0 + 0.617 C1

Predictor        Coef        Stdev    t-ratio        p
Constant        56.01        16.48       3.40    0.004
C1             0.6175       0.1396       4.42    0.001

s = 6.341        R-sq = 58.3%      R-sq(adj) = 55.3%

Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     1      786.99      786.99     19.57    0.001
Error         14      562.95       40.21
Total         15     1349.94


MTB > stop
```

Hypothesis about the correlation Coefficient

The correlation coefficient, r, which we calculate from a sample, is an estimate of a population parameter namely the correlation coefficient in the population that was sampled. This parameter is denoted by ρ. If we wish to ask whether there is, in fact, a correlation between Y and X in the population, we can test:

$H_0: \rho = 0$

$H_A: \rho \neq 0$ Test statistic $t_{calc} = \dfrac{r\sqrt{N-2}}{\sqrt{1-r^2}}$

$\sim t_{N-2}$ d.f.

e.g. Calculate the sample correlation coefficient, and testing $H_0: \rho = 0$. The data are wing and tail lengths among birds of a particular species

Wing length (cm) X	Tail length (cm) Y
10.4	7.4
10.8	7.6
11.1	7.9
10.2	7.2
10.3	7.4
10.2	7.1
10.7	7.4
10.5	7.2
10.8	7.8
11.2	7.7
10.6	7.8
11.4	8.3

8.2a

Pearson Corelation for P8.2 .

```
. use minitab
initiating minitab (MiniTab 7.2) for use
Variable set: MTBMACRO=/pl2a/minitab7.2/macros
acs4:/u5/tfung/stat601.14--> minitab
 MINITAB RELEASE 7.2.3 *** IBM Risc System 6000 Version
 (C) Copyright 1989 Minitab, Inc. - ALL RIGHTS RESERVED
 U.S. FEDERAL GOVERNMENT USERS SEE HELP FGU
 MARCH 31, 1999 *** UNIV OF CALGARY
 STORAGE AVAILABLE 248000
      Type NEWS for information on new features in
      Minitab Release 7.2
> read 'data1' c1 c2
      12 ROWS READ

  ROW     C1     C2

    1   10.4    7.4
    2   10.8    7.6
    3   11.1    7.9
    4   10.2    7.2
   .   .   .

 MTB > desc c1 c2

                 N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
C1              12   10.683   10.650   10.660    0.395    0.114
C2              12    7.567    7.500    7.540    0.350    0.101

               MIN      MAX       Q1       Q3
C1          10.200   11.400   10.325   11.025
C2           7.100    8.300    7.250    7.800

MTB > let c3=c1-10.683
MTB > let c4=c2-7.567
MTB > let c5=c3*c4
MTB > let c6=c3*c3
MTB > let c7=c4*c4
MTB > let k1=sum(c5)
MTB > let k2=sum(c6)
MTB > let k3=sum(c7)
MTB > let k4=sqrt(k2)
MTB > let k5=sqrt(k3)
MTB > let k6=k1/(k4*k5)
MTB > prin k6
K6        0.870353
MTB > corr c1 c2

Correlation of C1 and C2 = 0.870
```

$r = 0.866$

$H_0: \rho = 0, \quad H_A: \rho \neq 0$ use $\alpha = 0.05$

$$t_{calc} = \frac{0.866\sqrt{12-2}}{\sqrt{1-(0.866)^2}} = \frac{0.866\sqrt{10}}{\sqrt{0.25}}$$

$$= 5.48$$

$$t_{\frac{\alpha}{2}}(10) = t_{0.025}(10) = 2.228$$

$t_{calc} = 5.48 > 2.228$ reject H_0 at 0.05 level

If one desires to test $H: \rho = \rho_0$ for any $\rho_0 \neq 0$
$H_A: \rho \neq \rho_0$ & $\rho_0 \neq \pm 1$,
we have to use Fisher's z-transformation,
which is $z(r) = \frac{1}{2} \log_e \frac{1+r}{1-r}$

Under H_0, $z(r) \sim N\left(z(\rho_0), \frac{1}{N-3}\right)$

The test statistic is then computed as:

$$\zeta = \frac{z(r) - z(\rho_0)}{\left(\frac{1}{\sqrt{N-3}}\right)}$$

If $|\zeta| > \zeta_{\frac{\alpha}{2}}$, then reject H_0 at α level.

$\therefore$ $100(1-\alpha)\%$ confidence interval for $z(\rho)$:

$$-z_{\alpha/2} < \frac{z(r) - z(\rho)}{\sigma_{z(r)}} < z_{\alpha/2}$$

$$\Rightarrow z(r) - z_{\alpha/2}\,\sigma_{z(r)} < z(\rho) < z(r) + z_{\alpha/2}\,\sigma_{z(r)}$$

$$\Rightarrow z(r) - z_{\alpha/2}\frac{1}{\sqrt{N-3}} < z(\rho) < z(r) + z_{\alpha/2}\frac{1}{\sqrt{N-3}} \quad (1)$$

Use the inverse Fisher transformation:

$$r = \frac{e^{2z(r)} - 1}{e^{2z(r)} + 1}$$

, the confidence interval for ρ are obtained.

e.g. (a) Use the data 'wing & tail lengths' example

$H_0: \rho = 0.75$, $\quad r = 0.866$

$H_A: \rho \neq 0.75$ $\quad N = 12$ $\quad \alpha = 0.05$

(b) construct 95% C.I. of ρ.

Solution: $z(r) = z(.866) = \frac{1}{2}\log_e \frac{1+0.866}{1-0.866}$

$$= \frac{1}{2}\log_e \frac{1.866}{0.134} = 1.3169$$

$$z(\rho_0) = z(0.75) = \frac{1}{2}\log_e \frac{1+0.75}{1-0.75} = \frac{1}{2}\log_e \frac{1.75}{0.25}$$

$$= 0.9729$$

$$z_{calc} = \frac{z(.866) - z(.75)}{\left(\frac{1}{\sqrt{9}}\right)} = (1.3169 - .9729) \times 3$$

$$= 1.032$$

$z_{0.025} = 1.96$, $z_{calc} = 1.032 < 1.96$

DO NOT reject H_0 at 0.05 level.

(b) $Z(r) = 1.3169$, $N = 12$, $z_{0.025} = 1.96$

from eqt (1), we have

$$1.3169 - \frac{1.96}{3} \leq Z(\rho) \leq 1.3169 + \frac{1.96}{3}$$

$$\Rightarrow \quad 0.664 \leq Z(\rho) \leq 1.970$$

$$\Rightarrow \quad 0.664 \leq \frac{1}{2} \log_e \frac{1+\rho}{1-\rho} \leq 1.970$$

$\therefore$ 95% C.I. for $Z(\rho) = (0.664, 1.970)$

$$\Rightarrow 95\% \text{ C.I. for } \rho = \left(\frac{e^{2 \times 0.664} - 1}{e^{2 \times 0.664} + 1}, \frac{e^{2 \times 1.97} - 1}{e^{2 \times 1.97} + 1} \right)$$

$$= \left(\frac{e^{1.328} - 1}{e^{1.328} + 1}, \frac{e^{3.94} - 1}{e^{3.94} + 1} \right) = \left(\frac{3.773 - 1}{3.773 + 1}, \frac{51.419 - }{51.419 + 1} \right)$$

$$= (0.581, 0.962)$$

Comparing Two Correlation Coefficients :

Hypothesis about two correlation coefficients may be approached by the use of

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}}$$

e.g. For a sample of 15 bird wing & tail length a correlation coeff. of 0.843 was calculated. A sample of 18 such measurements from a second bird species yielded a correlation coefficient of 0.78 Let us test the equality of two population coefficients.

$H_0: \rho_1 = \rho_2$

$H_A: \rho_1 \neq \rho_2$ $\qquad \alpha = 0.05$

$r_1 = .843$

$n_1 = 15$

$z_1 = z(r_1) = \frac{1}{2} \log_e \frac{1 + .843}{1 - 0.843} = \frac{1}{2} \log_e \frac{1.843}{0.157} = 1.2314$

$r_2 = 0.784$

$n_2 = 18$

$z_2 = z(r_2) = \frac{1}{2} \log_e \frac{1 + .784}{1 - .784} = \frac{1}{2} \log_e \frac{1.784}{0.216} = 1.0557$

$$Z = \frac{1.2314 - 1.0557}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0.1758}{0.3873} = 0.454$$

$z_{\alpha/2} = 1.96$ $\qquad 0.454 < 1.9$

DO NOT reject H_0 at 0.05 level

Spearman Rank Correlation

Spearman's Rank correlation is a nonparametric analog of Pearson's r. A correlation coefficient that makes use of the rank-order information in the observations, and not their actual values, was proposed by the British psychologist Charles Spearman in 1904. It is typically used when the dependent variable is measured on the ordinal scale, or when one or both variables being correlated is severely skewed or has an outlier. It is also preferred by some researchers when the number of case is less than 30.

If we assign ranks (1, 2, 3 ... n) to the observational units with respect to the values of each variable separately and then compute the Pearson correlation of the ranks, the result is Spearman's rank correlation coefficient:

$$r_s = r_{ranks}$$

Spearman's r_s is a Pearson correlation coefficient calculated on ranks (r_{ranks}) instead of on raw scores (r_{xy}). It follows that Spearman's r_s is an index of the linear relationship of the ranks of the variables, and that its value is limited to the interval from -1 to 1, inclusive.

In assigning ranks to the n values of the two variables, it is immaterial whether the ranks (1, 2, ... n) are assigned from largest to smallest or from smallest to largest, as long as the direction of ranking

is the same for both variables. However, it is conventional, and generally less conducive to confusion, to assign the rank of 1 to the smallest observation and successively larger rank numbers to successively larger observations. If there are tied observations, they should be assigned the mean of the ranks they would have received if they had not been exactly tied.

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where d_i is the difference between the two ranks assigned to the two variables for the ith observational unit.

e.g. Owners of show jumping competition horses generally insure their horses, in view of the unpredictability of injury and disease. To what extent does the amount of insurance depend on the # of competitions won by horses in recent yrs?

Horse	# winning (X)	X-rank	insured value Y	Y-rank	d_i	d_i^2
1	19	5	7.5	5	0	0
2	33	9	10	7	2	4
3	35	10	15	9	1	1
4	20	6.5	25	10	-3.5	12.25
5	4	1	3.5	1	0	0
6	16	4	5	3.5	0.5	.25
7	15	3	5	3.5	-0.5	.25
8	12	2	4	2	0	0
9	20	6.5	10	7	-.5	.25
10	29	8	10	7	1	1
					0	19

$$r_s = 1 - \frac{6 \times 19}{10\,(10^2-1)} = 0.883$$

The null hypothesis when using Spearman's Correlation is that there is <u>NO</u> linear relationship between the two sets of ranks. i.e. that the population correlation (ρ_s) is 0. Table B-8 (textbook) of Appendix B, with a sample size of 10, the critical value for $\alpha = 0.05$ (for a two-sided test) is 0.648. $0.883 > 0.648$ reject H_0 at 0.05 level. Note that the table of critical values for Spearman correlation can only used when $N \leq 30$ and $N \geq 5$. When there are more than 30 cases, a reasonably good approximation of the significance test can be obtained by using the table of critical value for Pearson's r (Table b-6) with $N-2$ degree of freedom.

8.8c

```
execute.
list variables=id x rx y ry.
```

List:spearman.spo

id	x	Rx	y	Ry
1.00	19.00	5.000	7.50	5.000
2.00	33.00	9.000	10.00	7.000
3.00	35.00	10.000	15.00	9.000
4.00	20.00	6.500	25.00	10.000
5.00	4.00	1.000	3.50	1.000
6.00	16.00	4.000	5.00	3.500
7.00	15.00	3.000	5.00	3.500
8.00	12.00	2.000	4.00	2.000
9.00	20.00	6.500	10.00	7.000
10.00	29.00	8.000	10.00	7.000

Number of cases read: 10 Number of cases listed: 10

```
correlation rx with ry.
```

Correlations

Correlations

		Ry RANK of y
Rx RANK of x	Pearson Correlation	.883
	Sig. (2-tailed)	.001
	N	10

```
nonpar corr x with y.
```

Nonparametric Correlations

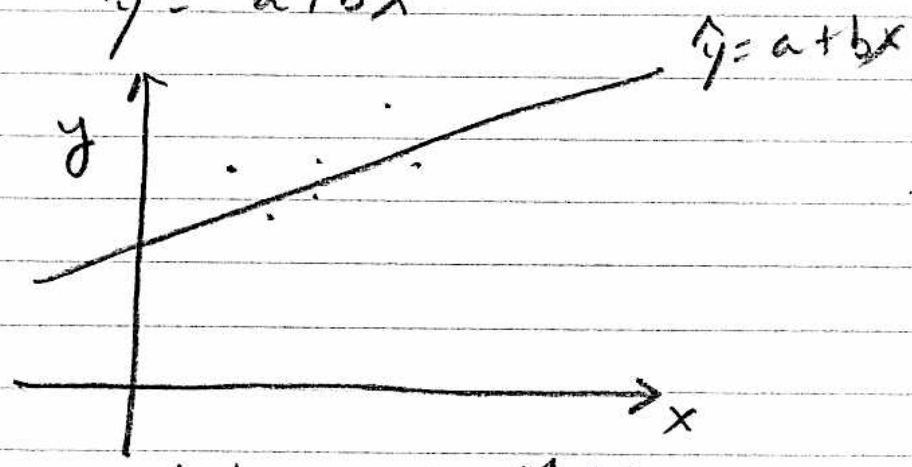
Correlations

			y
Spearman's rho	x	Correlation Coefficient	.883
		Sig. (2-tailed)	.001
		N	10

Simple regression

In research studies, relationships between variables are often linear, but they are never perfect. When a correlation between two variables is not perfect, the statistical technique of regression can be used to identify the straight line that runs through the data points with the best possible fit, as determined by statistical criterion known as least squares. The regression equation is the formula for the best-fitting straight line to characterize the linear relationships between X and y. The basic linear equation is:

$$\hat{y} = a + bX$$



Find a, & b such that

$$\sum (y_i - \hat{y}_i)^2 \text{ is minimized.}$$

By Calculus, $$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$a = \bar{y} - b\bar{x}$$

e.g.

Mid term	$x_i - \bar{x}$	$(x_i-\bar{x})^2$	Final	$y_i - \bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
2	-3	9	3	-3	9
6	1	1	7	1	1
5	0	0	6	0	0
9	4	16	8	2	8
7	2	4	9	3	6
9	4	16	10	4	16
3	-2	4	4	-2	4
4	-1	1	6	0	0
1	-4	16	2	-4	16
4	-1	1	5	-1	1
$\bar{x} = 5.0$	0	68	$\bar{y} = 6.0$	0	61

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{61}{68} = 0.897$$

$$a = \bar{y} - b\bar{x} = 6.0 - (0.897)(5.0) = 1.515$$

$$\hat{y} = 1.515 + 0.897x$$

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

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

$$= SS_{error} + SS_{Reg}$$

SS_{error} : Sum of square due to error/residual

SS_{Reg} : Sum of square due to regression

$$R^2 = \frac{SS_{reg}}{SS_T}$$

R^2 is defined as the coefficient of determination which is the proportion of the total variation in y that can be accounted for by the linear relation with x.

The coefficient of determination takes on values from 0 (or 0%) when x and y are <u>NOT</u> linearly related, to 1 (or 100%) when there is a perfect straight-line relationship between x and y in the data.

SPSS syntax for regression:

```
regression  variables= y  x / dependent = y / enter x.
```

Discussion of computer printout in Polit's book p247.

8.11a

This example from Polit's book Page 247:

```
-> get file='d:\stat601.14\p247.sav'.

-> regression variables=final midterm/
->          descriptive=MEAN STDDEV CORR COV/
->          dependent=final/enter midterm.
```

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

Listwise Deletion of Missing Data

	Mean	Std Dev	Label
FINAL	6.000	2.582	
MIDTERM	5.000	2.749	

N of Cases = 10

Correlation, Covariance:

	FINAL	MIDTERM
FINAL	1.000	.955
	6.667	6.778
MIDTERM	.955	1.000
	6.778	7.556

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

Equation Number 1 Dependent Variable.. FINAL

Descriptive Statistics are printed on Page 2

Block Number 1. Method: Enter MIDTERM

Variable(s) Entered on Step Number
1.. MIDTERM

Multiple R	.95499
R Square	.91201
Adjusted R Square	.90101
Standard Error	.81236

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	1	54.72059	54.72059
Residual	8	5.27941	.65993

8.11b

F = 82.91922 Signif F = .0000

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

Variable	B	SE B	Beta	T	Sig T
MIDTERM	.897059	.098513	.954992	9.106	.0000
(Constant)	1.514706	.555529		2.727	.0260

End Block Number 1 All requested variables entered.

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 differences in the units of measure, 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:

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

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

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

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

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.

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

$$F = \frac{(SS_{reg}/d.f._{reg})}{SS_{residual}/d.f._{residual}}$$

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

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

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 all independent variables are of the same

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

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

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%] ← R²     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 : <u>Hierarchical multiple regression</u>, independent 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 independent variables after the effect of other variables have been controlled.

Method III : <u>Stepwise Multiple Regression</u>

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.

<u>Nature of the independent variables</u>

Multiple regression is used to predict a dependent 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 X1	African Amer. X2	Hispanic X3
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 one way 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.

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

Examine gre-V, gre-Q effect on gpa-G controlling for motiv, gpa-U

```
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


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.


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

$R^2_{Y k2}$

Block Number 2. Method: Enter GRE_V GRE_Q

Variable(s) Entered on Step Number
3.. GRE_Q GRE-Quant
4.. GRE_V GRE-Verbal

2
YK_1 →

Multiple R	.94062		
R Square	.88476	R Square Change	.10447
Adjusted R Square	.85403	F Change	6.79918
Standard Error	.16901	Signif F Change	.0079

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	4 → $k1$	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 2 All requested variables entered.

Test for Added Variables : $H_0: \beta_3 = \beta_4 = 0$ (refer to notes)

$$F = \frac{\left(\frac{.88476 - 0.78029}{2}\right)}{\left(\frac{1 - 0.88476}{20-4-1}\right)} = \frac{0.0522}{0.00768} = 6.795$$

at $\alpha = 0.05$, $F_{.05}(2, 15) = 3.68$, $6.795 > 3.68$ Reject H_0 at 0.05

8.24 ⑧

```
-> regression variables=gpa_u gre_v gre_q motiv gpa_g/
->     descriptive=corr/
->     statistics=default,cha/
->     dependent=gpa_g/method=stepwise.
```

Stepwise

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

Listwise Deletion of Missing Data

N of Cases = 20

Correlation:

y

	GPA_U	GRE_V	GRE_Q	MOTIV	GPA_G
GPA_U	1.000	.617	.452	.725	.855
GRE_V	.617	1.000	.708	.403	.761
GRE_Q	.452	.708	1.000	.199	.466
MOTIV	.725	.403	.199	1.000	.773
GPA_G	.855	.761	.466	.773	1.000

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

Equation Number 1 Dependent Variable.. GPA_G Graduate GPA

Descriptive Statistics are printed on Page 7

Block Number 1. Method: Stepwise Criteria PIN .0500 POUT .

Variable(s) Entered on Step Number
1.. GPA_U Undergrad GPA

Multiple R	.85453		
R Square	.73022	R Square Change	.73022
Adjusted R Square	.71524	F Change	48.72178
Standard Error	.23606	Signif F Change	.0000

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	1	2.71497	2.71497
Residual	18	1.00303	.05572

F = 48.72178 Signif F = .0000

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

9
8.5

Variable	B	SE B	Beta	T	Sig T
GPA_U	1.275232	.182695	.854531	6.980	.0000
(Constant)	-1.004642	.606111		-1.658	.1147

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

Variable	Beta In	Partial	Min Toler	T	Sig T
GRE_V	.377034	.571467	.619764	2.871	.0106
GRE_Q	.100119	.171963	.795870	.720	.4815
MOTIV	.324641	.430778	.475014	1.968	.0656

* *

Variable(s) Entered on Step Number
2.. GRE_V GRE-Verbal

Multiple R	.90461		
R Square	.81833	R Square Change	.08810
Adjusted R Square	.79695	F Change	8.24406
Standard Error	.19933	Signif F Change	.0106

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	2	3.04253	1.52127
Residual	17	.67547	.03973

F = 38.28691 Signif F = .0000

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

Variable	B	SE B	Beta	T	Sig T
GPA_U	.928281	.195962	.622039	4.737	.0002
GRE_V	.002562	8.9244E-04	.377034	2.871	.0106
(Constant)	-1.317267	.523264		-2.517	.0221

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

Variable	Beta In	Partial	Min Toler	T	Sig T
GRE_Q	-.165425	-.273843	.387686	-1.139	.2715
MOTIV	.361886	.583243	.349125	2.872	.0111

8.26 (10)

Variable(s) Entered on Step Number
3.. MOTIV Movitation

Multiple R	.93815		
R Square	.88013	R Square Change	.06180
Adjusted R Square	.85765	F Change	8.24876
Standard Error	.16690	Signif F Change	.0111

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	3	3.27231	1.09077
Residual	16	.44569	.02786

F = 39.15783 Signif F = .0000

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

Variable	B	SE B	Beta	T	Sig T
GPA_U	.513375	.218612	.344012	2.348	.0320
GRE_V	.002737	7.4970E-04	.402688	3.650	.0022
MOTIV	.016120	.005613	.361886	2.872	.0111
(Constant)	-1.169684	.441130		-2.652	.0174

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

Variable	Beta In	Partial	Min Toler	T	Sig T
GRE_Q	-.098479	-.196582	.340776	-.777	.4495

End Block Number 1 PIN = .050 Limits reached.

Logistic Regression :

Predicting whether an event will or will not occur, as well as identifying the variables useful in making the prediction, is important in most academic disciplines as well as "real" world. Why do some people develop coronary heart disease and others not? Why do some business succeed, while others fail?

When the dependent variable can have only two values, the assumptions necessary for hypothesis testing in regression analysis are necessarily violated. Another difficulty with multiple regression is that predicted values cannot be interpreted as probabilities. They are not constrained to fall in the interval between 0 and 1.

In logistic regression we directly estimate the probability of an event occurring. For the case of a single independent variable, the logistic regression model can be written as:

$$\log \frac{p}{1-p} = b_0 + b_1 x_1$$

where b_0, b_1 can be estimated from the data

$$\Rightarrow \frac{p}{1-p} = e^{b_0 + b_1 x_1}$$

$$\Rightarrow p = e^{b_0 + b x_1} - p e^{b_0 + b_1 x_1}$$

$$\Rightarrow p + e^{b_0+b_1x_1} = e^{b_0+b_1x_1}$$

$$\Rightarrow p\left(1+e^{b_0+b_1x_1}\right) = e^{b_0+b_1x_1}$$

$$\Rightarrow p = \frac{e^{b_0+b_1x_1}}{1+e^{b_0+b_1x_1}}$$

For a set of multiple predictors, we have

$$p = \frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_nx_n}}{1+e^{b_0+b_1x_1+b_2x_2+\cdots+b_nx_n}}$$

(1) Hierarchical logistic regression

(2) Stepwise logistic regression.

(3) Simultaneous logistic regression

See computer printout for discussion.

(11)

8:29

Logistic Regression

```
Total number of cases:       20 (Unweighted)
Number of selected cases:    20
Number of unselected cases:  0

Number of selected cases:                    20
Number rejected because of missing data:     0
Number of cases included in the analysis:    20


Dependent Variable Encoding:

Original        Internal
Value           Value
     .00        0
    1.00        1

Dependent Variable..   FINISH     Finish Grad School

Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood   26.920467

* Constant is included in the model.


Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..        GPA_U      Undergrad GPA
           GRE_V      GRE-Verbal
           GRE_Q      GRE-Quant
           MOTIV      Movitation

Estimation terminated at iteration number 6 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood          9.033
 Goodness of Fit           10.461
 Cox & Snell - R^2           .591
 Nagelkerke - R^2            .799

                      Chi-Square    df Significance

 Model                   17.888     4        .0013
 Block                   17.888     4        .0013
 Step                    17.888     4        .0013

Classification Table for FINISH
The Cut Value is .50
                    Predicted
                   no       yes     Percent Correct
                     n  I    y
Observed         +-------+-------+
   no       n    I    7  I    1  I   87.50%
                 +-------+-------+
   yes      y    I    0  I   12  I  100.00%
                 +-------+-------+
                            Overall  95.00%

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

Variable            B       S.E.      Wald     df       Sig        R     Exp(B)
```

Simultaneous Logistic Regression

8.30 (12)

GPA_U	3.7613	5.0064	.5644	1	.4525	.0000	43.0041
GRE_V	.0395	.0289	1.8664	1	.1719	.0000	1.0402
GRE_Q	-.0311	.0271	1.3172	1	.2511	.0000	.9694
MOTIV	.3339	.1790	3.4774	1	.0622	.2343	1.3964
Constant	-39.5097	22.0664	3.2059	1	.0734		

Logistic Regression

```
Total number of cases:       20 (Unweighted)
Number of selected cases:    20
Number of unselected cases:  0

Number of selected cases:                    20
Number rejected because of missing data:     0
Number of cases included in the analysis:    20


Dependent Variable Encoding:

Original        Internal
Value           Value
     .00        0
    1.00        1

Dependent Variable..   FINISH     Finish Grad School

Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood   26.920467

* Constant is included in the model.


Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..        MOTIV     Movitation

Estimation terminated at iteration number 5 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood          13.059
 Goodness of Fit            18.730
 Cox & Snell - R^2            .500
 Nagelkerke - R^2             .676

                        Chi-Square     df Significance

 Model                    13.862      1          .0002
 Block                    13.862      1          .0002
 Step                     13.862      1          .0002

Classification Table for FINISH
The Cut Value is .50
                    Predicted
                   no       yes      Percent Correct
                     n  I    y
Observed         +-------+-------+
   no       n    I    7  I    1  I   87.50%
                 +-------+-------+
   yes      y    I    2  I   10  I   83.33%
                 +-------+-------+
                            Overall  85.00%
```

Hierarchical logistic Regr.

8.31 (13)

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

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
MOTIV	.2948	.1227	5.7768	1	.0162	.3746	1.3429
Constant	-19.5085	8.1600	5.7156	1	.0168		

Beginning Block Number 2. Method: Enter

Variable(s) Entered on Step Number
1.. GPA_U Undergrad GPA
GRE_V GRE-Verbal
GRE_Q GRE-Quant

Estimation terminated at iteration number 6 because
Log Likelihood decreased by less than .01 percent.

-2 Log Likelihood	9.033
Goodness of Fit	10.461
Cox & Snell - R^2	.591
Nagelkerke - R^2	.799

	Chi-Square	df	Significance
Model	17.888	4	.0013
Block	4.026	3	.2587
Step	4.026	3	.2587

Classification Table for FINISH
The Cut Value is .50

Observed		Predicted no (n)	Predicted yes (y)	Percent Correct
no	n	7	1	87.50%
yes	y	0	12	100.00%
			Overall	95.00%

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

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
MOTIV	.3339	.1790	3.4774	1	.0622	.3364	1.3964
GPA_U	3.7613	5.0064	.5644	1	.4525	.0000	43.0041
GRE_V	.0395	.0289	1.8664	1	.1719	.0000	1.0402
GRE_Q	-.0311	.0271	1.3172	1	.2511	.0000	.9694
Constant	-39.5097	22.0664	3.2059	1	.0734		

Logistic Regression

Total number of cases: 20 (Unweighted)
Number of selected cases: 20
Number of unselected cases: 0

Number of selected cases: 20
Number rejected because of missing data: 0
Number of cases included in the analysis: 20

Dependent Variable Encoding:

8.32 (14)

Original Value	Internal Value
.00	0
1.00	1

Dependent Variable.. FINISH Finish Grad School

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 26.920467

* Constant is included in the model.

Estimation terminated at iteration number 2 because
Log Likelihood decreased by less than .01 percent.

Classification Table for FINISH
The Cut Value is .50

		Predicted no n	Predicted yes y	Percent Correct
Observed no	n	0	8	.00%
yes	y	0	12	100.00%
			Overall	60.00%

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

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Constant	.4055	.4564	.7891	1	.3744		

Beginning Block Number 1. Method: Forward Stepwise (COND)

---------- Variables not in the Equation ----------
Residual Chi Square 11.423 with 4 df Sig = .0222

Variable	Score	df	Sig	R
GPA_U	7.5566	1	.0060	.4543
GRE_V	3.6628	1	.0556	.2485
GRE_Q	.3753	1	.5402	.0000
MOTIV	10.2827	1	.0013	.5547

Variable(s) Entered on Step Number
1.. MOTIV Movitation

Estimation terminated at iteration number 5 because
Log Likelihood decreased by less than .01 percent.

-2 Log Likelihood	13.059
Goodness of Fit	18.730
Cox & Snell - R^2	.500
Nagelkerke - R^2	.676

Stepwise logistic reg.

	Chi-Square	df	Significance
Model	13.862	1	.0002
Block	13.862	1	.0002
Step	13.862	1	.0002

8.33 (15)

Classification Table for FINISH
The Cut Value is .50

		Predicted no (n)	Predicted yes (y)	Percent Correct
Observed no	n	7	1	87.50%
yes	y	2	10	83.33%
			Overall	85.00%

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

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
MOTIV	.2948	.1227	5.7768	1	.0162	.3746	1.3429
Constant	-19.5085	8.1600	5.7156	1	.0168		

---------- Model if Term Removed ----------
Based on Conditional Parameter Estimates

Term Removed	Log Likelihood	-2 Log LR	df	Significance of Log LR
MOTIV	-13.804	14.549	1	.0001

---------- Variables not in the Equation ----------
Residual Chi Square 3.339 with 3 df Sig = .3423

Variable	Score	df	Sig	R
GPA_U	.8338	1	.3612	.0000
GRE_V	1.9148	1	.1664	.0000
GRE_Q	.0906	1	.7634	.0000

No more variables can be deleted or added.