Coefficient of **Partial** Determination

As the R^2 provides information about the $SSR(X_1,X_2,X_3)$, there are also

Coefficients of PARTIAL Determination : this measures "how much variation a variable accounts for out of the variation available to that variable when it enters". This gives a proportional measure of the contribution of each variable after all other variables are in the model.

eg. $Y = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + e_i$

Take X_2

What is the Coefficient of Partial Determination?

1) How much did the variable account for? (after other variables→partial)
$SSR(X_2|X_1,X_3) = 1.502621$

2) What SS was available to it when it entered the model.

$SSE(X_1,X_3) = SSE(X_1,X_2,X_3) + SSR(X_2|X_1,X_3)$

$= 61.443 + 230.62548 = 292.06848$

Partial $R^2_{X_2|X_1,X_3} = r^2_{2.13} = \frac{230.62548}{292.06848} = 0.7896281 = 78.96281\%$

These calculations are available from SAS PROC REG with the / PCORR2 option on the MODEL statement.

SAS will also produce a Partial Correlation of the TYPE I SS.

Output from PROC REG

Parameter Estimates

Standardized

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.84693	2.00188	8.92	<.0001
X1	1	1.10313	0.32957	3.35	0.0032
X2	1	0.32152	0.03711	8.66	<.0001
X3	1	1.28894	0.29848	4.32	0.0003

Variable	DF	Type I SS	Type II SS	Standardized Estimate	Squared Semi-partial Corr Type I
INTERCEP	1	37446	244.171679	0.00000000	.
X1	1	306.732328	34.418508	0.26023468	0.44501687
X2	1	263.794445	230.625476	0.65915439	0.38272125
X3	1	57.290222	57.290222	0.30693999	0.08311845

Variable	DF	Squared Partial Corr Type I	Squared Semi-partial Corr Type II	Squared Partial Corr Type II	Tolerance
INTERCEP	1	.	.	.	.
X1	1	0.44501687	0.04993545	0.35904408	0.73735836
X2	1	0.68960879	0.33459866	0.78962809	0.77010493
X3	1	0.48251214	0.08311845	0.48251214	0.88224762

Standardized Regression Coefficients

This technique addresses two aspects of estimating β_k values

1) There is some potential difficulty with rounding errors in the calculations, particularly for the $(X'X)^{-1}$ matrix calculations.

    These roundoff errors are aggravated by (1) more variables in the model, (2) multicolinearity and (3) b values of very different magnitudes.

    Standardized regression coefficients can help with the last problem.

2) The magnitude of the regression coefficients cannot be compared.

    Since the regression coefficients have units which are $\frac{\text{Y units}}{\text{X unit}}$, they will vary with the units of X and Y.

    eg. If different people do the same study and various investigators take measurements on X_1 in (1) inches, (2) feet, (3) meters and (4) mm, then the same study will very different values for b_1.

    The same is true of if a dependent variable (Y) is measured in (1) dollars, (2) thousands of dollars, or (3) median family income units (multiples of about 18 thousand).

    As a result of these scaling factors, the regression coefficients have an interpretation in terms of the regression coefficients, but the regression coefficients will differ for different units, and must be examined within the context of those units.

    Standardized Regression Coefficients, however, have no units, but their size can be interpreted as a measure of impact or importance of each variable on the calculation of the predicted value.

There are several ways to calculated Standardized Regression Coefficients

1) The variables can be “standardized" prior to doing the regression

$$Y' = \frac{1}{\sqrt{n-1}} \frac{Y_i - \bar{Y}}{s_Y}$$

$$X'_{ik} = \frac{1}{\sqrt{n-1}} \frac{X_{ik} - \bar{X_k}}{s_{x_k}}$$

where s_Y and s_{x_k} are ordinary standard deviations

regression on these variables gives the *standardized regression model*

$$Y_i = \beta'_1 X'_{1i} + \beta'_2 X'_{2i} + \beta'_3 X'_{3i} + \epsilon_i$$

where $\beta_0 = 0$

2) If the matrix calculations are done with the standardized values of X and Y, then the $X'X$ and and $X'Y$ matrices are

$$X'X = \begin{bmatrix} 1 & r_{12} & \dots & r_{1,p-1} \\ r_{21} & 1 & \dots & r_{2,p-1} \\ \vdots & & & \\ r_{p-1,1} & r_{p-1,2} & \dots & 1 \end{bmatrix} \qquad X'Y = \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Y,p-1} \end{bmatrix}$$

Note that there is no row for the intercept

so, another way to get the standardized regression coefficients is to calculated the matrix formula for $B = (X'X)^{-1}X'Y$ using the correlation matrices, or $B' = (R_{XX})^{-1}R_{XY}$

3) There is also a relationship between the standardized regression coefficient and the ordinary least squares regression coefficient. The relationship is

$$\beta_k = \beta'_k \left(\frac{s_Y}{s_X} \right)$$

The interpretation of the standardized regression coefficient is as a measure of relative impact on the calculations or as relative importance of the variable to the model.

The size of the variable is not longer influenced by units, and standardized regression coefficients are unitless.

The SIGN of the regression coefficient is retained, so negative and positive effects can still be interpreted.

Example : The standard deviations are given by (for the mathematician example)

$$s_Y = \sqrt{\frac{\Sigma Y_i^2 - \frac{(\Sigma Y)^2}{n}}{n-1}} = \sqrt{\frac{38135.26 - \frac{(948)^2}{24}}{23}} = 5.47429$$

$$s_X = \sqrt{\frac{\Sigma X_{ik}^2 - \frac{(\Sigma X_k)^2}{n}}{n-1}}$$

for X_3

$$s_3 = \sqrt{\frac{\Sigma X_{i3}^2 - \frac{(\Sigma X_3)^2}{n}}{n-1}} = \sqrt{\frac{899.49 - \frac{(128.6)^2}{24}}{23}} = 1.303$$

$$\beta_3 = \beta_3' \left(\frac{s_Y}{s_X}\right) \text{ so } \beta_3' = \beta_3 \left(\frac{s_X}{s_Y}\right) = 1.2889 \left(\frac{1.303}{5.472}\right) = 0.30694$$

all values are available from the $X'X$, $X'Y$ and $Y'Y$ matrices

Interpretation

1) Size of value (magnitude, regardless of sign) is important. This is an indicator of "importance", or impact in the calculation of the predicted value. This would generally agree with observations and evaluations made by P>|t| and SSII and Partial R^2, but not always.

2) The SIGN is important, and will match the sign on the regression coefficient.

Effect of Correlation among the X_k variables

1) If the X_k variables are uncorrelated, then they will describe a certain variation whether alone or in concert with other variables,

and the variables describe the same variation no matter which other variables they are adjusted for.

Also the regression coefficients will stay the same, will be stable

There are several ways of creating this type of design.

a) Orthogonal variables : result from transformation which extract the attributes of a variable while retaining a 0 correlation with other variables.

orthogonal polynomial multipliers are a good example of this (from any table)

Orthogonal Polynomial Multipliers (equally spaced X) variable with 5 levels

X =	1	2	3	4	5
Linear	-2	-1	0	1	2
Quadratic	2	-1	-2	-1	2
Cubic	-1	2	0	-2	1
Quartic	1	-4	6	-4	1

Note that all crossproducts sum to zero

b) Some multivariate analyses (such as PCA) will create orthogonal variables, these can be used as independent variables

c) many designed experiments are orthogonal, factorials are a good example

eg. 2x2x2 factorial

Test	A	B	AB	C	AC	BC	ABC	
X =	-1	-1	1	-1	1	1	-1	abc
	1	-1	-1	-1	-1	1	1	Abc
	-1	1	-1	-1	1	-1	1	aBc
	1	1	1	-1	-1	-1	-1	ABc
	-1	-1	1	1	-1	-1	1	abC
	1	-1	-1	1	1	-1	-1	AbC
	-1	1	-1	1	-1	1	-1	aBC
	1	1	1	1	1	1	1	ABC

What happens to the EXTRA SS? If X_1 and X_2 are uncorrelated, then

$$SSR(X_1) = SSR(X_1|X_2)$$
$$SSR(X_2) = SSR(X_2|X_1)$$

each variable is uninfluenced by the other in terms of its SSR.

Another type of uncorrelated example is given in the text where each level of one variable occurs at each level of another variable. These will be uncorrelated even though the variables are quantitative.

Example

```
DATA ONE; INFILE CARDS MISSOVER;
     TITLE1 'EXST7034 - Example NWK Table 8.7 :
      Uncorrelated variables';
     LABEL Y = 'Crew Productivity Score';
   INPUT TRIAL CREWSIZE BONUSPAY Y;
CARDS; RUN;
 1   4   2   42
 2   4   2   39
 3   4   3   48
 4   4   3   51
 5   6   2   49
 6   6   2   53
 7   6   3   61
 8   6   3   60
;

PROC REG DATA=ONE; TITLE2 'All models in PROC REG';
        MODEL  Y = BONUSPAY;
        MODEL  Y = CREWSIZE;
        MODEL  Y = CREWSIZE BONUSPAY / SS2; RUN;
```

Note from the handout that:

1) The two fitted together account for the sum of the SS of each individually
2) The regression coefficients of the two together do not change
3) EVEN THOUGH THE TWO ARE INDEPENDENT,
the two alone were not significant (0.0885) or barely sig (0.0351)

but together both were highly significant. This is due entirely to the reduction of the error variance term.

EXST7034 - Example NWK Table 8.7 : Uncorrelated variables
All models in PROC REG

Model: MODEL1
Dependent Variable: Y   Crew Productivity Score

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	171.12500	171.12500	4.128	0.0885
Error	6	248.75000	41.45833		
C Total	7	419.87500			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|	Variable Label
INTERCEP	1	27.250000	11.60773808	2.348	0.0572	Intercept
BONUSPAY	1	9.250000	4.55292946	2.032	0.0885	

Model: MODEL2
Dependent Variable: Y   Crew Productivity Score

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	231.12500	231.12500	7.347	0.0351
Error	6	188.75000	31.45833		
C Total	7	419.87500			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|	Variable Label
INTERCEP	1	23.500000	10.11135912	2.324	0.0591	Intercept
CREWSIZE	1	5.375000	1.98300067	2.711	0.0351	

Model: MODEL3
Dependent Variable: Y   Crew Productivity Score

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	402.25000	201.12500	57.057	0.0004
Error	5	17.62500	3.52500		
C Total	7	419.87500			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|	Type II SS	Variable Label
INTERCEP	1	0.375000	4.74045093	0.079	0.9400	0.022059	Intercept
CREWSIZE	1	5.375000	0.66379590	8.097	0.0005	231.125000	
BONUSPAY	1	9.250000	1.32759180	6.968	0.0009	171.125000	

Multicolinearity : strong relationship between two variables (high correlation)

2) Strong correlations are easy to detect if a single X is correlated to another, however, one variable may be correlated to a linear combination of other variables. (eg. $X_1 \approx X_2 + X_3 - X_4$)

When variables are highly correlated, the effects may not adversely effect our predictive ability,

but, the regressions coefficients are usually way off (unbiased, but off).

As a result, they are not useful as estimates of the rates we often desire, and holding one constant while varying another to examine the effect is not a fruitful exercise.

Standardization of the variables may help in stabilizing the variance

This is not usually a serious problem until correlations are “quite high".

Perfect correlations among the X variables results in a matrix which cannot be inverted (the determinant is 0)

this is referred to as Singularity
Ill condition matrix
Matrix not of full rank
(ie. cannot fit as many variables as there are columns)

Examples of some perfectly correlated variables

```
DATA TWO; INFILE CARDS MISSOVER;
     TITLE1 'EXST7034 - Example NWK Table 8.8 :
      Perfectly correlated variables';
   INPUT CASE X1 X2 Y;
CARDS; RUN;
1   2    6     23
2   8    9     83
3   6    8     63
4  10   10    103
;

PROC REG DATA=TWO; TITLE2 'Generic example';
        MODEL  Y = X1;
        MODEL  Y = X2;
        MODEL  Y = X1 X2 / SS2; RUN;

DATA TWO; INFILE CARDS MISSOVER;
     TITLE1 'EXST7034 - Modified example NWK Table 8.8
      : Perfectly correlated independent variables';
   INPUT CASE X1 X2 Y;
CARDS; RUN;
1   2    6     23
2   8   12     83
3   7   11     63
4  10   14    103
;
PROC REG DATA=TWO; TITLE2 'Modified generic example';
        MODEL  Y = X1;
        MODEL  Y = X2;
        MODEL  Y = X1 X2 / SS2; RUN;
```

EXST7034 - Example NWK Table 8.8 : Perfectly correlated variables

```
Generic example

Model: MODEL1            Dependent Variable: Y
Analysis of Variance
                         Sum of         Mean
Source          DF      Squares       Square      F Value       Prob>F
Model            1   3500.00000   3500.00000         .            .
Error            2      0.00000      0.00000
C Total          3   3500.00000
Parameter Estimates
                 Parameter      Standard    T for H0:
Variable  DF      Estimate         Error   Parameter=0    Prob > |T|
INTERCEP   1      3.000000    0.00000000          .            .
X1         1     10.000000    0.00000000          .            .


Generic example
Model: MODEL2            Dependent Variable: Y
Analysis of Variance
                         Sum of         Mean
Source          DF      Squares       Square      F Value       Prob>F
Model            1   3500.00000   3500.00000         .            .
Error            2      0.00000      0.00000
C Total          3   3500.00000
Parameter Estimates
                 Parameter      Standard    T for H0:
Variable  DF      Estimate         Error   Parameter=0    Prob > |T|
INTERCEP   1    -97.000000    0.00000000          .            .
X2         1     20.000000    0.00000000          .            .


Generic example
Model: MODEL3            Dependent Variable: Y
Analysis of Variance
                         Sum of         Mean
Source          DF      Squares       Square      F Value       Prob>F
Model            1   3500.00000   3500.00000         .            .
Error            2      0.00000      0.00000
C Total          3   3500.00000
NOTE: Model is not full rank. Least-squares solutions for the parameters are
        not unique. Some statistics will be misleading. A reported DF of 0 or
        B means that the estimate is biased.  The following parameters have
        been set to 0, since the variables are a linear combination of other
        variables as shown.
       X2       = +5.0000 * INTERCEP +0.5000 * X1
Parameter Estimates
               Parameter    Standard    T for H0:
Variable  DF   Estimate        Error   Parameter=0    Prob > |T|    Type II SS
INTERCEP   B   3.000000  0.00000000          .             .         6.176471
X1         B  10.000000  0.00000000          .             .      3500.000000
X2         0          0  0.00000000          .             .                .
```

EXST7034 - Modified example NWK Table 8.8 : Perfectly correlated independent variables

Modified generic example

Model: MODEL1 Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	3425.17986	3425.17986	91.558	0.0107
Error	2	74.82014	37.41007		
C Total	3	3500.00000			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.985612	7.64217078	0.129	0.9092
X1	1	9.928058	1.03756871	9.569	0.0107

Modified generic example

Model: MODEL2 Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	3425.17986	3425.17986	91.558	0.0107
Error	2	74.82014	37.41007		
C Total	3	3500.00000			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-38.726619	11.56551739	-3.348	0.0788
X2	1	9.928058	1.03756871	9.569	0.0107

Modified generic example

Model: MODEL3 Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	3425.17986	3425.17986	91.558	0.0107
Error	2	74.82014	37.41007		
C Total	3	3500.00000			

NOTE: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

X2 = +4.0000 * INTERCEP +1.0000 * X1

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|	Type II SS
INTERCEP	B	0.985612	7.64217078	0.129	0.9092	0.622252
X1	B	9.928058	1.03756871	9.569	0.0107	3425.179856
X2	0	0	0.00000000	.	.	.

Note that with perfect correlation between X_1 and X_2 and Y,

1) No error terms (ie. =0) , perfect fits every time , → no tests

2) Only one needed to fit when the two are put together

3) SAS warns “not full rank" when the two are put together (but not alone).

Note that with perfect correlation between several X variables, but not with Y

1) You may get decent fits of each variable alone, but if there are two perfectly correlated variables the fits are identical

2) Only one fitted when the two are put together, and this matches the fits alone

3) SAS warns “not full rank".

The text makes an issue of the fact that with perfect correlation, an infinite number of models can be obtained. In practice, most software will bomb or detect the problem. We will see various diagnostics later (Ch 11, we are in 8) which will detect the problem.

sample problem

Obs	X1	X2	Y
	1	1	2
	2	2	4
	3	3	6

all perfectly correlated,

this could be fitted by

Y = 1*X1 + 1*X2
Y = 0*X1 + 2*X2
Y = 2*X1 + 0*X2
Y = 0.5*X1 + 1.5*X2
Y = 1.5*X1 + 0.5*X2
Y = 1.3*X1 + 0.7*X2
Y = 102*X1 - 100*X2

or any other model where $b_1 + b_2 = 2$

This results whenever two variables are perfectly correlated and there is a perfect fit with no error.

It is clear that the regression coefficients cannot be interpreted

What if the correlations are just high, not perfect?

1) We have no problem getting a good fit, but regressions coefficients will not be stable (they will vary widely from sample to sample).
Also, the fact that the reg coeff for each X are unstable makes prediction outside the range of that X untenable.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error
X1	1	0.857187	0.12878079
X1|X2	1	0.222353	0.30343892
X1|X3	1	1.000585	0.12823209
X1|X2,X3	1	4.334092	3.01551136
X2	1	0.856547	0.11001562
X2|X1	1	0.659422	0.29118728
X2|X3	1	0.850882	0.11244824
X2|X1,X3	1	-2.856848	2.58201527
X3	1	0.199429	0.32662975
X3|X1	1	-0.431442	0.17661556
X3|X2	1	0.096029	0.16139267
X3|X1,X2	1	-2.186060	1.59549900

Note that as more variables are added to the model, the regression coefficients vary greatly, and the standard errors generally increase.
However, even as the standard errors increase, the MSE decreases and the precision on the predicted value may be quite acceptable.
Recall that we do not assume that the covariance is 0 when calculating $s_{\hat{Y}}$, so the strong correlation between variables may also be influenced by strong negative or positive covariances

2) The whole idea of “holding one X constant" while varying another goes against the “high correlation" between variables. If we vary one, the other should vary in a predictable fashion as well.
Suppose the variables “surface temperature" and “bottom temperature" are used to predict the abundance of shrimp. Since these vary together, how far can we realistically vary one while holding the other constant?

The text book recommends simple correlations, this is a useful diagnostic for many situations,
but this will not detect the most insidious Multicolinearity problems
We will later discuss some more serious diagnostics.

Pearson Correlation Coefficients / Prob>|R| under Ho: Rho=0/N = 20

	X1	X2	X3
X1	1.00000	0.92384	0.45778
Triceps skinfold thickness		0.0001	0.0424
X2	0.92384	1.00000	0.08467
Thigh circumference	0.0001		0.7227
X3	0.45778	0.08467	1.00000
Midarm circumference	0.0424	0.7227	

Correlations of linear combinations among independent variables in Body Fat Example (Neter, Wasserman & Kuttner, 1989).

Dependent Variable: X1		Triceps skinfold thickness	
Root MSE	0.19946	R^2 = 0.9986	r = 0.9993
Dependent Variable: X2		Thigh circumference	
Root MSE	0.23295	R^2 = 0.9982	r = 0.9991
Dependent Variable: X3		Midarm circumference	
Root MSE	0.37699	R^2 = 0.9904	r = 0.9952

The effect of Multicolinearity on a model is a serious one, and one which will require additional techniques to address.
The problem adversely effects
Estimates of regression coefficients
Variance of the reg coeff

We will return to this problem later with several ways of addressing it directly (this problem is so serious that we may even be willing ot accept a “biased estimator")
or ways of getting around it through variable selection techniques in building the model