Ideas unique to Multiple Regression - no analogy with Simple Linear Regression.
A number of important sub-topics are introduced in this Chapter (*see also "Chapter 9" notes in c678*)

--Formal tests for <u>Addition</u>/<u>Deletion</u> of 1 or more terms in a regression model (nicely done)

-- <u>Collinearity</u> -- very important -- should be in a chapter by itself!

-- <u>Interaction</u> terms -- again, should be in a chapter by itself ("Effect Modification" in Epidemiology)!

7.1   Extra Sums of Squares

Extra SS if focus on SS_{reg}: *Reduced* SS if focus on $SS_{residual}$:

Same idea whether adding or removing 1 (or more) term(s)

EXAMPLE

Y: % Body Fat (by underwater weighing !!)

X_1 Skinfold Thickness (by calipers:- some discomfort)

X_2 Skinfold Thickness (by tape measure:- painless)

X_3 Midarm Circumference (by tape measure:- painless)

*IN COMPACT NOTATION (**2 "X" term**s)*

SSR (short for SS_{Reg})		**SSE** (short for SS_{Error})		SST (short for SS_{total})
$SSR(X_1)$	+	$SSE(X_1)$	=	SST
$SSR(X_1,X_2)$	+	$SSE(X_1,X_2)$	=	SST
----------------		--------------		--------
$SSR(X_1,X_2) - SSR(X_1)$	=	$SSE(X_1) - SSE(X_1,X_2)$		
"Extra SSR due to X_2"	=	"Reduced SSE due to X_2"		
$SSR(X_2| X_1)$				

*IN COMPACT NOTATION (**3 "X" 's**)*

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

$=>$ $SSR(X_1)$ $= SST$ $-$ $SSE(X_1)$ $\vert$ *Note*: **order matters!**

$SSR(X_2| X_1)$ $= SSE(X_1)$ $-$ $SSE(X_1,X_2)$ $\vert$ $SSR(X_2| X_1)$ not same

$SSR(X_3| X_1,X_2)$ $= SSE(X_1,X_2) -$ $SSE(X_1,X_2,X_3)$ $\vert$ as $SSR(X_1| X_2)$

Different decompositions have different meanings

e.g.   a   b   c

   $SSR(X_1)$ $SSR(X_1)$ $SSR(X_3)$

   $SSR(X_2|X_1)$ $SSR(X_3|X_1)$ $SSR(X_2|X_3)$

   $SSR(X_3|X_1,X_2)$ $SSR(X_2|X_1,X_3)$ $SSR(X_1|X_2,X_3)$

*These decompositions are referred to as "**variables added in orde**r SS" or "**Type I** SS"*
*(to distinguish them from "**variables added last** SS" or "**Type III** SS" used below)*

If labeling is not obvious, or if you forget what Type I means (I don't blame you, given the
imaginative choice of the label), one way to recognize it for what it is is that Type I SS count each
separate contribution just once .. so that the successive Type I SSR's add to the overall SSR.

Graphical Depiction of these decompositions ... see Figure 7.1 on page 265

(I find this diagram so helpful that I scanned it and put it on the www page for c678 -- where we
used a different text)

Using SS decomposition in ANOVA Table   => Decomposition of Degrees of Freedom

The above decomposition is for the Sums of Squares.

When the decomposition is done one term at a time, the $SSR(X_1)$, $SSR(X_2|X_1)$, $SSR(X_3|X_1,X_2)$ ...
and the corresponding $MS(X_1)$, $MS(X_2|X_1)$, $MS(X_3|X_1,X_2)$ are the same -- since each divisor
involves 1 df.

7.2
and
7.3   **"Extra Sums of Squares" to test coefficients of Multiple Regression**

From Cochran's Theorem [partitioning SS; each (1/ β^2) MS ~ indep. central/non-central X^2]

Single   [H_0: **this** β's = 0]

  $t^* = b / SE[b]$ with df = df for residuals (e's) when this X & other X's in the model

    or   (*equivalently*, since $t^{*2} = F^*$)

  $$F^* = \frac{SSR[\text{this X | other X's}] / 1\text{ df}}{SSE[\text{this X & other X's}] / \text{df for e's if this X & other X's}} = \frac{MSR[\text{this X | otherX's}]}{MSE[\text{this X & other X's}]}$$

Several β's [H_0: **these** β's = 0]

  $$F^* = \frac{SSR[\text{these X's | other X's}] / \text{\# of these X's}}{SSE[\text{these & other X's}] / \text{df for E if these X's & other X's}} = \frac{MSR[\text{these X's | other X's}]}{MSE[\text{these & other X's}]}$$

All β's [H_0: **all the** β's = 0]

  $$F^* = \frac{SSR[\text{all these X's }] / \text{\# of X's}}{SSE[\text{these X's }] / \text{df for e's if these X's}} = \frac{MSR[\text{ these X's}]}{MSE[\text{ these X's}]}$$

IN GENERAL [H_0: **set of (linear) constraints on the** β's]

  $$F^* = \frac{SSR[\text{constrained }\beta\text{'s}] / \text{\# of constraints}}{SSE[\text{no constraints}] / \text{df for e's if no constraints}} = \frac{MSR[\text{"due to" constrained }\beta\text{'s}]}{MSE[\text{no constraints}]}$$

  i.e.

  $$F^* = \frac{\{SS_{Reg}[\text{larger model}] - SS_{Reg}[\text{smaller model}] \} / \text{\# of constraints}}{SSE[\text{no constraints}] / \text{df for e's in larger model}}$$

  $$= \frac{MSR[\text{"due to" constrained }\beta\text{'s}]}{MSE[\text{larger model}]}$$

Q: Why use the MSE from the larger model as the denominator of each test?

Even if some terms in larger model are unnecessary, this MSE is an unbiased estimator of σ^2[ ].

7.4   Coefficients of Partial Determination

*Recall***:** Coefficient of <u>Multiple</u> <u>Determination</u>:

Reduction in Var(Y) through use of $X_1, X_2, ..$

$$r^2_{Y \text{ with } X1, X2, X3, ...} = \frac{\text{Var}[Y] - \text{Var}[Y \mid X_1 \ X_2 \ ... \ ]}{\text{Var}[Y]} = \frac{\text{SSR}[X_1 \ X_2 \ ... \ ]}{\text{SSE}[\text{model with just } b_0]}$$

By analogy: Coefficient of <u>Partial</u> <u>Determination</u> (by X_1 -- after already fitting X_2):

Reduction in Var(residual $Y \mid X_2$)

$$r^2_{Y \text{ with } X1 \text{ given } X2} = \frac{\text{Var}[Y \mid X_2] - \text{Var}[Y \mid X_1 \ X_2]}{\text{Var}[Y \mid X_2]}$$

$$= \frac{\text{SSR}[X_1 \mid X_2 \ ]}{\text{SSE}[\text{model with just } \{X_0 \text{ and}\} \ X_1]}$$

likewise...

$$r^2_{Y \text{ with } X2 \text{ given } X1} = \frac{\text{Var}[Y \mid X_1] - \text{Var}[Y \mid X_1 \ X_2]}{\text{Var}[Y \mid X_1]}$$

$$= \frac{\text{SSR}[X_2 \mid X_1 \ ]}{\text{SSE}[\text{model with just } \{X_0 \text{ and}\} \ X_2]}$$

note:

$$r^2_{Y \text{ with } X1 \text{ given } X2} = r^2_{Y' \text{ with } X1'} ,$$

where Y' = residual of Y after regressing Y on X_2, and

X_1' = residual of X_1 after regressing X_1 on X_2 .

Going on to more X's ...

$$r^2_{Y \text{ with } X4 \text{ given } X1 , X2 \text{ and } X3} = \frac{\text{SSR}[X_4 \mid X_1 \ X_2 \ X_3 \ ]}{\text{SSE}[\text{model with } X_1 \ X_2 \ X_3 \ ]}$$

Coefficients of Partial Correlation (Square Root of Coefficient of Partial determination)

(Useful to those who think in correlations rather than regresssion coefficients)

$$r_{Y \text{ with } X4 \text{ given } X1 , X2 \text{ and } X3}$$

$$= \text{sign}[b_4 \text{ in model of Y on } X_1 \text{ to } X_4] \times \sqrt{r^2 \text{ of Y with } X_4 \text{ given } X_1 \text{ to } X_3}$$

(see eqns. 7.41 & 7.42 on how to calculate them from lower order partial correlation coefficients)

7.5   Standardized Multiple Regression Model

- relevant also for Simple Regression model:- <u>eliminates</u> b's in <u>different</u> Y/X_1 , Y/X_2,... <u>units</u>
  (so don't have to worry if X_1 is in cm or inches or metres, X_2 in Kg or lbs or grams-- or \$'s)

- issues of <u>numerical accuracy</u> no longer quite as relevant with high-precision facilities
  (accuracy issues if X's of very different magnitudes, or highly correlated so det[$\mathbf{X^TX}$] close to 0)

Correlation Transformation

- transform each X variable so that each entry in $\mathbf{X^TX}$ matrix is a correlation of 2 X's , i.e.

$$X\,' = \frac{X\,'\text{- Xbar}}{\sqrt{\;(X - Xbar)^2}}$$

... $X'^T X'$ matrix = correlations of pairs of X's

(text formula makes transf. unnec. complex)

- transform Y variable in same way , i.e.

$$Y\,' = \frac{Y\,'\text{- Ybar}}{\sqrt{\;(Y - Ybar)^2}}$$

(text formula makes transf. unnec. complex)

- fit model $Y\,' = {}_1'\,X_1\,' + {}_2'\,X_2\,' + ... + e\,'$ (*no intercept* since ave[$Y\,'$] now $= 0$)

work back ...

$$_k = {}_k{}'\,\frac{SD[Y]}{SD[X_k]}$$

$$_0 = \text{Ybar - } (\,{}_1\,X_1 + {}_2\,X_2 + ... \,)$$

$X'^T X'$ matrix = correlation matrix of original X's = R_{XX}

$X'^T Y'$ vector = correlation of Y with each original X = R_{YX}

LS Solution for $\mathbf{b}\,'$, the column vector $(\,{}_1',\;{}_2',\,...\,)^T$

$$\mathbf{b}\,' = (R_{XX})^{-1}\,R_{YX}$$

For worked example, see text

7.6   Multicollinearity and its Effects (see also item from Graybill on Ill-Conditioning)

Meaning: (high) intercorrelation among some or all of the X terms in a multiple regression

Implications:

- easiest to understand by examining just 2 Xs and just the two extremes:

- see "Confounding[in pictures and numbers]" in Chapter 8 material in c678 www page;
- see also the "hammock" spreadsheet

X1 & X2 uncorrelated	X1 & X2 perfectly (+ or -) correlated
• b_1 (X1-only model) = b_1 (X1 & X2 model)	b_1 (X1 only model) doesn't have same <u>meaning</u> or <u>value</u> as b_1 (X1&X2 model)
• SSR(X1) = SSR(X1 | X2)	SSR(X1 | X2) = SSR(X2 | X1) = 0
and vice versa (can fit b1 and b2 "marginally")	different (b_1, b_2) pairs give same fit (see Fig 7.2 in NWNW4)
uncorrelated estimates of $_1$ and $_2$	cannot "separate" $_1$ and $_2$ estimates:- (b_1, b_2) are unstable • $(b_1 +ve , b_2 -ve)$ in one sample • $(b_1 -ve , b_2 +ve)$ in another
	Like hammock, fitted surface rests on a "knife-edge"
	BUT: can make predictions within (X1,X2) data region

In practice, inter-correlations of X's (and effects of these) are usually somewhere in between. but, more difficult to "see" if more than 2 X's. (***Large SE's for b's can be a warning***)

Chapter 9 will describe methods for detecting multicollinearity in "higher-D" X data.
Chapter 10 explains remedial methods (including "Ridge Regression").

7.7   Polynomial Regression

(Use of higher powers of X (or products of two or more X's) as terms in a multiple regression)
Can be helpful for fitting Response Functions of a single X, or a Response Surface for 2 X's.

• INSIGHT (in SAS) has a dangerously-simple interface for fitting a polynomial in a single X.

• Polynomial regression is simply a multiple regression where the terms are powers of same basic
X variable; BUT one needs to be extra careful about overfitting and about extrapolation This will
remind me to relate one physician's use of a fitted polynomial of time (fitted in the original Lotus1-
2-3 software) to monitor (and ?? anticipate) a patient's White Blood Cell (WBC) count over time.

• *On a related note*: concerning Fig1 of the article " ***Changes in Alcohol Consumption With Age***"
in Can J Public Health Vol. 82, July/August 1991 pp231-4 elsewhere on this www page. (see also
excerpts from article)

RESULTS (from text of the article)

Measures of consumption (based on a total of 3,304 interviews)

*The top panel of Figure 1 shows mean alcohol consumption in drinks per month for males,
females and all respondents by age. The middle panel presents in a similar manner the
frequency of drinking occasions per month, and the bottom panel shows the quantity (mean
number of drinks consumed per drinking occasion).*

These figures were derived by rank ordering all respondents by age. **Each data point
represents the mean of successively older groups of 25 respondents**. *Plotting data in this
fashion provides information on the relative density of observations according to age and sex.*
**In order to reduce the scatter which could obscure trends, a 4253H, twice compound
smoother with endpoint adjustment was used. This consists of a series of running median
smoothers and the Hanning running weighted average smoother applied twice**.

*The top panel shows an age-related decline in total alcohol consumption per month for all
respondents. This line has a slope of -0.12. It is clear from these figures that the age-related
decrease seen in the top panel is largely due to a decrease in quantity which shows a rather
steady decline with age (slope = -0.26), rather than any change in frequency, which has a slope
of only -.04. The correlation between frequency and dose is -0.11 which is small, but statistically
significant (p < 0.0001).*

• I question the choice of "smoothing" that the authors carried out. Although they are non-
parametric,*they look like high order polynomials that seem to "follow" every little random
twist and turn in the raw data*. To my eye, the patterns in the bottom panel are quite linear---
the "join the dots" approach (even with each dot being a running median) over-accentuates
the random components -- what one wants first is the BIG PICTURE .. the clear
downwards "close to linear" trend. I believe the little ups and downs along the way are
random noise -- and that they are being over-emphasized. I cannot imagine that the
population medians actually behave like this.

• Polynomial models are often used for prediction. Thus, the meaning of individual   's is less
critical than when they are associated with different X's. Nevertheless, they are a good example of
the benefits of "centering" X variables in any multiple regression, and of the **induced
collinearity if one uses "raw" powers of X, or -- for that matter --products of two different
X variables**. If X is a positive RV, then X^2 can be strongly correlated with it.

7.8   Regression Models involving Interactions

See also: www material, Session 5, course 678: Interaction(Effect Modification) in Regression

Definitions ...

> ***Interaction*** *(statistical)*
> - "Non-additivity" of "effects" in regression
> - need for product term in regression analysis (miettinen)
>   *note that need for product term may be* <u>*scale-dependent (e.g. Y vs log Y scale)*</u>
>
> ***(Effect) Modification*** *(epidemiological)*
> - Inconstancy of a parameter of a relation over other subject characteristic (miettinen)
> - "Different slopes for different folks" (jh)
>
> ***"Modifier*** (of a relation)
> - A characteristic (of individuals) on which a parameter of a relation depends (osm)

Examples... *(first 4 are on course 678 www page)*

Equation for Ideal Weight as function of Height	- modification by Gender
Average Earnings as function of Education / Age	- modification by Gender
Decline in Bone Density with Age	- Different in 19th and 20th Centuries
Hit further with aluminum than wood baseball bat?	- Depends on *where on bat* one hits
Changes over time in injury rates	- Different in intervn. & ref. areas?

Comments on NKNW4 (and most statisticians') terminology:

> "regression model <u>is not additive</u>, or. equivalently, it contains an ***interaction effect*** "
>
> "non-additive" is a more informative phrase; it avoids possible over-interpretation of the word interaction; it highlights the <u>non constancy</u> (over X_2) of effect of X_1 on (a specific function of) μ_Y. It also plays down the interpretation of phrases like "... interaction effect is of a *reinforcement or synergistic* type" or "of an *interference* or *antagonistic* type". (p311)
>
> **Oxford English Dictionary** ...
>
> > *interact*:-   "to act reciprocally, to act on each other"
> >
> > *interaction*:-   "reciprocal action, action of persons or things on each other"

Interpretation of models that include interaction (product) terms

$$E[Y \mid X_1 , X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

$$= (\beta_0 + \beta_2 X_2) + (\beta_1 + \beta_3 X_2) X_1 \ \text{... if wish to think of E[Y] vs } X_1$$

$$= (\beta_0 + \beta_1 X_1) + (\beta_2 + \beta_3 X_1) X_2 \ \text{... if wish to think of E[Y] vs } X_2$$

mathematically symmetric in X_1 and X_2 (although seldom so in practice)

Interpretation ...

If one of the two X's is the natural "modifier" of the "Y - other X" relation, easier to refer to modifier by another symbol (M) and write equation as

$$E[Y \mid X, M] = \beta_0 + \beta_1 X + \beta_2 M + \beta_3 X M$$

$$= (\beta_0 + \beta_2 M) + (\beta_1 + \beta_3 M) X$$

Don't try to interpret the β_3 "in isolation". And, although the several different (M-specific) X-Y relations can be represented in a single equation, remember that they must be separated out when describing them.. i.e.

interaction   <--->   "Different Y vs X story" for each level of M
"no single summary that applies to all levels of M"

Think of β for product term as "additional Y-X slope for each unit difference in M"

If β for product term is small (in sense that Y-X slope isn't that different from one end of the M range to the other) , then the Y-X slope obtained by dropping M from model is not that misleading

("105 lbs. + 5.5 lbs. for every inch over 5 feet" ... for adults of either sex??)

"Removable" vs. "Non-Removable" Interactions
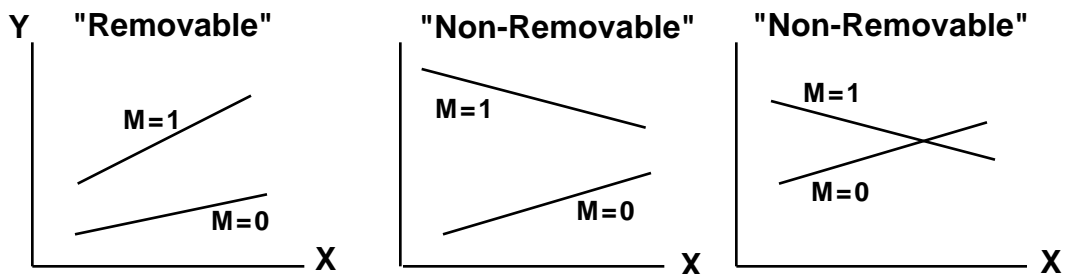
X & M both Binary => β associated with XM product is a"double difference"

$$[\ \overline{y}_{X=1} - \overline{y}_{X=0} \ ]_{M=1} - [\ \overline{y}_{X=1} - \overline{y}_{X=0} \ ]_{M=0}$$

Statistical precision/power to measure/detect a double difference considerable lower than that to measure a single difference. We often end up not being able to adequately statistically test if X differences in response are M-specific, and so depend on analogy or other outside information judgment when deciding whether to report them separately (M-specifically).

Warning in most textbooks: if put product term in model, then must also include each component of the product (i.e. X_1 and X_2 as well as X_1X_2)
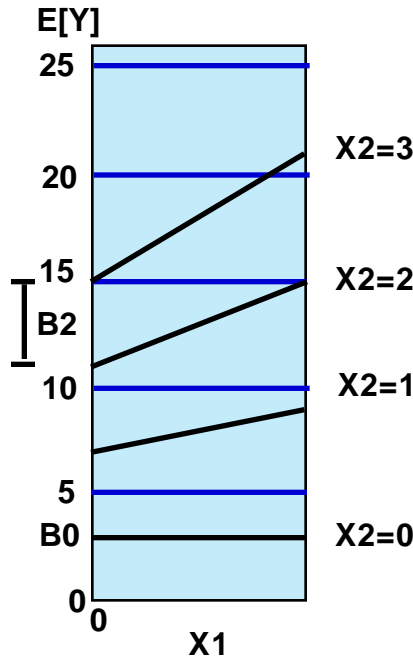
Absolutely must? No. BUT be careful to interpret coefficients carefully! Helps to draw the lines e.g.

$E[Y \mid X_1, X_2] =$

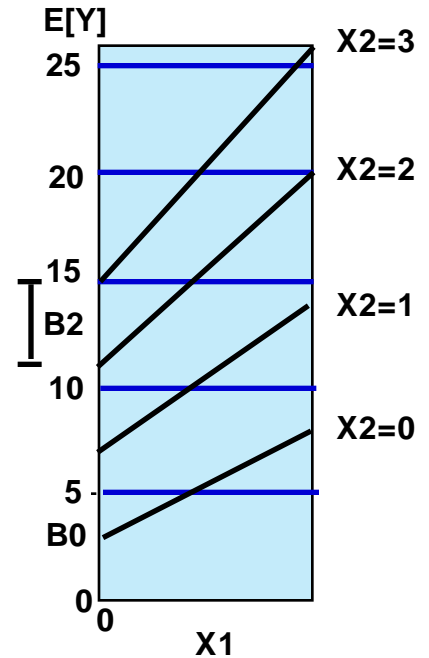$B_0 + B_3X_1X_2$
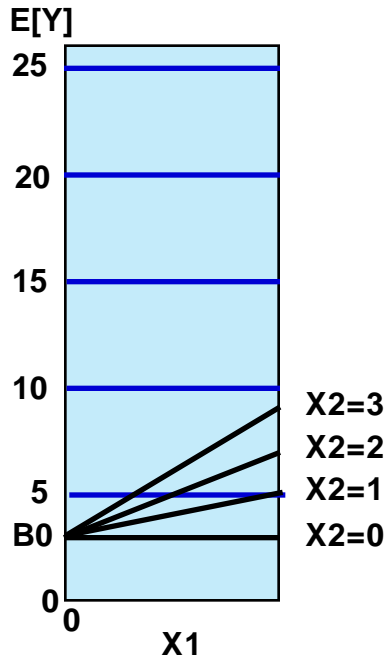
$B_0 + B_2X_2 + B_3X_1X_2$

$B_0 + B_1X_1 + B_2X_2 + B_3X_1X_2$

Product Terms to test change in Y level or Y-time slope (or both) when changes introduced (serially)

   Examples

      - Prescriptions filled before and after the introduction of the Quebec Drug Plan
      - Motor vehicle fatality rates before/after change back to 65 mph limits
      - Asthma deaths before/after removal of certain asthma drug in New Zealand
      - Numbers of Marriage Licences issued before/after HIV tests became mandatory
      - What Does It Take to Heat a New Room? (see datasets on 697 www page)

Product Terms to test change in Y level or Y-time slope (or both) when changes introduced (parallel groups)

   Example

      -The Lidkoping Accident Prevention Programme -- a community approach to preventing childhood injuries in Sweden (see www material in "datasets" in course 626)

Reducing Collinearity of Product Term and its Components ... by Centering

   Example

      -The Lidkoping Accident Prevention Programme   (X=Time   M = Program)

7.9   Constrained regression

- See text