Applied Regression Analysis

Week 5

1. Homework week 4: highlights
2. Multiple regression
   - Graphical interpretation / Assumptions I
   - Assumptions II / Least squares estimation
   - Computer output
   - Step by step review
3. Hypothesis testing: F-test / partial F-test
4. Homework

Stanley Lemeshow, Professor of Biostatistics

College of Public Health, The Ohio State University

THE OHIO STATE UNIVERSITY

WEEK 5: MULTIPLE REGRESSION

Suppose we wish to predict one variable, y, from k independent variables $x_1, x_2, \ldots, x_k,\ k > 1$

$y =$ "dependent" variable

$x_1, x_2, \ldots, x_k =$ "independent" variables

The general form of the regression model for k independent variables is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

error

Regression coefficients
That need to be estimated

Independent variables

Note: in the 2nd order model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

if we let

$$\left.\begin{aligned} x_1 &= x \\ x_2 &= x^2 \end{aligned}\right\}$$

here we really have 1 independent variable. x_2 is a function of that variable.

Then we can write this as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \varepsilon$$

In the multiple regression model some of the x_i may be functions of a few basic variables.

It should be noted that, with respect to what has come before

(1) It is sometimes difficult to determine the best choice of model.

- There will sometimes be several reasonable candidates to choose from.

(2) It is difficult (if not impossible) to visualize what the fitted model looks like.

- not possible to plot the data or the model when $k > 3$.

(3) Sometimes the best-fitting model will be difficult to interpret in real-life terms.

(4) Computations can't be done by hand

- high-speed computers are necessary
- reliable packaged computer program is necessary

Example:

$$y = \text{weight} \quad (\text{WGT})$$

$$x_1 = \text{height} \quad (\text{HGT})$$

$$x_2 = \text{age} \quad (\text{AGE})$$

There are n = 12 children available, each having a particular kind of nutritional deficiency

The data are:

	y	x_1	x_2
Child	WGT	HGT	AGE
1	64	57	8
2	71	59	10
3	53	49	6
4	67	62	11
5	55	51	8
6	58	50	7
7	77	55	10
8	57	48	9
9	56	42	10
10	51	42	6
11	76	61	12
12	68	57	9

Many models are possible. For example

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

or

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where $x_3 = x_1^2$

or

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

where $x_3 = x_1^2,\ x_4 = x_2^2,\ x_5 = x_1 x_2$

so, this is equivalent to

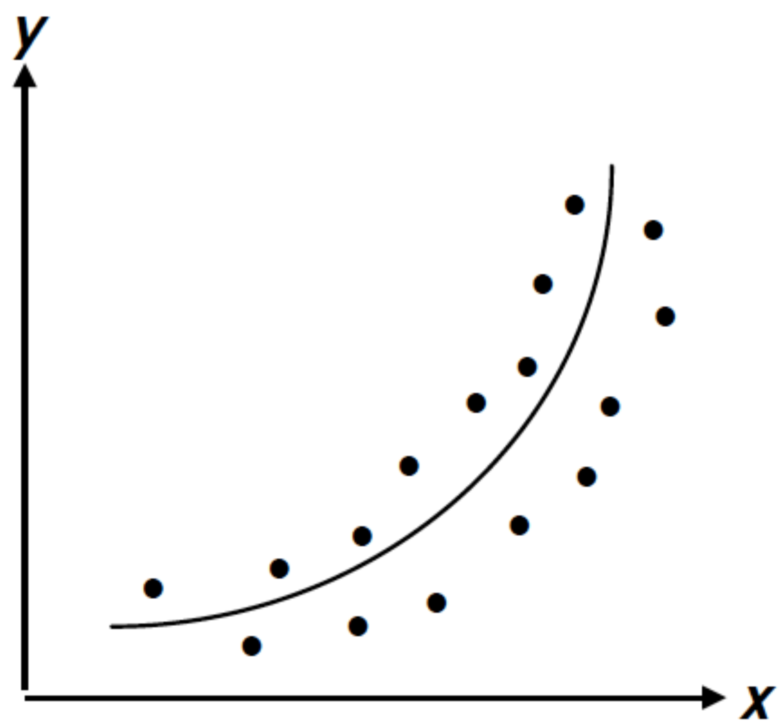
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$$

choice of best model is a topic to be considered later

One reasonable criterion might be to choose the one with the max R^2.

Graphical Interpretation

If we had a single independent variable our lives would be quite simple (even if we have higher-order polynomial models



The regression equation is the path described by the mean values of the distribution of *y* when x is allowed to vary.

When $k \geq 2$ our problems increase significantly

We no longer deal with a line or a curve but, rather, with a hyper surface in (k + 1) - dimensional space.

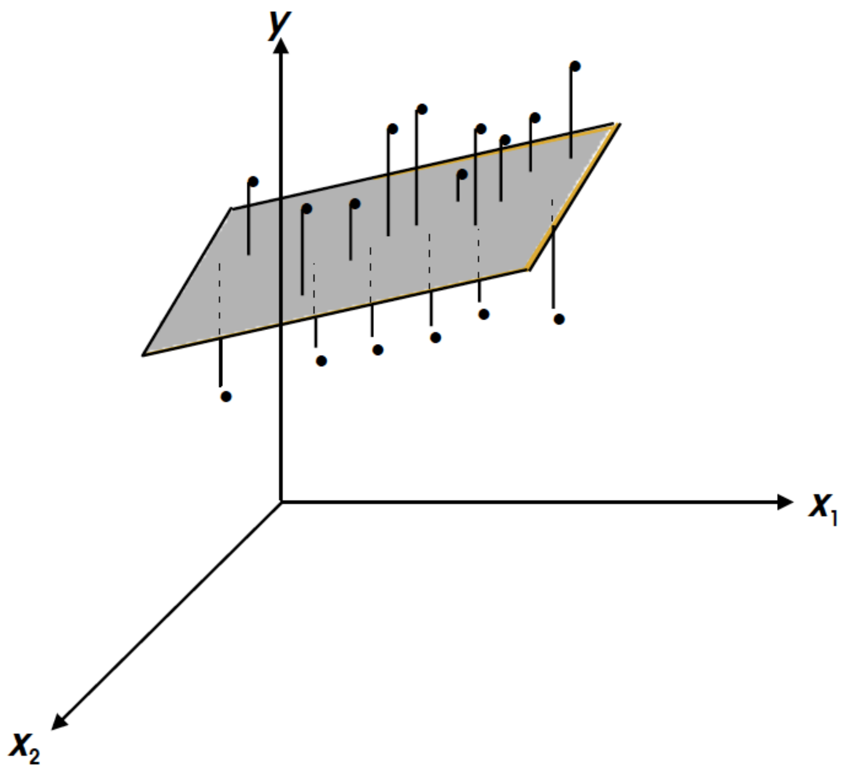
If $k > 2$, we can't plot the scatter of points or the regression equation.

For $k = 2$ we seek the surface in 3-dimensional space that best fits the scatter of points $\left(x_{11}, x_{21}, y_1\right), \left(x_{12}, x_{22}, y_2\right), \ldots, \left(x_{1n}, x_{2n}, y_n\right)$

In this case, the regression equation is the surface described by the mean values of y at various combinations of x_1, x_2.

i.e., at each distinct pair of values x_1 and x_2 there is a distribution of y values with mean $\mu_{y|x_1,x_2}$ and variance $\sigma^2_{y|x_1,x_2}$.

- The simplest curve in two-dimensional space is the straight line.
- The simplest surface in three-dimensional space is a plane that has the statistical model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



In the three dimensional case, the least squares solution giving the best fitting plane is determined by minimizing the sum of squares of distances between the observed y_i and the predicted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$ based on the fitted plane.

i.e., minimize $\quad SSE = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)^2$

Assumptions of Multiple Regression

(1) For each specific combination of $x_1, x_2, \ldots, x_k$, y is a (univariable) random variable with a certain probability distribution.

(2) The y observations are statistically independent.

(3) The mean value of y at $x_1, x_2, \ldots, x_k$ is a linear function of $x_1, \ldots, x_k$.

i.e., $\mu_{y|x_1, x_2, \ldots, x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

or $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$

Note:

(a) The surface $\mu_{y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ is called the *regression equation* or *response surface* or *regression surface*.

(b) If some of the independent variables are higher-order functions of a few basic independent variables (e.g., $x_3 = x_1^2$, $x_5 = x_1 x_2$) then $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ is really nonlinear in the basic variables. Hence we use the term "surface" rather than "plane".

We can use the multiple regression techniques so long as the model is inherently linear in the regression coefficients.

e.g., $\mu_{y|x} = \beta_0 e^{\beta_1 x}$ is inherently linear

since $\ln\left(\mu_{y|x}\right) = \ln\left(\beta_0\right) + \beta_1 x_1$

$$\mu^*_{y|x} = \beta^*_0 + \beta_1 x_1$$

However,

For this we need nonlinear regression procedures

$$\mu_{y|x_1,x_2} = e^{\beta_1 x_1} + e^{\beta_2 x_2}$$

cannot be transformed directly into a form that is linear in β_1 and β_2

(c) ε is the error component in the model. It is the amount by which any individual's observed response deviates from the response surface.

Assumptions (cont'd)

(4) $\sigma^2_{y|x_1,x_2,\ldots,x_k} = \mathrm{var}\left(y \middle| x_1, x_2, \ldots, x_k\right) \equiv \sigma^2$

i.e., homoskedasticity

In general, mild departures from this assumption will not adversely affect the results.

(5) For any fixed $x_1, x_2, \ldots, x_k$ y is normally distributed

i.e., $Y \sim N\left(\mu_{y|x_1,\ldots,x_k}, \sigma^2\right)$

These assumptions are not necessary for obtaining least squares estimates but are necessary for hypothesis testing and other inferential techniques.

Fortunately, usual parametric techniques used in regression analysis are "robust" in the sense that only extreme departures from the assumptions may yield spurious results.

Least Squares Estimates of Parameters

Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

denote the fitted least squares regression model

The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are chosen so that

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki} \right)^2$$

is smaller than would be the case with any other value of $\hat{\beta}_i$

This minimum sum of squares is generally called the

"residual sum of squares"

"error sum of squares"

"sum of squares about regression"

The $\hat{\beta}_i$ determined with the method of least squares are also the minimum variance unbiased estimates of β_i.

The least-squares regression equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

is that unique linear combination of the independent variables $x_1, x_2, \ldots, x_k$ that has maximum possible correlation with the dependent variable.

i.e.,

$r_{y,\hat{y}}$ is greater than $r_{y,\hat{y}'}$ where $\hat{y}'$ is any other linear combination of the x's

Also note:

- each $\hat{\beta}_i$ is a linear function of the y values
- since y is assumed to be normally distributed, each of the estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ will be normally distributed
- computer programs will provide us with these as well as their estimated variances. $t-$ tests and confidence intervals would be carried out in the usual manner.

Example

For the WGT, HGT and AGE data

$$\text{WGT} = \beta_0 + \beta_1\text{HGT} + \beta_2\text{AGE} + \beta_3\text{AGE}^2 + \varepsilon$$

The least squares estimates are

$$\hat{\beta}_0 = 3.438 \qquad \hat{\beta}_1 = 0.724$$

$$\hat{\beta}_2 = 2.777 \qquad \hat{\beta}_3 = -0.042$$

so

$$\widehat{WGT} = 3.438 + .724(HGT) + 2.77(AGE) - .042(AGE)^2$$

The ANOVA table for this model is:

ANOVA

Source	df	SS	MS	F
Regression	3	SSY − SSE = 693.06	231.02	9.47
Residual	8	SSE = 195.19	24.40	
Total	11	SSY = 888.25		

$R^2 = 0.7802$

To get the ANOVA table we use the familiar partitioning

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

SSY

SSY-SSE

SSE

Total SS = total variability in *y* before accounting for the joint effect of using the independent variables HGT, AGE, AGE^2

Residual SS = SS due error
= amount of *y* variation left unexplained after the independent variables have been used in the regression equation to predict *y*.

Regression SS = reduction in variation (or variation explained) due to the independent variables in the regression equation.

now, to test H_0: all k independent variables considered together do not explain a significant amount of the variation in y,

or $\quad H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

vs. $\quad H_a : \text{some } \beta_i \neq 0$

we use

$$F = \frac{\text{MS regression}}{\text{MS residual}}$$

and $F \sim F(k, n-1-k)$

The hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

vs. H_a : not all $\beta_i = 0$

can also be tested by an equivalent expression

$$F = \frac{R^2}{1 - R^2} \frac{(n-1-k)}{k}$$

which is also compared to $F(k, n-1-k)$

note: $\boxed{R^2 = \frac{\text{SSY} - \text{SSE}}{\text{SSY}}}$

Example

In the HGT, WGT, AGE example, from the ANOVA table

$$F = \frac{\text{MS regression}}{\text{MS residual}} = \frac{231.02}{24.40} = 9.47$$

and $F_{.99}(3,8) = 7.59 \therefore p < .01$

$\therefore$ reject H_0

also

$$F = \frac{R^2}{1-R^2}\frac{(n-1-k)}{k} = \frac{.7802}{1-.7802}\frac{(12-1-3)}{3} = 9.47$$

note:

MS residual $= \dfrac{1}{n-1-k}\text{SSE} = \dfrac{1}{n-1-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is an unbiased estimate of σ^2 under the assumed model.

MS regression is an independent estimate of σ^2 only if H_0 is true. Otherwise it overestimates σ^2.

Hence we always reject if F gets too large.

```
. gen agesq=age*age

. regress wgt hgt age agesq

  Source |       SS       df       MS              Number of obs =      12
---------+------------------------------           F(  3,      8) =    9.47
   Model |  693.060463     3  231.020154           Prob > F      =  0.0052
Residual |  195.189537     8  24.3986921           R-squared     =  0.7803
---------+------------------------------           Adj R-squared =  0.6978
   Total |      888.25    11       80.75           Root MSE      =  4.9395

------------------------------------------------------------------------------
     wgt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     hgt |   .7236902   .2769632      2.613   0.031       .085012    1.362368
     age |   2.776875   7.427279      0.374   0.718     -14.35046    19.90421
   agesq |  -.0417067   .4224071     -0.099   0.924     -1.015779    .9323659
   _cons |   3.438426   33.61082      0.102   0.921     -74.06826    80.94512
------------------------------------------------------------------------------

. vif

Variable |       VIF       1/VIF
---------+----------------------
   agesq |      89.97    0.011115
     age |      89.68    0.011150
     hgt |       1.61    0.620927
---------+----------------------
Mean VIF |      60.42
```

```
. regress wgt hgt age

  Source |       SS       df       MS              Number of obs =      12
---------+------------------------------           F(  2,      9) =   15.95
   Model |  692.822607     2  346.411303           Prob > F      =  0.0011
Residual |  195.427393     9  21.7141548           R-squared     =  0.7800
---------+------------------------------           Adj R-squared =  0.7311
   Total |      888.25    11       80.75           Root MSE      =  4.6598

------------------------------------------------------------------------------
     wgt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     hgt |    .722038   .2608051      2.768   0.022        .1320559    1.31202
     age |   2.050126   .9372256      2.187   0.056       -.0700253   4.170278
   _cons |   6.553048   10.94483      0.599   0.564       -18.20587   31.31197
------------------------------------------------------------------------------

. vif

Variable |       VIF       1/VIF
---------+----------------------
     age |      1.60    0.623202
     hgt |      1.60    0.623202
---------+----------------------
Mean VIF |      1.60
```

```
. regress wgt hgt

  Source |       SS       df       MS              Number of obs =      12
---------+------------------------------           F(  1,    10) =   19.67
   Model |  588.922523     1  588.922523           Prob > F      =  0.0013
Residual |  299.327477    10  29.9327477           R-squared     =  0.6630
---------+------------------------------           Adj R-squared =  0.6293
   Total |      888.25    11       80.75           Root MSE      =  5.4711

------------------------------------------------------------------------------
     wgt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     hgt |    1.07223    .241731      4.436   0.001       .5336202    1.610841
   _cons |   6.189849   12.84875      0.482   0.640      -22.43894    34.81864
------------------------------------------------------------------------------

. vif

Variable |       VIF       1/VIF
---------+----------------------
     hgt |       1.00    1.000000
---------+----------------------
Mean VIF |       1.00
```

The previous ANOVA Table may be presented as follows.

Source		df	SS	MS	F
Regression	x_1	1	588.92	588.92	19.67 ***
	$x_2 \mid x_1$	1	103.90	103.90	4.78 *
	$x_3 \mid x_1, x_2$	1	0.24	0.24	0.01 NS
Residual		8	195.19	24.40	
Total		11	888.25		

*** p<.01   * .05<p<.10

Here

$SS(x_1) =$ SS explained just using x_1 = HGT alone

$SS(x_2 \mid x_1) =$ extra SS explained by using x_2 = AGE in addition to x_1 in predicting y

$SS(x_3 \mid x_1, x_2) =$ extra SS explained by using $x_3 = AGE^2$ in addition to x_1 and x_2 in predicting y

Questions:

1. Does x_1 = HGT alone significantly aid in predicting *y* ?
2. Does the addition of x_2 = AGE significantly contribute to the prediction *of y* after controlling for the contribution of x_1 ?
3. Does the addition of x_3 = AGE^2 significantly contribute to the prediction of *y* after controlling for the contribution of x_1 and x_2 ?

Let us consider these one at a time

Question 1: Does x_1 = HGT alone significantly aid in predicting y ?

Fit the straight line model $y = \beta_0 + \beta_1 \times \text{HGT}$

$$SS(x_1) = 588.92 = \text{regression SS for this straight line model}$$

$$\text{SSE} = SS(x_2 | x_1) + SS(x_3 | x_1, x_2) + \text{SS resid}$$

$$= 103.90 + 0.24 + 195.19 = \boxed{299.33}$$

$$\text{df} = \text{df}(x_2 | x_1) + \text{df}(x_3 | x_1, x_2) + \text{df resid}$$

$$= 1 + 1 + 8 = \boxed{10}$$

$$\therefore \text{MS resid} = 299.33 / 10 = 29.933$$

and

$$F = \frac{\text{MS regression}}{\text{MS resid}} = \frac{588.92}{29.933} = \boxed{19.67} \text{ as in the table}$$

$$F \sim F(1,10) \text{ here } p < .01$$

i.e., x_1 contributes significantly to the linear prediction of y

Question 2: Does the addition of x_2 = AGE significantly contribute to the prediction of y after controlling for the contribution of x_1?

To answer this we use a "partial F-test". This test allows for the elimination of variables that are of no help in predicting y and thus enables one to reduce the set of possible independent variables to an economical set of "important" predictors.

To perform a partial F-test concerning a variable x^*, say, given that $x_1, x_2, \ldots, x_p$ are already in the model we:

(1) Compute the "extra SS from adding x^*, given $x_1, x_2, \ldots, x_p$"

- This is placed into the ANOVA table under the source heading "Regression $x^* | x_1, x_2, \ldots, x_p$"

Extra SS from adding x^* given $x_1, x_2, \ldots, x_p$ = regression SS when $x_1, x_2, \ldots, x_p$ and x^* are all in the model − regression SS when $x_1, x_2, \ldots, x_p$ and not x^* are all in the model

or

$$\boxed{\begin{aligned} SS\left(x^* \middle| x_1, x_2, \ldots, x_p\right) = {} & \text{regression SS}\left(x_1, x_2, \ldots, x_p, x^*\right) \\ & - \text{regression SS}\left(x_1, x_2, \ldots, x_p\right) \end{aligned}}$$

In our example

$$SS\left(x_2 \middle| x_1\right) = \text{regression } SS\left(x_1, x_2\right) - \text{regression } SS\left(x_1\right)$$

$$= 692.82 - 588.92$$

$$= 103.90$$

$$SS\left(x_3 \middle| x_1, x_2\right) = \text{regression } SS\left(x_1, x_2, x_3\right) - \text{regression } SS\left(x_1, x_2\right)$$

$$= 693.06 - 692.82$$

$$= 0.24$$

To test

H_0 : The addition of x^* to a model already containing $x_1, x_2, \ldots, x_p$ does not significantly improve the prediction of y

we compute

$$\boxed{F\left(x^* \middle| x_1, x_2, \ldots, x_p\right) = \frac{\mathrm{SS}\left(x^* \middle| x_1, x_2, \ldots, x_p\right)}{\mathrm{MS\ residual}\left(x_1, x_2, \ldots, x_p, x^*\right)}}$$

and

$$F\left(x^* \middle| x_1, x_2, \ldots, x_p\right) \sim F\left(1, n-p-2\right)$$

In our example

$$F\left(x_2 \middle| x_1\right) = \frac{SS\left(x_2 \middle| x_1\right)}{\text{MS residual}\left(x_1, x_2\right)} = \frac{103.90}{\left(\dfrac{.24 + 195.19}{1+8}\right)} = 4.78$$

$F_{.90}(1,9) = 3.36;\ F_{.95}(1,9) = 5.12$

and

$$F\left(x_3 \middle| x_1, x_2\right) = \frac{SS\left(x_3 \middle| x_1, x_2\right)}{\text{MS residual}\left(x_1, x_2, x_3\right)} = \frac{0.24}{24.40} = 0.01$$

Hence, the addition of x_2 after accounting for x_1 significantly adds to the prediction of y at the $\alpha = 0.10$ level.

Had we used $\alpha = 0.05$ we would not add x_2.

Once x_1 = HGT and x_2 = AGE are in the model, the addition of $x_3 = AGE^2$ is superfluous.

There is an alternative (but equivalent) way to perform the partial F-test. That involves a test of

$$H_0 : \beta^* = 0$$

where β^* is the coefficient of x^* in

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \beta^* x^* + \varepsilon$$

Here

$$t = \frac{\hat{\beta}^*}{s_{\hat{\beta}^*}} \begin{array}{l} \leftarrow \text{estimated coefficient} \\ \leftarrow \text{estimated standard error of } \hat{\beta}^* \end{array}$$

printed by computer programs

$$\left.\begin{array}{l} \text{reject } H_0 \text{ if } t > t_{1-\alpha/2}(n-p-2) \\ \text{or if } t < t_{\alpha/2}(n-p-2) \end{array}\right\} \text{2- sided test of } H_0$$

$$H_a : \beta^* \neq 0$$

similarly, one sided tests can be constructed

e.g., for $H_a : \beta^* > 0$ $\quad$ (reject if $t > t_{1-\alpha}(n-p-2)$)

In our example

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

in the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Then

$$t = \frac{\hat{\beta}_2}{s_{\hat{\beta}_2}} = \frac{2.050}{0.937} = 2.188$$

and $t_{.95}(9) = 1.833$, $t_{.975}(9) = 2.2622$

Hence $.05 < p < .10$ since 2-sided

Note $t^2 = 2.188^2 = 4.79 =$ partial $F\left(x_2 \middle| x_1\right)$

in ANOVA table

and

$$t^2_{1-\alpha/2}\left(9\right) = F_{1-\alpha}\left(1,9\right)$$

Similarly, when testing

$$H_0 : \beta_3 = 0$$

vs $\quad H_a : \beta_3 \neq 0$

in the model $\quad y = \beta_0 + \beta_1 x_1 + \beta_3 x_2 + \beta_3 x_3 + \varepsilon$

we compute

$$t = \frac{\hat{\beta}_3}{s_{\hat{\beta}_3}} = \frac{-.042}{.422} = -.0995$$

and

$$t^2 = \left(-.0995\right)^2 = .01 = \text{ Partial } F\left(x_3 \middle| x_1, x_2\right)$$

in ANOVA table