Multiple Linear Regression

- is used to relate a *continuous* response (or dependent) variable Y to several explanatory (or independent) (or predictor) variables $X_1, X_2, \ldots, X_k$

- assumes a linear relationship between mean of Y and the X's with additive normal errors

- X_{ij} is the value of independent variable j for subject i.

- Y_i is the value of the dependent variable for subject i, $i = 1, 2, \ldots, n$.

- Statistical model

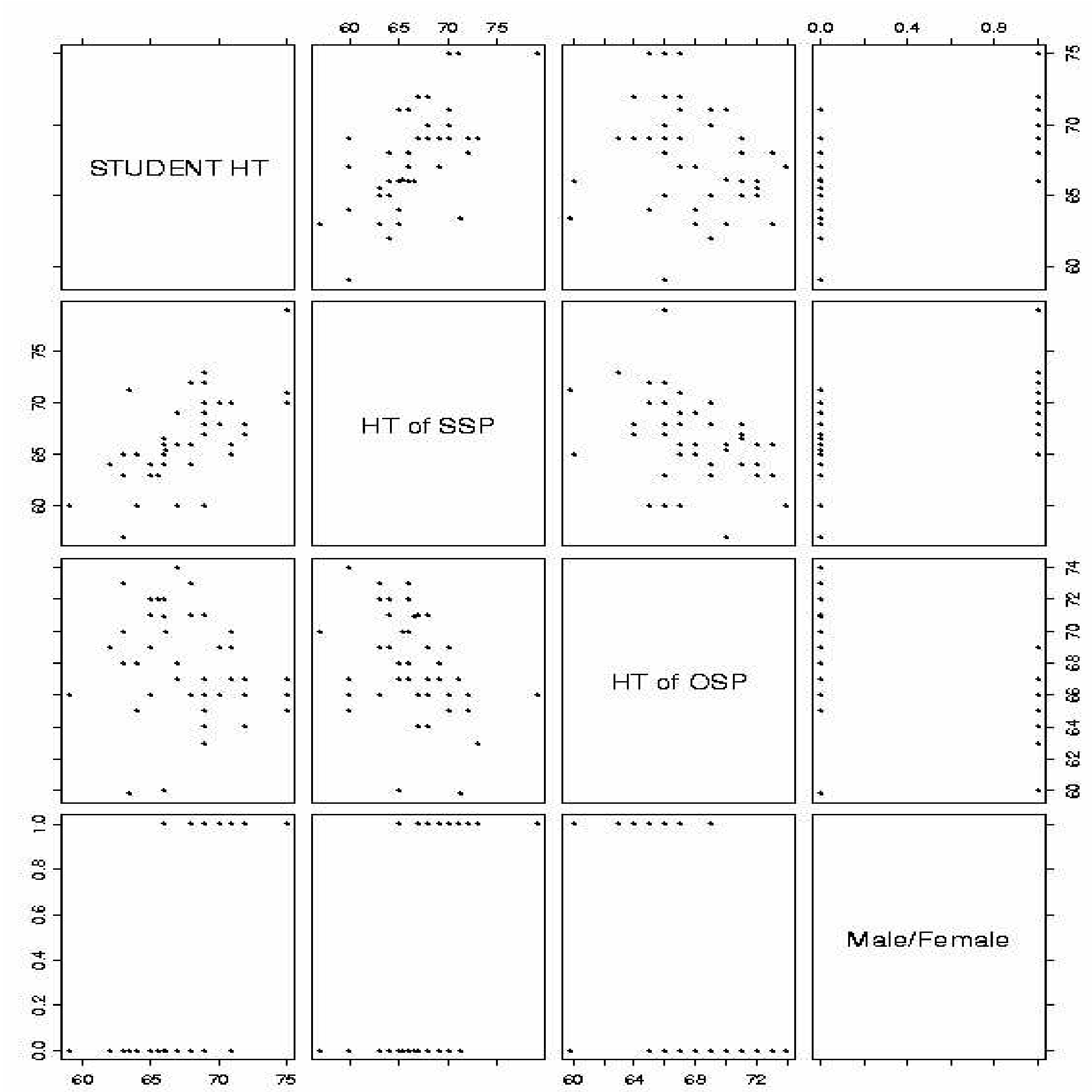
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$$

the errors are assumed to be a sample from $N(0, \sigma^2)$

Example 1: The data set **height.mtw** can be found on WebCT, in the data files folder. The variables in the data set include Y (height of student), X_1 (height of same sex parent), X_2 (height of opposite sex parent), and sex of the student ($X_3 = 1$ for males; $X_3 = 0$ for females).

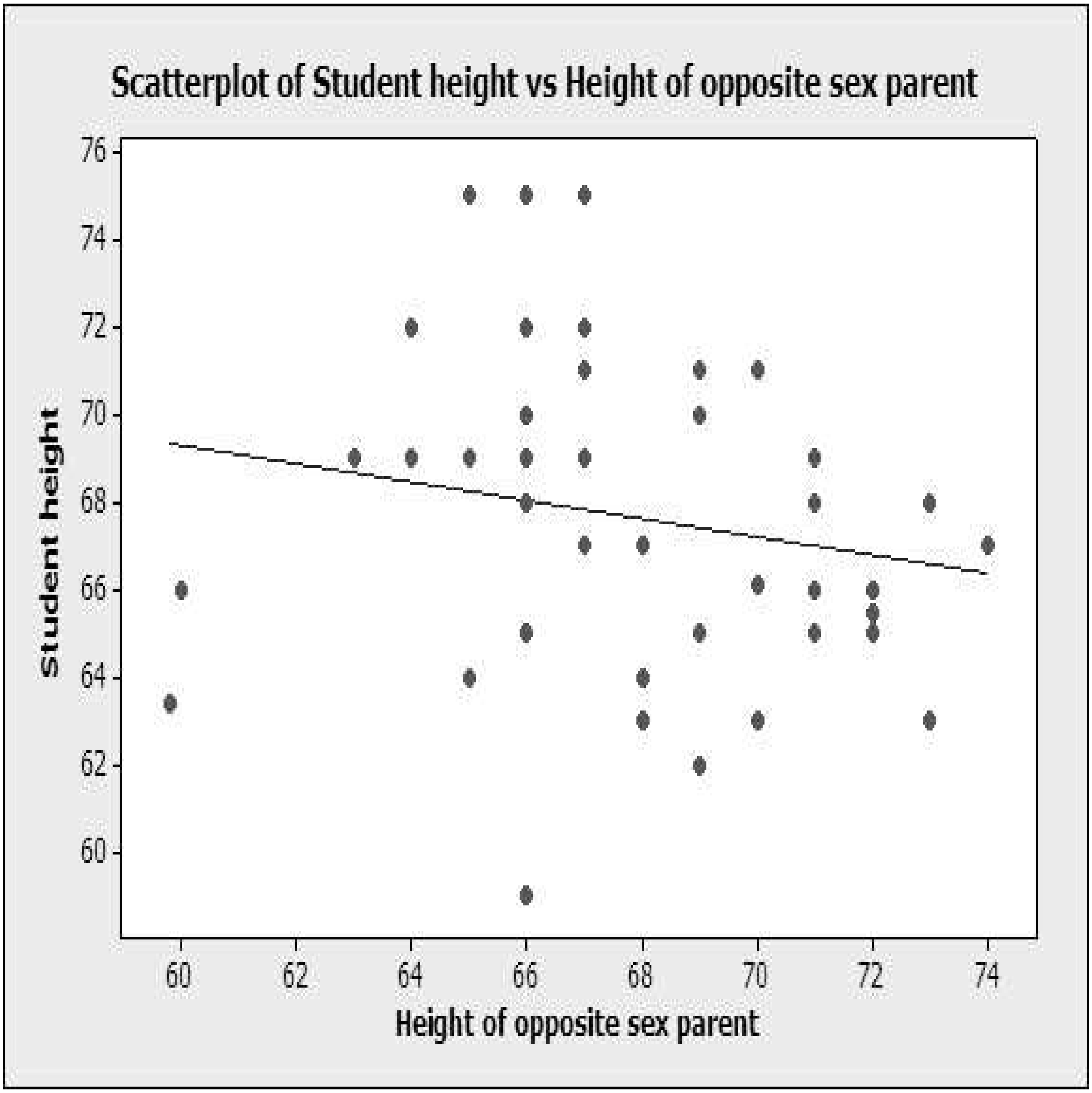
The following minitab output gives the various Pearson correlations for different pairs.

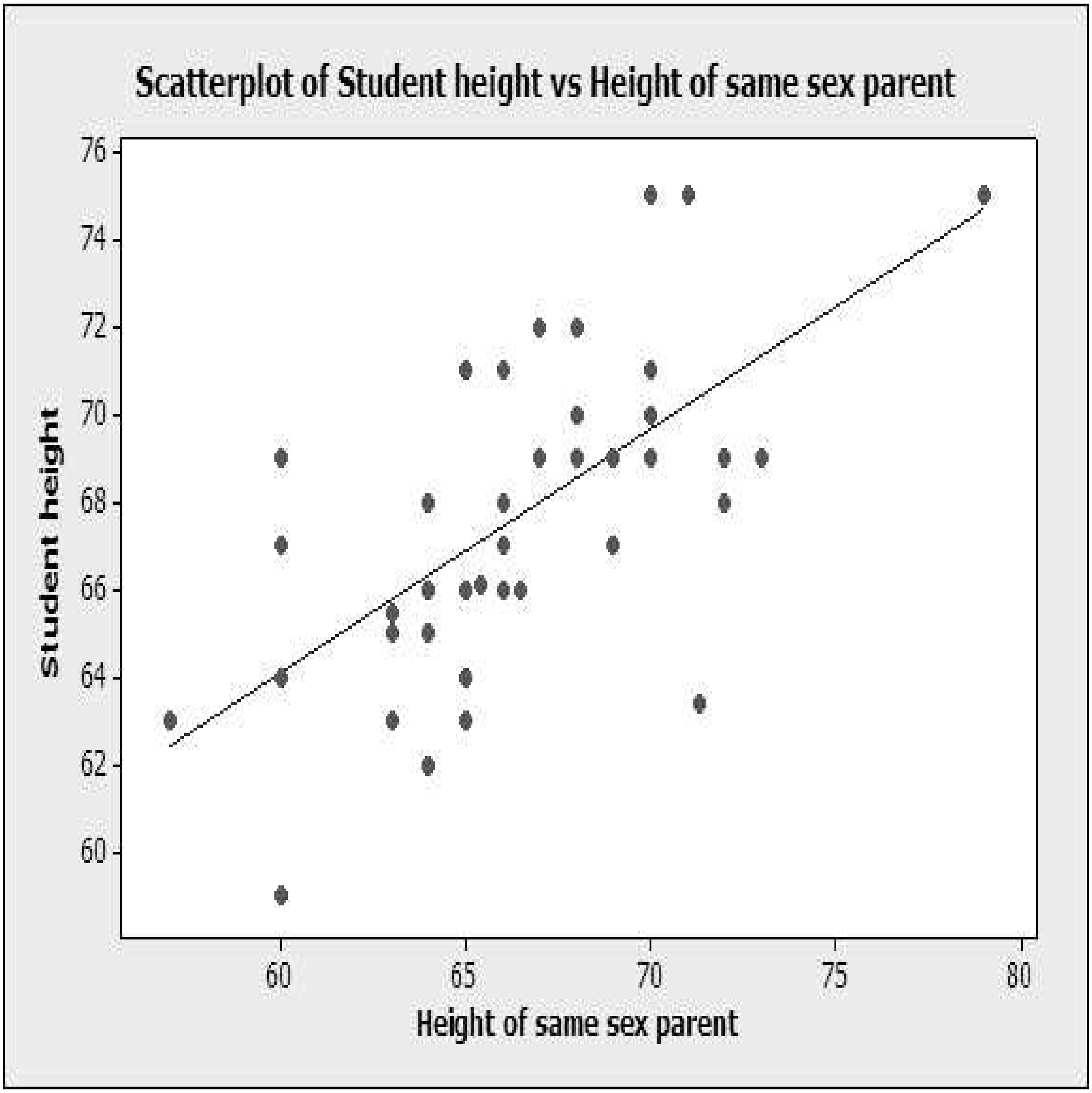
```
Results for: HEIGHTS.MTW
MTB > corr c1-c4

                Student heig  Height of sa  Height of op
Height of sa           0.641
                       0.000

Height of op          -0.193        -0.408
                       0.205         0.005

Male or Fema           0.692         0.606        -0.539
                       0.000         0.000         0.000

Cell Contents: Pearson correlation
               P-Value
```

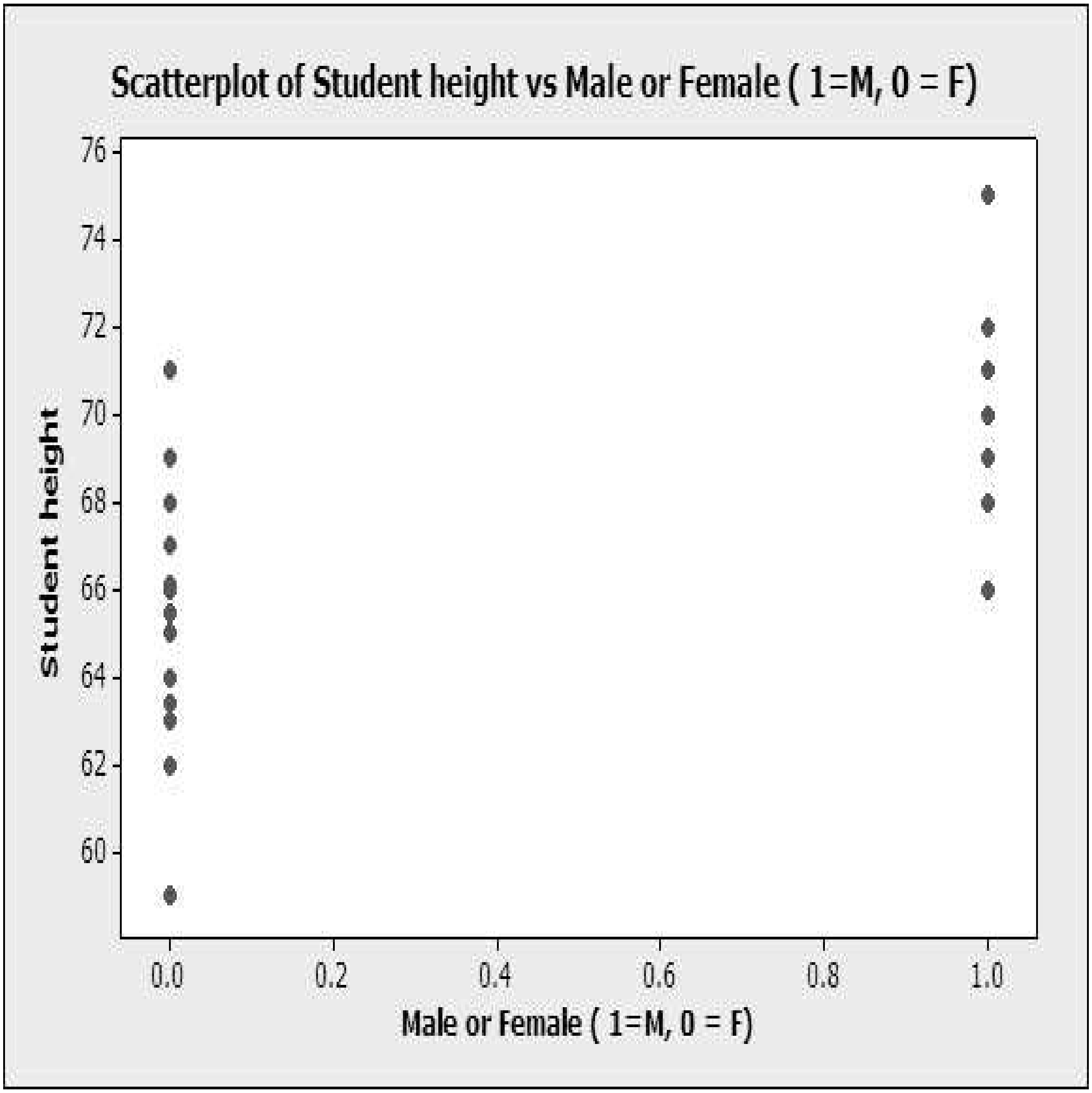


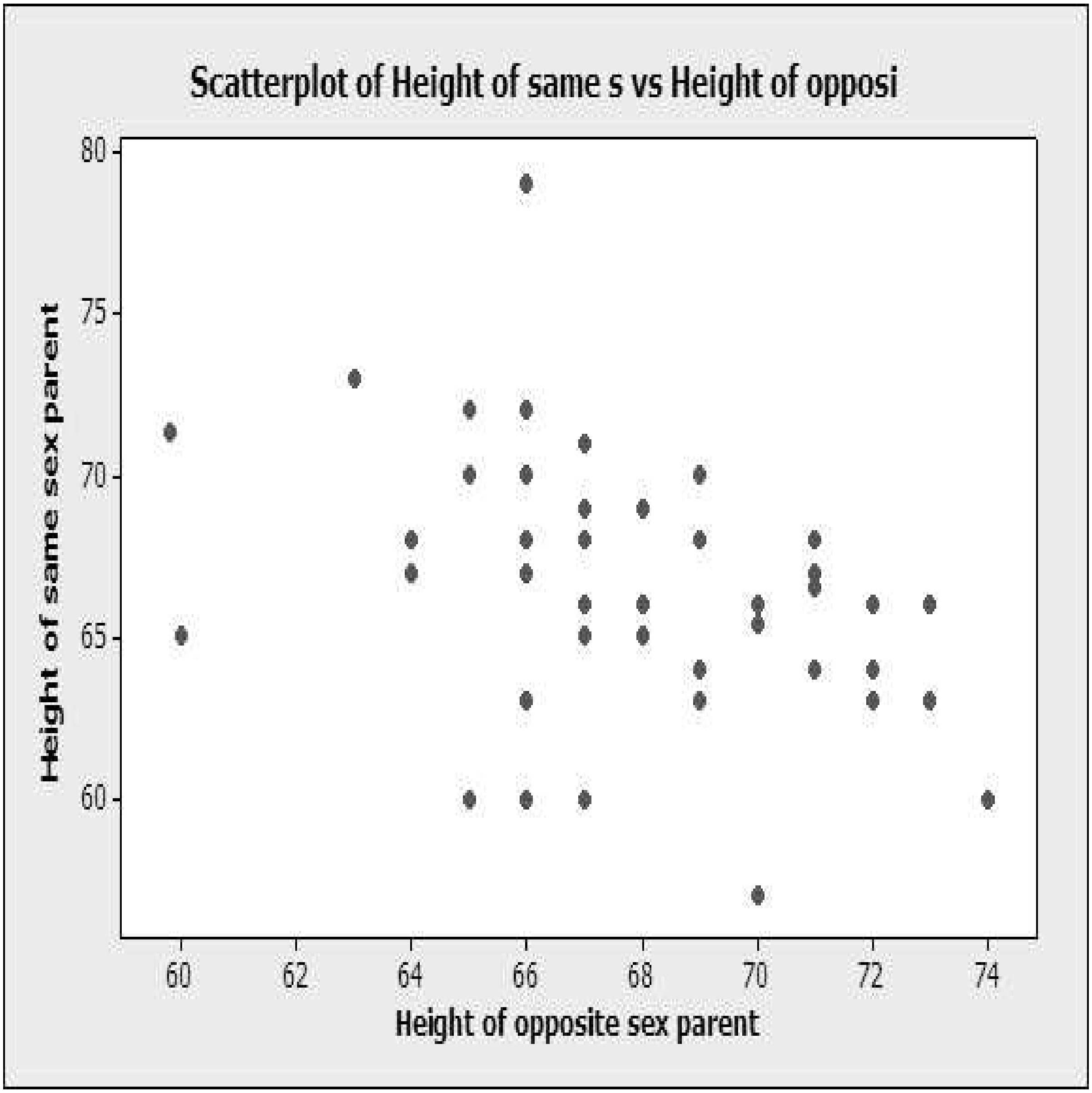
There are a number of possible regression models, including regressions on single or multiple independent variables.

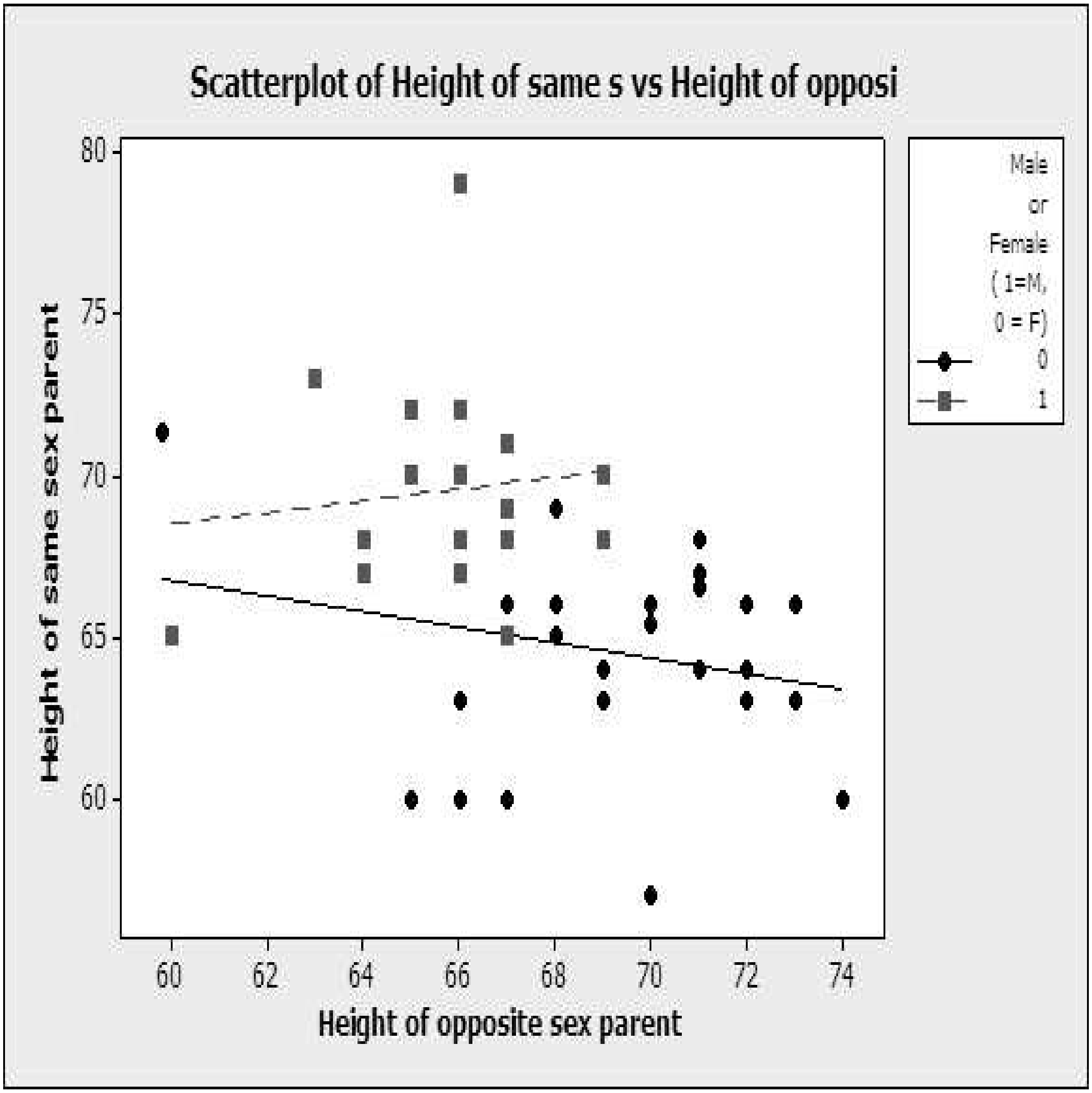
As with simple linear regression, parameters are estimated by least squares. Details of estimation and statistical inference to be discussed later.

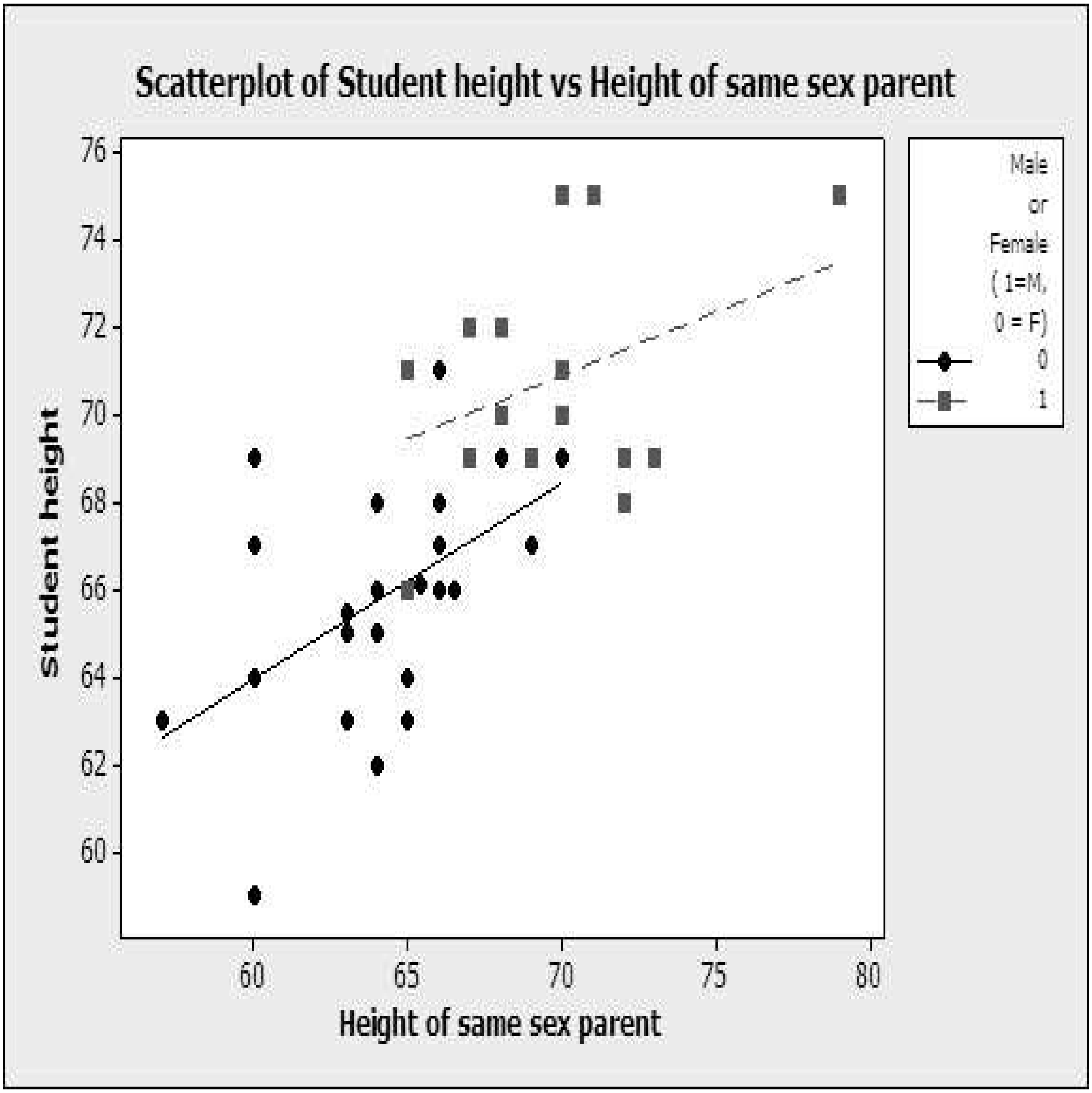

Scatterplot of Student height vs Height of opposite sex parent
Student height
76
74
72
70
68
66
64
62
60
60
62
64
66
68
70
72
74
Height of opposite sex parent

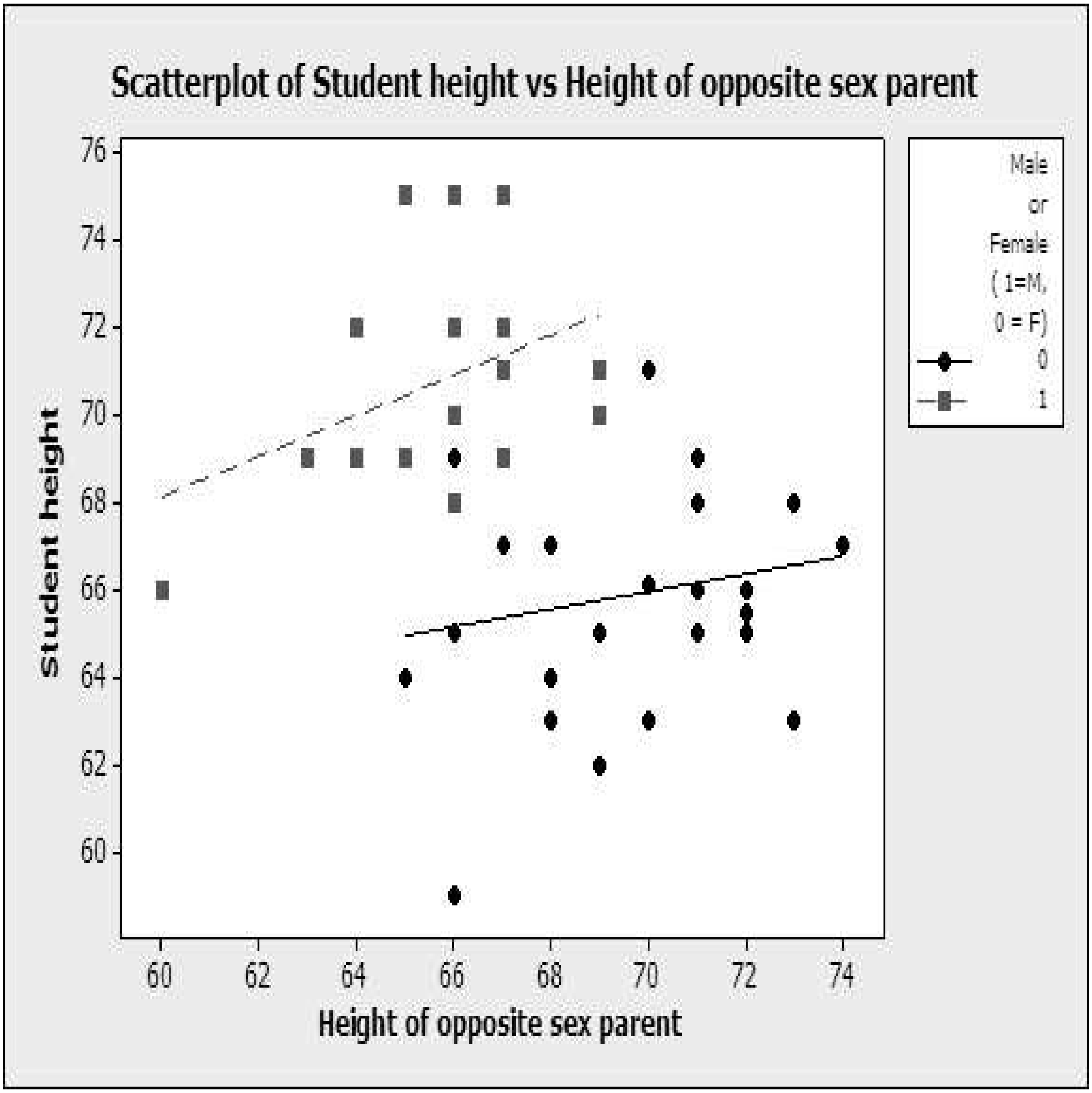

Scatterplot of Student height vs Height of same sex parent
Student height
76
74
72
70
68
66
64
62
60
60
65
70
75
80
Height of same sex parent


Scatterplot of Student height vs Male or Female (1=M, 0 = F)
Student height
76
74
72
70
68
66
64
62
60
0.0
0.2
0.4
0.6
0.8
1.0
Male or Female (1=M, 0 = F)


Scatterplot of Height of same s vs Height of opposi
Height of same sex parent
80
75
70
65
60
60
62
64
66
68
70
72
74
Height of opposite sex parent


Scatterplot of Height of same s vs Height of opposi
Height of same sex parent
Height of opposite sex parent
80
75
70
65
60
60
62
64
66
68
70
72
74
Male
or
Female
(1=M,
0 = F)
0
1


Scatterplot of Student height vs Height of same sex parent
Student height
Height of same sex parent
76
74
72
70
68
66
64
62
60
60
65
70
75
80
Male or Female (1=M, 0 = F)
0
1


Scatterplot of Student height vs Height of opposite sex parent
Student height
76
74
72
70
68
66
64
62
60
60
62
64
66
68
70
72
74
Height of opposite sex parent
Male or Female (1=M, 0 = F)
0
1

Student height versus height of same sex parent

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

```
MTB > regress c1 1 c2

The regression equation is
Student height = 30.7 + 0.557 Height of same sex parent

Predictor                     Coef  SE Coef     T      P
Constant                    30.702    6.765  4.54  0.000
Ht of same sex parent       0.5567   0.1017  5.47  0.000

S = 2.74051   R-Sq = 41.1%   R-Sq(adj) = 39.7%

Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  225.07  225.07  29.97  0.000
Residual Error  43  322.95    7.51
Total           44  548.02
```

Student height versus height of opposite sex parent

$$Y_i = \beta_0 + \beta_1 X_{i2} + \epsilon_i$$

```
MTB > regress c1 1 c3

The regression equation is
Student height = 81.9 - 0.209 Height of opposite sex par

Predictor                     Coef   SE Coef       T       P
Constant                     81.89     11.06    7.40   0.000
Ht opposite sex parent     -0.2094    0.1626   -1.29   0.205

S = 3.50309   R-Sq = 3.7%   R-Sq(adj) = 1.5%

Analysis of Variance

Source          DF      SS     MS     F       P
Regression       1   20.34  20.34  1.66   0.205
Residual Error  43  527.68  12.27
Total           44  548.02
```

The regression equation is
Student height = 22.7 + 0.586 Height of same sex parent + 0.090 Height of opposite sex parent

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

```
MTB > regress c1 2 c2 c3

Predictor                    Coef  SE Coef     T      P
Constant                    22.66    14.30  1.59  0.120
Ht same sex parent         0.5859   0.1122  5.22  0.000
Ht opposite sex parent     0.0897   0.1403  0.64  0.526

S = 2.75953   R-Sq = 41.6%   R-Sq(adj) = 38.9%

Analysis of Variance

Source          DF      SS      MS      F      P
Regression       2  228.19  114.09  14.98  0.000
Residual Error  42  319.83    7.62
Total           44  548.02
```

The regression equation is Student height = 21.5 + 0.338 Height of same sex parent + 0.325 Height of opposite sex parent + 4.44 Male or Female (1=M, 0 = F)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

```
MTB > regress c1 3 c2 c3 c4

Predictor                       Coef   SE Coef     T      P
Constant                       21.50     11.70  1.84  0.073
Ht same sex parent            0.3375    0.1061  3.18  0.003
Ht opposite sex parent        0.3249    0.1254  2.59  0.013
Male/Female (1=M,0=F)         4.4446    0.9534  4.66  0.000

S = 2.25790   R-Sq = 61.9%   R-Sq(adj) = 59.1%

Analysis of Variance

Source          DF      SS      MS      F      P
Regression       3  339.00  113.00  22.16  0.000
Residual Error  41  209.02    5.10
Total           44  548.02
```

Multiple Regression Model

- **Statistical model:**

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \ldots + \beta_1 X_{ik} + \epsilon_i$$

- **Assumptions:**

  The errors are a sample from $N(0, \sigma^2)$

- **Estimation:**

  The regression parameters are estimated using least squares. That is, we choose β_0, $\beta_1, \ldots, \beta_k$ to minimize

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2$$

- **Inferences:**

  The ANOVA table similar to that for simple linear regression, with changes to degrees of freedom to match the number of predictor variables.

Source	d.f.	SS	MS	F
Regression	k	SSR	MSR=SSR/k	F_{obs}=MSF
Residual	n-k-1	SSE	MSE=SSE/(n-k-1)	
Total	n-1	SST		

- SST=SSR+SSE
- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

  H_A: one or more of β_1, β_2, ..., β_k are non-zero
- p-value $= P(F_{k,n-k-1} > F_{obs})$
- The goodness of fit of the linear regression line is measured by the **coefficient of determination** R^2.

$$R^2 = \frac{SSR}{SST}$$

- R^2 is the fraction of the total variability in y accounted for by the regression line, and ranges between 0 and 1. $R^2 = 1.00$ indicates a perfect (linear) fit, while $R^2 = 0.00$ is a complete lack of linear fit.

Matrix Methods for Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$$

X_{ij} the value of explanatory variable j for subject i.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & ... & X_{1k} \\ 1 & X_{21} & X_{22} & ... & X_{2k} \\ 1 & X_{31} & X_{32} & ... & X_{3k} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 1 & X_{n1} & X_{n2} & ... & X_{nk} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

or in matrix form

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix}$$

$$
\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & ... & X_{1k} \\ 1 & X_{21} & X_{22} & ... & X_{1k} \\ 1 & X_{31} & X_{32} & ... & X_{1k} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 1 & X_{n1} & X_{n2} & ... & X_{nk} \end{bmatrix}
$$

$$
\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{pmatrix}
$$

and

$$
\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}
$$

- $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$ denote estimates of the true regression coefficients β_0, β_1,..., β_k. The estimates minimize the sum of squared residuals, or error sum of squares(SSE). Hence, they are then called least squares estimates. The error sum of squares is:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik}))^2 = e'e$$

- SSE is minimized when all $\frac{\partial SSE}{\partial \hat{\beta}_i} = 0$. These conditions result in a system of normal equations which are solved simultaneously to obtain estimates of β_0, β_1, ..., β_k. In matrix form the Normal Equations are

$$(\mathbf{X}'\mathbf{X})\hat{\beta} = (\mathbf{X}'\mathbf{Y})$$

- A unique solution (when it exists) to the normal equations is

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ . \\ . \\ . \\ \hat{\beta}_k \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}(\mathbf{X}'\mathbf{Y})$$

Least squares estimates are calculated using this equation.

- The vector

$$\hat{\mathbf{Y}} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ . \\ . \\ . \\ \hat{Y}_n \end{pmatrix}$$

contains the predicted values, where

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_k X_{ik}$$

In matrix form, the predicted values are given by

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$$

- The error variance, σ^2, is estimated as follows

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-(k+1)} = \frac{\sum(y_i - \hat{y}_i)^2}{n-(k+1)}$$

which is the sum of squares for error (SSE) divided by the error degrees of freedom (n-(k+1)).

- As before, the total sum of squares is $\sum(y_i - \bar{y})^2$, and the total degrees of freedom is $n-1$.
- With these quantities, an ANOVA table can be constructed in the same manner as for simple linear regression.

Source	df	SS	MS	F
Regression	k	SSR	MSR	F_{obs} = MSR/MSE
Error	n-k-1	SSE	MSE	
Total	n-1	SST		

- The observed value of F is used to test the overall utility of the regression model. The formal hypotheses are $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, vs H_A : at least 1 of $\beta_1, \beta_2, \ldots, \beta_k$ is non-zero.
- The p-value is $P(F_{k,n-k-1} > F_{obs})$.

- Example: Consider the following data set with 8 values of the response, Y, and four explanatory variables, X1,X2,X3, and X4.

Y	X1	X2	X3	X4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44

- Regression model using all four explanatory variables is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

- In MINITAB the data are assembled as shown below.

```
MTB > set c1             # Y
DATA> 78.5 74.3 104.3 87.6 95.9 109.2 102.7 72.5
DATA> set c2             # X1
DATA> 7 1 11 11 7 11 3 1
DATA> set c3             # X2
DATA> 26 29 56 31 52 55 71 31
DATA> set c4             # X3
DATA> 6 15 8 8 6 9 17 22
DATA> set c5             # X4
```

```
DATA> 60 52 20 47 33 22 6 44
DATA> end
MTB > print c1-c5
        C1    C2   C3   C4   C5
Row      Y    X1   X2   X3   X4
  1   78.5     7   26    6   60
  2   74.3     1   29   15   52
  3  104.3    11   56    8   20
  4   87.6    11   31    8   47
  5   95.9     7   52    6   33
  6  109.2    11   55    9   22
  7  102.7     3   71   17    6
  8   72.5     1   31   22   44

MTB > copy c1 m1  # Y vector
MTB > set c6
DATA> 8(1)
DATA> end
```

```
MTB > copy c6 c2 c3 c4 c5 m2
MTB > print m2

 Matrix M2    # m2 = X matrix

1   7  26   6  60
1   1  29  15  52
1  11  56   8  20
1  11  31   8  47
1   7  52   6  33
1  11  55   9  22
1   3  71  17   6
1   1  31  22  44
```

- To determine the value of the regression coefficients, the above matrix equation must be solved. To do so, first requires determination of the elements of three other important matrices, $(\mathbf{X}'\mathbf{X})$, $(\mathbf{X}'\mathbf{Y})$, and $(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$.

```
MTB > transpose m2 m3
MTB > multiply m3 m2 m4
```

```
MTB > inverse m4 m5
MTB > print m5

 Matrix M5     # m5 = inv(X'X)

1880.08 -22.6974 -19.0160 -20.9861 -18.5740
 -22.70   0.2914   0.2263   0.2622   0.2223
 -19.02   0.2263   0.1931   0.2100   0.1882
 -20.99   0.2622   0.2100   0.2425   0.2059
 -18.57   0.2223   0.1882   0.2059   0.1839

MTB > multiply m3 m1 m6
MTB > multiply m5 m6 m7
MTB > print m7

 Matrix M7   # m7 = b

48.5583
 1.6154
 0.6921
 0.0588
 0.0150
```

- The resulting regression model is

$$\hat{Y} = 48.5583 + 1.6154X_1 + 0.6921X_2 + 0.0588X_3$$
$$+0.0150X_4$$

- The error variance, σ^2, is then estimated as MSE below.

```
MTB > multiply m2 m7 m8
MTB > subtract m8 m1 m9
MTB > transpose m9 m10
MTB > multiply m10 m9 m11

Answer = 28.6406   # SSE
```

This is $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}$.

```
MTB > let k1 = 8-(4+1)   # k1 = n-(k+1)
MTB > let k2 = 1/k1
MTB > multiply k2 m11 m12

Answer = 9.5469    # MSE
```

- The regression coefficient values estimated above are point estimates.
- To make confidence intervals, we need to first estimate their standard errors.
- In matrix form, the standard errors are the square-roots of the diagonal elements of the **covariance matrix** of the regression coefficients.

- The covariance matrix of the regression coefficients contains the variances of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ on its diagonal, and related statistics call covariances off of the diagonals.

$$Cov(\hat{\beta}) = \sigma^2[\mathbf{X}'\mathbf{X}]^{-1}$$

The diagonal elements of this covariance matrix, beginning from the (1,1) position, and proceeding to the $(k+1, k+1)$ position, are, respectively, $V(\hat{\beta}_0)$, $V(\hat{\beta}_1)$, ..., $V(\hat{\beta}_k)$.

The covariance matrix is estimated by replacing σ^2 in the previous expression by MSE. The square roots of the diagonal elements in this estimated covariance matrix are the standard errors of the associated elements of $\hat{\beta}$.

```
MTB > multiply m12 m5 m13
MTB > print m13

 Matrix M13     # Cov(b)

17948.9 -216.689 -181.543 -200.351 -177.324
 -216.7    2.782    2.161    2.504    2.122
 -181.5    2.161    1.844    2.005    1.797
 -200.4    2.504    2.005    2.315    1.965
 -177.3    2.122    1.797    1.965    1.756

MTB > let k3 = sqrt(17948.9)
```

```
MTB > let k4 = sqrt(2.782)
MTB > let k5 = sqrt(1.844)
MTB > let k6 = sqrt(2.315)
MTB > let k7 = sqrt(1.756)
MTB > print k3 - k7

K3    133.974  # SEbo
K4    1.66793  # SEb1
K5    1.35794  # SEb2
K6    1.52151  # SEb3
K7    1.32514  # SEb4
```

- To test the hypothesis

  H_0 : variable X_j has no effect, **given that all all other predictor variables are in the model**

  H_A : variable X_j has an effect, **given that all all other predictor variables are in the model**

  alternatively

  H_0 : variable X_j provides no significant additional reduction in SSE **given that all all other predictor variables are in the model**

  H_A : variable X_j provides a significant additional reduction in SSE, **given that all all other predictor variables are in the model**

  – Calculate the test statistic $t_{obs} = \hat{\beta}_j / s.e.(\hat{\beta}_j)$

  – p-value $= 2P(t_{n-1-k} > t_{obs})$

- Confidence intervals for the β_i are constructed using the standard errors computed above, as

$$\hat{\beta}_i \pm t_{\alpha/2,n-(k+1)} s.e(\hat{\beta}_i)$$

  (The degrees of freedom for t is the error degrees of freedom in the ANOVA table, or $n-1-k$)

- Example: 95% confidence intervals for regression coefficients are computed using $t_{.025,3} = 3.182$.

```
MTB > print m7

 Matrix M7   # m7 = b

48.5583
 1.6154
 0.6921
 0.0588
 0.0150

MTB > let k10 = 48.5583 + 3.182*k3
MTB > let k11 = 48.5583 - 3.182*k3
MTB > print k10 k11

K10    474.862      # 95% CI for Beta 0
K11    -377.745

MTB > let k12 = 1.6154 + 3.182*k4
MTB > let k13 = 1.6154 - 3.182*k4
```

```
MTB > print k12 k13

K12    6.92276     # 95% CI for Beta 1
K13    -3.69196

MTB > let k14 = 0.6921 + 3.182*k5
MTB > let k15 = 0.6921 - 3.182*k5
MTB > print k14 k15

K14    5.01306     # 95% CI for Beta 2
K15    -3.62886

MTB > let k16 = 0.0588 + 3.182*k6
MTB > let k17 = 0.0588 - 3.182*k6
MTB > print k16 k17

K16    4.90025    # 95% CI for Beta 3
K17    -4.78265

MTB > let k18 = 0.0150 + 3.182*k7
MTB > let k19 = 0.0150 - 3.182*k7
MTB > print k18 k19

K18    4.23160    # 95% CI for Beta 4
K19    -4.20160
```

The various calculations are tedious in minitab, which isn't well set up for matrix arithmetic. Here are the commands for R or Splus. (The course Stat2050 teaches selected aspects of the R and/or Splus programming languages. These are the programs of choice for statisticians in North America and Europe, and the programs have superb graphics capabilities.) You can download the freeware program R at *cran.r-project.org*.

```
X<-matrix(c(
1,7,26,6,60,
1,1,29,15,52,
1,11,56,8,20,
1,11,31,8,47,
1,7,52,6,33,
1,11,55,9,22,
1,3,71,17,6,
1,1,31,22,44),byrow=T,ncol=5)

y<-c(78.5,74.3,104.3,87.6,95.9,109.2,102.7,72.5)

XpX<-t(X)%*%X    # contains Xprime X

XpXi<-solve(XpX) #(Xprime  X) inverse
Xpy<-t(X)%*%y    #Xprime y

bhat<-XpXi%*%Xpy #least squares estimates
```

```
yhat<-X%*%bhat   #predicted values

ymyhat<-y-yhat   #residuals
SSE<-sum(ymyhat^2) #error sum of squares
MSE<-SSE/(8-1-4) #error mean square

COV<-MSE*XpXi    #covariance matrix of beta hat
sqrt(diag(COV))  #standard errors of beta hats
```

cutting and pasting into R or Splus leads to:

```
> XpX<-t(X)%*%X
> XpX
     [,1] [,2]  [,3] [,4]  [,5]
[1,]    8   52   351   91   284
[2,]   52  472  2381  447  1744
[3,]  351 2381 17345 3983 10361
[4,]   91  447  3983 1279  3142
[5,]  284 1744 10361 3142 12458
>
> XpXi<-solve(XpX)
> XpXi
     [,1]      [,2]     [,3]     [,4]     [,5]
[1,] 1880.07 -22.697 -19.015 -20.986 -18.574
[2,]  -22.69   0.291   0.226   0.262   0.222
[3,]  -19.01   0.226   0.193   0.210   0.188
[4,]  -20.98   0.262   0.210   0.242   0.205
[5,]  -18.57   0.222   0.188   0.205   0.183
```

```
> Xpy<-t(X)%*%y
>
> bhat<-XpXi%*%Xpy
> bhat
           [,1]
[1,] 48.5583488
[2,]  1.6153626
[3,]  0.6920995
[4,]  0.0587741
[5,]  0.0149964
>
> yhat<-X%*%bhat

> ymyhat<-y-yhat
> SSE<-sum(ymyhat^2)
> MSE<-SSE/(8-1-4)
> MSE
[1] 9.546873
>
> COV<-MSE*XpXi
> sqrt(diag(COV))
[1] 133.973  1.667  1.357  1.521  1.325
```

Confidence intervals for the mean, and prediction intervals

- Where $\mathbf{x}^* = (1, x_1^*, x_2^*, \ldots, x_k^*)'$, a $100(1-\alpha)\%$ **confidence interval for the mean of Y** at $\mathbf{x}^*$ is given by

$$\mathbf{x}^{*'}\hat{\beta} \pm t_{\alpha/2, n-k-1}\sqrt{MSE}\sqrt{\mathbf{x}^{*'}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{x}^*}$$

- A $100(1-\alpha)\%$ **prediction interval for Y** when $\mathbf{x} = \mathbf{x}^*$ is

$$\mathbf{x}^{*'}\hat{\beta} \pm t_{\alpha/2, n-k-1}\sqrt{MSE}\sqrt{1 + \mathbf{x}^{*'}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{x}^*}$$

Example:

In the four variable multiple regression, suppose that we wish to find a 95% confidence interval for the mean, and a 95% prediction interval for y, when $x^* = (1, 8, 30, 10, 40)$.

The syntax of the appropriate minitab "regress" call is:

```
MTB > regress c20 4 c21-c24;
SUBC> pred 8 30 10 40.

...

Predicted Values for New Observations

New
Obs    Fit  SE Fit       95% CI            95% PI
  1  83.43   12.59  (43.36, 123.51)  (42.17, 124.69)XX
```

In R or Splus the appropriate commands, with output, are:

```
 #95% CI for mean at x=xstar
 xstar<-c(1,8,30,10,40)
 ypred<-sum(xstar*bhat)
 tcrit<-qt(.975,3)
 pm<-tcrit*sqrt(MSE)*sqrt(xstar%*%(XpXi%*%xstar))
 ucl<-ypred+pm
 lcl<-ypred-pm
 lcl
#        [,1]
#[1,] 43.3576
 ypred
#[1] 83.43183
 ucl
#         [,1]
#[1,] 123.5061

 #95% prediction interval for y at x=xstar
 pm<-tcrit*sqrt(MSE)*sqrt(1+ xstar%*%(XpXi%*%xstar))
 ucl<-ypred+pm
 lcl<-ypred-pm
 lcl
#         [,1]
#[1,] 42.16884
 ypred
%[1] 83.43183
 ucl
%         [,1]
%[1,] 124.6948
```

Partial F test for comparing nested models:

Model 1 - k independent variables

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$$

Model 2 - $m < k$ independent variables

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{im} + \epsilon_i$$

is obtained from model 1, by setting

$$\beta_{m+1} = \beta_{m+2} = \ldots = \beta_k = 0$$

That is, $k - m$ of the model 1 parameters are set to 0 in model 2.

We say that model 2 is **nested** within the **full model**, model 1.

Model	SSE	MSE	Regression df	Error df
1	SSE_1	MSE_1	k	n-k-1
2	SSE_2	MSE_2	m	n-m-1

To test the hypothesis $H_0 : \beta_{m+1} = \beta_{m+2} = \ldots = \beta_k = 0$ against the alternative that at least one of $\beta_{m+1}, \beta_{m+2}, \ldots, \beta_k$ is non-zero, calculate the test statistic:

$$F_{obs} = \frac{(SSE_2 - SSE_1)/(k - m)}{MSE_1}$$

– The p-value for the test is $P(F_{k-m,n-k-1} > F_{obs})$.

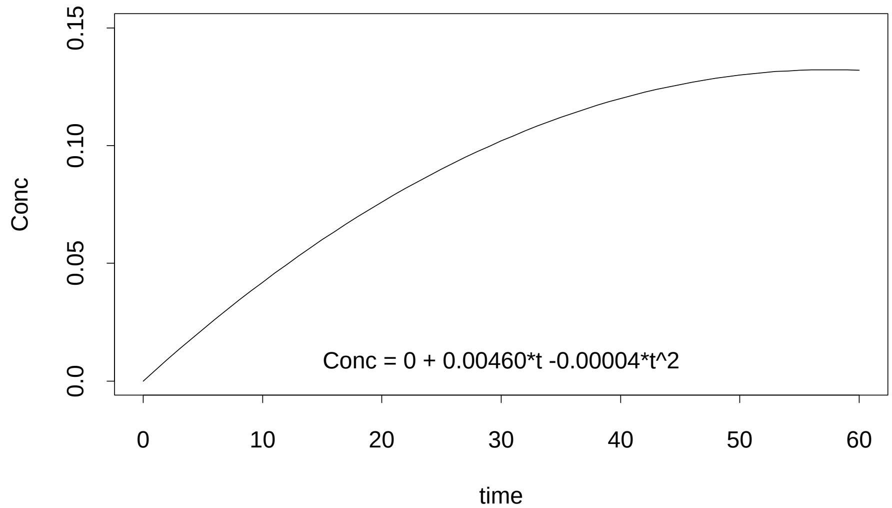
– The numerator degrees of freedom for F is the number of parameters set to 0 under H_0.

– The denominator degrees of freedom for F is the error degrees of freedom for the “bigger” model.

Types of (Linear) Regression Models

Curve

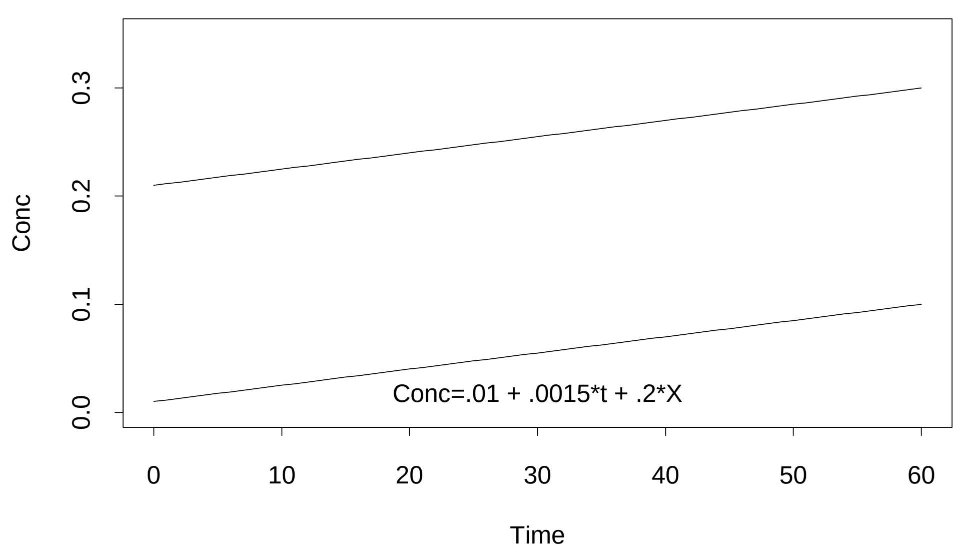
- $Conc = \beta_0 + \beta_1 t + \beta_2 t^2$



One continuous, one binary predictor

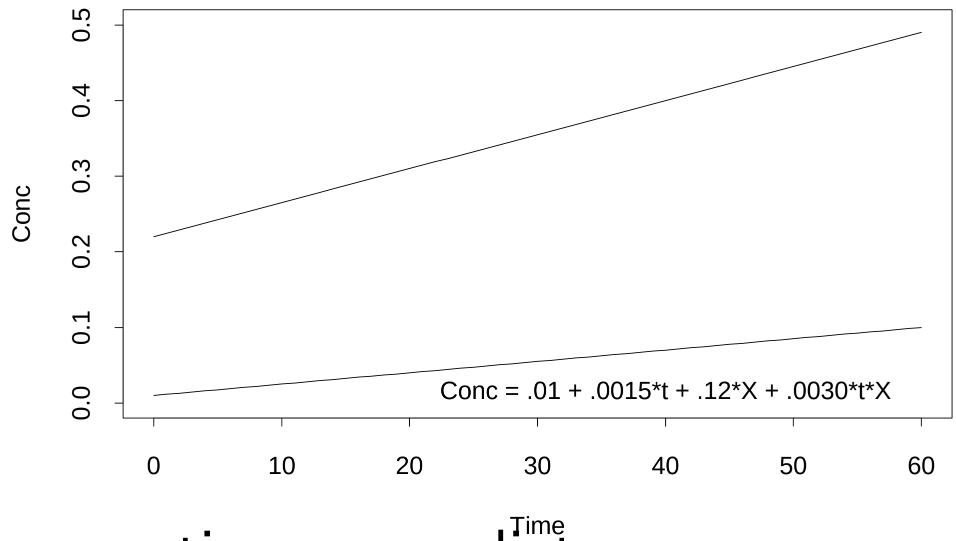
Two parallel lines

- $Conc = \beta_0 + \beta_1 time + \beta_2 X$, where X $=$ 0 for Males, 1 for Females



Two nonparallel lines

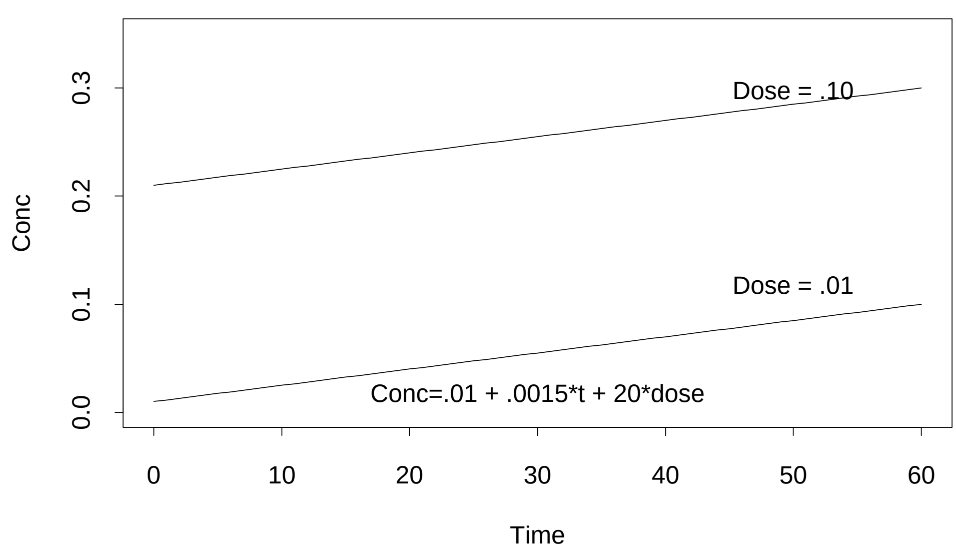
- $Conc = \beta_0 + \beta_1 time + \beta_2 X + \beta_3 time * X$, where X = 0 for Males, 1 for Females



Two continuous predictors

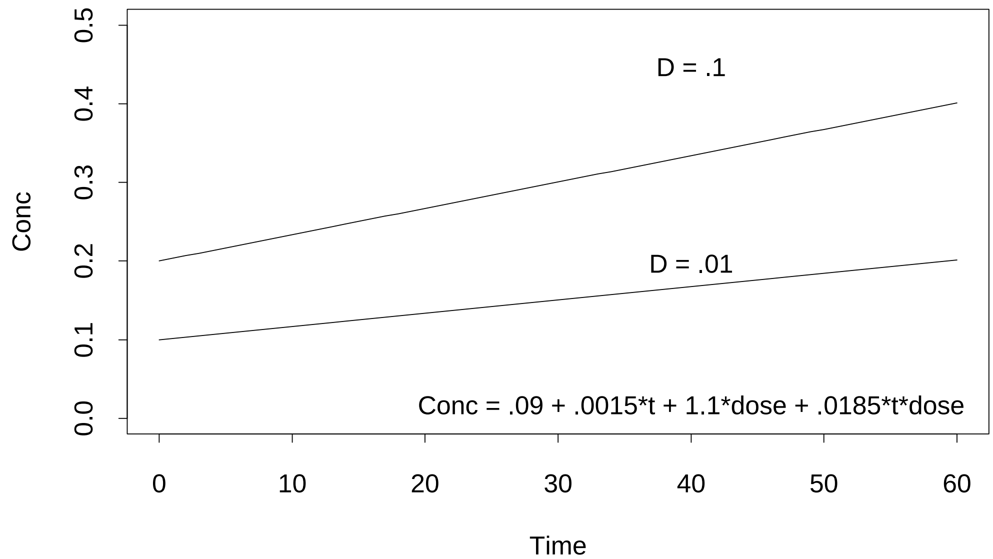
First order

- $Conc = \beta_0 + \beta_1 time + \beta_2 Dose$
- effect of dose constant over time



Interaction

- $Conc = \beta_0 + \beta_1 time + \beta_2 Dose + \beta_3 * time * dose$
- effect of dose changes with time



One-Way Analysis of Variance using Regression

We can carry out any Analysis of Variance using multiple regression, by suitable definition of the independent variables.

Recall Example: A group of 32 rats were randomly assigned to each of 4 diets labelled (A,B,C,and D). The response is the liver weight as a percentage of body weight. Two rats escaped and another died, resulting in the following data

A	B	C	D
3.42	3.17	3.34	3.65
3.96	3.63	3.72	3.93
3.87	3.38	3.81	3.77
4.19	3.47	3.66	4.18
3.58	3.39	3.55	4.21
3.76	3.41	3.51	3.88
3.84	3.55		3.96
	3.44		3.91

Anova Model:

$$X_{ij} = \mu_i + e_{ij}$$

Assumptions e_{ij} are independent $N(0, \sigma^2)$, $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, n_i$

Indicator Variables

Define $a-1$ indicator variables to identify the a groups

Index the subjects using a single index i, $i = 1, \ldots, n$

Let

$X_{i1} = 1$ if subject i is in group 1, and 0 otherwise.

$X_{i2} = 1$ if subject i is in group 2, and 0 otherwise.

$X_{i3} = 1$ if subject i is in group 3, and 0 otherwise.

...

$X_{i,a-1} = 1$ if subject i is in group $a-1$, and 0 otherwise.

Regression Model :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{a-1} X_{i,a-1} + \epsilon_i$$

Table of Means

Group	ANOVA mean	Regression mean
1	μ_1	$\beta_0 + \beta_1$
2	μ_2	$\beta_0 + \beta_2$
	...	
$a-1$	μ_{a-1}	$\beta_0 + \beta_{a-1}$
a	μ_a	β_0

Equivalent ANOVA and regression hypotheses.

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_a$$

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{a-1}$$

```
MTB > oneway c1 c2
One-way ANOVA: C1 versus C2

Source  DF      SS      MS      F      P
C2       3  1.1649  0.3883  10.84  0.000
Error   25  0.8954  0.0358
Total   28  2.0603

S = 0.1893   R-Sq = 56.54%   R-Sq(adj) = 51.32%

MTB > indicator c2 c3-c6
MTB > print c2 c3-c6
Data Display

Row  C2  C3  C4  C5  C6
  1   1   1   0   0   0
  2   1   1   0   0   0
  3   1   1   0   0   0
...
  7   1   1   0   0   0
  8   2   0   1   0   0
  9   2   0   1   0   0
 10   2   0   1   0   0
...
 18   3   0   0   1   0
 19   3   0   0   1   0
 20   3   0   0   1   0
...
 28   4   0   0   0   1
```

```
 29   4   0   0   0   1

MTB > regress c1 3 c3-c5

Regression Analysis: C1 versus C3, C4, C5

The regression equation is
C1 = 3.94 - 0.133 C3 - 0.506 C4 - 0.338 C5

Predictor        Coef  SE Coef      T      P
Constant      3.93625  0.06691  58.83  0.000
C3           -0.13339  0.09795  -1.36  0.185
C4           -0.50625  0.09463  -5.35  0.000
C5            -0.3379   0.1022  -3.31  0.003

S = 0.189253   R-Sq = 56.5%   R-Sq(adj) = 51.3%

Analysis of Variance

Source          DF       SS       MS      F      P
Regression       3  1.16490  0.38830  10.84  0.000
Residual Error  25  0.89541  0.03582
Total           28  2.06032
```