Multiple Regression Examples

Example: Tree data.

- we have seen that a simple linear regression of usable volume on diameter at chest height is not suitable, but that a quadratic model $y = \beta_0 + \beta_1 x + \beta_2 x^2$ explains the curvature
- simple fit:

```
MTB > regress c2 1 c1;
SUBC> residuals c3.
The regression equation is
volume = - 191 + 11.0 diameter
Predictor       Coef       Stdev    t-ratio         p
Constant     -191.12       16.98     -11.25     0.000
diameter     11.0413      0.5752      19.19     0.000

s = 20.33       R-sq = 95.3%     R-sq(adj) = 95.1%

Analysis of Variance

SOURCE       DF          SS          MS          F
Regression    1      152259      152259     368.43     0
Error        18        7439         413
Total        19      159698
```

- quadratic fit

```
MTB > let c3=c1**2
MTB > name c3 'diameter^2'
MTB > Regress c2 2 c1 c3;
SUBC>   Constant;
SUBC>   Predict 30 900;
SUBC>   Brief 2.

Regression Analysis: volume versus diameter,
diameter^2

The regression equation is
volume = 29.7 - 5.62 diameter + 0.290 diameter^2

Predictor      Coef  SE Coef      T      P
Constant      29.74    51.39   0.58  0.570
diameter     -5.620    3.792  -1.48  0.157
diameter^2  0.29037  0.06572   4.42  0.000

S = 14.2715   R-Sq = 97.8%   R-Sq(adj) = 97.6%

Analysis of Variance

Source           DF      SS     MS       F      P
Regression        2  156235  78118  383.54  0.000
Residual Error   17    3463    204
Total            19  159698
```

```
Source       DF  Seq SS
diameter      1  152259
diameter^2    1    3976

Predicted Values for New Observations

New
Obs     Fit  SE Fit        95% CI              95% PI
  1  122.46    5.15  (111.59, 133.33) (90.45, 154.47)

Values of Predictors for New Observations

New
Obs  diameter  diameter^2
  1      30.0         900
```

- note that the estimates of β_0 and β_1 change when the quadratic term is included
- the DF for regression has increased to two, and the DF for residual has decreased by one
- the residual sum of squares for the quadratic fit is much smaller, 3463 versus 7439
- the $MSE = s^2$ in also much smaller, 204 versus 413, (sometimes SSE decreases but

MSE does not, because the decrease in the numerator does not offset the decrease in the denominator)

- confidence intervals will be narrower, inferences will be more precise
- SST is the same, it depends on the responses only, not on the model
- R^2 is larger, .978 versus .953
- the quadratic term is highly significant ($T = 4.42$, $P = .000$)
- the SEQ SS table shows the breakdown of SSR into $SSR(diameter) = 152259$, the same as before, and $SSR(diameter^2|diameter) = 3976$, the amount explained by the quadratic term, given that the linear term is already in the model
- the increase in SSR between the two models is the same as the drop in SSE
- formerly the prediction interval for a new response at $diameter = 30$ was (96.31, 183.92), now it is (90.45, 154.47)

- this is narrower, but also shifted because the prediction has changed from 140.11 to 122.46

Example: The data set **height.mtw** can be found on BLS, in the Minitab 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).

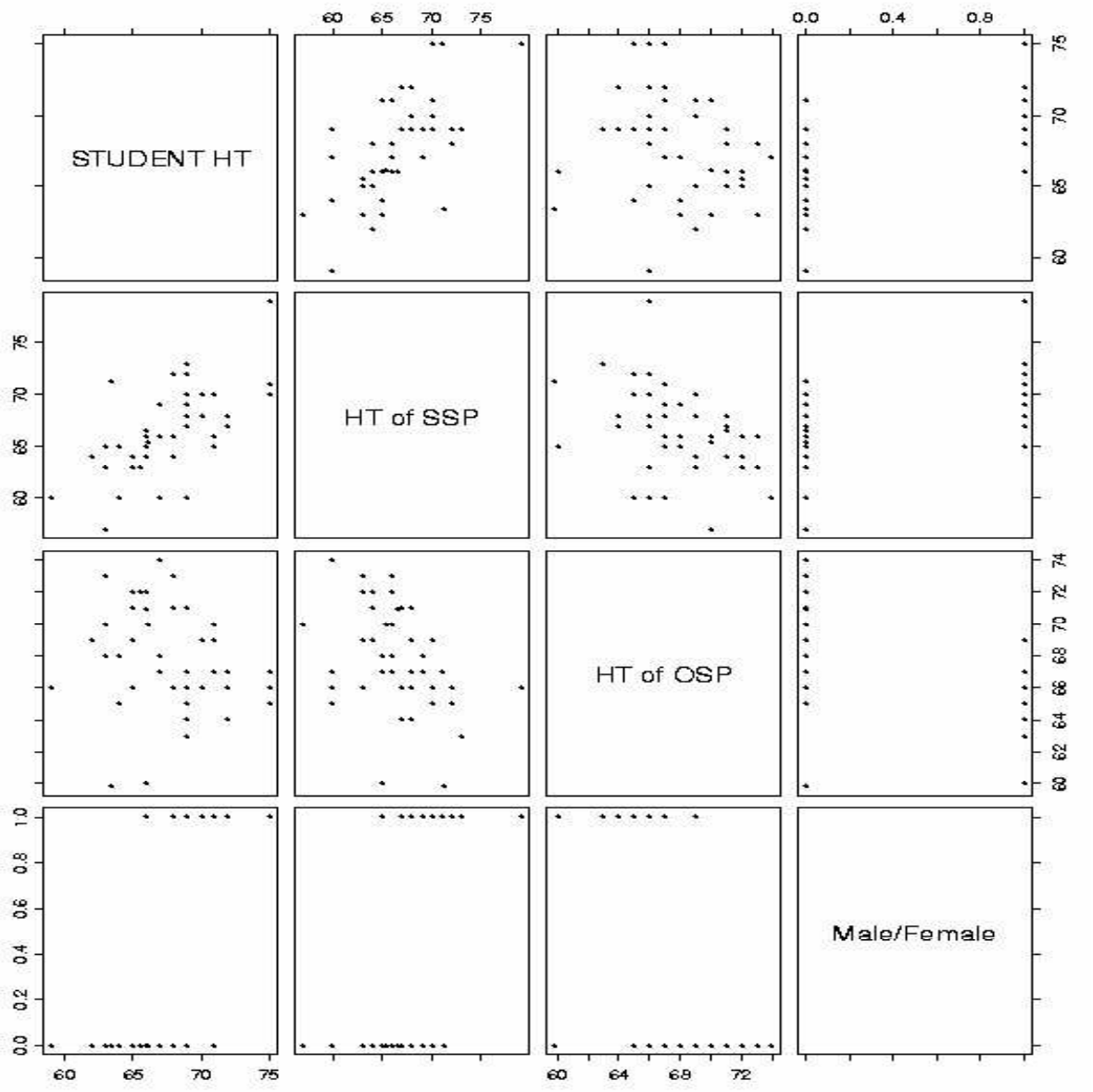
- The following Minitab output gives the various Pearson correlations for different pairs.

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

                  Student   Height SSP    Height OSP
Height SSP         0.641
                   0.000

Height OSP        -0.193       -0.408
                   0.205        0.005

M/F                0.692        0.606       -0.539
                   0.000        0.000        0.000
Cell Contents: Pearson correlation
               P-Value
```

- the height of student is positively correlated with the height of the same sex

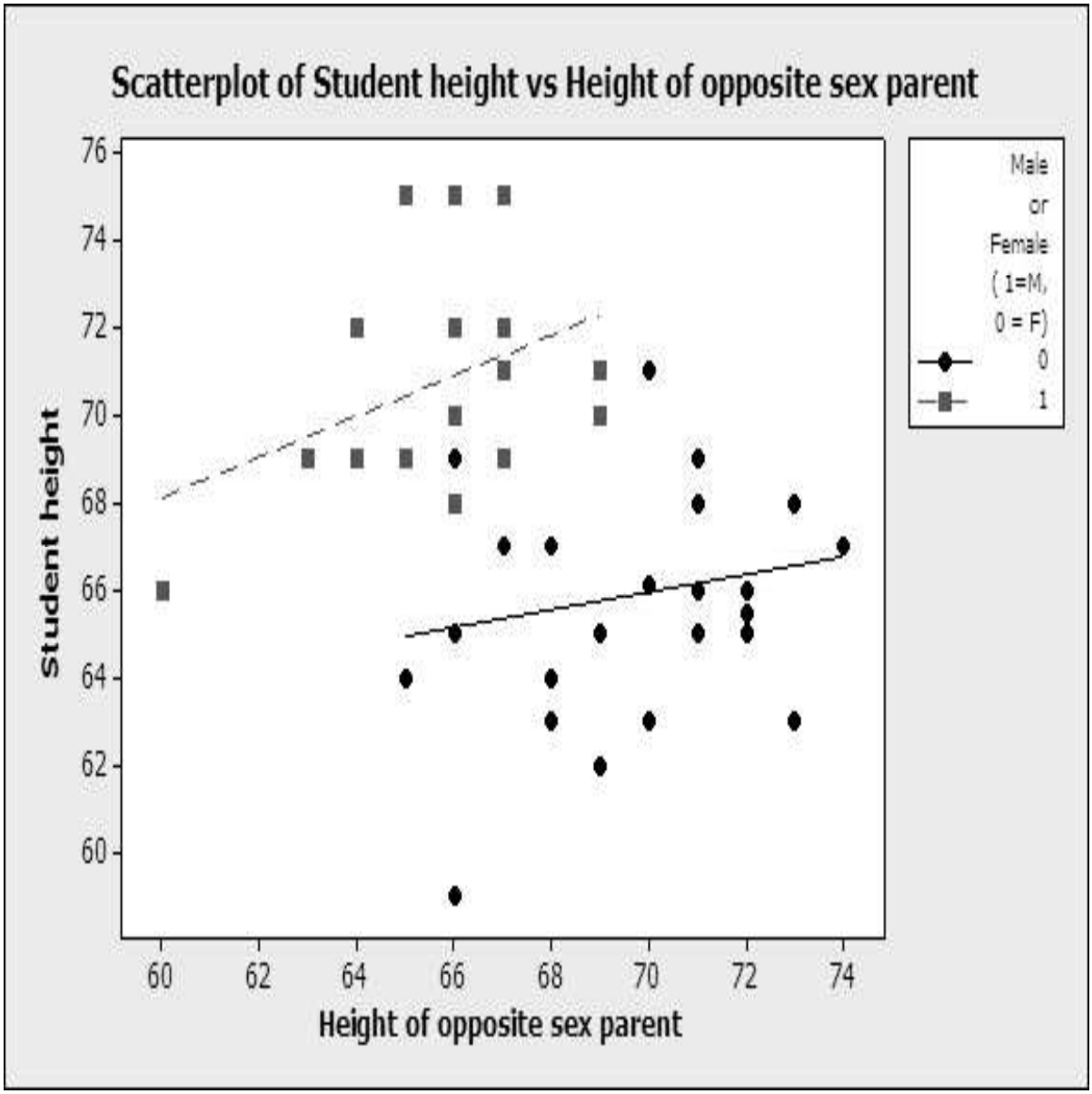
parent, and slightly negatively correlated with the height of the opposite sex parent

- the positive correlation with sex indicates that the males have greater heights than the females
- a 'pairs plot' gives all pairwise scatterplots - the first row shows how student height relates to the predictors

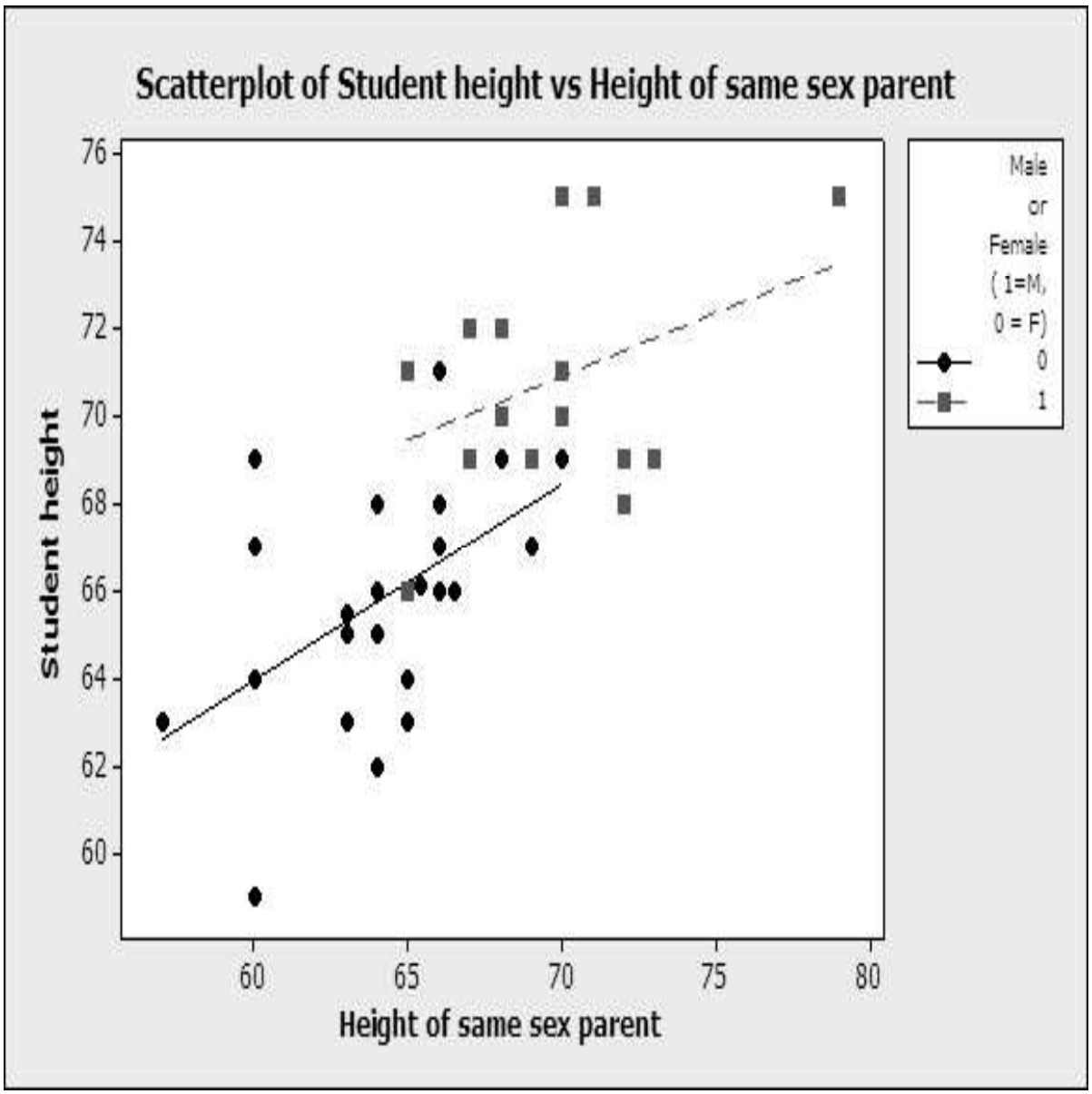


- the negative association of student height with height of opposite sex parent is surprising
- using separate labels for males and females, and superimposing separate regression lines shows positive associations for both males

and females!!!


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

- the similar plot of student heights versus height of same sex parent shows nearly equal association for both males and females



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

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 SSP

Predictor       Coef  SE Coef     T      P
Constant     30.702    6.765  4.54  0.000
SSP          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
```

- only 41% of the variation in student height

is explained by the parent of the same height

- the estimate of σ is $s = 2.74$, so prediction intervals are very roughly ± 5.5 inches - very wide!

Student height versus height of opposite sex parent

```
MTB > regress c1 1 c3

The regression equation is
Student height = 81.9 - 0.209 OSP

Predictor            Coef  SE Coef       T      P
Constant            81.89    11.06    7.40  0.000
OSP                -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
```

Student height versus height of same sex parent and 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
SSP           0.5859   0.1122  5.22  0.000
OSP           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
```

- note that the extra amount explained by the opposite sex parent is small $SSR(OSP|SSP) = 228.19 - 225.17 = 3.02$ - the T value for this variable is small and not significant

Student height versus heights of both parents taking into account their sex

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

```
MTB > regress c1 3 c2 c3 c4

Predictor       Coef  SE Coef     T      P
Constant       21.50    11.70  1.84  0.073
SSP           0.3375   0.1061  3.18  0.003
OSP           0.3249   0.1254  2.59  0.013
M/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

Source                            DF  Seq SS
Height of same sex parent          1  225.07
Height of opposite sex parent      1    3.11
Male or Female ( 1=M, 0 = F)       1  110.81
```

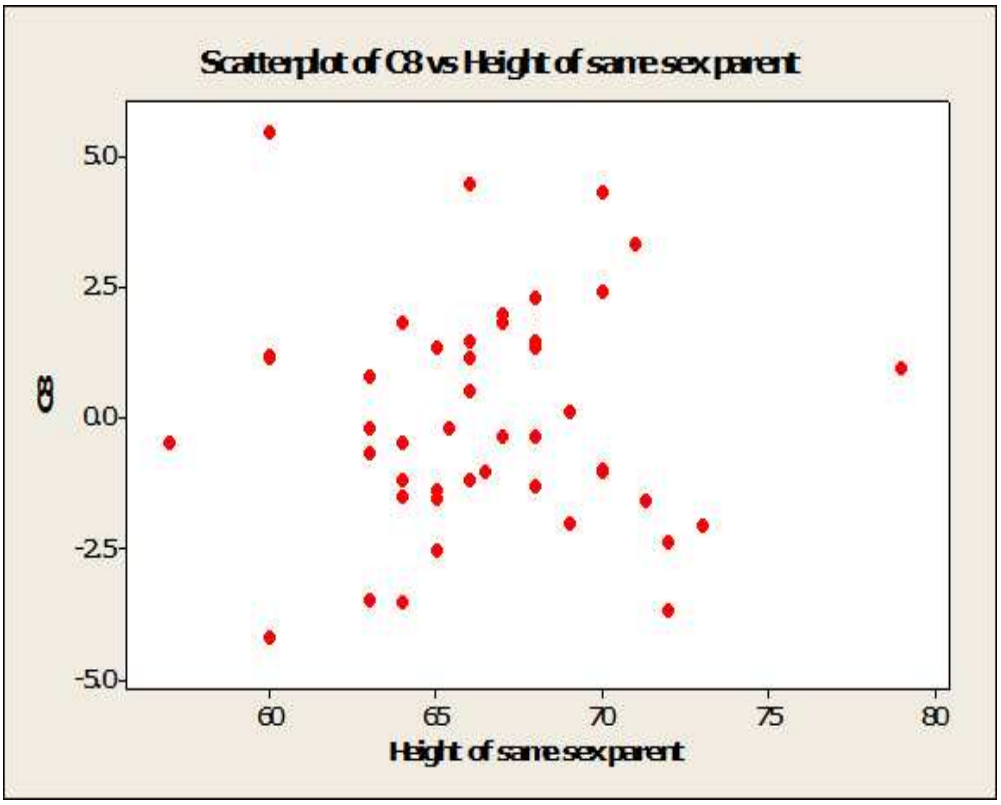
```
Unusual Observations

     Height of
      same sex  Student
Obs     parent   height     Fit  SE Fit  Residual  St Resid
 11       66.0   71.000  66.524   0.468     4.476      2.03R
 28       60.0   69.000  63.524   0.725     5.476      2.56R
 44       71.3   63.400  64.999   1.381    -1.599     -0.89 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

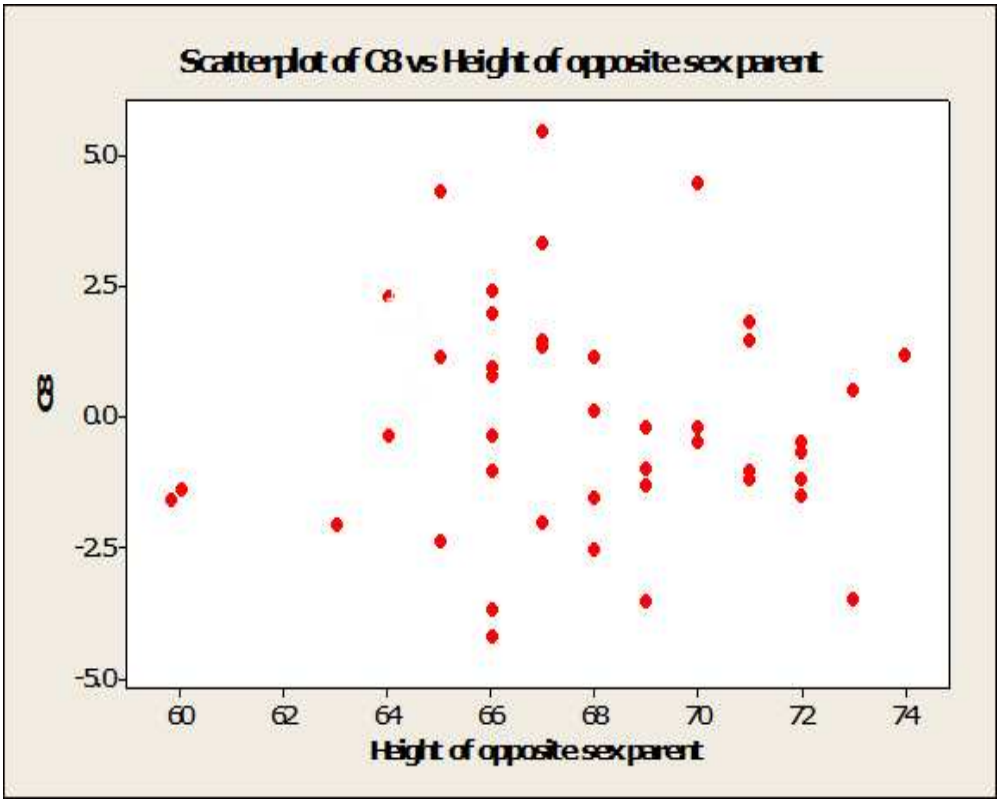
- the student's sex greatly improves the fit
- the R^2 has increased to 61.9
- all variables are significant
- s has dropped to $s = 2.26$ - prediction intervals are still wide
- two outliers are large (St. Resid denotes a standardized residual, so most should be between -2 and 2
- one case is has the potential to influence the fit, because it is an outlier in the space of the predictors
- investigation reveals this is a female student with mother 71.3 inches tall and father 59.8 inches short!
- it is quite possible that the sex of the parents was recorded incorrectly

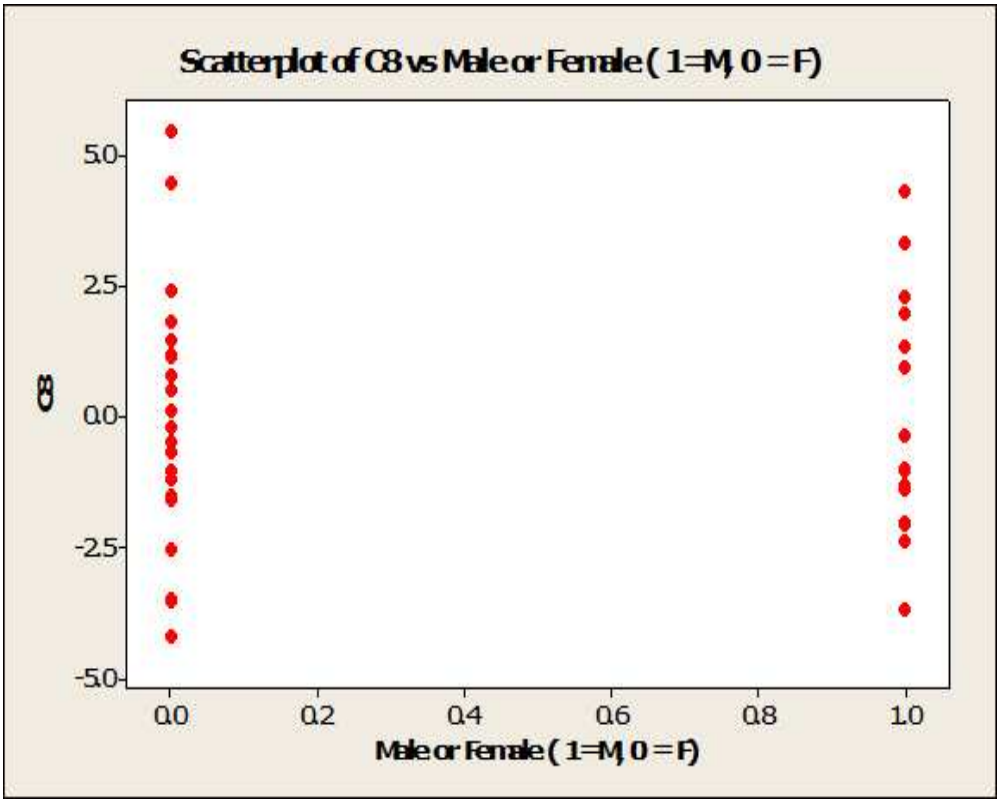
- note how the sequential sum of squares tables changes if the order in which the variables is entered into the model is changed

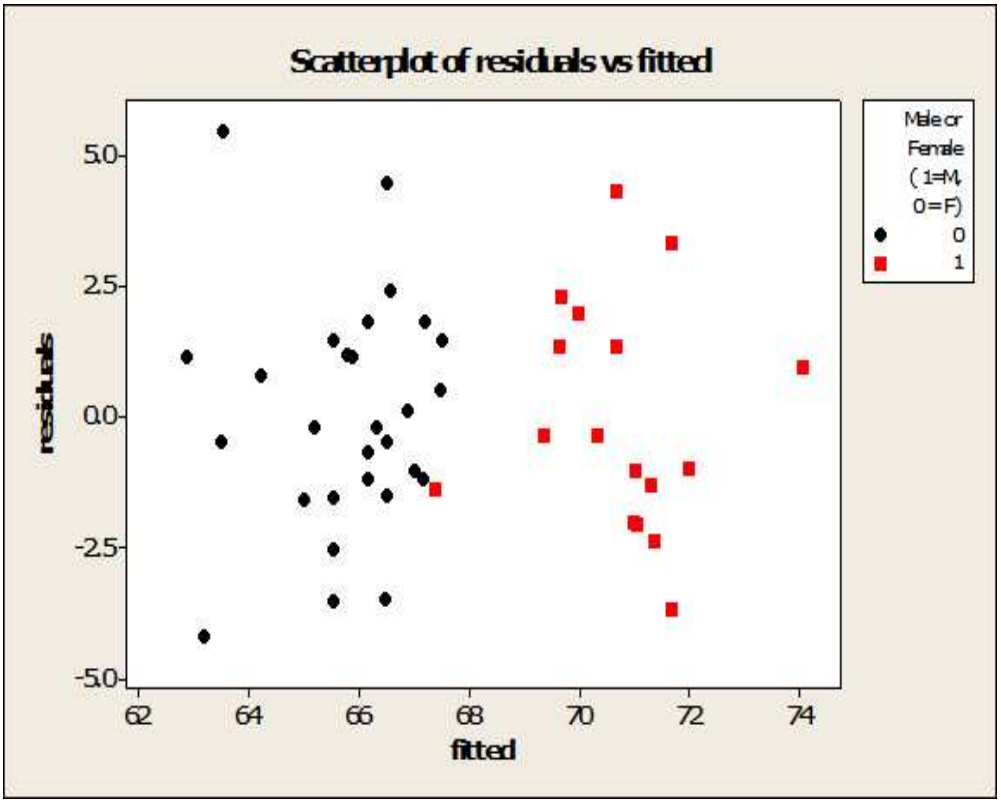
```
Source                                DF  Seq SS
Male or Female ( 1=M, 0 = F)           1  262.23
Height of same sex parent              1   42.54
Height of opposite sex parent          1   34.23

Source                                DF  Seq SS
Male or Female ( 1=M, 0 = F)           1  262.23
Height of opposite sex parent          1   25.18
Height of same sex parent              1   51.59

Source                                DF  Seq SS
Height of opposite sex parent          1   20.34
Male or Female ( 1=M, 0 = F)           1  267.07
Height of same sex parent              1   51.59
```

- plots of the residuals versus each of the predictors show no problems




Scatterplot of C8 vs Height of opposite sex parent
C8
5.0
2.5
0.0
-2.5
-5.0
60
62
64
66
68
70
72
74
Height of opposite sex parent


Scatterplot of C8 vs Male or Female (1=M, 0 = F)
C8
5.0
2.5
0.0
-2.5
-5.0
0.0
0.2
0.4
0.6
0.8
1.0
Male or Female (1=M, 0 = F)



- in this plot the males, with the larger fitted values, are indicated with squares, the females are circles

- the coefficients in the model for SSP and OSP are nearly the same
- this suggests that one can add or average the heights of the two parents
- the following model uses the average, and with one fewer predictor the fit to the data is nearly as good as before

```
 let c10 = (c2+c3)/2
MTB > regress c1 2 c4 c10

Regression Analysis: Student height versus Male or Female (, C10

The regression equation is
Student height = 21.3 + 4.50 Male or Female ( 1=M, 0 = F) + 0.665 C10


Predictor                      Coef  SE Coef     T      P
Constant                      21.31    11.33  1.88  0.067
Male or Female ( 1=M, 0 = F)  4.4971   0.6969  6.45  0.000
C10                           0.6649   0.1693  3.93  0.000


S = 2.23104   R-Sq = 61.9%   R-Sq(adj) = 60.0%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       2  338.96  169.48  34.05  0.000
Residual Error  42  209.06    4.98
Total           44  548.02


Source                        DF  Seq SS
Male or Female ( 1=M, 0 = F)   1  262.23
C10                            1   76.74
```

Unusual Observations

Obs	Male or Female (1=M, 0 = F)	Student height	Fit	SE Fit	Residual	St Resid
3	0.00	59.000	63.198	0.782	-4.198	-2.01R
11	0.00	71.000	66.522	0.461	4.478	2.05R
28	0.00	69.000	63.530	0.713	5.470	2.59R
33	1.00	66.000	67.362	1.022	-1.362	-0.69 X

- the R^2 for this model is the same as before, and the MSE and its square root s have been reduced