Example of multiple regression

main topic interpreting results session command see also

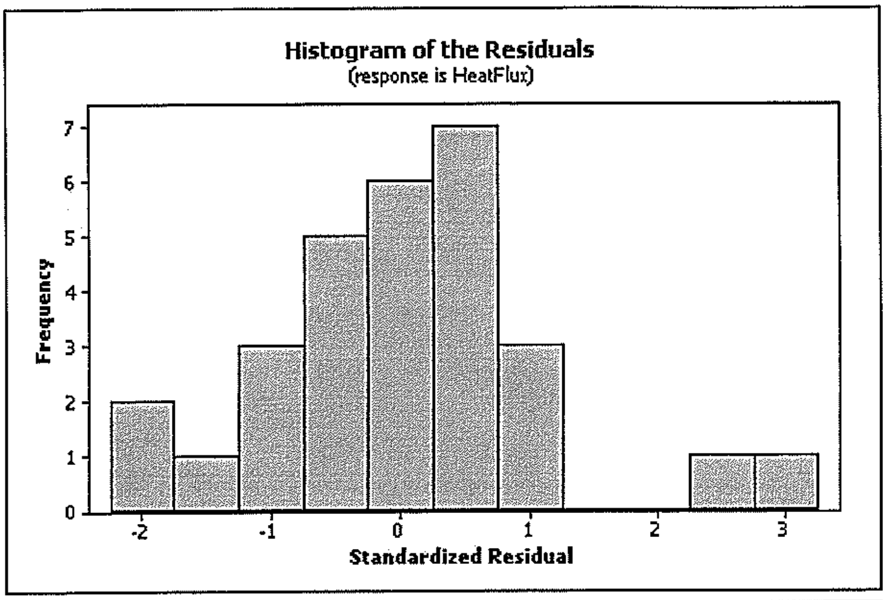
As part of a test of solar thermal energy, you measure the total heat flux from homes. You wish to examine whether total heat flux (HeatFlux) can be predicted by insulation, by the position of the focal points in the east, south, and north directions, and by the time of day. Data are from [27]. You found, using best subsets regression, that the best two-predictor model included the variables North and South and the best three-predictor added the variable East. You evaluate the three-predictor model using multiple regression.

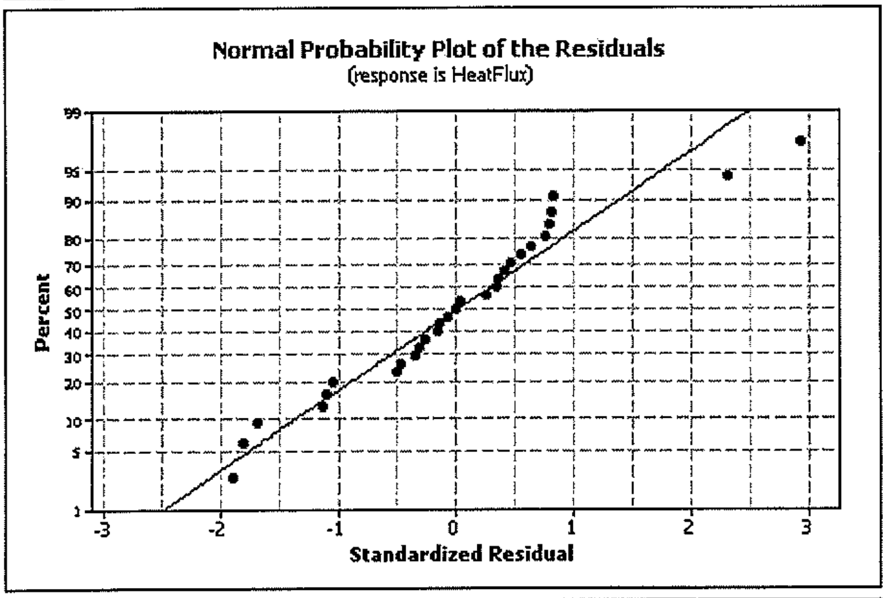
1 Open the worksheet EXH_REGR.MTW.
2 Choose **Stat > Regression > Regression.**
3 In **Response**, enter *HeatFlux.*
4 In **Predictors**, enter *East South North.*
5 Click **Graphs.**
6 Under **Residuals for Plots**, choose **Standardized.**
7 Under **Residual Plots**, choose **Individual Plots.** Check **Histogram of residuals, Normal plot of residuals**, and **Residuals versus fits.** Click **OK.**
8 Click **Options.** Under **Display**, check **PRESS** and **predicted R-square.** Click **OK** in each dialog box.

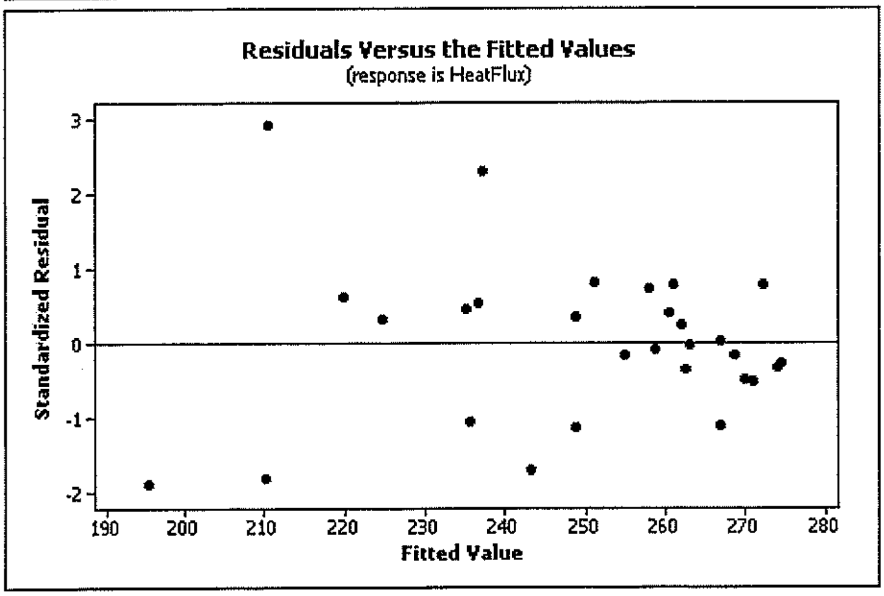
Session window output

Regression Analysis: HeatFlux versus East, South, North

```
The regression equation is
HeatFlux = 389 + 2.12 East + 5.32 South - 24.1 North


Predictor      Coef  SE Coef       T      P
Constant     389.17    66.09    5.89  0.000
East          2.125    1.214    1.75  0.092
South        5.3185   0.9629    5.52  0.000
North       -24.132    1.869  -12.92  0.000


S = 8.59782   R-Sq = 87.4%   R-Sq(adj) = 85.9%

PRESS = 3089.67   R-Sq(pred) = 78.96%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       3  12833.9  4278.0  57.87  0.000
Residual Error  25   1848.1    73.9
Total           28  14681.9


Source  DF   Seq SS
East     1    153.8
South    1    349.5
North    1  12330.6


Unusual Observations

Obs  East  HeatFlux     Fit  SE Fit  Residual  St Resid
  4  33.1    230.70  210.20    5.03     20.50     2.94R
 22  37.8    254.50  237.16    4.24     17.34     2.32R

R denotes an observation with a large standardized residual.
```







Interpreting the results

Session window output

- The p-value in the Analysis of Variance table (0.000) shows that the model estimated by the regression procedure is significant at an α-level of 0.05. This indicates that at least one coefficient is different from zero.
- The p-values for the estimated coefficients of North and South are both 0.000, indicating that they are significantly related to HeatFlux.

The p-value for East is 0.092, indicating that it is not related to HeatFlux at an α-level of 0.05. Additionally, the sequential sum of squares indicates that the predictor East doesn't explain a substantial amount of unique variance. This suggests that a model with only North and South may be more appropriate.

- The R^2 value indicates that the predictors explain 87.4% of the variance in HeatFlux. The adjusted R^2 is 85.9%, which accounts for the number of predictors in the model. Both values indicate that the model fits the data well.
- The predicted R^2 value is 78.96%. Because the predicted R^2 value is close to the R^2 and adjusted R^2 values, the model does not appear to be overfit and has adequate predictive ability.
- Observations 4 and 22 are identified as unusual because the absolute value of the standardized residuals are greater than 2. This may indicate they are outliers. See Checking your model, Identifying outliers, and Choosing a residual type.

Graph window output

- The histogram of the residuals exhibits a pattern consistent with a normal distribution. The histogram is most effective with data sets with more than 50 observations. The normal probability plot is easier to interpret for smaller samples.
- The normal probability plot shows an approximately linear pattern consistent with a normal distribution. The two points in the upper-right corner of the plot may be outliers. Brushing the graph identifies these points as 4 and 22, the same points that are labeled unusual observations in the output. See Checking your model and Identifying outliers.
- The plot of residuals versus the fitted values shows that the residuals get smaller (closer to the reference line) as the fitted values increase, which may indicate the residuals have nonconstant variance. See [9] for information on nonconstant variance.

Example of best subsets regression

main topic interpreting results session command see also

Total heat flux is measured as part of a solar thermal energy test. You wish to see how total heat flux is predicted by other variables: insulation, the position of the focal points in the east, south, and north directions, and the time of day. Data are from Montgomery and Peck [27], page 486.

1 Open the worksheet EXH_REGR.MTW.

2 Choose **Stat > Regression > Best Subsets**.

3 In **Response**, enter **Heatflux**.

4 In **Free Predictors**, enter **Insolation–Time**. Click **OK**.

Session window output

Best Subsets Regression: HeatFlux versus Insolation, East, ...

```
Response is HeatFlux

                                              I
                                              n
                                              s
                                              o
                                              l
                                              a   S N
                                              t E o o T
                                              i a u r i
                              Mallows         o s t t m
Vars  R-Sq  R-Sq(adj)     C-p       S  n t h h e
   1  72.1       71.0    38.5  12.328        X
   1  39.4       37.1   112.7  18.154  X
   2  85.9       84.8     9.1  8.9321      X X
   2  82.0       80.6    17.8  10.076        X X
   3  87.4       85.9     7.6  8.5978    X X X
   3  86.5       84.9     9.7  8.9110  X   X X
   4  89.1       87.3     5.8  8.1698  X X X X
   4  88.0       86.0     8.2  8.5550  X   X X X
   5  89.9       87.7     6.0  8.0390  X X X X X
```

Interpreting the results

Each line of the output represents a different model. Vars is the number of variables or predictors in the model. R^2 and adjusted R^2 are converted to percentages. Predictors that are present in the model are indicated by an X.

In this example, it isn't clear which model fits the data best. The model with all the variables has the highest adjusted R^2 (87.7%), a low Mallows Cp value (6.0), and the lowest S value (8.0390). The four-predictor model with all variables except Time has a lower Cp value (5.8), although S is slightly higher (8.16) and adjusted R^2 is slightly lower (87.3%). The best three-predictor model includes North, South, and East, with a slightly higher Cp value (7.6) and a lower adjusted R^2(85.9%). The best two-predictor model might be considered the minimum fit. The multiple regression example indicates that adding the variable East does not improve the fit of the model.

Before choosing a model, you should always check to see if the models violate any regression assumptions using residual plots and other diagnostic tests. See Checking your model.

Example of stepwise regression

main topic   interpreting results   session command   see also

Students in an introductory statistics course participated in a simple experiment. Each student recorded his or her height, weight, gender, smoking preference, usual activity level, and resting pulse. They all flipped coins, and those whose coins came up heads ran in place for one minute. Afterward, the entire class recorded their pulses once more. You wish to find the best predictors for the second pulse rate.

1 Open the worksheet ~~PULSE.MTW~~. EXH_REGR.MTW
2 Press [CTRL] + [M] to make the Session window active.
3 Choose **Editor > Enable Commands** so Minitab displays session commands.
4 Choose **Stat > Regression > Stepwise.**
5 In **Response**, enter ~~*Pulse2*~~. Heatflux
6 In **Predictors**, enter ~~*Pulse1 Ran Weight*~~. Insolation – Time
7 Click **Options.**
8 **In Number of steps between pauses,** type *2*. Click **OK** in each dialog box.
9 In the Session window, at the first **More**? prompt, type *Yes*.
10 In the Session window, at the first **More**? prompt, type *No*.

Session window output

03.09.2004 12:17:28

Welcome to Minitab, press F1 for help.

Stepwise Regression: HeatFlux versus Insolation; East; ...

Forward selection. Alpha-to-Enter: 0,15

Response is HeatFlux on 5 predictors, with N = 29

```
Step                1        2
Constant        607,1    483,7

North           -21,4    -24,2
T-Value         -8,34   -12,48
P-Value         0,000    0,000

South                     4,80
T-Value                   5,04
P-Value                  0,000

S                12,3     8,93
R-Sq            72,05    85,87
R-Sq(adj)       71,02    84,78
Mallows C-p      38,5      9,1
PRESS         4855,93  2736,51
R-Sq(pred)      66,93    81,36
```

More? (Yes, No, Subcommand, or Help)

SUBC> Yes

```
Step                3        4
Constant        389,2    270,2

North           -24,1    -21,1
T-Value        -12,92    -8,91
P-Value         0,000    0,000

South            5,32     5,34
T-Value          5,52     5,83
P-Value         0,000    0,000

East              2,1      3,0
T-Value          1,75     2,40
P-Value         0,092    0,025

Insolation               0,052
T-Value                   1,92
P-Value                  0,067

S                8,60     8,17
R-Sq            87,41    89,09
R-Sq(adj)       85,90    87,27
Mallows C-p       7,6      5,8
PRESS         3089,67  2847,17
R-Sq(pred)      78,96    80,61
```

More? (Yes, No, Subcommand, or Help)

SUBC> Yes

No variables entered or removed

More? (Yes, No, Subcommand, or Help)

SUBC>

SUBC>

Stepwise Regression: HeatFlux versus Insolation; East; ...

Backward elimination. Alpha-to-Remove: 0,15

Response is HeatFlux on 5 predictors, with N = 29

Step	1	2
Constant	325,4	270,2
Insolation	0,068	0,052
T-Value	2,33	1,92
P-Value	0,029	0,067
East	2,6	3,0
T-Value	2,04	2,40
P-Value	0,053	0,025
South	3,80	5,34
T-Value	2,60	5,83
P-Value	0,016	0,000
North	-22,9	-21,1
T-Value	-8,49	-8,91
P-Value	0,000	0,000
Time	2,4	
T-Value	1,34	
P-Value	0,194	
S	8,04	8,17
R-Sq	89,88	89,09
R-Sq(adj)	87,68	87,27
Mallows C-p	6,0	5,8
PRESS	3109,95	2847,17
R-Sq(pred)	78,82	80,61

MTB > print c3-c8

Data Display

Row	HeatFlux	Insolation	East	South	North	Time
1	271,8	783,35	33,53	40,55	16,66	13,20
2	264,0	748,45	36,50	36,19	16,46	14,11
3	238,8	684,45	34,66	37,31	17,66	15,68
4	230,7	827,80	33,13	32,52	17,50	10,53
5	251,6	860,45	35,75	33,71	16,40	11,00
6	257,9	875,15	34,46	34,14	16,28	11,31
7	263,9	909,45	34,60	34,85	16,06	11,96
8	266,5	905,55	35,38	35,89	15,93	12,58
9	229,1	756,00	35,85	33,53	16,60	10,66
10	239,3	769,35	35,68	33,79	16,41	10,85
11	258,0	793,50	35,35	34,72	16,17	11,41
12	257,6	801,65	35,04	35,22	15,92	11,91
13	267,3	819,65	34,07	36,50	16,04	12,85
14	267,0	808,55	32,20	37,60	16,19	13,58
15	259,6	774,95	34,32	37,89	16,62	14,21
16	240,4	711,85	31,08	37,71	17,37	15,56
17	227,2	694,85	35,73	37,00	18,12	15,83
18	196,0	638,10	34,11	36,76	18,53	16,41
19	278,7	774,55	34,79	34,62	15,54	13,10
20	272,3	757,90	35,77	35,40	15,70	13,63
21	267,4	753,35	36,44	35,96	16,45	14,51
22	254,5	704,70	37,82	36,26	17,62	15,38
23	224,7	666,80	35,07	36,34	18,12	16,10
24	181,5	568,55	35,26	35,90	19,05	16,73
25	227,5	653,10	35,56	31,84	16,51	10,58
26	253,6	704,05	35,73	33,16	16,02	11,28
27	263,0	709,60	36,46	33,83	15,89	11,91
28	265,8	726,90	36,26	34,89	15,83	12,65
29	263,8	697,15	37,20	36,27	16,71	14,06

MTB >