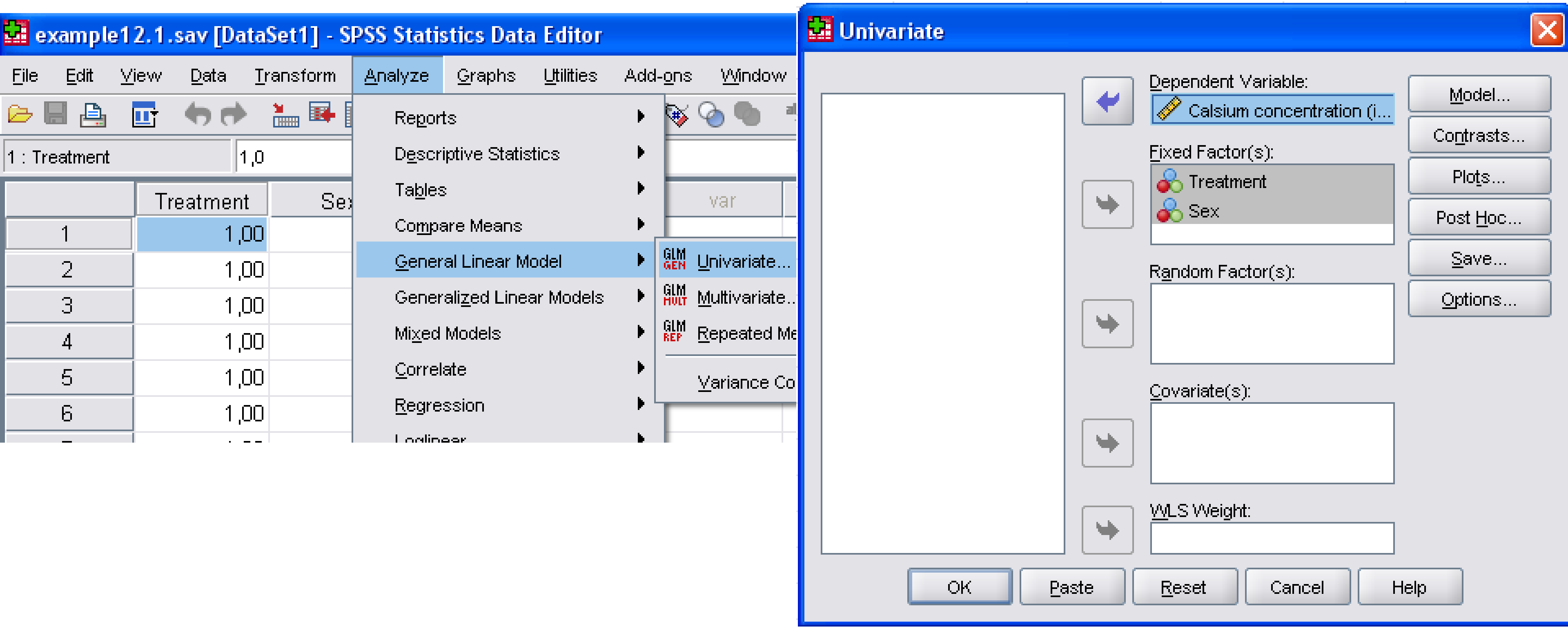
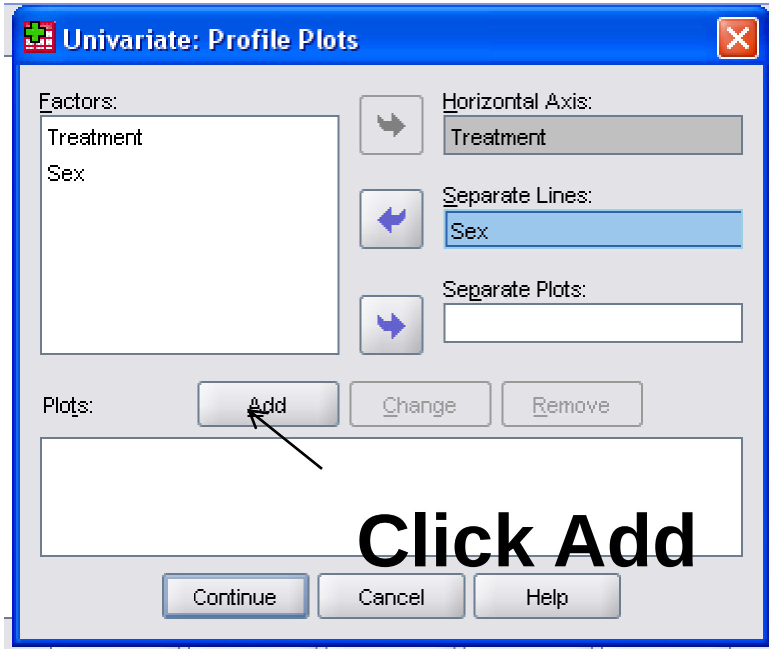
General linear models

One and Two-way ANOVA in SPSS
Repeated measures ANOVA
Multiple linear regression

2-way ANOVA in SPSS Example 14.1



2-way ANOVA in SPSS



Repeated measures

The stroop test

BLUE

Table 7.1 Field Independence and a Reverse Stroop Task

	Form Names			Color Names		
Subject	*Normal Condition*	*Congruent Condition*	*Incongruent Condition*	*Normal Condition*	*Congruent Condition*	*Incongruent Condition*
Field-independent						
1	191	206	219	176	182	196
2	175	183	186	148	156	161
3	166	165	161	138	146	150
4	206	190	212	174	178	184
5	179	187	171	182	185	210
6	183	175	171	182	185	210
7	174	168	187	167	160	178
8	185	186	185	153	159	169
9	182	189	201	173	177	183
10	191	192	208	168	169	187
11	162	163	168	135	141	145
12	162	162	170	142	147	151
Field-dependent						
13	277	267	322	205	231	255
14	235	216	271	161	183	187
15	150	150	165	140	140	156
16	400	404	379	214	223	216
17	183	165	187	140	146	163
18	162	215	184	144	156	165
19	163	179	172	170	189	192
20	163	159	159	143	150	148
21	237	233	238	207	225	228
22	205	177	217	205	208	230
23	178	190	211	144	155	177
24	164	186	187	139	151	163

Note: Response variable is time in milliseconds.

The model

$$y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl} + u_i + \varepsilon_{ijkl}$$

where y_{ijkl} represent the observation for the ith subject in the lth group (l = 1, 2), under the jth type condition (j = 1, 2), and the kth cue condition (k = 1, 2, 3), α_j, β_k, γ_l represent the main effects of type, cue, and group, $(\alpha\beta)_{jk}$, $(\alpha\gamma)_{jl}$, $(\beta\gamma)_{kl}$ the first order interactions between these factors and $(\alpha\beta\gamma)_{jkl}$ the second order interaction. These effects are known as *fixed effects*. The u_i represents the effect of subject i and ε_{ijkl} the residual

Table 7.1 Field Independence and a Reverse Stroop Task

	Form Names			*Color Names*		
Subject	*Normal Condition*	*Congruent Condition*	*Incongruent Condition*	*Normal Condition*	*Congruent Condition*	*Incongruent Condition*
Field-independent						
1	191	206	219	176	182	196
2	175	183	186	148	156	161
3	166	165	161	138	146	150
4	206	190	212	174	178	184
5	179	187	171	182	185	210
6	183	175	171	182	185	210
7	174	168	187	167	160	178
8	185	186	185	153	159	169
9	182	189	201	173	177	183
10	191	192	208	168	169	187
11	162	163	168	135	141	145
12	162	162	170	142	147	151
Field-dependent						
13	277	267	322	205	231	255
14	235	216	271	161	183	187
15	150	150	165	140	140	156
16	400	404	379	214	223	216
17	183	165	187	140	146	163
18	162	215	184	144	156	165
19	163	179	172	170	189	192
20	163	159	159	143	150	148
21	237	233	238	207	225	228
22	205	177	217	205	208	230
23	178	190	211	144	155	177
24	164	186	187	139	151	163

Note: Response variable is time in milliseconds.

Assumptions

1. *Normality*: the data arise from populations with normal distribution.
2. *Homogeneity of variance*: the variances of the assumed normal distributions are equal.
3. *Sphericity*: the variances of the differences between all pairs of the repeated measurements are equal. This requirement implies that the covariances between pairs of repeated measures are equal and that the variances of each repeated measurement are also equal, i.e., the covariance matrix of the repeated measures must have the so-called *compound symmetry* pattern.

Table 7.1 Field Independence and a Reverse Stroop Task

	Form Names			*Color Names*		
Subject	*Normal Condition*	*Congruent Condition*	*Incongruent Condition*	*Normal Condition*	*Congruent Condition*	*Incongruent Condition*
Field-independent						
1	191	206	219	176	182	196
2	175	183	186	148	156	161
3	166	165	161	138	146	150
4	206	190	212	174	178	184
5	179	187	171	182	185	210
6	183	175	171	182	185	210
7	174	168	187	167	160	178
8	185	186	185	153	159	169
9	182	189	201	173	177	183
10	191	192	208	168	169	187
11	162	163	168	135	141	145
12	162	162	170	142	147	151
Field-dependent						
13	277	267	322	205	231	255
14	235	216	271	161	183	187
15	150	150	165	140	140	156
16	400	404	379	214	223	216
17	183	165	187	140	146	163
18	162	215	184	144	156	165
19	163	179	172	170	189	192
20	163	159	159	143	150	148
21	237	233	238	207	225	228
22	205	177	217	205	208	230
23	178	190	211	144	155	177
24	164	186	187	139	151	163

Note: Response variable is time in milliseconds.

If the sphericity does not hold…

a) The trick: Greenhouse and Geisser
b) Repeated MANOVA
c) More complex models

Table 7.1 Field Independence and a Reverse Stroop Task

	Form Names			Color Names		
Subject	*Normal Condition*	*Congruent Condition*	*Incongruent Condition*	*Normal Condition*	*Congruent Condition*	*Incongruent Condition*
Field-independent						
1	191	206	219	176	182	196
2	175	183	186	148	156	161
3	166	165	161	138	146	150
4	206	190	212	174	178	184
5	179	187	171	182	185	210
6	183	175	171	182	185	210
7	174	168	187	167	160	178
8	185	186	185	153	159	169
9	182	189	201	173	177	183
10	191	192	208	168	169	187
11	162	163	168	135	141	145
12	162	162	170	142	147	151
Field-dependent						
13	277	267	322	205	231	255
14	235	216	271	161	183	187
15	150	150	165	140	140	156
16	400	404	379	214	223	216
17	183	165	187	140	146	163
18	162	215	184	144	156	165
19	163	179	172	170	189	192
20	163	159	159	143	150	148
21	237	233	238	207	225	228
22	205	177	217	205	208	230
23	178	190	211	144	155	177
24	164	186	187	139	151	163

Note: Response variable is time in milliseconds.

And now in SPSS

Table 7.1 Field Independence and a Reverse Stroop Task

	Form Names			*Color Names*		
Subject	*Normal Condition*	*Congruent Condition*	*Incongruent Condition*	*Normal Condition*	*Congruent Condition*	*Incongruent Condition*
Field-independent						
1	191	206	219	176	182	196
2	175	183	186	148	156	161
3	166	165	161	138	146	150
4	206	190	212	174	178	184
5	179	187	171	182	185	210
6	183	175	171	182	185	210
7	174	168	187	167	160	178
8	185	186	185	153	159	169
9	182	189	201	173	177	183
10	191	192	208	168	169	187
11	162	163	168	135	141	145
12	162	162	170	142	147	151
Field-dependent						
13	277	267	322	205	231	255
14	235	216	271	161	183	187
15	150	150	165	140	140	156
16	400	404	379	214	223	216
17	183	165	187	140	146	163
18	162	215	184	144	156	165
19	163	179	172	170	189	192
20	163	159	159	143	150	148
21	237	233	238	207	225	228
22	205	177	217	205	208	230
23	178	190	211	144	155	177
24	164	186	187	139	151	163

Note: Response variable is time in milliseconds.

Multiple linear regression

- Regression: One variable is considered dependent on the other(s)
- Correlation: No variables are considered dependent on the other(s)
- Multiple regression: More than one independent variable
- Linear regression: The independent factor is scalar and linearly dependent on the independent factor(s)
- Logistic regression: The independent factor is categorical (hopefully only two levels) and follows a s-shaped relation.

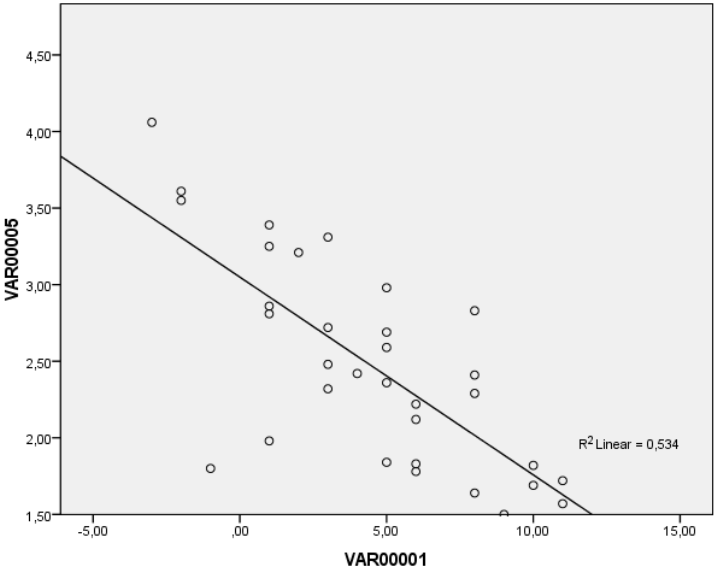
Remember the simple linear regression?

If Y is linearly dependent on X, simple linear regression is used:

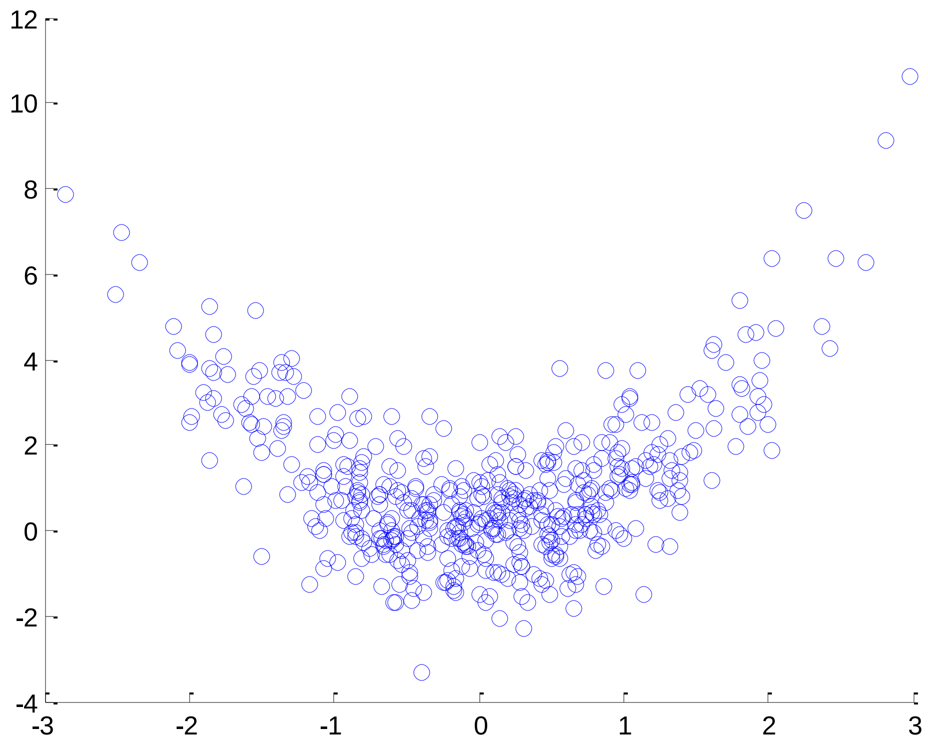
$$Y_j = \alpha + \beta X_j$$

α is the intercept, the value of Y when X = 0

β is the slope, the rate in which Y increases when X increases



I the relation linear?

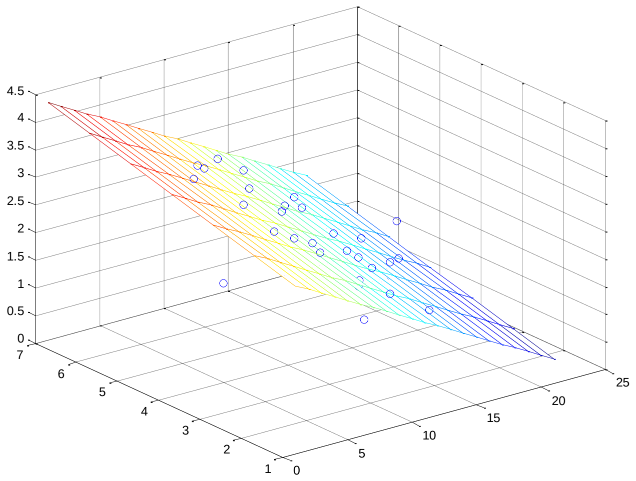


Multiple linear regression

If Y is linearly dependent on more than one independent variable:

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j}$$

α is the intercept, the value of Y when X_1 and X_2 = 0
β_1 and β_2 are termed partial regression coefficients
β_1 expresses the change of Y for one unit of X when β_2 is kept constant



Multiple linear regression – residual error and estimations

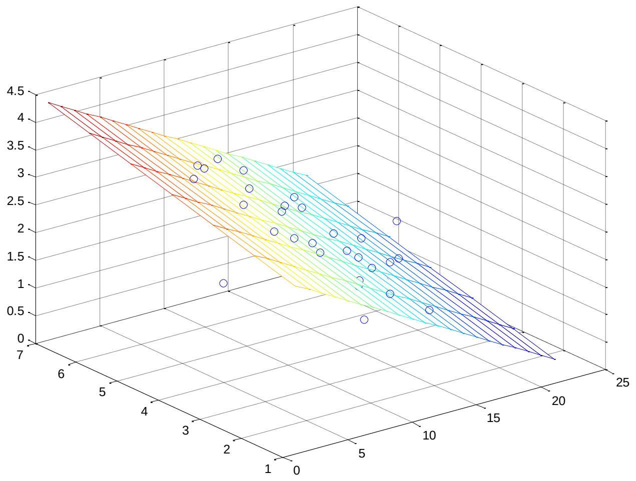
As the collected data is not expected to fall in a plane an error term must be added

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \varepsilon_j$$

The error term sums up to be zero.

Estimating the dependent factor and the population parameters:

$$\hat{Y}_j = a + b_1 X_{1j} + b_2 X_{2j}$$



Multiple linear regression – general equations

In general an finite number (m) of independent variables may be used to estimate the hyperplane

$$Y_j = \alpha + \sum_{i=1}^{m} \beta_i X_{ij} + \varepsilon_j$$

The number of sample points must be two more than the number of variables

Multiple linear regression – least sum of squares

The principle of the least sum of squares are usually used to perform the fit:

$$\sum_{j=1}^{n}\left(Y_j - \hat{Y}_j\right)^2$$

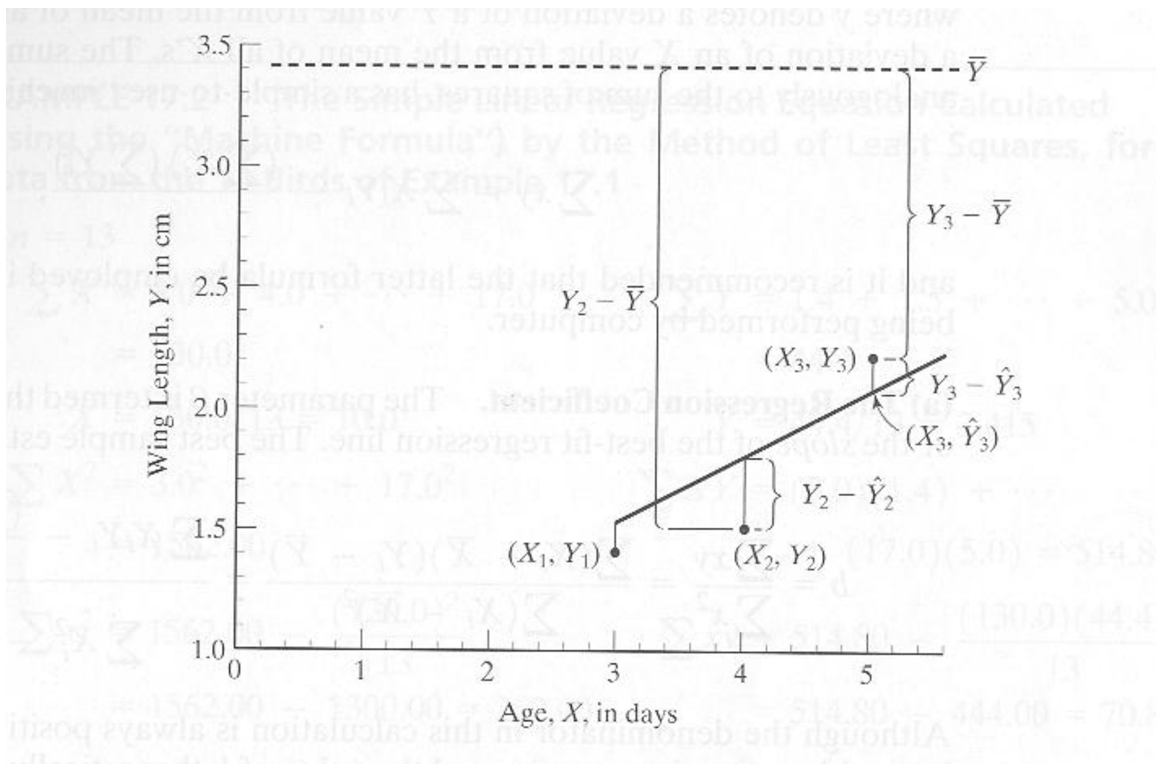


FIGURE 17.2: An enlarged portion of Figure 17.1, showing the partitioning of Y deviations.

Multiple linear regression – An example

EXAMPLE 20.1a The $n \times M$ Data Matrix for a Hypothetical Multiple Regression or Correlation ($n = 33$; $M = 5$)

	Variable (i)				
j	1 (°C)	2 (cm)	3 (mm)	4 (min)	5 (ml)
1	6	9.9	5.7	1.6	2.12
2	1	9.3	6.4	3.0	3.39
3	−2	9.4	5.7	3.4	3.61
4	11	9.1	6.1	3.4	1.72
5	−1	6.9	6.0	3.0	1.80
6	2	9.3	5.7	4.4	3.21
7	5	7.9	5.9	2.2	2.59
8	1	7.4	6.2	2.2	3.25
9	1	7.3	5.5	1.9	2.86
10	3	8.8	5.2	0.2	2.32
11	11	9.8	5.7	4.2	1.57
12	9	10.5	6.1	2.4	1.50
13	5	9.1	6.4	3.4	2.69
14	−3	10.1	5.5	3.0	4.06
15	1	7.2	5.5	0.2	1.98
16	8	11.7	6.0	3.9	2.29
17	−2	8.7	5.5	2.2	3.55
18	3	7.6	6.2	4.4	3.31
19	6	8.6	5.9	0.2	1.83
20	10	10.9	5.6	2.4	1.69
21	4	7.6	5.8	2.4	2.42
22	5	7.3	5.8	4.4	2.98
23	5	9.2	5.2	1.6	1.84
24	3	7.0	6.0	1.9	2.48
25	8	7.2	5.5	1.6	2.83
26	8	7.0	6.4	4.1	2.41
27	6	8.8	6.2	1.9	1.78
28	6	10.1	5.4	2.2	2.22
29	3	12.1	5.4	4.1	2.72
30	5	7.7	6.2	1.6	2.36
31	1	7.8	6.8	2.4	2.81
32	8	11.5	6.2	1.9	1.64
33	10	10.4	6.4	2.2	1.82

Multiple linear regression – The fitted equation

EXAMPLE 20.1c A Computer Fit of a Multiple-Regression Equation to the Data of Example 20.1a, Where Variable 5 Is the Dependent Variable

Regression model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

For each i (where $i = 1, 2, 3, 4$),

H_0: $\beta_i = 0$

H_A: $\beta_i \neq 0$

Variable i	b_i	s_{b_i}	t	ν	b'_i
X_1	−0.12932	0.021287	−6.075	28	−0.73176
X_2	−0.018785	0.056278	−0.334	28	−0.41108
X_3	0.046215	0.20727	0.223	28	−0.26664
X_4	0.20876	0.067034	3.114	28	0.36451

Y intercept: $a = 2.9583$

Multiple linear regression – Are any of the coefficients significant?

F = regression MS / residual MS

EXAMPLE 20.1d A Computer Analysis of Variance for the Multiple Regression Data of Example 20.1a

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_A: \beta_1$ and/or β_2 and/or β_3 and/or $\beta_4 \neq 0$

Source of Variation	SS	DF	MS
Total	14.747	32	
Multiple regression	9.7174	4	2.4294
Residual	5.0299	28	0.17964

$F = 13.5$, with DF of 4 and 28

$F_{0.05(1),4,28} = 2.71$, so reject H_0.

$P \ll 0.0005 \quad [P = 2.99 \times 10^{-6}]$

Coefficient of determination: $R^2 = 0.65893$

Adjusted coefficient of determination: $R_a^2 = 0.61021$

Multiple correlation coefficient: $R = 0.81175$

Standard error of estimate: $s_{Y \cdot 1,2,3,4} = 0.42384$

Multiple linear regression – Is it a good fit?

R^2 = 1-regression SS / total SS

- Is an expression of how much of the variation can be described by the model
- When comparing models with different numbers of variables the adjusted R-square should be used:

R_a^2 = 1 – regression MS / total MS

The multiple regression coefficient:

R = sqrt(R^2)

The standard error of the estimate = sqrt(residual MS)

EXAMPLE 20.1d A Computer Analysis of Variance for the Multiple Regression Data of Example 20.1a

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_A: \beta_1$ and/or β_2 and/or β_3 and/or $\beta_4 \neq 0$

Source of Variation	SS	DF	MS
Total	14.747	32	
Multiple regression	9.7174	4	2.4294
Residual	5.0299	28	0.17964

$F = 13.5$, with DF of 4 and 28

$F_{0.05(1),4,28} = 2.71$, so reject H_0.

$P \ll 0.0005 \quad [P = 2.99 \times 10^{-6}]$

Coefficient of determination: $R^2 = 0.65893$

Adjusted coefficient of determination: $R_a^2 = 0.61021$

Multiple correlation coefficient: $R = 0.81175$

Standard error of estimate: $s_{Y \cdot 1,2,3,4} = 0.42384$

Multiple linear regression – Which of the coefficient are significant?

- s_{bi} is the standard error of the regression parameter b_i
- t-test tests if b_i is different from 0
- $t = b_i / s_{bi}$
- ν is the residual DF
- p values can be found in a table

EXAMPLE 20.1c A Computer Fit of a Multiple-Regression Equation to the Data of Example 20.1a, Where Variable 5 Is the Dependent Variable

Regression model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

For each i (where $i = 1, 2, 3, 4$),

H_0: $\beta_i = 0$

H_A: $\beta_i \neq 0$

Variable i	b_i	s_{b_i}	t	ν	b'_i
X_1	−0.12932	0.021287	−6.075	28	−0.73176
X_2	−0.018785	0.056278	−0.334	28	−0.41108
X_3	−0.046215	0.20727	−0.223	28	−0.26664
X_4	0.20876	0.067034	3.114	28	0.36451

Y intercept: $a = 2.9583$

Multiple linear regression – Which of the are most important?

- The standardized regression coefficient , b' is a normalized version of b

$$b_i' = b_i \sqrt{\frac{\sum x_i^2}{\sum y^2}}$$

EXAMPLE 20.1c A Computer Fit of a Multiple-Regression Equation to the Data of Example 20.1a, Where Variable 5 Is the Dependent Variable

Regression model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

For each i (where $i = 1, 2, 3, 4$),

H_0: $\beta_i = 0$

H_A: $\beta_i \neq 0$

Variable i	b_i	s_{b_i}	t	ν	b_i'
X_1	−0.12932	0.021287	−6.075	28	−0.73176
X_2	−0.018785	0.056278	−0.334	28	−0.41108
X_3	−0.046215	0.20727	−0.223	28	−0.26664
X_4	0.20876	0.067034	3.114	28	0.36451

Y intercept: $a = 2.9583$

Multiple linear regression - multicollinearity

- If two factors are well correlated the estimated b's becomes inaccurate.
- Collinearity, intercorrelation, nonorthogonality, illconditioning
- Tolerance or variance inflation factors can be computed

- Extreme correlation is called singularity and on of the correlated variables must be removed.

Multiple linear regression – Pairvise correlation coefficients

EXAMPLE 20.1b A Matrix of Simple Correlation Coefficients, as It Might Appear as Computer Output (from the Data of Example 20.1a)

	1	2	3	4	5
1	1.00000	0.32872	0.16767	0.05191	−0.73081
2	0.32872	1.00000	−0.14550	0.18033	−0.21204
3	0.16767	−0.14550	1.00000	0.24134	−0.05541
4	0.05191	0.18033	0.24134	1.00000	0.31267
5	−0.73081	−0.21204	−0.05541	0.31267	1.00000

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}; \sum xy = \sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right); \sum x^2 = \sum \left(X_i - \bar{X}\right)^2$$

Multiple linear regression – Assumptions

The same as for simple linear regression:

1. Y's are randomly sampled
2. The residuals are normal distributed
3. The residuals have equal variance
4. The X's are fixed factors (their error are small).
5. The X's are not perfectly correlated