Multivariate Genetic Analysis

Bivariate Heritability

The genetic and environmental components of covariation can be separated by a technique which is analogous to **factor analysis** [8]. Just as **structural equation modeling (SEM)** can be used to analyze the components of variance influencing a single variable in the case of univariate analysis, SEM can also be used to analyze the sources and structure of covariation underlying multiple variables [8]. When based on genetically informative relatives such as twins, this methodology allows researchers to estimate the extent to which genetic and environmental influences are shared in common by several traits or are trait specific. Information not only comes from the covariance between the variables but also from the cross-twin cross-trait covariances. More precisely, a larger cross-twin cross-trait correlation between monozygotic twin pairs as compared with dizygotic twin pairs suggests that covariance between the variables is partially due to genetic factors. A typical starting point in bivariate and multivariate analysis is the **Cholesky decomposition**. Other methods include common and independent genetic pathway models, as well as genetic simplex models, which will also be discussed.

Multivariate Genetic Analysis

Cholesky Decomposition

The most commonly used multivariate technique in the Classical Twin design (*see* **Twin Designs**) is the Cholesky decomposition. As shown in Figure 1, the Cholesky is a method of triangular decomposition where the first variable (y_1) is assumed to be caused by a latent factor (*see* **Latent Variable**) (η_1) that can explain the variance in the remaining variables ($y_2, \ldots, y_n$). The second variable (y_2) is assumed to be caused by a second latent factor (η_2) that can explain variance in the second as well as remaining variables ($y_2, \ldots, y_n$). This pattern continues until the final observed variable (y_n) is explained by a latent variable (η_n), which is constrained from explaining the variance in any of the previous observed variables. A Cholesky decomposition is specified for each latent source of variance A, D, C, or E, and as in the univariate case, ACE, ADE, AE, DE, CE, and E models are fitted to the data (*see* **ACE Model**).

The expected variance–covariance matrix in the Cholesky decomposition is parameterized in terms of n latent factors (where n is the number of variables). All variables load on the first latent factor, $n - 1$ variables load on the second factor and so on, until

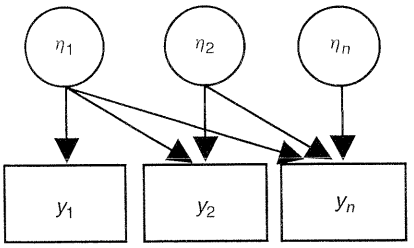

Figure 1 Multivariate Cholesky triangular decomposition, $y_1, \ldots, y_n$ = observed phenotypic variables, η_{1-n} = latent factors

Table 1 Phenotypic factor correlations for the factor analytic dimensions of depression, phobic anxiety, and somatic distress. Male correlations appear below the diagonal (reproduced from [3, p. 455])

	Females (n = 2219)		
	1	2	3
1 Depression		0.64	0.52
2 Phobic anxiety	0.63		0.58
3 Somatic distress	0.54	0.58	
	Males (n = 1418)		

the final variable loads on the nth latent factor only. Each source of phenotypic variation (i.e., A, C or D, and E) is parameterized in the same way. Therefore, the full factor Cholesky does not distinguish between common factor and specific factor variance and does not estimate a specific factor effect for any variable except the last.

Although symptoms of fatigue and somatic distress are frequently comorbid with anxiety and depressive disorders, a number of studies [6, 7, 14] have demonstrated that a significant proportion of patients with somatic disorders do not meet the criteria for other psychological disorders. As an example of how the Cholesky decomposition can be used to resolve these sorts of questions, we administered self-report measures of anxiety, depression, and somatic distress to a community-based sample of 3469 Australian twin individuals aged 18 to 28 years. As shown in Table 1, the phenotypic correlations between somatic distress, depression, and phobic anxiety, for males and females alike, are all high.

Table 2 Univariate model-fitting for the factor analytic dimensions of depression, phobic anxiety, and somatic distress. The table includes standardized proportions of variance attributable to genetic and environmental effects (reproduced from [4, p. 1056])

A	C	E	$-2LL$	df	$\Delta-2LL$	Δdf	p
Depression							
0.33	0.00	0.67	10720.59	7987			
0.33		0.67	10720.59	7988	0.05	1	.82
	0.24	0.76	10729.87	7988	9.28	1	[b]
		1.00	10791.02	7989	70.44	2	[c]
Phobic Anxiety							
0.37	0.03	0.59	7968.50	7979			
0.41		0.59	7968.62	7980	0.11	1	.74
	0.30	0.70	7976.54	7980	8.04	1	[b]
		1.00	8051.14	7981	82.64	2	[c]
Somatic Distress							
0.11	0.17	0.72	9563.28	7969			
0.32		0.68	9566.50	7970	3.22	1	.07
	0.25	0.75	9564.08	7970	0.79	1	.37
		1.00	9624.57	7971	61.29	2	[c]

Note: A, C & E = additive genetic, shared/common environment, and nonshared environment variance.
Results based on Maximum Likelihood.
$\Delta-2LL = -2$ log-likelihood.
[a] $p < 0.05$,
[b] $p < 0.01$,
[c] $p < 0.001$.

The results for the univariate genetic analyses in Table 2 reveal that an additive genetic (*see* **Additive Genetic Variance**) and **nonshared environmental effects** model best explains individual differences in depression and phobic anxiety scores, for male and female twins alike. The same could not be said for somatic distress because there is insufficient power to choose between additive genetic or shared environment effects as the source of familial aggregation in somatic distress. This limitation can be overcome using multivariate genetic analysis, which has greater power to detect genetic and environmental effects by making use of all the covariance terms between variables. Moreover, it will allow us to determine whether somatic distress is etiologically distinct from self-report measures of depression and anxiety.

As shown in Table 3, an additive genetic and nonshared environment (AE) model best explained the sources of covariation between the three factors.

Table 3 Multivariate Cholesky decomposition model-fitting results. Results are based on combined male and female data adjusted for sex differences in the prevalence of depression, phobic anxiety, and somatic distress (reproduced from [4, p. 1056])

Model			$-2LL$	df	$\Delta -2LL$	Δdf	p
A	C	E	26 000.29	15 285			
A		E	26 003.82	15 291	3.53	6	.74
	C	E	26 015.20	15 291	14.91	6	[a]
		E	26 147.97	15 297	147.68	12	[c]

Note: A, C & E = additive genetic, shared/common environment, and nonshared environment variance.
Results based on Maximum Likelihood.
$\Delta -2LL = -2$ log-likelihood.
[a] $p < .05$,
[b] $p < .01$,
[c] $p < .001$.

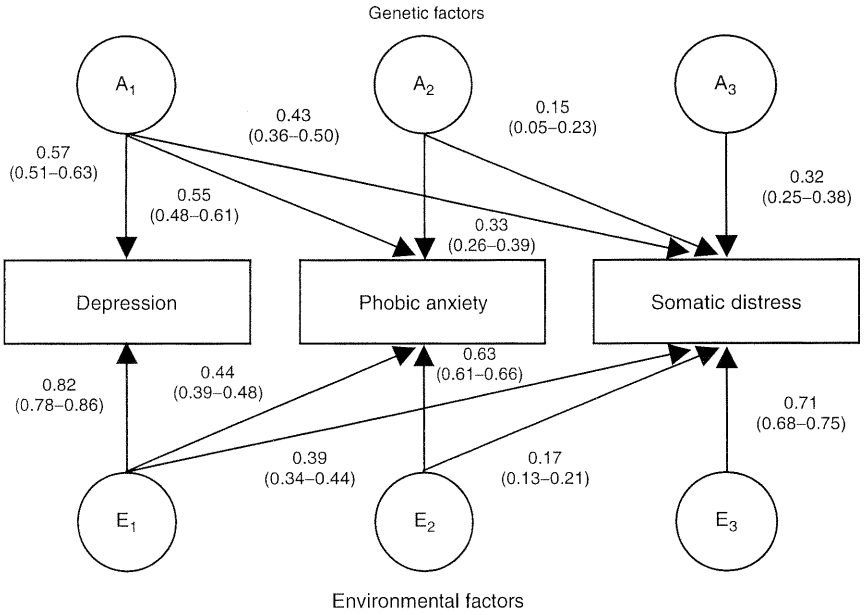

Figure 2 Path diagram showing standardized path coefficients and 95% confidence intervals for the latent genetic (A_1 to A_3) and environmental (E_1 to E_3) effect (reproduced from [4] p. 1057)

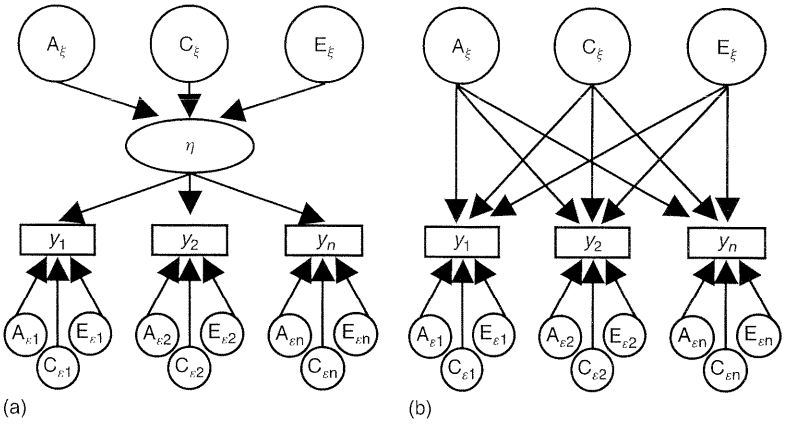

Figure 3 Common (a) and independent (b) genetic pathway models. η = common factor, $y_1, \ldots, y_n$ = observed phenotypic variables, A_ξ, C_ξ & E_ξ = latent genetic & environmental factors, $A_{\varepsilon 1-n}$, $C_{\varepsilon 1-n}$ & $E_{\varepsilon 1-n}$ = latent genetic & environmental residual factors

This is illustrated in Figure 2, where 33% (i.e., $0.32^2/0.43^2 + 0.15^2 + 0.32^2$) of the genetic variance in somatic distress is due to specific gene action unrelated to depression or phobic anxiety. In addition, 74% of the individual environmental influence on somatic distress is unrelated to depression and phobic anxiety. These results support previous findings that somatic symptoms are partly etiologically distinct, both genetically and environmentally from the symptoms of anxiety and depression.

Common and Independent Genetic Pathway Models

Alternate multivariate methods can be used to estimate common factor and specific factor variance (*see* **Factor Analysis: Exploratory**). For instance, the common pathway model in Figure 3a assumes that the genetic and environmental effects (A, C and E) (*see* **ACE Model**) contribute to one or more latent intervening variables (η), which in turn are responsible for the observed patterns of covariance between symptoms ($y_1, \ldots, y_n$).

This is in contrast to the independent pathway model in Figure 3b, which predicts that genes and environment have different effects on the covariance between symptoms. Because it can be shown algebraically that the common pathway is nested within the independent pathway model, the two models can be compared using a likelihood ratio chi-squared statistic (*see* **Goodness of Fit**).

Parker's 25-item Parental Bonding Instrument (PBI) [13] was designed to measure maternal and paternal parenting along the dimensions of Care and

Table 4 Best fitting univariate models with standardized proportions of variance attributable to genetic and environmental variance (reproduced from [5, p. 390])

PBI dimensions	A	C	E	$-2LL$	df
Coldness	.61	–	.39	5979.01	3603
Overprotection	.22	.24	.54	7438.84	3602
Autonomy	.33	.17	.51	9216.17	3599

Note: A, C & E = additive genetic, shared/common environment, and nonshared environment variance.
Results based on Maximum Likelihood.
$\Delta -2LL = -2$ log-likelihood.
[a] $p < .05$,
[b] $p < .01$,
[c] $p < .001$.

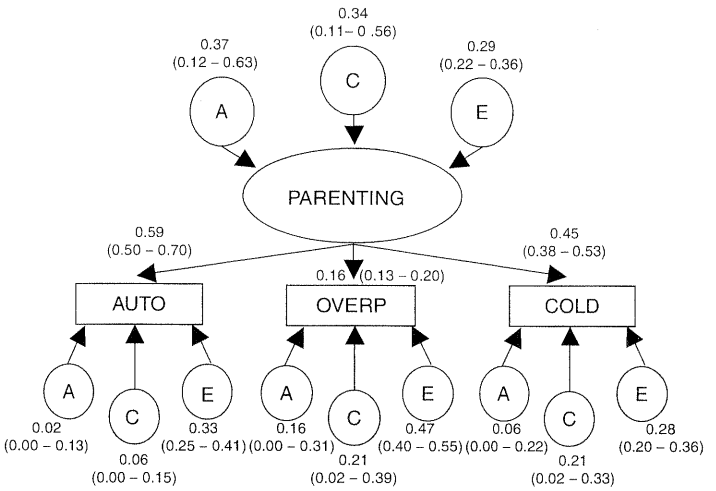

Figure 4 Common pathway genetic model (saturated) for the PBI dimensions with standardized proportions of variance and 95% confidence intervals (reproduced from 5, p. 391). AUTO = Autonomy, OVERP = Overprotection, COLD = Coldness, Results based on Weighted Least Squares

Table 5 Comparison of the common and independent pathway models for the PBI dimensions of Autonomy, Coldness, and Overprotection [5]

Model	χ^{2a}	df	p	*AIC*
Independent pathway	9.17	15	.87	−20.83
Common pathway	13.39	19	.82	−24.61

Note: Results based on Weighted Least Squares.
AIC = Akaike Information Criterion.

Overprotection [11, 12]. However, factor analysis of the short 14-item version based on 4514 females, aged 18 to 45, has yielded three correlated factors: Autonomy, Coldness, and Overprotection [5]. Univariate analyses of the three dimensions, which are summarized in Table 4, reveal that variation in parental Overprotection and Autonomy can be best explained by additive genetic, shared, and nonshared environmental effects, whereas the best fitting model for Coldness includes additive genetic and nonshared environmental effects. As is shown in Table 5, when compared to an independent pathway model, a common pathway genetic model provided a more parsimonious fit to the three PBI dimensions. The common pathway model is illustrated in Figure 4.

Genetic Simplex Modeling

When genetically informative longitudinal data are available, a multivariate Cholesky can again be fitted to determine the extent to which genetic and environmental influences are shared in common by a trait measured at different time points. However, this approach is limited in so far as it does not take full advantage of the time-series nature of the data, that is, that causation is unidirectional through time [1].

One solution is to fit a simplex model, which explicitly takes into account the longitudinal nature of the data. As shown in Figure 5, simplex models are autoregressive, whereby the genetic and environmental latent variables at time i are causally related to the immediately preceding latent variables (η_{i-1}).

Eta (η_i) is a latent variable (i.e., A, C, E, or D) at time i, β_i is the regression of the latent factor on the immediately preceding latent factor η_{i-1}, and ζ_i is the new input or innovation at time i. When using data from MZ and DZ twin pairs, structural equations can be specified for additive genetic sources of variation (A), common environmental (C), nonadditive genetic sources of variation such as dominance

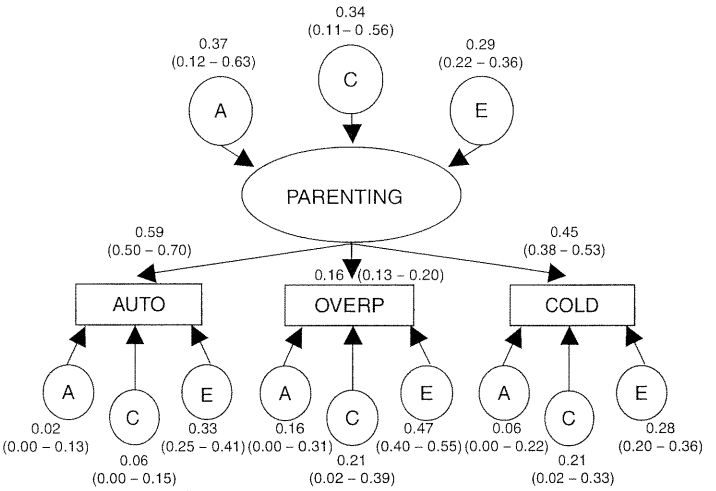

Figure 4 Common pathway genetic model (saturated) for the PBI dimensions with standardized proportions of variance and 95% confidence intervals (reproduced from 5, p. 391). AUTO = Autonomy, OVERP = Overprotection, COLD = Coldness, Results based on Weighted Least Squares

Table 5 Comparison of the common and independent pathway models for the PBI dimensions of Autonomy, Coldness, and Overprotection [5]

Model	χ^{2a}	df	p	*AIC*
Independent pathway	9.17	15	.87	−20.83
Common pathway	13.39	19	.82	−24.61

Note: Results based on Weighted Least Squares.
AIC = Akaike Information Criterion.

Overprotection [11, 12]. However, factor analysis of the short 14-item version based on 4514 females, aged 18 to 45, has yielded three correlated factors: Autonomy, Coldness, and Overprotection [5]. Univariate analyses of the three dimensions, which are summarized in Table 4, reveal that variation in parental Overprotection and Autonomy can be best explained by additive genetic, shared, and nonshared environmental effects, whereas the best fitting model for Coldness includes additive genetic and nonshared environmental effects. As is shown in Table 5, when compared to an independent pathway model, a common pathway genetic model provided a more parsimonious fit to the three PBI dimensions. The common pathway model is illustrated in Figure 4.

Genetic Simplex Modeling

When genetically informative longitudinal data are available, a multivariate Cholesky can again be fitted to determine the extent to which genetic and environmental influences are shared in common by a trait measured at different time points. However, this approach is limited in so far as it does not take full advantage of the time-series nature of the data, that is, that causation is unidirectional through time [1].

One solution is to fit a simplex model, which explicitly takes into account the longitudinal nature of the data. As shown in Figure 5, simplex models are autoregressive, whereby the genetic and environmental latent variables at time i are causally related to the immediately preceding latent variables (η_{i-1}).

Eta (η_i) is a latent variable (i.e., A, C, E, or D) at time i, β_i is the regression of the latent factor on the immediately preceding latent factor η_{i-1}, and ζ_i is the new input or innovation at time i. When using data from MZ and DZ twin pairs, structural equations can be specified for additive genetic sources of variation (A), common environmental (C), nonadditive genetic sources of variation such as dominance

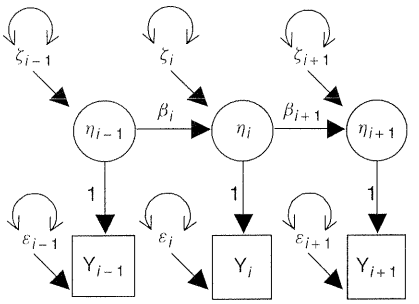

Figure 5 General simplex model. β = regression of the latent factor on the previous latent factor ζ = new input or innovation at time i η = latent variable at time i

or epistasis (D), and unique environmental sources of variation (E).

Because measurement error does not influence observed variables at subsequent time points, simplex designs therefore permit discrimination between transient factors effecting measurement at one time point only, and factors that are continuously present or exert a long-term influence throughout the time series [1, 9]. Although denoted as error variance, the error parameters will also include variance attributable to short-term nonshared environmental effects.

We have used this model-fitting approach to investigate the stability and magnitude of genetic and environmental effects underlying major dimensions of adolescent personality across time [2]. The junior eysenck personality questionnaire (JEPQ) was administered to over 540 twin pairs at ages 12, 14, and 16 years. Results for JEPQ Neuroticism are presented here.

The additive genetic factor correlations based on a Cholesky decomposition are shown in Table 6. These reveal that the latent additive genetic factors are highly correlated. This is consistent with a pleiotropic model of gene action, whereby the same genes explain variation across different time points. As shown in Table 7, the fit of the ACE simplex models

Table 6 Additive genetic (above diagonal) and nonshared environmental latent factor correlations for JEPQ Neuroticism (reproduced from [2])

	Females			Males		
	1	2	3	1	2	3
1 12 years		0.76	0.68		0.86	0.79
2 14 years	0.40		0.94	0.24		0.74
3 16 years	0.36	0.41		0.12	0.53	

Table 7 Multivariate model-fitting results for JEPQ Neuroticism based on twins aged 12, 14, and 16 years (reproduced from [2])

	Neuroticism									
	Females					Males				
	$-2LL$	df	$\Delta 2LL$	Δdf	P	$-2LL$	df	$\Delta 2LL$	Δdf	p
Cholesky										
ACE	10424.88	1803				10016.48	1753			
Simplex										
ACE	10424.95	1805	0.07	2	.96	10016.88	1755	0.40	2	.82
AE	10425.34	1810	0.39	5	1.00	**10021.26**	**1760**	**4.38**	**5**	**.50**
Drop ζ_{a3}	**10426.57**	**1811**	**1.23**	**1**	**.27**	10058.59	1761	37.32	1	[c]
CE	10432.45	1810	7.50	5	.19	10032.32	1760	15.44	5	[b]
E	10663.09	1815	238.14	10	[c]	10650.99	1765	634.11	10	[c]

Note: Results based on Maximum Likelihood.
ζ_{a3} = Genetic innovation at time 3.
Best fitting models in bold.

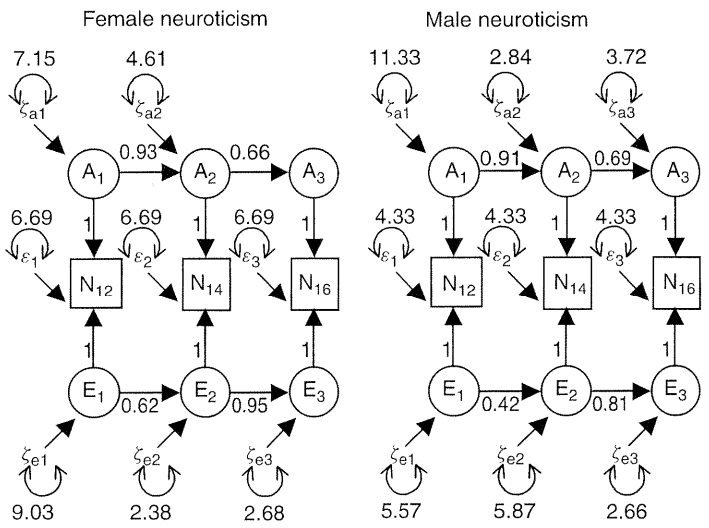

Figure 6 Best fitting genetic simplex model for female and male Neuroticism (reproduced from [2]). N_{12-16} = Neuroticism 12 to 16 yrs, A_{1-3}, E_{1-3} = additive genetic & nonshared environmental effects, ζ_{a1-3}, ζ_{e1-3} = additive genetic innovations & nonshared environmental innovations, ε_{1-3} = error parameters 12 to 16 yrs double/single headed arrows = variance components/path coefficients

provided a better explanation of the Neuroticism time-series data, in so for as the fit was no worse than the corresponding Cholesky decompositions. The final best-fitting AE simplex models for male and female Neuroticism are shown in Figure 6.

It is difficult to imagine that genetic variation in personality is completely determined by age 12. As shown in Figure 6, smaller genetic innovations are observed for male Neuroticism at 14 and 16, as well as female Neuroticism at 14. These smaller genetic innovations potentially hint at age-specific genetic effects related to developmental or hormonal changes during puberty and psychosexual development.

When data are limited to three time points, a common genetic factor model will also provide a comparable fit when compared to the genetic simplex model. Other possible modeling strategies include biometric growth models (see [10]). Despite these limitations, time-series data even when based on three time points still provides an opportunity to test explicit hypotheses of genetic continuity. Moreover, the same data are ideal for fitting univariate and multivariate linkage models to detect quantitative trait loci of significant effect.

The above sections have provided an introduction to bivariate and multivariate analyses and how these methods can be used to estimate the genetic and environmental covariance between phenotypic measures. This should give the reader an appreciation for the flexibility of SEM approaches to address more complicated questions beyond univariate decompositions. For a more detailed treatment of this subject, see [9].

References

[1] Boomsma, D.I., Martin, N.G. & Molenaar, P.C. (1989). Factor and simplex models for repeated measures: application to two psychomotor measures of alcohol sensitivity in twins, *Behavior Genetics* **19**, 79–96.

[2] Gillespie, N.A., Evans, D.E., Wright, M.M. & Martin, N.G. (submitted). Genetic simplex modeling of Eysenck's dimensions of personality in a sample of young Australian twins, *Journal of Child Psychology and Psychiatry*.

[3] Gillespie, N., Kirk, K.M., Heath, A.C., Martin, N.G. & Hickie, I. (1999). Somatic distress as a distinct psychological dimension, *Social Psychiatry and Psychiatric Epidemiology* **34**, 451–458.

[4] Gillespie, N.G., Zhu, G., Heath, A.C., Hickie, I.B. & Martin, N.G. (2000). The genetic aetiology of somatic distress, *Psychological Medicine* **30**, 1051–1061.

[5] Gillespie, N.G., Zhu, G., Neale, M.C., Heath, A.C. & Martin, N.G. (2003). Direction of causation modeling between measures of distress and parental bonding, *Behavior Genetics* **33**, 383–396.

[6] Hickie, I.B., Lloyd, A., Wakefield, D. & Parker, G. (1990). The psychiatric status of patients with chronic fatigue syndrome, *British Journal of Psychiatry* **156**, 534–540.
[7] Kroenke, K., Wood, D.R., Mangelsdorff, D., Meier, N.J. & Powell, J.B. (1988). Chronic fatigue in primary care, *Journal of the American Medical Association* **260**, 926–934.
[8] Martin, N.G. & Eaves, L.J. (1977). The genetical analysis of covariance structure, *Heredity* **38**, 79–95.
[9] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families. NATO ASI Series*, Kluwer Academic Publishers, Dordrecht.
[10] Neale, M.C. & McArdle, J.J. (2000). Structured latent growth curves for twin data, *Twin Research* **3**, 165–177.
[11] Parker, G. (1989). The parental bonding instrument: psychometric properties reviewed, *Psychiatric Developments* **4**, 317–335.
[12] Parker, G. (1990). The parental bonding instrument. A decade of research, *Social Psychiatry and Psychiatric Epidemiology* **25**, 281–282.
[13] Parker, G., Tupling, H. & Brown, L.B. (1979). A parental bonding instrument, *British Journal of Medical Psychology* **52**, 1–10.
[14] Wessely, S. & Powell, R. (1989). Fatigue syndromes: a comparison of chronic postviral with neuromuscular and affective disorders, *Journal of Neurology, Neurosurgery, and Psychiatry* **52**, 940–948.

NATHAN A. GILLESPIE AND NICHOLAS G. MARTIN

Multivariate Kurtosis *see* Multivariate Normality Tests

Multivariate Multiple Regression

The multivariate linear model is used to explain and to analyze the relationship between one or more explanatory variables and $p > 1$ quantitative dependent or response variables that have been observed at n subjects. In case all the explanatory variables are qualitative, the multivariate linear model is called the **Multivariate Analysis of Variance** (*MANOVA*) *model*. When all the explanatory variables are quantitative, that is a multivariate system of quantitative variables is given in which the relationships between p dependent quantitative variables and, say q, independent quantitative are of interest, then the model is referred to as a *multivariate regression model*. Multivariate regression analysis is used to investigate the relationships when the p dependent variables are correlated. In contrast, when the dependent variables are uncorrelated, relationships can be assessed by carrying out p univariate regression analyses (*see* **Regression Models**). Often, there is an implied predictive aim in the investigation, and the formulation of appropriate and parsimonious relationships among the variables is a necessary prerequisite.

Consider the small example given in [2]. The data in Table 1 show the four measurements: chest circumference (CC), midupper arm circumference (MUAC), height and age (in months) for a sample on nine girls. One practical objective would be to develop a predictive model for CC and MUAC from knowledge of height and age.

The dependent variables CC and MUAC are highly correlated with each other and Pearson's correlation coefficient is 0.77, so they should be incorporated in a single multivariate regression model for maximum efficiency as multiple regression analyses for each variable separately will ignore this correlation in the construction of hypothesis tests or confidence intervals.

Let y_{i1} denote the CC of the ith girl, y_{i2} the MUAC, x_{i1} the height, and x_{i2} the age; then, the univariate regression models for each variable are

$$y_{i1} = \beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2} + e_{i1}, \quad i = 1, \ldots, n = 9, \tag{1}$$

and

$$y_{i2} = \beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2} + e_{i2}, \quad i = 1, \ldots, n = 9. \tag{2}$$

In the multivariate case, the observations y_{i1} and y_{i2} are put in a row vector so that the model has the form

$$(y_{i1} \quad y_{i2}) = (1 \quad x_{i1} \quad x_{i2}) \begin{pmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{12} & \beta_{22} \end{pmatrix} + (e_{i1} \quad e_{i2}), \quad i = 1, \ldots, n = 9. \tag{3}$$