Analysis of Functional Abilities for Elderly Danish Twins Using GEE Models

Maria Iachina[1,2], Bent Jørgensen[1], Kaare Christensen[2], and Ivan Iachine[1,2]

[1] Department of Statistics and Demography, University of Southern Denmark

[2] The Danish Twin Registry, Institute of Public Health, University of Southern Denmark

In this work we present a new method for genetic analysis of twin data which is based on generalized estimating equations and allows for analysis of various response types (e.g., continuous, binary, counts) combined with estimation of residual correlations. The new approach allows for control of covariates of any kind (e.g., continuous, counts) by modeling the dependence of mean and variance on background variables. The proposed method was applied to identify the covariates that have a significant influence on elderly people's functional abilities, and find the estimates for the correlation coefficients of residuals for MZ and DZ twins in a sample of 2401 Danish twin 75 years of age or older. The bootstrap method was used to obtain standard errors for correlation coefficients. It was shown, that the chosen covariates have similar effects on MZ and DZ twins, and that the residual correlation in MZ twins is significantly higher than in DZ twins, which indicates that genetic factors play an etiological role in the determination of physical status of elderly people, controlled for 10 background variables.

In this study we investigate the functional abilities of elderly people using twin data, which were obtained in an interview survey among Danish twins aged 75 and older. In this age group, most twins have a deceased co-twin. Since studying pairs in which both twins are alive might introduce an oversampling of healthy twin individuals, we included all twins who were 75 years and older, regardless of whether the co-twin was alive (Christensen et al., 2000).

In humans, two types of twinning occur: monozygotic twins share all their genetic material, and dizygotic twins, like ordinary siblings, share, on average, 50% of their genes. This difference in genetic similarities of twins provides a basis for a variety of methods of genetic analysis of phenotypes using twin data (Neale & Cardon, 1992). A crucial assumption of this methodology is that MZ and DZ are no different from ordinary individuals in the general population. Some studies might suggest that this is not necessarily the case and explain this by differences in the intra-uterine development (Barker, 1994). Therefore, a question which is worthwhile to investigate is that of similarity of monozygotic and dizygotic twins with regard to the phenotype studied.

Another interesting question concerns the roles of genetic and environmental factors in the determination of health indicators in the elderly. A prevailing assumption in gerontology is that the accumulation of unique environmental exposures during life is the key determinant of health at older ages (Harris et al., 1992). Alternatively, evolutionary biologists have argued that there is less selective pressure against deleterious genetic mutations first expressed late in life than against mutations expressed early in life. This hypothesis predicts an increase in genetic variance among the oldest individuals (Charlesworth, 1900).

The present study investigates the roles of genetic and environmental factors in the variation of functions important for functional abilities among elderly people, controlled for a number of background variables. Simultaneously we will determine which of these background variables have a significant influence on the person's functional abilities.

Usually, a likelihood-based model is applied in twin studies. The validity of classical maximum likelihood inference for such models depends upon the correct choice of distribution for responses. In this work we propose to use Generalized Estimating Equations (GEE; Liang & Zeger, 1986) for the analysis of twin data. The advantage of the GEE approach is that it requires weaker distributional assumptions and maintains the properties of consistency and asymptotic normality of parameter estimates (Fitzemaurice et al., 1993).

Materials and Methods

Study Population

In this work we study the first wave of the Longitudinal Study of Aging Danish Twins which includes twin pairs born in Denmark between 1870 and 1910 and same-sex pairs born between 1911 and 1930 (Christensen et al., 2000; Hauge et al., 1968). The study comprised all registered Danish twins who were 75 years or older on January 1, 1995, regardless of whether the co-twin was alive, a total of 3099 individuals. Face-to-face interviews were completed during a 3-month period (February–April 1995) by 100 interviewers, and a total of 2401 interviews were conducted, corresponding to a participation rate of 77%. This dataset is hereafter referred to as LSADT95.

The response rate was significantly higher for men (81%) than for women (74%). The responders and nonresponders were similar in terms of age distribution and monozygotic-dizygotic ratio. The mean age for responders

Address for correspondence: Maria Iachina, Department of Statistics and Demography, University of Southern Denmark, Campusvej 55 DK-5230 Odense M, Denmark. Email: mia@statdem.sdu.dk

https://doi.org/10.1375/twin.5.4.289 Published online by Cambridge University Press

and nonresponders of both sexes were within 0.6 year of 81 year. The previous 18 years of hospital admission patterns were nearly identical for female responders and nonresponders, although the male nonresponders tended to have slightly fewer hospitalizations than did the responders (Christensen et al., 2000). A large number of the twins from LSADT95 were those whose co-twins had died before the survey. Thus the given twin study included 351 intact twin pairs. Of the 2401 individual twins who completed the interview assessment, 75 were excluded from the present sample because they were assessed by proxy and 431 individuals were excluded because they did not complete the interview.

The Model Construction

Christensen et al. (2000) analyzed LSADT95 dataset using linear regression to control for age and sex and constructed a "Strength" score. This score is based on the nature of the functional abilities. The "Strength" score takes values from 1 to 4, and it grows from "bad" to "good" for example: a person who has "strength" = 3 has it physically better than a person who has "strength" = 1. A more detailed description of "Strength" score may be found in Appendix A.

We want to use the "Strength" score as a dependent variable in our analysis, but we want that our dependent variable takes values from 0 to 1. Therefore we transformed the "Strength" score by subtracting 1 and thereafter dividing by 3, the new transformed score was called simply "strength". From all questions to participants, 29 were chosen to be independent variables in the model, because they known from previous studies to have influence on functional abilities. The list of the covariates is shown in Table 1.

To simplify the interpretation of the results of this analysis all independent variables for which it is possible to define a "good" and a "bad" status were coded so that increase of the variable corresponds to change of status from "good" to "bad". To investigate the effect of sex on the outcome, the independent variable "woman" was introduced (coded 1 for males and 2 for females). The respective regression coefficient indicates how women perform in comparison to men.

Table 1
Covariates

Number	Name	Answers
01	Age	75–102
02	Zygosity	1,2 (mz,dz)
03	Sex	1,2 (m,f)
04	Is your twin alive?	1,2 (y,n)
05	BMI	13–42
06	How many biological children do you have?	0–11
07	How many adopted children do you have?	0–4
08	What type of elementary school education did you receive?	1–4 (< 7,7–8, 9–10, 11)
09	Did you get any education after elementary school?	1–7 (none, SE, PE, < 3,3–4, > 4, other)
10	Are you or have you been married?	1–5 (never, m, dev, sep, en)
11	Have you lost any close relatives or close friends during the last 5 years?	1,2 (y,n)
12	Do you participate in any joint activities?	1,2 (y,n)
13	Do you smoke?	1–3 (n, less then one c. a day,y)
14	Do you drink?	1,2 (n,y)
15	Do you live alone?	1,2 (alone, not alone)
16	How do you consider your health in general?	1–5 (vg, g, ok, b, vb)
17	Did you ever hit your head so seriously that you became unconscious?	0–3 (n, 1, 2, >2 times)
18	Have you been taking pain killing medications during long periods of time?	1,2 (n,y)
19	Did a doctor ever tell you that you have/had any of the following diseases: Diabetes	1,2(n,y)
20	Irregular heart rhythm	1,2 (n,y)
21	Cancer	1,2 (n,y)
22	Angina pectoris	1,2 (n,y)
23	Wet lungs	1,2 (n,y)
24	Gastric ulcer	1,2 (n,y)
25	Heart attack	1,2 (n,y)
26	Stroke	1,2 (n,y)
27	Hypertension	1,2 (n,y)
28	Other heart problems	1,2 (n,y)
29	Parkinson's disease	1,2 (n,y)

Let Y_{ij} ,j = 1,2,i = 1,…,K represent the jth (first or second) twin in the ith pair of the total of K pairs, then μ_{ij} will be the corresponding vector of means; and let the vector of independent variables for the jth twin on the ith pair be $x_{ij} = (x_{ij1},\ldots,x_{ijp})$.

The GEE model will be used in this twin analysis. The term Generalized Estimating Equations was first introduced in the context of longitudinal data analysis by Liang and Zeger (1986). These authors introduced a class of estimating equations that are based on second moment assumption only.

The main assumption of this analysis is that twins from different pairs are independent given the covariates and that responses of members of a twin pair may be dependent. Estimation of correlation between twins in a pair is computed using Unstructured working correlation matrix structure, which is

$$Corr(Y_{ij},Y_{ik}) = \left\{\begin{matrix} 1, j = k \\ \rho_{jk}, j \neq k \end{matrix}\right\},$$

$$\text{where } \hat{\rho}_{jk} = \frac{1}{\quad\quad} \sum_{i=1}^{K} e_{ij}e_{ik} .$$

Her is e_{ij} a Pearson's residual given by

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{Var(\mu_{ij})}},$$

and ρ is the correlation between e_{i1} and e_{i2}, Pearson's residuals for the first and second twin respectively.

In this analysis we will use binomial distribution's variance function given by

$$Var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})\varphi;$$

and logit link function given by

$$\mu_{ij} = \frac{e^{x_{ij}\beta}}{\quad\quad},$$

where β is a vector of regression parameters; and φ is the dispersion parameter estimated by

$$\hat{\phi} = \frac{1}{2K - p} \sum_{i=1}^{K} (e_{i1}^2 + e_{i2}^2);$$

For a more detailed description of GEE see Pickles (Pickles, 1998).

To eliminate all nonsignificant parameters from vector β Simes's method was used.

Simes (Simes, 1986) presents a step-down multiple testing procedure, which he refers to as a sequentially rejective Bonferroni test, based on the ordered P values. Let $P_1,\ldots P_k$ be the ordered p-values for testing hypotheses $H_{0,1},\ldots,H_{0,k}$. Then $H_{0,i}$ is rejected if $P_j \leq j\ \alpha/k$ for any j = 1,…k. Using this procedure the probability of falsely rejecting one or more hypotheses is at most α. The k is a number of parameters in the biggest model.

In our case α = 0.05, k = 29, and Simes's method will be: first, find a p-value for every parameter; second, eliminate the parameter with the largest nonsignificant p-value from the model, and find a new estimate for β in the reduced model. Repeating these steps until no more parameters can be eliminated we get the model with only significant parameters.

Results

We will start with a separate analysis of monozygotic (MZ) and dizygotic (DZ) twins, because of hypotheses of dissimilarities between MZ and DZ twins due to the differences in the intra-uterine development (Barker, 1994). Using SAS/GENMOD we can obtain the estimates for the regression coefficient $\hat{\beta}_{MZ}$, $\hat{\beta}_{DZ}$, and the estimates for correlation coefficient $\hat{\rho}_{MZ}$, $\hat{\rho}_{DZ}$ .Comparing β_{MZ} and β_{DZ} can be done by applying the Wald test to the hypothesis

$$H_0 : \beta_{MZ} = \beta_{DZ} , \qquad (1)$$

which yields a p-value = 0.4897. Since the respective p-value is higher than the usual significance level of 5%, the hypothesis (1) cannot be rejected. In the following it will therefore be assumed that $\beta_{MZ} = \beta_{DZ}$.

Now we want to find one common estimate for regression coefficient $\hat{\beta}_{MDZ}$ and still two estimates for correlation coefficient $\hat{\rho}_{MZ}$, $\hat{\rho}_{DZ}$ this can be done using SAS coded procedure GENMOD_MDZ, which is described in (Iachina, 2001). As it was mentioned earlier we have 29 parameters in the model. Some of them have a significant influence on the response, while other do not. Table 2 shows the results of the Simes method, which eliminate all nonsignificant parameters from vector $\beta_{MDZ.}$

The dispersion parameter $\hat{\varphi}$ was estimated to about 0.21.

Using GEE we get vector of estimated regression coefficients and two estimates for the correlation coefficients $\hat{\rho}_{MZ}$ and $\hat{\rho}_{DZ}$. To find the standard errors of $\hat{\rho}_{MZ}$ and $\hat{\rho}_{DZ}$ the parametric bootstrap algorithm of Efron and Tibshirani (Efron & Tibshirani, 1986) was used.

In Table 3 are shown estimates, standard errors and 95% confidence intervals for ρ_{MZ}, ρ_{DZ} and $\rho_{MZ} - \rho_{DZ}$. Since the confidence interval for $\rho_{MZ} - \rho_{DZ}$ does not include zero, we can conclude that there is a significant difference between ρ_{MZ} and ρ_{DZ}.

Discussion

In this work we investigated functional abilities of persons aged 75 years or older by analyzing the variable "strength", which was defined earlier, using the LSADT95 data set.

Christensen et al. (2000) analyzed the same data set by applying a two-step procedure to estimate the residual correlation. First, the estimate of the mean in each age and sex group was found, which can be described as almost equivalent to fitting a linear regression model with age and sex as dependent variables. Then, the correlation estimates were found using the residuals obtained from the previous step. It was found that monozygotic twins show significantly higher residual correlation than dizygotic twins, indicating a genetic influence on "strength".

In our analysis, residual correlation coefficients were estimated using a GEE logistic regression model with logit link

https://doi.org/10.1375/twin.5.4.289 Published online by Cambridge University Press

Table 2

Results of GEE Analysis of Strength Score for MZ and DZ Twins

Parameter	Estimate	*SE*	*p*-value
Intercept	8.76	0.55	< 0.0001
Bmi	–0.025	0.0061	< 0.0001
Age	–0.093	0.0052	< 0.0001
Woman	–0.29	0.053	< 0.0001
Educ after school	0.066	0.016	< 0.0001
Alcohol	0.035	0.0067	< 0.0001
Self-reported health	–0.65	0.025	< 0.0001
Stroke	–0.69	0.093	< 0.0001
Heart problem	–0.34	0.085	< 0.0001
Hypertension	0.13	0.055	0.015
Wet lungs	–0.21	0.085	0.010

Table 3

Correlation Coefficient Estimates

Parameter	Estimate	*SE*	95% confidence interval
ρ_{MZ}	0.2846	0.03	(0.2258; 0.3434)
ρ_{DZ}	0.0256	0.025	(–0.0224; 0.0736)
$\rho_{MZ} - \rho_{DZ}$	0.2591	0.033	(0.1941; 0.3238)

function with 29 background variables. The results confirm the findings of (Christensen et al., 2000); namely that the residual correlation difference between monozygotic and dizygotic twins is statistically significant (see Table 3). Note, however, that in our case we found residual genetic effects after controlling for a much larger number of background variables (see Table 2).

Another important result of our analysis is that the Wald test for H_0: $\beta_{MZ} = \beta_{DZ}$ shows that there are no statistically significant differences between the regression coefficients for MZ and DZ twins. This means, that the covariates have the same effects on MZ and DZ twins, thereby offering no evidence for the hypotheses of dissimilarities between monozygotic and dizygotic twins in this regard.

We have also determined which of the chosen covariates have a significant influence on the strength score (see Table 2). This analysis shows, that the functional ability of a person worsens with age, functional abilities of women are worse than those of men, and that increased BMI has an adverse effect on a person's functional abilities, in line with what was shown in previous studies. An interesting result of this analysis is that the presence of education after school is related to an improvement of a person's physical status. The diagnoses of stroke, heart problems, and wet lungs have an adverse effect on the person's functional abilities, as would be expected.

The use of alcohol and the presence of hypertension diagnosis are significant factors too. Notice, however, that the two corresponding coefficients have a positive sign (see Table 2), so that alcohol consumption and presence of diagnosis of hypertension have a beneficial influence on the physical health of persons that are 75 years old or older. These findings may be explained by the fact, that moderate alcohol consumption has been shown to be beneficial in several studies, and a diagnosis of hypertension may be the basis for an adequate treatment. The negative sign before the covariate "self reported health" (see Table 2) indicates, that for the majority of participants the estimation of their own physical status corresponds to the computed variable strength.

The method of Generalized Estimating Equations has several advantages compared with the traditional method for twin data analysis. The GEE method allows for analysis of various types of responses (e.g., continuous, binary, counts). In particular, the traditional linear regression model is a special case of the GEE method, obtained by choosing the constant variance function and identity link function. Contrary to the traditional approach, the GEE logistic regression model used in the present paper takes into account the so-called "ceiling" effect, observed by Christensen (Christensen et al., 2000), by choosing a variance function which is zero at the end points of the scale. Moreover, the new method uses a single-step procedure to obtain the residual correlation estimates, yielding standard errors and confidence intervals, that take the variation of regression coefficient estimates into account, resulting in a more precise statistical inference.

https://doi.org/10.1375/twin.5.4.289 Published online by Cambridge University Press

Acknowledgment

Supported by U.S. National Institute on Aging (Grant NIA-P01-AG08761)

References

Barker, D. J. P. (1994). *Mothers, babies, and disease in later life.* London: BMI Publishing Group.

Charlesworth, B. (1990). Optimization models, quantitative genetics, and mutation. *Evolution, 44,* 520–538.

Christensen, K., McGue, M., Yashin, A., Iachine, I., Holm, N. V., & Vaupel, J. W. (2000). Genetic and environmental influences on functional abilities in Danish twins aged 75 years and older. *Medical Sciences, 55*A, M446–M452.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1,* 54–77.

Fitzmaurice, G. M., Laird, N., & Rotnitzky, A. (1993). Regression models for discrete longitudinal responses. *Statistical Science, 8,* 284–309.

Harris, J.R., Pedersen, N. L., McClearn, G. E., Plomin, R., & Nesselroade, J. R (1992). Age differences in genetic and environmental influences for health from the Swedish adoption/twin study of aging. *Journal of Gerontology, 47,* P213–P220.

Hauge, M., Harvald, B., & Fischer, M. (1968). The Danish twin register. *Acta Geneticae Medicae et Gemellologiae, 17*(2), 315–332.

Iachina, M. (2001). *Analysis of functional abilities for older Danish twins using GEE models (Monographs Vol. 6).* Odense, Denmark: Department of Statistics and Demography.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13–22

Neale, M. C., & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pickles, A. (1998). Generalized estimating equations. *Encyclopedia of Biostatistics, 2,* 1626–1637.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika, 73,* 751–754.

Appendix A

The functional abilities section comprised 26 questions, which are summarized in Table 4. All items refer to what the participant was able to do on the day of the interview on a 1 to 4 scale: 4 = can do without fatigue; 3 = can do with fatigue or minor difficulties; 2 = can do with aid or major difficulties; 1 = cannot do. To identify meaningful quantitative subscales, Christensen et al. (2000) factor analyzed the 26 items in the LSADT95 data.

In the factor analysis, three factors had an eigenvalue of more than 1, but few of the items loaded on the third factor. Therefore, a two-factor solution was adopted (Table 4). The first factor loaded highest on items dealing with ability to walk, run, climb stairs, and carry weights and was interpreted to reflect a dimension of strength. The second factor loaded highest on items dealing with ability to dress and wash oneself and get in and out of bed, and was interpreted to reflect a dimension of agility.

The score "Strength" was calculated by taking the average response of items loading highest on the factor or having judged to be relevant for that dimension; they are marked as * in Table 4. If a respondent did not answer one or two of the questions, the mean value for same-sex twins with the similar answers for others questions was substituted for these missing items. If a respondent did not answer in more than two questions the result is a missing in the scale score.

Table 4

Functional Abilities and Factor Pattern Matrix

Functional Ability	Factor 1	Factor 2
Get up from a chair and bed	0.45	0.68
Walk around in the house *	0.57	0.63
Able to go the toilet	0.46	0.61
Walk up down stairs one floor *	0.72	0.45
Walk up stairs to the second floor *	0.76	0.36
Able to get outdoors *	0.73	0.42
Able to walk 400 meters without resting *	0.75	0.41
Do light exercise *	0.67	0.23
Do hard exercise *	0.50	–0.10
Walk in nice weather for 0.5 to 1 hour *	0.82	0.29
Walk in bad weather for 0.5 to 1 hour *	0.80	0.18
Run 100 meters *	0.61	0.00
Carry 5 kilos *	0.70	0.26
Wash upper part of body	0.43	0.62
Wash lower part of body	0.46	0.68
Wash hair	0.50	0.52
Dress upper part of body	0.19	0.84
Dress lower part of body	0.22	0.84
Take socks and shoes on and off	0.35	0.75
Comb hair	0.60	0.72
Cut toenails	0.49	0.23
Cut fingernails	0.21	0.59
Chew hard food	0.26	0.20
Eat without help	–0.10	0.67
Read ordinary newspaper text	0.12	0.27
Hear conversation between three or more persons	0.21	0.06

https://doi.org/10.1375/twin.5.4.289 Published online by Cambridge University Press