Regression analysis of multiple outcomes

Annibale Biggeri,1 Elena Stanghellini,2 Mirella Ruggeri3

1 Dipartimento di statistica «G. Parenti», Università di Firenze, viale Morgagni 59, 50134 Firenze
2 Dipartimento di scienze statistiche, Università di Perugia
3 Istituto di psichiatria, Università di Verona

Corrispondenza a: Annibale Biggeri, e-mail: abiggeri@ds.unifi.it

Riassunto

In epidemiologia il controllo dei fattori di confondimento è di solito condotto nella fase di analisi statistica dei dati. Questa prevede la stratificazione (quando il numero di confondenti è limitato) o l’uso di modelli di regressione. Questa strategia tuttavia non è facilmente estendibile al caso in cui si vogliono analizzare più variabili di risposta contemporaneamente. Un esempio è quello dei disturbi respiratori (sintomi asmatici e/o sintomi bronchitici) rispetto a fattori individuali, sociali e ambientali (inquinamenti indoor e/o outdoor), un altro riguarda studi longitudinali sugli esiti (in termini di psicopatologia, disabilità sociale, capacità relazionale) in epidemiologia psichiatrica. Il presente lavoro passa in rassegna i metodi di analisi usati per questo tipo di dati e illustra una nuova classe di modelli di regressione per risposte multivariate.

Introduction

The introduction of multiple regression analyses in Epidemiology became popular in the late seventies and it was justified by the need to take into account more than one determinant in the analysis of observational studies such as prospective cohort or retrospective case-control investigations. Indeed the non experimental feature of the epidemiological research is a strong demand for balanced comparisons between, say, groups exposed and not exposed to be achieved a posteriori by statistical methods. This is not the case in experimental studies where randomization will assure a priori comparability between the treatment (exposed) and control (not exposed) groups. The simplest method of adjustment consists in splitting the data in a series of strata (corresponding to levels of a covariate to be adjusting for) and in summarizing the results of the within strata comparisons; the researcher should check internal consistency of the results by inspecting the within strata estimates and performing statistical test of homogeneity. The main limitation of this approach is that only few covariates can be considered because enough sample size (i.e. number of subjects) should be assured in each stratum. Moreover stratification allows one to evaluate only one determinant at a time. Regression models, whose coefficients can be interpreted as relative effect measures (relative risks, odds ratios or rate ratios), represent a major tool for epidemiologists, in that they allow the simultaneous evaluation of several covariates. There is a strict analogy between stratification and regression analysis: in both we have a response variable and, at least, two explanatory variables, one of them considered the determinant under study and the other a potential confounder or effect modifier. The analysis usually consists in the estimation of the relationship between the determinant and the response conditional to the confounder; in the stratified analysis we obtain an estimate averaging the effect estimates obtained for each discrete value of the confounder (stratum), in the regression analysis the conditioning is carried out by augmentation of the regression equation of a term for the confounder. What it is not immediately clear is that in both instances we evaluate the relationship between the determinant and the response given the joint distribution of the determinant and the confounder; i.e. the correlation among the explanatory variables is always assumed even if not explicitly stated. This fact is highlighted in the following example: suppose we want to study the effect of air pollution on the occurrence of a given disease and we want to adjust for temperature. The neat effect of air pollution can be obtained either by regressing disease frequency on air pollution and temperature using a multiple regression analysis, or by considering as response the variable containing the residuals of the regression of disease frequency on average daily temperature and as determinant the variable containing the residuals of the regression of air pollution on temperature: that is the variation in disease frequency not accounted for temperature is regressed on the variation in air pollution not explained by its association with temperature. We should remember this when explaining graphical models.

In this paper we present an extension of the regression models to more than one response and more than one group of covariates, in a way that purely explanatory and intermediate variables can be modelled simultaneously. While we limit our exemplification to continuous gaussian data, the method discussed is quite general and allows continuous, categorical and a mixture of quantitative and qualitative variables. This topic appears of utmost importance in Psychiatric Epidemiology and in particular in the evaluation of the outcomes of psychiatric care. In fact, in psychiatry the effects of treatments should be evaluated using multiple measures exploring many areas, for example including psychopathology, social disability, quality of life, service satisfaction and service utilisation (1); (2); (3); (see reference 4, 5 for an application to different fields and
variables are relevant. A non-zero regression coefficient of Y if the marginal effects of the explanatory variables on the outcome variables. In principle, it can be justified.*

The first way of studying such a complex situation is to regress each covariate in the model (Coef.), their standard error (Std. Err.), the t-test and its p-value, the confidence intervals of the coefficient are shown.

| Equation | Obs |Parms  |RMSE |“R-sq” |F |P |  
|----------|-----|------|-----|-------|---|---|---
| costs    | 194 |3     | 1.167 |0.2886|38.73 <0.0001 |  
| Coef.    | Std. Err.|t |P>|t| |[95% Conf.Interval]|  
| costs    |    |  |  |   |   |   |  
| base GAF | -0.036 |0.006 | -5.829 <0.001 | -0.049 -0.024 |  
| base DAS | 0.737 |0.313 | 2.535 0.200 | 0.119 1.355 |  

### B- Outcome GAF and DAS vs Total costs, baseline GAF and DAS

| Equation | Obs |Parms  |RMSE |“R-sq” |F |P |  
|----------|-----|------|-----|-------|---|---|---
| outcome GAF | 194 |4 | 11.432 |0.4784|58.099 <0.0001 |  
| Coef.    | Std. Err.|t |P>|t| |[95% Conf INTERVAL]|  
| outcome GAF |    |  |  |   |   |   |  
| base GAF | 0.479 |0.066 | 7.202 <0.001 | 0.348 0.610 |  
| base DAS | -5.063 |3.113 | -1.626 0.105 | -11.184 1.059 |  
| costs | -2.472 |0.709 | -3.487 0.001 | -3.865 -1.078 |  

### Outcome DAS

| Equation | Obs |Parms  |RMSE |“R-sq” |F |P |  
|----------|-----|------|-----|-------|---|---|---
| base GAF | -0.008 |0.002 | -4.013 <0.001 | -0.011 -0.004 |  
| base DAS | 0.345 |0.088 | 3.933 <0.001 | 0.172 0.517 |  
| costs | 0.057 |0.020 | 2.841 0.005 | 0.017 0.096 |  

Correlation between outcome GAF and outcome DAS: -0.6388

**Table 1. Separate univariate regression approach.**

A: total costs of psychiatric care as function of baseline Global Functioning Status (GAF) and Social Disability (DAS). B: outcome GAF and DAS as function of baseline GAF, DAS and total costs.

The dependent variable in the regression equation, the number of subjects (Obs), the number of terms in the regression equation (Parms), residual mean square error (RMSE), coefficient of determination (R-sq), F test and p-value are reported in the first part; in the second part the regression coefficient for each covariate in the model (Coef.), their standard error (Std. Err.), the t-test and its p-value, the confidence intervals of the coefficient are shown.

6 for a general discussion of graphical models and causal relationships. A glossary with the more unusual or specialistic terms has been added at the end of the paper.

**The univariate approach:**

**a simple reductionistic but false approach**

We begin our discussion by considering situations where the set of variables can be grouped into two separate sets: the explanatory variables and the outcome variables. As usual, we indicate with Xi the i-th explanatory variable and with Yj the j-th outcome variable. The first way of studying such a complex situation is to regress each outcome variable separately on the whole set of the explanatory variables. This approach deliberately disregards potential associations among the response variables. In principle, it can be justified if the marginal effects of the explanatory variables on the outcome variables are relevant. A non-zero regression coefficient of Yj against each Xi expresses a dependency between Yj and Xi given all the other explanatory variables. A zero value implies independency among the couple conditioning on the value of other explanatory variables. However, depending on the nature of the dependencies among the variables, this approach can be misleading. As the regression coefficient of the dependent variable Yj against each explanatory variable is the sum of direct and indirect effects (7), the indirect effects being via the other outcome variables not included in the model, there can be situations where this marginal coefficient is not significantly different from zero even when each term of the sum is. Viceversa, there can be situations where the marginal coefficient is significantly different from zero, only due to correlations induced by marginalising over the other outcome variables.

**Example**

We use as an example a subset of data derived from the South-Verona Outcome Project. This is a naturalistic, longitudinal study aimed at assessing the outcome of care provided by a Community Mental Health Service (8); (9). More extensive analyses of these data can be found in (1) and a complete analysis based on graphical chain models is in (2).

Patients were followed up after a baseline assessment of each subject of his/her Global Functioning Status (GAF) and Social Disability (DAS). During the follow-up total costs of psychiatric care were measured and at the end of the study period each subject was again evaluated for GAF and DAS. In this framework we have two response variables: outcome GAF and outcome DAS, measured at the end of follow-up; one intermediate variable: total costs; and two predictors: baseline GAF and DAS, assessed at the start of the follow-up time. An intuitive simple way to analyze such data is to conduct a series of multiple regressions: first total costs on baseline GAF and DAS, then the outcomes GAF and, separately, DAS on all the predictors. As we can see from table 1 baseline DAS appeared not significantly different from zero even when each predictor was included in the model, there can be situations where the marginal coefficient is significantly different from zero, only due to correlations induced by marginalising over the other outcome variables.
gressions, just make several independent analyses. It is reductionistic because each response is isolated from the other and any relationship among them is not considered in the modelling phase. Some more sophisticated data analysts could evaluate the correlation among the responses using the residuals of the separate regressions (see bottom of table 1). When there is a very low correlation among the responses and we are considering different unrelated outcome phenomena this approach could be justified, but otherwise two situations could arise. First, each response depends on a different unrelated set of predictors: in this case the simple approach described is less efficient and the precision of the estimates is diminished. Second, some predictors could be relevant to more than one response: in this case not only efficiency but, even more important, accuracy of the estimates is affected, as we will show below.

**A more complex multiple regression approach**

A better way to analyse the above situation is to perform a series of univariate regressions in which each outcome variable $Y_j$ is regressed against both the explanatory variables and the other outcome variables. Though this method can be quite complicated, the advantage consists in the fact that the marginal coefficient of the univariate regression of $Y_j$ against $X_i$ is decomposed into its terms, therefore permitting a separated evaluation of the significance of each of them. Another advantage is in the interpretation of the regression coefficient of the outcome variables against the explanatory variables, representing a conditional dependency given all the other variables in the model. A zero value therefore implies independence among the two variables conditional to all the other variables in the model.

**Example**

In our example the outcome GAF is regressed on the other outcome DAS plus the three predictors, total costs baseline GAF and DAS and likely for the outcome DAS. The results show that DAS is predicted only by baseline level of DAS, which is a very different conclusion from that obtained in the previous simple analysis (table 2).

However, the above approach also has some inefficiencies due to the inclusions in each regression of some non-significant relationships. For instance, in regressing DAS against all the other variables we are forced to assume that a relationship between the outcome GAF and baseline DAS holds. This contradicts the findings of regressing GAF against all the other variables. As it will be shown in the next Section, the two variables are conditionally independent given the other two. The consequence is that the above sequence of regressions does not allow us to take explicitly into account this independence, therefore causing a loss of efficiency of our estimates.

**Modelling the concentration matrices**

The next step is to investigate the structure of independencies in the joint distribution of the variables. We will consider the causal ordering among the variables, such that some variables are explanatory of other outcome variables. According to such ordering, the joint distribution $f(X,Y)$ factorises into the product $f(X)f(Y|X)$. Then we will investigate the structure of independencies in the joint distribution of the $X$ variables first, and then the structure of independencies in the joint distribution of the $Y$ variables conditioning on $X$.

To summarise information concerning the independencies/associations among the variables we use a *chain graph* model. As a chain graph model is univocally determined by its graph (that is a pictorial representation of the qualitative structure of the relation-
ships among the variables) we will make no distinction among the model and the graph. In a chain graph, the variables are represented by nodes or vertices. The nodes/variables are included into boxes, according to their nature of explanatory or outcome variables. In the figure 1 two outcomes (Y) are represented as nodes in a box and three predictors (X) as nodes within another box. More generally, we will assume that the nodes/variables can be divided into subsets and that these subsets can be ordered in a way that all the variables in a subset are potentially explanatory for the variables in the following subsets. Variables in the same box are considered to be on the same footing, that is either the relationships among them do not exhibit a clear ordering or the order is not relevant for the analysis. The ordering among the boxes (here taken from left to right) reflects the causal order, and should be specified by a priori knowledge. It is important that such an order does not contain cycles, that is, is not allowed for a variable to be (directly or indirectly) self-explanatory. In this sense a chain graph differs from structural equation models (such as LISREL models, (10)), where cycles among the variables are allowed. In particular it is not always possible to interpret the coefficients of the relationships represented in the LISREL models as regression coefficients, i.e. changes in the response(s) for unit change in the explanatory variable(s) keeping all the other explanatory variables constant.

In a chain graph, some pairs of variables are joined by edges. Edges are undirect among variables in the same box and direct (i.e. arrows) otherwise, reflecting the causal ordering. An edge missing among a couple of variables in the first box (that is, the one with only explanatory variables) should imply a conditional independence among the two variables giving all the variables in the first box (in figure 1 X1 is independent from X2 given X3). If the box is not the first one, the variables in all the preceding boxes should be added to the conditioning set. A missing arrow between one variable in the first box and another in a later box implies that the couple are independent conditioning on all the variables in the later box and in the box(es) preceding (Y1 is independent from X1 given Y2 and X2, X3). These criteria can be sharpened, in the sense that the conditioning set can, in principle, be reduced further, depending on the structure of the graph. The interested reader is referred to reference 11 (Ch. 8).

To begin with, we study the conditional independence in the X variables. One way to characterize the notion of conditional independence is via the Factorisation Theorem (see reference 12, pag. 4). Let X, X1, X2, a partition of X. If the joint distribution of X factorises into the product of two terms g(X1, X2) h(X3) then X2 is independent from X conditioning on X1 and vice versa. Application of the above theorem leads to the following characterisation for jointly Gaussian variables. Let Σ be the covariance matrix of a set X of jointly Gaussian variables. Let Σ1 be the concentration matrix and σ the generic entry in the matrix. As a consequence of the factorisation theorem, Xi is independent from X conditioning on all the other variables if and only if σ = 0. Let’s recall the relationship between σ and ρ, the partial correlation coefficient between X1, X2 given all the other variables (see reference 13, Ch. 5, for details): ρ = σ / (σσ)1/2.

Therefore, studying the structure of zeros in the concentration matrix allows us to elicit conditional independence statements. The conditional independence graph in the first box will be such that an edge between two variables will be missing whenever the corresponding entry in the concentration matrix of the variables in such a box is zero.

More generally, the study of the structure of independencies in a chain graph for jointly Gaussian random variables, may be carried out via the analysis of a sequence of concentration matrices of opportune defined joint and conditional distributions. For instance, to analyse the independence structure between a response variable,
say $Y_1$, and the $X$ variables we investigate the corresponding entries in the concentration matrix of $(Y_1, X)$; to analyse the independencies between two response variables, say $Y_1$ and $Y_2$, given the $X$ variables and all the other response variables $Y \setminus \{Y_1, Y_2\}$ we investigate the corresponding entry in the concentration matrix of $(Y, X)$. Again, to understand this step let’s recall the relationship between $\beta_{ji}$, the regression coefficient of $X_j$ on $X_i$, given all the other predictive variables, and $\sigma^2$ (see reference 13, Ch. 5 for details): $\beta_{ji} = -\sigma^2 / \sigma^2_{jj}$ (equation 1).

Again, we stress the point that, depending on the structure of the graph, efficient estimates of such regression coefficients may be performed only by jointly estimating the correlations among the variables, therefore using results for ML estimates of $\Sigma^{-1}$ for an arbitrary structure of zeros. These are summarised in (13) or (12) and implemented in the computer program MIM (12). Also the theory of testing for presence/absence of an edge have been developed via the formalisation of likelihood ratio test.

Example

Let’s return to our example: now we represent the relationships among the five variables (baseline GAF and DAS, total costs, outcome GAF and DAS) in the chain graph of figure 2.

The analysis proceeds as follows:

Step 1: Relationships in the first box. The empirical (observed) covariance, concentration and correlation matrices for the two variables base GAF and base DAS, are shown in table 3. It is evident that the two variables are not marginally independent. Sometimes, as the variables in the first box are merely exogenous, (considered as fixed by design) this step is omitted.

Step 2: Relationships between the response “total costs of psychiatric care” and the two predictors base GAF and DAS. The empirical (observed) concentration matrix for the three variables considered is summarised in table 4. The reader may check the consistency of these analyses with the regression analysis of costs against base GAF and base DAF highlighted by equation (1) and with the regression analysis reported in table 1. Actually the coefficient for base GAF is obtained from the matrix of concentrations (table 4): the off diagonal value for base GAF / Costs is equal to 0.027 and the diagonal value for Costs is 0.746; from the formula reported before ($\beta_{ji} = -\sigma^2 / \sigma^2_{jj}$) the regression coefficient becomes $\beta_{\text{base GAF}/\text{Costs}} = -0.027/0.746 = -0.036$ (see table 1). The same calculation for baseline DAS gives $-0.55/0.746 = 0.737$ (a value identical to that reported in table 1).

Step 3: Relationships among the variables in the third box and dependencies between the variables in the third box on the variables in the first and second boxes. The empirical (observed) concentration matrix for the five variables considered, GAF and DAS at baseline, total costs and GAF and DAS outcomes, is shown in the table 5a together with the partial correlation matrix. It is evident inspecting the last matrix that outcome DAS is independent from baseline GAF (r=0.049) and from total costs (r=0.056) keeping constant the other variables, while outcome GAF is independent from baseline DAS (r=0.080), again keeping constant the other variables.

We point out that the entries of the last matrix are the partial correlations coefficients given all the other variables, i.e. -0.325 is the correlation coefficient of base GAF and base DAS given all the other variables, including the ones in the later boxes. We already said that the ordering among the variables renders this coefficient not relevant for the analysis. On the contrary, of great relevance is, for example, the correlation coefficient –0.639 between outcome GAF and outcome DAS giving all the other variables, as these variables are all in the previous boxes. The meaning of this coefficient is that there is a residual association between the two variables also after taking into account the associations induced by the common regressors.

The modelling strategy consists in setting to zero the appropriate entries of the concentration matrix. The results in table 5b show the fitted concentration and partial concentration matrices. The model has a good fit (deviance 2.1 with 3 df, p=0.55) and the regression coefficients for the two outcomes GAF and DAS are reported in table 6.

The results highlight the gain in efficiency of this approach: the standard errors of the estimates are systematically lower than those obtained by separate multiple regressions. Moreover they are more accurate since non relevant association among variables are dropped. Indeed here we point out that for a particular class of model the proposed methodology is the most efficient. These are the non decomposable models. With this expression we denote the relationships among several variables characterized by chordless cycles in the graph: the conditional independence structure cannot be expressed in the right box the outcomes (GAF and DAS). The reader may check the fitted concentration and partial concentration matrices.

Figure 2. Graphical Chain Blocks representation of conditional dependencies among several variables. In the left box the predictors (baseline GAF and DAS), in the middle box the intermediate process variable (total cost of psychiatric care), in the right box the outcomes (GAF and DAS). The lines denote a correlation between variables, the arrows denote a dependence relationship between two variables.

Table 6. Estimate regression coefficients using the concentration matrix modelling approach.

| Variable | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|----------|-------|-----------|---|-------|-------------------|
| outcome GAF | base GAF | 0.308 | 0.045 | 6.7800 | <0.0001 | 0.219 | 0.397 |
| costs | -1.462 | 0.433 | -3.3756 | 0.0007 | -2.310 | -0.613 |
| outcome DAS | base GAF | 0.219 | 0.053 | 4.0939 | <0.0001 | 0.114 | 0.323 |
pressed in any simple form by separate regression analyses. For example, the conditional dependencies among GAF and DAS at baseline and at the end of the follow-up are represented by a graph with four nodes and four edge but no inner chords, outcome GAF being conditionally independent from baseline DAS and outcome DAS from baseline GAF. To model the relationships between baseline and outcome GAF, baseline and outcome DAS taking into account the correlation between the two outcomes requires special methods. Regressing one outcome (e.g. GAF) at a time on the appropriate predictor (baseline GAF) and the other outcome (DAS) would imply the existence of a relationship between outcome DAS and baseline GAF, which is false. It is clear now that the inefficiency of the regression strategy comes from the assumption of unnecessary associations.

Model criticism and sensitivity analysis
When dealing with continuous variables it is assumed that a multivariate gaussian distribution will apply. An important consequence of this would be the absence of non linearity and second order interactions among variables. This is usually done when building up a regression model by checking: 1- if there is an important non linear effect of the covariate of interests; 2- if the effect of a covariate is modified by another covariate.

Example
In the multivariate analysis of global functioning, social disability and total costs of psychiatric care we have to model the relationships among five variables. Regressing each variable in turn on any other plus the respective quadratic term we have a total of 20 (5 times 4) tests; to avoid the problem of multiple comparisons a graphical assessment has been suggested by (14). The ordered values of each test are then plotted against the quantiles of the reference distribution: deviations from the straight line indicate non linearity for some variables (Q-Q plot: figure 3). According to most outlying tests, baseline DAS appeared to be non linearly related to baseline and outcome GAF. Lower values of social disability seemed not to predict global functioning at outcome. The presence of second order interactions are evaluated regressing each variables against each pair of covariates plus their interaction term. A total of 30 (5 times the number of combination of 2 out of 4) tests are performed and plotted against the quantile of the reference distribution. The most deviating tests regarded the association between baseline DAS and baseline and outcome GAF modified by outcome DAS (figure not shown). The reader should note that we have checked interactions in regression models with only two regressors: maybe there exists a relationship between baseline DAS and outcome GAF modified by outcome DAS but this interaction disappears when considering a model with four variables (baseline GAF and DAS; outcome GAF and DAS). In summary, it seems reasonable to assume absence of important non linearity in the relationships among the variables analyzed, since the outlying squared and interaction terms highlighted in the Q-Q plot are effects of the complex pattern of dependencies among the four variables baseline and outcome GAF and DAS (see the 4-cycle chordless graph of conditional independence in figure 2).

In any case the Q-Q plot results would suggest some model inadequacy: the analysis should be replicated either after having transformed baseline DAS (e.g. by taking logarithms) or after having categorized it and the results of these different strategies should be presented and discussed. A further problem consists in the effect of asymmetry on the regression coefficient estimates. This could be avoided through some data transformation (e.g. logarithmic) but this strategy could result in less interpretable regression coefficients. Anyway it is important to do some sensitivity analysis, for example replicating the results using Box-Cox transformed data or after deleting some outlying or influential observations.
Conclusions
The multidimensional nature of the outcome of psychiatric care implies a strong demand for more sophisticated methods of analysis. The familiar regression approach has been widely used in the epidemiological literature to evaluate complex dependencies among one response and several predictors. The natural extension of such an approach to multivariate responses has been illustrated using quantitative variables. The methods proposed are applied to discrete categorical variables as well as to a mixture of discrete and continuous ones. Moreover the strategy outlined allows the researcher to design a series of regression analyses to cope with sets of naturally ordered variables (predictors, intermediate and response variables) in a sensible way. Last but not least the one-to-one correspondence between a model and its graphical representation facilitates the interpretation and the interfacing of researchers from different disciplines.

Glossary

**Adjusted effect**: a relationship between two variables not accounted for the effect of other variables (confounders). It can be called a relationship conditional to the set of variable considered.

**Confounder**: a strong predictor of a response which is associated with the determinant under study.

**Concentration**: an entry of the inverse of the covariance matrix; it is also called precision following the usual meaning of precision as the reciprocal of the variance.

**Effect modifier**: a variable which changes the magnitude and/or the sign of the relationship between a response and a determinant.

**Graph**: a pictorial representation where the qualitative structure of the relationships among variables is visually represented.

**Intermediate variable**: a variable involved in the pathway from a cause and an effect; a cause can be masked by an intermediate variable, such as sexual life styles are not risk factors for AIDS given the intermediate variable seropositivity to HIV; or not, like sunburns which are risk factors for malignant melanoma directly and indirectly via atypical nevi (intermediate variable).

**Marginal effect**: a crude relationship between two variables (e.g. a response and a determinant) not taking into account other relevant variables (confounders or effect modifiers).

**Partial correlation**: a correlation conditional to all the other variables in the model; it reflects the amount of variability shared by two variables not accountable to associations with other covariates.

Bibliography