

## Chapter 5

# Multiple correlation and multiple regression

The previous chapter considered how to determine the relationship between two variables and how to predict one from the other. The general solution was to consider the ratio of the covariance between two variables to the variance of the predictor variable (*regression*) or the ratio of the covariance to the square root of the product the variances (*correlation*). This solution may be generalized to the problem of how to predict a single variable from the weighted linear sum of multiple variables (*multiple regression*) or to measure the strength of this relationship (*multiple correlation*). As part of the problem of finding the weights, the concepts of *partial covariance* and *partial correlation* will be introduced. To do all of this will require finding the variance of a composite score, and the covariance of this composite with another score, which might itself be a composite.

Much of psychometric theory is merely an extension, an elaboration, or a generalization of these concepts. Almost all tests are composites of items or subtests. An understanding how to decompose test variance into its component parts, and conversely, an understanding how to analyze tests as composites of items, allows us to analyze the meaning of tests. But tests are not merely composites of items. Tests relate to other tests. A deep appreciation of the basic Pearson correlation coefficient facilitates an understanding of its generalization to multiple and partial correlation, to factor analysis, and to questions of validity.

### 5.1 The variance of composites

If  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are vectors of  $N$  observations centered around their mean (that is, deviation scores) their variances are  $V_{x_1} = \sum x_{i1}^2 / (N - 1)$  and  $V_{x_2} = \sum x_{i2}^2 / (N - 1)$ , or, in matrix terms  $V_{x_1} = \mathbf{x}'_1 \mathbf{x}_1 / (N - 1)$  and  $V_{x_2} = \mathbf{x}'_2 \mathbf{x}_2 / (N - 1)$ . The variance of the composite made up of the sum of the corresponding scores,  $\mathbf{x} + \mathbf{y}$  is just

$$V_{(\mathbf{x}_1 + \mathbf{x}_2)} = \frac{\sum (x_i + y_i)^2}{N - 1} = \frac{\sum x_i^2 + \sum y_i^2 + 2\sum x_i y_i}{N - 1} = \frac{(\mathbf{x} + \mathbf{y})'(\mathbf{x} + \mathbf{y})}{N - 1}. \quad (5.1)$$

Generalizing 5.1 to the case of  $n$  xs, the composite matrix of these is just  ${}_N \mathbf{X}_n$  with dimensions of  $N$  rows and  $n$  columns. The matrix of variances and covariances of the individual items of this composite is written as  $\mathbf{S}$  as it is a sample estimate of the population variance-covariance matrix,  $\Sigma$ . It is perhaps helpful to view  $\mathbf{S}$  in terms of its elements,  $n$  of which are variances

and  $n^2 - n = n * (n - 1)$  are covariances:

$$\mathbf{S} = \begin{pmatrix} v_{x1} & c_{x1x2} & \cdots & c_{x1xn} \\ c_{x1x2} & v_{x2} & & c_{x2xn} \\ \vdots & & \ddots & \vdots \\ c_{x1xn} & c_{x2xn} & \cdots & v_{xn} \end{pmatrix}$$

The diagonal of  $\mathbf{S} = \text{diag}(\mathbf{S})$  is just the vector of individual variances. The *trace* of  $\mathbf{S}$  is the sum of the diagonals and will be used a great deal when considering how to estimate reliability. It is convenient to represent the sum of all of the elements in the matrix,  $\mathbf{S}$ , as the variance of the composite matrix.

$$V_{\mathbf{X}} = \sum \frac{\mathbf{X}'\mathbf{X}}{N-1} = \frac{\mathbf{1}'(\mathbf{X}'\mathbf{X})\mathbf{1}}{N-1}.$$

## 5.2 Multiple regression

The problem of the optimal linear prediction of  $\hat{\mathbf{y}}$  in terms of  $\mathbf{x}$  may be generalized to the problem of linearly predicting  $\hat{\mathbf{y}}$  in terms of a composite variable  $\mathbf{X}$  where  $\mathbf{X}$  is made up of individual variables  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ . Just as  $b_{y \cdot x} = \text{cov}_{xy} / \text{var}_x$  is the optimal slope for predicting  $y$ , so it is possible to find a set of weights ( $\beta$  *weights* in the standardized case,  $b$  *weights* in the unstandardized case) for each of the individual  $\mathbf{x}_i$ s.

Consider first the problem of two predictors,  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , we want to find the find weights,  $b_i$ , that when multiplied by  $\mathbf{x}_1$  and  $\mathbf{x}_2$  maximize the covariances with  $\mathbf{y}$ . That is, we want to solve the two simultaneous equations

$$\begin{cases} v_{x1}b_1 + c_{x1x2}b_2 = c_{x1y} \\ c_{x1x2}b_1 + v_{x2}b_2 = c_{x2y} \end{cases}.$$

or, in the standardized case, find the  $\beta_i$ :

$$\begin{cases} \beta_1 + r_{x1x2}\beta_2 = r_{x1y} \\ r_{x1x2}\beta_1 + \beta_2 = r_{x2y} \end{cases}. \quad (5.2)$$

We can directly solve these two equations by adding and subtracting terms to the two such that we end up with a solution to the first in terms of  $\beta_1$  and to the second in terms of  $\beta_2$ :

$$\begin{cases} \beta_1 = r_{x1y} - r_{x1x2}\beta_2 \\ \beta_2 = r_{x2y} - r_{x1x2}\beta_1 \end{cases} \quad (5.3)$$

Substituting the second row of (5.3) into the first row, and vice versa we find

$$\begin{cases} \beta_1 = r_{x1y} - r_{x1x2}(r_{x2y} - r_{x1x2}\beta_1) \\ \beta_2 = r_{x2y} - r_{x1x2}(r_{x1y} - r_{x1x2}\beta_2) \end{cases}$$

Collecting terms and rearranging :

$$\begin{cases} \beta_1 - r_{x_1x_2}^2\beta_2 = r_{x_1y} - r_{x_1x_2}r_{x_2y} \\ \beta_2 - r_{x_1x_2}^2\beta_1 = r_{x_2y} - r_{x_1x_2}r_{x_1y} \end{cases}$$

leads to

$$\begin{cases} \beta_1 = (r_{x_1y} - r_{x_1x_2}r_{x_2y}) / (1 - r_{x_1x_2}^2) \\ \beta_2 = (r_{x_2y} - r_{x_1x_2}r_{x_1y}) / (1 - r_{x_1x_2}^2) \end{cases} \quad (5.4)$$

Alternatively, these two equations (5.2) may be represented as the product of a vector of unknowns (the  $\beta$ s) and a matrix of coefficients of the predictors (the  $r_{xi}$ s) and a matrix of coefficients for the criterion ( $r_{xiy}$ ):

$$(\beta_1 \beta_2) \begin{pmatrix} r_{x_1x_1} & r_{x_1x_2} \\ r_{x_1x_2} & r_{x_2x_2} \end{pmatrix} = (r_{x_1y} \quad r_{x_2y}) \quad (5.5)$$

If we let  $\beta = (\beta_1 \beta_2)$ ,  $\mathbf{R} = \begin{pmatrix} r_{x_1x_1} & r_{x_1x_2} \\ r_{x_1x_2} & r_{x_2x_2} \end{pmatrix}$  and  $\mathbf{r}_{xy} = (r_{x_1y} \quad r_{x_2y})$  then equation 5.5 becomes

$$\beta \mathbf{R} = \mathbf{r}_{xy} \quad (5.6)$$

and we can solve Equation 5.6 for  $\beta$  by multiplying both sides by the inverse of  $\mathbf{R}$ .

$$\beta = \beta \mathbf{R} \mathbf{R}^{-1} = \mathbf{r}_{xy} \mathbf{R}^{-1} \quad (5.7)$$

Similarly, if  $\mathbf{c}_{xy}$  represents the covariances of the  $\mathbf{x}_i$  with  $\mathbf{y}$ , then the  $\mathbf{b}$  weights may be found by

$$\mathbf{b} = \mathbf{c}_{xy} \mathbf{S}^{-1}$$

and thus, the predicted scores are

$$\hat{\mathbf{y}} = \beta \mathbf{X} = \mathbf{r}_{xy} \mathbf{R}^{-1} \mathbf{X}. \quad (5.8)$$

The  $\beta_i$  are the *direct effects* of the  $\mathbf{x}_i$  on  $\mathbf{y}$ . The *total effects* of  $\mathbf{x}_i$  on  $\mathbf{y}$  are the correlations, the *indirect effects* reflect the product of the correlations between the predictor variables and the direct effects of each predictor variable.

Estimation of the  $\mathbf{b}$  or  $\beta$  vectors, with many diagnostic statistics of the quality of the regression, may be found using the `lm` function. When using categorical predictors, the linear model is also known as *analysis of variance* which may be done using the `anova` and `aov` functions. When the outcome variables are dichotomous, *logistic regression* using the *generalized linear model* function `glm` and a binomial error function. A complete discussion of the power of the generalized linear model is beyond any introductory text, and the interested reader is referred to e.g., Cohen et al. (2003); Dalgaard (2002); Fox (2008); Judd and McClelland (1989); Venables and Ripley (2002).

Diagnostic tests of the regressions, including plots of the residuals versus estimated values, tests of the normality of the residuals, identification of highly weighted subjects are available as part of the graphics associated with the `lm` function.

### 5.2.1 Direct and indirect effects, suppression and other surprises

If the predictor set  $vecx_i, x_j$  are uncorrelated, then each separate variable makes a unique contribution to the dependent variable,  $y$ , and  $R^2$ , the amount of variance accounted for in  $y$ , is the sum of the individual  $r^2$ . In that case, even though each predictor accounted for only 10% of the variance of  $y$ , with just 10 predictors, there would be no unexplained variance. Unfortunately, most predictors are correlated, and the  $\beta$ s found in 5.5 or 5.7 are less than the original correlations and since

$$R^2 = \sum \beta_i r_{x_i y} = \beta' r_{xy}$$

the  $R^2$  will not increase as much as it would if the predictors were less or not correlated.

An interesting case that occurs infrequently, but is important to consider, is the case of *suppression*. A *suppressor* does not correlate with the criterion variable, but, because it does correlate with the other predictor variables, removes variance from those other predictor variables. This has the effect of reducing the denominator in equation 5.5 and thus increasing the  $\beta_i$  for the other variables. Consider the case of two predictors of stock broker success: self reported need for achievement and self reported anxiety (Table 5.1). Although Need Achievement has a modest correlation with success, and Anxiety has none at all, adding Anxiety into the multiple regression increases the  $R^2$  from .09 to .12. An explanation for this particular effect might be that people who want to be stock brokers are more likely to say that they have high Need Achievement. Some of this variance is probably legitimate, but some might be due to a tendency to fake positive aspects. Low anxious scores could reflect a tendency to fake positive by denying negative aspects. But those who are willing to report being anxious probably are anxious, and are telling the truth. Thus, adding anxiety into the regression removes some misrepresentation from the Need Achievement scores, and increases the multiple  $R^2$

### 5.2.2 Interactions and product terms: the need to center the data

In psychometric applications, the main use of regression is in predicting a single criterion variable in terms of the linear sums of a predictor set. Sometimes, however, a more appropriate model is to consider that some of the variables have multiplicative effects (i.e., interact) such the effect of  $x$  on  $y$  depends upon a third variable  $z$ . This can be examined by using the product terms of  $x$  and  $z$ . But to do so and to avoid problems of interpretation, it is first necessary to *zero center* the predictors so that the product terms are not correlated with the additive terms. The default values of the `scale` function will center as well as standardize the scores. To just center a variable,  $x$ , use `scale(x, scale=FALSE)`. This will preserve the units of  $x$ . `scale` returns a matrix but the `lm` function requires a data.frame as input. Thus, it is necessary to convert the output of `scale` back into a data.frame.

A detailed discussion of how to analyze and then plot data showing interactions between experimental variables and subject variables (e.g., manipulated positive affect and extraversion) or interactions of subject variables with each other (e.g., neuroticism and extraversion)

---

<sup>1</sup> Although the correlation values are enhanced to show the effect, this particular example was observed in a high stakes employment testing situation.

**Table 5.1** An example of suppression is found when predicting stockbroker success from self report measures of need for achievement and anxiety. By having a suppressor variable, anxiety, the multiple R goes from .3 to .35.

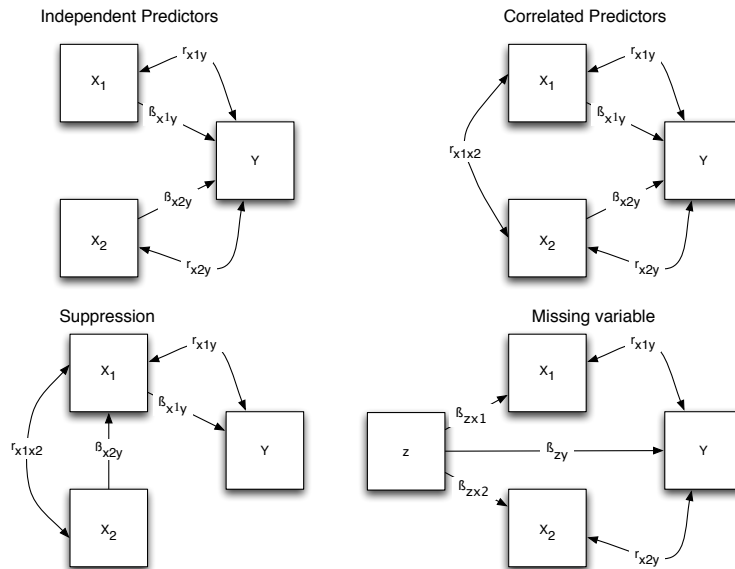
```
> stock
> mat.regress(stock,c(1,2),3)

      Nach Anxiety Success
achievement  1.0   -0.5   0.3
Anxiety     -0.5    1.0   0.0
Success      0.3    0.0   1.0

$beta
  Nach Anxiety
0.4    0.2

$R
Success
0.35

$R2
Success
0.12
```



**Fig. 5.1** There are least four basic regression cases: The independent predictor where the  $\beta_i$  are the same as the correlations; the normal, correlated predictor case, where the  $\beta_i$  are found as in 5.7; the case of suppression, where although a variable does not correlate with the criterion, because it does correlate with a predictor, it will have useful  $\beta_i$  weight; and the case where the model is misspecified and in fact a missing variable accounts for the correlations.

is beyond the scope of this text and is considered in great detail by Aiken and West (1991) and Cohen et al. (2003), and in less detail in an online appendix to a chapter on experimental approaches to personality Revelle (2007), <http://personality-project.org/r/simulating-personality.html>. In that appendix, simulated data are created to show additive and interactive effects. An example analysis examines the effect of Extraversion and a movie induced mood on positive affect. The regression is done using the `lm` function on the centered data (Table 5.2). The graphic display shows two regression lines, one for the simulated “positive mood induction”, the other for a neutral induction.

**Table 5.2** Linear model analysis of simulated data showing an interaction between the personality dimension of extraversion and a movie based mood induction. Adapted from Revelle (2007).

```
> # a great deal of code to simulate the data
> mod1 <- lm(PosAffect ~ extraversion*reward,data = centered.affect.data) #look for interactions
> print(summary(mod1,digits=2))
```

Call:

```
lm(formula = PosAffect ~ extraversion * reward, data = centered.affect.data)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-2.062 -0.464  0.083  0.445  2.044
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.8401    0.0957   -8.8    6e-14 ***
extraversion    -0.0053    0.0935   -0.1     0.95
reward1         1.6894    0.1354   12.5   <2e-16 ***
extraversion:reward1  0.2529    0.1271    2.0    0.05 *
```

---

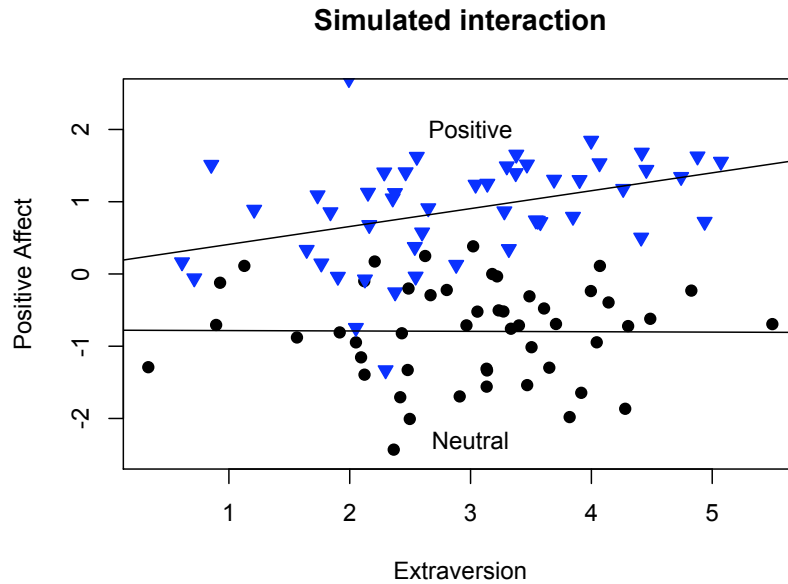
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.68 on 96 degrees of freedom
Multiple R-squared:  0.63,    Adjusted R-squared:  0.62
F-statistic: 54 on 3 and 96 DF,  p-value: <2e-16
```

### 5.2.3 Confidence intervals of the regression and regression weights

The multiple correlation finds weights to best fit the particular sample. Unfortunately, it is biased estimate of the population values. Consequently, the value of  $R^2$  is likely to shrink when applied to another sample. Standard estimates for the amount of *shrinkage* consider the size of the sample as well as the number of variables in the model. For  $N$  subjects and  $k$  predictors, estimated  $R^2$ ,  $\tilde{R}^2$ , is

$$\tilde{R}^2 = 1 - (1 - R^2) \frac{N-1}{N-k-1}.$$



**Fig. 5.2** The (simulated) effect of extraversion and movie induced mood on positive affect. Adapted from Revelle (2007). Detailed code for plotting interaction graphs is available in the online appendix to that article.

The *confidence interval of  $R^2$*  is, of course, a function of the variance of  $R^2$  which is (taken from Cohen et al. (2003) and Olkin and Finn (1995))

$$SE_{R^2}^2 = \frac{4R^2(1-R^2)(N-k-1)^2}{(N^2-1)(N+3)}.$$

Because multiple R is partitioning the observed variance into modeled and residual variance, testing the hypothesis that the multiple R is zero may be done by analysis of variance and leads to an F ratio with k and (N-k-1) degrees of freedom:

$$F = \frac{R^2(n-k-1)}{(1-R^2)k}.$$

The standard errors of the beta weights is

$$SE_{\beta_i} = \sqrt{\frac{1-R^2}{(N-k-1)(1-R_i^2)}}$$

where  $R_i^2$  is the multiple correlation of  $\mathbf{x}_i$  on all the other  $\mathbf{x}_j$  variables. (This term, the *squared multiple correlation*, is used in estimating communalities in factor analysis, see 6.2.1. It may be found by the `smc` function.).

### 5.2.4 Multiple regression from the covariance/correlation matrix

Using the raw data allows for error diagnostics and for the inclusion of interaction terms. But since Equation 5.7 is expressed in terms of the correlation matrix, the regression weights can be found from the correlation matrix. This is particularly useful if one does not have access to the raw data (e.g., when reanalyzing a published study), or if the correlation matrix is synthetically constructed. The function `mat.regress` allows one to extract subsets of variables (predictors and criteria) from a matrix of correlations and find the multiple correlations and beta weights of the x set predicting each member of the y set.

### 5.2.5 The robust beauty of linear models

Although the  $\beta$  weights 5.7 are the optimal weights, it has been known since Wilks (1938) that differences from optimal do not change the result very much. This has come to be called “the robust beauty of linear models” Dawes and Corrigan (1974); Dawes (1979) and follows the principal of “it don’t make no nevermind” Wainer (1976). That is, for standardized variables predicting a criterion with  $.25 < \beta < .75$ , setting all  $\beta_i = .5$  will reduce the accuracy of prediction by no more than 1/96th. Thus the advice to standardize and add. (Clearly this advice does not work for strong negative correlations, but in that case standardize and subtract. In the general case weights of -1, 0, or 1 are the robust alternative.)

A graphic demonstration of how a very small reduction in the  $R^2$  value can lead to an infinite set of “fungible weights” that are all equally good in predicting the criterion is the paper by Waller (2008) with associated R code. This paper reiterates the skepticism that one should have for the interpretability of any particular pattern of  $\beta$  weights.

## 5.3 Partial and semi-partial correlation

Given three or more variables, an interesting question to ask is what is the relationship between  $\mathbf{x}_i$  and  $\mathbf{y}$  when the effect of  $\mathbf{x}_j$  has been removed? In an experiment it is possible to answer this by forcing  $\mathbf{x}_i$  and  $\mathbf{x}_j$  to be independent by design. Then it is possible to decompose the variance of  $\mathbf{y}$  in terms of effects of  $\mathbf{x}_i$  and  $\mathbf{x}_j$  and possibly their interaction. However, in the correlational case, it is likely that  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are correlated. A solution is to consider linear regression to predict  $\mathbf{x}_i$  and  $\mathbf{y}$  from  $\mathbf{x}_j$  and to correlate the residuals. That is, we know from linear regression that it is possible to predict  $\mathbf{x}_i$  and  $\mathbf{y}$  from  $\mathbf{x}_j$ . Then the correlation of the residuals  $\mathbf{x}_{i.} = \mathbf{x}_i - \hat{\mathbf{x}}_i$  and  $\mathbf{y}_{.j} = \mathbf{y} - \hat{\mathbf{y}}_j$  is a measure of the strength of the relationship between  $\mathbf{x}_i$  and  $\mathbf{y}$  when the effect of  $\mathbf{x}_j$  has been removed. This is known as the *partial correlation*, for it has partialled out the effects on both the  $\mathbf{x}_i$  and  $\mathbf{y}$  of the other variables.

In the process of finding the appropriate weights in the multiple regression, the effect of each variable  $\mathbf{x}_i$  on the criterion  $\mathbf{y}$  was found with the effect of the other  $\mathbf{x}_j (j \neq i)$  variables removed. This was done explicitly in Equation 5.4 and implicitly in 5.7. The numerator in 5.4 is a covariance with the effect of the second variable removed and the denominator is a variance with the second variable removed. Just as in simple regression where  $\beta$  is a covariance divided by a variance and a correlation is a covariance divided by the square root

of the product of two variances, so is the case in multiple correlation where the  $\beta_j$  is a partial covariance divided by a partial variance and a partial correlation is a partial covariance divided by the square root of the product of two partial variances. The *partial correlation* between  $x_i$  and  $y$  with the effect of  $x_j$  removed is

$$r_{(x_i, x_j)(y, x_j)} = \frac{r_{x_i y} - r_{x_i x_j} r_{x_j y}}{\sqrt{(1 - r_{x_i x_j}^2)(1 - r_{y x_j}^2)}} \quad (5.9)$$

Compare this to 5.4 which is the formula for the  $\beta$  weight.

Given a data matrix,  $\mathbf{X}$  and a matrix of covariates,  $\mathbf{Z}$ , with correlations  $\mathbf{R}_{xz}$  with  $\mathbf{X}$ , and correlations  $\mathbf{R}_z$  with each other, the residuals,  $\mathbf{X}^*$  will be

$$\mathbf{X}^* = \mathbf{X} - \mathbf{R}_{xz} \mathbf{R}_z^{-1} \mathbf{Z}$$

To find the matrix of partial correlations,  $\mathbf{R}^*$  where the effect of a number of the  $\mathbf{Z}$  variables been removed, just express equation 5.9 in matrix form. First find the residual covariances,  $\mathbf{C}^*$  and then divide these by the square roots of the residual variances (the diagonal elements of  $\mathbf{C}^*$ ).

$$\mathbf{C}^* = (\mathbf{R} - \mathbf{R}_{xz} \mathbf{R}_z^{-1})$$

$$\mathbf{R}^* = (\sqrt{\text{diag}(\mathbf{C}^*)}^{-1} \mathbf{C}^* \sqrt{\text{diag}(\mathbf{C}^*)}^{-1}) \quad (5.10)$$

Consider the correlation matrix of five variables seen in Table 5.3. The partial correlations of the first three with the effect of the last two removed is found using the `partial.r` function.

**Table 5.3** Using `partial.r` to find a matrix of partial correlations

```
> R.mat
      V1  V2  V3  V4  V5
V1 1.00 0.56 0.48 0.40 0.32
V2 0.56 1.00 0.42 0.35 0.28
V3 0.48 0.42 1.00 0.30 0.24
V4 0.40 0.35 0.30 1.00 0.20
V5 0.32 0.28 0.24 0.20 1.00

> partial.r(R.mat, c(1:3), c(4:5)) #specify the matrix for input, and the columns for the X and Z variables
      V1  V2  V3
V1 1.00 0.46 0.38
V2 0.46 1.00 0.32
V3 0.38 0.32 1.00
```

The *semi-partial correlation*, also known as the *part-correlation* is the correlation between  $\mathbf{x}_i$  and  $\mathbf{y}$  removing the effect of the other  $\mathbf{x}_j$  from the predictor,  $\mathbf{x}_i$ , but not from the criterion,  $\mathbf{y}$ . It is just

$$r_{(x_i, x_j)(y)} = \frac{r_{x_i y} - r_{x_i x_j} r_{x_j y}}{\sqrt{(1 - r_{x_i x_j}^2)}} \quad (5.11)$$

### 5.3.1 *Alternative interpretations of the partial correlation*

Partial correlations are used when arguing that the effect of  $\mathbf{x}_i$  on  $\mathbf{y}$  either does or does remain when other variables,  $\mathbf{x}_j$  are statistically “controlled”. That is, in Table 5.3, the correlation between V1 and V2 is very high, even when the effects of V4 and V5 are removed. But this interpretation requires that each variable is measured without error. An alternative model that corrects for error of measurement (unreliability) would show that when the error free parts of V4 and V5 are used as covariates, the partial correlation between V1 and V2 becomes 0.. This issue will be discussed in much more detail when considering models of *reliability* as well as *factor analysis* and *structural equation models*.

## 5.4 Alternative regression techniques

That the linear model can be used with categorical predictors has already been discussed. Generalizations of the linear model to outcomes that are not normally distributed fall under the class of the *generalized linear model* and can be found using the `glm` function. One of the most common extensions is to the case of *dichotomous outcomes* (pass or fail, survive or die) which may be predicted using *logistic regression*. Another generalization is to non-normally distributed *count data* or *rate data* where *Poisson regression* is used. These models are solved by iterative maximum likelihood procedures rather than ordinary least squares as used in the linear model.

The need for these generalizations is that the normal theory of the linear model is inappropriate for such dependent variables. (e.g., what is the meaning of a predicted probability higher than 1 or less than 0?) The various generalizations of the linear model transform the dependent variable in some way so as to make linear changes in the predictors lead to linear changes in the dependent variable. For a very complete discussion of when to apply the linear model versus generalizations of these models, consult Cohen et al. (2003).

### 5.4.1 *Logistic regression*

Consider, for example, the case of a binary outcome variable. Because the observed values can only be 0 or 1, it is necessary to predict the probability of the score rather than the score itself. But even so, probabilities are bounded (0,1) so regression estimates less than 0 or greater than 1 are meaningless. A solution is to analyze not the probabilities themselves, but rather a monotonic transformation of probabilities, the logistic function:

$$p(Y|X) = \frac{1}{1 + e^{-(\beta_0 + \beta x)}}.$$

Using deviation scores, if the likelihood,  $p(\mathbf{y})$ , of observing some binary outcome,  $\mathbf{y}$ , is a continuous function of a predictor set,  $\mathbf{X}$ , where each column of  $\mathbf{X}$ ,  $\mathbf{x}_i$ , is related to the outcome probability with a *logistic* function where  $\beta_0$  is the predicted intercept and  $\beta_i$  is the effect of  $\mathbf{x}_i$

$$p(y|x_1 \dots x_i \dots x_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_i x_i + \dots + \beta_n x_n)}}$$

and therefore, the likelihood of not observing  $y$ ,  $p(\tilde{y})$ , given the same predictor set is

$$p(\tilde{y}|x_1 \dots x_i \dots x_n) = 1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_i x_i + \dots + \beta_n x_n)}} = \frac{e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_i x_i + \dots + \beta_n x_n)}}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_i x_i + \dots + \beta_n x_n)}}$$

then the *odds ratio* of observing  $y$  to not observing  $y$  is

$$\frac{p(y|x_1 \dots x_i \dots x_n)}{p(\tilde{y}|x_1 \dots x_i \dots x_n)} = \frac{1}{e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_i x_i + \dots + \beta_n x_n)}} = e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_i x_i + \dots + \beta_n x_n)}.$$

Thus, the logarithm of the odds ratio (the *log odds*) is a linear function of the  $\mathbf{x}_i$ :

$$\ln(odds) = \beta_0 + \beta_1 x_1 + \dots + \beta_i x_i + \dots + \beta_n x_n = \beta_0 + \beta \mathbf{X} \quad (5.12)$$

Consider the probability of being a college graduate given the predictors of age and several measures of ability. The data set `sat.act` has a measure of education (0 = not yet finished high school, ..., 5 have a graduate degree). Converting this to a dichotomous score (education >3) to identify those who have finished college or not, and then predicting this variable by a **logistic regression** shows that age is positively related to the probability of being a college graduate (not an overly surprising result) as is a higher ACT (American College Testing program) score. The results are expressed as changes in the logarithm of the odds for unit changes in the predictors. Expressing these as odds ratios may be done by taking the anti-log (i.e., the exponential) of the parameters. The confidence intervals of the parameters or of the Odds Ratios may be found by using the `confint` function (Table 5.4).

### 5.4.2 Poisson regression

If the underlying process is thought to be binary with a low probability of one of the two alternatives (e.g., scoring a goal in a football tournament, speaking versus not speaking in a classroom, becoming sick or not, missing school for a day, dying from being kicked by a horse, a flying bomb hit in a particular area, a phone trunk line being in use, etc.) sampled over a number of trials and the measure is the discrete counts (e.g., 0, 1, ... n= number of responses) of the less likely alternative, one appropriate distributional model is the *Poisson*. The Poisson is the limiting case of a *binomial* over  $N$  trials with probability  $p$  for small  $p$ . For a random variable,  $Y$ , the probability that it takes on a particular value,  $y$ , is

$$p(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}$$

where both the expectation (mean) and variance of  $Y$  are

$$E(Y) = \text{var}(Y) = \lambda$$

and  $y$  factorial is

$$y! = y * (y - 1) * (y - 2) \dots * 2 * 1.$$

**Table 5.4** An example of logistic regression using the `glm` function. The resulting coefficients are the parameters of the logistic model expressed in the logarithm of the odds. They may be converted to odds ratios by taking the exponential of the parameters. The same may be done with the confidence intervals of the parameters and of the odds ratios.

```
> data(sat.act)
> college <- (sat.act$education > 3) + 0 #convert to a binary variable
> College <- data.frame(college,sat.act)
> logistic.model <- glm(college~age+ACT,family=binomial,data=College)
> summary(logistic.model)

Call:
glm(formula = college ~ age + ACT, family = binomial, data = College)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8501 -0.6105 -0.4584  0.5568  1.7715
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.78855    0.79969  -9.739  <2e-16 ***
age           0.23234    0.01912  12.149  <2e-16 ***
ACT           0.05590    0.02197   2.544   0.0109 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 941.40  on 699  degrees of freedom
Residual deviance: 615.32  on 697  degrees of freedom
AIC: 621.32
Number of Fisher Scoring iterations: 5

> round(exp(coef(logistic.model)),2)
> round(exp(confint(logistic.model)),digits=3)

(Intercept)      age      ACT
           0.00      1.26      1.06
           2.5 % 97.5 %
(Intercept) 0.000 0.002
age         1.217 1.312
ACT         1.014 1.105
```

The sum of independent Poisson variables is itself distributed as a Poisson variable, so it is possible to aggregate data across an independent grouping variable.

*Poisson regression* models the mean for  $Y$  by modeling  $\lambda$  as an exponential function of the predictor set  $(x_i)$

$$E(Y) = \lambda = e^{\alpha + \beta_1 x_1 + \dots + \beta_p x_p}$$

and the log of the mean will thus be a linear function of the predictors.

Several example data sets are available in R to demonstrate the advantages of Poisson regression over simple linear regression. `epil` in **MASS** reports the number of epileptic seizures before and after administration of an anti-seizure medication or a placebo as a function of age and other covariates, `quine` (also in **MASS**) reports the rate of absenteeism in a small town in Australia as a function of culture, age, sex, and learning ability.

**Table 5.5** Using the general linear model `glm` to do Poisson regression for the effect of an anti-seizure drug on epilepsy attacks. The data are from the `epil` data set in **MASS**. Compare this analysis with a simple linear model or with a linear model of the log transformed data.

```
> data(epil)
> summary(glm(y~trt+base,data=epil,family=poisson))

Call:
glm(formula = y ~ trt + base, family = poisson, data = epil)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6157 -1.5080 -0.4681  0.4374 12.4054

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.278079   0.040709  31.396 < 2e-16 ***
trtprogabide -0.223093   0.046309  -4.817 1.45e-06 ***
base          0.021754   0.000482  45.130 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2517.83  on 235  degrees of freedom
Residual deviance:  987.27  on 233  degrees of freedom
AIC: 1759.2

Number of Fisher Scoring iterations: 5

> summary(lm(y~trt+base,data=epil))

lm(formula = y ~ trt + base, data = epil)

Residuals:
    Min       1Q   Median       3Q      Max
-19.40019 -3.29228  0.02348  2.11521  58.88226

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.27396   0.96814  -2.349  0.0197 *
trtprogabide -0.91233   1.04514  -0.873  0.3836
base          0.35258   0.01958  18.003 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.017 on 233 degrees of freedom
Multiple R-squared:  0.582,    Adjusted R-squared:  0.5784
F-statistic: 162.2 on 2 and 233 DF,  p-value: < 2.2e-16
```

### *5.4.3 Robust regression using $M$ estimators*

Robust techniques estimate relationships trying to correct for unusual data (outliers). A number of packages include functions that apply robust techniques to estimate correlations, covariances, and linear regressions. The **MASS** package, **robust**, **robustbase** all include robust estimation procedures. An interesting demonstration of the power of the human eye to estimate relationships was presented by Wainer and Thissen (1979) who show that visual displays are an important part of the data analytic enterprise. Students shown figures representing various pure cases of correlation were able to estimate the underlying correlation of contaminated data better than many of the more classic robust estimates.