7

How Much Does the Model Explain?

OVERVIEW OF THE CHAPTER

This chapter starts by noting that, counter to what might intuitively be expected, estimated variance parameters may go up when explanatory variables are added to the model. This leads to issues in the definition of the concept of explained variance, commonly referred to as R^2. A definition of R^2, defined at level one, is given which does not have this problem in its population values, although it still may sometimes have the problem in values calculated from data.

Next an exposition is given of how different components of the hierarchical linear model contribute to the total observed variance of the dependent variable. This is a rather theoretical section that may be skipped by the reader.

7.1 Explained variance

The concept of 'explained variance' is well known in multiple regression analysis: it gives an answer to the question how much of the variability of the dependent variable is accounted for by the linear regression on the explanatory variables. The usual measure for the *explained proportion of variance* is the squared multiple correlation coefficient, R^2. For the hierarchical linear model, however, the concept of 'explained proportion of variance' is somewhat problematic. In this section, we follow the approach of Snijders and Bosker (1994) to explain the difficulties and give a suitable multilevel version of R^2.

One way to approach this concept is to transfer its customary treatment, well-known from multiple linear regression, straightforwardly to the hierarchical random effects model: treat proportional reductions in the estimated variance components, σ^2 and τ_0^2 in the random-intercept model for two levels, as analogs of R^2-values. Since there are several variance components in the hierarchical linear model, this approach leads to several R^2-values, one for each variance component. However, this definition of R^2 now and then leads to unpleasant surprises: it sometimes happens that adding explanatory variables *increases* rather than decreases some of the variance components. Even negative values of R^2 are possible. Negative values of R^2 are clearly undesirable and are not in accordance with its intuitive interpretation.

In the discussion of R^2-type measures, it should be kept in mind that these measures depend on the distribution of the explanatory variables. This implies that these variables, denoted in this section by X, are supposed to be drawn at random from the population at level one and the population at level two, and not determined by the experimental design or the researcher's whim. In order to stress the random nature of the X-variable, the values of X are denoted X_{ij}, instead of by x_{ij} as in earlier chapters.

7.1.1 Negative values of R^2?

As an example, we consider data from a study by Vermeulen and Bosker (1992) on the effects of part-time teaching in primary schools. The dependent variable Y is an arithmetic test score; the sample consists of 718 grade 3 pupils in 42 schools. An intelligence test score X is used as predictor variable. Group sizes range from 1 to 33 with an average of 20. In the following, it sometimes is desirable to present an example for balanced data (i.e., with equal group sizes). The balanced data presented below are the data artificially restricted to 33 schools with 10 pupils in each school, by deleting schools with fewer than 10 pupils from the sample and randomly sampling 10 pupils from each school from the remaining schools. For demonstration purposes, three models were fitted: the empty Model A; Model B, with a group mean, $\overline{X}_{.j}$, as predictor variable; and Model C, with a within-group deviation score $(X_{ij} - \overline{X}_{.j})$ as predictor variable. Table 7.1 presents the results of the analyses both for the balanced and for the entire data set. The residual variance at level one is denoted σ^2; the residual variance at level two is denoted τ_0^2.

Table 7.1: Estimated residual variance parameters $\hat{\sigma}^2$ and $\hat{\tau}_0^2$ for models with within-group and between-group predictor variables.

	$\hat{\sigma}^2$	$\hat{\tau}_0^2$
I. Balanced design		
A. $Y_{ij} = \beta_0 + U_{0j} + E_{ij}$	8.694	2.271
B. $Y_{ij} = \beta_0 + \beta_1 \overline{X}_{.j} + U_{0j} + E_{ij}$	8.694	0.819
C. $Y_{ij} = \beta_0 + \beta_2(X_{ij} - \overline{X}_{.j}) + U_{0j} + E_{ij}$	6.973	2.443
II. Unbalanced design		
A. $Y_{ij} = \beta_0 + U_{0j} + E_{ij}$	7.653	2.798
B. $Y_{ij} = \beta_0 + \beta_1 \overline{X}_{.j} + U_{0j} + E_{ij}$	7.685	2.038
C. $Y_{ij} = \beta_0 + \beta_2(X_{ij} - \overline{X}_{.j}) + U_{0j} + E_{ij}$	6.668	2.891

From Table 7.1 we see that in the balanced as well as in the unbalanced case, $\hat{\tau}_0^2$ increases as a within-group deviation variable is added as an explanatory variable to the model. Furthermore, for the balanced case, $\hat{\sigma}^2$ is not affected by adding a group-level variable to the model. In the unbalanced case, $\hat{\sigma}^2$ increases slightly when adding the group variable. When R^2 is defined as the proportional reduction in the residual variance parameters, as discussed above, then R^2 on the group level is negative for Model C, while for the entire data set R^2 on the pupil level is negative for Model B. Estimating σ^2 and τ_0^2 using the REML method results in slightly different parameter estimates. The pattern, however, remains the same. It is argued below that defining R^2 as the proportional reduction in residual variance parameters $\hat{\sigma}^2$ and $\hat{\tau}_0^2$, respectively, is not the best way to define a measure analogous to R^2 in the linear regression model; and that the problems mentioned can be solved by using other definitions, leading to the measure denoted below by R_1^2.

7.1.2 Definitions of the proportion of explained variance in two-level models

In multiple linear regression, the customary R^2 parameter can be introduced in several ways: for example, as the maximal squared correlation coefficient between the dependent variable and some linear combination of the predictor variables, or as the proportional reduction in the residual variance parameter due to the joint predictor variables. A very appealing principle to define measures of modeled (or explained) variation is the principle of *proportional reduction of prediction error*. This is one of the definitions of R^2 in multiple linear regression, and can be described as follows. A population of values is given for the explanatory and the dependent variables $(X_{1i}, \ldots, X_{qi}, Y_i)$, with a known joint probability distribution; β is the value for which the expected squared error

$$\mathcal{E}\Big(Y_i - \sum_{h=0}^{q} \beta_h X_{hi}\Big)^2 \tag{7.1}$$

is minimal. (This is the definition of the ordinary least squares (OLS) estimation criterion. In this equation, β_0 is defined as the intercept and $X_{0i} = 1$ for all i.) If, for a certain case i, the values of $X_{1i}, \ldots, X_{qi}$ are unknown, then the best predictor for Y_i is its expectation $\mathcal{E}(Y)$, with mean squared prediction error $\text{var}(Y_i)$; if the values $X_{1i}, \ldots, X_{qi}$ are given, then the linear predictor of Y_i with minimum squared error is the regression value $\sum_h \beta_h X_{hi}$. The difference between the observed value Y_i and the predicted value $\sum_h \beta_h X_{hi}$ is the prediction error. Accordingly, the mean squared prediction error is defined as the value of (7.1) for the optimal (i.e., estimated) value of β.

The proportional reduction of the mean squared error of prediction is the same as the proportional reduction in the unexplained variance, due to the use of the variables $X_1, \ldots, X_q$. Mathematically, it can be expressed as

$$R^2 = \frac{\text{var}(Y_i) - \text{var}\big(Y_i - \sum_h \beta_h X_{hi}\big)}{\text{var}(Y_i)} = 1 - \frac{\text{var}\big(Y_i - \sum_h \beta_h X_{hi}\big)}{\text{var}(Y_i)}.$$

This formula expresses one of the equivalent ways to define R^2.

The same principle can be used to define 'explained proportion of variance' in the hierarchical linear model. For this model, however, there are several options with respect to what one wishes to predict. Let us consider a two-level model with dependent variable Y. In such a model, one can choose between predicting an individual value Y_{ij} at the lowest level, or a group mean $\overline{Y}_{.j}$. On the basis of this distinction, two concepts of explained proportion of variance in a two-level model can be defined. The first, and most important, is the *proportional reduction of error for predicting an individual outcome*. The second is the *proportional reduction of error for predicting a group mean*. We treat the first concept here; the second is of less practical importance and is discussed in Snijders and Bosker (1994).

First consider a two-level random effects model with a random intercept and some predictor variables with fixed effects but no other random effects:

$$Y_{ij} = \gamma_0 + \sum_{h=1}^{q} \gamma_h X_{hij} + U_{0j} + R_{ij}. \tag{7.2}$$

Since we wish to discuss the definition of 'explained proportion of variance' as a population parameter, we assume temporarily that the vector γ of regression coefficients is known.

Explained variance at level one

We define the explained proportion of variance at level one. This is based on how well we can predict the outcome of Y_{ij} for a randomly drawn level-one unit i within a randomly drawn level-two unit j. If the values of the predictors X_{ij} are unknown, then the best predictor for Y_{ij} is its expectation; the associated mean squared prediction error is var(Y_{ij}). If the value of the predictor vector X_{ij} for the given unit is known, then the best linear predictor for Y_{ij} is the regression value $\sum_{h=0}^{q} \gamma_h X_{hij}$ (where X_{h0j} is defined as 1 for all h,j.) The associated mean squared prediction error is

$$\mathrm{var}\Big(Y_{ij} - \sum_h \gamma_h X_{hij}\Big) = \sigma^2 + \tau_0^2.$$

The level-one explained proportion of variance is defined as the proportional reduction in mean squared prediction error:

$$R_1^2 = 1 - \frac{\mathrm{var}\big(Y_{ij} - \sum_h \gamma_h X_{hij}\big)}{\mathrm{var}(Y_{ij})}. \tag{7.3}$$

Now let us proceed from the population to the data. The most straightforward way to estimate R_1^2 is to consider $\hat{\sigma}^2 + \hat{\tau}_0^2$ for the empty model,

$$Y_{ij} = \gamma_0 + U_{0j} + E_{ij}, \tag{7.4}$$

as well as for the fitted model (7.2), and compute 1 minus the ratio of these values. In other words, R_1^2 is just the proportional reduction in the value of $\hat{\sigma}^2 + \hat{\tau}_0^2$ due to including the X-variables in the model. For a sequence of nested models, the contributions to the estimated value of (7.3) due to adding new predictors can be considered to be the contribution of these predictors to the explained variance at level one.

To illustrate this, we once again use the data from the first (balanced) example, and estimate the proportional reduction of prediction error for a model where within-group and between-groups regression coefficients may be different.

From Table 7.2 we see that $\hat{\sigma}^2 + \hat{\tau}_0^2$ for model A amounts to 10.965, and for model D to 7.964. R_1^2 is thus estimated to be $1-(7.964/10.965) = 0.274$.

Population values of R_1^2 are nonnegative

What happens to R_1^2 when predictor variables are added to the multilevel model? Is it possible that adding predictor variables leads to smaller values of R_1^2? Can we even be sure at all that these quantities are positive?

Table 7.2: Estimating the level-one explained variance (balanced data).

	$\hat{\sigma}^2$	$\hat{\tau}_0^2$
A. $Y_{ij} = \beta_0 + U_{0j} + E_{ij}$	8.694	2.271
D. $Y_{ij} = \beta_0 + \beta_1(X_{ij} - \overline{X}_{.j}) + \beta_2 \overline{X}_{.j} + U_{0j} + E_{ij}$	6.973	0.991

It turns out that a distinction must be made between the population parameter R_1^2 and its estimates from data. *Population* values of R_1^2 in correctly specified models, with a constant group size n, become smaller when predictor variables are deleted, provided that the variables U_{0j} and E_{ij} on the one hand are uncorrelated with all the X_{ij} variables on the other hand (the usual model assumption).

For *estimates* of R_1^2, however, the situation is different: these estimates sometimes do increase when predictor variables are deleted. When it is observed that an estimated value for R_1^2 becomes smaller by the addition of a predictor variable, or larger by the deletion of a predictor variable, there are two possibilities: either this is a chance fluctuation, or the larger model is misspecified. Whether the first or the second possibility is more likely will depend on the size of the change in R_1^2, and on the subject-matter insight of the researcher. In this sense changes in R_1^2 in the 'wrong' direction serve as a *diagnostic* for possible misspecification. This possibility of misspecification refers to the fixed part of the model, that is, the specification of the explanatory variables having fixed regression coefficients, and not to the random part of the model. We return to this in Section 10.2.3.

7.1.3 Explained variance in three-level models

In three-level random intercept models (Section 4.9), the residual variance, or mean squared prediction variance, is the sum of the variance components at the three levels, $\sigma^2 + \tau_0^2 + \varphi_0^2$. Accordingly, the level-one explained proportion of variance can be defined here as the proportional reduction in the sum of these three variance parameters.

Example 7.1 ***Variance in maths performance explained by IQ.***
In Example 4.8, Table 4.5 exhibits the results of the empty model (Model 1) and a model in which IQ has a fixed effect (Model 2). The total variance in the empty model is 7.816 + 1.746 + 2.124 = 11.686 while the total unexplained variance in Model 2 is 6.910 + 0.701 + 1.109 = 8.720. Hence the level-one explained proportion of variance is $1-(8.720/11.686) = 0.25$.

7.1.4 Explained variance in models with random slopes

The idea of using the proportional reduction in the prediction error for Y_{ij} and $\overline{Y}_{.j}$, respectively, as the definitions of explained variance at either level, can be extended to two-level models with one or more random regression coefficients. The formulas to calculate R_1^2 can be found in Snijders and Bosker (1994). However, the estimated values for R_1^2 usually change only very little when random regression coefficients are included in the model. The reason is that this definition of explained variance is based on prediction of the dependent variable from *observed* variables, and the knowledge of the precise specification of the random part only gives minor changes to the quality of this prediction.

The formulas for estimating R_1^2 in models with random intercepts only are very easy. Estimating R_1^2 in models with random slopes is more tedious. The simplest possibility for estimating R_1^2 in random slope models is to re-estimate the models as random intercept models with the same fixed parts (omitting the random slopes), and use the resulting parameter estimates to calculate R_1^2 in the usual (simple) way for random intercept models. This will normally yield values that are very close to the values for the random slopes model.

Example 7.2 ***Explained variance for language scores.***
In Table 5.4, a model was presented for the data set on language scores in elementary schools used throughout Chapters 4 and 5. When a random intercept model is fitted with the same fixed part, the estimated variance parameters are $\hat{\tau}_0^2 = 8.10$ for level two and $\hat{\sigma}^2 = 38.01$ for level one. For the empty model, Table 4.1 shows that the estimates are $\hat{\tau}_0^2 = 18.12$ and $\hat{\sigma}^2 = 62.85$. This implies that explained variance at level one is $1-(38.01 + 8.10)/(62.85 + 18.12) = 0.43$. This is a quite high explained variance. It can be deduced from the variances in Table 4.4.2 that the main part of this is due to IQ.

7.2 Components of variance[1]

The preceding section focused on the total amount of variance that can be explained by the explanatory variables. In these measures of explained variance, only the fixed effects contribute. It can also be theoretically illuminating to decompose the observed variance of Y into parts that correspond to the various constituents of the model. This is discussed in this section for a two-level model.

For the dependent variable Y, the level-one and level-two variances in the empty model (4.6) are denoted by σ_E^2 and τ_E^2, respectively. The total variance of Y is therefore $\sigma_E^2 + \tau_E^2$, and the components of variance are the parts into which this quantity is split. The first split, obviously, is the split of $\sigma_E^2 + \tau_E^2$ into σ_E^2 and τ_E^2, and was extensively discussed in our treatment of the intraclass correlation coefficient.

To obtain formulas for a further decomposition, it is necessary to be more specific about the distribution of the explanatory variables. It is usual in single-level as well as in multilevel regression analysis to condition on the values of the explanatory variables, that is, to consider those as given values. In this section, however, all explanatory variables are regarded as random variables with a given distribution.

7.2.1 Random intercept models

For the random intercept model, we divide the explanatory variables into level-one variables X and level-two variables Z. Deviating from the notation in other parts of this book, matrix notation is used, and X and Z denote vectors.

The explanatory variables $X_1, \ldots, X_p$ at level one are collected in the vector X with value X_{ij} for unit i in group j. It is assumed more specifically that X_{ij} can be decomposed into independent level-one and level-two parts,

$$X_{ij} = X_{ij}^W + X_j^B \tag{7.5}$$

[1] This is a more advanced section which may be skipped by the reader.

(i.e., a kind of multivariate hierarchical linear model without a fixed part). The expectation is denoted by $\mathcal{E}X_{ij} = \mu_X$, the level-one covariance matrix is

$$\mathrm{cov}\big(X_{ij}^W\big) = \Sigma_X^W,$$

and the level-two covariance matrix is

$$\mathrm{cov}\big(X_j^B\big) = \Sigma_X^B.$$

This implies that the overall covariance matrix of X is the sum of these,

$$\mathrm{cov}(X_{ij}) = \Sigma_X^W + \Sigma_X^B = \Sigma_X.$$

Further, the covariance matrix of the group average for a group of size n is

$$\mathrm{cov}(\overline{X}_{.j}) = \frac{1}{n}\,\Sigma_X^W + \Sigma_X^B.$$

It may be noted that this notation deviates slightly from the common split of X_{ij} into

$$X_{ij} = \big(X_{ij} - \overline{X}_{.j}\big) + \overline{X}_{.j}. \tag{7.6}$$

The split (7.5) is a population-based split, whereas the more usual split (7.6) is sample-based. In the notation used here, the covariance matrix of the within-group deviation variable is

$$\mathrm{cov}\big(X_{ij} - \overline{X}_{.j}\big) = \frac{n-1}{n}\,\Sigma_X^W,$$

while the covariance matrix of the group means is

$$\mathrm{cov}\big(\overline{X}_{.j}\big) = \frac{1}{n}\,\Sigma_X^W + \Sigma_X^B.$$

For the discussion in this section, the present notation is more convenient.

The split (7.5) is not a completely innocuous assumption. The independence between X_{ij}^W and X_j^B implies that the covariance matrix of the group means is at least as large[2] as $1/(n-1)$ times the within-group covariance matrix of X.

The vector of explanatory variables $Z = (Z_1, \ldots, Z_q)$ at level two has value Z_j for group j. The vector of expectations of Z is denoted

$$\mathcal{E}Z_j = \mu_Z,$$

and the covariance matrix is

$$\mathrm{cov}(Z_j) = \Sigma_Z.$$

In the random intercept model (4.8), denote the vector of regression coefficients of the Xs by

$$\gamma_X = (\gamma_{10}, \ldots, \gamma_{p0})',$$

[2]The word 'large' is meant here in the sense of the ordering of positive definite symmetric matrices.

and the vector of regression coefficients of the Zs by

$$\gamma_Z = (\gamma_{01}, \ldots, \gamma_{0q})' .$$

Taking into account the stochastic nature of the explanatory variables then leads to the following expression for the variance of Y:

$$\begin{aligned} \text{var}(Y_{ij}) &= \gamma_X' \, \Sigma_X \, \gamma_X + \gamma_Z' \, \Sigma_Z \, \gamma_Z + \tau_0^2 + \sigma^2 \\ &= \gamma_X' \, \Sigma_X^W \, \gamma_X + \gamma_X' \, \Sigma_X^B \, \gamma_X + \gamma_Z' \, \Sigma_Z \, \gamma_Z + \tau_0^2 + \sigma^2 . \end{aligned} \tag{7.7}$$

This equation only holds if Z and $\bar{X}$ are uncorrelated. For the special case where all explanatory variables are uncorrelated, this expression is equal to

$$\text{var}(Y_{ij}) = \sum_{h=1}^{p} \gamma_{h0}^2 \, \text{var}(X_h) + \sum_{h=1}^{q} \gamma_{0h}^2 \, \text{var}(Z_h) + \tau_0^2 + \sigma^2 .$$

(This holds, for example, if there is only one level-one and only one level-two explanatory variable.) This formula shows that, in this special case, the contribution of each explanatory variable to the variance of the dependent variable is given by the product of the regression coefficient and the variance of the explanatory variable.

The decomposition of X into independent level-one and level-two parts allows us to indicate precisely which parts of (7.7) correspond to the unconditional level-one variance σ_E^2 of Y, and which parts to the unconditional level-two variance τ_E^2:

$$\begin{aligned} \sigma_E^2 &= \gamma_X' \, \Sigma_X^W \, \gamma_X + \sigma^2 , \\ \tau_E^2 &= \gamma_X' \, \Sigma_X^B \, \gamma_X + \gamma_Z' \, \Sigma_Z \, \gamma_Z + \tau_0^2 . \end{aligned}$$

This shows how the within-group variation of the level-one variables eats up some part of the unconditional level-one variance; parts of the level-two variance are eaten up by the variation of the level-two variables, and also by the between-group (composition) variation of the level-one variables. Recall, however, the definition of Σ_X^B, which implies that the between-group variation of X is taken net of the 'random' variation of the group mean, which may be expected given the within-group variation of X_{ij}.

7.2.2 Random slope models

For the hierarchical linear model in its general specification given by (5.12), a decomposition of the variance is very complicated because of the presence of the cross-level interactions. Therefore the decomposition of the variance is discussed for random slope models in the formulation (5.15), repeated here as

$$Y_{ij} = \gamma_0 + \sum_{h=1}^{q} \gamma_h \, x_{hij} + U_{0j} + \sum_{h=1}^{p} U_{hj} \, x_{hij} + R_{ij} , \tag{7.8}$$

without bothering about whether some of the x_{hij} are level-one or level-two variables, or products of a level-one and a level-two variable.

Recall that in this section the explanatory variables X are stochastic. The vector $X = (X_1, \ldots, X_q)$ of all explanatory variables has mean $\mu_{X(q)}$ and covariance matrix $\Sigma_{X(q)}$. The subvector $(X_1, \ldots, X_p)$ of variables that have random slopes has mean $\mu_{X(p)}$ and covariance matrix $\Sigma_{X(p)}$. These covariance matrices could be split into within-group and between-group parts, but that is left to the reader.

The covariance matrix of the random slopes $(U_{1j}, \ldots, U_{pj})$ is denoted by T_{11} and the $p \times 1$ vector of the intercept–slope covariances is denoted by T_{10}.

With these specifications, the variance of the dependent variable can be shown to be given by

$$\begin{aligned} \text{var}(Y_{ij}) = \gamma' \, \Sigma_{X(q)} \, \gamma \; + \; \tau_0^2 \; + \; 2 \, \mu'_{X(q)} \, T_{10} \; + \; \mu'_{X(q)} \, T_{11} \, \mu_{X(q)} \\ + \; \text{trace}\left(T_{11} \, \Sigma_{X(p)}\right) \; + \; \sigma^2. \end{aligned} \tag{7.9}$$

(A similar expression, but without taking the fixed effects into account, is given by Snijders and Bosker (1993) as formula (21).) A brief discussion of all terms in this expression is as follows.

1. The first term, $\gamma' \, \Sigma_{X(q)} \, \gamma$, gives the contribution of the fixed effects and may be regarded as the 'explained part' of the variance. This term could be split into a level-one and a level-two part as in the preceding subsection.

2. The next few terms,

$$\tau_0^2 \; + \; \mu'_{X(q)} \, T_{10} \; + \; \mu'_{X(q)} \, T_{11} \, \mu_{X(q)}, \tag{7.10}$$

should be seen as one piece. One could rescale all variables with random slopes to have a zero mean (cf. the discussion in Section 5.1.2); this would lead to $\mu_{X(q)} = 0$ and leave of this piece only the intercept variance τ_0^2. In other words, (7.10) is just the intercept variance after subtracting the mean from all variables with random slopes.

3. The penultimate term, $\text{trace}\left(T_{11} \, \Sigma_{X(p)}\right)$, is the contribution of the random slopes to the variance of Y. In the extreme case where all variables $X_1, \ldots, X_q$ would be uncorrelated and have unit variances, this expression reduces to the sum of squared random slope variances. This term also could be split into a level-one and a level-two part.

4. Finally, σ^2 is the residual level-one variability that can neither be explained on the basis of the fixed effects, nor on the basis of the latent group characteristics that are represented by the random intercept and slopes.

7.3 Glommary

Estimated variance parameters. These may go up when variables are added to the hierarchical linear model. This seems strange but it is a known property of the hierarchical linear model, indicating that one should be careful with the interpretation of the fine details of how the variance in the dependent variable is partitioned across the levels of the nesting structure.

Explained variance. Denoted by R_1^2, this was defined as the proportional reduction in mean squared error for predicting the dependent variable, due to the knowledge of the values of the explanatory variables. This was elaborated for two-level and three-level random intercept models.

Explained variance in random slope models. In this approach, this is so little different from R_1^2 in random intercept models that it is better for this definition of explained variance to pay no attention to the distinction between random slope and random intercept models.

Components of the variance of the dependent variable. These show how not only the fixed effects of explanatory variables, reflected in R_1^2, but also the random effects contribute to the variance of the dependent variable.