20

INTRODUCTION TO CLUSTERED DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Determine if clustering is likely to be present in your data.
2. Draw a diagram to represent a hierarchical data structure, and identify repeated measures and spatial data structures as well as non-hierarchical structures.
3. Understand why clustering might be a problem, particularly as related to estimating standard errors of coefficients and to confounding by cluster effects.
4. Understand what impact clustering might have on your analysis of either continuous or discrete data.
5. Understand the uses, advantages and limitations of simpler methods to deal with clustering, such as fixed effects and stratified modelling, correction factors, robust variance estimation and survey estimation procedures.

20.1 INTRODUCTION

In common usage, a cluster denotes a set of objects (*eg* individuals) in a small group (see also the definition in Section 2.7). In statistics, cluster analysis aims to identify clusters among the observations, based on the similarity of their outcomes and possibly their physical distance. Our usage of **clustered data** is similar but does not pertain to cluster analysis. We think of clusters as observations that share some common features (that are not explicitly taken into account by explanatory variables in a model). This type of clustering is always derived from the data structure, of which the most common example is a hierarchical data structure. It is usually expected to lead to dependence between the responses of observations in a group (or cluster) because the shared feature makes the outcomes 'more similar' than otherwise. Thus, two alternative and occasionally encountered terms for these data are **hierarchical data** and **correlated data** (although the latter term is more general and may refer to other data structures as well).

Before proceeding, recall that statistical dependence between observations (for example, Y_1 and Y_2) is measured by covariance or correlation (which equals the covariance divided by the respective standard deviations):

$$\rho = \mathrm{corr}(Y_1, Y_2) = \frac{\mathrm{cov}(Y_1, Y_2)}{\mathrm{SD}(Y_1)\mathrm{SD}(Y_2)}, \qquad \text{where } -1 \le \rho \le 1$$

Eq 20.1

Similarity between observations corresponds to positive values of ρ and the dependence increases the further the value is from zero.

20.2 CLUSTERING ARISING FROM THE DATA STRUCTURE

In this section, we discuss the clustering which arises from animals sharing a common environment, clustering in space (*eg* geographical proximity) and repeated measurements within the same individual.

Common environment

Cows within a herd, puppies within a litter, and quarters within a cow are all examples of clustering in an environment. We usually assume that the degree of similarity among all pairs of observations within such a cluster are equal. Clustering is not necessarily restricted to a single level. For example, pigs might be clustered within a litter which might be clustered within a pen of pigs, which might be clustered in a farm which might be clustered in a region, as shown in the Fig. 20.1. Such data are called hierarchical or multilevel data. The hierarchy may also be expressed by saying that pigs are 'nested' within farms, and farms are nested within regions. The structure shown in Fig. 20.1 is a 5-level structure. In practice, we deal more often with data that have a 2-level or a 3-level structure.

The defining property of a hierarchical structure is that units together at some (low) level must also be together at all higher levels. In Fig. 20.1, this, for example, requires piglets from the same litter to be kept also in the same pen (and farm). Sometimes the data have two (or more) distinct hierarchies that cannot be merged into a single hierarchy. For example, in a salmon aquaculture study, each sea-water pen (production cage) was hierarchically nested within both the sea-water site and the fresh-water plant, or hatchery, from which the fish originated (Aunsmo *et al*, 2009). If, in this example, it was not true that all pens within a site originated

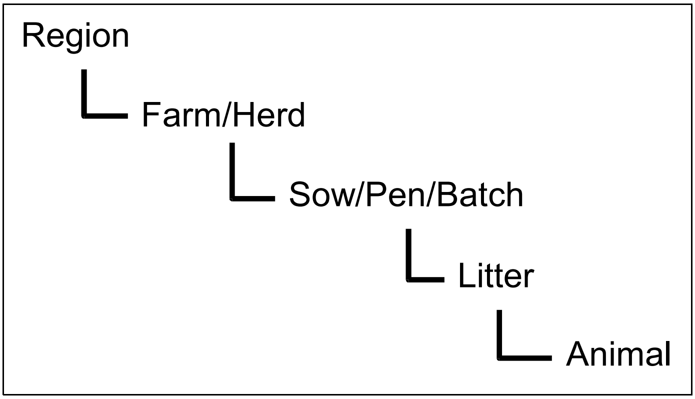

Fig. 20.1 A typical hierarchical data structure in veterinary epidemiology

from the same hatchery, and nor was it true (unrealistically) that every hatchery only delivered to a single site, the full data structure is no longer hierarchical and is instead called a **cross-classification**. Minor deviations from a strict hierarchical structure may be dealt with by suitably restricting the data. In the salmon aquaculture example, if only A few sites obtained fish from multiple hatcheries, one could restrict the data to one hatchery per site. Data structures with one hierarchy relating to the physical location of the animals and another hierarchy relating to their origin or slaughter are not uncommon (see also Example 24.6 for data on chicken which contains another more complicated and less common data structure called **multiple membership**). The term cross-classification is also used about factorial structures among categorical predictors and is here the rule rather than the exception; *eg* sex and breed are cross-classified in a dataset when several breeds are represented within each sex, and vice versa. Fig. 20.2 shows classification diagrams of the form used in the multilevel literature (Browne *et al*, 2001) to represent hierarchical and cross-classified data structures for the two versions of the salmon aquaculture example.

Spatial clustering

The hierarchy in Fig. 20.1 suggests that farms in the same region are similar. It sometimes seems natural to replace or extend this relationship by one where the dependence between farms

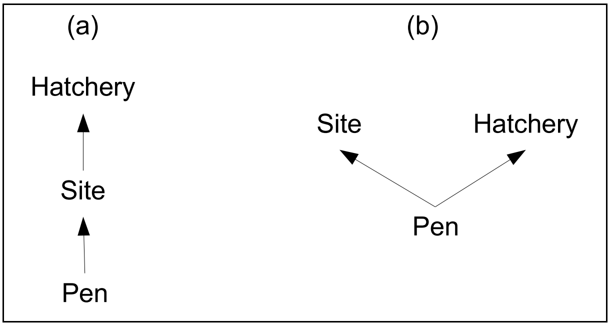

Fig. 20.2 Classification diagrams for hierarchical (a) and cross-classified (b) data structures in aquaculture. Hierarchical structure requires that every site receives fish from only one hatchery

is directly related (inversely proportional) to the distance between them. Spatial models incorporate the actual locations of study subjects (in this example the subjects are farms but they could also be the actual locations of cows in a tie-stall barn). Spatial data and analysis are reviewed in Chapters 25-26 of the book. If accurate spatial information is not available or detailed spatial modelling is not desirable (*eg* due to sparse data), spatial clustering might be accounted for by hierarchical level(s).

Repeated measurements
Repeated measures arise when several measurements of a variable are taken on the same animal (or other unit of observation) over a period of time. Daily milk weights in a cow are highly correlated because the level of milk production on one day is likely to be quite close to the production on the day before and the day after. Multiple measurements of lactation total milk production across lactations within a cow are also repeated measurements, but would not be so highly correlated. We might think of repeated measures as a special type of hierarchical clustering (*eg* in Fig. 20.1, an additional level could be added at the bottom of the hierarchy for repeated measurements on the animal). Note that for the data to be strictly hierarchical, no animals should move between herds (pens *etc*) during the study. However, just as with spatial clustering, several special considerations apply to repeated measures. Observations close together in time are likely to be more highly correlated than measurements with a longer time span between them. Also, repeated measurements might occur at any level in the hierarchy, not just at the lowest level. For example, if a study on pig production involved several batches within a farm, the batch level would then correspond to repeated measures over time on the farm. Analysis of repeated measures data is reviewed in Chapter 23 of the book.

Diagrams such as Figs. 20.1-2 (and Fig. 20.4) are highly recommended to determine and present data structures, as long as their defaults with regard to spatial and repeated structures are kept in mind. In certain situations, we may decide to disregard some hierarchical levels in the analysis, as discussed in Section 20.2.4. It is important to realise that the data structure pertains not only to the outcome but also to the predictor variables, and so it is useful to know whether predictors vary or were applied at particular levels. We elaborate on this idea in the context of the simplest two-level experimental design: the split-plot design. Section 20.2.3 briefly discusses how the effects of predictors vary in their interpretation at the different levels of a hierarchy.

20.2.1 Split-plot design

The split-plot concept and terminology dates back to the early 20th century when statistical methods were developed in the context of agricultural field trials. Consider the planning of an experiment involving two factors A and B with a and b levels, respectively. The special feature of the design is that factor B is practically applicable to smaller units of land (plots) than factor A. In the field trial context, we might think of A as a large-scale management factor such as pesticide spraying by plane and B as a small-scale factor such as plant variety. The experimental units for factor A are called **whole-plots**. The design needs some replication, and we assume we have c blocks of size a at our disposal, giving a total of ac whole-plots. The blocks would typically be separate pieces of land or experimental sites. **Note** Split-plot designs can also be constructed without blocks, but for simplicity, we describe the design with blocks. Within each block, the design would now be laid out in a two-step procedure, as illustrated in Fig. 20.3.

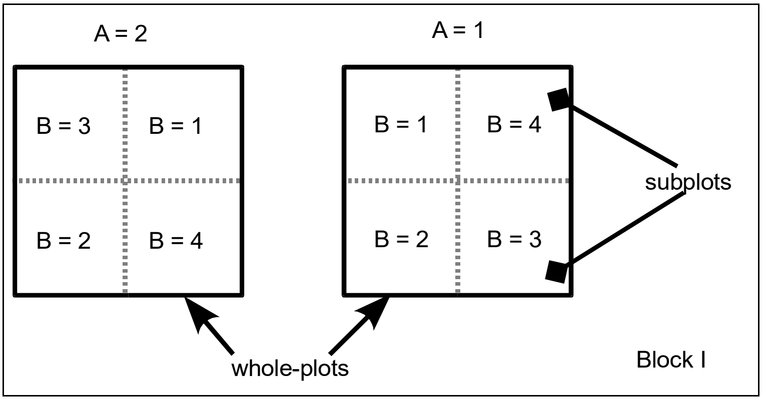

Fig. 20.3 Split-plot layout within one block, with a=2 whole-plots and b=4 subplots

1. randomly distribute the levels of factor A onto the *a* whole-plots within each of the *c* blocks,
2. divide each whole-plot into *b* **subplots**, and randomly distribute the levels of factor B onto the subplots.

As an animal-production example, we might have a herd-management factor A (*eg* tie-stall versus free-stall barns) and a treatment B applicable to individual animals (*eg* vaccination with one of four vaccines). Thus, the whole-plots would be the herds, and the subplots the animals. The blocks could be groups (possibly pairs) of similar herds, *eg* in the same region or of approximately the same size. In this example, the subplot treatment (vaccine) should be randomly assigned to animals, whereas the whole-plot treatments can hardly be randomised to herds (farmers will not be sympathetic to the idea of changing barn type). A split-plot design corresponds to a 2-level hierarchy with whole-plots as the upper level and subplots as the bottom level.

In the analysis of a split-plot experiment, the two factors A and B cannot be expected to be treated equally because they are applied to different experimental units. In particular, effects of the whole-plot factor A should be compared with the variation between whole-plots (corresponding to the first step of the design construction), and effects of the subplot factor B to the variation between subplots. It follows that it is necessary to split the total variation (or specifically the variance) in the data into variations between and within whole-plots. These variations are estimated independently from each other and with different accuracy (degrees of freedom). Usually the whole-plot variation will be considerably larger than the subplot variation, and factor A is estimated with less precision than factor B. The interaction between A and B 'belongs to' the subplot variation because differences between B-levels within any A-level can be determined within the whole-plots. This makes the split-plot design particularly attractive in situations where the principal interest is in the main effect of factor B and its interaction with factor A. In the example above, this would correspond to estimating the effects of the vaccines and determining if the vaccines worked differently in tie-stall compared with free-stall barns. The split-plot design is discussed in most statistical texts on experimental design (*eg* Mead , 1990)).

20.2.2 Variation at different levels

The split-plot design with its 2-level structure (*eg* cows within herds) illustrates how variation in the outcome of interest resides at the different levels of the hierarchy and how predictor variables explain variation at these different levels. One important implication is that the amount of unexplained variation at the different levels indicates what can be achieved by a detailed study of the units at the different levels. For example, a large unexplained variation between herds might indicate a substantial room for improvement in the outcome of interest, if we were able to understand why some herds do better than others. Generally, interventions targeted at the level where the greatest variation resides would seem to have the greatest chance of success. Explorative studies prior to interventions are one example of when the clustering of the data within the hierarchical structure is of primary interest (Dohoo *et al*, 2001).

20.2.3 Clustering of predictor variables

While the focus of our discussion to this point has been in the variation in the outcome of interest, we have also noted that predictor variables occur at various levels and might also be clustered. There is a wealth of potential relationships that can be examined when the hierarchical structure of the data is taken into consideration. For example, if data are recorded at the cow level, but clustered at the herd level we can examine:

- cow-level factors (*eg* lactation number) that affect a cow-level outcome (lactation total milk production),
- herd-level factors (*eg* barn type) that affect a cow-level outcome,
- herd-level factors (*eg* barn type) that affect a herd-level outcome (*eg* average lactation total milk production for the herd),
- cow-level factors (*eg* lactation number) that affect a herd-level outcome (*eg* average lactation total milk production for the herd), where the cow-level factors could either be recorded individually or aggregated to the herd-level (*eg* average lactation number for the herd),
- herd-level factors (*eg* barn type) that might alter a cow-level relationship (*eg* is effect of lactation number on milk production different in tie-stall and free-stall barns?) or vice versa.

Correctly evaluating the potential range of effects outlined above requires correct identification of the hierarchical structure of both outcome and predictors in the data. Fig. 20.4 illustrates how the predictor levels may be added to the hierarchical diagram.

20.2.4 Aggregation of levels

The hierarchical structure in a dataset might contain many levels, as shown in the 5-level structure of Fig. 20.1. However, sometimes we decide to exclude some levels from our analysis, and in this section, we give a few comments related to two common scenarios. In order to estimate the variation and the effects at the different levels, a certain minimal amount of replication is necessary at all levels. This is intuitively obvious because, if, for example, all batches contained only a single litter, then there would be no way of distinguishing between batch and litter effects. Another potential problem for the analysis is a strongly variable replication at one of the hierarchical levels (*eg* if some batches contain only one litter while

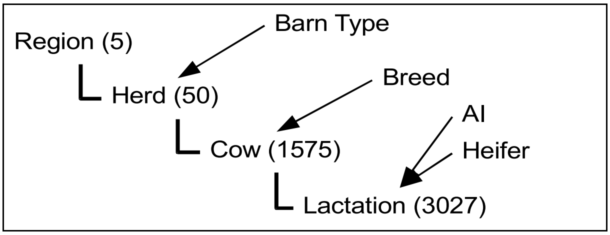

Fig. 20.4 Hierarchical structure of -reu_cfs- data on reproductive performance in Reunion Island

others contain up to 10 litters). To detect such problems, it is worthwhile to compute the range and average of the number of replications at each hierarchical level. There is no definitive rule as to the minimal replication but, whenever the average number of replicates is less than 2 and/or more than half of the units are unreplicated, problems can be anticipated. To illustrate the arbitrariness of such a rule, the Reunion Island dataset of Example 20.1 had, on average, only 1.9 lactations per cow without substantially impeding the analysis (Examples 21.3, 22.7-9, 22.11). If some levels need to be omitted in the hierarchy, it is useful to keep those at which principal predictors reside and those showing a lot of variation in a null model (*ie* without fixed effects) or based on descriptive statistics.

Example 20.1 Hierarchical data structure of Reunion Island data

data = reu_cfs

In a study of reproductive measures on Reunion Island (Dohoo *et al*, 2001), the calving to first service interval (-cfs-) and success or failure of first-service conception (-fscr-) were two of the outcomes evaluated. The study comprised 3,027 lactations distributed on 1,575 cows in 50 herds on Reunion Island. The data are strictly hierarchical because all cows remained on the same herds throughout the study period. The table below gives the number of units at each level and descriptive statistics for the replication at the level above.

Level	Number of units	Replication at level above	
		Mean	Range
Region	5	-	-
Herd	50	10	3–16
Cow	1575	31.5	8–105
Lactation	3027	1.9	1–5

Strictly speaking, the different lactations of each cow were repeated measures over time. However, the very short series of repeated measures per cow (1.9 observations on the average) does not realistically allow any complex repeated measures modelling.

The data included two dichotomous, lactation-level predictors: -heifer- (primiparous vs multiparous) and -ai- (artificial insemination vs natural breeding), as well as additional predictors not included in the reu_cfs dataset, *eg* breed and barn type. The levels of variation of each predictor are indicated in the hierarchical diagram in Fig. 20.4.

For discrete data, in particular binary data, some methods for clustered data have difficulty dealing with strong clustering even when there is ample replication; a possible solution for such cases is to aggregate the data at the lowest level (*eg* from the pig level to the batch level, when pigs are clustered in batches). We defer discussion of this issue to Section 22.5.7 and Example 22.12 in which 3-level data are aggregated to 2 levels.

20.3 EFFECTS OF CLUSTERING

Aside from any interest we might have in questions pertaining to the data structure, the reason for our interest in clustering is that it must be taken into account to obtain valid estimates of the effects of interest. This is because the assumption of independence inherent in most of the statistical models reviewed up till now in the book will be invalidated by the clustering.

To start with, let's address 2 common questions: 1. what happens if clustering is ignored?, and 2. if the data show no dependence, can clustering be ignored? If the presumption of these questions is whether one can escape the nuisance of accounting for clustering if it is not 'influential', we must raise a warning sign. Today's standard statistical software offers a variety of easily accessible options to account for clustering, and we find it hard to justify scientifically the use of a flawed method (even if only slightly) when better methods are readily available. If 'no dependence' means that a significance test of correlation turned out non-significant, it might be worthwhile to recall that the data showing no (significant) evidence against independence is by no means a proof of independence (by the distinction between Type I and Type II errors of statistical tests). Remember, "absence of evidence is not evidence of absence" (Carl Sagan).

Having said that, it might be fruitful for the understanding of the concept of clustering to examine the consequences of ignoring it. Perhaps not too surprisingly, the answer to the question to some extent depends on the statistical model used. Linear and logistic regression are discussed in more detail in the sections below. However, one general effect of ignoring clustering is that the standard errors (SEs) of parameter estimates will be wrong and often too small. This is particularly true if the factor of interest is a group-level factor (*eg* a herd-level factor such as barn type), or if it is an individual-level factor that is also highly clustered within groups (*eg* breed which may vary within a herd but most animals in a herd are of a single breed).

For a 2-level structure and a group-level predictor, it is possible to compute a **variance adjustment factor** (Section 20.3.3) for a cluster-adjusted analysis relative to an unadjusted analysis. Unfortunately, the simple variance adjustment leads to a widespread, but incorrect, belief that clustering always and only causes variance inflation. The discussion of the split-plot design illustrated the separation of the total variation into variation between and within whole-plots, with different values and degrees of freedom for each level. Therefore, if the data show these variations to be respectively large and small, the cluster-adjusted (split-plot) analysis will actually give smaller standard errors for subplot predictors—and larger standard errors for whole-plot predictors. It also follows that in a dataset with only a few herds (even if there is little clustering within herds), ignoring the hierarchical structure will lead you to grossly overestimate the power for evaluation of herd-level factors because it is the number of herds that determines the appropriate degrees of freedom, not the number of animals within herds. However, accounting for the data structure in the analysis might lead to smaller SEs for an animal-level factor. A final, less clear-cut effect of ignoring clustering is in the weighting of observations from different clusters. If the number of cows in different herds is highly variable, an unadjusted analysis gives unreasonably large weight to large herds. In summary, ignoring clustering can lead to other deficiencies than variance inflation, and in answer to question 2. above, even when 'no dependence' is seen, one would usually want to use cluster-adjusted methods to properly take into account the data structure.

20.3.1 Clustering for continuous data

Least squares estimation for linear (regression) models yields unbiased estimates of the regression coefficients, even if clustering is present and ignored (Liang & Zeger, 1993). This, perhaps intuitively surprising, fact is however of limited practical use because the corresponding SEs might be strongly affected by ignoring clustering. Thus, without reliable standard errors and test statistics to assess the precision and significance of the estimates, the statistical analysis does not go very far. Also, even if the estimates are unbiased, they might be very inefficient. By means of two simulated datasets, Example 20.2 illustrates how clustering might affect the standard errors. In this example, we use a linear mixed model (Chapter 21) to account for clustering, but other approaches are presented in this chapter and Chapter 23.

20.3.2 Clustering for discrete data

Estimation procedures in regression models for discrete data (*eg* logistic and Poisson regression models) are asymptotically unbiased which means that with infinitely large samples, they produce correct parameter estimates (Liang & Zeger, 1993). However, with limited sample sizes, some bias in the estimates may be present. If the data are clustered and the clustering is ignored in the analysis, the variance (or the SEs) of the estimates will (in most cases) be underestimated as was seen in models for continuous data. The larger 'true' variance in the parameter estimate means that the parameter estimate might be far from the true value, but this will not be readily apparent, resulting in (apparently) more biased estimates. Example 20.3 illustrates the practical implication of ignoring clustering for two simulated datasets. In this example, we use a logistic (generalised linear) mixed model to account for the clustering (Chapter 22), but other approaches are described in this chapter and Chapter 23.

20.3.3 Variance inflation as a result of clustering

The effect of clustering on variance estimates can most easily be seen in the situation in which a group (*eg* herd) level factor is being evaluated, but the outcome (*eg* milk production) is measured at the individual (*eg* cow) level. In this case, it is the variance of the herd mean milk production which is important for statistical testing. The magnitude of the effect of clustering on this variance (estimate) depends on both the **intra-class correlation** (*ICC*), and the size of the clusters. The *ICC* is the correlation between two observations within a cluster, assuming that the correlation is the same in all pairs of observations. It has all the usual properties of a correlation coefficient (*eg* takes values between -1 and 1, with a value of zero corresponding to independence). Methods for estimating the *ICC* depend on the type of outcome and model, and will be discussed in subsequent Chapters 21-23. If we assume that this correlation (*ICC* or ρ) is the same in all herds, then the variance of a herd mean milk production $(\text{var}(\overline{Y}))$ for a herd of size m is:

Example 20.2 Clustered continuous data

data = simcont_clustherd and simcont_clustcow

Two simulated datasets, each consisting of cows in 100 herds, were created. Herd sizes ranged from 20 to 311 cows (μ=116). Herd mean daily milk production varied randomly between herds (μ=30 kg/day, σ_h=7 kg/day), with larger herds tending to have higher production. Individual cow daily milk production values were normally distributed around the herd average (with σ=8 kg/day) unless the factor -X- was present, in which case the milk production was 5 kg higher. The single predictor -X- was added to each dataset with the herd prevalence of -X- varying between datasets. In the first dataset (simcont_clustherd), -X- was a herd-level factor so all cows in 50 herds had X=1 and all cows in 50 herds had X=0. In the second dataset (simcont_clustcow), -X- was a cow-level factor, present in half of the cows in each herd.

For each dataset, 2 or 3 models were fit. In the first, an ordinary linear model (a simple 2-sample comparison) ignoring herd was fit. In the second, a linear mixed model was used to account for the clustering within herds. In the third, herd average values of milk production were computed and analysed with respect to -X- (also a two-sample comparison); this was only appropriate for dataset 1 in which -X- was a herd-level variable.

Regression coefficients and SEs for analyses of two simulated datasets

		Linear model		Linear mixed model		Herd average linear model	
Dataset	Parameter	Estimate	SE	Estimate	SE	Estimate	SE
1:-X- at herd level	-X-	3.557	0.200	3.796	1.496	3.779	1.497
	constant	30.021	0.146	31.137	1.058	31.166	1.059
2:-X- at cow level	-X-	4.982	0.199	4.968	0.149		
	constant	29.257	0.141	30.646	0.728		

In dataset 1, all of the estimates for -X- are a long way from the true value (5) but this is due to random variation in the generation of the data. Most importantly, ignoring clustering produces SEs that are much lower than they should be. Controlling for clustering by computing herd average values for milk production and analysing those with respect to presence/absence of -X- produces almost exactly the same values as those observed from the linear mixed model.

In dataset 2, both estimates for -X- are close to the true value because estimation of a cow-level effect is much more precise than a herd-level effect. The linear mixed model gives a reduced SE for -X-, because the SE is derived from the within-herd variation which is smaller than both the between-herd variation and the total variation. For the constant (average milk production for cows with X=0 across herds), the correct SE involves the between-herd variation, and when clustering is ignored, the SE is, again, far too small.

$$\mathrm{var}(\overline{Y})=\frac{\sigma^2}{m}[1+(m-1)\rho] \qquad \textit{Eq 20.2}$$

where σ^2 is the variance among individual cow milk production values. **Note** If there is no clustering (*ie* ρ=0), then this formula is the usual one for the variance of a group mean (σ^2/m). In the literature, the quantity [1+(m-1)ρ] is referred to as the **variance inflation factor** or design effect (Okoumunne *et al*, 2002; Wears, 2002). In order to avoid confusion with the variance inflation factor for multicollinearity (Chapter 14), we use the term **design effect** (deff) from Section 2.10.4. The deff is the ratio between the variance (of the mean) in a clustered data structure relative to a data structure with independence. In Section 2.11.6 you saw how this

Example 20.3 Clustered binary data

data = simbin_clustherd and simbin_clustcow

To the same (fictitious) 100 herds as in Example 20.1, a binary outcome -dis- (disease) was added. In both datasets, the effect of -X- corresponded to an *OR* of 2, or a regression coefficient of ln(2)=0.693 on the logistic scale. The disease level of non-exposed cows was set at p=0.2, corresponding to a value of ln(0.2/0.8)=-1.4 on logistic scale. Herd effects varied on the logistic scale with a standard deviation of 1. As before, the first dataset (simbin_clustherd) had -X- as a herd-level factor (with -X- present in 50 herds), and the second dataset (simbin_clustcow) had -X- as a cow-level factor (with -X- present in 50% of the cows in each herd).

For each dataset, two models were fit: an ordinary logistic regression ignoring herd clustering (essentially a 2X2-table analysis), and a logistic mixed model to account for herd clustering.

Regression coefficients and SEs for analyses of two simulated binary datasets

		Logistic model		Logistic mixed model	
Dataset	Parameter	Estimate	SE	Estimate	SE
1:-X- at herd level	-X-	0.529	0.042	0.620	0.204
	constant	-1.242	0.033	-1.305	0.146
2:-X- at cow level	-X-	0.586	0.042	0.697	0.046
	constant	-1.250	0.032	-1.361	0.111

In both datasets, the most conspicuous difference between the two analyses is that the simple logistic model underestimates the standard errors for all parameters except the cow-level predictor. The parameter estimates of the mixed logistic model are somewhat closer to the true value in this case, but the SEs show that it could easily have been the other way around. Note that the SEs for the logistic mixed model in dataset 2 are less than in dataset 1 because a within-herd design is more powerful than a between-herd design.

quantity can be used to adjust sample size estimates for clustering when computing sample sizes. Table 20.1 shows how both the group size and the magnitude of the *ICC* affect how much the variance needs to be inflated to adequately account for clustering. *ICC*s have been computed for herd-level clustering of a number of infectious diseases and were found to range from 0.04 (*Anaplasma marginale* in cattle) to 0.42 (bovine viral diarrhea in cattle), but most were less than 0.2 (Otte & Gumm, 1997).

Table 20.1 Effect of group size (m) and ICC (ρ) on variance of group means when dealing with clustered data (from Eq 20.2)

ρ	m	deff	Comment
0	any	1	No within-group clustering = no variance inflation
1	m	m	Complete within-group clustering effectively makes the sample size equal to the number of groups
0.1	6	1.5	A low ICC with a moderate group size can have as much impact as a high ICC with a very small group size
0.5	2	1.5	
0.1	101	11	Large group sizes, even with a low ICC, result in a very high variance inflation (deff)

Finally, a few notes on the use of deffs. First, they apply to cluster means and, therefore, more generally to between-cluster effects, but not to within-cluster effects. Second, Eq. 20.2 depends only on the variance, *ICC* and cluster size, so it also applies to discrete outcomes. However, due to the relation between the mean and variance in discrete data, the variance will not be constant within a cluster if there are (additional) within-cluster predictors, and nor will it be constant between clusters with different values of between-cluster predictors. Consequently, for discrete data, a single deff value can only be seen as a rough approximation of the variance inflation. Third, if group sizes are unequal, you can use the minimum, average and maximum group sizes in Eq. 20.2 and the deff formula to assess the impact of clustering.

20.4 SIMULATION STUDIES ON THE IMPACT OF CLUSTERING

Examples 20.2-3 illustrate the effects of clustering in single (simulated) datasets. In order to explore the effects of clustering systematically, a series of simulation studies were carried out. We will present some results for a binary outcome (Sections 20.4.1-2), but another part of the simulation studies involved a continuous (normally distributed) outcome (Dohoo & Stryhn, 2006). One finding unique to a continuous outcome was the absence of any bias in estimation of a regression coefficient by ignoring the clustering in the analysis. The impact of clustering on the SEs of the estimated regression coefficient was qualitatively similar to those for a binary outcome described below, but quantitatively more pronounced.

A recent simulation study explored the impact of ignoring clustering in a 2-level setting when there is little replication within clusters (Clarke, 2008). It was concluded that, at an average of (at least) 5 observations per cluster, valid and reliable estimates can be obtained with mixed models with either continuous or binary outcomes. Some biases were observed with very sparse replication (2 observations per cluster), but ignoring clustering was concluded to be associated with increased risk of Type I errors, even when there were few observations per cluster.

20.4.1 Simulation study on the impact of clustering for binary outcome

The simulation study is presented in the framework of a 2-level hierarchy consisting of cows within 100 herds of variable sizes (mean herd size=150; see Dohoo & Stryhn, 2006). One binary outcome (*Y*) was generated for each cow within each herd. The log-odds of the baseline herd prevalence was generated from N(-1.4,1), resulting in a baseline prevalence of approximately 25%. The cow-level predictor *X* was generated from a standard normal distribution, and set to have one of 17 levels of within-herd clustering (*ICC*(*X*)), ranging from 0 (complete independence) to 1 (a herd-level predictor). The effect of *X* was linear on the logit scale with a regression coefficient of 0.69, equivalent to an *OR* of 2. Each of the 17 scenarios were simulated 1000 times. Within each iteration, the dataset was created as described above. Subsequently, a simple logistic regression with *X* as the sole predictor and a logistic mixed model were carried out for each of the simulated datasets. Bias in estimates of the coefficient β of *X* and in SE(β) were computed by dividing the means of the respective estimates from the simple logistic model with the means from the logistic mixed model (left panel of Fig. 20.5). The effect of clustering on the variability of the individual estimates was evaluated by determining the standard deviation (SD) of β among the simulations for both models (right panel of Fig. 20.5).

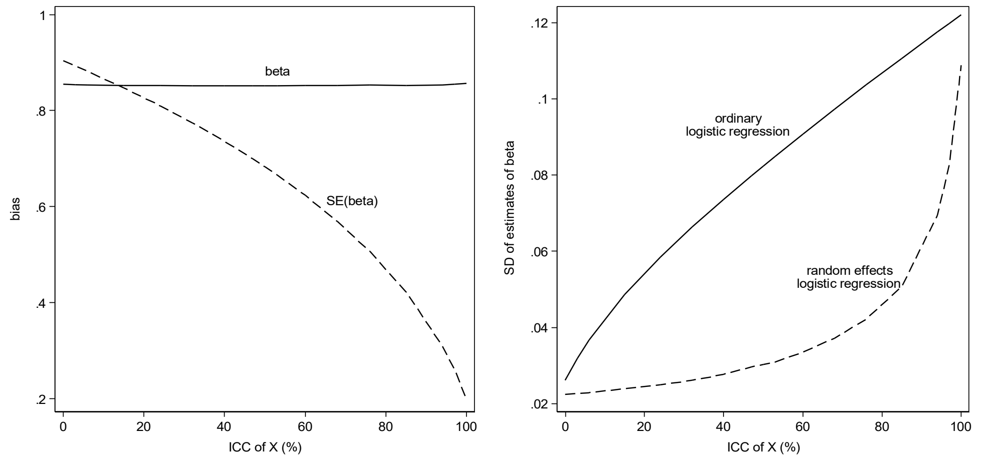

Fig. 20.5 Results of simulation study to assess effect of clustering in binary data: bias in estimates of treatment effect (beta) and of SE(beta) in left panel; standard deviation among simulations of individual beta estimates in right

The left panel of Fig. 20.5 shows the ratio between estimates of the simple logistic and logistic mixed models to be constant around a value of 0.85 for all values of the *ICC* of *X*. It is well-known that estimates (β) of logistic mixed models are numerically larger than those of simple logistic regression (cluster-specific (SS) and population-averaged (PA) estimates, respectively, in the terminology of Chapter 22), and their ratio agrees well with the value ($1/\sqrt{(1+0.346*1)}$ $=0.86$) obtained from Eq. 22.2 in Section 22.2.2. When *X* is a herd-level predictor ($ICC(X)=1$), the SE of its coefficient is grossly underestimated, as would be expected from Eq. 20.2 with a large m and $\rho\approx0.23$ (from Eq. 22.4 in Section 22.2.3). When *X* is a cow-level predictor that doesn't cluster at all within herds ($ICC(X)=0$), the SE(β) is slightly overestimated by the logistic mixed model (when the difference in scale is taken into account). At intermediate *ICC*(*X*) values, the ratio between SE(β)s decreases as smooth function of *ICC*(*X*); note that the actual point where the two curves cross depends on the specific settings of the design and parameters.

The right panel shows that even without any systematic biases between the estimates (beyond the distinction between SS and PA estimates), the individual estimates from the simple logistic regression are much more variable than those from the logistic mixed model, especially for intermediate values of *ICC*(*X*). As a consequence, individual estimates derived from an ordinary logistic model may be badly off the true value, but it would be impossible to predict the direction or magnitude of the bias because the estimated SE(β)s are underestimated. At extreme values of *ICC*(*X*) (0 or 1), the variability of the ordinary regression estimates are much closer to what would be expected.

20.4.2 Clustering and confounding

It has often been assumed that controlling for clustering also serves to control for unmeasured confounders which may be associated with the groups in which the animals are clustered. In order to investigate this assumption, the simulated data described above were extended to include a confounding variable (*Z*) which had a standard normal distribution, which also doubled the odds of the outcome and that was correlated with the predictor (ρ=0.5). As with *X*, *Z* was created to have one of 17 levels of within-herd clustering (*ICC*(*Z*)), ranging from 0 (complete independence) to 1 (a herd-level predictor). The bias in estimates of β and SE(β) for *X* and *Z* were computed by dividing their respective means of estimates from a logistic mixed model which did not include *Z* by those from a model which did include *Z*. The results are shown in Fig. 20.6.

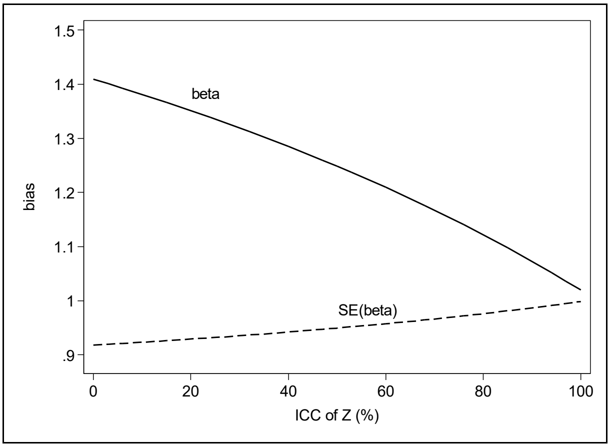

Fig. 20.6 Evaluation of ability of mixed model to control for unmeasured confounders

If the confounder is a true group-level factor (*ie ICC*=1) adjustment for clustering in a logistic mixed model completely eliminates any confounding effect of *Z*. On the other hand, if there is some variation in *Z* among individuals within a group (*ie ICC*<1), then there is residual confounding due to *Z* and not including *Z* in the model will produce a biased estimate—in this case an estimate of β that is up to 1.4 times larger than it should be. The actual magnitude and direction of the confounding bias will depend on the strength and direction of the relationships between *Z* and *Y* and between *Z* and *X*. While the estimate of the β is biased if *ICC*<1, the estimate of the SE(β) is virtually unbiased regardless of the *ICC* of *Z*, because both models had a correct specification of the variance structure. This demonstrates why estimates (and not just their SEs) often change when clustering is accounted for in an analysis. Depending on the *ICC* of an unmeasured confounder, some (0<*ICC*<1) or all (*ICC*=1) of its confounding effect will be removed by accounting for clustering in the analysis.

20.5 INTRODUCTION TO METHODS FOR DEALING WITH CLUSTERING

The primary focus in the book among methods for the analysis of clustered data is on **mixed**, or **random effects**, **models**, which are reviewed for continuous and discrete data in Chapters 21 and 22, respectively. In addition, mixed models for survival data, **frailty models**, were introduced in Section 19.11. Mixed models are also used for repeated measures and spatial data (Chapters 23 and 26). Another widely used approach to clustered data, estimation by **generalised estimation equations** (GEE), will be introduced in the context of repeated measures data (Chapter 23) for which it was originally developed, although GEE is more widely applicable, *eg* to hierarchical data. The present section contains some introductory remarks on detection of clustering, and a discussion of simpler, traditional approaches to dealing with clustering using fixed effects, stratification, scaling of variance estimates and **robust variance** estimates, as well as a discussion of the connection with methods for analysis of survey data (Section 2.10).

20.5.1 Detection of clustering

The primary resource for detection of clustering is the researcher's awareness. Whenever data are collected from individuals that are managed in a group, we should suspect that the data might be clustered. More generally, this is the case whenever animals share common features of potential impact that are not accounted for by explanatory variables. Any hierarchical structure in the origin or management of individuals might introduce clustering, as shown in Figs. 20.1 and 20.4. Also, repeated measures and spatial data structures should always be noticed and examined.

One might expect some general statistical test for clustering to be 'standard' (in common use), but this is not so. We offer two explanations. One is that clustering is dealt with differently in discrete and continuous data, and in different statistical models. A general approach is to extend a statistical model with an additional parameter (or effect) for clustering, estimate that parameter and test whether it differs significantly from zero (no clustering). This approach has been introduced in Chapter 18 where addition of an extra variance parameter to the Poisson model produced a negative binomial model. In discrete models such as logistic and Poisson regression, one might also compare the actual variation in the data with the expected variation (from the binomial or Poisson distributions) by a goodness-of-fit statistic, which, if significant, indicates **overdispersion**, potentially a result of clustering (see Sections 16.12.4, 18.5.3 and 20.5.3). A second reason why testing for clustering is less common than one might expect, is that even small amounts of clustering might have substantial impact on variance estimates, as illustrated in Section 20.3.3. Therefore, one is often inclined to keep a clustering effect in the statistical model even if it is not statistically significant, particularly if it shows 'some effect' and is strongly suggested by the data structure.

20.5.2 Fixed effects and stratified models

We will first discuss one simple and previously common approach to dealing with clustering which has also occasionally been used in previous chapters of this book—that is to include the group identifier as a **fixed effect** in the regression model. Let us for the sake of the discussion, in this and the following sections, refer to the groups as herds and to the within-group subjects as cows. In fixed effects models, dummy (indicator) variables representing the 'group' (*eg* herd) are included in the model. The fixed effects analysis then effectively estimates a separate parameter for each herd. This has the effect of separating the variation between herds from the residual variation and results in more appropriate tests of significance for within-herd factors.

There are several major drawbacks to this approach. The first is that one cannot include any herd-level predictors (*eg* barn type) in the model because they will be absorbed into the herd effects. The second drawback is that the model does not contain any dependence between cows in the same herd (*ie* the model contains only the within-herd variance as the between-herd variance is removed by the fixed effects), and therefore does not properly inflate the variance on means across herds (*eg* the calving to conception interval for heifers treated with a certain vaccine). Another way of saying this is that any inferences made are specific to the actual herds, where very often one would want conclusions to refer to a more general population of herds. A third drawback is that with many herds it requires the fitting of a large number of parameters (one for each herd), and the parameter estimates in the model might become unstable if there are relatively few observations per group. As we are not usually interested in the actual effects of each herd, these fixed effects are often considered 'nuisance' parameters. The consequences of having a large number of 'nuisance' parameters in the model might be more serious for discrete than normal distribution models because asymptotic properties of estimation procedures used to fit discrete distribution models break down if there is a large number of parameters to estimate (relative to the number of observations).

On the other hand, the obvious benefit of a fixed-effects approach is a simpler statistical analysis, because the fixed effects can be added to the model (*eg* linear or logistic) without changing its structure and without extra software requirements. In particular, with limited model-checking facilities for mixed models in some software, you might be tempted to base part of the model-checking on the fixed effects version of a mixed model, although, strictly speaking, this is incorrect. More substantively, a fixed-effects approach may be preferable when the herds are specific to the study and do not represent a more general population. The fixed-effects approach is illustrated in Examples 20.4 and 20.5 for the previously used simulated datasets with a cow-level predictor and a binary and continuous outcome, respectively.

Another simple approach to dealing with clustered binary data and a dichotomous within-herd factor is to carry out a **stratified analysis** using the **Mantel-Haenszel** procedure described in Chapter 13, with strata defined by the clustering variable. Mantel-Haenszel-type stratified analyses are limited to binary outcomes and a single categorical within-group predictor; for multifactorial problems, they are mainly used for descriptive purposes. A stratified analysis is also included in Example 20.4.

20.5.3 Factors to correct for clustering

This section summarises two ways of correcting an analysis in which clustering has not been taken into account in the model. These involve an estimate of the **intra-class correlation coefficient** (**ICC**) (Section 20.3.3) or an estimate of the **overdispersion**, and using one of these to adjust the standard error (SE) of regression coefficients. Note that these methods rely on the simplistic premise that clustering affects only the SEs of estimates (and, generally, when not taken into account, leads to SEs that are too small by the same amount (factor) for all regression coefficients). Our previous examples have shown that this is not always the case. Therefore, not all uncorrected analyses might be 'repaired' by increasing the SEs, and the researcher must pay particular attention to the requirements for these correction factors to be meaningful.

Adjustment by the design effect
Eq. 20.2 shows how the effect of clustering on the variance of herd means, in terms of the design effect (deff), depends on the *ICC* and the herd size. If the *ICC* is known and both the *ICC* and deff are the same in all herds, an analysis involving only herd-level factors, but ignoring clustering in herds, might be corrected by inflating the standard errors of regression coefficients by the square root of the deff. In practice, herds are rarely of the same size, and the *ICC* is not known nor constant between herds except for special cases (although constant *ICC*s are often assumed for normally distributed data). The method might be acceptable as an approximation even without these conditions (Donner, 1993) but the validity of the approximation is difficult to assess. In summary, this approach is largely obsolete but may still be of use for descriptive purposes.

Overdispersion as a result of clustering
The concept of overdispersion was introduced in Chapter 16 for models based on the binomial

Example 20.4 Summary of analyses for simulated binary data

data = simbin_clustherd and simbin_clustcow

The simulated binary datasets from Example 20.3 with -X- as either a herd-level or a cow-level factor were analysed by the simpler methods of dealing with clustering of this section, as well as a logistic mixed model (Chapter 22) and the generalised estimating equation (GEE) procedure (Chapter 23; an exchangeable correlation structure was used). Some methods are only applicable to one of the datasets; for example, stratification by clusters only works for a cow-level predictor. All results are summarised in Table 20.2.

Table 20.2 Results for simulated binary datasets with predictor at herd or cow level

Dataset	-X- herd level				-X- cow level				
Parameter	**Constant**		**-X-**		**Constant**		**-X-**		**SS**
Statistic	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**	**PA[a]**
Uncond. logistic model	-1.242	0.033	0.529	0.042	-1.250	0.032	0.586	0.042	PA
Fixed effects	-	-	-	-	-2.130	0.632	0.704	0.046	SS
Mantel-Haenszel	-	-	-	-	-	-	0.698	0.046	SS
Variance adj. (overdisp)	-1.242	0.140	0.529	0.181	-	-	-	-	PA
Variance adj. (Williams)	-1.108	0.139	0.558	0.186	-	-	-	-	PA
Robust variance[b]	-1.242	0.146	0.529	0.211	-1.250	0.114	0.586	0.044	PA
Logistic mixed model	-1.305	0.146	0.620	0.204	-1.361	0.111	0.697	0.046	SS
Generalised estim. eq.	-1.110	0.125	0.559	0.177	-1.112	0.095	0.587	0.042	PA

[a] Subject specific or population average estimate
[b] Same results obtained for analysis by survey method clustered at herds

In both datasets, the PA estimates were markedly closer to zero than the SS ones (as expected because estimates by these methods are on different scales and have different interpretations—see Chapters 22-23 for further discussion).

In the dataset with -X- as a cow-level factor, the fixed effects estimate for the PA estimates of X were all quite close, as were the SS estimates. The SEs were also very similar. The estimated constant from the fixed effects model corresponds to the proportion of positives (on logit scale) in herd 1, and its estimate therefore differed from the overall proportion (across all herds) from the other models. The 99 coefficients for herds 2-100 are not shown.

In the dataset with -X- as a herd-level factor, there was considerable variation in the estimates of the constant and the SE from the unconditional model was grossly underestimated. The PA estimates of X from the William's method and GEE were somewhat higher than those from the unconditional model. The estimated overdispersion parameter was 18.38, computed by dividing the Pearson χ^2 (1801.50) by its degrees of freedom (100-2=98). This corresponded to a standard deviation 4.29 times larger than from the binomial distribution. The SEs were therefore multiplied by this factor (and were comparable to those from GEE).

The fairly large differences between the parameter estimates indicate that the choice of procedure has appreciable impact on the results, and illustrates that the simpler methods (fixed effects, stratification and variance adjustments) often fail to give the same answers as the 2 main approaches—the logistic mixed model and GEE.

Example 20.5 Summary of analyses for simulated continuous data

data = simcont_clustherd and simcont_clustcow

The simulated continuous datasets from Example 20.2 with -X- as either a herd-level or a cow-level factor were analysed by the simpler methods of dealing with clustering of this section, as well as a linear mixed model (Chapter 21) and the generalised estimating equation (GEE) procedure (Chapter 23; an exchangeable correlation structure was used). Fixed effects modelling only works for a cow-level predictor. All results are summarised in Table 20.3.

Table 20.3 Results for simulated continuous datasets with predictor at herd or cow level

Dataset	-X- herd level				-X- cow level			
Parameter	Constant		-X-		Constant		-X-	
Statistic	Est	SE	Est	SE	Est	SE	Est	SE
Uncond. linear model	30.021	0.146	3.557	0.200	29.257	0.141	4.982	0.199
Fixed effects	--	--	--	--	24.324	1.800	4.968	0.149
Robust variance[a]	30.021	1.092	3.557	1.712	29.257	0.874	4.982	0.142
Linear mixed model	31.137	1.058	3.796	1.496	30.646	0.728	4.968	0.149
Generalised estim. eq.	31.135	1.040	3.797	1.488	30.648	0.722	4.968	0.141

[a] same results obtained for analysis by survey method clustered at herds

In the dataset with -X- as a cow-level factor, fixed effects modelling gave identical results to the linear mixed model for the coefficient of -X-, whereas the constant corresponded to a different parameter (the mean for herd 1 with -X- absent). The agreement in estimates and SEs between different methods was generally better than for the binary datasets in Example 20.4. The linear mixed model and GEE estimates were almost identical, so the main difference was between these estimates and those of the simple linear model. The robust SEs were larger than the unadjusted SEs, except for the coefficients of -X- as a cow-level predictor; this behaviour fits nicely with our discussion in Examples 20.2-3. The linear mixed model is the generally preferred choice, see Chapter 21 for further discussion of this model and its assumptions.

distribution and in Chapter 18 for the Poisson distribution. Table 16.5 shows a hypothetical example of overdispersion caused by clustering. Overdispersion may occur in all (discrete) models with a relation between the distribution's mean and variance, and intuitively means that the dispersion in the data is larger than expected from the mean (and the relation between the mean and variance). (**Note** Overdispersion does therefore not occur in normal distribution models!) The converse situation, that the dispersion in the data is less than expected from the mean, is possible as well, and is called **underdispersion**. Underdispersion is less common in practice and more difficult to interpret, the one standard example being a negative correlation between observations in a cluster caused by competition for limited resources (*eg* feed).

Adjustment by overdispersion factor

Generalised linear models (Section 16.11) allow for an additional **dispersion** (or scale) parameter (φ) to take into account if the 'natural dispersion' in the data does not match the distribution used (*eg* binomial or Poisson). As before, for a 2-level model with only herd-level predictors, this might be used to adjust for the inflation in variance at the herd level. In principle, this correction is valid for unequal herd sizes as well. Let's look at a binomial model with an overdispersion parameter in the context of the data in Table 16.5 to see how that would work. Denote by Y_i and n_i the number of positive outcomes and the total number of animals sampled in herd *i*, respectively. Then the model's assumptions are:

$$\mathrm{E}(Y_i)=n_i p_i \qquad \text{and} \qquad \mathrm{var}(Y_i)=\varphi\, n_i p_i(1-p_i) \qquad \textbf{\textit{Eq 20.3}}$$

Here φ is assumed to be independent of the group sizes n_i , and clearly that is not necessarily true! It makes us realise that using an overdispersion parameter assumes a particular form of the variance inflation across groups. As for the Poisson distribution (Section 18), several methods to estimate φ exist, and the Pearson estimate is generally considered preferable (Hilbe, 2009; McCullagh & Nelder, 1989). Note that for binary data, replication within the groups is necessary for the method to work, *ie* the data are essentially **grouped binary** (*ie* **binomial**) data with no within-group predictors. In this situation, other scaling possibilities than the relation in Eq 20.3 exist, *eg* the Williams method (Collett, 2002) which also affects the parameter estimates. For moderately varying herd sizes, the two methods do not differ much (they are identical for equal herd sizes). The Williams method (and also the beta-binomial model for grouped binary data discussed in Section 22.4.5) assumes

$$\mathrm{var}(Y_i)=[1+(n_i-1)\rho]n_i p_i(1-p_i), \qquad \textbf{\textit{Eq 20.4}}$$

where ρ is the *ICC*, and the overdispersion factor, $\varphi=1+(n_i-1)\rho$, therefore depends on the group size (n_i). The logistic mixed model of Chapter 22 assumes yet another relation where the inflation also depends on the probability p_i. We demonstrate the methods by the simulated binary data with a herd-level predictor (Example 20.4) where there is no within-herd predictor so that the data can therefore be aggregated to the herd level (grouped binary data) without any loss of information.

The advantages of the simple overdispersion approach are its numerical simplicity and its relatively weak assumptions (involving only the variance). You can also use ordinary regression model diagnostics after fitting the model. The size of the overdispersion parameter provides an estimate of the severity of the clustering problem. The disadvantages are: a potential problem in estimating φ (when there is sparse replication), the assumption that the overdispersion is constant (or takes the Williams form) when herd sizes (n_i) differ strongly, the lack of likelihood-based inference and, for binary data, its limitation to grouped (binomial) data (*ie* no within-herd predictors are possible). As noted in the introduction of Section 20.3, using overdispersion more generally to compensate for non-modelled hierarchical clustering is not recommended, particularly because there is little reason to believe that the only effect of clustering is to increase the standard errors.

20.5.4 Robust variance estimation

In a 'usual' regression model (linear, logistic *etc*), the SEs of the coefficients in the model are based upon the assumption that the model is true in all respects and that the errors are independent and follow the appropriate distribution (Gaussian for a linear model) or binomially distributed (for a logistic model). If these assumptions are met and you had an infinite sample, the estimated β would be correct and you would have an SE of zero.

There is an alternative approach to computing the variance (and hence the SE) of β that is referred to as robust variance estimation, or Huber-White variance estimation (Huber, 1967; White, 1980). It also goes under the name 'sandwich' variance estimation because, in matrix notation, the formula for the variance matrix of the βs looks like a sandwich (for the mathematically conversant reader: $A^{-1}BA^{-1}$, where A^{-1} denotes the inverse matrix, and the actual form of the matrices depends on the model where the matrix A^{-1} is usually the model-based variance matrix and B is the correction term (Hardin & Hilbe, 2007). These estimates are less sensitive to the assumptions on which model-based estimation is built (*eg* homoscedasticity in a linear regression model) but they also have a slightly different interpretation. The SEs simply estimate the expected variability in the βs if repeated samples of the same size as the dataset were drawn from the original population, and thus, are somewhat analogous to bootstrap SEs (Guan, 2003). As such, they are more robust to violations of any of the assumptions on which the model is based and usually (but not necessarily) produce larger SEs (and hence, wider CIs) than the usual variance estimates. While robust SEs might also be computed for discrete data, the 'robustness' is less obvious with discrete data because model misspecifications might affect not only the variances but also the estimates themselves.

The robust variance estimate can also be allowed to vary across clusters, which is important when dealing with clustered data, because in this variant, it additionally relaxes the assumption of independence to require only independence of observations across clusters, not within clusters (Froot, 1989; Williams, 2000). A more complete discussion of alternative variance estimation procedures (including sandwich estimators and others) can be found in (Hardin & Hilbe, 2007). We illustrate the cluster-adjusted robust variance method by analysing all simulated datasets from Examples 20.2-3 with this method (Examples 20.4 and 20.5).

The advantages of cluster-adjusted robust variance estimation are that it is simple to use (if implemented by your software) and it does not require specific assumptions about the nature of the clustering. For linear models, it provides SEs that are robust to different violations of the model assumptions (*eg* distribution of errors and heteroscedasticity). One disadvantage of this approach is that it provides no information about the magnitude or origin of clustering. Further, it has no impact on the point estimates of the parameters, which might be considered particularly critical for non-normal data, and the SEs differ in their interpretation from usual SEs. Finally, it should also be said that cluster-adjusted robust variance estimation is part of the generalised estimating equations (GEE) approach to clustered data (Section 23.5) which offers more control over the modelling without requiring additional assumptions.

20.5.5 Survey methods

In Chapter 2, we saw how survey design can be incorporated into estimation of means and proportions from a dataset obtained by a complex sampling design, and noted that the survey analysis approach extends to some regression models, including those covered in Chapters 14-19. Multistage survey designs involve one or several levels of clustering (Section 2.10.3), so it is natural to ask how the methods to account for clustering in the survey and multilevel analysis frameworks relate to each other. A theoretical comparison is beyond the scope of the book, but we will give some pointers and illustrate by the -dairy_dis- example from Chapter 2. A recent review of the multilevel (mixed model) approach for multistage survey data pointed out that caution needs to exercised when the design involves sampling weights (Rabe-Hesketh & Skrondal, 2006). Conversely, the variance linearisation method for survey analysis corresponds to **robust variance** estimation for a two-level structure with clustering at the highest level (Section 20.5.4). Some implementations of survey procedures allow for inclusion of additional levels, but the results can generally be expected to be close to those for robust variance estimation clustered at the highest level.

Other variance estimation procedures exist for survey analysis which have no immediate analogue in multilevel modelling. In particular, for variance estimates based on **replication**, the data are repeatedly split into subsamples and the variance among the subsamples is calculated. (Sampling weights are taken into account in the subsamples.) There are multiple ways of forming the subsamples (random groups, jackknife replication, balanced repeat replication, bootstrap methods). This approach to variance estimation has the advantage that it can be used for any statistics being estimated (no analytical solutions required), but it is computationally intensive.

Survey method analysis of the simulated 2-level datasets we used to illustrate the other approaches to dealing with clustered data, gives the same results as robust variance estimation (Examples 20.4 and 20.5). We illustrate in Example 20.6 the effects of features specific to survey design (sampling weights, stratification and finite population corrections, see Chapter 2) on the results of a logistic regression to evaluate the effect of two predictors (lactation number and herd-prevalence of leukosis) on the risk of a cow testing positive for *Neospora caninum.* It is intuitively obvious that incorporation of sampling weights will change both estimates and standard errors whenever observations represent different numbers of sampling units in the population. Ignoring sampling weights yields estimates for the study sample instead of the source population. Some (non-survey) statistical software will allow for weights in the estimation procedure, but one needs to be careful that these weights have the same meaning (effect) as the sampling weights. Stratification is essentially another weighting scheme whereby strata are given weights relative to their proportion of the total population. This will generally not give the same results as including the strata as fixed effects. Stratification is exclusive to survey analysis, as is finite population corrections, which will only affect the standard errors.

In summary, survey analysis procedures offer unique possibilities to incorporate features of a complex sampling design. For a design involving only clustering, the results are comparable to those of robust variance estimation, and the survey methodology offers no substantial added advantages.

20.5.6 Summary of clustered data methods

A variety of approaches for dealing with clustered data has been presented (with mixed models to follow in Chapters 21-22 and the GEE procedure in Chapter 23). We have illustrated some of the differences between the methods by the comparative tables of estimates for the simulated datasets in Examples 20.4 and 20.5, and conclude with a summary table for the methods covered. Bayesian methods are reviewed in Chapter 24. Table 20.4 gives only a very brief summary; consult the respective sections for details.

Example 20.6 Survey and multilevel analysis for complex survey data

data = dairy_dis

The probability of a cow testing positive for *Neospora caninum* was evaluated using a logistic regression model with both a cow-level predictor (lactation number) and a herd-level predictor (proportion of the herd positive for leukosis). The estimated coefficient and SE are presented for versions of survey analysis incrementally incorporating more features of the sampling design.

Survey analysis	**Lactation number**		**Leukosis prevalence**	
	Estimate	**SE**	**Estimate**	**SE**
(1): herd clustering	-0.077	0.032	0.889	0.442
(2): (1) + sampling weights	-0.103	0.038	0.885	0.417
(3): (2) + fixed province effects	-0.095	0.041	0.771	0.419
(4): (2) + province stratification	-0.103	0.038	0.885	0.420
(5): (4) + finite population corr.	-0.103	0.037	0.885	0.397

The first 3 analyses can also be obtained by robust variance estimation in logistic regression. However, the P-values for tests will differ slightly, because the survey procedures in Stata use reference *F*- or *t*-distributions whereas, the logistic regression procedure uses chi-square and Z-distributions. It is seen that ignoring the sampling weights has a major impact on the estimate for the cow-level predictor (lactation number). Modelling province by fixed effects strongly affects the estimate for the herd-level predictor. The finite population corrections reduce the standard errors slightly. Analysis (5) is preferred because it incorporates all features of the sampling design.

Table 20.4 Summary of approaches for clustered data

	Properties/Features				
Method to account for clustering [VER coverage]	**Adjusted SE**	**Adjusted β**	**>1 level of clustering**	**Estimate of ρs**	**Comments on scope or use of method**
Fixed effects [20.5.2]	yes[a]	yes	no	no	no cluster-level predictors
Stratification [20.5.2]	yes	yes	no	no	specific designs (binary data)
Overdispersion correction [20.5.3]	yes	no	no	no	no within-cluster predictors, not for continuous (normal distribution) data
Robust SE (clustered) [20.5.4]	yes	no	(no)	no	also adjusts for other model violations (continuous data)
Survey methods [20.5.5]	yes	(no)	yes	no	additional features (sampling weights, stratification)
Linear mixed model [21]	yes	yes	yes	yes	continuous (normal distribution) data
Discrete mixed model (GLMM) [22]	yes	yes	yes	yes	cluster-specific (SS) parameters
Generalised estimating equations (GEE) [23]	yes	yes	(no)	(yes)	population-averaged (PA) parameters (discrete data)
Bayesian mixed model (continuous/discrete) [24]	yes	yes	yes	yes	different statistical approach, additional components of analysis

[a] Table contents reflect attributes of each method. For example, fixed effects models do adjust SEs and coefficients (β) but cannot handle more than one level of clustering or provide an estimate of ICC (ρ)

Note GEE method yields correlations as part of working correlation matrix, alternating logistic regression version of GEE for binary data allows for two levels of clustering, and both GEE and robust variance methods can adjust for multiple levels provided a sufficient number of units at highest level

REFERENCES

Aunsmo A, Ovretveit S, Valle P, Larssen R, Sandberg M. Modelling sources of variation and risk factors for spinal deformity in farmed Atlantic salmon using hierarchical and cross-classified multilevel model. Prev Vet Med. 2009;90 137-145: .

Browne W, Goldstein H, Rasbach J. Multiple membership multiple classification (MMMC) models Stat Modelling. 2001; 1: 103-24.

Clarke P. When can group level clustering be ignored? Multilevel models versus single-level models with sparse data J Epidemiol Community Health. 2008; 62: 752-8.

Collett D. Modelling Binary Data, 2nd Ed. Chapman & Hall/CRC; London. 2002.

Dohoo I, Stryhn H. Simulation studies on the effects of clustering. In: Proceedings of the 11th Symposium of the International Society for Veterinary Epidemiology and Economics, Cairns. 2006. p. 505.

Dohoo IR, Tillard E, Stryhn H, Faye B. The use of multilevel models to evaluate sources of variation in reproductive performance in dairy cattle in Reunion Island Prev Vet Med. 2001; 50: 127-44.

Donner A. The comparison of proportions in the presence of litter effects Prev Vet Med. 1993; 18: 17-26.

Froot K. Consistent covariance matrix estimation with cross-sectional dependence and heteroskedasticity in financial data Journal of Financial and Quantitative Analysis. 1989; 24: 333-55.

Guan W. From the help desk: Bootstrapping standard errors The Stata J. 2003; 3: 71-80.

Hardin J, Hilbe J. Generalized Linear Models and Extensions, 2nd Ed. Stata Press; College Station. 2007.

Hilbe J. Logistic Regression Models. CRC Press; Boca Raton. 2009.

Huber P. The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: University of California Press, Berkeley. 1967. p. 221-223.

Liang KY, Zeger SL. Regression analysis for correlated data Annu Rev Public Health. 1993; 14: 43-68.

McCullagh P, Nelder J. Generalized Linear Models, 2nd ED. Chapman & Hall; London. 1989.

Mead R. The Design of Experiments: Statistical Principles for Practical Applications. Cambridge University Press; Cambridge. 1990.

Okoumunne O, Gulliford M, Chinn S. A note on the use of the variance inflation factor for determining sample size in cluster randomized trials The Statistician. 2002; 51: 479-84.

Otte MJ, Gumm ID. Intra-cluster correlation coefficients of 20 infections calculated from the results of cluster-sample surveys Prev Vet Med. 1997; 31: 147-50.

Rabe-Hesketh S, Skrondal A. Multilevel modelling of complex survey data J Royal Stat Soc, Series A. 2006; 169: 805-27.

Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data Acad Emerg Med. 2002; 9: 330-41.

White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity Econometrica 1980; 48:817-830.

Williams RL. A note on robust variance estimation for cluster-correlated data Biometrics. 2000; 56: 645-6.