MEASUREMENT ERROR IN SLICING ANALYSIS

JAN DE LEEUW

CONTENTS

1. SLICING ANALYSIS

In a *slicing analysis* we look at the distribution of an outcome variable in various slices of the population of students. In the example we are interested in here, for instance, slices are defined by *test score* and *race*, and the outcome variable is some form of *placement* of the student. Somewhat more generally, we will look at examples in which we study relationships between an outcome variable Y, a covariate X, and a grouping variable Z. Both X and Y can be multivariate. Thus slicing analysis is a particular (often tabular) form of the *analysis of covariance*.

More formally, we study the conditional distribution[1] of $\underline{y}$ given $\underline{x}$ and $\underline{z}$, or, equivalently, the regression of $\underline{y}$ on $\underline{x}$ and $\underline{z}$. The joint probability distribution[2] $\pi(xyz)$ is decomposed as

$$\pi(yxz) = \pi(y|xz)\pi(xz) \tag{1.1}$$

We can also present the model in a simple path diagram.

Date: March 11, 2001.

[1]We use the convention of underlining random variables.

[2]This can be either a discrete distribution – a table with probabilities –, a continuous density, or a combination of the two.

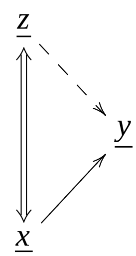

FIGURE 1. Placement as function of test and race

For the moment, ignore the fact that one of the single arrows is dashed and the other one is solid.

The path diagram, and the equivalent decomposition in Equation (1.1), define a *saturated* model, i.e. a model which always fits the data perfectly, no matter what the data are. Thus we do not have to fit or test, we just present the data in the form of the joint marginal distribution of $\underline{x}$ and $\underline{z}$ and the conditional distribution of $\underline{y}$ in each of the "cells" defined by $\underline{x}$ and $\underline{z}$.

2. THE DATA

Data for the example are taken from Oakes [2000]. They were provided by the Rockford, Illlinois, school district in connection with a desegregation case. The table on the left gives the marginal distribution (of student-placements) for deciles of the SAT9 test and for the minority and majority groups. The table on the right gives the probability of placement in advanced classes in 1999-2000 for each of the cells.

Dec	Maj	Min
01	633	1212
02	941	1207
03	1166	1103
04	1492	1109
05	1367	667
06	1819	650
07	1788	446
08	2346	377
09	2271	227
10	2124	98

Dec	Maj	Min
01	0.03	0.02
02	0.05	0.04
03	0.07	0.07
04	0.15	0.13
05	0.20	0.15
06	0.31	0.24
07	0.46	0.37
08	0.61	0.59
09	0.74	0.70
10	0.86	0.81

TABLE 1. Joint marginal (left) and conditional distribution (right) for Rockford data

It is important to see that no assumptions are involved in this "analysis". It is descriptive, it merely presents the actual data. One can make the argument that the tables are misleading, because they do not clearly show sampling distributions of the probabilities, and some of them are measured much more precisely than others. One could also argue that these "raw" tables could perhaps be smoothed somewhat to surpress the effect of random fluctuations, especially in the cells with a small number of observations. We shall address some of these concerns below.

The actual data can also be presented graphically. This is done in Figure 2, which shows the proportions of placement in more advanced courses for the 10 deciles of the SAT9.

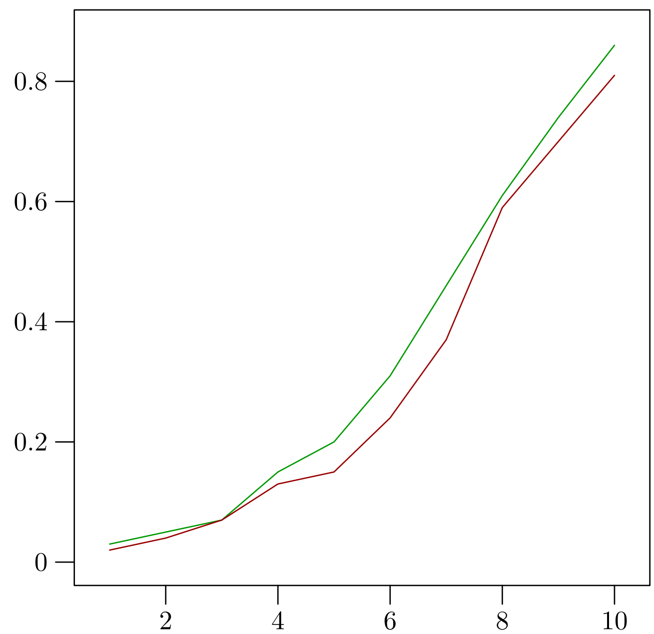

FIGURE 2. Conditional distributions for majority (green) and minority (red) students

3. TESTING

The arrow from $\underline{z}$ to $\underline{y}$ in Figure 1 is dashed. This is because in our example we are interested in the (more restrictive) model in which group membership $\underline{z}$ is not relevant for placement, i.e.

$$\pi(xyz) = \pi(y|x)\pi(xz). \tag{3.1}$$

In this model, the dashed arrow is missing.

In specific situations the model implies various testable consequences.

- If the regressions are linear, then the coefficient of $\underline{z}$ in the regression of $\underline{y}$ on $\underline{x}$ and $\underline{z}$ should be zero.

- If variables are discrete, then we can form cross tables of $\underline{y}$ and $\underline{x}$ for each value of $\underline{z}$. Suppose $\underline{x}$ defines rows and $\underline{y}$ defines columns of these tables, and normalize the tables such that rows add up to one. Then each of the tables should be the same.

In our example the missing arrow means that the two columns of the right part of Table 1 are equal. We can test this is many ways, either by testing the equality of the probabilities themselves, or by testing the equality of some transform of the probabilities. Because of other computations in this report, we compute the inverse probit transform $\Phi^{-1}(p)$ and its estimated standard errors. The delta-method approximation to the sampling variance of a probit is given by

$$\mathbf{V}(\Phi^{-1}(\underline{p})) = \frac{\pi(1-\pi)}{n\phi^2(\Phi^{-1}(\pi))}, \tag{3.2}$$

where π is the true probability of success and n is the number of trials. We can estimate these variances by substituting sample proportions p for π, and then compute estimated standard errors by taking square roots.

The results are given in Table 2.

Dec	Maj	Min
01	-1.88	-2.05
02	-1.64	-1.75
03	-1.48	-1.48
04	-1.04	-1.13
05	-0.84	-1.04
06	-0.50	-0.71
07	-0.10	-0.33
08	0.28	0.23
09	0.64	0.52
10	1.08	0.88

Dec	Maj	Min
01	0.0996	0.0830
02	0.0689	0.0655
03	0.0557	0.0572
04	0.0398	0.0477
05	0.0386	0.0593
06	0.0307	0.0539
07	0.0297	0.0605
08	0.0262	0.0652
09	0.0284	0.0875
10	0.0338	0.1460

TABLE 2. Probits (left) with standard errors (right) for Rockford Data

We can now also compute the difference of the probits for majority and minority students, and their standard errors, z-values, and significance probabilities. These are given in Table 3. If we summarize the information in this table, and test the hypothesis that the majority and minority columns are the same, we find a chi-square of 40.01 with 10 degrees of freedom, which gives a p-value less that .0001.

Dec	Diff	StErr	z-value	p-value
01	0.17	0.130	1.33	0.091
02	0.11	0.095	1.11	0.132
03	0.00	0.080	0.00	0.500
04	0.09	0.062	1.45	0.074
05	0.19	0.071	2.75	0.003
06	0.21	0.062	3.39	0.000
07	0.23	0.067	3.43	0.000
08	0.05	0.070	0.74	0.231
09	0.12	0.092	1.29	0.098
10	0.20	0.150	1.35	0.088

TABLE 3. Probit comparisons per slice

4. PROBIT ANALYSIS

Working with saturated models has the advantage that we stay close to the data, and we do not have to make assumptions that we do not feel comfortable about. On the other hand, using restrictive models often gives us more power, fewer parameters, and more ease of interpretation.

In this section we reduce the number of parameters from twenty (the number of probits) to eleven, by using the model

$$\pi(1|xz) = \Phi(\alpha(z) + \gamma(x)). \tag{4.1a}$$

This model says that the 10×2 Table 2 of probits is additive, i.e. there is not interaction between test score and ethnicity.

We now further reduce the number of parameters to four by using the model

$$\pi(1|xz) = \Phi(\alpha(z) + \beta(z)x), \tag{4.1b}$$

and then to three by using

$$\pi(1|xz) = \Phi(\alpha(z) + \beta x), \tag{4.1c}$$

Models (4.1b) and (4.1c) say that probits are linear with deciles. In (4.1b) the two ethnicity groups have different slopes and intercepts, in (4.1c) they have different intercepts but the same slope. This is the probit version of the analysis of covariance.

5. MEASUREMENT ERROR

Slice analyses (and similar group comparison techniques) have been criticized because they do not take measurement error in the predictors into account [Lord, 1960; Cronbach et al., 1977]. In our example, there is no error in the grouping variable, but the test score, of course, supposedly measures an underlying achievement variable. And it does so imperfectly. Once again, we capture this idea in a path diagram.

Suppose $\underline{\xi}$ is the *true score* of a test, and $\underline{x}$ is the observed score. Moreover $\underline{y}$ is the placement variable, and $\underline{z}$ is the grouping variable. We now draw the saturated model

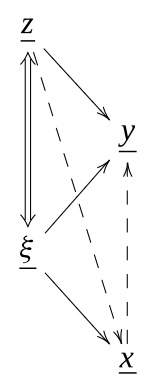

FIGURE 3. Placement and test as functions of true score and race

This corresponds with the decomposition

$$\pi(yx\xi z) = \pi(y|x\xi z)\pi(x|\xi z)\pi(\xi|z)\pi(z). \tag{5.1}$$

Although we refer to this model as a saturated model, it is different from the model in Section 1. One of the variables in Figure 3 is unobserved (missing), and although the model is saturated, it is sadly under-identified. In fact, we can take whatever we want for $\underline{\xi}$ and still fit the saturated model perfectly. This means we have to fill in the missing information by making various restrictive assumptions. They are needed to identify the model and to make estimation possible.

First, there are some dotted arrows in Figure 3. We are interested in the unsaturated model in which these arrows disappear. The arrow from $\underline{x}$ to $\underline{y}$ is missing, which means we assume there is no *direct* effect of observed test score on placement. Knowing the true score is enough. In systems in which the teachers or counselors use the actual test score for placement this assumption is probably false. In our example, and in general, there is no way to falsify the assumption – we merely can talk about its plausability. The most we can say, perhaps, is that it is in the psychometric tradition.

The same thing is true for the second dashed arrow, from $\underline{z}$ to $\underline{x}$. This arrow, about which we will have to say more below, is very critical. It says that $\pi(x|\xi z) = \pi(x|\xi)$, which means that the observed test score only depends on the true score, not on the group. This is obviously again solidly in the psychometric tradition, but the plausibility of the assumption is highly debatable. The whole issue of cultural or racial bias comes into play here. Nevertheless we do make the assumption to identify our model.

With the two arrows deleted, our model becomes

$$\pi(yx\xi z) = \pi(y|\xi z)\pi(x|\xi)\pi(\xi|z)\pi(z). \tag{5.2}$$

This is still not enough in terms of assumptions. We have to make parametruic assumptions on the distribution of the latent achievement variables, and again we are guided by the psychometric tradition. We suppose $\pi(x\xi|z)$ is bivariate normal[3] for each group z. And we suppose $\pi(y|\xi z)$ is given by a linear probit model, i.e.

$$\pi(1|\xi z) = \Phi(\alpha(z) + \beta(z)\xi), \tag{5.3}$$

which $\Phi(\bullet)$ the normal ogive. Both slope $\beta(z)$ and intercept $\alpha(z)$ of the regression of the probit of placement on the true score can differ for the two groups.

Again, these assumptions are mostly made to make estimation possible. They are classical, and they are the basis of test and latent trait theory, but whether they are plausible or not is a matter of opinion.

We now study the model in a bit more detail. It follows from (5.2) that

$$\pi(1|xz) = \int_{-\infty}^{+\infty} \pi(1|\xi z)\pi(\xi|xz)d\xi. \tag{5.4}$$

Since $\pi(x\xi|z)$ is bivariate normal, it follows that $\pi(\xi|xz)$ is normal as well. After some routine, but tedious, manipulation we find

$$\pi(1|xz) = \Phi\left(\frac{\alpha(z) + \beta(z)\mathbf{E}(\underline{\xi}|xz)}{\sqrt{1 + \beta^2(z)\mathbf{V}(\underline{\xi}|xz)}}\right). \tag{5.5}$$

Because of normality, we know that we can write

$$\mathbf{E}(\underline{\xi}|xz) = \lambda(z) + \theta(z)x, \tag{5.6a}$$

$$\mathbf{V}(\underline{\xi}|xz) = \omega^2(z), \tag{5.6b}$$

which implies

$$\pi(1|xz) = \Phi\left\{\frac{\alpha(z) + \beta(z)\lambda(z)}{\sqrt{1 + \beta^2(z)\omega^2(z)}} + \frac{\beta(z)\theta(z)}{\sqrt{1 + \beta^2(z)\omega^2(z)}}x\right\}. \tag{5.7}$$

This is still a linear probit model, but the two slopes and two intercepts are now functions of ten parameters. Thus we still have a serious case of

[3] We ignore the fact, for the moment, that $\underline{x}$ in our data has been discretized.

under-identification. The only way out of this problem is, of course, to make additional assumptions.

Appendix A. Evaluation of an Integral

In this appendix we evaluate

$$\mathcal{I}(\alpha,\beta) \triangleq \int_{-\infty}^{+\infty} \Phi(\alpha+\beta x)\phi(x)dx, \tag{A.1}$$

where $\Phi(\bullet)$ and $\phi(\bullet)$ are the standard normal cdf and pdf, respectively.

First,

$$\Phi(\alpha+\beta x) = \int_{-\infty}^{\alpha+\beta x} \phi(y)dy = \int_{-\infty}^{0} \phi(y+\alpha+\beta x)dy. \tag{A.2}$$

Thus

$$\mathcal{I}(\alpha,\beta) = \int_{-\infty}^{+\infty}\int_{-\infty}^{0} \phi(y+\alpha+\beta x)dy\phi(x)dx. \tag{A.3}$$

The quadratic form in the exponent is

$$\begin{aligned}(y+\alpha+\beta x)^2 + x^2 = \\ (\beta^2+1)x^2 + 2\beta(y+\alpha)x + (y+\alpha^2) = \\ (\beta^2+1)\left(x+\frac{\beta(y+\alpha)}{\beta^2+1}\right)^2 + \frac{(y+\alpha)^2}{\beta^2+1}.\end{aligned} \tag{A.4}$$

Thus we arrive at

$$\mathcal{I}(\alpha,\beta) = \frac{1}{\sqrt{\beta^2+1}}\int_{-\infty}^{0}\phi\left(\frac{y+\alpha}{\sqrt{\beta^2+1}}\right)dy = \Phi\left(\frac{\alpha}{\sqrt{\beta^2+1}}\right). \tag{A.5}$$

Also, by a simple linear change of variables, we have the more general result

$$\int_{-\infty}^{+\infty} \Phi(\alpha+\beta x)\phi_{\mu,\sigma^2}(x)dx = \Phi\left(\frac{\alpha+\beta\mu}{\sqrt{\beta^2\sigma^2+1}}\right). \tag{A.6}$$

Appendix B. Selection

In the previous section we made the assumptions

$$\pi(x|\xi z) = \pi(x|\xi), \tag{B.1}$$

$$\pi(x\xi|z) \text{ is bivariate normal.} \tag{B.2}$$

In this section we work out the consequences of these assumptions. We do this in more generality than is actually necessary, and we specialize to the bivariate normal situation that interests us here at the end.

From (B.1)

$$\pi(x\xi|z) = \pi(x|\xi)\pi(\xi|z). \tag{B.3}$$

This implies for the expected values

$$\mathbf{E}(\underline{x}|z) = \mathbf{E}(\mathbf{E}(\underline{x}|\underline{\xi})|z), \tag{B.4a}$$

and for the variances

$$\mathbf{V}(\underline{x}|z) = \mathbf{E}(\mathbf{V}(\underline{x}|\underline{\xi})|z) + \mathbf{V}(\mathbf{E}(\underline{x}|\underline{\xi})|z). \tag{B.4b}$$

Finally. for the covariances.

$$\mathbf{C}(\underline{x},\underline{\xi}|z) = \mathbf{C}(\underline{\xi},\mathbf{E}(\underline{x}|\underline{\xi})|z). \tag{B.4c}$$

Now assume linear homoscedastic regression of $\underline{x}$ on $\underline{\xi}$. Thus

$$\mathbf{E}(\underline{x}|\underline{\xi}) = A\underline{\xi} + b, \tag{B.5a}$$

and

$$\mathbf{V}(\underline{x}|\underline{\xi}) = D, \tag{B.5b}$$

where D does not depend on $\underline{\xi}$. This is weaker than assuming joint multivariate normality, but it leads to important simplifications. We find

$$\mathbf{E}(\underline{x}|z) = A\mu(z) + b, \tag{B.6a}$$

$$\mathbf{V}(\underline{x}|z) = D + A\Sigma(z)A', \tag{B.6b}$$

$$\mathbf{C}(\underline{x},\underline{\xi}|z) = A\Sigma(z), \tag{B.6c}$$

where we have used $\mu(z) \stackrel{\Delta}{=} \mathbf{E}(\underline{\xi}|z)$ and $\Sigma(z) \stackrel{\Delta}{=} \mathbf{V}(\underline{\xi}|z)$.

We can rewrite our results by introducing means and dispersions for the total (marginal, unselected) population. Then

$$A = \Sigma_{x\xi}\Sigma_{\xi\xi}^{-1}, \tag{B.7a}$$

$$b = \mu_x - A\mu_\xi, \tag{B.7b}$$

$$D = \Sigma_{xx} - \Sigma_{x\xi}\Sigma_{\xi\xi}^{-1}\Sigma_{\xi x}, \tag{B.7c}$$

and thus

$$\text{(B.8a)} \qquad \mathbf{E}(\underline{x}|z) = \mu_x + \Sigma_{x\xi}\Sigma_{\xi\xi}^{-1}(\mu(z) - \mu_\xi),$$

$$\text{(B.8b)} \qquad \mathbf{V}(\underline{x}|z) = \Sigma_{xx} - \Sigma_{x\xi}\Sigma_{\xi\xi}^{-1}(\Sigma_{\xi\xi} - \Sigma(z))\Sigma_{\xi\xi}^{-1}\Sigma_{\xi x},$$

$$\text{(B.8c)} \qquad \mathbf{C}(\underline{x}, \underline{\xi}|z) = \Sigma_{x\xi}\Sigma_{\xi\xi}^{-1}\Sigma(z).$$

This is the *Pearson-Aitkin-Lawley Selection Theorem*. The classical references are Pearson [1903]; Aitkin [1934]; Lawley [1943-44]. Usually the theorem is used in a slightly different fashion, to deduce the unknown moments of the unselected population from the known moments of the selected population. Compare Birnbaum et al. [1950], for instance, for a clear discussion of its standard applications.

If we apply the theorem to a standard bivariate normal distribution, with correlation ρ, then

$$\text{(B.9a)} \qquad \mathbf{E}(\underline{x}|z) = \rho\mu(z),$$

$$\text{(B.9b)} \qquad \mathbf{V}(\underline{x}|z) = (1 - \rho^2) + \rho^2\sigma^2(z)),$$

$$\text{(B.9c)} \qquad \mathbf{C}(\underline{x}, \underline{\xi}|z) = \rho\sigma^2(z).$$

In our example ρ is the *reliability index*, the square root of the *reliability*.

Thus the joint conditional distribution of $\underline{x}$ and $\underline{\xi}$ is

$$\text{(B.10)} \qquad \begin{bmatrix} \underline{x}|z \\ \underline{\xi}|z \end{bmatrix} \sim \mathcal{N}\left\{\begin{bmatrix} \rho\mu(z) \\ \mu(z) \end{bmatrix}, \begin{bmatrix} (1-\rho^2) + \rho^2\sigma^2(z)) & \rho\sigma^2(z) \\ \rho\sigma^2(z) & \sigma^2(z) \end{bmatrix}\right\},$$

and $\pi(\xi|xz)$ is normal with mean

$$\text{(B.11a)} \qquad \mathbf{E}(\underline{\xi}|xz) = \frac{(1-\rho^2)\mu(z) + \rho^2\sigma^2(z)x}{(1-\rho^2) + \rho^2\sigma^2(z))},$$

and variance

$$\text{(B.11b)} \qquad \mathbf{V}(\underline{\xi}|xz) = \frac{(1-\rho^2)\sigma^2(z)}{(1-\rho^2) + \rho^2\sigma^2(z))}.$$

REFERENCES

A.C. Aitkin. Note on selection from a multivariate normal population. *Proceedings Edinburgh Mathematical Society*, 4:106–110, 1934.

Z.W. Birnbaum, E. Paulson, and F.C. Andrews. On the ffect of selection performed on some coordinates of a multi-dimensional population. *Psychometrika*, 15:191–204, 1950.

L.J. Cronbach, D.R. Rogosa, R.E. Floden, and G.G. Price. Analysis of covariance in nonrandomized experiments: parameters affecting bias. Technical report, Stanford Evaluation Consortium, 1977.

D.N. Lawley. A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh, Section A*, 62:28–30, 1943-44.

F.M. Lord. Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, 55:307–321, 1960.

J. Oakes. Within-school integration, grouping practices, and educational quality in Rockford schools. Technical report, Graduate School of Education and Information Studies UCLA, 2000.

K. Pearson. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society, Section A*, 200:1–66, 1903.

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES
E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu