Chapter 9
Statistical Editing

Statistical editing or data editing is a big and important part of a survey process. If raw data has not been edited in any way, editing may require much time and resources. This is typically the case if a paper questionnaire is used, but all other modes give opportunity to make some pre-editing and hence final editing is obviously easier. This chapter presents the core methods and tools for editing, but we start from the main purpose of it

Statistical editing thus is a crucial part of quality assurance of the survey data and also the survey process. The first point is close to quality control and improvement of the data, and we focus here on it. However, it is also needed to look forward and to document strong and weak points of the survey process, that is, to learn about them for future surveys. This has been too often forgotten and hence the same errors or mistakes are repeated. It is good also to look at other similar types of surveys and how they have edited and to use this information in a best way.

Specific tasks of statistical editing are as follows:

• Evaluate and develop the survey process for the future, as learning by doing. It is good to follow what happens in other surveys similar to this particular one.

• Develop the system that helps in reducing manual work in editing, using selective editing, for instance. Its aim is to concentrate on detecting most fatal errors of the data, such whose impact on data quality is big.

• Detect, check and correct errors of micro level so that the results from a macro (aggregate) level are also plausible and reliable.

• Pay special attention to missing values so that they are coded with as many codes as it is possible, or left as missing. At the same time it is should be decided preliminarily what to do for missing values in data analysis. If you decide to impute those partially or entirely, this decision is good to do now since you also need to think which auxiliary variables could be used in imputation.

• Provide indicators that tell about changes made in editing and how some core estimates have been revised ('improved') due to editing. Estimate also the workload of the editing (and imputation).

Edit Rules

Edit rules are the rules for checking the correctness of individual variable values.

These may be more or less strict; this can be determined by gates given for each value checking. If the gate is narrow, it thus is more strictly checked than in the other cases, thus if the gate is broad. The workload of editing thus much depends on the broadness of gates in editing.

On the other hand, the quality is expected to be better if the gate is narrower. But: all depends on how well a possibly erroneous value can be corrected. If all suspect values are possible to check from the respondent (using the list of suspect values), it is very fine but it is not possible in most human surveys. Fortunately, it is often possible in the case of big businesses of business surveys.

If suspect values cannot be checked against true values, the only strategy is to make them believable, plausible or logical. This means first that they should be at a correct level or **(i) within the predetermined range of each variable.** This is fairly easy to do at data entry already, thus in pre-editing. This is the first edit rule and should always be followed but it is not clear which values only can be accepted.

It is possible to give different acceptable values for different sub-groups such as gender, age or education. This is **(ii) the second edit rule, thus so that one value may depend on the value of another variable**. It is a bit more demanding in computer-assisted surveys but possible. It is danger that a respondent does not like if his/her answer has not accepted due to his/her another answer. Hence it should be used carefully in pre-editing, but in the stage of post-fieldwork editing it should be done but the solution is not maybe always nice since it is needed to change one of both values in order to pass this edit rule.

To generalize, the third level edit rule thus can be one-, two-, three-, and multidimensional. The number of suspect values is obviously growing at the same time, consequently the workload for checking and correcting implausible values. Usually, the logics of different values seems to be most important than that the value is right absolutely.
For example:

- If the age of a person is 10, and his/her has a child, it is maybe best to change the age but to keep the child there.
- If the age of a person is 20 and he/she is university professor, either the age is wrong or the occupation is wrong.
- If a person is unemployed but the wage is 5000 €, one of these values is obviously wrong. It is good to look other answers as well before correcting one value.
- ...

Other edit checks

Identifiers may have a big role in surveys, for survey institution people in particular. These are unique identifiers such as personal identity code or business entity code. An identifier may consists of several variables as well such as firstname, secondname, birthdate or birthyear. All these are confidential and should not be given outsiders without the permission. On the other hand, they should be correct, and hence checked and corrected if needed. All correct identifiers should be maintained in the survey institution as long as needed.

Moreover, the confidential identifiers are converted to a protected form into the file given outsiders. This conversion rule should be saved since it is possibly needed later. Some type of randomization is good to use in this conversion.

Due to mistakes in data entry it is possible that after merging two data for example, two same identifiers = duplications are in a new file. One of these should be deleted. This is part of editing work.

Extreme or other exceptional values may be awkward. They are often called **Outliers**. On the other hand, the data file consists also of **Inliers** that look as ordinary values but if controlling them by one or more other variables, they do not look anymore ordinary. In editing, it is most important to detect those of these that are not correct. An erroneous outlier is called **Out-Error**, and the inlier **In-Error**, respectively. Hopefully, such errors do not exist in the cleaned data.

Unless the questionnaire already and the data entry has not been given good codes for missing values, this coding should be made at the editing stage.

Selective editing

Selective editing is much used in business surveys where the fact variables are common. It is possible to use also for other variables. There are several approaches to selective editing but the basic idea is to construct a model (that can be a statistical model or a mathematical model as a function) that predicts the probability that a certain value is erroneous (called error-localization). In editing, the values with highest error probabilities are first checked and corrected most carefully, and those with low probability less carefully or even left as such or corrected automatically.

When developing the selective editing model, it is good to train it with the so-called training data set in which the workability of the model has been first checked against a smaller data set, often more manually. And when it has been found that the model works, it has been run over the whole data.

Although a data file such as the ESS micro file should be edited and thus cleaned, do not believe it completely. For example, this editing has only be based on one-dimensional looking/checking that is easist. Two examples below.

Figure 9.1 Graphical presentation

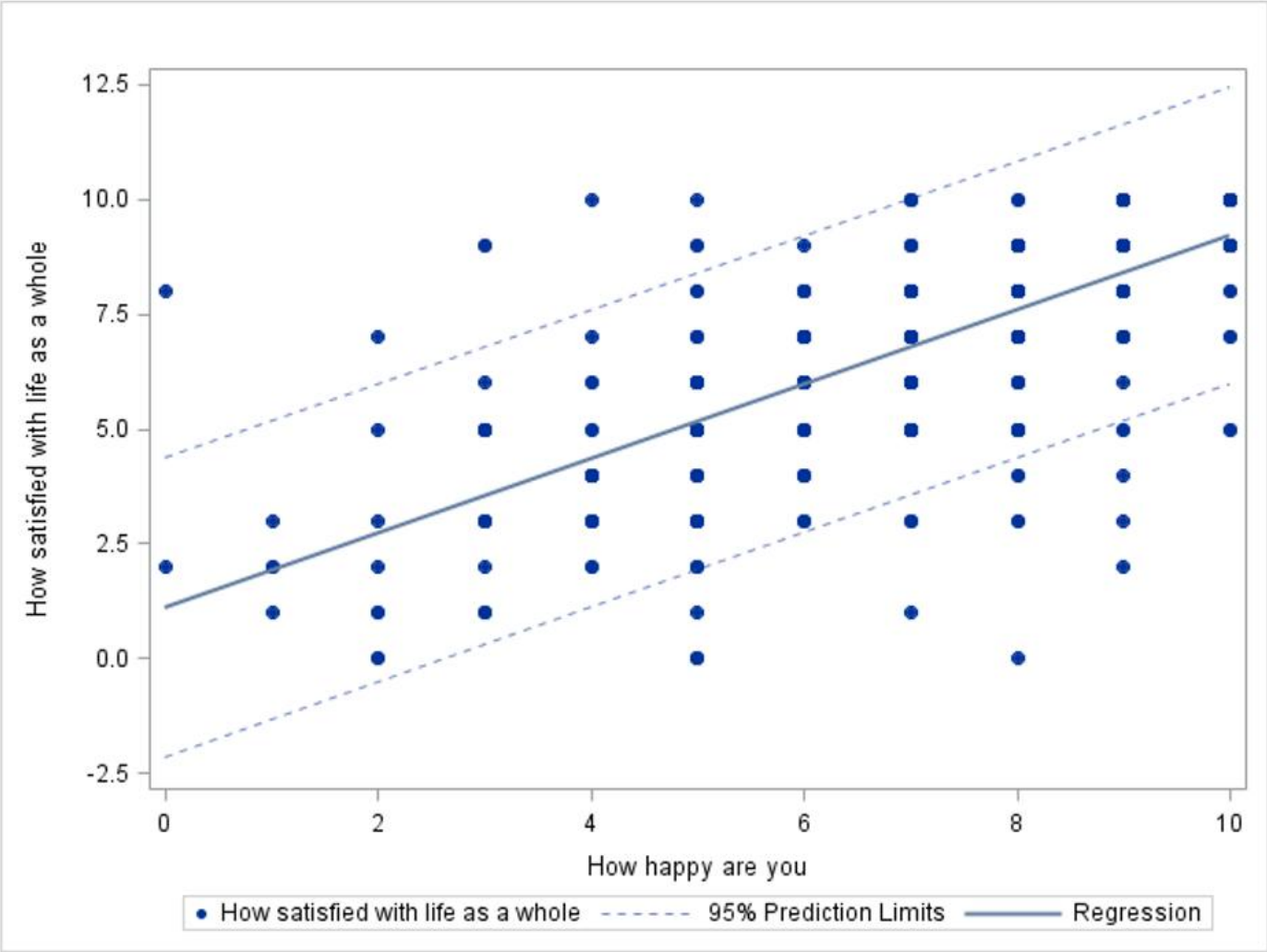
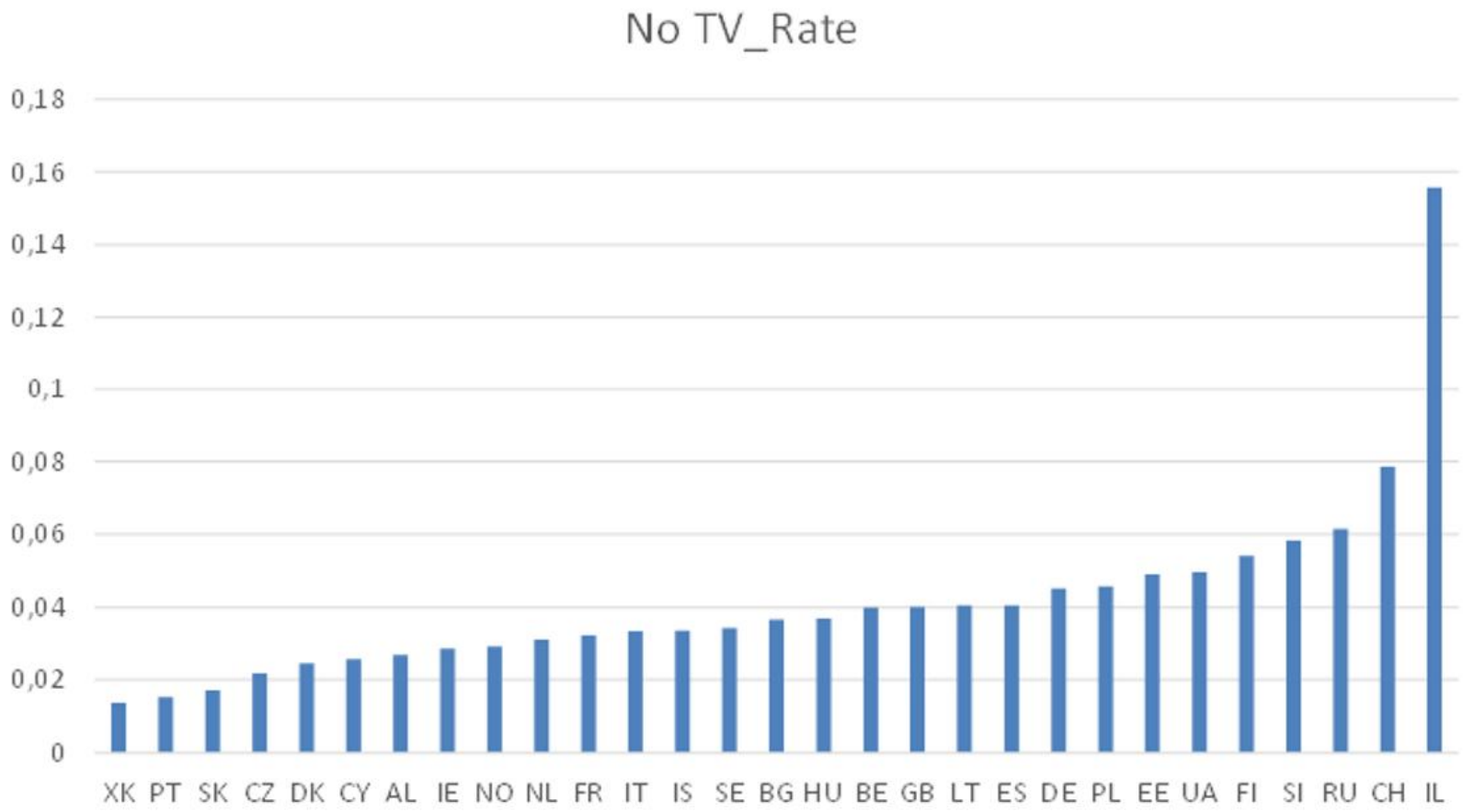


Table 9.3 Cross-tabulation of the ESS data by TV watching

TV watching, total time on average weekday)	TV watching, news/politics/current affairs on average weekday)										
	0 Not at all	1 Less than 0.5 hours	2 0.5 to 1 hour	3 More than 1 hour up to 1.5 hours	4 More than 1.5 hours up to 2 hours	5 More than 2 hours up to 2.5 hours	6 More than 2.5 hours up to 3 hours	7 More than 3hours	66 Not appli-cable	77 Refusal	88 Don't know
0	0	0	0	0	0	0	0	0	2342	0	0
1	664	2204	85	24	14	3	3	9	0	1	15
2	826	3350	2552	84	25	5	11	13	0	0	16
3	605	2887	2867	948	58	17	5	8	0	0	14
4	617	2744	3670	1075	581	26	15	8	0	0	21
5	426	1784	2882	1283	489	313	31	16	0	0	15
6	330	1378	2605	1225	618	257	277	27	0	0	14
7	688	1801	3780	2272	1375	663	454	971	0	0	45
77	0	0	4	0	0	0	0	0	0	1	0
88	20	46	39	11	6	2	0	4	0	0	57
99	2	6	11	2	1	1	0	1	1	0	0
Total	4178	16200	18495	6924	3167	287	796	1057	2343	2	197

Figure 9.4 ***ESS countries by the answer 'Not applicable' rate (No TV rate) for the question on watching TV news and politics and current affairs.***



Most rates are rather low but that for Israel is like an outlier. We do not try to interpret this result but it is good to do in proper survey analysis.