Table 2.1 Small example showing a case of successful attribute disclosure

OBS	Key variables			Sensitive variable
	Race	Age	Region	Religion
1	Black	50–60	1234	Roman/catholic
2	Black	50–60	1234	Roman/catholic
3	Black	50–60	1234	Roman/catholic
4	Black	50–60	1234	Roman/catholic

Another example: If a hospital publishes data showing that all female patients aged 56 to 60 have cancer, an intruder then knows the medical condition (=cancer) of any female patient aged 56 to 60 staying in this hospital without having to identify the specific individual.

2.2.3 Inferential Disclosure

Inferential disclosure occurs when the intruder is able to determine the value of some sensitive characteristic of an individual more accurately with the released data than it would have been possible otherwise. Inferential disclosure happens when individual's sensitive characteristics can be well predicted from a good model applied on the released data.

For example, with a highly predictive regression model, an intruder will be able to infer a respondents sensitive income information using attributes recorded in the data, leading to inferential disclosure.

In practice, a model would be fit onto the released data and external information on an individual would be used to predict attributes of this individual. For example, assume that *age*, *gender*, *region*, *economic status*, *economic status* and *income* are available in released data. The intruder also has information on the first five mentioned variables on a particular person, say *A*. He can fit a regression model with income as response on the released data and he will receive the fitted regression coefficients. A linear combination of these coefficients with the values on person A predicts the income of this individual. Depending on the quality of the model, the income might be fitted accurately enough. If this is the case, the intruder successfully disclosed the income of person A.

As an example, we use the EU-SILC data set from R package **laeken** (Alfons and Templ, 2013). This data set is synthetically generated. However, to show inferential disclosure, we assume that the values are real except the income components. Let's also assume that an intruder is interested in the variable employee cash or near cash net income (variable `py010n`) of a person with the following attitudes:

```
intrudersKnowledge <- data.frame("hsize" = 3,
                                 "db040" = "Tyrol",
                                 "age"= 34,
                                 "pl030" = "2",
                                 "pb220a" = "AT",
                                 "eqIncome" = 16090)
intrudersKnowledge

  ##   hsize db040 age pl030 pb220a eqIncome
  ## 1     3 Tyrol  34     2     AT    16090
```

Note that this is already more information than typically available, since the intruder even knows the equivalized income of persons in the household.

He is interested in knowing the employee cash or near cash net income of a person. The intruder might try to find a good model using personal income on employee cash or near cash net income as response and certain other variables as predictors. For simplicity, assume that he is just using his predictors without any interaction terms (the inclusion of interaction terms do not improve the predictive power for this data set anyhow). Since the employee cash net income variable is right-skewed, the log is taken. Moreover, assume the intruder knows that the income is not 0, so we only take observations with cash net incomes greater than zero.

```
data(eusilc)
mod1 <- lm(log(py010n) ~ hsize + db040 + age +
                         pl030 + pb220a + eqIncome,
           data=eusilc[eusilc[,"py010n"] > 0, ])
s1 <- summary(mod1)
s1$r.squared

  ## [1] 0.3682565
```

We see that the predictive power of the model is not very high ($R^2 \sim 0.37$).

Let us assume that the real (unknown!) value of `py010n` for person 1 is 9500€. We predict this value from our given data set and estimated model.

```
exp(predict(mod1, intrudersKnowledge))

  ##        1
  ## 7242.003
```

We get an estimated value of 7.242003×10^3 on employee cash or near cash net income. We can ask ourselves if this value is far enough from the true unknown value of 9500€. However, we should also take the model uncertainty into account.

```
exp(predict(mod1, intrudersKnowledge, interval = "prediction"))

  ##        fit      lwr      upr
  ## 1 7242.003 1826.026 28721.73
```

We see that the prediction interval is rather large, inferential disclosure is hardly possible with this scenario.

Exercises:

Question 2.1 **Choice of variables (I)**
Have a brief look at a popular data set, the EU-SILC data in R-package **laeken** (Alfons and Templ 2013). Read the help for this data set by typing in R:

```
install.packages("laeken")
data(eusilc)
?eusilc
```

Determine which of the variables should be defined as
(a) direct identifiers (if any)
(b) categorical key variables
(c) continuous key variables
(d) sensitive variables

Question 2.2 **Choice of variables (II)**
Please have a brief look at another popular data set, the Structural Earnings Statistics in R-package **laeken** (Alfons and Templ 2013). Read the help for this data set by typing in R:

```
data(ses)
?ses
```

Determine which of the variables should be defined as
(a) direct identifiers (if any)
(b) categorical key variables
(c) continuous key variables
(d) sensitive variables

Question 2.3 **Frequencies of key variables**
Use again the EU-SILC data set from package **laeken**. Is a re-identification of observation 8 possible assuming region (*db040*), age, gender (*rb090*) and economic status (*pl030*) as categorical key variables. Is re-identification of observation 3 easily possible?

If you do not have any experience in R, please read Chap. 1 first, before starting this exercise.

2.3 Disclosure Risk Versus Information Loss and Data Utility

Applying SDC techniques to the original microdata will result in information loss and hence affect data utility. Data utility describes the value of data as an analytical resource, comprising analytical completeness and analytical validity.