REGRESSION PROBLEMS FOR FINITE POPULATIONS

Julio M. Singer

1 Definition of the regression problems

In this document I try to define what are the regression problems for finite populations. We define a finite population as a collection of \( N \) identifiable units labeled \( s = 1, \ldots, N \), using notation similar to that employed in Stanek, Singer & Lencina (2004). Suppose that associated to unit \( s \) there are two quantities \( y_s \) and \( x_s \), where \( y_s \) is a response and \( x_s \) is an explanatory variable. We assume that when unit \( s \) is observed, both the parameter \( y_s \) and the explanatory variable \( x_s \) are observed, the former possibly with error and the latter without error.

Based on a simple random sample of size \( n \), we are interested in

A) Predicting \( y_0 \) given \( x_0 \) (which may or may not correspond to one of the sampled units).

B) Estimating some measure of association between the response and the explanatory variable, like

\[
B = \sum_{s=1}^{N} \frac{(y_s - \mu_y)(x_s - \mu_x)}{\sum_{s=1}^{N} (x_s - \mu_x)^2},
\]

with \( \mu_y = N^{-1} \sum_{s=1}^{N} y_s \), \( \mu_x = N^{-1} \sum_{s=1}^{N} x_s \).

C) Estimate \( A \) and \( B \) and predict \( y_0 \) given \( x_0 \) (which may or not be one of the sampled values) under the model \( y_s = A + B x_s \), \( s = 1, \ldots, N \). We will comment on this below.

2 Solution to the problems when there is NO response error

A) This is something we haven’t done. I think this is what you had in mind when you said that we should try to solve the problem without transforming the response. If you think of a straight line as an ”approximation” to the relation between the response and the explanatory variable the difference between each \( y_s \) and its projection on the line is ”lack of fit”. I believe that the parameters \( A \) and \( B \) are defined as the values of the coefficients of the least squares line. I don’t think they can be
defined in the infinite population case and this is why response error is included in the usual regression problem definition.

B) This is what Luz Mery is doing in the regression paper. In Table 3, this corresponds to the case $\sigma_e = 0$.

C) This does not make sense when there is no response error, since $A$ and $B$ may be computed exactly for any sample of size $\geq 2$.

3 Solution to the problems when there is response error

A) Again this is something we haven’t done. It is similar to item A) above with the addition of an extra term for response error.

B) This is what Luz Mery is doing in the regression paper. In Table 3, this corresponds to the case $\sigma_e \neq 0$.

C) This is what is usually meant by linear regression. The expected values (over response error) of the response given the values of the explanatory variable lie on the regression line. We have not attacked this problem.