

Liang, Hua

Working Paper

Errors in variables models

SFB 373 Discussion Paper, No. 1999,77

Provided in Cooperation with:

Collaborative Research Center 373: Quantification and Simulation of Economic Processes,
Humboldt University Berlin

Suggested Citation: Liang, Hua (1999) : Errors in variables models, SFB 373 Discussion Paper, No. 1999,77, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin,
<https://nbn-resolving.de/urn:nbn:de:kobv:11-10046652>

This Version is available at:

<https://hdl.handle.net/10419/61768>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Errors in Variables Models

Hua Liang
October 13, 1999

Error-in-variables (EIV) models are regression models in which the regressors are observed with errors. These models include the linear EIV models, the non-linear EIV models, and the partially linear EIV models. Suppose that we want to investigate the relationship between the yield (Y) of corn and available nitrogen (X) in the soil. A common approach is to assume that Y depends upon X linearly. To evaluate the degree of dependence, it's necessary to sample the soil of the experimental plot and to perform an analysis. We can not observe X , but rather an estimate of X . Therefore, we represent the observed nitrogen by W , also called the surrogate of X . The model thus studied is an errors-in-variables model.

This chapter surveys the basic results and explains how errors in variables models are implemented in XploRe. The first part covers the class of ordinary linear errors-in-variables models, which has been studied in detail by Fuller (1987). The second part focuses on the nonlinear errors-in-variables or measurement error models surveyed in Carroll, Ruppert and Stefanski (1995). In the third part, we give an overview of partially linear errors-in-variables models. All chapters contain practical examples. The corresponding quantlets are contained in the quantlib `eiv`.

1 Linear EIV Models

```
gest = eivknownatt(w,y,kww)
      estimates the parameters with known reliability ratio

gest = eivknownratue(w,y,delta)
      estimates the parameters with known ratio of the two variances of the
      two measurement errors

gest = eivknownvaru(w,y,sigmau)
      estimates the parameters with known variance of the measurement
      error U

gest = eivknownvarumod(omega,w,y,sigmau)
      calculates modified estimators of the parameters with known variance
      of the measurement error U

gest = eivlinearinstr(w,z,y)
      estimates the parameters with the instrumental variable z

gest = eivvec1(w,y,sigue,siguu)
      estimates the parameters with multi-dimensional variables x with
      known variance and covariance of  $\varepsilon$  and  $U$ 

gest = eivvec2(w,y,gamma)
      estimates the parameters for multi-dimensional variables x with
      known covariance of the measurement error U

gest = eivlinearinstrvec(w,z,y)
      estimates the parameters for multi-dimensional variables with the in-
      strumental variable z
```

A linear errors-in-variables model is defined as:

$$\begin{aligned} Y &= \alpha + \beta^T X + \varepsilon \\ W &= X + U, \end{aligned} \quad (1)$$

where Y is the dependent variable, X is the matrix of regressors, and U is a random term. In this model, the regressors X are observed with error, i.e., only the variable $W = X + U$, called **the manifest variable**, is directly observed. The unobserved variable X is called a **latent variable** in some areas of application, while U is called **the measurement error**. Models with fixed X are called **functional models**, while models with random X are called **structural models**.

We assume that the random variables (X, ε, U) are independent with mean $(\mu_x, 0, 0)$ and covariance matrix $\text{diag}(\Sigma_{xx}, \sigma_{ee}, \sigma_{uu}I_p)$. In the `eiv` quantlib the method of moments is used to estimate the parameters. In the literature, it's generally assumed that (X, ε, U) are jointly normally distributed, and that (W, Y) follows a bivariate normal distribution (Fuller (1987)). Even without the normality assumption, various moment methods may be used to estimate all parameters. Furthermore, we assume that $\sigma_{ee} = \delta\sigma_{uu}$. Thus, the mean and the variance of the joint distribution of (Y, W) are

$$\mu = \begin{pmatrix} \alpha + \beta^T \mu_x \\ \mu_x \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \delta\sigma_{uu} + \beta^T \Sigma_{xx} \beta & \beta^T \Sigma_{xx} \\ \Sigma_{xx} \beta & \Sigma_{xx} + \sigma_{uu} I_p \end{pmatrix}. \quad (2)$$

We write $m_{yy} = \sum_{t=1}^n (Y_t - \bar{Y})^2 / (n-1)$, $m_{wy} = \sum_{t=1}^n (W_t - \bar{W})(Y_t - \bar{Y}) / (n-1)$ and $m_{ww} = \sum_{t=1}^n (W_t - \bar{W})(W_t - \bar{W})^T / (n-1)$. Using the method of moments, we define the solutions of the following equations as the estimators of $\beta, \Sigma_{xx}, \sigma_{uu}$.

$$\begin{cases} m_{yy} &= \delta\sigma_{uu} + \beta^T \Sigma_{xx} \beta \\ m_{wy} &= \Sigma_{xx} \beta \\ m_{ww} &= \Sigma_{xx} + \sigma_{uu} I_p \end{cases} \quad (3)$$

1.1 A Single Explanatory Variable

Let's first investigate the case of single explanatory variable, i.e., $p = 1$. The least squares estimator based on the observed variables is biased towards zero because of the disturbance of the measurement error. In fact, let

$$\hat{\gamma}_1 = \left\{ \sum_{t=1}^n (W_t - \bar{W})^2 \right\}^{-1} \sum_{t=1}^n (W_t - \bar{W})(Y_t - \bar{Y}) \quad (4)$$

be the regression coefficient computed from the observed variables. $\hat{\gamma}_1$ would be an unbiased estimator of β if there were no measurement error U . By the properties of the bivariate normal,

$$E\hat{\gamma}_1 = \sigma_{ww}^{-1} \sigma_{yx} = \beta_1 (\sigma_{xx} + \sigma_{uu})^{-1} \sigma_{xx}. \quad (5)$$

The least squares regression coefficient is biased towards zero because of the disturbance of the measurement error U ; the measurement error attenuates the regression coefficient. The ratio $k_{ww} = \sigma_{xx} / (\sigma_{xx} + \sigma_{uu})$, which defines the degree of attenuation, is called the **reliability** of W , or the **reliability ratio**. As pointed out above, ignoring measurement error leads to the least squares slope as an estimator of βk_{ww} , not of β .

In this section, we consider several estimators for the linear `eiv` models. These estimators have different forms based on the corresponding assumption on the variances. A complete account is given in Fuller (1987).

Assume that the degree of attenuation k_{ww} is known. In this case, the estimators of β and α are defined as $\hat{\beta} = k_{ww}^{-1} \hat{\gamma}_1$ and $\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{W}$. Moreover, their variances are estimated by

$$\widehat{\text{var}}(\hat{\beta}) = n^{-1} S_{vv} \left\{ \sum_{i=1}^n (W_i - \bar{W})^2 \right\}^{-1} S_l^2$$


and

$$\widehat{\text{var}}(\hat{\alpha}) = n^{-1} S_{vv} + \bar{W}^2 \widehat{\text{var}}(\hat{\beta}),$$

where $S_l^2 = (n-2)^{-1} \sum_{i=1}^n \{Y_i - \bar{Y} - \hat{\gamma}_1(W_i - \bar{W})\}^2$ and $S_{vv} = (n-2)^{-1} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} W_i)^2$. Incidentally, the estimators of α and $\text{var}(\hat{\alpha})$ always have the same forms, whatever the estimators of β and $\text{var}(\hat{\beta})$.

The quantlet `eivknownatt` evaluates the moment estimates of the parameters $\mu_x, \beta, \alpha, \sigma_{xx}, \sigma_{uu}, \text{var}(\hat{\alpha})$ and $\text{var}(\hat{\beta})$. Its syntax is the following:

```
gest = eivknownatt(w,y,kww)
```

 `eivknownatt-example.xpl`

where

`w`

the observed regressors,

`y`

the response,

`kww`

the degree of attenuation.

This quantlet returns a list `gest`, which contains the followings estimates:

`gest.mux`

estimate of the mean of X ,

`gest.beta1` $\hat{\beta}$,

`gest.beta0` $\hat{\alpha}$,

`gest.sigmax`

estimate of the variance of X ,

`gest.sigmau`

estimate of the variance of U ,

`gest.sigmae`

estimate of the variance of ε ,

`gest.varbeta1`

the estimate of the variance of $\hat{\beta}$,

`gest.varbeta0`

the estimate of the variance of $\hat{\alpha}$.

We consider the following example, based on simulated data, in which the distribution of the measurement error U is normal with mean 0 and standard deviation 0.9, the latent variable X having the same distribution, so that the reliability ratio equals $k_{ww} = 0.5$.

```

library("eiv")
n = 100
randomize(n)
x=0.9*normal(n)           ; latent variables
w=x+0.9*normal(n)         ; manifest variables
y=0.9+0.8*x+0.01*normal(n)
kww =0.5                  ; reliability ratio
gest=eivknownnatt(w,y,kww)

```

The parameter estimates are the following:

```

gest.mux=-0.093396
gest.beta1=0.79286
gest.beta0=0.8425
gest.sigmax=0.72585
gest.sigmau=0.72585
gest.sigmae=0.074451
gest.varbeta1=0.0085078
gest.varbeta0=0.0054358

```

The true values are $\mu_{x0} = 0$, $\beta_0 = 0.8$, $\alpha_0 = 0.9$, $\sigma_{u0} = 0.81$ and $\sigma_{e0} = 0.81$.

Assume that the ratio of two variances of the two measurement errors, $k_{ww} = \sigma_{uu}^{-1}\sigma_{ee}$, is known. Then the estimators of the parameters of the most interest are defined as

$$\hat{\beta} = \frac{m_{yy} - \delta m_{ww} + \{(m_{yy} - \delta m_{ww})^2 + 4\delta m_{wy}^2\}^{1/2}}{2m_{wy}}$$


and

$$\widehat{\text{var}}(\hat{\beta}) = (n - 1)^{-1} \hat{\sigma}_{ww}^{-2} (\hat{\sigma}_{ww} S_{vv} + \hat{\sigma}_{uu} S_{vv} - \hat{\sigma}_{uv}^2),$$

where $\hat{\sigma}_{ww} = m_{wy} \hat{\beta}$, $\hat{\sigma}_{uu} = m_{ww} - \hat{\sigma}_{ww}$, $\hat{\sigma}_{uv} = -\hat{\beta} \hat{\sigma}_{uu}$, $S_{vv} = (n - 2)^{-1} \sum_{i=1}^n \{Y_i - \bar{Y} - \hat{\beta}(W_i - \bar{W})\}^2$.

The quantlet `eivknownratue` estimates the parameters in this situation. Its syntax is similar to that of the quantlet `eivknownnatt`:

```
gest = eivknownratue(w,y,delta)
```

 `eivknownratue-example.xpl`

where `delta` is the ratio of the two variances.

For the purpose of illustration, we use the data which Fuller (1987) originally analyzed. The variables Y and W are the numbers of hen pheasants in Iowa at August and spring in the period from 1962 to 1976. Both measurement are subjected to the measurement errors. The ratio of σ_{ee} to σ_{uu} is supposed to be 1/6. We use the following XploRe code:

```

v=read("pheasants.dat")
n=rows(v)
y=v[,2]
x=v[,3]
delta=1/6

```

The data set available in XploRe. Running

```

library("eiv")
gest=eivknownratue(x,y,delta)

```

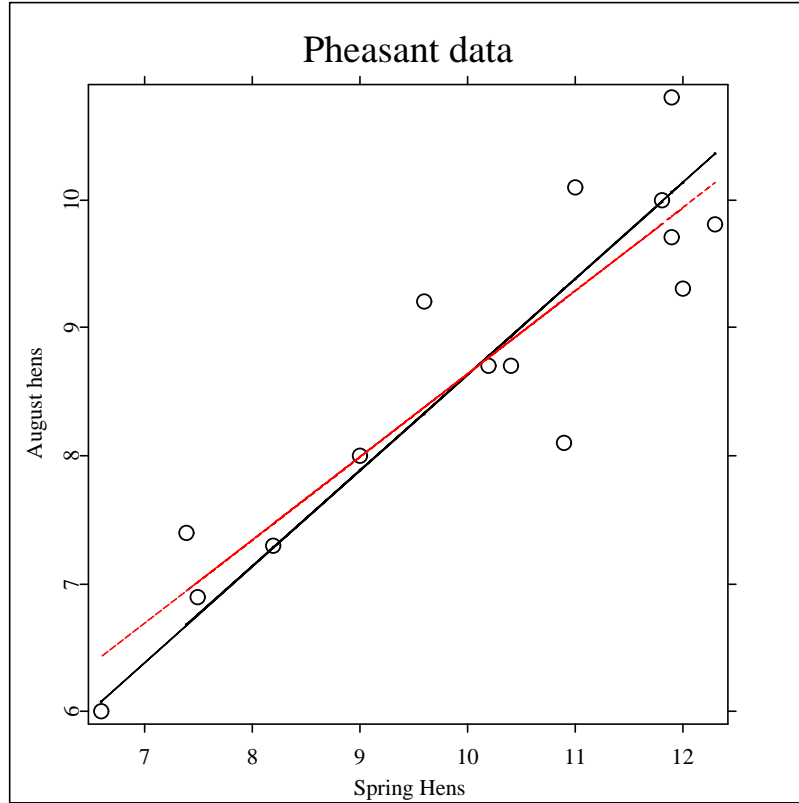


Figure 1: Pheasant data and estimated structural lines

we obtain the estimates of slope and intercept as 0.75158 (s.e. 0.0962) and 1.1158 (s.e. 0.9794). In Figure 1, the empty circles represents the observation data, the solid line is based on the ordinary least squares estimator, and the dashed line is the fit based on the moment estimator. Even in this small-sample data set, different conclusions are obvious if we are ignoring the measurement errors.

When the variance of measurement error σ_{uu} is known, we define the estimators of β and the variance of this estimator as

$$\hat{\beta} = (m_{ww} - \sigma_{uu})^{-1} m_{wy} \quad (6)$$


and

$$\widehat{\text{var}}(\hat{\beta}) = (n - 1)^{-1} \hat{\sigma}_{ww}^{-2} (m_{ww} S_{vv} + \hat{\beta}^2 \sigma_{uu}^2),$$

where $\hat{\sigma}_{ww} = m_{ww} - \sigma_{uu}$ and $S_{vv} = (n - 2)^{-1} \sum_{i=1}^n \{Y_i - \bar{Y} - \hat{\beta}(W_i - \bar{W})\}^2$.

The quantlet `eivknownvaru` evaluates the moment estimates stated above. Its syntax is similar to that of the two previous quantlets:

```
gest = eivknownvaru(w,y,sigmau)
```

 `eivknownvaru-example.xpl`

where `sigmau` is the variance of the error U .

We now use the quantlet `eivknownvaru` to analyze a real example from Fuller (1987). In this example, we study the relationship between the yield of corn (Y) and soil nitrogen (X), the latter of which cannot be measured exactly. The variance arising from these two variables has been estimated to be $\sigma_{uu} = 57$. We assume that σ_{uu} is known and compute the related estimates using the quantlet `eivknownvaru`. The ordinary least squares estimates are $\hat{\beta}_{LS} = 0.34404$ and $\hat{\alpha}_{LS} = 73.152$, ignoring the measurement errors. We use the XploRe code:

```
z=read("corn.dat")
n=rows(z)
y=z[,1]
x=z[,2:3]
w=x[,2]
sigmau=57
gest=eivknownvaru(w,y,sigmau)
```

The moment estimates are $\hat{\beta}_{MM} = 0.42316$ (s.e. 0.1745) and $\hat{\alpha}_{MM} = 73.152$ (s.e.12.542), $\hat{\sigma}_{xx} = 247.85$ and $\hat{\sigma}_{ee} = 43.29$. So, the reliability ratio is $247.85/304.85 = 0.81$. In Figure 2, the circles represent the observation data, the solid line is based on the ordinary least squares estimator, and the dashed line is the fit based on the moment estimator.

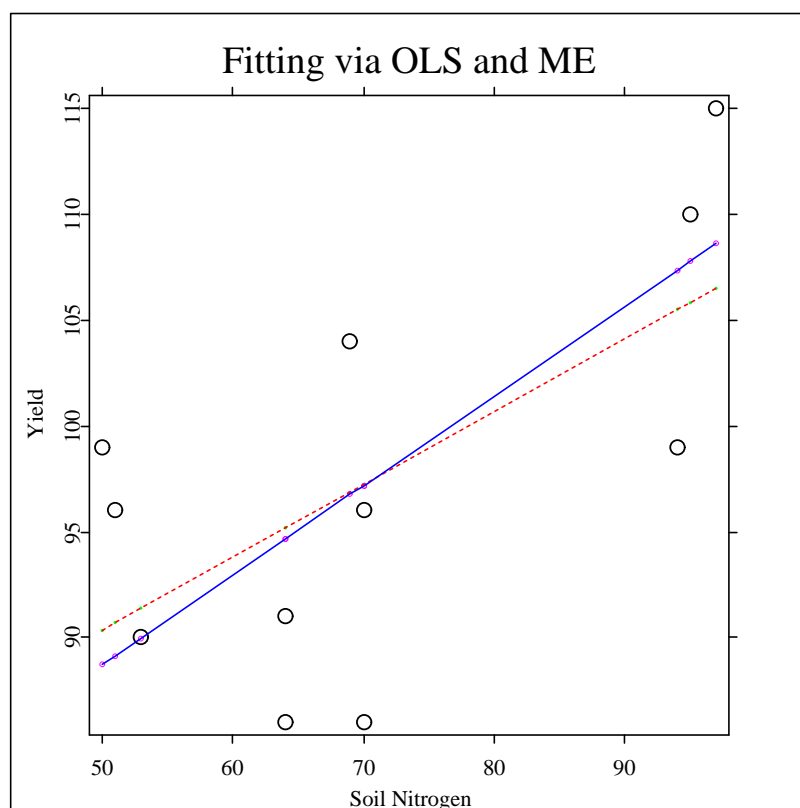


Figure 2: Output display for Yields of Corn

Theoretical study and empirical evidence have shown that the method of mo-

ments estimator given in (6) performs poorly in small samples, since such ratios are typically biased estimators of the ratio of the expectations. For this reason, we consider the modification proposed by Fuller (1987) of this estimator. Define an estimator of β by

$$\tilde{\beta} = \{\tilde{H}_{ww} + \omega(n-1)^{-1}\sigma_{uu}\}^{-1}m_{wy},$$

where $\omega > 0$ is a constant to be determined later, and

$$\tilde{H}_{ww} = \begin{cases} m_{ww} - \sigma_{uu} & \text{if } \hat{\lambda} \geq 1 + (n-1)^{-1} \\ m_{ww} - \{\hat{\lambda} - (n-1)^{-1}\}\sigma_{uu} & \text{if } \hat{\lambda} < 1 + (n-1)^{-1} \end{cases}$$

with $\hat{\lambda}$ being the root of


$$\det\{m_{(y,w)(y,w)} - \lambda \text{diag}(0, \sigma_{uu})\} = 0.$$

This estimator has been shown to be almost unbiased for β . Its variance is estimated by

$$\widehat{\text{var}}(\hat{\beta}) = (n-1)^{-1}\{\tilde{H}_{ww}^{-1}\tilde{\sigma}_{vv} + \tilde{H}_{ww}^{-2}(\sigma_{uu}\tilde{\sigma}_{vv} + \tilde{\beta}^2\sigma_{uu}^2)\}$$

where $\tilde{\sigma}_{vv} = (n-2)^{-1}(n-1)(m_{yy} - 2\tilde{\beta}m_{wy} + \tilde{\beta}^2m_{ww})$. For detailed theoretical discussions see Section 2.5 of Fuller (1987).

The quantlet `eivknownvarumod` implements the calculating procedure.

```
gest = eivknownvarumod(omega, w, y, sigmau)
 eivknownvarumod-example.xpl
```

input

```
omega
    scalar,
w
    n x 1 matrix, the design variables,
y
    n x 1 matrix, the response,
sigmau
    the variance of measurement error.
```

output

```
gest.mux    the mean value of X,
gest.beta1   $\hat{\beta}$ ,
gest.beta0   $\hat{\alpha}$ ,
gest.sigmax
    the estimate of the variance of X,
gest.sigmae
    the estimate of the variance of error  $\varepsilon$ ,
gest.varbeta1
    the estimate of the variance of  $\hat{\beta}$ ,
gest.varbeta0
    the estimate of the variance of  $\hat{\alpha}$ .
```


We return to consider the data set “corn”. Calculating the different choices of ω , we obtain the following results. A comparison with the results shown by the quantlet `eivknownvaru` indicates that $\tilde{\beta}$ is the same as $\hat{\beta}_{MM}$ when ω takes 0.

ω	$\hat{\beta}_{LS}$	$\tilde{\beta}$	$\widehat{\text{var}}(\tilde{\beta})$
0	0.34404	0.42316	0.030445
1	0.34404	0.41365	0.030165
2	0.34404	0.40455	0.029927
$2 + 2m_{ww}^{-1}\sigma_{uu}$	0.34404	0.40125	0.029847
5	0.34404	0.37952	0.029419
10	0.34404	0.34404	0.029072

The estimates $\tilde{\beta}$ and $\widehat{\text{var}}(\tilde{\beta})$ decrease with ω , and $\tilde{\beta}$ is equivalent to $\hat{\beta}_{LS}$ when $\omega=10$. The linear fitting for $\omega=2 + 2m_{ww}^{-1}\sigma_{uu}$ is shown in Figure 3.

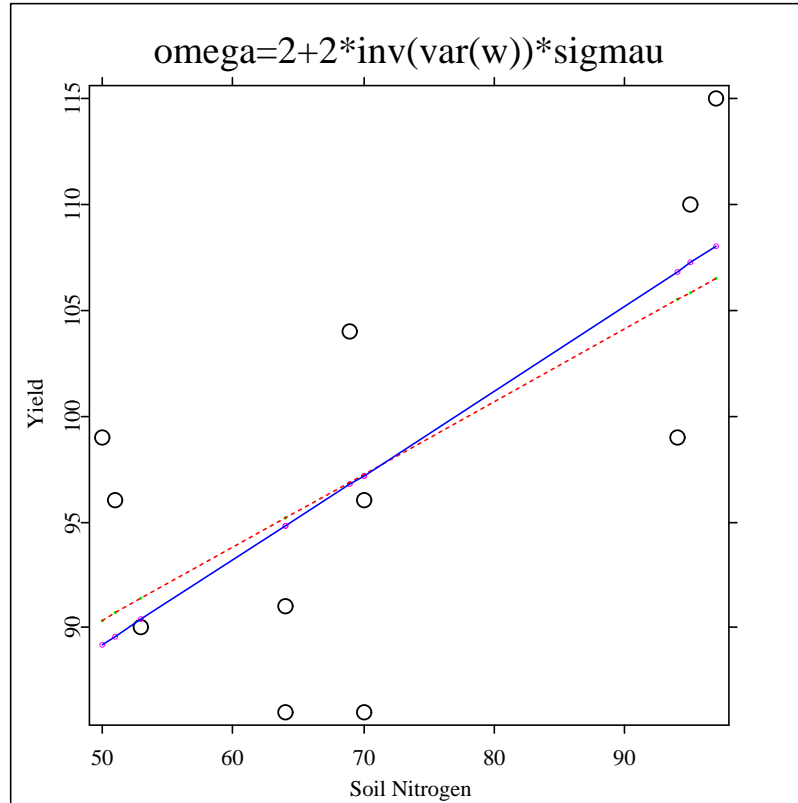


Figure 3: Output display for Yields of Corn

In this paragraph, we assume that we can observe a third variable Z which is correlated with X . The variable Z is called as an **instrumental variable** for X if

$$E \left\{ n^{-1} \sum_{i=1}^n (Z_i - \bar{Z})(\varepsilon_i, U_i) \right\} = (0, 0),$$

$$E \left\{ n^{-1} \sum_{i=1}^n (Z_i - \bar{Z}) X_i \right\} \neq 0.$$

Although we do not assume that k_{ww} and σ_{eu} are known or that σ_{xu} is zero, we can still estimate α and β by the method of moments as follows: Let $\hat{\beta} = m_{wz}^{-1} m_{yz}$ and $\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{W}$, where $m_{yz} = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})(Z_i - \bar{Z})$ and $m_{wz} = (n-1)^{-1} \sum_{i=1}^n (W_i - \bar{W})(Z_i - \bar{Z})$. Furthermore, we estimate the variances of the approximate distributions of $\hat{\beta}$ and $\hat{\alpha}$ by

$$\widehat{\text{var}}(\hat{\beta}) = (n-1)^{-1} m_{wz}^{-2} m_{zz} S_{vv} \quad \text{and} \quad \widehat{\text{var}}(\hat{\alpha}) = n^{-1} S_{vv} + \bar{W}^2 \widehat{\text{var}}(\hat{\beta})$$

with $S_{vv} = (n-2)^{-1} \sum_{i=1}^n \{Y_i - \bar{Y} - \hat{\beta}(W_i - \bar{W})\}^2$.

The quantlet `eivlinearinstr` accomplishes the implementation. Its syntax is the following:

```
gest = eivlinearinstr(w,z,y)  eivlinearinstr-example.xpl
```

The estimates of α and β are returned in the list `gest`

```
gest.beta1
  the estimate of  $\beta$ ,

gest.beta0
  the estimate of  $\alpha$ .
```

Before ending this section, we use the quantlet `eivlinearinstr` to study a practical data-set, in which we study Alaskan earthquakes for the period from 1969-1978. The data are from Fuller (1987). In the data structure, we have the logarithm of the seismogram amplitude of 20 second surface waves, denoted by Y , the logarithm of the seismogram amplitude of longitudinal body waves, denoted by W and the logarithm of maximum seismogram trace amplitude at short distance, denoted by Z .

```
gest = eivlinearinstr(w,z,y)  eivlinearinstr-example.xpl
```

The estimates are

```
gest.beta0=-4.2829 (s.e.1.1137)
gest.beta1=1.796 (s.e.0.2131)
```

Figure 4 shows the fitted results, in which the circles represent the data Y , the solid line is based on the above estimates, and the dashed line contains the estimated values based on the regression of Y on W . This means that if we ignore the measurement errors, then it shows an obvious difference.

1.2 Vector of Explanatory Variables

Suppose that X is a p -dimensional row vector with $p > 1$, β is a p -dimensional column vector, and the $(1+p)$ -dimensional vectors $e = (\varepsilon, U)^T$ are independently normal $N(0, \Sigma_{ee})$ random vectors.

Assume that the covariance between ε and U , $\Sigma_{\varepsilon u}$ and the covariance matrix of U , Σ_{uu} are known. Then the other parameters, such as β , α and others, are estimated by

$$\begin{aligned} \hat{\beta} &= (m_{ww} - \Sigma_{uu})^{-1} (m_{wy} - \Sigma_{\varepsilon u}), \\ \hat{\alpha} &= \bar{Y} - \bar{W} \hat{\beta}, \\ \hat{\sigma}_{\varepsilon} &= m_{yy} - 2m_{wy} \hat{\beta} + \hat{\beta}^T m_{ww} \hat{\beta} + 2\Sigma_{\varepsilon u} \hat{\beta} - \hat{\beta}^T \Sigma_{uu} \hat{\beta} \end{aligned}$$

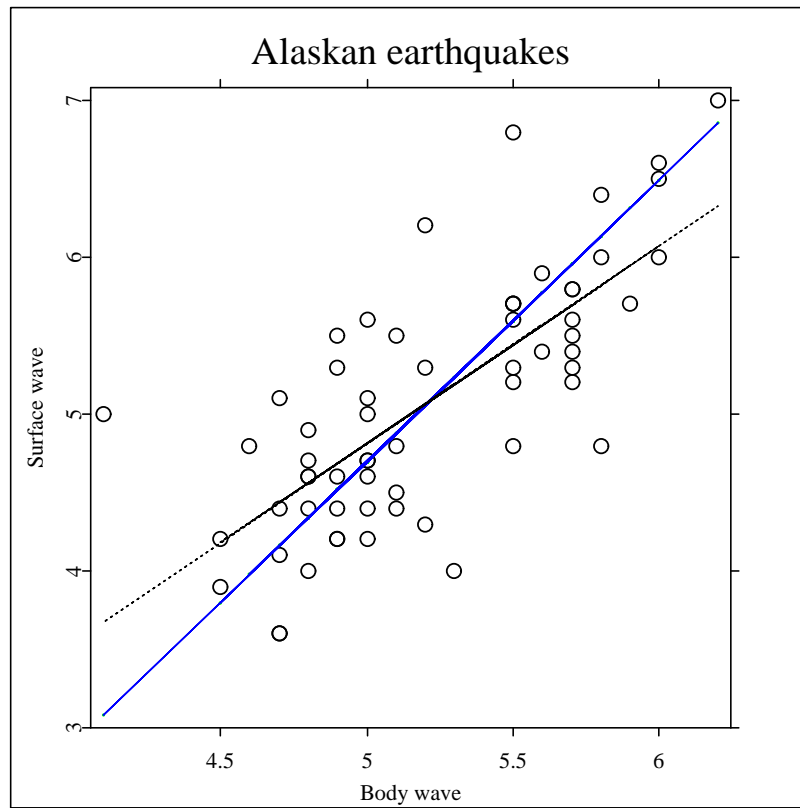



Figure 4: Output display for Alaskan Earthquakes

and $\hat{\Sigma}_{xx} = m_{ww} - \Sigma_{uu}$, provided $\hat{\Sigma}_{xx}$ is positive definite and $\hat{\sigma}_\varepsilon \geq \Sigma_{\varepsilon u} \Sigma_{uu}^+ \Sigma_{u\varepsilon}$, where Σ_{uu}^+ denotes the generalized inverse of Σ_{uu} . If either of these conditions is violated, the estimators fall on the boundary of the parameter space, and the above forms must be modified. For a detailed discussion, see Section 2.2 of Fuller (1987).

The quantlet `eivvec1` evaluates these estimates. Its syntax is the following:

```
gest = eivvec1(w, y, sigue, siguu)
```

 `eivvec1-example.xpl`

The estimates are listed in the variable `gest` as follows:

`gest.mux`

scalar, the estimate of the mean of X ,

`gest.hatbeta0`

scalar, the estimate of α ,

`gest.hatbeta1`

vector, the estimate of β ,

`gest.hatsigmax`

$p \times p$ matrix, the estimate of the covariance of X ,

`gest.hatsigae`

scalar, the estimate of the variance of ε .

We calculate a simulated data set with the quantlet `eivvec1` as follows:

```
library("xplore")
library("eiv")
n = 100
randomize(n)
nu = #(2,3,4)
sig=0*matrix(3,3)
sig[,1]=#(0.25, 0.9, 0.1)
sig[,2]=#(0.9, 1, 0.2)
sig[,3]=#(0.1, 0.2, 4)
x=normal(n,3)*sig+nu'
w=x+0.01*normal(n,3)
a1=#(1.2, 1.3, 1.4)
y=0.75+x*a1+0.09*normal(n)
sigue=#(0.11, 0.09, 0.45)
siguu=0*matrix(3,3)
siguu[,1]=#(1.25, 0.009, 0.01)
siguu[,2]=#(0.009,0.081, 0.02)
siguu[,3]=#(0.01, 0.02, 1.96)
gest=eivvec1(w,y,sigue,siguu)
```

The estimates are: $\mu_x = (2.024, 2.9106, 3.9382)^T$, $\hat{\beta} = (0.011384, 0.013461, 0.013913)^T$,
 $\hat{\beta}_0 = 12.362$, $\hat{\sigma}_{ee} = 1034.9$, $\hat{\Sigma}_{xx} = \begin{pmatrix} 0.84466, & 1.0319, & 0.43677 \\ 1.0319, & 1.664, & 1.0941 \\ 0.43677, & 1.0941, & 19.781 \end{pmatrix}$.

In this paragraph, our aim is to estimate the parameters in p -dimensional measurement error models when the entire error covariance structure is either known, or known up to a multiple scalar. Assume that the errors (U, ε) obey normal distribution with mean zero vector and covariance $cov(U, \varepsilon)$, which can be represented as $\Gamma_{(u,\varepsilon)(u,\varepsilon)}\sigma^2$, where $\Gamma_{(u,\varepsilon)(u,\varepsilon)}$ is known. Then the maximum likelihood estimators of β and σ^2 are


$$\begin{aligned}\hat{\beta} &= (m_{ww} - \hat{\lambda}\Gamma_{uu})^{-1}(m_{wy} - \hat{\lambda}\Gamma_{\varepsilon u}), \\ \hat{\sigma}_m^2 &= (p+1)^{-1}\hat{\lambda},\end{aligned}$$

where Γ_{uu} and $\Gamma_{\varepsilon u}$ are the submatrices of $\Gamma_{(u,\varepsilon)(u,\varepsilon)}$, and $\hat{\lambda}$ is the smallest root of

$$|m_{(y,w)(y,w)} - \lambda\Gamma_{(u,\varepsilon)(u,\varepsilon)}| = 0.$$

The quantlet `eivvec2` evaluates these likelihood estimators. This case is an extension of the case discussed in the context for the quantlet `eivknownvaru`. The theoretical details are given in Fuller (1987). The syntax of this quantlet is the following:

```
gest = eivvec2(w, y, gamma)
```

 `eivvec2-example.xpl`

where `gamma` is a known matrix.

The following simulated example shows us how to run the quantlet `eivvec2`.

```
library("xplore")
library("eiv")
n=100
randomize(n)
```

```

sig=0*matrix(3,3)
sig[,1]=#(0.25, 0.09, 0.1)
sig[,2]=#(0.09, 1, 0.2)
sig[,3]=#(0.1, 0.2, 0.4)
x=sort(uniform(n,3)*sig)
w=x+0.03*normal(n,3)
beta0=#(0.5, 0.6, 0.7)
y=x*beta0+0.05*normal(n)
gamma=(#(0.03,0,0,0))'|(#(0,0,0)~0.05*unit(3))
gest=eivvec2(w,y,gamma)

```

The estimates are the following:

```

gest.hatbeta=(0.18541, 0.0051575, -0.088003)
gest.sigmam=0.015424

```

Consider the method of instrumental variables for the p -dimensional case. Assume that the q -dimensional vector of instrumental variables Z is available and that $n > q \geq p$. In addition, assume that $\sum_{i=1}^n Z_i^T Z_i$ is nonsingular with probability one, $E\{Z_i^T(\varepsilon_i, U_i)\} = (0, 0)$, and the rank of $(\sum_{i=1}^n Z_i^T Z_i)^{-1} \sum_{i=1}^n Z_i^T W_i$ is q with probability one. When $q = p$, we define the estimator of β as

$$\hat{\beta} = (\mathbf{Z}^T \mathbf{W})^{-1} (\mathbf{Z}^T \mathbf{Y}).$$

Otherwise, write

$$S_{aa} = (n - q)^{-1} \{(\mathbf{Y}, \mathbf{W})^T (\mathbf{Y}, \mathbf{W}) - (\mathbf{Y}, \mathbf{W})^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{Y}, \mathbf{W})\}$$

and define the estimator as

$$\hat{\beta} = (\mathbf{W}^T \mathbf{W} - \tilde{\gamma} S_{aa22})^{-1} (\mathbf{W}^T \mathbf{Y} - \tilde{\gamma} S_{aa21}),$$


where S_{aa21} and S_{aa22} are the submatrices of S_{aa} , and $\tilde{\gamma}$ is the smallest root of

$$|(\mathbf{Y}, \mathbf{W})^T (\mathbf{Y}, \mathbf{W}) - \gamma S_{aa}| = 0.$$

Its statistical inferences refer to Section 2.4 of Fuller (1987).

The quantlet `eivlinearinstrvec` achieves the calculation procedure in XploRe. This generalizes the quantlet `eivlinearinstr` to the p -dimensional case.

```
gest =eivlinearinstrvec(w,z,y)
```

 `eivlinearinstrvec-example.xpl`

We end this section with an example, in which we randomly produce variables w , instrumental variables z , and response y . Then we execute the quantlet `eivlinearinstrvec` and get the estimates.

```

library("xplore")
library("eiv")
n=100
randomize(n)
w=floor(6*uniform(n,3)+#(4,5,5)')
z=floor(8*uniform(n,4)+#(3,3,2,2)')
y=floor(9*uniform(n)+2)
gest=eivlinearinstrvec(w,z,y)

```

The estimate of the parameter vector is

```
gest=(0.19413, 0.24876, 0.37562)
```

2 Non-linear EIV Models

```
res = reca(y,w,z,su2)
      implementation of regression calibration

res = simex(y,w,z,su2,lam,b)
      implementation of simulation extrapolation
```

When the relationship between response and the covariates is nonlinear and the covariates are measured with errors, the models are called nonlinear EIV models. There is a numerous body of literature on the nonlinear EIV models (the monograph by Carroll, Ruppert and Stefanski (1995) give a good overview of the nonlinear methods). In this section we mainly describe two simple approximate techniques for handling measurement error in the analysis of nonlinear EIV models. The presentation here is based on Carroll, Ruppert and Stefanski (1995).

We denote the dependent variable by Y , the variables observed with error by X , the variables measured without error by Z , and the manifest variable by W . We define a nonlinear errors-in-variables model as:

$$\begin{aligned} E(Y|X) &= g(X) \\ W &= X + U \end{aligned} \tag{7}$$

Two classes of nonlinear `eiv` models are considered:

- Error models, including Classical Measurement Error models and Error Calibration models, where the conditional distribution of W given (Z, X) is modeled;
- Controlled-variable or Berkson error models, where the conditional distribution of X given (Z, W) is modeled.

From the viewpoint of measurement error construction, the usual model is typically restricted on the classical additive measurement error model:

$$W = X + u \text{ with } E(u|X, Z) = 0.$$

In the controlled variable model, the measurement error model has the form:

$$X = W + U' \text{ with } E(U'|W) = 0.$$

The example considered in this section is an occupational study on the relationship between dust concentration and chronic bronchitis. In the study, $N = 499$ workers of a cement plant in Heidelberg were observed from 1960 to 1977. The response Y is the appearance of chronic bronchitis, and the correctly measured covariates Z are smoking and duration of exposure. The effect of the dust concentration in the individual working area X is of primary interest in the study. This concentration was measured several times in a certain time period and averaged, leading to the surrogate W for the concentration.

Ignoring the ME, we conducted a logistic regression with the response chronic bronchitis and the regressors $\log(1+\text{dust concentration})$, duration (in years), and smoking. The calculations were conducted by XploRe with the following commands:

```
dat = read("heid.dat")
y = dat[,1]
```

```

w = dat[,2]
z = dat[,3]
library("glm")
doglm(w~z y)

```

In interactive modeling, the binomial distribution and the logistic link have to be chosen for the GLM. The output table from XploRe for the logistic model is given in Figure 5.

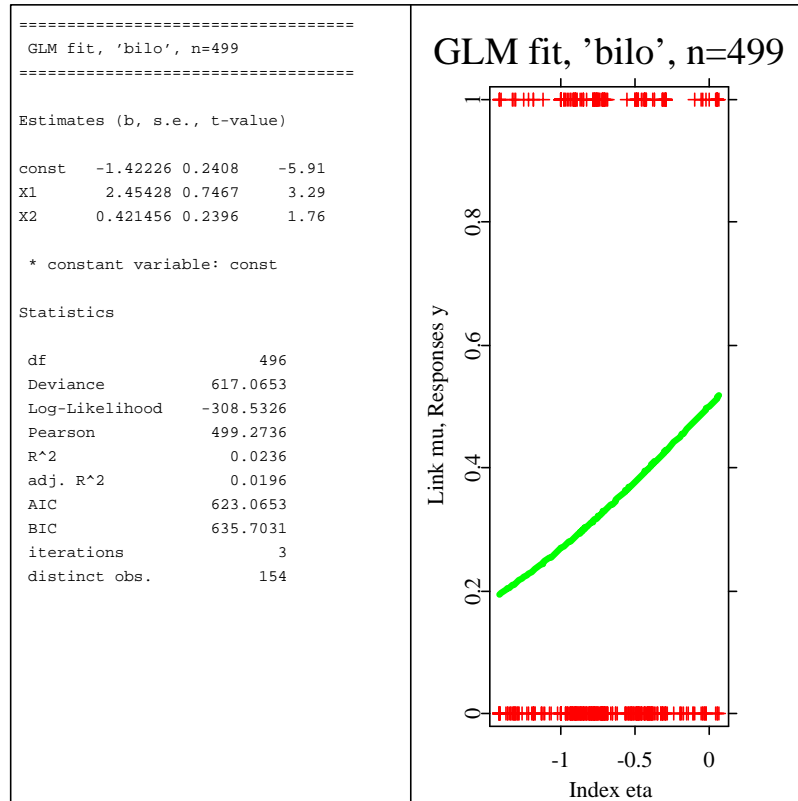



Figure 5: XploRe output display the Heidelberg data

2.1 Regression Calibration

Regression calibration was suggested as a general approach by Carroll and Stefanski (1990) and Gleser (1992). The idea of this method is to replace the unobserved X by its expected value $E(X|W, Z)$ and then to perform a standard EIV analysis, since the latent variable X is approximated by the regression $E(X|W, Z)$. The corresponding XploRe quantlet is called `reca` and is discussed below.

```
res = reca(y, w, z, su2)
```

 `reca-example.xpl`

input

`y`
 $n \times 1$ matrix, the design variables,
`w`
 $n \times 1$ matrix,
`z`
 $n \times 1$ matrix,
`su2`
 scalar, the variance of measurement error.

output

`res.beta`
 vector, the estimate,
`res.bv`
 matrix, the variance of the estimate.

We give an example to explain this code. Let's come back to the Heidelberg data.

```

library("xplore")
library("eiv")
v=read("heid.dat")
y=v[,1]
w=v[,2]
z=v[,3]
su2=var(w)/4
res=reca(y,w,z,su2)
  
```


The estimate of the slope parameter of the dust concentration is 2.9193 with standard error 0.9603, compared to the naive estimates 2.54428 (s.e. 0.8641). Here, the shape of the curve is similar to that obtained by the naive model.

2.2 Simulation Extrapolation

Simulation extrapolation is a complementary approximate method that shares the simplicity of regression calibration and is well suited to problems with additive measurement error. This is a simulation-based method for estimating and reducing bias due to measurement error. The estimates are obtained by adding additional measurement error to the data in a resampling-like stage, establishing a trend of measurement error, and extrapolating this trend back to the case of no measurement error. For a detailed explanation of this method, see Carroll, Ruppert and Stefanski (1995). The quantlet `simex` implements calculation in XploRe. Its syntax is

```

library("eiv")
gest = simex(y,w,z,su2,lam,b)
  
```

 `simex-example.xpl`

where

input

`y`
 $n \times 1$ matrix, the design variables,
`w`
 $n \times 1$ matrix,

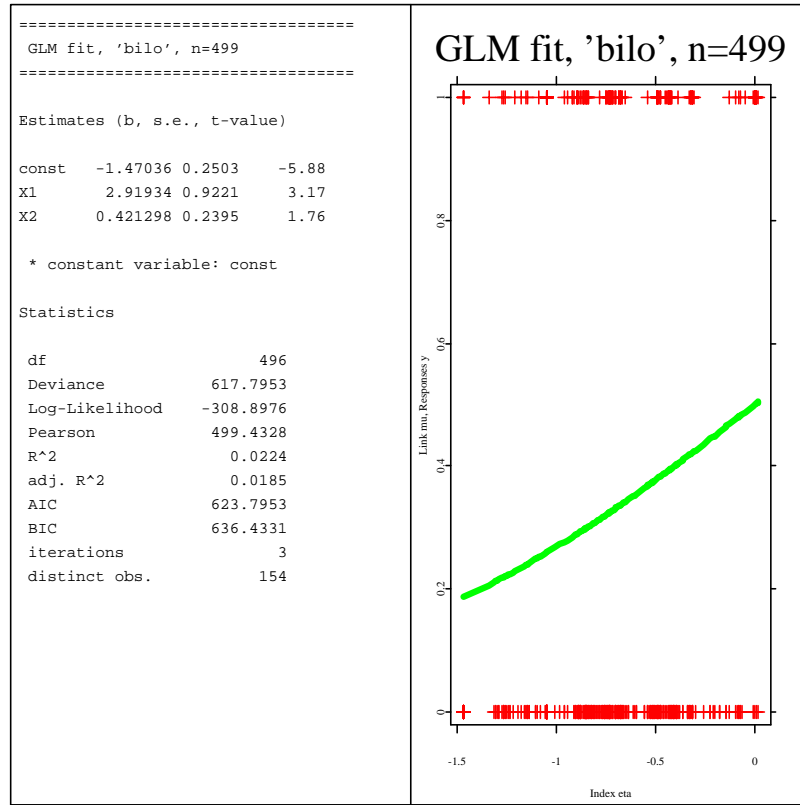


Figure 6: RECA estimation

- z**
 $n \times 1$ matrix,
- su2**
the variance of the measurement error,
- lam**
pseudo-parameter for generating pseudo-errors,
- b**
the number of replication in each simulation.

output

The list variable **gest** contains:

- gest.simexl**
the estimate based on linear extrapolant function
- gest.simexq**
the estimate based on quadratic extrapolant function

Consider the Heidelberg data again. As before, we assume that the ME is normal with variance $\sigma_u^2 = 0.25 * \sigma_z^2$. The results for $\hat{\beta}_{\text{SIMEX}}$ were 2.8109 (linear) and 3.0051 (quadratic).

3 Partially Linear EIV Models

`sf = eivplmnr (w,t,y,sigma,h)`
 computes statistical characteristics for the partially linear EIV models

Partially linear eiv models relate a response Y to predictors (X, T) with mean function $\beta^T X + g(T)$, where the regressors X are measured with additive errors, that is,

$$\begin{aligned} Y &= X^T \beta + g(T) + \varepsilon \\ W &= X + U, \end{aligned} \quad (8)$$

where the variable U is independent of (Y, X, T) with mean zero and $\text{Var}(U) = \Sigma_{uu}$, $E(\varepsilon|X, T) = 0$ and $E(\varepsilon^2|X, T) = \sigma^2(X, T) < \infty$.

Here, we only introduce the conclusions. The related proofs and discussions can be found in Liang, Härdle and Carroll (1997).

3.1 The Variance of Error Known

In EIV linear regression, inconsistency caused by the measurement error can be overcome by applying the so-called correction for attenuation. In our context, this suggests that we use the estimator

$$\hat{\beta}_n = (\widetilde{\mathbf{W}}^T \widetilde{\mathbf{W}} - n \Sigma_{uu})^{-1} \widetilde{\mathbf{W}}^T \widetilde{\mathbf{Y}}. \quad (9)$$

In some cases, we assume that the model errors ε_i are homoscedastic with common variance σ^2 . In this event, since $E\{Y_i - X_i^T \beta - g(T_i)\}^2 = \sigma^2$ and $E\{Y_i - W_i^T \beta - g(T_i)\}^2 = E\{Y_i - X_i^T \beta - g(T_i)\}^2 + \beta^T \Sigma_{uu} \beta$, we define

$$\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n (\widetilde{Y}_i - \widetilde{W}_i^T \hat{\beta}_n)^2 - \hat{\beta}_n^T \Sigma_{uu} \hat{\beta}_n \quad (10)$$

as the estimator of σ^2 .

Theorem 3.1 *Suppose that certain conditions hold and $E(\varepsilon^4 + \|U\|^4) < \infty$. Then $\hat{\beta}_n$ is an asymptotically normal estimator, i.e.,*

$$n^{1/2}(\hat{\beta}_n - \beta) \longrightarrow^{\mathcal{L}} N(0, \Sigma^{-1} \Gamma \Sigma^{-1}),$$

where $\Sigma = E\{X - E(X|T)\}^{\otimes 2}$, $\Gamma = E[(\varepsilon - U^T \beta)\{X - E(X|T)\}]^{\otimes 2} + E\{(UU^T - \Sigma_{uu})\beta\}^{\otimes 2} + E(UU^T \varepsilon^2)$. Note that $\Gamma = E(\varepsilon - U^T \beta)^2 \Sigma + E\{(UU^T - \Sigma_{uu})\beta\}^{\otimes 2} + \Sigma_{uu} \sigma^2$ if ε is homoscedastic and independent of (X, T) , where $A^{\otimes 2} = A \cdot A^T$.

Theorem 3.2 *Under the same conditions as that of Theorem 3.1, if the ε 's are homoscedastic with variance σ^2 , and independent of (X, T) . Then*

$$n^{1/2}(\hat{\sigma}_n^2 - \sigma^2) \longrightarrow^{\mathcal{L}} N(0, \sigma_*^2),$$

where $\sigma_*^2 = E\{(\varepsilon - U^T \beta)^2 - (\beta^T \Sigma_{uu} \beta + \sigma^2)\}^2$.

3.2 The Variance of Error Unknown

The technique of partial replication is adopted when Σ_{uu} is unknown and must be estimated. That is, we observe $W_{ij} = X_i + U_{ij}$, $j = 1, \dots, m_i$.

We consider here only the usual case that $m_i \leq 2$, and assume that a fraction δ of the data has such replicates. Let \bar{W}_i be the sample mean of the replicates. Then a consistent, unbiased method of moments estimate for Σ_{uu} is

$$\hat{\Sigma}_{uu} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} (W_{ij} - \bar{W}_i)^{\otimes 2}}{\sum_{i=1}^n (m_i - 1)}.$$

The estimator changes only slightly to accommodate the replicates, becoming

$$\begin{aligned} \hat{\beta}_n &= \left[\sum_{i=1}^n \{ \bar{W}_i - \hat{g}_{w,h}(T_i) \}^{\otimes 2} - n(1 - \delta/2) \hat{\Sigma}_{uu} \right]^{-1} \\ &\quad \times \sum_{i=1}^n \{ \bar{W}_i - \hat{g}_{w,h}(T_i) \} \{ Y_i - \hat{g}_{y,h}(T_i) \}, \end{aligned} \quad (11)$$

where $\hat{g}_{w,h}(\cdot)$ is the kernel regression of the \bar{W}_i 's on T_i .

The limit distribution of (11) is $N(0, \Sigma^{-1} \Gamma_2 \Sigma^{-1})$, with

$$\begin{aligned} \Gamma_2 &= (1 - \delta) E \left[(\varepsilon - U^T \beta) \{ X - E(X|T) \} \right]^{\otimes 2} \\ &\quad + \delta E \left[(\varepsilon - \bar{U}^T \beta) \{ X - E(X|T) \} \right]^{\otimes 2} \\ &\quad + (1 - \delta) E \left(\left[\{ UU^T - (1 - \delta/2) \Sigma_{uu} \} \beta \right]^{\otimes 2} + UU^T \varepsilon^2 \right) \\ &\quad + \delta E \left(\left[\{ \bar{U} \bar{U}^T - (1 - \delta/2) \Sigma_{uu} \} \beta \right]^{\otimes 2} + \bar{U} \bar{U}^T \varepsilon^2 \right). \end{aligned} \quad (12)$$

In (12), \bar{U} refers to the mean of two U 's. In the case that ε is independent of (X, T) , the sum of the first two terms simplifies to $\{\sigma^2 + \beta^T (1 - \delta/2) \Sigma_{uu} \beta\} \Sigma$.

3.3 XploRe Calculation and Practical Data


The quantlet `eivplmnor` estimates the parameters of partially linear `eiv` model, with the assumption that the conditional distribution of Y given X and T is normally distributed. Its principle is similar to the quantlet `gplmnoid` in the `GPLM` quantlib. We show the following example:

```
library("xplore")
library("eiv")
n = 100
randomize(n)
sigma = 0.0081
b = 1|2
p = rows(b)
x = 2.*uniform(n,p)-1 ; latent variable
t = sort(2.*uniform(n)-1,1) ; observable variable
w = x+sqrt(sigma)*uniform(n) ; manifest variable
m = 0.5*cos(pi.*t)+0.5*t
y = x*b+m+normal(n)./2
h=0.5
sf = eivplmnor(w,t,y,sigma,h)
b~sf.b ; estimates of b and g(t)
```

```

dds = createdisplay(1,1)
datah1=t~m
datah2=t~sf.m
part=grid(1,1,rows(t))'
setmaskp(datah1,1,0,1)
setmaskp(datah2,4,0,3)
setmaskl(datah1,part,1,1,1)
setmaskl(datah2,part,4,1,3)
show(dds,1,1,datah1,datah2)

```

 eivplmnor-example.xpl

A partially linear fit for $E(y|x, t)$ is computed. sf.b contains the coefficients for the linear part. sf.m contains the estimated nonparametric part evaluated at observations t , see Figure 7. There the thin curve line represents true data and the thick one does the nonparametric estimates.

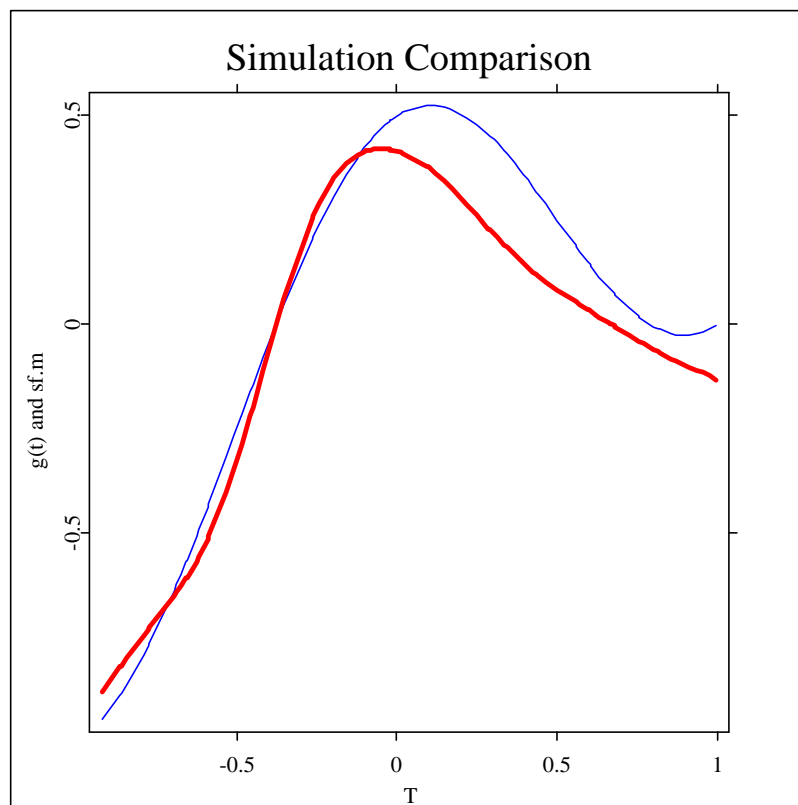


Figure 7: Output display for partially linear EIV example

We now use the quantlet `eivplmnor` to calculate practical data from the Framingham Heart Study. In this data set, the response variable Y is the average blood pressure in a fixed 2-year period, T , the age and W , the logarithm of the observed cholesterol level, for which there are two replicates.

For the purpose of illustration, we only use the first cholesterol measurement. The measurement error variance is obtained in the previous analysis. The estimate of β is 9.438 with the standard error 0.187. For nonparametric fitting, we choose

the bandwidth using cross-validation to predict the response. Precisely we compute the squared error using a geometric sequence of 191 bandwidths ranging in $[1, 20]$. The optimal bandwidth is selected to minimize the square error among these 191 candidates. An analysis ignoring measurement error found some curvature in T , see Figure 8 for the estimate of $g(T)$.

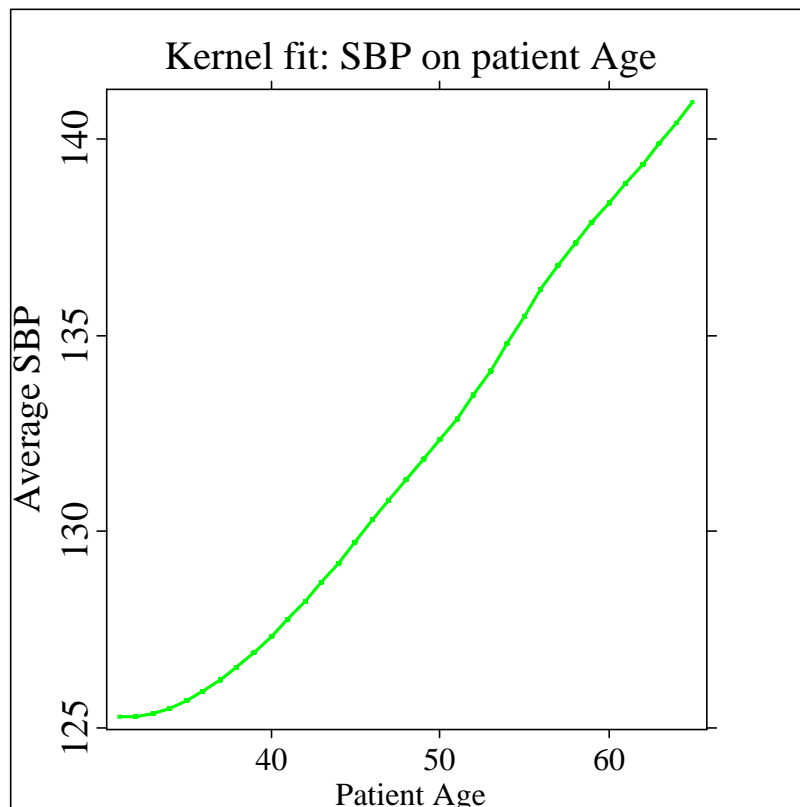


Figure 8: Framingham data study

References

- Carroll, R.J. and Stefanski, L.A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors, *Journal of the American Statistical Association*, **85**: 652-663.
- Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Nonlinear Measurement Error Models*, Vol. 63 of Monographs on Statistics and Applied Probability, Chapman and Hall, New York.
- Fuller, W.A. (1987). *Measurement Error Models*, Wiley and Sons, New York.
- Gleser, L.J. (1992). The importance of assessing measurement reliability in multivariate regression, *Journal of the American Statistical Association*, **87**: 696-707.
- Härdle, W., Liang, H. and Gao, J. T. (2000). *Partially Linear Models*. Physica-Verlag, Heidelberg.

Liang, H., Härdle, W. and Carroll, R. (1997). *Large sample theory in a semiparametric partially linear errors-in-variables model*, SFB DP No. 27. Humboldt University of Berlin.