

A test for the slope in the functional measurement error model

KAI BRUCHLOS

ABSTRACT. There are two measurement error models in linear regression, the structural and the functional. Theoretical investigations and applications are concentrated on the structural model with jointly normally and identically distributed observations because there is no test for the slope in the functional model so far. This gap will be closed here for the model with one independent variable. Furthermore it is stated that the functional model is a natural extension of the classical linear regression model with one independent variable if there are errors of measurement in both variables. Moreover it is not postulated in the functional model that the expectations are equal. So the functional model is more realistic than the structural.

1. Introduction

In the classical linear regression model

$$\tilde{Y}_i = \alpha + \beta \tilde{x}_i + \varepsilon_i$$

it is assumed that the explanatory variables \tilde{x}_i have fixed values in repeated sampling ([6], p. 19; [10], p. 194). This assumption is valid e.g. for time series but it does not apply to a variety of other cases, especially if one considers the linear relationship of two economic variables like consumption and income. In the latter one uses the measurement error model which regards the explained variable and the explanatory variable as measured with errors i.e. stochastic.

Received September 2, 2013.

2010 *Mathematics Subject Classification.* 62J05, 62F03.

Key words and phrases. Error in variables, functional measurement error, linear regression, slope, test.

<http://dx.doi.org/10.12697/ACUTM.2015.19.08>

The theory of measurement error models features two types of models, the functional model

$$\begin{aligned} Y_i &= \alpha + \beta x_i + \varepsilon_i, \quad V_i = x_i + \delta_i, \quad i = 1, \dots, n, \\ x_1, \dots, x_n &\text{ are fixed values,} \\ \varepsilon_1, \dots, \varepsilon_n, \delta_1, \dots, \delta_n &\text{ are independent random variables,} \\ \varepsilon_i &\sim N(0, \sigma_\varepsilon^2), \delta_i \sim N(0, \sigma_\delta^2), \end{aligned}$$

and the structural model

$$\begin{aligned} \check{Y}_i &= \alpha + \beta \check{X}_i + \varepsilon_i, \quad \check{V}_i = \check{X}_i + \delta_i, \quad i = 1, \dots, n, \\ \check{X}_1, \dots, \check{X}_n, \varepsilon_1, \dots, \varepsilon_n, \delta_1, \dots, \delta_n &\text{ are independent random variables,} \\ \check{X}_i &\sim N(\mu, \sigma^2), \varepsilon_i \sim N(0, \sigma_\varepsilon^2), \delta_i \sim N(0, \sigma_\delta^2), \end{aligned}$$

where Y_i, \check{Y}_i correspond with \check{Y}_i , and V_i, \check{V}_i with \tilde{x}_i added to the error.

The crucial difference between the functional and the structural model is that in the structural model the random variables \check{V}_i are i.i.d. which is not the case in the functional model. In the functional model, each random variable V_i has its own expected value, namely x_i . This corresponds to the classical linear regression model.

One aim of this proposal is to formulate a model in which the classical regression model is a special case of the functional model. Why not of the structural model? In the structural model all explained variables \check{Y}_i have the same expected value $\mathbb{E}(\check{Y}_i) = \alpha + \beta\mu$ while in the classical model the explained variables \check{Y}_i have different expected values $\mathbb{E}(\check{Y}_i) = \alpha + \beta\tilde{x}_i$. So because of pure mathematical reasons the classical regression model cannot be a special case of the structural model! Additionally the structural model condition of \check{V}_i being i.i.d. is not that close to reality. Examples for which this model can be used are rather specific (*cp.* [7], p. 34; [9], p. 198).

In the structural model there is a test for the slope β ([7], p. 45), in the functional model we only have asymptotic tests ([8], p. 412 *et sqq.*). This paper proposes a test for the slope in the functional model. This test is the counterpart to the test in the structural model.

The test for the slope cannot be transferred exactly from the structural to the functional model due to a special correlation – which is calculated for the test – being zero in the functional model and therefore is not an appropriate test statistic.

The model will work if the following conditions are satisfied.

- (1) The classical linear regression model is designed with conditional expectations, whereas a two-dimensional random variable (X, Y) is implied (a standard mathematical view, *cp.* [5]).
- (2) The functional measurement error model is defined as an extension of the classical linear model.

- (3) The random variables for the correlation are constructed at the level of (X, Y) .

The values x_1, \dots, x_n are realisations of X because otherwise the regression model cannot be formulated with conditional expectations. This approach presumes the special feature of the functional model, the x_i being non-random.

In the following we have two levels: the level of the regression model and the level of (X, Y) . After defining the functional model we construct a similar model for (X, Y) . In order to distinguish between the two levels we introduce the notation $|_{x_i}$ for the regression model which is only a special kind to number serially.

The aim of the test is to construct two random variables L, M so that the null hypothesis $H_0^{(1)} : \beta = \beta_0$ is equivalent to the null hypothesis $H_0^{(2)} : \varrho_{L,M} = 0$ where $\varrho_{L,M}$ is the correlation coefficient of L and M . The null hypothesis $H_0^{(2)}$ is checked with a t-test for zero correlation. For the calculation of the test statistic it is necessary that the ratio of error variances is known.

A link between the correlation analysis and the regression analysis in the context of random variables (Theorem 1) is essential for the test of the slope β . Consequently a sample can be used for the test statistic as well as for the estimator of β .

It is necessary that (X, Y) is normally distributed. This property follows from the symmetry of the functional model (Proposition 2): if the conditional distribution of Y_i given $X = x_i$ is a normal distribution it is necessary that the conditional distribution of V_i given $Y = y_i$ is a normal distribution as well so that one can change the variables:

$$V_i = \alpha' + \beta' y_i + \delta_i, \quad Y_i = y_i + \varepsilon_i.$$

The test for the slope in the functional model can be applied to average claim in non-life insurance ([2]).

2. Regression model

We begin with the environment of the classical linear regression model (cp. [5]).

Let (X, Y) be a two-dimensional real random variable with continuous density $f(x, y)$, marginal densities f_X, f_Y , expectations $\mathbb{E}(X), \mathbb{E}(Y)$, variances $\sigma_X^2 := \text{Var}(X)$, $\sigma_Y^2 := \text{Var}(Y) > 0$, with the correlation coefficient $\varrho_{X,Y}$ of X and Y , and the conditional expectation $\mathbb{E}(Y|X = x)$.

Let the conditional distribution of Y given $X = x$ be a normal distribution, more precisely, the distribution $N(\alpha + \beta x, \sigma_\varepsilon^2)$ with $\alpha, \beta \in \mathbb{R}$,

$\beta \neq 0, \sigma_\varepsilon^2 > 0$. In particular,

$$\mathbb{E}(Y|X = x) = \alpha + \beta x .$$

Let x_1, \dots, x_n be realisations of X , and $Y_{|x_1}, \dots, Y_{|x_n}$ statistically independent random variables, which have as distribution the conditional distribution of Y given $X = x_i$, thus $Y_{|x_i} \sim N(\alpha + \beta x_i, \sigma_\varepsilon^2)$. These are the explained variables. Still missing for the functional model are the measurement errors: first of all let $E_{|x_i} := Y_{|x_i} - (\alpha + \beta x_i)$ for $i = 1, \dots, n$.

Lemma 1. (i) *There holds $E_{|x_i} \sim N(0, \sigma_\varepsilon^2)$ for $i = 1, \dots, n$.*

(ii) *$E_{|x_1}, \dots, E_{|x_n}$ are statistically independent.*

(iii) *Let $\varepsilon_1 := E_{|x_1}, \dots, \varepsilon_n := E_{|x_n}$. If the x_i have fixed values in repeated sampling then $\tilde{Y}_i = \alpha + \beta x_i + \varepsilon_i$ is the classical linear regression model (cp. [6], p. 93; [10], p. 194).*

Remark 1. $Y_{|x_i}$ takes the place of the explained variable.

Next let $D_{|x_1}, \dots, D_{|x_n}$ be statistically independent random variables with the properties that $D_{|x_1}, \dots, D_{|x_n}, E_{|x_1}, \dots, E_{|x_n}$ are statistically independent and $D_{|x_i} \sim N(0, \sigma_\delta^2)$ with $\sigma_\delta^2 := (1 - \varrho_{X,Y}^2)\sigma_X^2$ for $i = 1, \dots, n$.

Remark 2. The definition $\sigma_\delta^2 := (1 - \varrho_{X,Y}^2)\sigma_X^2$ guarantees the symmetry of the functional model – take notice of Lemma 3. The symmetric form is

$$V_i = \alpha' + \beta' y_i + \delta_i, \quad Y_i = y_i + \varepsilon_i$$

with y_1, \dots, y_n fixed values, $\alpha' := -\alpha/\beta$, $\beta' := 1/\beta$ and $\delta_1 := D_{|x_1}, \dots, \delta_n := D_{|x_n}$. The pairs of true values are $(x_1, y_1), \dots, (x_n, y_n)$ which are not observed directly. The observed sample is

$$(v_1, w_1), \dots, (v_n, w_n) := (x_1 + \delta_1(x_1), y_1 + \varepsilon_1(y_1)), \dots, (x_n + \delta_n(x_n), y_n + \varepsilon_n(y_n)).$$

The particular value for σ_δ^2 is not necessary for the following.

Proposition 1. *The random variable $(D_{|x_i}, E_{|x_i})$ is normally distributed for $i = 1, \dots, n$.*

The measurement errors for the explained variable and the explanatory variable have the same properties as in the classical model: they are normally distributed with expected value zero, and the measurement error of one special measuring has no influence on any other.

To complete the functional model we still need the random variables for the explanatory variable.

Lemma 2. *Let $V_i := V_{|x_i} := D_{|x_i} + x_i$ and $Y_i := \alpha + \beta x_i + \varepsilon_i$ for $i = 1, \dots, n$.*

(i) *There holds $V_{|x_i} \sim N(x_i, \sigma_\delta^2)$ and*

$$\mathbb{E}(Y_i) = \mathbb{E}(Y_{|x_i}) = \alpha + \beta \mathbb{E}(V_{|x_i}) = \alpha + \beta \mathbb{E}(V_i), \quad i = 1, \dots, n.$$

- (ii) With $x_i, \delta_i, \varepsilon_i, V_i,$ and Y_i we have the functional measurement error model ([8], p. 407; [3], p. 25).
- (iii) If we set $\delta_i \equiv 0$ for $i = 1, \dots, n$ and the x_i have fixed values in repeated sampling, then we get again the classical linear regression model.

Remark 3. (i) v_i is a realisation of $V_i = V_{|x_i}, w_i$ is a realisation of Y_i and $Y_{|x_i}$ respectively.

(ii) The x_1, \dots, x_n are only realisations of X . The x_1, \dots, x_n have nothing to do with $\check{X}_1, \dots, \check{X}_n$. And the random variable X has nothing to do with the random variables $\check{X}_1, \dots, \check{X}_n$. This is a consequence of the attribute $\mathbb{E}(\check{Y}_i) = \alpha + \beta\mu$ of the structural model.

Now a crucial step follows. Theorem 1 carries the linearity of the expectations from the level $Y_{|x_i}, V_{|x_i}$ to the level X, Y .

Theorem 1. *There holds $\mathbb{E}(Y) = \alpha + \beta\mathbb{E}(X)$.*

Proof. From the equation ([5], p. 84, (3.6.23'))

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} f_X(x)\mathbb{E}(Y|X = x)dx$$

we have

$$\begin{aligned} \mathbb{E}(Y) &= \int_{-\infty}^{\infty} f_X(x) \cdot (\alpha + \beta \cdot x) dx \\ &= \int_{-\infty}^{\infty} f_X(x) \cdot \alpha + f_X(x) \cdot \beta \cdot x dx \\ &= \alpha \cdot \int_{-\infty}^{\infty} f_X(x) dx + \beta \cdot \int_{-\infty}^{\infty} f_X(x) \cdot x dx \\ &= \alpha + \beta \cdot \mathbb{E}(X) . \end{aligned}$$

□

Now let (X, Y) be normally distributed, which is a consequence of the symmetry of the functional model ([1], p. 401, (III)).

Proposition 2. *If the conditional distribution of X given $Y = y$ is a normal distribution, then (X, Y) is normally distributed.*

With the normality assumption we can calculate the expectation and the variance of $Y_{|x_i}$, which have the usual form (cp. [5]).

Lemma 3. *There hold*

$$\mathbb{E}(Y_{|x_i}) = \mathbb{E}(Y) + \frac{\sigma_Y^2}{\sigma_X^2} \varrho_{X,Y}(x_i - \mathbb{E}(X))$$

and

$$\text{Var}(Y_{|x_i}) = (1 - \varrho_{X,Y}^2) \sigma_Y^2 .$$

In the same way as the measurement errors have been set on the level $x_i, Y_{|x_i}$, they now will be introduced on the level (X, Y) .

Let E be a random variable with the property $E \sim N(0, \sigma_\varepsilon^2)$ and $\mathcal{W} := \mathbb{E}(Y) + E$, thus we have by Theorem 1

$$\mathcal{W} \sim N(\alpha + \beta \mathbb{E}(X), \sigma_\varepsilon^2) .$$

Let D be a random variable with the properties that D and E are statistically independent and $D \sim N(0, \sigma_\delta^2)$. Let $\mathcal{V} := \mathbb{E}(X) + D$, thus $\mathcal{V} \sim N(\mathbb{E}(X), \sigma_\delta^2)$. So we have

$$\mathbb{E}(\mathcal{W}) = \alpha + \beta \mathbb{E}(\mathcal{V}) .$$

3. Test for the slope

Our aim now is to test the slope in the functional model, namely

$$H_0^{(1)} : \beta = \beta_0 \quad \text{versus} \quad H_1^{(1)} : \beta \neq \beta_0 .$$

There is no test yet for these hypotheses, so we consult a t-test for zero correlation: supposed the random variables L and M are bivariate normally distributed the corresponding hypotheses are

$$H_0^{(2)} : \varrho_{L,M} = 0 \quad \text{versus} \quad H_1^{(2)} : \varrho_{L,M} \neq 0 .$$

In the following we will prove that the null hypothesis $H_0^{(1)}$ is equivalent to the null hypothesis $H_0^{(2)}$ if one defines L and M suitably. For that purpose we pull up $Y_{|x_i}, V_{|x_i}$ to the level (X, Y) .

Let D_1, \dots, D_n be statistically independent sample random variables of D , E_1, \dots, E_n statistically independent sample random variables of E , $\mathcal{V}_i := \mathbb{E}(X) + D_i$ and $\mathcal{W}_i := \mathbb{E}(Y) + E_i$ for $i = 1, \dots, n$. Let $L := E - \beta_0 D$, $M := a_1 \mathcal{W} + a_2 \mathcal{V}$ with $a_1, a_2 \in \mathbb{R} \setminus \{0\}$.

A normal distribution of (D, E) and the validity of the equation

$$\text{Cov}(L, M) = \mathbb{E} [a_1 E^2 - a_2 \beta_0 D^2]$$

lead to the following result.

Proposition 3. *Let $\lambda := \sigma_\varepsilon^2 / \sigma_\delta^2$.*

- (i) $\text{Cov}(L, M) = (a_1 \lambda - a_2 \beta_0) \sigma_\delta^2$.
- (ii) *Let $f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, $(a_1, \beta_0) \mapsto a_1 \lambda - \beta_0$, be a map. If $(\zeta^{-1} \eta, \eta)$ is a zero of f , then $(-\eta^{-1}, -\zeta \eta^{-1})$ is also a zero of f .*
- (iii) (L, M) is normally distributed.

Proposition 3 (iii) supplies the assumption for the t-test for zero correlation.

Now we formulate the basic statement.

Theorem 2. *For $\beta \geq 0$, $a_1 := \lambda^{-1}\beta_0$, and $a_2 := 1$ the statement “ $\beta = \beta_0$ ” is equivalent to “ $\varrho_{L,M} = 0$ ”, so $H_0^{(1)} \iff H_0^{(2)}$.*

Proof. By Proposition 3 (i) $\varrho_{L,M} = 0$ follows from $\beta = \beta_0$.

$\text{Cov}(L, M) = 0$ follows from $\varrho_{L,M} = 0$ because of $\sigma_\delta^2, \sigma_\varepsilon^2 > 0$. This leads to the following two possible conclusions.

1. $\beta = \beta_0$ by Proposition 3 (i).
2. $\beta = -\lambda\beta_0^{-1}$ follows from Proposition 3 (ii) with $a_1 = \lambda^{-1}(-\lambda\beta_0^{-1})$, which is impossible because of the assumption $\beta \geq 0$ (cp. [7], p. 45). \square

Theorem 2 is also valid with $\beta \leq 0$. So for the test we only need to know the sign of β .

To reject $H_0^{(2)}$ with

$$\left| \frac{R(n-2)^{1/2}}{(1-R^2)^{1/2}} \right| > t_{n-2; 1-\alpha/2}$$

for a level of significance α we finally have to calculate the sample correlation coefficient R .

Proposition 4. *Suppose that λ is known. Let $a_1 := \lambda^{-1}\beta_0$ and $a_2 := 1$. Then*

$$R = \left(\sum_{i=1}^n \mathcal{A}_i \mathcal{B}_i \right) \left(\sum_{i=1}^n \mathcal{A}_i^2 \sum_{i=1}^n \mathcal{B}_i^2 \right)^{-1/2}$$

with

$$\mathcal{A}_i := \mathcal{W}_i - \left(\frac{1}{n} \sum_{j=1}^n \mathcal{W}_j \right) - \beta_0 \left(\mathcal{V}_i - \frac{1}{n} \sum_{j=1}^n \mathcal{V}_j \right)$$

and

$$\mathcal{B}_i := \lambda^{-1}\beta_0 \mathcal{W}_i + \mathcal{V}_i - \frac{1}{n} \sum_{j=1}^n (\lambda^{-1}\beta_0 \mathcal{W}_j + \mathcal{V}_j).$$

Proof. The result follows from

$$\begin{aligned}
 L_i - \bar{L} &= \alpha + \beta_0 \mathbb{E}(X) + E_i - \frac{1}{n} \sum_{j=1}^n (\alpha + \beta_0 \mathbb{E}(X) + E_j) \\
 &\quad - \beta_0 (\mathbb{E}(X) + D_i) + \frac{\beta_0}{n} \sum_{j=1}^n (\mathbb{E}(X) + D_j) \\
 &= \mathcal{W}_i - \left(\frac{1}{n} \sum_{j=1}^n \mathcal{W}_j - \beta_0 (\mathcal{V}_i - \frac{1}{n} \sum_{j=1}^n \mathcal{V}_j) \right).
 \end{aligned}$$

□

Remark 4. Because of $\mathcal{V}_i \sim N(\mathbb{E}(X), \sigma_\delta^2)$ and $\mathcal{W}_i \sim N(\mathbb{E}(Y), \sigma_\varepsilon^2)$ we consider the observed value v_i as a realisation of \mathcal{V}_i and the observed value w_i as a realisation of \mathcal{W}_i .

4. Example

Let us consider mean income and mean consumption of German private households in Euro ([4], table 6.6.1):

Year	2004	2005	2006	2007	2009	2010	2011	2012
Income	3368	3496	3489	3584	3711	3758	3871	3989
Consumption	1989	1996	2089	2067	2156	2168	2252	2310

We have the observed sample

$$(v_1, w_1) = (3368, 1989), \dots, (v_8, w_8) = (3989, 2310).$$

Figure 1 shows the linearity of the data:

We can assume that the measurement error of income is the same as the measurement error of consumption. So we have $\lambda = 1$. Let $\alpha = 0.01$, $r := R((v_1, w_1), \dots, (v_8, w_8))$ and

$$t := \left| \frac{\sqrt{6}r}{\sqrt{1-r^2}} \right|.$$

We get the following results:

β_0	t	$t_{8;0.995}$
2	10.83	3.71
1	6.36	3.71
0.5	0.65	3.71

So the mean consumption grows less than the mean income of German private households.

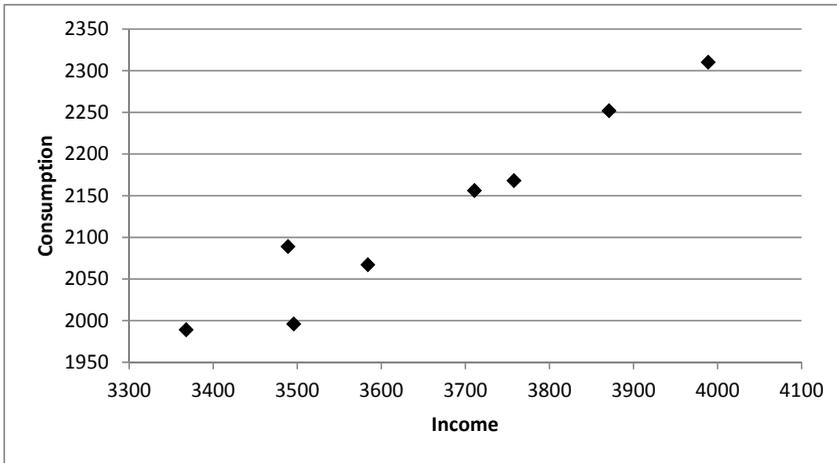


FIGURE 1. Dataset of private household income and consumption

Acknowledgements

The author is grateful to the referee for many valuable comments and suggestions.

References

- [1] A. Bhattacharyya, *On some sets of sufficient conditions leading to the normal bivariate distribution*, *Sankhyā* **6** (1943), 399–406.
- [2] K. Bruchlos, *Estimator and test for the schadenparameter in health insurance tariffs with cost sharing*, *Schweiz. Aktuarver. Mitt.* **1** (2005), 11–28. (German)
- [3] R. J. Carroll, A. Ruppert, L. A. Stefanski, and C. Crainiceanu, *Measurement Error in Nonlinear Models*, 2nd ed., Chapman&Hall, London, 2006.
- [4] *Federal Statistical Office of Germany*, *Statistical Yearbook 2014*, Wiesbaden, 2014. (German)
- [5] M. Fisz, *Probability Theory and Mathematical Statistics*, 3rd ed., Wiley, New York, 1976.
- [6] D. N. Gujarati, *Basic Econometrics*, 2nd ed., McGraw-Hill, New York, 1988.
- [7] W. A. Fuller, *Measurement Error Models*, Wiley, New York, 1987.
- [8] M. Kendall and A. Stuart, *The Advanced Theory of Statistics. Volume 2: Inference and Relationship*, 4th ed., Griffin, London, 1979.
- [9] A. Madansky, *The fitting of straight lines when both variables are subject to error*, *J. Amer. Statist. Assoc.*, **54** (1959), 173–205.
- [10] W. R. Pestman, *Mathematical Statistics*, de Gruyter, Berlin, 1998.

DEPARTMENT OF MATHEMATICS AND INFORMATICS, UNIVERSITY OF APPLIED SCIENCE
MITTELHESSEN, WILHELM-LEUSCHNER-STR. 13, 61169 FRIEDBERG, GERMANY
E-mail address: kai.bruchlos@mnd.thm.de